PERFORMATIVE MULTI-OBJECTIVE OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Performative prediction studies learning under model-induced distribution shift, where the deployed model shapes the very data it aims to predict. Existing theory and algorithms largely focus on a single objective or collapse multiple criteria into one aggregated loss, while practical systems may involve several potentially conflicting objectives such as accuracy, fairness, robustness, and welfare. We introduce performative multi-objective optimization (PMOO), a framework for jointly optimizing several performative objectives, and define Performative Pareto Stationary (PPS) points as solutions satisfying the first-order Pareto stationarity condition under the environment induced by the deployed model. A central challenge in PMOO is the coupling between *environmental drift*, induced by the deployed model, and *multiplier drift*, reflecting evolving trade-offs among objectives. Deployment affects the gradient geometry used to select the multiplier weights, while these weights shape model updates and subsequent environments. For smooth nonconvex objectives, we analyze multi-gradient descent (MGD) and show that its best iterate attains approximate PPS. To stabilize the multiplier dynamics, we propose proximal MGD (PMGD). In the strongly convex setting, we further develop a coupled convergence analysis that makes explicit how performative sensitivity and multiplier geometry influence the dynamics. Together, these results extend performative prediction to a Pareto framework for learning with conflicting objectives.

1 INTRODUCTION

Machine learning systems are often deployed in environments that react to them. A credit-scoring model alters applicants' strategic behavior (Zhang et al., 2022; Xie & Zhang, 2024); a recommendation system reshapes user participation and retention (Zhang et al., 2019; Jin et al., 2026); a hiring rule incentivizes applicants to invest in interview preparation (Zhang et al., 2020; Jin et al., 2022; Xie et al., 2024). In each case, the data the system later sees is shaped by the system itself. Performative prediction (PP) formalizes optimization in this feedback loop by treating the target data distribution $D(\theta)$ as a function of the deployed model $\theta \in \Theta$ (Perdomo et al., 2020; Brown et al., 2022; Hardt & Mendler-D"unner, 2025).

Within this framework, the ideal target is the **performative optimal** (PO) point

$$\theta_{\mathrm{PO}} = \arg\min_{\theta\in\Theta} V(\theta), \quad V(\theta) := \mathbb{E}_{Z\sim D(\theta)}[\ell(\theta; Z)],$$

the model that minimizes the expected loss on the very distribution it induces. Because the distribution being averaged over is itself a function of the variable being optimized, finding θ_{PO} generally requires differentiating through the map $\theta \to D(\theta)$, which is rarely tractable. A more accessible solution concept is therefore the **performative stable** (PS) point

$$\theta_{\mathrm{PS}} = \arg\min_{\theta\in\Theta} J(\theta; \theta_{\mathrm{PS}}), \quad J(\tilde{\theta}; \theta) := \mathbb{E}_{Z\sim D(\theta)}[\ell(\tilde{\theta}; Z)],$$

defined through the *decoupled performative risk* $J(\tilde{\theta}; \theta)$, in which a candidate model $\tilde{\theta}$ is scored on data drawn from the environment fixed by a separate deployed model θ. The decoupled view is what makes PS algorithmically tractable. Differentiating $V(\theta) = J(\theta; \theta)$ requires, by the chain rule, a derivative through $D(\theta)$, which is rarely available. The decoupled form sidesteps this: one freezes the current environment and uses only the partial gradient with respect to the candidate,

$$g(\theta) := \nabla_1 J(\theta; \theta), \tag{1}$$

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

obtained by deploying θ, sampling from $D(\theta)$, and differentiating the empirical loss as in standard ERM. Approaches such as repeated risk minimization (RRM) and repeated gradient descent (RGD) build on this *decoupled gradient*: at each round, θ_t minimizes (or take a gradient step on) $J(\cdot; \theta_{t-1})$, and under certain conditions these iterates converge to θ_{PS} (Mendler-D"unner et al., 2020; Perdomo et al., 2020; Li & Wai, 2024; Khorsandi et al., 2025; Xie et al., 2025).

The deployment settings often involve several objectives, such as predictive risks across population groups, fairness, and participation(Peet-Pare et al., 2023; Jin et al., 2026; Jia et al., 2024). For example, a credit model may be evaluated through group-specific risks while different groups respond differently to its deployment. Multi-objective optimization (MOO) (Ge et al., 2022; Fliege & Svaiter, 2000) provides methods for combining objective gradients adaptively. However, standard static-MOO analyses do not directly cover the additional distribution changes induced by deployment. This motivates studying the stability of multi-objective learning within a performative feedback loop.

The difficulty lies in the coupling between *performative drift*, induced by the deployed model, and *multiplier drift*, relecting evolving trade-offs across objectives. Specifically, deployment changes the environment in which objective gradients are evaluated, thereby changing the gradient geometry used to select the combination multipliers. Moreover, these multipliers determine the next model update and hence the next environment. Adaptive multipliers already occur in static MOO (Fliege et al., 2019); performative feedback embeds their evolution within an additional deployment loop. The resulting interaction can amplify perturbations, so contraction of a fixed-weight model update alone does not establish stability of the adaptive dynamics.

We develop a theoretical framework for performative multi-objective optimization (PMOO). Our target is a **performative Pareto stationary** (PPS) point, defined by the absence of a common first-order descent direction when the model's induced environment is held fixed. We analyze repeated multi-gradient descent (MGD) and introduce proximal MGD (PMGD) to stabilize the multiplier dynamics, which penalizes multiplier changes across iterations. Our analysis connects environmental sensitivity and multiplier response to stationarity and convergence. Additional related work is discussed in Appendix A. Our main contributions are:

- We formulate PMOO for learning with multiple objectives in model-dependent environments and define PPS as the frozen-environment first-order stationarity concept for Pareto compromise under performative feedback (Sec. 2).
- For smooth nonconvex objectives, we establish a best-iterate approximate-PPS bound for MGD, separating optimization progress from a performative-sensitivity term (Sec. 3).
- We analyze the coupling between model and multiplier updates identifying *environment drift* and *multiplier drift*, introduce temporal proximal regularization through PMGD, and derive sufficient convergence conditions under strong convexity (Sec. 4).
- Through synthetic examples and a credit-risk task with strategic responses, we examine model convergence, multiplier behavior, and sensitivity to deployment feedback (Sec. 5).

2 Problem Setup

Let $\Theta \subseteq \mathbb{R}^d$ be a closed compact parameter space and let $m \geq 2$ be the number of objectives. Recall from Sec. 1 that the *post-deployment environment* $D(\theta)$ is the data distribution induced by deploying model $\theta \in \Theta$. We extend the scalar decoupled performative risk to a vector of m objectives. For each $i \in [m]$, let $J_i(\tilde{\theta}; \theta) = \mathbb{E}_{Z \sim D(\theta)}[\ell_i(\tilde{\theta}; Z)]$ denote *decoupled performative objective* that scores a candidate model $\tilde{\theta}$ under the environment induced by a deployed model θ, and define

$$V_i(\theta) := J_i(\theta; \theta).$$

Beyond the standard interpretation as a *prediction loss* (Perdomo et al., 2020; Izzo et al., 2021; Perdomo, 2023), ℓ_i in our setting can target a wide range of objectives, including *group fairness gap* (differences across protected groups under $D(\theta)$) (Hardt et al., 2016b; Franzese et al., 2025; Agarwal et al., 2018), *participation rates* (the fraction of users who continue to engage with the system given θ) (Raab et al., 2024), and *social welfare scores* (aggregate utility over the induced population) (Pardeshi et al., 2024; Hu & Chen, 2020).

Performative Multi-Objective Optimization (PMOO). Stacking the objectives gives the vector optimization problem

$$\min_{\theta\in\Theta} V(\theta) := \big(V_1(\theta),\dots,V_m(\theta)\big).$$

We study this setting through a frozen-environment stationarity notion based on decoupled gradients. As in scalar PP, the full gradient of V_i is unavailable. It would require differentiating through the distribution map $\theta \to D(\theta)$, which we do *not* assume access to. We thus work with the per-objective decoupled gradients

$$g_i(\theta) := \nabla_1 J_i(\theta;\theta), \quad G(\theta) := [g_1(\theta),\dots,g_m(\theta)], \quad H(\theta) := G(\theta)^\top G(\theta),$$

the multi-objective analogue of the decoupled gradient (Eq. equation 1) used in repeated retraining. To combine these m gradients into a single update, we follow the multiple-gradient descent algorithm (MGDA) (Désidéri, 2012) and define the *multi-gradient descent direction map*

$$d(\theta) \in \arg\min_{d\in\mathbb{R}^d} \Big\{ \max_{i\in[m]} \langle g_i(\theta), d\rangle + \frac{1}{2}\|d\|^2 \Big\}. \tag{2}$$

$d(\theta)$ is the steepest common first-order descent direction under the current performative environment. For any candidate direction d, the term $\max_{i\in[m]} \langle g_i(\theta), d\rangle$ measures the *worst* first-order change across objectives; minimizing this together with $\frac{1}{2}\|d\|^2$ yields the smallest-norm direction that improves all objectives simultaneously as much as possible. Crucially, no preference weighting across objectives is committed to in advance—the multiplier emerges endogenously from the convex geometry of $\{g_i(\theta)\}_{i\in[m]}$—so $d(\theta)$ is the canonical *preference-free* direction at θ.

The endogenous multiplier referenced above can be made explicit through the dual of equation 2 below.

Proposition 2.1 (MGDA dual formulation)**.** *Given $G = [g_1,\dots,g_m]$, the dual problem of equation 2 is $\lambda^\star \in \arg\min_{\lambda\in\Delta_m} \frac{1}{2}\|G\lambda\|^2$, and every primal–dual optimal pair satisfies $d^\star = -G\lambda^\star$. Moreover, the primal optimizer $d^\star$ is unique.*

With $d(\theta)$ defined, we can now formalize the corresponding notion of stationarity for PMOO.

Definition 2.2 (Performative Pareto stationary point)**.** A point $\theta \in \Theta$ is *performative Pareto stationary* (PPS) if $d(\theta) = 0$. More generally, θ is ε-PPS if $\|d(\theta)\| \le \varepsilon$.

PPS concerns first-order stationarity with the induced environment held fixed; it does not generally imply Pareto stationarity or optimality of the full performative risk vector $V(\theta)$. The condition $d(\theta) = 0$ says that no direction decreases all m objectives simultaneously to first order under the environment $D(\theta)$ that θ itself induces. This is exactly the multi-objective analogue of performative stability in scalar PP, where a stable point θ_{PS} admits no first-order improvement under its own induced environment $D(\theta_{\mathrm{PS}})$. Accordingly, $\|d(\theta)\|$ measures how far θ is from PPS and serves as the natural criticality measure for PMOO. The next lemma characterizes the structure of solutions to equation 2, the performative counterpart of a standard result in static MOO (Désidéri, 2012; Fliege & Svaiter, 2000; Lei et al., 2025).

Lemma 2.3 (MGDA geometry)**.** *Let $g_1,\cdots,g_m \in \mathbb{R}^d$ be doupled gradients and let d be a solution to equation 2. There exists $\lambda \in \Delta_m := \{\lambda \in \mathbb{R}^m_+ : \sum_i \lambda_i = 1\}$ such that*

$$d = -\sum_{i=1}^m \lambda_i g_i, \qquad \textit{and} \qquad \langle g_i, d\rangle \le -\frac{1}{2}\|d\|^2 \quad \forall i \in [m]. \tag{3}$$

Moreover, $d = 0$ if and only if $0 \in \mathrm{conv}\{g_1,\dots,g_m\}$.

Lemma 2.3 implies that the descent direction always takes the form $d(\theta) = -G(\theta)\lambda(\theta)$ for a simplex multiplier $\lambda(\theta) \in \triangle_m$, and the inequality $\langle g_i, d\rangle \le -\frac{1}{2}\|d\|^2$ confirms that $d(\theta)$ is a strict descent direction for every objective whose gradient is nonzero. The active components of $\lambda(\theta)$ identify which objectives *bind* in the current local trade-off, and $d(\theta)$ is the most balanced common descent generated by their gradients. Moreover, the equivalence $d = 0 \iff 0 \in \mathrm{conv}\{g_1,\dots,g_m\}$ shows that PPS coincides with the classical convex-hull first-order condition of static MOO, but evaluated with the decoupled performative gradients $J_i(\theta;\theta)$ rather than with the full derivatives of V_i. This decoupling is what makes the condition reachable from a deployed model, and is what distinguishes PMOO from a static MOO problem laid over a performative loss.

Why PMOO is nontrivial. We study the stability of adaptive multi-gradient updates under deployment feedback. The deployed model changes the environment in which subsequent gradients are evaluated, while the gradient-combination multipliers evolve with this geometry. This produces two coupled sources of nonstationarity:

- *Environment drift.* The descent direction at iteration t is computed under $D(\theta_t)$, while the next iterate θ_{t+1} is evaluated under shifted environment $D(\theta_{t+1})$. This invalidates the standard static-descent argument and introduces residual performative terms in one-step progress bound (Thm. 3.4);
- *Multiplier drift.* Preference-free MGD does not commit to any fixed scalarization. The simplex multiplier λ_t is recomputed at every iteration from the current gradient geometry $G(\theta_t)$, and so evolves with the iterate.

These two drifts do not act independently. At each step, θ_t induces an environment $D(\theta_t)$, which determines the gradient matrix $G(\theta_t)$ and the multiplier λ_t, which together drive the update direction $-G(\theta_t)\lambda_t$ and hence the next environment. The model, the environment, and multiplier all move one coupled trajectory. PMOO is therefore nontrivial not only because the environment moves with deployment, which already happens in scalar PP, but because the Pareto descent direction is itself selected through a multiplier that moves with the iterate. The analysis that follows is organized accordingly. The nonconvex result (Sec. 3) isolates environmental drift and yields an approximate-PPS guarantee on the best iterate of MGD. The strongly convex analysis (Sec. 4) additionally control multiplier drift, upgrading approximate criticality to convergence to a stable Pareto object.

3 (Proximal) Multi-Gradient Descent and Nonconvex Convergence Analysis

In scalar PP, the standard way to reach a performative stable point is *repeated gradient descent* (RGD): at each round, deploy the current model, sample data from the induced environment, and take a gradient step on the empirical loss as if the environment were static. The natural algorithm for PMOO is the multi-objective analogue of this loop: **multi-gradient descent (MGD)**, which at each deployment computes the per-objective decoupled gradients, combines them through the MGDA subproblem equation 2, and steps along the resulting common descent direction (Algorithm 1). The MGDA subproblem itself is the canonical first-order construction in static MOO for approaching Pareto stationary points without committing to a scalarization (Désidéri, 2012; Fliege & Svaiter, 2000); MGD applies it within a deployment loop, treating each induced environment as static for one step. Our first question is whether this principle remains useful when the objectives are performative.

Algorithm 1 Repeated multi-gradient descent (MGD)

Require: initial $\theta_0 \in \Theta$, stepsize $\eta > 0$, horizon T

1: **for** $t = 0, 1, \ldots, T-1$ **do**
2: Compute $g_i^t = \nabla_1 J_i(\theta_t; \theta_t)$ for all $i \in [m]$
3: Compute

$$d_t \in \arg\min_{d \in \mathbb{R}^d} \max_{i \in [m]} \langle g_i^t, d \rangle + \frac{1}{2} \|d\|^2$$

4: Deploy $\theta_{t+1} = \theta_t + \eta d_t$
5: **end for**
6: **return** the iterate with the smallest $\|d_t\|$

Nonconvex analysis of MGD. For smooth nonconvex PMOO, convergence to fixed PPS point is not generally guaranteed. Thm. 3.4 instead establishes a decay property of the criticality measure $\|d_t\|$.

Assumption 3.1 (Lipschitz loss)**.** *For each $i \in [m]$, $\ell_i(\theta; z)$ is S_i-Lipschitz in z for every $\theta \in \Theta$.*

Assumption 3.2 (Joint smoothness)**.** *For each $i \in [m]$, $\nabla_\theta \ell_i(\theta; z)$ is L_i-Lipschitz in both θ and z.*

Assumption 3.3 (ε-sensitivity)**.** *$\exists\, \varepsilon > 0$ such that $W_1(D(\theta), D(\theta')) \le \varepsilon \|\theta - \theta'\|$, $\forall \theta, \theta' \in \Theta$.*

Theorem 3.4 (Nonconvex convergence of MGD)**.** *Suppose Assumptions 3.1, 3.2, and 3.3 hold. Let $\kappa_i = \varepsilon S_i$ and $L = \max_i L_i$. If the stepsize $\eta \in (0, 1/L)$ and there exists $\underline{J}_i$ with $J_i(\theta; \theta) \ge \underline{J}_i$ for all $\theta \in \Theta$ and $i \in [m]$, then for every objective $i \in [m]$ and $T \ge 1$,*

$$\min_{0 \le t \le T-1} \|d_t\|^2 \le \frac{4\big(J_i(\theta_0; \theta_0) - \underline{J}_i\big)}{\eta(1 - L\eta)T} + \frac{4\kappa_i^2}{(1 - L\eta)^2}. \tag{4}$$

That is, MGD returns an ε_T-PPS iterate with $\varepsilon_T = \mathcal{O}(1/\sqrt{T}) + \mathcal{O}(\kappa_i)$.

The bound exposes a structural feature of performative MOO. Unlike static MGDA, which drives the criticality measure to zero, MGD with one update per environment change drives it only into a *sensitivity-determined neighborhood* whose radius scales with $\kappa_i = \varepsilon S_i$. Moreover, the above analysis also applies to state-dependent performative prediction (Brown et al., 2022), and we discuss the details in Appendix E.12 as a supplement to the main paper.

A natural follow-up question is whether the sensitivity floor can be removed under additional structure. The next corollary shows that it can, when at least one objective is unaffected by deployment.

Corollary 3.5 (Non-performative anchoring mechanism)**.** *Under the conditions of Theorem 3.4, if there exists $i \in [m]$ such that $J_i(\hat{\theta}; \theta) = J_i(\hat{\theta}; \theta')$ for any $\hat{\theta}, \theta, \theta' \in \Theta$, then MGD returns an $\mathcal{O}(1/\sqrt{T})$-PPS iterate.*

A non-performative objective acts as an *anchor*. By the MGDA geometry, every nonzero d_t decreases this objective by a quadratic amount in its frozen environment, and the decrease cannot be offset by performative drift. The bias floor therefore vanishes, and we recover $\mathcal{O}(1/\sqrt{T})$ best-iterate rate.

Vanishing $\|d_t\|$ is not enough. Thm. 3.4 guarantees that the criticality measure $\|d_t\|$ becomes small along the MGD trajectory. A natural follow-up is whether the iterates $\{\theta_t\}$ themselves converge to a fixed PPS point. The answer is no in general, and the reason illuminates the design of the algorithm we develop next.

Algorithm 2 Proximal MGD

Require: initial $\theta_0 \in \Theta$ and λ_0, $\tau > 0$, $\eta > 0$, T

1: **for** $t = 0, 1, \ldots, T-1$ **do**
2: Compute $g_i^t = \nabla_1 J_i(\theta_t; \theta_t)$ for all $i \in [m]$
3: Compute $H_t = G_t^\top G_t$ with $G_t = [g_1^t, \ldots, g_m^t]$
4: Compute

$$\widehat{\lambda}_t \in \arg\min_{\lambda \in \Delta_m} \left\{ \frac{1}{2} \lambda^\top H_t \lambda + \frac{\tau}{2} \|\lambda - \lambda_t\|^2 \right\}$$

5: Deploy $\theta_{t+1} = \theta_t - \eta G_t \widehat{\lambda}_t$, $\lambda_{t+1} = \widehat{\lambda}_t$
6: **end for**
7: **return** (θ_T, λ_T)

One obstacle is well-known from static optimization: vanishing first-order residuals do not imply iterate convergence. Even for smooth nonconvex single-objective minimization, the standard guarantee bounds only the *gradient norm* along the trajectory, while the iterates may oscillate in a neighborhood of stationary points without settling. PMOO inherits this difficulty, amplified by the two coupled drifts described in Section 2. Environmental drift means that a small $\|d_t\|$ only certifies stationarity under the current environment $D(\theta_t)$; it does not tell us whether the environment itself has settled. Multiplier drift adds a second source of instability: by Lemma 2.3, the same d_t can correspond to multiple simplex multipliers λ, and the one MGD implicitly selects, λ_t, is determined endogenously by the gradient geometry $G(\theta_t)$. As the iterate moves, λ_t moves with it, and so does the descent direction $-G(\theta_t)\lambda_t$. Even if the iterates approach a region of parameter space, the multiplier may continue to fluctuate, preventing them from settling.

The dual formulation of MGDA (Proposition 2.1) suggests a direct remedy. Since the multiplier λ is the primary variable in the dual, we can regularize its dynamics across iterations to slow multiplier drift. We call the resulting algorithm **proximal MGD (PMGD)**.

Proximal MGD. Algorithm 2 augments the dual subproblem of MGDA with a temporal proximal term $\frac{\tau}{2}\|\lambda - \lambda_t\|^2$, which penalizes deviation of the new multiplier from the previous one and thereby anchors multiplier dynamics across iterations. The strength of this anchoring is controlled by $\tau \geq 0$: when $\tau = 0$, the dual subproblem reduces to $\min_\lambda \frac{1}{2}\|G\lambda\|^2$ and PMGD recovers MGD exactly (by Prop. 2.1); when $\tau > 0$, the multiplier responds more slowly to changes in the gradient geometry.

4 Strongly Convex Convergence Analysis

The nonconvex analysis in Sec. 3 showed that MGD reaches only an approximate PPS, with iterates that may continue to oscillate even after $\|d_t\|$ becomes small. We next analyze the coupled model-multiplier dynamics under strong convexity. In this section, We derive sufficient conditions for convergence. These conditions control the worst-case interaction between deployment feedback and adaptive gradient weighting. Moreover, we assume that Θ is nonempty, compact, and convex, and the MGD/PMGD updates remain in Θ.

Assumption 4.1 (Strong convexity)**.** *For each $i \in [m]$, $\ell_i(\theta; z)$ is μ_i-strongly convex in θ for every z.*

Assumption 4.2 (Curvature dominates sensitivity)**.** *For each $i \in [m]$, $\frac{\mu_i}{L_i} > \varepsilon$, where L_i and ε are the smoothness and sensitivity constants of Sec. 3.*

These two assumptions are standard in scalar performative prediction setting, where together with Assumptions 3.2 and 3.3 they guarantee convergence to performative stable point θ_{PS} (Sec. 1) (Perdomo et al., 2020). Assumption 4.2 ensures that each $J_i(\cdot;\theta)$ has enough curvature to overcome the perturbation introduced by environment shifts. Concretely, define the *effective curvature* $\alpha = \min_{i\in[m]}(\mu_i - \varepsilon L_i)$,which is positive by Assumption 4.2. The quantity α measures the net curvature available for contraction after subtracting the sensitivity-induced perturbation, and will appear directly in the convergence rates.

PPS as a multi-objective performative stable point. Under strong convexity, the PPS condition is not merely a first-order criticality requirement; it is a global self-consistency property, generalizing performative stability from scalar PP to the multi-objective setting.

Proposition 4.3 (PPS as a Pareto-weighted performative stable point)**.** *Under Assumption 4.1, every PPS point $\theta^\star \in \Theta$ admits a multiplier $\lambda^\star \in \Delta_m$ such that $\theta^\star \in \arg\min_{\theta\in\Theta} \sum_{i=1}^m \lambda_i^\star J_i(\theta;\theta^\star)$.*

Proposition 4.3 identifies a PPS point as a best response to its own induced environment under a certifying simplex weighting. It does not mean that each objective is individually minimized. Moreover, uniqueness of the multiplier at a fixed model and uniqueness of the PPS point are distinct properties. For example, the strongly convex objectives $f_1(x) = x^2$ and $f_2(x) = (x-1)^2$ admit an entire interval of Pareto stationary models, each with a unique certifying multiplier. We therefore analyze model convergence through the joint update rather than infer it from multiplier identifiability alone.

4.1 Contraction relative to certifying multiplier fibers

Using the dual formulation (Prop. 2.1), define the *active multiplier set* and the *PPS fixed-point set* as

$$\Lambda(\theta) := \arg\min_{\lambda\in\Delta_m} \frac{1}{2}\lambda^\top H(\theta)\lambda, \quad \mathcal{M} := \{(\theta,\lambda) \in \Theta \times \Delta_m : \lambda \in \Lambda(\theta),\ G(\theta)\lambda = 0\}.$$

The set $\Lambda(\theta)$ collects all simplex multipliers that solve the dual subproblem at θ; $\mathcal{M}$ collects the joint (θ,λ)-configurations that simultaneously solve the dual and produce zero descent direction. Let $\mathcal{S} := \mathrm{proj}_\theta(\mathcal{M})$ denote the projection of $\mathcal{M}$ onto parameter space, i.e., the set of PPS points in Θ. By Prop. 4.3, every $\theta \in S$ is a multi-objective performative stable point, so $\mathcal{S}$ is the set of stable Pareto equilibria the algorithms could plausibly target. $\forall \bar\theta \in \mathcal{S}$, denote $\mathcal{F}(\bar\theta) = \{\lambda \in \triangle_m : (\bar\theta,\lambda) \in \mathcal{M}\}$ as the set of multipliers that certify $\bar\theta$ as a PPS point. Lemma E.2 shows that $\mathcal{M}$ and $\mathcal{S}$ are nonempty.

To control multiplier drift relative to the multipliers that certify a PPS point, without requiring the certifying multiplier itself to be unique, we impose the following transverse curvature condition.

Assumption 4.4 (Curvature transverse to the PPS set)**.** *$\exists\ \sigma_s > 0$ such that for any $\bar\theta \in \mathcal{S}$, $\lambda \in \triangle_m$ and $\bar\lambda \in \arg\min_{z\in\mathcal{F}(\bar\theta)} \|z - \lambda\|$, $(\lambda-\bar\lambda)^T H(\bar\theta)(\lambda-\bar\lambda) \geq \sigma_s \left\|\lambda - \bar\lambda\right\|^2$.*

For any $(\bar\theta,\bar\lambda) \in \mathcal{M}$ the inequality is trivial (both sides are zero, since $\bar\lambda$ is itself the nearest point in $\mathcal{F}(\bar\theta)$ to itself). The substantive content of the assumption is what happens *off* the PPS set: as soon as a multiplier λ moves away from its nearest certifying multiplier on $\mathcal{F}(\bar\theta)$, the quadratic form induced by $H(\bar\theta)$ grows at least quadratically in the deviation. The assumption therefore controls deviations transverse to the certifying multiplier fiber, while permitting flat directions within that fiber. Importantly, Assumption 4.4 by itself does not require the certifying multiplier of a PPS point to be unique.

The convergence rate depends on three additional constants that summarize the geometry of the iteration. Let $\Gamma = \max_{i\in[m]}(1+\varepsilon)L_i$ where L_i and ε are the smoothness and sensitivity constants.

$$r := \sqrt{1 - 2\eta\alpha + \eta^2\Gamma^2}, \qquad G_\infty := \sup_{\theta\in\Theta} \max_{i\in[m]} \|g_i(\theta)\| < \infty, \qquad c := \eta\sqrt{m}\, G_\infty.$$

The factor r captures the one-step contraction of the model: for $\eta \in (0, 2\frac{\alpha}{\Gamma^2})$, it satisfies $r \in (0,1)$, and a single MGD step contracts $\|\theta_t - \theta^\star\|$ by at most r when $\theta^\star$ is the target. The constant G_∞ bounds the size of any single objective gradient; it is finite because Θ is compact and g_i are Lipschitz (Lemma E.3). $c := \eta\sqrt{m}\, G_\infty$ then bounds the per-step displacement of one MGD update in parameter space. Finally, let L_H denote the Lipschitz constant of $H(\theta)$ on Θ (Lemma E.5).

Theorem 4.5 (Model convergence under transverse multiplier curvature)**.** *Under Assumptions 3.2, 3.3, 4.1, 4.2 and 4.4, for $\tau \geq 0$ let $A_s := \begin{pmatrix} r + c\frac{L_H}{\sigma_s+\tau} & c\frac{\tau}{\sigma_s+\tau} \\ \frac{L_H}{\sigma_s+\tau} & \frac{\tau}{\sigma_s+\tau} \end{pmatrix}$. If $r + c\frac{L_H}{\sigma_s} < 1$, then:* (i) ***PMGD****: the spectral radius $\rho(A_s) < 1$,and $\exists\, \theta_1^\star \in \mathcal{S}$ such that θ_t converges to $\theta_1^\star$ with $\|\theta_t - \theta_1^\star\| = \mathcal{O}\big(\rho(A_s)^t\big)$, and $\min_{\omega \in \mathcal{M}} \|(\theta_t, \lambda_t) - \omega\| = \mathcal{O}\big(\rho(A_s)^t\big)$;* (ii) ***MGD****: $\rho(A_s) = r + c\frac{L_H}{\sigma_s} < 1$, and $\exists\, \theta_2^\star \in \mathcal{S}$ such that θ_t converges to $\theta_2^\star$ with $\|\theta_t - \theta_2^\star\| = \mathcal{O}\big(\rho(A_s)^t\big)$;* (iii) ***Both*** *MGD and PMGD converge into the set $\mathcal{S} = \{\theta^*\}$ while its certifying multiplier need not be unique.*

The contraction matrix A_s encodes how the model error and the multiplier error interact across one iteration: the top row tracks how each error contributes to the next-step model error, the bottom row tracks how each contributes to the next-step multiplier error. The spectral radius $\rho(A_s)$ is therefore the joint contraction factor for the coupled (θ_t, λ_t) dynamics. The condition $r + c\,L_H/\sigma_s < 1$ is the requirement that this joint dynamics contract.

Theorem 4.5 separates model convergence from multiplier identifiability. Under the joint contraction condition, the PPS point is unique and both MGD and PMGD converge to the same θ^*. However, Assumption 4.4 only controls multiplier deviations transverse to the certifying fiber; it does not require this fiber to be a singleton. Therefore, model convergence can hold even when multiple multipliers certify the same PPS point. We next ask when stronger multiplier-side geometry also yields a unique certifying multiplier and hence a unique joint model-multiplier fixed point.

4.2 Contraction under uniform multiplier curvature

We next strengthen the multiplier-side geometry to obtain uniqueness and convergence of the joint model-multiplier state, rather than model convergence alone.

Assumption 4.6 (Curvature on the entire simplex tangent space)**.** *There exists a constant $\sigma > 0$ such that for any $\theta \in \Theta$ and $u \in \mathbb{R}^m$ with $\mathbf{1}^\top u = 0$, $u^\top H(\theta) u \geq \sigma \|u\|^2$.*

Unlike Assumption 4.4 where curvature is only required in directions transverse to $\mathcal{F}(\bar{\theta})$, Assumption 4.6 requires uniform curvature on *all* feasible multiplier perturbations (those orthogonal to $\mathbf{1}$). Since $u^\top H(\theta) u = \|G(\theta)u\|^2$, the condition says that any nontrivial feasible perturbation of the multiplier produces a nontrivial change in the combined gradient, i.e., the gradients $\{g_i(\theta)\}_{i\in[m]}$ remain sufficiently distinct on the simplex tangent space at every θ. This eliminates flat feasible directions and renders the multiplier subproblem uniformly curved, so that the dual minimizer λ_t is unique at every iteration. Uniqueness of the PPS point requires the full contraction conditions below, not multiplier uniqueness alone.

Theorem 4.7 (Joint model–multiplier convergence under uniform multiplier curvature)**.** *Under Assumptions 3.2, 3.3, 4.1, 4.2 and 4.6, for $\tau \geq 0$ let $A_\tau := \begin{pmatrix} r + c\frac{L_H}{\sigma+\tau} & c\frac{\tau}{\tau+\sigma} \\ \frac{L_H}{\sigma+\tau} & \frac{\tau}{\tau+\sigma} \end{pmatrix}$. If $r + c\frac{L_H}{\sigma} < 1$, then* (i) ***PMGD****: the spectral radius $\rho(A_\tau) < 1$ and PMGD has a unique fixed point $(\theta_1^\star, \lambda^\star) \in \Theta \times \triangle_m$ with $\theta_1^\star$ a PPS point, satisfying $\|\theta_t - \theta_1^\star\| + \|\lambda_t - \lambda^\star\| = \mathcal{O}\big(\rho(A_\tau)^t\big)$;* (ii) ***MGD****: $\rho(A_\tau) = r + c\frac{L_H}{\sigma} < 1$, and MGD has a unique fixed point $\theta_2^* \in \Theta$ which is a PPS point, satisfying $\|\theta_t - \theta_2^*\| = \mathcal{O}\big(\rho(A_\tau)^t\big)$;* (iii) *MGD and PMGD share the same unique joint PPS certificate (θ^*, λ^*) and $\theta^* = \theta_1^* = \theta_2^*$.*

Matrix A_τ has the same structure as A_s from Thm. 4.5, with the weaker constant σ_s replaced by σ, using uniform multiplier curvature. The uniqueness conclusion follows from contraction of the joint update under all stated assumptions. The stronger curvature assumption alone does not imply a faster rate or a unique PPS point.

5 Experiments

This section presents experiments on both synthetic and real data that illustrate the theoretical phenomena and the convergence behavior of MGD and PMGD. Detailed settings, additional results (e.g., comparisons with scalar methods), and discussion are provided in Appendix D.

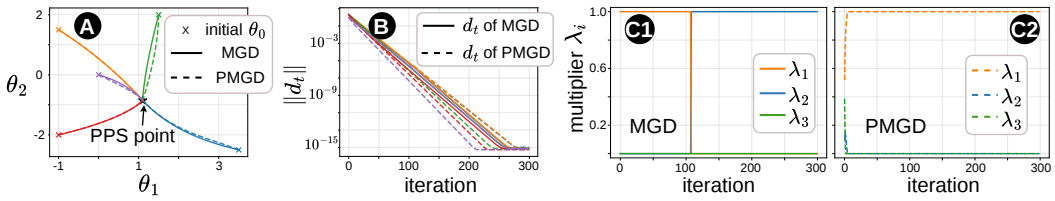


Figure 1: (A): MGD and PMGD converge to the unique PPS point under different initializations. (B): d_t during the iterations of the algorithms in (A) with different initializations. (C1)(C2): Multiplier $\lambda = (\lambda_1, \lambda_2, \lambda_3)$ during the iterations under one initialization in (A).

Synthetic data. We construct a family of synthetic PMOO instances designed to isolate the algorithmic phenomena identified by the theory. Throughout, we set the number of objectives to $m = 3$ and use the quadratic decoupled objective $J_i(\hat{\theta}; \theta) = \frac{1}{2}\|Q_i^{1/2}(\hat{\theta} - B_i\theta - b_i)\|^2$ where $\theta, \hat{\theta}, b_i \in \mathbb{R}^2$ and $Q_i, B_i \in \mathbb{R}^{2\times 2}$. The decoupled gradient is then $g_i(\theta) = \nabla_1 J(\theta; \theta) = Q_i(I - B_i)\theta - Q_i b_i$; for brevity, let $A_i = Q_i(I - B_i)$ and $c_i = Q_i b_i$.

We first examine convergence to a unique PPS point. Take each A_i positive definite and $c_i = A_i(1.1, -0.85)^T$. Then $\theta^* = (1.1, -0.85)$ is the unique solution to $0 \in \text{conv}\{g_1(\theta^*), .., g_3(\theta^*)\}$, hence the unique PPS point. Under various initializations, both MGD and PMGD converge to θ^*, with $\|d_t\| \to 0$ (Fig. 1 (A)-(B)).

The same setup also illustrates multiplier drift. Fig. 1(C1) shows that MGD's multiplier λ_t becomes discontinuous near $\theta^\star$: as $g_i(\theta_t) = A_i(\theta_t - \theta^*) \approx 0$, the matrix $G(\theta_t)$ vanishes, the multiplier subproblem $\arg\min_{\lambda\in\triangle_m} \frac{1}{2}\lambda^T H(\theta_t)\lambda$ becomes degenerate, and the solution oscillates wildly on the simplex (different λ values yield nearly identical $\|G(\theta_t)\lambda\|$). This confirms that (i) vanishing $\|d_t\| = \|G(\theta_t)\lambda\|$ does not imply multiplier convergence, and multiplier drift is a real phenomenon; (ii) PMGD's proximal regularization stabilizes the multiplier dynamics (Fig. 1(C2)).

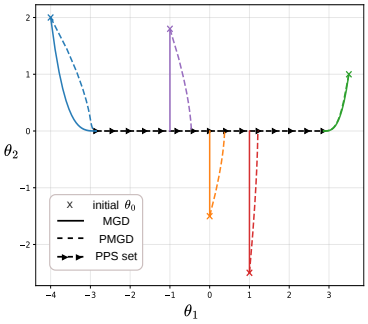


Figure 2: MGD and PMGD converge to different points within the same PPS set under different initializations.

To illustrate that strong convexity alone does not guarantee a unique PPS point, we next construct a two-objective instance with $g_1(\theta_1, \theta_2) = (\kappa\theta_1 + 1, \theta_2)$ and $g_2(\theta_1, \theta_2) = (\kappa\theta_1 - 1, \theta_2)$, where $\kappa = 0.35$. The PPS set is $\mathcal{S} = \{(x, 0) : x \in [-1/\kappa, 1/\kappa]\}$. Fig. 2 shows different initializations can converge to different PPS points. Since this instance has a non-singleton PPS set, it necessarily lies outside the contraction regime of Thm. 4.5. The example therefore illustrates why strong convexity alone is insufficient and why an additional joint contraction condition is needed for model-level uniqueness.

Credit data. We evaluate PMOO on a credit-risk prediction task based on the *Give Me Some Credit* dataset (Fusion & Cukierski, 2011) with a logistic classifier as the prediction model. We split individuals into three age groups by age quantiles and treat the group-wise logistic risks as three separate objectives. The three group-specific logistic risks define the objectives, allowing us to examine adaptive gradient weighting under heterogeneous simulated responses to deployment.

Following Perdomo et al. (2020), performativity is introduced through strategic feature response: after deploying θ, a subset of credit-related features M is shifted according to $x'_M = x_M - \epsilon_g \theta_M$, where $\epsilon_g, g \in \{1, 2, 3\}$ is the group-dependent response strength and θ_M is the model coefficient restricted to features M. We set the default response strengths to $\epsilon_1 = 0.04, \epsilon_2 = 0.08, \epsilon_3 = 0.12$, reflecting that different groups experience the deployment-induced feature shift to different degrees.

Fig. 3 reports the limiting group-loss vectors, multiplier trajectories, and stationarity residuals for MGD and PMGD. The two algorithms exhibit decreasing residuals and different final group-loss vectors : their final group-loss vectors illustrate the trade-off geometry directly, i.e., improving one group's loss often worsens another's, and confirm that the choice of algorithm (and initialization) affects which Pareto compromise is reached.

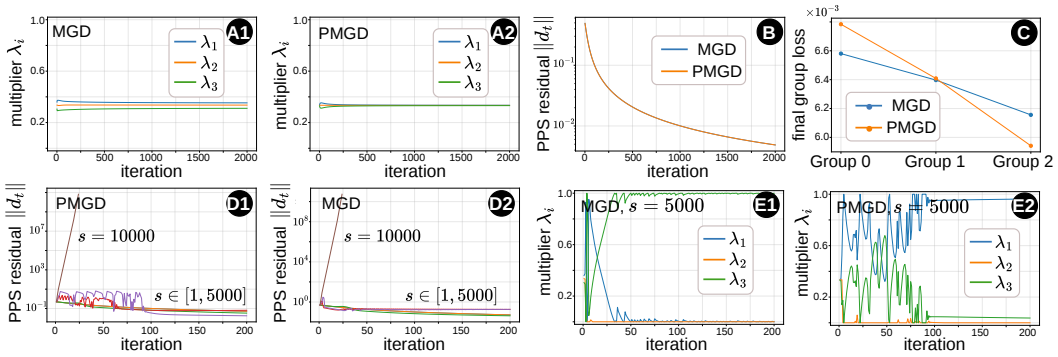


Figure 3: (A1)(A2): The multiplier λ during the iterations for MGD and PMGD with $s = 1$. (B): $\|d_t\|$ for both algorithms with $s = 1$. (C): The group losses at iteration $t = 2000$. (D1)(D2): $\|d_t\|$ with different factor s for MGD and PMGD. (E1)(E2): The multiplier λ for MGD and PMGD with factor $s = 5000$ in the first 200 iterations.

We next examine how the strength of performative feedback affects the analysis. Scaling each response strength by a factor s, so that $\hat{\epsilon}_g = s\epsilon_g$, we increase the deployment-induced distribution shift. Fig. 3(D) shows that as s grows, the criticality residual $\|d_t\|$ diverges and the parameter norm grows substantially, consistent with the system leaving the regime covered by the sufficient contraction conditions in Thm. 4.5 and 4.7. Comparing Figs. 3(A) and (E), the multiplier trajectories also become more variable at larger s, illustrating the coupled behavior predicted by our analysis: environmental and multiplier drift are coupled in PMOO, and stronger model-induced distribution shift destabilizes both the model iterates and the endogenous trade-off multiplier.

6 Conclusion

We study the stability of adaptive multi-gradient learning under performative feedback. We study performative multi-objective optimization (PMOO), where multiple objectives are evaluated under environments that change with the deployed model. The central interaction is between deployment-induced changes in gradient geometry and the endogenous multipliers that determine subsequent updates. We propose PMGD to stabilize the multiplier drift. In the smooth nonconvex regime, we establish a best-iterate approximate-PPS guarantee for MGD. We next derive sufficient convergence conditions for the coupled dynamics under strong convexity. PMGD provides temporal regularization of the multiplier update. Synthetic and credit-risk experiments illustrate model convergence, multiplier behavior, and sensitivity to deployment feedback.

AI USE STATEMENT

In this work, we used generative AI tools for limited assistance. The theoretical framework, mathematical claims, and proofs were developed and completed by the authors. Generative AI was used only to review the proofs and help identify potential issues. For the experiments, generative AI was used to assist with writing and editing code. Generative AI was also used to improve the clarity and readability of the manuscript. We did not use generative AI for the other required-disclosure tasks, or the remaining required-disclosure tasks were not applicable to this work. We have reviewed all AI-assisted work before incorporating them into the final work. We take responsibility for the final content of this work, including all text, claims and other artifacts produced with the aid of generative AI.

REFERENCES

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International conference on machine learning*, pp. 60–69. PMLR, 2018.

Omer Ben-Porat, Fedor Sandomirskiy, and Moshe Tennenholtz. Protecting the protected group: Circumventing harmful fairness. In *Proceedings of the aaai conference on artificial intelligence*, volume 35, pp. 5176–5184, 2021.

Gavin Brown, Shlomi Hod, and Iden Kalemaj. Performative prediction in a stateful world. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, pp. 6045–6061. PMLR, 2022.

Jean-Antoine Désidéri. Multiple-gradient descent algorithm (mgda) for multiobjective optimization. *Comptes Rendus. Mathématique*, 350(5-6):313–318, 2012.

Dmitriy Drusvyatskiy and Lin Xiao. Stochastic optimization with decision-dependent distributions. *Mathematics of Operations Research*, 48(2):954–998, 2023.

Joerg Fliege, LM Grana Drummond, and Benar Fux Svaiter. Newton's method for multiobjective optimization. *SIAM Journal on Optimization*, 20(2):602–626, 2009.

J"org Fliege and Benar F. Svaiter. Steepest descent methods for multicriteria optimization. *Mathematical Methods of Operations Research*, 51(3):479–494, 2000.

J"org Fliege, A. Ismael F. Vaz, and Luís Nunes Vicente. Complexity of gradient descent for multiobjective optimization. *Optimization Methods and Software*, 34(5):949–959, 2019.

Olive Franzese, Ali Shahin Shamsabadi, Carter Luck, and Hamed Haddadi. Secure and confidential certificates of online fairness. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Credit Fusion and Will Cukierski. Give me some credit. `https://kaggle.com/competitions/GiveMeSomeCredit`, 2011. Kaggle.

Yingqiang Ge, Xiaoting Zhao, Lucia Yu, Saurabh Paul, Diane Hu, Chu-Cheng Hsieh, and Yongfeng Zhang. Toward pareto efficient fairness-utility trade-off in recommendation through reinforcement learning. *arXiv preprint arXiv:2201.00140*, 2022.

Moritz Hardt and Celestine Mendler-D"unner. Performative prediction: Past and future. *Statistical Science*, 40(3):417–436, 2025.

Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In *Proceedings of the 2016 ACM conference on innovations in theoretical computer science*, pp. 111–122, 2016a.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016b.

Lily Hu and Yiling Chen. Fair classification and social welfare. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 535–545, 2020.

Zachary Izzo, Lexing Ying, and James Zou. How to learn when data reacts to your model: Performative gradient descent. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4641–4650. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/izzo21a.html.

Zhuangzhuang Jia, Yijie Wang, Roy Dong, and Grani A. Hanasusanto. Distributionally robust performative optimization. *arXiv preprint arXiv:2407.01344*, 2024.

Kun Jin, Xueru Zhang, Mohammad Mahdi Khalili, Parinaz Naghizadeh, and Mingyan Liu. Incentive mechanisms for strategic classification and regression problems. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 760–790, 2022.

Kun Jin, Tian Xie, Yang Liu, and Xueru Zhang. Addressing polarization and unfairness in performative prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 22408–22416, 2026.

Pedram Khorsandi, Rushil Gupta, Mehrnaz Mofakhami, Simon Lacoste-Julien, and Gauthier Gidel. Tight lower bounds and improved convergence in performative prediction. In *Advances in Neural Information Processing Systems*, 2025.

William G La Cava. Optimizing fairness tradeoffs in machine learning with multiobjective metamodels. In *Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 511–519, 2023.

Tosca Lechner, Ruth Urner, and Shai Ben-David. Strategic classification with unknown user manipulations. In *International Conference on Machine Learning*, pp. 18714–18732. PMLR, 2023.

Yuan-Zheng Lei, Yaobang Gong, and Xianfeng Terry Yang. A two-stage search framework for constrained multi-gradient descent. *arXiv preprint arXiv:2502.14104*, 2025.

Qiang Li and Hoi-To Wai. Stochastic optimization schemes for performative prediction with nonconvex loss. In *Advances in Neural Information Processing Systems*, 2024.

Qiang Li, Chung-Yiu Yau, and Hoi-To Wai. Multi-agent performative prediction with greedy deployment and consensus seeking agents. *Advances in Neural Information Processing Systems*, 35:38449–38460, 2022.

Annie Liang, Jay Lu, and Xiaosheng Mu. Algorithmic design: Fairness versus accuracy. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pp. 58–59, 2022.

Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 6755–6764. PMLR, 2020.

Celestine Mendler-D"unner, Juan Perdomo, Tijana Zrnic, and Moritz Hardt. Stochastic optimization for performative prediction. In *Advances in Neural Information Processing Systems*, volume 33, pp. 4929–4939, 2020.

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

Smitha Milli, John Miller, Anca D Dragan, and Moritz Hardt. The social cost of strategic classification. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 230–239, 2019.

Alan Mishler and Nicole Dalmasso. Fair when trained, unfair when deployed: Observable fairness measures are unstable in performative prediction settings. *arXiv preprint arXiv:2202.05049*, 2022.

Mehrnaz Mofakhami, Ioannis Mitliagkas, and Gauthier Gidel. Performative prediction with neural networks. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, pp. 11079–11093. PMLR, 2023.

Kanad S Pardeshi, Itai Shapira, Ariel D Procaccia, and Aarti Singh. Learning social welfare functions. *Advances in Neural Information Processing Systems*, 37:41733–41766, 2024.

Garnet Liam Peet-Pare, Nidhi Hegde, and Alona Fyshe. Long term fairness via performative distributionally robust optimization. *OpenReview preprint*, 2023.

Juan C. Perdomo. Performative prediction: Theory and practice. 2023. URL https://api.semanticscholar.org/CorpusID:272930686.

Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-D"unner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 7599–7609. PMLR, 2020.

Georgios Piliouras and Fang-Yi Yu. Multi-agent performative prediction: From global stability and optimality to chaos. In *Proceedings of the 24th ACM Conference on Economics and Computation*, pp. 1047–1074, 2023.

Reilly Raab, Ross Boczar, Maryam Fazel, and Yang Liu. Fair participation via sequential policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14758–14766, 2024.

Mitas Ray, Lillian J Ratliff, Dmitriy Drusvyatskiy, and Maryam Fazel. Decision-dependent risk minimization in geometrically decaying dynamic environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 8081–8088, 2022.

Esther Rolf, Max Simchowitz, Sarah Dean, Lydia T Liu, Daniel Bjorkegren, Moritz Hardt, and Joshua Blumenstock. Balancing competing objectives with noisy data: Score-based classifiers for welfare-aware machine learning. In *International Conference on Machine Learning*, pp. 8158–8168. PMLR, 2020.

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

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

Tian Xie, Ding Zhu, Jia Liu, Mahdi Khalili, and Xueru Zhang. Sprint: Stochastic performative prediction with variance reduction. *arXiv preprint arXiv:2509.17304*, 2025.

Man-Jie Yuan and Wei Gao. Learning with interactive models over decision-dependent distributions. In *Asian Conference on Machine Learning*, pp. 1229–1244. PMLR, 2023.

Xueru Zhang, Mahdi M. Khalili, Cem Tekin, and Mingyan Liu. Group retention when using machine learning in sequential decision making: The interplay between user dynamics and fairness. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Xueru Zhang, Ruibo Tu, Yang Liu, Mingyan Liu, Hedvig Kjellstrom, Kun Zhang, and Cheng Zhang. How do fair decisions fare in long-term qualification? *Advances in Neural Information Processing Systems*, 33:18457–18469, 2020.

Xueru Zhang, Mohammad Mahdi Khalili, Kun Jin, Parinaz Naghizadeh, and Mingyan Liu. Fairness interventions as (dis) incentives for strategic manipulation. In *International Conference on Machine Learning*, pp. 26239–26264. PMLR, 2022.

A Related work

Performative prediction and stateful performativity. The original performative prediction framework (Perdomo et al., 2020) introduced performative stability and repeated risk minimization for scalar objectives. This line was extended to stochastic optimization (Mendler-D"unner et al., 2020), stateful environments (Brown et al., 2022), decision-dependent stochastic approximation (Drusvyatskiy & Xiao, 2023), and nonconvex scalar objectives (Li & Wai, 2024; Mofakhami et al., 2023). Very recent work sharpens scalar convergence bounds and studies variance-reduced or history-aware

methods, but still for a single objective (Khorsandi et al., 2025; Xie et al., 2025). Our work is orthogonal: we ask what happens when several performative criteria must be optimized simultaneously. Additionally, recent work also move beyond performative stability itself, studying performative gradient methods and direct optimization of the performative risk (Miller et al., 2021). Broader decision-dependent learning models analyze interactive or dynamically evolving distributions beyond the original PP framework (Ray et al., 2022; Yuan & Gao, 2023). More recently, multi-agent variants of performative prediction have shown that once one moves beyond a single learner, the resulting dynamics can exhibit consensus effects, instability, and even chaotic behavior (Li et al., 2022; Piliouras & Yu, 2023).

Fairness and welfare under performative or strategic feedback. Recent work documents that scalar performatively stable solutions can be unfair and can amplify polarization (Jin et al., 2026; Mishler & Dalmasso, 2022). Existing remedies mostly modify a scalar objective through regularization, reweighting, or distributional robustness (Peet-Pare et al., 2023; Jin et al., 2026; Jia et al., 2024; Ben-Porat et al., 2021). Related strategic-learning work studies fairness and welfare under behavioral responses but likewise optimizes scalar objectives or fixed trade-offs (Xie & Zhang, 2024; Zhang et al., 2019; Hardt et al., 2016a). Broader literature includes work on the social cost of strategic classification, on fairness interventions as incentives or disincentives for manipulation, and on strategic classification under uncertain or unknown user responses (Milli et al., 2019; Rolf et al., 2020; Lechner et al., 2023). These methods are natural when a scalar surrogate or a fixed trade-off is specified in advance. Our paper studies a complementary regime in which the gradient-combination weights are selected adaptively from the current multi-objective geometry and therefore evolve jointly with the performative environment.

Multi-objective optimization. MGDA and steepest-descent methods provide first-order algorithms for Pareto stationarity in smooth MOO (Fliege & Svaiter, 2000) as well as higher-order methods such as multi-objective Newton schems (Fliege et al., 2009), and their worst-case rates match the single-objective nonconvex and convex complexity theory (Fliege et al., 2019). In particular, the classical strongly convex MGDA rate is already formulated relative to a convergent trajectory and the scalarization implicitly selected at the limit (Fliege et al., 2019). MOO has also become a natural framework for fairness–utility trade-offs outside performative settings, including welfare-aware decision rules, Pareto-efficient fairness-utility trade-offs in the recommendation systems(Martinez et al., 2020; Ge et al., 2022; La Cava, 2023; Liang et al., 2022). Our analysis adapts the MGDA viewpoint to performative shifts and combines it with stable-point arguments from performative prediction.

Taken together, these literatures address complementary parts of the problem. Performative prediction provides stability tools for learning under model-dependent distributions, while classical MOO provides adaptive multi-gradient methods for Pareto stationarity under static environments. Fixed-weight performative methods remain appropriate when a preference weighting is specified in advance. Our focus is the regime in which the gradient-combination weights are instead selected endogenously and evolve together with the performative environment. We formulate this regime as PMOO and analyze the resulting coupled model-multiplier dynamics.

B Discussion

Strength and plausibility of Assumption 4.4 and 4.6. The two multiplier-curvature conditions are nested: Assumption 4.6 implies Assumption 4.4. To see this, observe that for any $\lambda, \bar{\lambda} \in \triangle_m$, we have $\mathbf{1}^\top(\lambda - \bar{\lambda}) = 0$, so $(\lambda - \bar{\lambda})$ lies in the simplex tangent space and Assumption 4.6 immediately gives $(\lambda - \bar{\lambda})^T H(\bar{\theta})\lambda - \bar{\lambda} \geq \sigma \left\|\lambda - \bar{\lambda}\right\|^2$, which is the conclusion of Assumption 4.4 with $\sigma_s = \sigma$.

The two conditions control different levels of multiplier-side geometry. Assumption 4.4 requires curvature only transverse to the certifying multiplier fiber, and therefore allows multiple multipliers to certify the same PPS point. Under the additional contraction condition in Thm. 4.5, however, the PPS point itself is unique. Assumption 4.6 is stronger: it imposes curvature over the entire simplex tangent space, making the multiplier response identifiable and yielding a unique certifying multiplier. Together with the joint contraction condition, this gives uniqueness and convergence of the full model–multiplier state in Thm. 4.7.

Extension to anchored MGD methods. Our framework can also be extended to analyze anchored preference with a fixed weight. When a preference vector is prescribed in advance, weighted gradient descent (WGD) and anchored MGD methods act as PPS selection mechanisms rather than preference-free PMOO algorithms. WGD refers to the method as shown in Algorithm 3 which optimizes $\sum_{i=1}^{m} \omega_i J_i(\hat{\theta}; \theta)$ and $\omega \in \triangle_m$ is a specified preference vector. Anchored MGD methods, on the other hand, extends current MGD or PMGD by incorporating regularization $\|\lambda - \bar{\lambda}\|^2$ on the specified multiplier $\bar{\lambda}$ as an anchor in Algorithm 4. In this setting, the multiplier is no longer an endogenous state variable once a preference vector is prescribed, so multiplier drift disappears from the dynamics. It can be proven that under the standard performative assumptions, these methods converge to the same unique PPS point associated with the prescribed preference, without imposing the additional multiplier-dynamics Assumption 4.6 and 4.4 required by MGD and PMGD. We therefore view WGD and anchored variants as a complementary fixed-preference layer of our framework. Detailed analyses are in Section C.

Limitations. This work is best viewed as a first theoretical step toward performative multi-objective optimization (PMOO) rather than a complete treatment of the area. First, the current analysis is population-level and does not model statistical error, which is unavoidable in practice when gradients, multipliers, and environment-dependent quantities are estimated from finite samples. A key next step is therefore to quantify how finite-sample noise and algorithmic estimation error affect the PPS residual, the asymptotic neighborhood size, and the related convergence error. In addition, our theory results in Section 4 rely on strong convexity and related smoothness–sensitivity assumptions. While these assumptions are standard in performative prediction theory (Perdomo et al., 2020), they are restrictive for modern overparameterized models, especially large neural networks, which are typically nonconvex in their parameters. Recent scalar performative prediction work (Mofakhami et al., 2023) has begun to develop theory better suited to such settings, and extending those ideas to the multi-objective case is an important open problem. More broadly, the present paper isolates the basic geometry and dynamics of PMOO, but a fuller theory will need to incorporate statistical error, stochastic optimization, and nonconvex model classes in a unified framework.

Future directions. In addition to the directions discussed in Limitations, there are several other directions worth exploring. A first direction is stochastic PMOO, where multi-gradient methods are combined with stochastic or variance-reduced estimators under performative feedback. A second direction is delayed or lazy deployment, which may reduce not only environment drift but also scalarization drift. A third direction is empirical: PMOO should be instantiated on concrete performative fairness, robustness, and welfare benchmarks, where the PPS set or the supported stable frontier may itself be a meaningful policy object.

C WGD AND RELATED ANCHORED MGD METHODS

In the main text, we focus on *preference-free* PMOO methods, namely MGD and PMGD. In this section, we include two fixed-preference counterparts, weighted gradient descent (WGD) and anchored MGD-type methods. Rather than searching over stable Pareto compromises without external preference information, they inject a prescribed preference vector and therefore act as *PPS selection mechanisms*. In other words, once a multiplier preference is fixed in advance, the goal is no longer to characterize the full preference-free PPS geometry, but to select the supported stable point associated with that prescribed preference.

In MGD and PMGD, the multiplier is an endogenous state variable determined by the current gradient geometry, and controlling its evolution is a central difficulty of PMOO. By contrast, in WGD and anchored variants, the preference is fixed exogenously, so multiplier drift is eliminated largely from the dynamics. Under the same standard performative assumptions used in the main text, we can therefore prove convergence to the supported PPS point associated with the prescribed preference without imposing the additional multiplier-geometry conditions (Assumption 4.6 and 4.4) required for preference-free MGD and PMGD.

WGD. The weighted gradient descent (WGD) method assumes that the decision-maker specifies a preference vector $w \in \Delta_m$, which can be regared as a fixed multiplier $\lambda = w$. As shown in

Algorithm 3, it then optimizes the decoupled scalarization

$$F_w(\tilde{\theta};\theta) := \sum_{i=1}^{m} w_i J_i(\tilde{\theta};\theta), \tag{5}$$

through repeated gradient descent $\theta_{t+1} = \theta_t - \eta \nabla_1 F_w(\theta_t;\theta_t)$. When the scalarized objective satisfies the standard performative assumptions, this simpler update admits a global linear rate.

Algorithm 3 Repeated weighted gradient descent (WGD)

Require: initial model $\theta_0 \in \Theta$, weight vector $w \in \Delta_m$, stepsize $\eta > 0$, horizon T

1: **for** $t = 0, 1, \ldots, T-1$ **do**
2: compute $g_t^w = \nabla_1 F_w(\theta_t;\theta_t) = \sum_{i=1}^m w_i \nabla_1 J_i(\theta_t;\theta_t)$
3: deploy $\theta_{t+1} = \theta_t - \eta g_t^w$
4: **end for**
5: return θ_T

Theorem C.1 (Linear convergence of WGD)**.** *Under Assumption 3.2, 3.3, 4.1 and 4.2, if the stepsize $0 < \eta \leq \frac{2\alpha}{\Gamma^2}$, then WGD admits a unique fixed point $\theta_w^\star$ which is a PPS point and $\|\theta_t - \theta_w^\star\| \leq r^t \|\theta_0 - \theta_w^\star\|$.*

C.1 Anchored Methods as PPS Selection Mechanisms

Algorithm 4 Anchored MGD methods

Require: initial model $\theta_0 \in \Theta$, initial multiplier λ_0, anchor $\bar{\lambda} \in \Delta_m$, $\tau \geq 0$, $\gamma > 0$, stepsize $\eta > 0$, T

1: **for** $t = 0, 1, \ldots, T-1$ **do**
2: compute $g_i^t = \nabla_1 J_i(\theta_t;\theta_t)$ for all $i \in [m]$
3: form $G_t = [g_1^t, \ldots, g_m^t]$ and compute $H_t = G_t^\top G_t$
4: compute

$$\hat{\lambda_t} \in \operatorname*{arg\,min}_{\lambda \in \Delta_m} \left\{ \frac{1}{2}\lambda^\top H_t \lambda + \frac{\tau}{2}\|\lambda - \lambda_t\|^2 + \frac{\gamma}{2}\left\|\lambda - \bar{\lambda}\right\|^2 \right\}$$

5: deploy $\theta_{t+1} = \theta_t - \eta G_t \hat{\lambda_t},\ \lambda_{t+1} = \hat{\lambda_t}$
6: **end for**
7: **return** (θ_T, λ_T)

If $\tau = 0$, Algorithm 4 only restricts the multiplier from deviating the predetermined preference $\bar{\lambda}$; we refer to this algorithm as anchored MGD (**AMGD**). If $\tau > 0$, Algorithm 4 additionally constrains the dynamics of the multiplier iteration process; we refer to this algorithm as anchored proximal MGD (**PAMGD**). Anchored methods, AMGD and PAMGD, add a strictly convex regularizer around a prescribed $\bar{\lambda} \in \Delta_m$. This removes flat multiplier directions by design and selects one supported PPS point as following.

Theorem C.2 (Anchored MGD)**.** *Under Assumptions 3.2, 3.3, 4.1, 4.2, if $\rho_\gamma := r + c\frac{L_H}{\gamma} < 1$, AMGD has a unique fixed point $\theta^\star$ which is a PPS point, the corresponding multiplier $\operatorname{arg\,min}_{\mu \in \Delta_m} \left\{ \frac{1}{2}\mu^\top H(\theta^*)\mu + \frac{\gamma}{2}\|\mu - \bar{\lambda}\|^2 \right\} = \bar{\lambda}$ and $\|\theta_t - \theta^\star\| \leq \rho_\gamma^t \|\theta_0 - \theta^\star\|$.*

Let $p_{\tau,\gamma} := \frac{L_H}{\tau+\gamma}, q_{\tau,\gamma} := \frac{\tau}{\tau+\gamma}$, and $A_{\tau,\gamma} := \begin{pmatrix} r + cp_{\tau,\gamma} & cq_{\tau,\gamma} \\ p_{\tau,\gamma} & q_{\tau,\gamma} \end{pmatrix}$.

Theorem C.3 (Proximal Anchored MGD)**.** *Under Assumptions 3.2, 3.3, 4.1, 4.2, if $\rho(A_{\tau,\gamma}) < 1$, PAMGD has a unique fixed point $(\theta^\star, \lambda^\star)$ such that θ^* is PPS and $\lambda^* = \bar{\lambda}$, and there exists $C_{\tau,\gamma} > 0$ such that*

$$\|\theta_t - \theta^\star\| + \|\lambda_t - \lambda^\star\| \leq C_{\tau,\gamma} \rho(A_{\tau,\gamma})^t.$$

Lemma C.4 (Same anchor as WGD)**.** *Fix $w \in \Delta_m$ in WGD and set the anchored multiplier $\bar{\lambda} = w$, then WGD, AMGD and PAMGD share the same fixed PPS point.*

MGD and proximal MGD preserve the original PPS geometry and may converge to a PPS set. Anchored methods add an external multiplier preference, which selects one supported PPS point. Thus anchoring can be regarded as a PPS selection mechanism.

D Additional Experiments

In this section, we present the detailed settings for experiments in Section 5 and additional experimental results.

D.1 Experimental Settings in Section 5

In the synthetic experiment, iteration numbers is 300, and the stepsize is set to 0.25. The proximal multiplier regularization $\tau = 0.35$ in PMGD. The A_i used in the experiment to examine convergence to a unique PPS point are set as

$$A_1 = \begin{pmatrix} 1.1 & 0 \\ 0 & 0.8 \end{pmatrix}, \; A_2 = \begin{pmatrix} 1.2268 & 0.3262 \\ 0.3262 & 0.9732 \end{pmatrix}, A_3 = \begin{pmatrix} 1.2082 & -0.1683 \\ -0.1683 & 0.9919 \end{pmatrix}.$$

In the credit experiment, we divided the population into three age groups based on the one-third and two-thirds quantiles of the age distribution. The subset of credit-related features M is $\{DebtRatio,\ MonthlyIncome,\ RevolvingUtilizationOfUnsecuredLines\}$. We use logistic regression with an L2 regularization coefficient 10^{-3}. The model step size is 0.1, and the simplex multiplier subproblem is solved by projected gradient steps on the simple with multiplier step size 0.1 and 20 inner iterations. The proximal multiplier regularization strength $\tau = 5$ in PMGD. All of our experiments are executed on CPUs; therefore, computational resources are not a limiting factor.

D.2 Supplement to the experiments in Section 5

In the Credit semi-synthetic experiment, the PPS residual $\|d_t\|$ decreases noticeably more slowly than in the fully controlled synthetic experiments by comparing Figure 3(B) and Figure 1(B). This is expected because the Credit setting involves finite-sample group-wise logistic losses, heterogeneous age-group objectives, and model-dependent strategic feature responses, rather than a low-dimensional quadratic construction with closed-form geometry.

As a result, the local multi-objective geometry is less well conditioned, and the simplex multiplier subproblem only gradually reduces the common-descent residual. To verify that this behavior is not caused by non-convergence, we additionally ran the methods for 30000 rounds. As shown in Figure 4, the residual $\|d_t\|$ continues to decrease and reaches the order of 10^{-4}, while the group-wise losses had already become nearly flat much earlier. This indicates that the objective values stabilize before the first-order Pareto residual is fully minimized: after the main loss reduction phase, the remaining iterations mainly refine the local Pareto-stationarity certificate rather than substantially changing the achieved group-loss trade-off. Thus, the credit experiment should be interpreted as a realistic setting where convergence to approximate PPS is slower than in controlled synthetic problems, but the qualitative PPS-selection behavior remains stable.

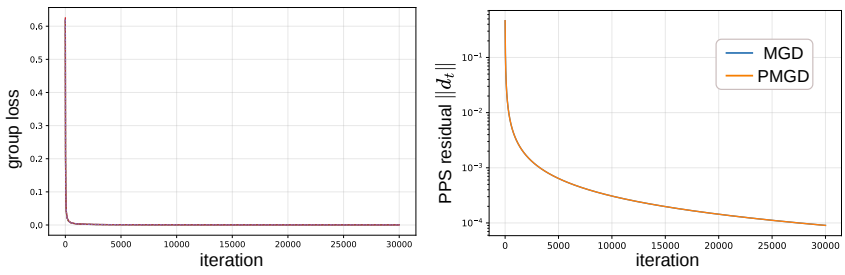


Figure 4: Left: group losses g_i for MGD and PMGD ($i \in \{1, 2, 3\}$) during the 30000 iterations. Right: The PPS residual $\|d_t\|$ for MGD and PMGD during the iterations.

D.3 Comparisons With Scalar Methods

In this section, we present the results of WGD and anchored MGD-type methods in Section C, and comparisons with PMGD and MGD.

Converge to the same unique PPS point. We examine the theory result in Lemma C.4. First, we use the same synthetic two-dimensional three-objective affine-gradient PMOO instance in Section 5 with diagonal gradients $g_i(\theta) = A_i\theta - c_i$ and

$$A_1 = \begin{pmatrix} 1.5 & 0 \\ 0 & 0.95 \end{pmatrix}, \; A_2 = \begin{pmatrix} 1.2425 & 0.1052 \\ 0.1052 & 1.1075 \end{pmatrix}, A_3 = \begin{pmatrix} 1.3472 & -0.4188 \\ -0.4188 & 1.2028 \end{pmatrix},$$

$$c_1 = (-1, 0.2), c_2 = (0.75, -0.9), c_3 = (1.375, 0.85).$$

The anchor $\bar{\theta}$ is set as $(0.5, 0.3, 0.2)$. Therefore, $\theta^* = (0, 0)$. Since each A_i is positive definite, this fixed point is unique. Based on this, WGD, AMGD, PAMGD iterate 300 rounds. As shown in Figure 5(Left), they all converge to the same $\theta^* = (0, 0)$ with different initializations, which verifies Lemma C.4.

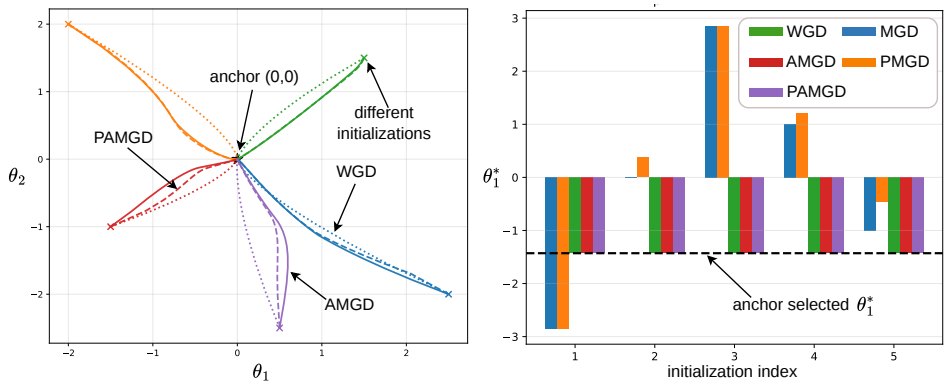


Figure 5: Left: WGD, AMGD,PAMGD converge to the same PPS point for different initializations. Different initializations correspond to different line color. Right: With different initializations, WGD, AMGD, PAMGD converge to the same PPS point while MGD and PMGD converge to different PPS points.

Next, we consider the case that PPS set is not singleton. We construct the same synthetic two-objective instance in Section 5 with $g_1(\theta_1, \theta_2) = (\kappa\theta_1 + 1, \theta_2)$ and $g_2(\theta_1, \theta_2) = (\kappa\theta_1 - 1, \theta_2)$, where $\kappa = 0.35$. The PPS set is $\mathcal{S} = \{(x, 0) : x \in [-1/\kappa, 1/\kappa]\}$ which is not singleton. The anchor $\bar{\lambda} = (0.75, 0.25)$. By its definition, $0.75(\kappa\theta_1^*(\bar{\lambda}) + 1) + 0.25(\kappa\theta_1^*(\bar{\lambda}) - 1) = 0$, therefore $\theta_1^*(\bar{\lambda}) = -\frac{10}{7} \approx -1.43$. As shown in Figure 5(Right), under different initializations, MGD and MGDA converge to different PPS points, while WGD, AMGD and PAMGD converge to the same PPS point $\theta^* = (-1.43, 0)$. We plot only θ_1^* because under this setting, $\theta_2^* = 0$.

Experiments on the Credit dataset. We adopt the same setting as the credit experiment in Section 5 and we set the update step size to 0.3. The anchor $\bar{\lambda}$ is set to $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Moreover, we increase the deployment-induced distribution shift by scaling response strength with s. $s \in \{1, 800, 2000, 5000\}$. As shown in the first and second column of Figure 6, all algorithms diverge and the parameter norm grows substantially as s grows. Notably, the anchor in WGD, AMGD and PAMGD exerts a profound dynamic influence on the multiplier, which leads AMGD and PAMGD approximate to WGD with a fixed preference vector $\bar{\lambda}$. Similar to the results in Section 5, by comparing the last two columns of Figure 6, the multiplier dynamics are more complex as s increases. This indicates again that environmental drift and multiplier drift are coupled in PMOO and the stronger performative effect destabilizes the multiplier dynamic.

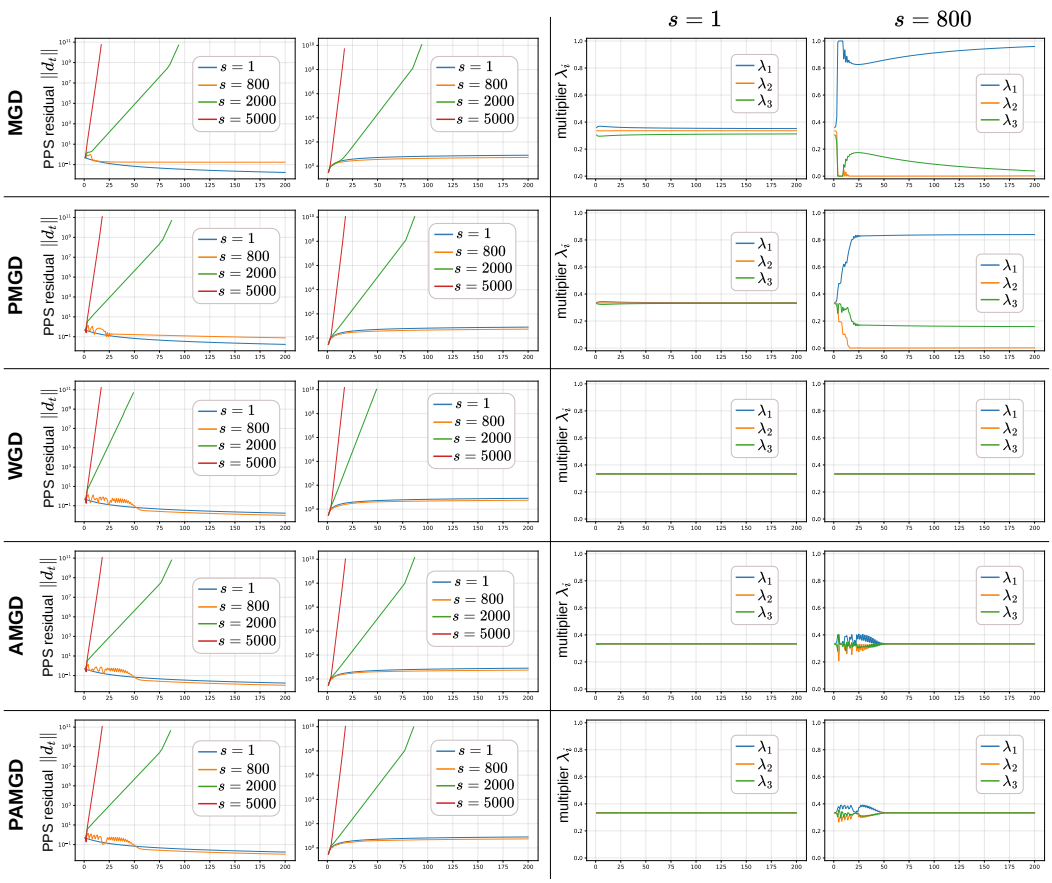


Figure 6: Each row in this figure represents a different algorithm. The figures in the first column represent the PPS residual $\|d_t\|$ with different response strength $s \in \{1, 800, 2000, 5000\}$. The figures in the second column represent the parameter drift norm defined as $\frac{\|\theta_t - \theta_0\|}{\max\{\|\theta_1 - \theta_0\|, 10^{-12}\}}$ for different algorithms. The figures in the third and last column are the multiplier λ_t during iterations with $s = 1, s = 800$, respectively.

E Additional preliminaries and proofs

E.1 Proof of Proposition 2.1

Proof. Introduce an auxiliary scalar variable α and rewrite Eq. equation 2 as

$$\min_{d \in \mathbb{R}^d,\, \alpha \in \mathbb{R}} \quad \alpha + \frac{1}{2}\|d\|^2 \quad \text{subject to} \quad \langle g_i, d\rangle \le \alpha \quad \forall i \in [m]. \tag{6}$$

This is a convex quadratic program with affine constraints. Its Lagrangian is

$$\mathcal{L}(d, \alpha, \lambda) := \alpha + \frac{1}{2}\|d\|^2 + \sum_{i=1}^{m} \lambda_i \big(\langle g_i, d\rangle - \alpha\big), \qquad \lambda_i \ge 0.$$

Collecting terms gives

$$\mathcal{L}(d, \alpha, \lambda) = \alpha\Big(1 - \sum_{i=1}^{m} \lambda_i\Big) + \frac{1}{2}\|d\|^2 + \Big\langle \sum_{i=1}^{m} \lambda_i g_i, d \Big\rangle.$$

For dual-feasible multipliers $\lambda \ge 0$, the infimum over $\alpha \in \mathbb{R}$ is finite only if

$$\sum_{i=1}^{m} \lambda_i = 1.$$

Thus the effective dual domain is Δ_m. Under this constraint, minimizing over d gives the first-order stationarity condition

$$d + \sum_{i=1}^{m} \lambda_i g_i = 0,$$

namely

$$d = -G\lambda.$$

Substituting this minimizer into the Lagrangian yields

$$\inf_{d,\alpha} \mathcal{L}(d, \alpha, \lambda) = -\frac{1}{2} \|G\lambda\|^2 .$$

Therefore the dual problem is

$$\max_{\lambda \in \Delta_m} -\frac{1}{2} \|G\lambda\|^2 .$$

Since the primal problem is convex and Slater's condition holds, e.g. by choosing any d and $\alpha > \max_i \langle g_i, d \rangle$, strong duality holds and the KKT conditions are necessary and sufficient. Therefore every primal–dual optimal pair satisfies the stationarity condition with respect to d, which gives $d^\star = -G\lambda^\star$.

Finally, the objective in Eq. equation 2 is the sum of the convex function $d \mapsto \max_i \langle g_i, d \rangle$ and the strictly convex function $d \mapsto \frac{1}{2} \|d\|^2$, hence it is itself strictly convex. Therefore the primal optimizer $d^\star$ is unique. □

Remark E.1. The dual variable λ is a simplex multiplier representation of the unique MGDA direction $d^\star$. Thus, nonuniqueness of λ does not contradict uniqueness of $d^\star$: different multipliers may represent the same direction whenever the map $\lambda \mapsto G\lambda$ is degenerate on the simplex.

E.2 Proof of Lemma 2.3

Proof. The primal MGDA subproblem

$$\min_d \max_i \langle g_i, d \rangle + \frac{1}{2} \|d\|^2$$

is equivalent to the quadratic program

$$\min_{d,\alpha} \alpha + \frac{1}{2} \|d\|^2 \quad \text{s.t.} \quad \langle g_i, d \rangle \le \alpha \qquad \forall i \in [m].$$

Its Lagrangian is

$$\mathcal{L}(d, \alpha, \lambda) = \alpha + \frac{1}{2} \|d\|^2 + \sum_{i=1}^{m} \lambda_i (\langle g_i, d \rangle - \alpha),$$

with multipliers $\lambda_i \ge 0$. Stationarity with respect to (d, α) gives

$$d + \sum_{i=1}^{m} \lambda_i g_i = 0, \qquad 1 - \sum_{i=1}^{m} \lambda_i = 0,$$

which proves the existence of $\lambda \in \Delta_m$ such that $d = -\sum_i \lambda_i g_i$.

Let $\alpha^\star := \max_i \langle g_i, d \rangle$. Since $(d, \alpha^\star)$ is feasible and $(0, 0)$ is feasible as well, optimality yields

$$\alpha^\star + \frac{1}{2} \|d\|^2 \le 0,$$

hence $\alpha^\star \le -\frac{1}{2} \|d\|^2$. Because every constraint satisfies $\langle g_i, d \rangle \le \alpha^\star$, we obtain

$$\langle g_i, d \rangle \le -\frac{1}{2} \|d\|^2 \qquad \forall i \in [m].$$

It remains to prove the convex-hull characterization. If $d = 0$, then the KKT representation gives $\sum_i \lambda_i g_i = 0$ for some $\lambda \in \Delta_m$, hence $0 \in \mathrm{conv}\{g_1, \ldots, g_m\}$. Conversely, if $0 \in \mathrm{conv}\{g_1, \ldots, g_m\}$, then there exists $\lambda \in \Delta_m$ with $\sum_i \lambda_i g_i = 0$. For any $d \in \mathbb{R}^d$,

$$\max_i \langle g_i, d \rangle \geq \sum_{i=1}^{m} \lambda_i \langle g_i, d \rangle = 0,$$

so the MGDA objective is always at least $\frac{1}{2}\|d\|^2 \geq 0$. The choice $d = 0$ attains value 0, hence it is optimal. □

E.3 Proof of Theorem 3.4

Proof. Fix any objective index $i \in [m]$. Let $g_i^t := \nabla_1 J_i(\theta_t; \theta_t)$. By the smoothness (Assumption 3.2), the descent lemma implies that for every fixed deployed model θ_t,

$$\begin{aligned} J_i(\theta_{t+1}; \theta_t) &\leq J_i(\theta_t; \theta_t) + \eta \langle g_i^t, d_t \rangle + \frac{L_i \eta^2}{2} \|d_t\|^2 \\ &\leq J_i(\theta_t; \theta_t) - \frac{\eta}{2} \|d_t\|^2 + \frac{L\eta^2}{2} \|d_t\|^2 \\ &= J_i(\theta_t; \theta_t) - \frac{\eta(1 - L\eta)}{2} \|d_t\|^2, \end{aligned}$$

where we used the MGD update $\theta_{t+1} = \theta_t + \eta d_t$ and the second inequality uses Lemma 2.3. Notice that the Kantorovich–Rubinstein dual representation of W_1 implies that for every fixed $\tilde{\theta}$,

$$\left|J_i(\tilde{\theta}; \theta) - J_i(\tilde{\theta}; \theta')\right| = \left|\mathbb{E}_{Z \sim D(\theta)}[\ell_i(\tilde{\theta}; Z)] - \mathbb{E}_{Z \sim D(\theta')}[\ell_i(\tilde{\theta}; Z)]\right| \leq G_i W_1(D(\theta), D(\theta')).$$

Using the assumed ε-sensitivity of D with $\kappa_i = \varepsilon G_i$ gives the following,

$$\left|J_i(\tilde{\theta}; \theta) - J_i(\tilde{\theta}; \theta')\right| \leq \kappa_i \|\theta - \theta'\| \qquad \forall \tilde{\theta}, \theta, \theta'.$$

Applying this bound at $\tilde{\theta} = \theta_{t+1}$ gives

$$J_i(\theta_{t+1}; \theta_{t+1}) \leq J_i(\theta_{t+1}; \theta_t) + \kappa_i \|\theta_{t+1} - \theta_t\| = J_i(\theta_{t+1}; \theta_t) + \kappa_i \eta \|d_t\|.$$

Combining the last two displays gives

$$J_i(\theta_{t+1}; \theta_{t+1}) \leq J_i(\theta_t; \theta_t) - c_\eta \|d_t\|^2 + \kappa_i \eta \|d_t\|,$$

where

$$c_\eta := \frac{\eta(1 - L\eta)}{2} > 0.$$

Moreover, notice that

$$\frac{c_\eta}{2} \|d_t\|^2 + \frac{\kappa_i^2 \eta^2}{2c_\eta} \geq 2\sqrt{\frac{c_\eta}{2} \|d_t\|^2 \frac{\kappa_i^2 \eta^2}{2c_\eta}} = \kappa_i \eta \|d_t\|.$$

Therefore,

$$\frac{c_\eta}{2} \|d_t\|^2 \leq J_i(\theta_t; \theta_t) - J_i(\theta_{t+1}; \theta_{t+1}) + \frac{\kappa_i^2 \eta^2}{2c_\eta}.$$

Summing from $t = 0$ to $T - 1$ and using the lower bound $J_i(\theta_T; \theta_T) \geq \underline{J}_i$ yields

$$\frac{c_\eta}{2} \sum_{t=0}^{T-1} \|d_t\|^2 \leq J_i(\theta_0; \theta_0) - \underline{J}_i + \frac{\kappa_i^2 \eta^2}{2c_\eta} T.$$

Since $\min_{t<T} \|d_t\|^2 \leq \frac{1}{T} \sum_{t=0}^{T-1} \|d_t\|^2$, we obtain

$$\min_{0 \leq t \leq T-1} \|d_t\|^2 \leq \frac{2\big(J_i(\theta_0; \theta_0) - \underline{J}_i\big)}{c_\eta T} + \frac{\kappa_i^2 \eta^2}{c_\eta^2}.$$

Substituting $c_\eta = \eta(1 - L\eta)/2$ gives exactly equation 4. □

E.4 Proof of Corollary 3.5

Proof. By assumption, there exists $i \in [m]$ such that $J_i(\hat{\theta}; \theta) = J_i(\hat{\theta}; \theta')$. Fix i and let $g_i^t := \nabla_1 J_i(\theta_t; \theta_t)$, similar to the proof in Section E.3, we can get that for every fixed deployed model θ_t,

$$J_i(\theta_{t+1}; \theta_{t+1}) = J_i(\theta_{t+1}; \theta_t) \leq J_i(\theta_t; \theta_t) - \frac{\eta(1 - L\eta)}{2} \|d_t\|^2 .$$

Therefore,

$$\frac{\eta(1 - L\eta)}{2} \sum_{t=0}^{T-1} \|d_t\|^2 \leq J_i(\theta_0; \theta_0) - \underline{J}_i$$

$$\min_{0 \leq t \leq T-1} \|d_t\|^2 \leq \frac{2\big(J_i(\theta_0; \theta_0) - \underline{J}_i\big)}{T\big(\eta(1 - L\eta)\big)} = \mathcal{O}(\frac{1}{T})$$

which means MGD returns an $\mathcal{O}(\frac{1}{\sqrt{T}})$-PPS iterate. □

E.5 Proof of Proposition 4.3

Proposition 4.3. Under Assumption 4.1, every PPS point $\theta^\star \in \Theta$ admits a multiplier $\lambda^\star \in \Delta_m$ such that $\theta^\star \in \arg\min_{\theta \in \Theta} \sum_{i=1}^m \lambda_i^\star J_i(\theta; \theta^\star)$.

Proof. Since $\theta^\star$ is a PPS point, by the first-order Pareto stationarity condition there exists a multiplier $\lambda^\star \in \Delta_m$ such that

$$\sum_{i=1}^m \lambda_i^\star g_i(\theta^\star) = 0.$$

Equivalently, using the definition of g_i,

$$\sum_{i=1}^m \lambda_i^\star \nabla_1 J_i(\theta^\star; \theta^\star) = 0.$$

Now fix the environment at $\theta^\star$ and define the scalarized decoupled objective

$$\Phi_{\lambda^\star}(\theta; \theta^\star) := \sum_{i=1}^m \lambda_i^\star J_i(\theta; \theta^\star), \qquad \theta \in \Theta.$$

Because $\lambda^\star \in \Delta_m$, we have $\lambda_i^\star \geq 0$ for all i and $\sum_{i=1}^m \lambda_i^\star = 1$. Under Assumption 4.1, each function $J_i(\cdot; \theta^\star)$ is convex on Θ. Hence $\Phi_{\lambda^\star}(\cdot; \theta^\star)$ is also convex on Θ as a nonnegative weighted sum of convex functions.

Moreover, differentiating $\Phi_{\lambda^\star}$ with respect to its first argument gives

$$\begin{aligned}
\nabla_\theta \Phi_{\lambda^\star}(\theta^\star; \theta^\star) &= \nabla_\theta \left(\sum_{i=1}^m \lambda_i^\star J_i(\theta; \theta^\star) \right) \Big|_{\theta = \theta^\star} \\
&= \sum_{i=1}^m \lambda_i^\star \nabla_1 J_i(\theta^\star; \theta^\star) \\
&= 0.
\end{aligned}$$

By convexity of $\Phi_{\lambda^\star}(\cdot; \theta^\star)$, for any $\theta \in \Theta$,

$$\begin{aligned}
\Phi_{\lambda^\star}(\theta; \theta^\star) &\geq \Phi_{\lambda^\star}(\theta^\star; \theta^\star) + \langle \nabla_\theta \Phi_{\lambda^\star}(\theta^\star; \theta^\star), \theta - \theta^\star \rangle \\
&= \Phi_{\lambda^\star}(\theta^\star; \theta^\star).
\end{aligned}$$

Therefore,

$$\Phi_{\lambda^\star}(\theta^\star; \theta^\star) \leq \Phi_{\lambda^\star}(\theta; \theta^\star), \qquad \forall \theta \in \Theta.$$

That is,

$$\theta^\star \in \arg\min_{\theta \in \Theta} \Phi_{\lambda^\star}(\theta; \theta^\star) = \arg\min_{\theta \in \Theta} \sum_{i=1}^m \lambda_i^\star J_i(\theta; \theta^\star).$$

This completes the proof. □

E.6 Proof of Theorem 4.5

Proof. We first prove the conclusion for PMGD, i.e. for a fixed $\tau \geq 0$. The MGD case will then follow by taking $\tau = 0$.

PMGD case. Define the proximal multiplier update as

$$M_\tau(\theta, \lambda) = \operatorname*{arg\,min}_{\mu \in \Delta_m} \left\{ \frac{1}{2}\mu^\top H(\theta)\mu + \frac{\tau}{2}\|\mu - \lambda\|^2 \right\}.$$

Therefor, the PMGD update can be written as

$$\lambda_{t+1} \in M_\tau(\theta_t, \lambda_t), \qquad \theta_{t+1} = \theta_t - \eta G(\theta_t)\lambda_{t+1}.$$

By Lemma E.2, the PPS fixed-point graph $\mathcal{M}$ and the PPS set $\mathcal{S}$ are nonempty. Moreover, for the 2*2 matrix A_s, its spectral radius $\rho(A_s) = \frac{r+c\frac{L_H}{\sigma_s+\tau}+\frac{\tau}{\sigma_s+\tau}+\sqrt{(r+c\frac{L_H}{\sigma_s+\tau}+\frac{\tau}{\sigma_s+\tau})^2-\frac{4r\tau}{\sigma_s+\tau}}}{2}$. It's easy to check that $\rho(A_s) < 1$ is equivalent to $r + c\frac{L_H}{\sigma_s} < 1$.

First, we prove the proximal multiplier stability relative to a fixed PPS set. Fix an arbitrary $\bar{\theta} \in \mathcal{S}$, and define

$$y_t(\bar{\theta}) := \operatorname{dist}(\lambda_t, \mathcal{F}(\bar{\theta})) := \min_{o \in \mathcal{F}(\bar{\theta})} \|\lambda_t - o\|.$$

We prove that $y_{t+1}(\bar{\theta}) \leq \frac{L_H}{\sigma_s+\tau}\|\theta_t - \bar{\theta}\| + \frac{\tau}{\sigma_s+\tau} y_t(\bar{\theta})$. Let

$$\mu := \lambda_{t+1} \in M_\tau(\theta_t, \lambda_t),$$

and choose $\bar{\lambda} \in \arg\min_{o \in \mathcal{F}(\bar{\theta})} \|\mu - o\|$. Set

$$u := \mu - \bar{\lambda}.$$

Since $\bar{\lambda} \in \mathcal{F}(\bar{\theta})$, we have $\bar{\lambda} \in \Lambda(\bar{\theta})$ and $G(\bar{\theta})\bar{\lambda} = 0$. By Assumption 4.4,

$$u^\top H(\bar{\theta})u \geq \sigma_s \|u\|^2. \tag{7}$$

Since $\bar{\lambda} \in \Lambda(\bar{\theta})$, the variational inequality for the convex problem defining $\Lambda(\bar{\theta})$ gives

$$\nabla_\lambda \frac{1}{2}\lambda^T H(\bar{\theta})\lambda\big|_{\lambda=\bar{\lambda}} = H(\bar{\theta})\bar{\lambda},\ \langle H(\bar{\theta})\bar{\lambda}, v - \bar{\lambda}\rangle \geq 0, \qquad \forall v \in \Delta_m.$$

Taking $v = \mu$, we obtain $u^\top H(\bar{\theta})\bar{\lambda} \geq 0$. Hence

$$u^\top H(\bar{\theta})u = u^\top H(\bar{\theta})\mu - u^\top H(\bar{\theta})\bar{\lambda} \leq u^\top H(\bar{\theta})\mu. \tag{8}$$

On the other hand, because $\mu \in M_\tau(\theta_t, \lambda_t)$, the first-order optimality condition in variational-inequality form is

$$\langle H(\theta_t)\mu + \tau(\mu - \lambda_t), v - \mu\rangle \geq 0, \qquad \forall v \in \Delta_m.$$

Taking $v = \bar{\lambda}$, we get

$$u^\top H(\theta_t)\mu + \tau u^\top(\mu - \lambda_t) \leq 0. \tag{9}$$

Therefore,

$$\begin{aligned} u^\top H(\bar{\theta})\mu &= u^\top \big(H(\bar{\theta}) - H(\theta_t)\big)\mu + u^\top H(\theta_t)\mu \\ &\leq u^\top \big(H(\bar{\theta}) - H(\theta_t)\big)\mu - \tau u^\top(\mu - \lambda_t) \\ &= u^\top \big(H(\bar{\theta}) - H(\theta_t)\big)\mu - \tau\|u\|^2 + \tau u^\top(\lambda_t - \bar{\lambda}), \end{aligned} \tag{10}$$

since $\mu - \lambda_t = u - (\lambda_t - \bar{\lambda})$. Let

$$\lambda_t^F \in \operatorname*{arg\,min}_{o \in \mathcal{F}(\bar{\theta})} \|\lambda_t - o\|.$$

Notice that $\mathcal{F}(\bar{\theta}) = \triangle_m \cap \{\lambda : G(\bar{\theta})\lambda = 0\} = \triangle_m \cap \ker G(\bar{\theta})$ and it's easy to check that $\triangle_m, \ker G(\bar{\theta})$ are both closed convex sets, therefore their intersection $\mathcal{F}(\bar{\theta})$ is a closed convex set. Because $\bar{\lambda}$ is the Euclidean projection of μ onto the closed convex set $\mathcal{F}(\bar{\theta})$, the projection inequality gives

$$\langle \mu - \bar{\lambda}, \lambda_t^F - \bar{\lambda}\rangle \leq 0 \ \Rightarrow\ u^\top(\lambda_t^F - \bar{\lambda}) \leq 0.$$

Consequently,

$$\begin{aligned} u^\top(\lambda_t - \bar{\lambda}) &= u^\top(\lambda_t - \lambda_t^F) + u^\top(\lambda_t^F - \bar{\lambda}) \\ &\le \|u\| \, \|\lambda_t - \lambda_t^F\| \\ &= \|u\| \, \operatorname{dist}(\lambda_t, \mathcal{F}(\bar{\theta})). \end{aligned} \tag{11}$$

Moreover, $\mu \in \Delta_m$, so $\|\mu\| \le 1$. Using the Lipschitz continuity of H by Lemma E.5,

$$\|H(\bar{\theta}) - H(\theta_t)\| \le L_H \|\theta_t - \bar{\theta}\|.$$

Combining Eq. equation 7-equation 11, we obtain

$$\begin{aligned} \sigma_s \|u\|^2 &\le u^\top H(\bar{\theta}) u \le u^\top H(\bar{\theta}) \mu \\ &\le L_H \|u\| \, \|\theta_t - \bar{\theta}\| - \tau \|u\|^2 + \tau \|u\| \operatorname{dist}(\lambda_t, \mathcal{F}(\bar{\theta})). \end{aligned}$$

If $u = 0$, the desired bound is trivial. Otherwise, dividing by $\|u\|$ yields

$$\|u\| \le \frac{L_H}{\sigma_s + \tau} \|\theta_t - \bar{\theta}\| + \frac{\tau}{\sigma_s + \tau} \operatorname{dist}(\lambda_t, \mathcal{F}(\bar{\theta})).$$

Since $\|u\| = \left\|\lambda_{t+1} - \bar{\lambda}\right\| = \min_{o \in \mathcal{F}(\bar{\theta})} \|\lambda_{t+1} - o\| = \operatorname{dist}(\lambda_{t+1}, \mathcal{F}(\bar{\theta}))$ by definitions, this proves

$$y_{t+1}(\bar{\theta}) \le \frac{L_H}{\sigma_s + \tau} \|\theta_t - \bar{\theta}\| + \frac{\tau}{\sigma_s + \tau} y_t(\bar{\theta}). \tag{12}$$

Next, we fix the same $\bar{\theta} \in \mathcal{S}$. Let

$$\bar{\lambda}_{t+1} \in \operatorname*{arg\,min}_{o \in \mathcal{F}(\bar{\theta})} \|\lambda_{t+1} - o\| .$$

Because $\bar{\lambda}_{t+1} \in \mathcal{F}(\bar{\theta})$, we have

$$G(\bar{\theta}) \bar{\lambda}_{t+1} = 0,$$

and hence

$$\bar{\theta} = \bar{\theta} - \eta G(\bar{\theta}) \bar{\lambda}_{t+1}.$$

Then

$$\begin{aligned} \|\theta_{t+1} - \bar{\theta}\| &= \|\theta_t - \eta G(\theta_t) \lambda_{t+1} - \bar{\theta}\| \\ &\le \|\theta_t - \eta G(\theta_t) \bar{\lambda}_{t+1} - \bar{\theta} + \eta G(\bar{\theta}) \bar{\lambda}_{t+1}\| + \eta \|G(\theta_t)(\lambda_{t+1} - \bar{\lambda}_{t+1})\| \end{aligned}$$

The first term is bounded by the fixed-weight contraction in Lemma E.4:

$$\|\theta_t - \eta G(\theta_t) \bar{\lambda}_{t+1} - \bar{\theta} + \eta G(\bar{\theta}) \bar{\lambda}_{t+1}\| \le r \|\theta_t - \bar{\theta}\|.$$

The second term is bounded by

$$\eta \|G(\theta_t)(\lambda_{t+1} - \bar{\lambda}_{t+1})\| \le c \operatorname{dist}(\lambda_{t+1}, \mathcal{F}(\bar{\theta})) = c y_{t+1}(\bar{\theta}).$$

Using Eq. equation 12, we obtain

$$\|\theta_{t+1} - \bar{\theta}\| \le (r + c \frac{L_H}{\sigma_s + \tau}) \|\theta_t - \bar{\theta}\| + c \frac{\tau}{\sigma_s + \tau} y_t(\bar{\theta}). \tag{13}$$

We next show that the PPS point is unique under the contraction condition. Suppose $\theta_a, \theta_b \in \mathcal{S}$. Consider MGD ($\tau = 0$) initialized at $\theta_0 = \theta_a$. Since θ_a is PPS, its MGDA direction is 0, and hence $\theta_t = \theta_a$ for all t. Applying Eq. (13) with $\bar{\theta} = \theta_b$ gives

$$\|\theta_a - \theta_b\| \le (r + c \frac{L_H}{\sigma_s}) \|\theta_a - \theta_b\|,$$

which means $\theta_a = \theta_b$. Therefore $\mathcal{S} = \{\theta^*\}$.

Together, Eq. equation 12 and equation 13 give

$$\begin{pmatrix} \|\theta_{t+1} - \bar{\theta}\| \\ y_{t+1}(\bar{\theta}) \end{pmatrix} \le A_s \begin{pmatrix} \|\theta_t - \bar{\theta}\| \\ y_t(\bar{\theta}) \end{pmatrix}. \tag{14}$$

Since A_s is a positive matrix for $\tau > 0$, the Perron–Frobenius theorem implies that its spectral radius $\rho(A_s)$ is a simple positive eigenvalue and admits an eigenvector $e = (e_1, e_2) \in \mathbb{R}^2$ with $e_1, e_2 > 0$ such that $eA_s = \rho(A_s)e$. Next, we define the weighted norm on $\Theta \times \triangle_m$ and its induced distance to the PPS set as following

$$\|(\theta, \lambda)\|_e := e_1 \|\theta\| + e_2 \|\lambda\|, \; D_e((\theta, \lambda), \mathcal{M}) := \inf_{(\hat{\theta}, \hat{\lambda}) \in \mathcal{M}} \left\| (\theta - \hat{\theta}, \lambda - \hat{\lambda}) \right\|_e .$$

It's easy to check that $D_e((\theta, \lambda), \mathcal{M}) := \inf_{\bar{\theta} \in \mathcal{S}} \left\{ e_1 \|\theta - \bar{\theta}\| + e_2 \operatorname{dist}(\lambda, \mathcal{F}(\bar{\theta})) \right\}$. For any fixed $\bar{\theta} \in \mathcal{S}$, the one-step comparison (Eq. equation 14) gives

$$\begin{pmatrix} \|\theta_{t+1} - \bar{\theta}\| \\ \operatorname{dist}(\lambda_{t+1}, \mathcal{F}(\bar{\theta})) \end{pmatrix} \leq A_s \begin{pmatrix} \|\theta_t - \bar{\theta}\| \\ \operatorname{dist}(\lambda_t, \mathcal{F}(\bar{\theta})) \end{pmatrix}.$$

Multiplying both sides by e yields

$$e_1 \|\theta_{t+1} - \bar{\theta}\| + e_2 \operatorname{dist}(\lambda_{t+1}, \mathcal{F}(\bar{\theta})) \leq \rho(A_s) \left[e_1 \|\theta_t - \bar{\theta}\| + e_2 \operatorname{dist}(\lambda_t, \mathcal{F}(\bar{\theta})) \right].$$

Taking the inf over $\bar{\theta} \in \mathcal{S}$ gives

$$D_e((\theta_{t+1}, \lambda_{t+1}), \mathcal{M}) \leq \rho(A_s) D_e((\theta_t, \lambda_t), \mathcal{M}).$$

Hence

$$D_e((\theta_t, \lambda_t), \mathcal{M}) \leq \rho(A_s)^t D_e((\theta_0, \lambda_0), \mathcal{M}).$$

Since $e_1, e_2 > 0$, the weighted norm is equivalent to the Euclidean 2-norm which means that there exists constants $C_1, C_2 > 0$ such that for any $(\theta, \lambda), (\theta', \lambda') \in \Theta \times \triangle_m$,

$$C_1 \|(\theta, \lambda) - (\theta', \lambda')\| \leq e_1 \|\theta - \theta'\| + e_2 \|\lambda - \lambda'\| \leq C_2 \|(\theta, \lambda) - (\theta', \lambda')\|.$$

Taking the inf over $(\theta', \lambda') \in \mathcal{M}$ and we can get

$$\begin{aligned} &\inf_{\omega \in \mathcal{M}} \|(\theta_t, \lambda_t) - \omega\| \leq \frac{1}{C_1} D_e((\theta_t, \lambda_t), \mathcal{M}) \\ &\leq \frac{1}{C_1} \rho(A_s)^t D_e((\theta_0, \lambda_0), \mathcal{M}) \leq \frac{C_2}{C_1} \rho(A_s)^t \inf_{\omega \in \mathcal{M}} \|(\theta_0, \lambda_0) - \omega\| = \mathcal{O}\big(\rho(A_s)^t\big). \end{aligned} \tag{15}$$

Further, we prove that $\exists\, \theta_1^* \in \mathcal{S}$ such that θ_t converges to θ_1^* and $\|\theta_t - \theta_1^*\| = \mathcal{O}\big(\rho(A_s)^t\big)$. We first show that $\sum_t \|\theta_{t+1} - \theta_t\| < \infty$. For any $\bar{\theta}_t \in \mathcal{S}$, and choose

$$\bar{\lambda}_{t+1} \in \operatorname*{arg\,min}_{o \in \mathcal{F}(\bar{\theta}_t)} \|\lambda_{t+1} - o\|.$$

Since $G(\bar{\theta}_t)\bar{\lambda}_{t+1} = 0$,

$$G(\theta_t)\lambda_{t+1} = (G(\theta_t) - G(\bar{\theta}_t))\bar{\lambda}_{t+1} + G(\theta_t)(\lambda_{t+1} - \bar{\lambda}_{t+1}).$$

Therefore, by Lemma E.4 $\left\|G(\theta_t)\bar{\lambda}_{t+1} - G(\bar{\theta}_t)\bar{\lambda}_{t+1}\right\| \leq \Gamma \left\|\theta_t - \bar{\theta}_t\right\|$ and $\|G(\theta_t)\| \leq \sqrt{m} G_\infty$ by its definition,

$$\|G(\theta_t)\lambda_{t+1}\| \leq \Gamma \|\theta_t - \bar{\theta}_t\| + \sqrt{m} G_\infty \operatorname{dist}(\lambda_{t+1}, \mathcal{F}(\bar{\theta}_t)).$$

Since $\theta_{t+1} - \theta_t = -\eta G(\theta_t)\lambda_{t+1}$, for any $\bar{\theta}_t \in \mathcal{S}$

$$\|\theta_{t+1} - \theta_t\| \leq \eta\Gamma \left\|\theta_t - \bar{\theta}_t\right\| + c \operatorname{dist}(\lambda_{t+1}, \mathcal{F}(\bar{\theta}_t)), \quad c = \eta\sqrt{m} G_\infty. \tag{16}$$

By the definition of D_e, there exists $\bar{\theta}_t \in \mathcal{S}$ such that for any $\delta > 0$,

$$e_1 \left\|\theta_t - \bar{\theta}_t\right\| \leq D_e((\theta_t, \lambda_t), \mathcal{M}) + \delta, \; e_2 \operatorname{dist}(\lambda_t, \mathcal{F}(\bar{\theta}_t)) \leq D_e((\theta_t, \lambda_t), \mathcal{M}) + \delta. \tag{17}$$

Let the $\bar{\theta} = \bar{\theta}_t$ in Eq. equation 12 and notice that $y_t(\bar{\theta}) = y_t(\bar{\theta}_t) = \operatorname{dist}(\lambda_t, \mathcal{F}(\bar{\theta}_t))$ by y_t's definition, we can get that for any $\delta > 0$

$$\begin{aligned} y_{t+1}(\bar{\theta}_t) = \operatorname{dist}(\lambda_{t+1}, \mathcal{F}(\bar{\theta}_t)) &\leq \frac{L_H}{\sigma_s + \tau} \left\|\theta_t - \bar{\theta}_t\right\| + \frac{\tau}{\sigma_s + \tau} \operatorname{dist}(\lambda_t, \mathcal{F}(\bar{\theta}_t)) \quad \text{by Eq. equation 12} \\ &\leq \Big(\frac{L_H}{(\sigma_s + \tau)e_1} + \frac{\tau}{(\sigma_s + \tau)e_2} \Big) \big(D_e((\theta_t, \lambda_t), \mathcal{M}) + \delta \big) \quad \text{by Eq. equation 17} \\ \Rightarrow \quad y_{t+1}(\bar{\theta}_t) &\leq \Big(\frac{L_H}{(\sigma_s + \tau)e_1} + \frac{\tau}{(\sigma_s + \tau)e_2} \Big) D_e((\theta_t, \lambda_t), \mathcal{M}) \quad \text{by taking } \delta \to 0. \end{aligned} \tag{18}$$

Substitute Eq. equation 17-equation 18 into Eq. equation 16 and we can get for any $\delta > 0$,

$$\sum_{t=0}^{\infty} \|\theta_{t+1} - \theta_t\| \leq \sum_{t=0}^{\infty} \Big(\frac{\eta\Gamma}{e_1} D_e((\theta_t, \lambda_t), \mathcal{M}) + \big(\frac{cL_H}{(\sigma_s + \tau)e_1} + \frac{c\tau}{(\sigma_s + \tau)e_2} \big) D_e((\theta_t, \lambda_t), \mathcal{M}) \Big)$$

$$= \sum_{t=0}^{\infty} \Big(\frac{\eta\Gamma}{e_1} + \frac{cL_H}{(\sigma_s + \tau)e_1} + \frac{c\tau}{(\sigma_s + \tau)e_2} \Big) D_e((\theta_t, \lambda_t), \mathcal{M})$$

$$\leq \Big(\frac{\eta\Gamma}{e_1} + \frac{cL_H}{(\sigma_s + \tau)e_1} + \frac{c\tau}{(\sigma_s + \tau)e_2} \Big) \sum_{t=0}^{\infty} \rho(A_s)^t D_e((\theta_0, \lambda_0), \mathcal{M})$$

$$= \Big(\frac{\eta\Gamma}{e_1} + \frac{cL_H}{(\sigma_s + \tau)e_1} + \frac{c\tau}{(\sigma_s + \tau)e_2} \Big) D_e((\theta_0, \lambda_0), \mathcal{M}) \sum_{t=0}^{\infty} \rho(A_s)^t < \infty \quad (\rho(A_s) < 1).$$

Therefore, $\{\theta_t\}$ is a Cauchy sequence. Since $\Theta \subseteq \mathbb{R}^d$ is closed and $\mathbb{R}^d$ is a complete space, there exists $\theta_1^* \in \Theta$ such that $\theta_t \to \theta_1^*$.

Next, we prove that $\theta_1^* \in \mathcal{S}$. By Eq. equation 15,

$$\inf_{o \in \mathcal{S}} \|\theta_t - o\| \leq \inf_{\omega \in \mathcal{M}} \|(\theta_t, \lambda_t) - \omega\| \leq \mathcal{O}\big(\rho(A_s)^t\big),$$

therefore $\lim_{t \to \infty} \inf_{o \in \mathcal{S}} \|\theta_t - o\| = 0$. Notice that for any $o' \in \mathcal{S}$,

$$\inf_{o \in \mathcal{S}} \|\theta_t - o\| \leq \|\theta_t - o'\| \leq \|\theta_t - \theta_1^*\| + \|\theta_1^* - o'\|$$

$$\Rightarrow \inf_{o \in \mathcal{S}} \|\theta_t - o\| \leq \|\theta_t - \theta_1^*\| + \inf_{o' \in \mathcal{S}} \|\theta_1^* - o'\|.$$

and symmetrically, $\inf_{o' \in \mathcal{S}} \|\theta_1^* - o'\| \leq \|\theta_t - \theta_1^*\| + \inf_{o \in \mathcal{S}} \|\theta_t - o\|$. Therefore,

$$\Big| \inf_{o \in \mathcal{S}} \|\theta_1^* - o\| - \inf_{o \in \mathcal{S}} \|\theta_t - o\| \Big| \leq \|\theta_t - \theta_1^*\| \Rightarrow \inf_{o \in \mathcal{S}} \|\theta_1^* - o\| = \lim_{t \to \infty} \inf_{o \in \mathcal{S}} \|\theta_t - o\| = 0.$$

Based on the definitions of $\mathcal{S}$ and $\mathcal{M}$, it's easy to check that both of them are closed sets, thus $\inf_{o \in \mathcal{S}} \|\theta_1^* - o\| = 0$ implies $\theta_1^* \in \mathcal{S}$. Moreover, similar to previous steps, it's easy to prove that $\|\theta_1^* - \theta_t\| = \mathcal{O}\big(\rho(A_s)^t\big)$.

MGD case. MGD corresponds to $\tau = 0$. In this case

$$M_0(\theta, \lambda) = \Lambda(\theta),$$

so the multiplier update is independent of the previous multiplier. Moreover

$$A_s = \begin{pmatrix} r + c\frac{L_H}{\sigma_s} & 0 \\ \frac{L_H}{\sigma_s} & 0 \end{pmatrix}.$$

The spectral radius is

$$\rho(A_s) = r + c\frac{L_H}{\sigma_s}.$$

Thus $\rho(A_s) < 1$ is exactly

$$r + c\frac{L_H}{\sigma_s} < 1.$$

Similar to the proof in the PMGD result with $\tau = 0$, we can prove the MGD conclusions.

Converge into the same set $\mathcal{S}$. Similar to the proof of Theorem 4.7 and the definition of $\mathcal{S}$ and $\mathcal{M}$, this is proved. □

E.7 Proof of Theorem 4.7

Proof. **PMGD case.** We first prove the conclusion for PMGD. Define the proximal multiplier update map

$$M_\tau(\theta, \lambda) := \operatorname*{arg\,min}_{\mu \in \Delta_m} \left\{ \frac{1}{2} \mu^\top H(\theta) \mu + \frac{\tau}{2} \|\mu - \lambda\|^2 \right\}. \tag{19}$$

By Algorithm 2, let the corresponding PMGD update be

$$T_\tau(\theta,\lambda) := \Big(\theta - \eta G(\theta) M_\tau(\theta,\lambda),\ M_\tau(\theta,\lambda)\Big). \tag{20}$$

Fix $(\theta,\lambda) \in \Theta \times \Delta_m$. The optimization problem in equation 19 has objective

$$\mu \mapsto \frac{1}{2}\mu^\top H(\theta)\mu + \frac{\tau}{2}\|\mu-\lambda\|^2.$$

Because $H(\theta) = G(\theta)G(\theta)^T \succeq 0$ and $\tau > 0$, this objective is strictly convex in μ. Since Δ_m is compact and convex, a unique minimizer exists. Therefore M_τ is single-valued. Next, we prove the Lipschitz continuity of M_τ.

Take arbitrary (θ,λ) and (θ',λ'), and write

$$\mu := M_\tau(\theta,\lambda), \qquad \mu' := M_\tau(\theta',\lambda').$$

By first-order optimality in variational-inequality form,

$$\langle H(\theta)\mu + \tau(\mu-\lambda), \nu-\mu\rangle \geq 0, \qquad \forall \nu \in \Delta_m, \tag{21}$$

and

$$\langle H(\theta')\mu' + \tau(\mu'-\lambda'), \nu-\mu'\rangle \geq 0, \qquad \forall \nu \in \Delta_m. \tag{22}$$

Choose $\nu = \mu'$ in equation 21 and $\nu = \mu$ in equation 22, then add these two together, yielding

$$\langle H(\theta)\mu - H(\theta')\mu', \mu-\mu'\rangle + \tau\|\mu-\mu'\|^2 \leq \tau\langle \lambda-\lambda', \mu-\mu'\rangle. \tag{23}$$

Rewrite the first term as

$$H(\theta)\mu - H(\theta')\mu' = H(\theta)(\mu-\mu') + \big(H(\theta)-H(\theta')\big)\mu'.$$

Substituting this into equation 23, and we can get

$$\langle H(\theta)(\mu-\mu'), \mu-\mu'\rangle + \tau\|\mu-\mu'\|^2 \leq \big\|\big(H(\theta)-H(\theta')\big)\mu'\big\|\,\|\mu-\mu'\| + \tau\|\lambda-\lambda'\|\,\|\mu-\mu'\|.$$

Notice that $\mu,\mu' \in \triangle_m$ thus $\mathbf{1}^\top(\mu-\mu') = 0$, therefore by Assumption 4.6, the first term on the LHS is bounded below by $\sigma\|\mu-\mu'\|^2$. Moreover, because $\mu' \in \Delta_m$, $\|\mu'\| \leq 1$, hence

$$\big\|\big(H(\theta)-H(\theta')\big)\mu'\big\| \leq L_H\|\theta-\theta'\|, \quad \text{By Lemma E.5.}$$

Therefore

$$(\sigma+\tau)\|\mu-\mu'\|^2 \leq \Big(L_H\|\theta-\theta'\| + \tau\|\lambda-\lambda'\|\Big)\|\mu-\mu'\|.$$

If $\mu \neq \mu'$, divide by $\|\mu-\mu'\|$; if $\mu = \mu'$, the desired bound is trivial. Thus

$$\|\mu-\mu'\| \leq \frac{L_H}{\sigma+\tau}\|\theta-\theta'\| + \frac{\tau}{\sigma+\tau}\|\lambda-\lambda'\| = p_\tau\|\theta-\theta'\| + q_\tau\|\lambda-\lambda'\|,$$

which proves the Lipschitz continuity of M_τ.

Assume $T_\tau(\theta,\lambda) = (\theta_+,\lambda_+)$ and $T_\tau(\theta',\lambda') = (\theta'_+,\lambda'_+)$, then $u = \lambda_+$ and $u' = \lambda'_+$. Notice that

$$\begin{aligned}
\big\|\theta_+ - \theta'_+\big\| &= \|\theta - \eta G(\theta)\mu - (\theta' - \eta G(\theta')\mu')\| \\
&= \theta - \eta G(\theta)\mu - (\theta' - \eta G(\theta')\mu) - \eta G(\theta')(\mu'-\mu) \\
&\leq \underbrace{\|\theta - \eta G(\theta)\mu - (\theta' - \eta G(\theta')\mu)\|}_{A} + \eta\underbrace{\|G(\theta')(\mu'-\mu)\|}_{B}.
\end{aligned} \tag{24}$$

By Lemma E.4, because $\mu \in \Delta_m$ is fixed,

$$A \leq r\|\theta-\theta'\|.$$

Also,

$$\eta B \leq \eta\|G(\theta')\|_{\mathrm{op}}\|\mu'-\mu\| \leq c\|\mu'-\mu\|.$$

Combining these two bounds with Eq. equation 24 yields $\big\|\theta_+ - \theta'_+\big\| \leq (r+cp_\tau)\|\theta-\theta'\| + cq_\tau\|\lambda-\lambda'\|$. Therefore, in vecotr form we can get that for any $(\theta,\lambda),(\theta',\lambda') \in \Theta \times \triangle_m$,

$\begin{pmatrix} \|\theta_+ - \theta'_+\| \\ \|\lambda_+ - \lambda'_+\| \end{pmatrix} \leq A_\tau \begin{pmatrix} \|\theta - \theta'\| \\ \|\lambda - \lambda'\| \end{pmatrix}$ by the definition of A_τ. Moreover, since $r, p_\tau, c, q_\tau > 0$ and A_τ is a positive matrix, it's easy to check that the spectral radius of A_τ is its maximal positive real eigenvalue denoted as λ_e and $\lambda_e = \rho(A_\tau) = \frac{r+cp_\tau+q_\tau+\sqrt{(r+cp_\tau+q_\tau)^2-4rq_\tau}}{2} > q_\tau > 0$. By the Perron-Frobenius Theorem, there exists $(e_1, e_2) \in \mathbb{R}^2$ with $e_1, e_2 > 0$ as its corresponding eigenvector and $(e_1, e_2)A_\tau = \rho(A_\tau)(e_1, e_2)$. Moreover, it's easy to check that $\rho(A_\tau) < 1$ is equivalent to $r + c\frac{L_H}{\sigma} < 1$.

Consider the norm $\|(\theta, \lambda)\|_e = e_1 \|\theta\| + e_2 \|\lambda\|$ for $(\theta, \lambda) \in \Theta \times \triangle_m$ and the distance induced by this norm D_e, it's easy to check that $(\Theta \times \triangle_m, D_e)$ is a complete space. Also,

$$\begin{aligned} & \begin{pmatrix} \|\theta_+ - \theta'_+\| \\ \|\lambda_+ - \lambda'_+\| \end{pmatrix} \leq A_\tau \begin{pmatrix} \|\theta - \theta'\| \\ \|\lambda - \lambda'\| \end{pmatrix} \Rightarrow (e_1, e_2) \begin{pmatrix} \|\theta_+ - \theta'_+\| \\ \|\lambda_+ - \lambda'_+\| \end{pmatrix} \leq (e_1, e_2) A_\tau \begin{pmatrix} \|\theta - \theta'\| \\ \|\lambda - \lambda'\| \end{pmatrix} \\ \Rightarrow \quad & (e_1, e_2) \begin{pmatrix} \|\theta_+ - \theta'_+\| \\ \|\lambda_+ - \lambda'_+\| \end{pmatrix} \leq \rho(A_\tau)(e_1, e_2) \begin{pmatrix} \|\theta - \theta'\| \\ \|\lambda - \lambda'\| \end{pmatrix} \\ \Rightarrow \quad & D_e(T_\tau(\theta, \lambda), T_\tau(\theta', \lambda')) \leq \rho(A_\tau) D_e((\theta, \lambda), (\theta', \lambda')), \end{aligned}$$

which means T_τ is a contraction on $(\Theta \times \triangle_m, D_e)$. Banach fixed-point theorem implies that T_τ admits a unique fixed point $(\theta^*, \lambda^*) \in \Theta \times \triangle_m$. Moreover, $T_\tau(\theta^*, \lambda^*) = (\theta^*, \lambda^*)$ means that $G(\theta^*)\lambda^* = 0$ and by Lemma 2.3, θ^* is a PPS point.

Finally, take $(\theta', \lambda') = (\theta^*, \lambda^*)$ and we can get $\begin{pmatrix} \|\theta_{t+1} - \theta^*\| \\ \|\lambda_{t+1} - \lambda^*\| \end{pmatrix} \leq A_\tau \begin{pmatrix} \|\theta_t - \theta^*\| \\ \|\lambda_t - \lambda^*\| \end{pmatrix}$. Similar to the proof of Theorem 4.5 in Section E.6, we can get

$$\|\theta_t - \theta^*\| + \|\lambda_t - \lambda^*\| = \mathcal{O}\big(\rho(A_\tau)^t\big)$$

with $\rho(A_\tau) = \frac{r+cp_\tau+q_\tau+\sqrt{(r+cp_\tau+q_\tau)^2-4rq_\tau}}{2}$ which proves the conclusion for PMGD.

MGD case. We now prove the conclusion for MGD. In MGD, $\tau = 0$, therefore $q_\tau = 0$ and $A_\tau = \begin{pmatrix} r + cp_\tau & cq_\tau \\ p_\tau & q_\tau \end{pmatrix} = \begin{pmatrix} r + c\frac{L_H}{\sigma} & 0 \\ \frac{L_H}{\sigma} & 0 \end{pmatrix}$. The spectral radius $\rho(A_\tau) = r + c\frac{L_H}{\sigma}$. Let the multiplier update map in MGD be $M(\theta) = \arg\min_{\mu \in \Delta_m} \left\{ \frac{1}{2} \mu^\top H(\theta) \mu \right\}$ and the corresponding MGD update be $T(\theta) = \theta - \eta G(\theta) M(\theta)$. The key difference between MGD update maps and PMGD update maps is that MGD update maps are unrelated to the multiplier λ. Similar to the proof in PMGD, we can get for any $\theta, \theta' \in \Theta$,

$$\|M(\theta) - M(\theta')\| \leq \frac{L_H}{\sigma} \|\theta - \theta'\|, \ \|T(\theta) - T(\theta')\| \leq (r + c\frac{L_H}{\sigma}) \|\theta - \theta'\|.$$

Since $r + c\frac{L_H}{\sigma} < 1$, T is a contraction on Θ. By Banach fixed-point theorem, there exists a unique fixed point $\theta^* \in \Theta$ for T. Similarly, we can prove θ^* is a PPS point and $\|\theta_t - \theta^*\| = \mathcal{O}\big(\rho(A_\tau)^t\big)$ with $\rho(A_\tau) = r + c\frac{L_H}{\sigma}$.

Converge to the same unique PPS point. First, suppose $(\theta^\star, \lambda^\star)$ is a fixed point of MGD. Then, by definition,

$$\lambda^\star \in \arg\min_{\lambda \in \Delta_m} \frac{1}{2} \lambda^\top H(\theta^\star) \lambda, \qquad G(\theta^\star)\lambda^\star = 0.$$

Consider now the proximal objective centered at $\lambda^\star$:

$$\Psi(\lambda) := \frac{1}{2} \lambda^\top H(\theta^\star) \lambda + \frac{\tau}{2} \|\lambda - \lambda^\star\|^2.$$

Since $\lambda^\star$ minimizes the first term over Δ_m and the second term is nonnegative with value 0 at $\lambda = \lambda^\star$, it follows immediately that $\lambda^\star$ minimizes Ψ over Δ_m. Together with $G(\theta^\star)\lambda^\star = 0$, this exactly means $(\theta^\star, \lambda^\star)$ is a fixed point of PMGD.

Conversely, suppose $(\theta^\star, \lambda^\star)$ is a fixed point of PMGD. Then

$$\lambda^\star = M_\tau(\theta^\star, \lambda^\star), \qquad G(\theta^\star)\lambda^\star = 0.$$

The first relation gives the variational inequality

$$\langle H(\theta^\star)\lambda^\star + \tau(\lambda^\star - \lambda^\star), \nu - \lambda^\star\rangle \geq 0, \qquad \forall \nu \in \Delta_m,$$

which simplifies to

$$\langle H(\theta^\star)\lambda^\star, \nu - \lambda^\star\rangle \geq 0, \qquad \forall \nu \in \Delta_m.$$

This is the variational-inequality characterization of the dual minimizer of

$$\min_{\lambda\in\Delta_m} \frac{1}{2}\lambda^\top H(\theta^\star)\lambda.$$

Hence $\lambda^\star = \lambda_{\mathrm{MGD}}(\theta^\star)$. Together with $G(\theta^\star)\lambda^\star = 0$, this shows that $(\theta^\star, \lambda^\star)$ is a fixed point of MGD.

We have proved equivalence, so the fixed-point sets coincide. The final statement follows immediately when both algorithms have unique globally attractive fixed points. □

E.8 Proof of Theorem C.1

Proof. For fixed $w \in \triangle_m$, recall that $F_w(\tilde{\theta};\theta) = \sum_{i=1}^m w_i J_i(\tilde{\theta};\theta)$. By Assumption 4.1, it's easy to check that $F_w(\cdot;\theta)$, a linear combination of J_i, is μ_w-strongly convex for every θ where $\mu_w = \sum_{i=1}^m w_i\mu_i$, therefore the minimizer $T_w(\theta) := \arg\min_{\tilde{\theta}} F_w(\tilde{\theta};\theta)$ is unique. Let $u = T_w(\theta)$ and $v = T_w(\theta')$. By first-order optimality,

$$\nabla_1 F_w(u;\theta) = 0, \qquad \nabla_1 F_w(v;\theta') = 0.$$

Using strong monotonicity of $\nabla_1 F_w(\cdot;\theta')$ in the first argument,

$$\begin{aligned}\mu_w \|u - v\|^2 &\leq \langle \nabla_1 F_w(u;\theta') - \nabla_1 F_w(v;\theta'), u - v\rangle \\ &= \langle \nabla_1 F_w(u;\theta') - \nabla_1 F_w(u;\theta), u - v\rangle \\ &\leq \|\theta - \theta'\| \|u - v\| \sum_{i=1}^m w_i\varepsilon L_i.\end{aligned}$$

If $u \neq v$, cancel $\|u - v\|$ on both sides to get

$$\|T_w(\theta) - T_w(\theta')\| \leq \frac{\sum_{i=1}^m w_i\varepsilon L_i}{\mu_w}\|\theta - \theta'\| = \frac{\sum_{i=1}^m w_i\varepsilon L_i}{\sum_{i=1}^m w_i\mu_i}\|\theta - \theta'\|.$$

The same inequality is trivial when $u = v$. Thus T_w is $\frac{\sum_{i=1}^m w_i\varepsilon L_i}{\sum_{i=1}^m w_i\mu_i}$-Lipschitz. Notice that $\alpha = \min_{i\in[m]}\alpha_i = \min_{i\in[m]}(\mu_i - \varepsilon L_i) > 0$ by the assumptions which means $\frac{\sum_{i=1}^m w_i\varepsilon L_i}{\sum_{i=1}^m w_i\mu_i} < 1$ and Banach's fixed-point theorem gives a unique fixed point $\theta_w^\star$ with $T_w(\theta_w^\star) = \theta_w^\star$.

Let $\theta^\star := \theta_w^\star$. Define

$$G_t(\tilde{\theta}) := \tilde{\theta} - \eta\nabla_1 F_w(\tilde{\theta};\theta_t).$$

Then the WGD update is $\theta_{t+1} = G_t(\theta_t)$. Because $\nabla_1 F_w(\theta^\star;\theta^\star) = 0$,

$$\begin{aligned}\|\theta_{t+1} - \theta^\star\| &= \|\theta_t - \eta\nabla_1 F_w(\theta_t;\theta_t) - (\theta^* - \eta\nabla_1 F_w(\theta^*;\theta^*))\| \\ &\leq r\|\theta_t - \theta^*\|, \qquad \text{By Lemma E.4}\end{aligned}$$

and the stepsize $\eta \in (0, \frac{2\alpha}{\Gamma^2})$, therefore $r < 1$ and $\|\theta_t - \theta^*\| \leq r^t\|\theta_0 - \theta^*\|$. Moreover, since $\nabla_1 F_w(\theta^\star;\theta^\star) = 0 = \sum_{i=1}^m w_i g_i(\theta^*) = wG(\theta^*)$, θ^* is a PPS point by Lemma 2.3.

□

E.9 Proof of Theorem C.2

[Theorem C.2] Under Assumptions 3.2, 3.3, 4.1, 4.2, if $\rho_\gamma := r + c\frac{L_H}{\gamma} < 1$, *AMGD has a unique fixed point* $\theta^\star$ *which is a PPS point, the corresponding multiplier* $\arg\min_{\mu\in\Delta_m}\left\{\frac{1}{2}\mu^\top H(\theta^*)\mu + \frac{\gamma}{2}\|\mu - \bar{\lambda}\|^2\right\} = \bar{\lambda}$ *and* $\|\theta_t - \theta^\star\| \leq \rho_\gamma^t\|\theta_0 - \theta^\star\|$.

Proof. Define $M_\gamma(\theta) := \arg\min_{\mu\in\Delta_m} \left\{\frac{1}{2}\mu^\top H(\theta)\mu + \frac{\gamma}{2}\|\mu-\bar{\lambda}\|^2\right\}$. Since $\gamma > 0$ therefore the objective in $M_\gamma(\theta)$ is strictly convex in μ for any fixed θ. Moreover, combining that $\triangle_m$ is compact and convex, M_γ is single-valued. For any $\theta, \theta' \in \Theta$, let

$$\mu := M_\gamma(\theta), \qquad \mu' := M_\gamma(\theta').$$

The variational inequalities for μ and μ' are

$$\left\langle H(\theta)\mu + \gamma(\mu-\bar{\lambda}), \nu-\mu\right\rangle \geq 0, \qquad \forall \nu \in \Delta_m,$$

and

$$\left\langle H(\theta')\mu' + \gamma(\mu'-\bar{\lambda}), \nu-\mu'\right\rangle \geq 0, \qquad \forall \nu \in \Delta_m.$$

Choose $\nu = \mu'$ in the first inequality and $\nu = \mu$ in the second. Adding these two together gives

$$\begin{aligned}
&\langle H(\theta)\mu - H(\theta')\mu', \mu-\mu'\rangle + \gamma\|\mu-\mu'\|^2 \leq 0 \\
\Rightarrow\ &(\mu-\mu')^T H(\theta)(\mu-\mu') + (\mu-\mu')^T\big(H(\theta)-H(\theta')\big)\mu' + \gamma\|\mu-\mu'\|^2 \leq 0 \\
\Rightarrow\ &\gamma\|\mu-\mu'\|^2 \leq (\mu'-\mu)^T\big(H(\theta)-H(\theta')\big)\mu' \quad \text{since } (\mu-\mu')^T H(\theta)(\mu-\mu') \geq 0 \\
\Rightarrow\ &\gamma\|\mu-\mu'\|^2 \leq \|H(\theta)-H(\theta')\|\,\|\mu-\mu'\| \quad \text{since } \mu' \in \triangle_m,\ \|\mu'\| \leq 1 \\
\Rightarrow\ &\gamma\|\mu-\mu'\|^2 \leq L_H\|\theta-\theta'\|\,\|\mu-\mu'\| \quad \text{By Lemma E.5.}
\end{aligned} \tag{25}$$

Therefore, by μ, μ''s definition, we can get that for any $\theta, \theta' \in \Theta$

$$\|M_\gamma(\theta) - M_\gamma(\theta')\| \leq \frac{L_H}{\gamma}\|\theta-\theta'\|. \tag{26}$$

Next, we prove the contraction of the AMGD update. For simplicity, define $T_\gamma(\theta) := \theta - \eta G(\theta)M_\gamma(\theta)$ for $\theta \in \Theta$ as one-step update for AMGD based on θ. Take arbitrary $\theta, \theta' \in \Theta$ and write

$$\mu := M_\gamma(\theta), \qquad \mu' := M_\gamma(\theta').$$

Then

$$T_\gamma(\theta) - T_\gamma(\theta') = \underbrace{\theta - \eta G(\theta)\mu - (\theta' - \eta G(\theta')\mu)}_{A} - \eta\underbrace{G(\theta')(\mu'-\mu)}_{B}.$$

By Lemma E.4,

$$\|A\| \leq r\|\theta-\theta'\|.$$

Moreover, notice that by Eq. equation 26,

$$\eta\|B\| \leq \eta\sqrt{m}G_\infty\|\mu'-\mu\| \leq c\frac{L_H}{\gamma}\|\theta-\theta'\|.$$

Therefore

$$\|T_\gamma(\theta) - T_\gamma(\theta')\| \leq (r + c\frac{L_H}{\gamma})\|\theta-\theta'\| = \rho_\gamma\|\theta-\theta'\|.$$

By the assumption $\rho_\gamma = r + c\frac{L_H}{\gamma} < 1$, T_γ is a contraction.

Since Θ is complete and invariant, Banach's fixed-point theorem yields a unique $\theta^\star \in \Theta$ such that $T_\gamma(\theta^\star) = \theta^\star$, and

$$\|\theta_t - \theta^\star\| \leq \rho_\gamma^t\|\theta_0-\theta^\star\|.$$

Applying equation 26 at $(\theta_t, \theta^\star)$ gives

$$\|\lambda_t - \lambda^\star\| \leq \frac{L_H}{\gamma}\|\theta_t-\theta^\star\| \leq \frac{L_H}{\gamma}\rho_\gamma^t\|\theta_0-\theta^\star\|,$$

which means that λ_t converges to $\lambda^* := M_\gamma(\theta^*)$.

Finally, we prove that θ^* is PPS and $\lambda^* = \bar{\lambda}$. By definition, $\lambda^\star = M_\gamma(\theta^\star)$, and the primal fixed-point equation gives

$$G(\theta^\star)\lambda^\star = 0. \tag{27}$$

Now use the variational inequality for $M_\gamma(\theta^\star)$:

$$\left\langle H(\theta^\star)\lambda^\star + \gamma(\lambda^\star-\bar{\lambda}), \nu-\lambda^\star\right\rangle \geq 0, \qquad \forall \nu \in \Delta_m.$$

Because $H(\theta^\star)\lambda^\star = G(\theta^\star)^\top G(\theta^\star)\lambda^\star = 0$ by equation 27, this simplifies to

$$\left\langle \lambda^\star - \bar{\lambda}, \nu - \lambda^\star \right\rangle \geq 0, \qquad \forall \nu \in \Delta_m.$$

Choosing $\nu = \bar{\lambda} \in \Delta_m$ yields

$$-\left\|\lambda^\star - \bar{\lambda}\right\|^2 \geq 0,$$

so $\lambda^\star = \bar{\lambda}$. Returning to equation 27, we obtain

$$G(\theta^\star)\bar{\lambda} = 0.$$

Since $\bar{\lambda} \in \Delta_m$, this means $0 \in \operatorname{conv}\{g_1(\theta^\star), \ldots, g_m(\theta^\star)\}$, hence $\theta^\star$ is PPS. □

E.10 Proof of Theorem C.3

Proof. Define $M_{\tau,\gamma}(\theta, \lambda) := \arg\min_{o \in \triangle_m} \left\{ \frac{1}{2} o^T H(\theta) o + \frac{\tau}{2} \|o - \lambda\|^2 + \frac{\gamma}{2} \left\|o - \bar{\lambda}\right\|^2 \right\}$. where $\bar{\lambda}$ is the anchor.

First, we prove the lipschitz continuity of $M_{\tau,\gamma}$. Because $\tau + \gamma > 0$, the objective in $M_{\tau,\gamma}$ is strictly convex in μ, so the minimizer is unique (similar to Theorem 4.7).

Take arbitrary (θ, λ) and (θ', λ'), and write

$$\mu := M_{\tau,\gamma}(\theta, \lambda), \qquad \mu' := M_{\tau,\gamma}(\theta', \lambda').$$

The corresponding variational inequalities are

$$\left\langle H(\theta)\mu + \tau(\mu - \lambda) + \gamma(\mu - \bar{\lambda}), \nu - \mu \right\rangle \geq 0, \qquad \forall \nu \in \Delta_m,$$

and

$$\left\langle H(\theta')\mu' + \tau(\mu' - \lambda') + \gamma(\mu' - \bar{\lambda}), \nu - \mu' \right\rangle \geq 0, \qquad \forall \nu \in \Delta_m.$$

Choose $\nu = \mu'$ in the first inequality and $\nu = \mu$ in the second, then add them. After rearrangement,

$$\langle H(\theta)\mu - H(\theta')\mu', \mu - \mu' \rangle + (\tau + \gamma) \|\mu - \mu'\|^2 \leq \tau \langle \lambda - \lambda', \mu - \mu' \rangle .$$

Expanding the first term and using the positive semidefiniteness of $H(\theta)$ gives

$$(\tau + \gamma) \|\mu - \mu'\|^2 \leq \left\| \big(H(\theta) - H(\theta')\big)\mu' \right\| \|\mu - \mu'\| + \tau \|\lambda - \lambda'\| \|\mu - \mu'\| .$$

As before, $\|\mu'\| \leq 1$, hence $\left\| \big(H(\theta) - H(\theta')\big)\mu' \right\| \leq L_H \|\theta - \theta'\|$. Therefore,

$$\|\mu - \mu'\| \leq \frac{L_H}{\tau + \gamma} \|\theta - \theta'\| + \frac{\tau}{\tau + \gamma} \|\lambda - \lambda'\| .$$

Next, let

$$a_t := \|\theta_t - \theta^\star\| , \qquad b_t := \|\lambda_t - \lambda^\star\| .$$

For two arbitrary states (θ, λ) and (θ', λ'), denote

$$\mu := M_{\tau,\gamma}(\theta, \lambda), \qquad \mu' := M_{\tau,\gamma}(\theta', \lambda').$$

By Lemma E.4,

$$\|\mu - \mu'\| \leq p_{\tau,\gamma} \|\theta - \theta'\| + q_{\tau,\gamma} \|\lambda - \lambda'\| . \tag{28}$$

For the primal step,

$$\theta - \eta G(\theta)\mu - (\theta' - \eta G(\theta')\mu') = \underbrace{\theta - \eta G(\theta)\mu - (\theta' - \eta G(\theta')\mu)}_{A} - \eta \underbrace{G(\theta')(\mu' - \mu)}_{B} .$$

By Lemma E.4,

$$\|A\| \leq r \|\theta - \theta'\| , \text{ and } \eta \|B\| \leq c \|\mu' - \mu\| .$$

Combining with equation 28 yields

$$\|\theta - \eta G(\theta)\mu - (\theta' - \eta G(\theta')\mu')\| \leq (r + c p_{\tau,\gamma}) \|\theta - \theta'\| + c q_{\tau,\gamma} \|\lambda - \lambda'\| .$$

Therefore,

$$\begin{pmatrix} a_{t+1} \\ b_{t+1} \end{pmatrix} \leq A_{\tau,\gamma} \begin{pmatrix} a_t \\ b_t \end{pmatrix} .$$

Since the structure of $A_{\tau,\gamma}$ is the same with A_s or A_τ in Theorem 4.7 and 4.5. Therefore, similar to the proof in previous theorems, we proved our results. □

E.11 Proof of Lemma C.4

Proof. Assume the weight is w. We show that $(i) : \theta^*$ is the fixed point of WGD $\Leftrightarrow$ $(ii) : \theta^*$ is the fixed point of AMGD $\Leftrightarrow$ $(iii) : \theta^*$ is the fixed point of PAMGD.

(i) $\Rightarrow$ (ii). If $\theta^\star$ is a fixed point of WGD with weight w, then

$$G(\theta^\star)w = 0.$$

Consider the anchored multiplier objective at $\theta^\star$:

$$\Phi_{\mathrm{anc}}(\lambda) := \frac{1}{2}\lambda^\top H(\theta^\star)\lambda + \frac{\gamma}{2}\left\|\lambda - w\right\|^2.$$

At $\lambda = w$ we have

$$H(\theta^\star)w = G(\theta^\star)^\top G(\theta^\star)w = 0.$$

Hence, for every $\nu \in \Delta_m$,

$$\Phi_{\mathrm{anc}}(\nu) - \Phi_{\mathrm{anc}}(w) = \frac{1}{2}\left\|G(\theta^\star)\nu\right\|^2 + \frac{\gamma}{2}\left\|\nu - w\right\|^2 \geq 0.$$

Thus $w = M_\gamma(\theta^\star)$, and together with $G(\theta^\star)w = 0$ we conclude that $(\theta^\star, w)$ is a fixed point of AMGD.

(ii) $\Rightarrow$ (i). If $(\theta^\star, w)$ is a fixed point of AMGD, then by definition

$$G(\theta^\star)w = 0,$$

which is exactly the fixed-point condition of WGD with weight w.

(i) $\Rightarrow$ (iii). Assume again that $G(\theta^\star)w = 0$. Consider the anchored proximal multiplier objective centered at w:

$$\Phi_{\mathrm{AP}}(\lambda) := \frac{1}{2}\lambda^\top H(\theta^\star)\lambda + \frac{\tau}{2}\left\|\lambda - w\right\|^2 + \frac{\gamma}{2}\left\|\lambda - w\right\|^2.$$

Exactly as above,

$$\Phi_{\mathrm{AP}}(\nu) - \Phi_{\mathrm{AP}}(w) = \frac{1}{2}\left\|G(\theta^\star)\nu\right\|^2 + \frac{\tau+\gamma}{2}\left\|\nu - w\right\|^2 \geq 0.$$

Thus $w = M_{\tau,\gamma}(\theta^\star, w)$ and $(\theta^\star, w)$ is a fixed point of PAMGD.

(iii) $\Rightarrow$ (i). If $(\theta^\star, w)$ is a fixed point of PAMGD, then by definition

$$G(\theta^\star)w = 0,$$

which is again the fixed-point condition of WGD.

This proves the equivalence. The final conclusion follows immediately under uniqueness. □

E.12 Discussions of State-dependent Performative Prediction

This section explains why our reduced-form assumptions apply to state-dependent performative settings. In state-dependent settings, current distribution also influences the distribution at the next round. Suppose the environment is generated by a transition map $T : \Theta \times \mathcal{P}(\mathcal{Z}) \to \mathcal{P}(\mathcal{Z})$ and that, for some constants $a \geq 0$ and $0 \leq b < 1$,

$$\mathrm{dist}\big(T(\theta; D), T(\theta'; D')\big) \leq a\left\|\theta - \theta'\right\| + b\,\mathrm{dist}(D, D')$$

for some probability metric dist on $\mathcal{P}(\mathcal{Z})$. Let $D(\theta)$ denote the fixed point of $T(\theta; \cdot)$, so $D(\theta) = T(\theta; D(\theta))$. Then

$$\mathrm{dist}\big(D(\theta), D(\theta')\big) = \mathrm{dist}\big(T(\theta; D(\theta)), T(\theta'; D(\theta'))\big) \leq a\left\|\theta - \theta'\right\| + b\,\mathrm{dist}\big(D(\theta), D(\theta')\big).$$

Rearranging gives

$$\mathrm{dist}\big(D(\theta), D(\theta')\big) \leq \frac{a}{1-b}\left\|\theta - \theta'\right\|.$$

Thus the sensitivity also holds in the stateful setting after replacing the sensitivity constant by $a/(1-b)$.

E.13 AUXILIARY LEMMAS

Lemma E.2 (Non-emptiness of the PPS set)**.** *Under Assumption 3.2, 3.3 and 4.1, assume that $\Theta \subset \mathbb{R}^d$ is nonempty, compact, and convex, and that $G : \Theta \to \mathbb{R}^{d\times m}$ is continuous by Lemma E.3. Suppose that the unified update correspondence*

$$\mathcal{T}_\tau(\theta, \lambda) := \left\{ \big(\theta - \eta G(\theta)\mu, \mu\big) : \mu \in M_\tau(\theta, \lambda) \right\}$$

is a self-map on $\Theta \times \Delta_m$, i.e.

$$\mathcal{T}_\tau(\theta, \lambda) \subseteq \Theta \times \Delta_m, \qquad \forall (\theta, \lambda) \in \Theta \times \Delta_m.$$

Then $\mathcal{M} \neq \varnothing$ and $\mathcal{S} \neq \varnothing$.

Proof. Fix $(\theta, \lambda) \in \Theta \times \Delta_m$. The feasible set Δ_m is nonempty and compact. The objective

$$\mu \mapsto \frac{1}{2}\mu^\top H(\theta)\mu + \frac{\tau}{2}\|\mu - \lambda\|^2$$

is continuous in (μ, θ, λ). Therefore, by the Weierstrass theorem, $M_\tau(\theta, \lambda)$ is nonempty and compact. Moreover, since $H(\theta) = G(\theta)^\top G(\theta) \succeq 0$, the first term is convex in μ, and the proximal term is convex. Therefore the objective is convex in μ, and hence its argmin set over the convex set Δ_m is convex. Thus M_τ is nonempty, compact-valued, and convex-valued.

We next verify upper hemicontinuity. Since the feasible correspondence is the constant compact-valued correspondence $(\theta, \lambda) \mapsto \Delta_m$, and the objective above is continuous in all variables, Berge's maximum theorem implies that the argmin correspondence M_τ is upper hemicontinuous and compact-valued. For $\tau > 0$, the objective is strongly convex in μ, so M_τ is single-valued; the same conclusion is not needed here.

The map

$$(\theta, \lambda, \mu) \mapsto \big(\theta - \eta G(\theta)\mu, \mu\big)$$

is continuous and affine in μ for fixed (θ, λ). Hence the image of the nonempty compact convex set $M_\tau(\theta, \lambda)$ is again nonempty, compact, and convex. The upper hemicontinuity of M_τ, together with the continuity of this image map, implies that $\mathcal{T}_\tau$ is upper hemicontinuous.

Now $\Theta \times \Delta_m$ is nonempty, compact, and convex, and by Kakutani's fixed-point theorem, there exist $(\theta^\star, \lambda^\star) \in \Theta \times \Delta_m$ and $\mu^\star \in M_\tau(\theta^\star, \lambda^\star)$ such that

$$(\theta^\star, \lambda^\star) = \big(\theta^\star - \eta G(\theta^\star)\mu^\star, \mu^\star\big).$$

Consequently, $\lambda^\star = \mu^\star$ and

$$G(\theta^\star)\lambda^\star = 0.$$

Since

$$\frac{1}{2}\|G(\theta^\star)\lambda^\star\|^2 = 0$$

and the function $\lambda \mapsto \frac{1}{2}\|G(\theta^\star)\lambda\|^2$ is nonnegative on Δ_m, $\lambda^\star$ is a global minimizer of the MGDA dual problem. Thus

$$\lambda^\star \in \Lambda(\theta^\star), \qquad G(\theta^\star)\lambda^\star = 0.$$

Therefore $(\theta^\star, \lambda^\star) \in \mathcal{M}$. Hence $\mathcal{M} \neq \varnothing$, and consequently $\mathcal{S} = \mathrm{proj}_\theta(\mathcal{M}) \neq \varnothing$. □

Lemma E.3 (Diagonal fields are strongly monotone and Lipschitz)**.** *Under Assumption 3.2, 3.3, 4.1, for every $i \in [m]$ and every $\theta, \theta' \in \Theta$,*

$$\langle g_i(\theta) - g_i(\theta'), \theta - \theta' \rangle \geq (\mu_i - \varepsilon L_i) \left\|\theta - \theta'\right\|^2, \tag{29}$$

and

$$\|g_i(\theta) - g_i(\theta')\| \leq (L_i + \varepsilon L_i) \left\|\theta - \theta'\right\|. \tag{30}$$

Proof. Fix $i \in [m]$ and write

$$g_i(\theta) - g_i(\theta') = \underbrace{\nabla_1 J_i(\theta; \theta) - \nabla_1 J_i(\theta'; \theta)}_{A_i} + \underbrace{\nabla_1 J_i(\theta'; \theta) - \nabla_1 J_i(\theta'; \theta')}_{B_i}.$$

For the first term A_i, Assumption 4.1 and 3.2 implies that, for fixed deployed model θ, the map $\tilde{\theta} \mapsto J_i(\tilde{\theta}; \theta)$ is μ_i-strongly convex and L_i-smooth. Therefore

$$\langle A_i, \theta - \theta' \rangle \geq \mu_i \|\theta - \theta'\|^2, \qquad \|A_i\| \leq L_i \|\theta - \theta'\|.$$

For the second term B_i, Assumption 3.3 gives

$$\|B_i\| \leq \varepsilon L_i \|\theta - \theta'\|.$$

Hence

$$\langle B_i, \theta - \theta' \rangle \geq -\|B_i\| \, \|\theta - \theta'\| \geq -\varepsilon L_i \|\theta - \theta'\|^2.$$

Adding the two bounds yields equation 29. The Lipschitz bound equation 30 follows from the triangle inequality:

$$\|g_i(\theta) - g_i(\theta')\| \leq \|A_i\| + \|B_i\| \leq (L_i + \varepsilon L_i) \|\theta - \theta'\|.$$

□

Lemma E.4 (Fixed-weight fields are strongly monotone and Lipschitz)**.** *Under Assumption 3.2, 3.3, 4.1, for every fixed $\lambda \in \Delta_m$, define*

$$F_\lambda(\theta) := G(\theta)\lambda = \sum_{i=1}^m \lambda_i g_i(\theta).$$

For every $\theta, \theta' \in \Theta$,

$$\langle F_\lambda(\theta) - F_\lambda(\theta'), \theta - \theta' \rangle \geq \alpha \|\theta - \theta'\|^2, \tag{31}$$

and

$$\|F_\lambda(\theta) - F_\lambda(\theta')\| \leq \Gamma \|\theta - \theta'\|. \tag{32}$$

Consequently, if the stepsize $\eta \in (0, \frac{2\alpha}{\Gamma^2})$, then

$$\|\theta - \eta F_\lambda(\theta) - (\theta' - \eta F_\lambda(\theta'))\| \leq r \|\theta - \theta'\|. \tag{33}$$

Proof. The claims equation 31 and equation 32 follow by taking the convex combination, with weights λ_i, of equation 29 and equation 30. Specifically,

$$\langle F_\lambda(\theta) - F_\lambda(\theta'), \theta - \theta' \rangle = \sum_{i=1}^m \lambda_i \langle g_i(\theta) - g_i(\theta'), \theta - \theta' \rangle \geq \sum_{i=1}^m \lambda_i \alpha \|\theta - \theta'\|^2 = \alpha \|\theta - \theta'\|^2,$$

and similarly

$$\|F_\lambda(\theta) - F_\lambda(\theta')\| \leq \sum_{i=1}^m \lambda_i \Gamma \|\theta - \theta'\| = \Gamma \|\theta - \theta'\|.$$

Now expand

$$\begin{aligned}
&\|\theta - \eta F_\lambda(\theta) - (\theta' - \eta F_\lambda(\theta'))\|^2 \\
&\quad = \|\theta - \theta'\|^2 - 2\eta \langle F_\lambda(\theta) - F_\lambda(\theta'), \theta - \theta' \rangle + \eta^2 \|F_\lambda(\theta) - F_\lambda(\theta')\|^2 \\
&\quad \leq \left(1 - 2\eta\alpha + \eta^2\Gamma^2\right) \|\theta - \theta'\|^2.
\end{aligned}$$

If the stepsize $\eta \in (0, \frac{2\alpha}{\Gamma^2})$, then $1 - 2\eta\alpha + \eta^2\Gamma^2 \in (0, 1)$, and hence taking square roots yields equation 33 with $r = \sqrt{1 - 2\eta\alpha + \eta^2\Gamma^2}$. □

Lemma E.5 ($H(\theta)$ is Lipschitz on Θ)**.** *Under Assumption 3.2, 3.3, 4.1, there exists a constant $L_H > 0$ such that*

$$\|H(\theta) - H(\theta')\| \leq L_H \|\theta - \theta'\|, \qquad \forall \theta, \theta' \in \Theta. \tag{34}$$

One admissible choice is

$$L_H := 2mG_\infty\Gamma.$$

Proof. For each $i, j \in [m]$, define

$$h_{ij}(\theta) := \langle g_i(\theta), g_j(\theta) \rangle .$$

Then

$$\begin{aligned} |h_{ij}(\theta) - h_{ij}(\theta')| &= |\langle g_i(\theta), g_j(\theta) \rangle - \langle g_i(\theta'), g_j(\theta') \rangle| \\ &\leq |\langle g_i(\theta), g_j(\theta) - g_j(\theta') \rangle| + |\langle g_i(\theta) - g_i(\theta'), g_j(\theta') \rangle| \\ &\leq \|g_i(\theta)\| \, \|g_j(\theta) - g_j(\theta')\| + \|g_i(\theta) - g_i(\theta')\| \, \|g_j(\theta')\| \\ &\leq 2G_\infty \Gamma \, \|\theta - \theta'\| . \end{aligned}$$

Thus every entry of $H(\theta) - H(\theta')$ is bounded by $2G_\infty \Gamma \, \|\theta - \theta'\|$ in absolute value. Therefore its Frobenius norm satisfies

$$\|H(\theta) - H(\theta')\|_F \leq m \cdot 2G_\infty \Gamma \, \|\theta - \theta'\| .$$

Finally, the operator norm is bounded by the Frobenius norm, so equation 34 holds with $L_H = 2mG_\infty \Gamma$. □