First-Order Bilevel Optimization for Non-truthful Auction Design

Anonymous authors
Paper under double-blind review

Abstract

Automated mechanism design has enabled the learning of complex auction mechanisms using neural networks. However, designing non-truthful auctions remains challenging because bidders' equilibrium strategies depend on the auction mechanism and thus evolve as the mechanism is optimized. In this work, we formulate auction design as a Stackelberg game with multiple strategic followers and develop a fully first-order algorithm for revenue optimization in non-truthful auctions. We establish non-asymptotic guarantees for hypergradient approximation and show that, under standard smoothness and strong-monotonicity assumptions, our method reaches an $O(\epsilon)$-stationary point using $\tilde{\mathcal{O}}(\epsilon^{-2})$ first-order oracle calls. Experiments on symmetric and asymmetric single-item auctions, as well as multi-item auctions, show that our method converges faster than second- and zeroth-order baselines and scales better as the number of bidders increases. Our method also discovers auctions that achieve revenue close to truthful baselines. These results provide a scalable framework for automated mechanism design beyond truthful auctions.

1 Introduction

Automated mechanism design (AMD) uses computation and data to optimize auction mechanisms when analytical solutions are unavailable. Early AMD approaches formulate mechanism design as large-scale mathematical programs (Conitzer & Sandholm (2002; 2004)), while recent work uses neural networks to learn approximately revenue-optimal mechanisms from valuation samples (Dütting et al. (2019); Feng et al. (2018); Wang et al. (2024); Pham et al. (2025)). These methods primarily focus on truthful mechanisms, where incentive compatibility makes truthful reporting an equilibrium and therefore removes the need to compute bidders' strategic responses.

Many widely deployed auctions, however, are non-truthful (Bigler (2019); Goke et al. (2021); Galgana & Golrezaei (2025)). Although the revelation principle often permits mechanism designers to restrict their attention to truthful mechanisms without loss, it does not apply in many settings, such as when bidders have type-dependent constraints on their feasible actions (Berzack et al., 2026). Moreover, practical considerations may limit the mechanism to a prescribed family of non-truthful auctions, such as first-price auctions with reserves. Such auction formats are common in practice; for example, first-price auctions are widespread in online advertising due to their simplicity and transparency (Harada, 2018). However, optimizing non-truthful mechanisms is substantially harder: changing the mechanism changes bidders' equilibrium strategies, so the revenue must be optimized through the equilibrium of an induced multi-player game.

We formulate non-truthful AMD as a multi-follower Stackelberg problem, where the auctioneer is the leader and bidders are strategic followers. This formulation connects mechanism design to bilevel optimization. Existing approaches to non-truthful mechanism optimization either use zeroth-order optimization (Feng et al. (2021)) or assume access to a differentiable equilibrium oracle (Thoma et al. (2026); Wang et al. (2022)). Meanwhile, current fully first-order bilevel methods primarily consider a single-follower setting (Kwon et al. (2023); Kornowski et al. (2024)). Unlike the single-follower bilevel problem, however, a multi-follower equilibrium couples the responses of all followers: changing one follower's strategy affects the incentives of the others, and these cross-player effects are generally asymmetric, so standard first-order bilevel constructions do not extend directly. This leaves open how to develop a scalable fully first-order method for multiple strategic followers.

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.

Main contributions. Our contributions are threefold. **First**, we formulate non-truthful auction design as a Stackelberg game with $N > 1$ followers and develop a fully first-order method for symmetric-Jacobian games, achieving ϵ-stationarity with $\mathcal{O}(N\epsilon^{-2})$ gradient oracle complexity. We also show why this approach fails in asymmetric games: the true hypergradient requires Jacobian $J^{-\top}$ whereas our method yields J^{-1}. **Second**, we resolve this mismatch using a finite-difference potential function that induces an auxiliary game with a local Jacobian approximating J^{-T}. We prove local strong monotonicity and non-asymptotic hypergradient guarantees for the auxiliary game, provide a fully first-order algorithm that converges to ϵ-stationarity using $\tilde{\mathcal{O}}(N^2\epsilon^{-2})$ oracle calls, and extend the analysis to linear inequality-constrained follower games. **Lastly**, we formulate a neural-network-parameterized auction design problem in which the auctioneer optimizes revenue while bidders respond strategically. We develop a warm-start strategy for tracking differentiable branches of local Nash equilibria, allowing the first-order method to be applied without assuming a globally unique equilibrium. We evaluate the approach on symmetric and asymmetric single-item auctions and multi-item auctions, demonstrating its applicability to equilibrium-dependent mechanism design.

2 Related Works

Neural Approaches to Learning Truthful Auctions. Classical approaches to revenue-maximizing mechanism design face two obstacles in practice. First, they typically assume the designer knows the distribution of buyers' values in advance (e.g., Myerson, 1981)—information that is often unavailable. Second, a general closed-form characterization of revenue-optimal multi-item auctions is unknown (e.g., Daskalakis et al., 2014; Rochet & Choné, 1998). Automated mechanism design (AMD) aims to circumvent these challenges by optimizing for revenue-maximizing mechanisms using sample valuation profiles and computational search (Conitzer & Sandholm, 2002; 2004). The first neural approach to AMD is the work of Dütting et al. (2019), which represents allocation and payment rules using neural networks and penalizing bidders' gains from misreporting to approximately enforce dominant-strategy incentive compatibility. More recently, GemNet (Wang et al., 2024) uses bidder-independent menus and a post-training transformation to guarantee incentive compatibility. In contrast, we optimize neural pay-your-bid mechanisms whose revenue depends on bidders' strategic responses, requiring us to account for how equilibrium strategies change with the mechanism.

Non-truthful Auction Design: Optimizing non-truthful auctions requires accounting for how mechanism changes affect bidders' equilibrium strategies. Several papers show that gradient dynamics empirically converge to Nash equilibria in auction games, but focus on computing bidders' strategies rather than optimizing the auction itself (Bichler et al., 2021; 2023a; Kohring et al., 2023; Huang et al., 2025). More directly related, Feng et al. (2021) optimize first-price auction reserves via zeroth-order gradient descent. They compute gradient estimates based on bidders' responsiveness to experimental shocks in reserves, rather than differentiating through a computed equilibrium. Unlike reserve optimization, our framework learns the allocation rule itself, including for multi-item auctions. Meanwhile, Thoma et al. (2026) formulate mechanism design as a bilevel optimization problem with (coarse) correlated equilibria as the lower-level solution. They assume access to a neural network that outputs a differentiable equilibrium for discrete-action games and its gradient with respect to mechanism parameters. In contrast, we develop novel fully first-order bilevel optimization algorithms with convergence guarantees for Stackelberg games with multiple followers, and apply them to auction design. We design auction mechanisms parameterized by neural networks and compute Nash equilibria through gradient dynamics. Scalability is a key limitation of Thoma et al. (2026); their models are only trained on two-player games with at most 16 discrete actions for each player. In contrast, our action space is continuous.

Bilevel Optimization: Bilevel optimization (BO) is an optimization problem constrained by the solution of another optimization problem. It is essentially a Stackelberg game with a single follower. A body of work studies fully first-order methods for BO. These works typically assume a strongly convex lower-level problem to ensure a unique lower-level solution and an invertible KKT matrix. For an unconstrained lower-level problem, Kwon et al. (2023) derives an approximate hypergradient using a penalty function and achieves a complexity of $\mathcal{O}(\epsilon^{1.5})$ in deterministic BO. For linearly constrained lower-level problems, Kornowski et al. (2024); Zhao et al. (2025) achieves a complexity of $\mathcal{O}(\epsilon^{-3}\delta^{-1})$ to converge to a (ϵ, δ)-Goldstein stationary point. To the best of our knowledge,

existing first-order algorithm for solving Stackelberg games with multiple followers with implicit differentiation is only applicable when Jacobian is block-diagonal (Niu et al. (2025)).

3 Non-truthful Auction Design as a Stackelberg Game

Consider a sealed-bid auction of N bidders and J items. Bidder i has a private valuation $t_i \in \mathcal{T}_i \subseteq \mathbb{R}_+^J$, drawn independently from a distribution F_i, and submits a bid $a_i \in \mathcal{A}_i \subseteq \mathbb{R}_+^J$. Let $t = (t_1, ..., t_N)$ and $a = (a_1, .., a_N)$. An auction mechanism is parametrized by $x \in X \subseteq \mathbb{R}^p$ and consists of a randomized allocation rule ρ_x and a payment rule P_x. Given the auction x, a bid profile a, and valuation t_i for bidder i, bidder i's expected utility can be computed by $u_i(x, a, t_i)$.

For a given mechanism x, each bidder i chooses a parametrized bidding strategy $\beta_{y_i} : \mathcal{T}_i \to A_i$. This strategy maps the bidder's valuation t_i to a bid in A_i, and y_i is a parameter vector defining the strategy. Unless stated otherwise, $y_i \in Y_i(x) := \mathbb{R}^{n_i}$. Let $y := (y_1, \cdots, y_N) \in Y(x) := \Pi_{i=1}^N Y_i(x) \subseteq \mathbb{R}^n$ and write $y_{\neg i} := (y_j)_{j \neq i}$. The mechanism and bidding strategies induce an *ex-ante expected cost* for each bidder. We work with costs rather than utilities so that the follower game can be expressed as a minimization problem.

Definition 3.1 (Ex-ante expected cost)**.** Bidder i's ex-ante expected cost is

$$g_i(x, y_i, y_{\neg i}) := -\mathbb{E}_{t_1, \cdots, t_N}[u_i(x, \{\beta_{y_j}(t_j)\}_{j=1}^N, t_i)].$$

Thus, minimizing g_i is equivalent to maximizing bidder i's ex-ante expected utility.

For a fixed mechanism x, the bidders therefore play a normal-form game $\mathcal{G}(x) := (x, [N], \{Y_i\}_{i \in [N]}, \{g_i\}_{i \in [N]})$. We assume bidders follow a Nash equilibrium with respect to the ex-ante expected cost functions. A Nash equilibrium is a strategy profile in which no player can reduce their expected cost by unilaterally changing their strategy parameter while holding everyone else's strategies fixed.

Definition 3.2 (Nash equilibrium)**.** We say $y^* \in Y(x)$ is a Nash equilibrium of the game $\mathcal{G}(x)$ if and only if $g_i(x, y^*) \leq g_i(x, y_i, y^*_{\neg i})$, $\forall i \in [N]$, $\forall y_i \in Y_i(x)$. We denote the set of Nash equilibria as $\text{NE}(\mathcal{G}(x))$.

The auction design problem is to choose the expected payment rule P_x and allocation rule ρ_x that maximize some objective, assuming bidders play Nash equilibrium strategies. We focus on maximizing the auctioneer's expected revenue. Because our framework is written as a minimization problem, let $f(x, y)$ denote the negative expected revenue under mechanism x and bidding-strategy profile y. We first consider the setting in which the induced lower-level game has a unique Nash equilibrium $y^*(x)$ and formulate the auction design problem as a Stackelberg Game.

$$\min_{x \in X} F(x) := f(x, y^*(x)) \quad \text{s.t. } y^*(x) \in \text{NE}(\mathcal{G}(x)). \tag{P1}$$

Section 6 extends this formulation to auction games with multiple equilibria.

To solve Problem P1, we run gradient descent on $F(x)$ to optimize the parameter x. Our goal is to find an ϵ-stationary point, which is defined as any point $x \in \mathbb{R}^p$ such that $\|\nabla F(x)\| \leq \epsilon$.

Hypergradient and second-order information: Due to the nested structure in Problem P1, computing the gradient of the function $F(x)$ requires differentiating through the Nash equilibrium $y^*(x)$. To show how to compute $\frac{dy^*}{dx}$, we define the gradient operator:

Definition 3.3 (gradient operator)**.** $V(x, y) := (\nabla_{y_1} g_1(x, y), \cdots, \nabla_{y_N} g_N(x, y)) \in \mathbb{R}^n$.

We can write the optimality conditions of the Nash equilibrium y^* using the gradient operator, and apply the Implicit Function Theorem to solve $\frac{dy^*}{dx}$:

$$V(x, y^*) = 0 \text{ (optimality)} \implies \frac{\partial}{\partial x} V(x, y^*) + \frac{\partial}{\partial y} V(x, y^*) \frac{d}{dx} y^*(x) = 0 \tag{1}$$

Solving Equation 1 to get $\frac{dy^*}{dx}$, we get the following exact hypergradient:

$$\begin{aligned} \nabla F(x) &= \nabla_x f(x, y^*(x)) + (\frac{d}{dx} y^*(x))^\top \nabla_y f(x, y^*(x)) \\ &= \nabla_x f(x, y^*(x)) - (\frac{\partial}{\partial x} V(x, y^*))^\top J(x, y^*)^{-\top} \nabla_y f(x, y^*), \end{aligned} \tag{2}$$

where $J(x,y) := \frac{\partial}{\partial y} V(x,y)$ is the Jacobian of the gradient operator.

Equation 2 gives us the exact hypergradient to solve Problem P1. However, the hypergradient involves the Jacobian $J(x,y^*)$ of the gradient operator, which is second-order and can be computationally intractable. The main goal of this work is to construct an approximate hypergradient estimator that bypasses second-order information and solves Problem P1 efficiently.

Definitions and Assumptions: Below we summarize the definitions and assumptions needed for the later hypergradient and convergence analysis for solving Problem P1.

Assumption 3.4 (Lower-level game)**.** We assume the lower-level game $\mathcal{G}$ is μ-strongly monotone, C-smooth and S-Hessian smooth. We assume V to be

1. μ-strongly monotone: $\langle V(x,y') - V(x,y), y' - y\rangle \geq \mu\|y' - y\|^2$, $\forall y', y \in \mathbb{R}^n$,
2. Jointly C-Lipschitz: $\|V(x',y') - V(x,y)\| \leq C\|(x',y') - (x,y)\|$, $\forall (x',y'),(x,y) \in X \times \mathbb{R}^n$,
3. Jointly S-Hessian Lipschitz: $\|\frac{\partial}{\partial(x,y)} V(x',y') - \frac{\partial}{\partial(x,y)} V(x,y)\| \leq S\|(x',y') - (x,y)\|$, $\forall (x',y'),(x,y) \in X \times \mathbb{R}^n$.

The strong monotonicity assumption guarantees a unique lower-level Nash equilibrium (Facchinei & Pang, 2003).

Assumption 3.5 (Upper-level objective)**.** We assume f to be

1. Jointly L_f-Lipschitz $\|f(x',y') - f(x,y)\| \leq L_f\|(x',y') - (x,y)\|$, $\forall (x',y'),(x,y) \in X \times \mathbb{R}^n$,
2. C_f-gradient Lipschitz $\|\nabla_{(x,y)} f(x',y') - \nabla_{(x,y)} f(x,y)\| \leq C_f\|(x',y') - (x,y)\|$, $\forall (x',y'),(x,y) \in X \times \mathbb{R}^n$,
3. S_f-Hessian Lipschitz $\|\frac{\partial}{\partial(x,y)} \nabla_{(x,y)} f(x',y') - \frac{\partial}{\partial(x,y)} \nabla_{(x,y)} f(x,y)\| \leq S_f\|(x',y') - (x,y)\|$, $\forall (x',y'),(x,y) \in X \times \mathbb{R}^n$.

4 Stackelberg Games with Symmetric Jacobian

We first consider Stackelberg games where the Jacobian $J = \frac{\partial}{\partial y} V(x,y^*)$ is symmetric, i.e., $J = J^\top$. This condition holds in the special case where followers do not interact through their decision variables, or when the pairwise strategic interactions between every follower i and j are reciprocal. This section proposes a first-order method for solving Stackelberg games with a symmetric Jacobian.

Consider problem P1. Define a perturbed game $\mathcal{G}_\delta := (x, [N], \{Y_i\}_{i\in[N]}, \{g_i + \delta f\}_{i\in[N]})$, where the followers' objectives are perturbed by the leader's objective δf with a small number $\delta > 0$. Let y_δ^* be the new Nash equilibrium for the perturbed game $\mathcal{G}_\delta$. We construct the following finite difference:

$$v_\delta(x, y_\delta^*) = \frac{1}{\delta} \sum_{i=1}^{N} \nabla_x g_i(x, y_{\delta,i}^*, y_{\neg i}^*) - \nabla_x g_i(x, y_i^*, y_{\neg i}^*) \tag{3}$$

We first quantify the error of the finite-difference estimator in the idealized setting where the perturbed equilibrium y_δ^* is computed exactly.

Theorem 4.1. *If the Jacobian $J = \frac{\partial}{\partial y} V(x,y^*)$ is symmetric, then:*

$$\|v_\delta(x, y_\delta^*) + (\frac{\partial}{\partial x} V(x,y^*))^\top J(x,y^*)^{-\top} \nabla_y f(x,y^*)\| \leq \mathcal{O}(N\delta). \tag{4}$$

Proof sketch. The perturbed equilibrium y_δ^* satisfies $V(x, y_\delta^*) + \delta \nabla_y f(x, y_\delta^*) = 0$. Differentiating with respect to δ at 0 gives $\lim_{\delta\to 0} \frac{y_\delta^* - y^*}{\delta} = \frac{d}{d\delta} y_\delta^*|_{\delta=0} = -J(x,y^*)^{-1} \nabla_y f(x, y^*(x))$. By the fundamental Theorem of calculus we have $v_\delta(x, y_\delta^*) = \left(\frac{\partial V(x,y^*)}{\partial x}\right)^\top \frac{y_\delta^* - y^*}{\delta} + \mathcal{O}(N\delta)$ where the

remainder term comes from summing up the error $\|y_\delta^* - y^*\| = \mathcal{O}(\delta)$ for each player. Thus,

$$v_\delta(x, y_\delta^*) \approx -(\frac{\partial}{\partial x} V(x, y^*))^\top J(x, y^*)^{-1} \nabla_y f(x, y^*) + \mathcal{O}(N\delta) \tag{5}$$

$$= -(\frac{\partial}{\partial x} V(x, y^*))^\top J(x, y^*)^{-\top} \nabla_y f(x, y^*) + \mathcal{O}(N\delta) \tag{6}$$

The last equality in Equation (6) is due to the symmetry of J. □

This approximation guarantee suggests that we can run inexact gradient descent to obtain a fully first-order approximation of the hypergradient with a non-asymptotic rate. This intuition is formalized in Theorem C.6. The proof of Theorem 4.1 highlights that the naive finite difference construction fails when $J(x, y^*)^\top \neq J(x, y^*)$.

5 Stackelberg Games with Asymmetric Jacobian

As alluded to above, the current perturbed game cannot be used to solve the Stackelberg game without a symmetric Jacobian. We propose a novel approach to resolve the Jacobian mismatch in this section.

5.1 Finite difference approximation for unconstrained Stackelberg game

Consider differentiating $\langle V(x, y), y - y^* \rangle$ with respect to y. The product rule gives $\nabla_y \langle V(x, y), y - y^* \rangle - V(x, y) = J(x, y)^\top (y - y^*)$. This is exactly the transpose operator missing in (6). However, differentiating the scalar $\langle V(x, y), y - y^* \rangle$ through y requires second-order derivatives of the follower objectives. Thus, we construct a finite-difference *potential* whose gradient approximates $\nabla_y \langle V(x, y), y - y^* \rangle$.

Definition 5.1 (Potential function)**.** Given a small number $\epsilon > 0$, define

$$\phi_\epsilon(x, y) := \tfrac{1}{\epsilon} \textstyle\sum_{j=1}^N [g_j(x, y_j + \epsilon(y_j - y_j^*), y_{\neg j}) - g_j(x, y)].$$

By definition of directional derivative, each finite-difference term in $\phi_\epsilon(x, y)$ converges to $\langle \nabla_{y_i} g_i(x, y), y_i - y_i^* \rangle$. So $\phi_\epsilon(x, y)$ approaches $\langle V(x, y), y - y^* \rangle$ and $\nabla_y \phi_\epsilon(x, y)$ approaches to $\nabla_y \langle V(x, y), y - y^* \rangle$ as $\epsilon \to 0$. Therefore, $\nabla_y \phi_\epsilon(x, y) - V(x, y)$ approximates $J(x, y)^\top (y - y^*)$. Consider an unconstrained perturbed game $\bar{\mathcal{G}}_\delta := (x, [N], \{Y_i\}_{i \in [N]}, \{\bar{g}_i\}_{i \in [N]})$ with cost function $\bar{g}_i(x, y) = \phi_\epsilon(x, y) - g_i(x, y) + \delta f(x, y)$. Denote the gradient operator for the perturbed game $\bar{\mathcal{G}}_\delta$ as $\bar{V}(x, y) = (\nabla_{y_1} \bar{g}_1(x, y), \cdots, \nabla_{y_N} \bar{g}_N(x, y)) \in \mathbb{R}^n$. Then we have $\bar{V} = \nabla_y \phi_\epsilon - V + \delta \nabla_y f$ where the term $\nabla_y \phi_\epsilon - V$ is exactly what we need to recover the Jacobian. Now consider the perturbed equilibrium condition

$$0 = \bar{V}(x, \bar{y}_\delta^*) \approx J(x, y)^\top (\bar{y}_\delta^* - y^*) + \delta \nabla_y f(x, y^*) \implies \frac{\bar{y}_\delta^* - y^*}{\delta} \approx -J(x, y)^{-\top} \nabla_y f(x, y^*)$$

Thus, the perturbed equilibrium $\bar{y}_\delta^*$ satisfies $\bar{y}_\delta^* \approx y^*$ and thus $\frac{d}{d\delta} \bar{y}_\delta^* |_{\delta=0} \approx -J(x, y^*)^{-\top} \nabla_y f(x, y^*)$, as required by the Stackelberg hypergradient. As a consequence, we use the finite difference $v_\delta(x, \bar{y}_\delta^*)$ induced by the perturbed equilibrium to recover the correct Jacobian:

Even though the construction resolves the Jacobian mismatch, it does not yield a fully first-order algorithm immediately. This is because the auxiliary game operator $\bar{V}$ does not necessarily inherit strong monotonicity from V. Thus the next technical challenge is to prove that the auxiliary game is strongly monotone locally around the neighborhood of the true equilibrium y^* (Lemma F.1), and that the perturbed equilibrium $\bar{y}_\delta^*$ stays inside this neighborhood (Lemma F.2). The following theorem puts the ideas together and shows that $v_\delta(x, \bar{y}_\delta^*)$ is a fully first-order approximation with a non-asymptotic rate.

Theorem 5.2 (Non-Asymptotic hypergradient approximation)**.** *For $0 < \delta < \frac{\mu}{2C_f}$ and $0 < R < \frac{\mu - \delta C_f}{S(N+1)(2+\sqrt{N})}$ and $0 < \epsilon < 1$, $\bar{V}$ is locally $\bar{\mu}$-strongly monotone and $\bar{C}$-smooth on $\bar{B}_R(y^*) = \{y \in \mathbb{R}^n : \|y - y^*\| \leq R\}$ for some $\bar{\mu}, \bar{C} > 0$. Furthermore, if $\delta < \min(\frac{\mu}{2C_f}, \frac{\bar{\mu} R}{2L_f})$, then we have a fully first-order hypergradient approximation with a non-asymptotic rate:*

$$\|v_\delta(x, \bar{y}_\delta^*) - \tfrac{\partial}{\partial x} V(x, y^*)^\top J(x, y^*)^{-\top} (-\nabla_y f(x, y^*))\| \leq \mathcal{O}(\tfrac{N\delta}{\bar{\mu}^2} + \epsilon).$$

Proof Sketch. Lemma F.1 shows that the auxiliary operator $\bar{V}$ is locally $\bar{\mu}$-strongly monotone and $\bar{C}$-smooth on $\bar{B}_R(y^*)$. Under the stated bound on δ, Lemma F.2 shows that the map $T(y) = y - \eta\bar{V}(x, y)$ is a contraction on $\bar{B}_R(y^*)$. Hence the auxiliary game admits a unique Nash equilibrium $\bar{y}_\delta^*$ in this neighborhood with $\|\bar{y}_\delta^* - y^*\| \leq \frac{2L_f\delta}{\bar{\mu}}$. Now we relate the displacement of this auxiliary equilibrium to the exact hypergradient. Let $v_\epsilon^* := (J(x, y^*)^\top + \epsilon D(x, y^*))^{-1}(-\nabla_y f(x, y^*))$ and $v^* := J(x, y^*)^{-\top}(-\nabla_y f(x, y^*))$. Lemma F.3 shows that the normalized equilibrium displacement approximates v_ϵ^*: $\|\frac{\bar{y}_\delta^* - y^*}{\delta} - v_\epsilon^*\| \leq \mathcal{O}(\frac{\sqrt{N}\delta}{\bar{\mu}^2})$. Lemma F.4 controls the bias introduced by the ϵD regularization: $\|v_\epsilon^* - v^*\| \leq \mathcal{O}(\epsilon)$. It remains to relate the equilibrium displacement to the finite-difference expression. By Taylor expanding $v_\delta(x, \bar{y}_\delta^*)$ around y^*, we have $v_\delta(x, \bar{y}_\delta^*) = \frac{\partial}{\partial x}V(x, y^*)^\top \frac{\bar{y}_\delta^* - y^*}{\delta} + r_\delta$ where $\|r_\delta\| \leq \mathcal{O}(\frac{N\delta}{\bar{\mu}^2})$. The sum of these three terms gives an upper bound on $\|v_\delta(x, \bar{y}_\delta^*) - \frac{\partial}{\partial x}V(x, y^*)^\top v^*\|$ which yields the result. □

Algorithm 1 summarizes our fully first-order method for both symmetric and asymmetric Stackelberg games. At each outer iteration, we approximately solve the original follower game and the perturbed game, construct the finite-difference hypergradient estimator, and update the leader variable by inexact gradient descent. We now analyze the algorithm's convergence.

Theorem 5.3. *Consider Problem P1. Under Assumptions 3.4 and 3.5, Algorithm 1 outputs an $\mathcal{O}(\epsilon)$-stationary point of the hyperobjective F with $\tilde{\mathcal{O}}(N^2\epsilon^{-2})$ oracle calls to f and g_i.*

Proof. By the unconstrained version of Theorem B.7, the hyperobjective F is smooth. Choose $\delta \leq \frac{\epsilon}{2N}$ and use the potential function $\phi_{\epsilon/2}$. By Lemma D.1, if the approximate lower-level solutions $\hat{y}, \hat{y}_\delta$ satisfy $\|y^* - \hat{y}\| \leq \delta^2$ and $\|\bar{y}_\delta^* - \hat{y}_\delta\| \leq \delta^2$, then the inexact hypergradient $\hat{\nabla}F(x) = \nabla_x f(x, \hat{y}) + \hat{v}_\delta(x, \hat{y}_\delta)$ satisfies $\|\hat{\nabla}F(x) - \nabla F(x)\| \leq \mathcal{O}(N\delta + \epsilon/2) \leq \mathcal{O}(\epsilon)$. Therefore, by Lemma D.2, inexact gradient descent requires $\mathcal{O}(\epsilon^{-2})$ hypergradient evaluations to obtain an $\mathcal{O}(\epsilon)$-stationary point of F. It remains to bound the cost of each hypergradient evaluation. Since the potential function is constructed using the approximate equilibrium $\hat{y}$, its gradient operator is also inexact, with an error of $\mathcal{O}(\epsilon)$ (Lemma G.6). Moreover, the auxiliary game is only locally strongly monotone and smooth on $\bar{B}_R(y^*)$. We therefore choose $\delta \leq \min\{\epsilon/(2N), \bar{\mu}R/(4L_f), \sqrt{R/2}\}$, which ensures $\|\bar{y}_\delta^* - \hat{y}\| \leq \|\bar{y}_\delta^* - y^*\| + \|y^* - \hat{y}\| \leq R$. Hence, gradient descent initialized at $\hat{y}$ remains in the local neighborhood and converges linearly to the auxiliary equilibrium. For each call to $\hat{\nabla}F(x)$, $\mathcal{O}(\log(1/\delta^2))$ lower-level iterations are sufficient to obtain the required δ^2-accurate solutions. Each evaluation of the auxiliary game operator requires gradients of all N surrogate costs, and evaluating each surrogate cost requires $\mathcal{O}(N)$ calls to the original follower objectives g_i. Thus, each lower-level iteration uses $\mathcal{O}(N^2)$ first-order oracle calls. Combining this with $\mathcal{O}(\epsilon^{-2})$ outer iterations and $\mathcal{O}(\log(1/\delta^2))$ lower-level iterations gives $\mathcal{O}\big(N^2\epsilon^{-2}\log(1/\delta^2)\big) = \tilde{\mathcal{O}}(N^2\epsilon^{-2})$. □

Note that, compared to the symmetric case, there is an extra factor of N due to the sum over all N players' objectives in the potential function.

5.2 Constrained Stackelberg game

We now extend the framework to follower games with inequality constraints. For each player i, let $Y_i = \{y_i \in \mathbb{R}^{n_i} : e^{(i)}(x, y_i) := A^{(i)}x + B^{(i)}y_i + b^{(i)} \leq 0\}$. Linear inequality constraints are more challenging because the hyperobjective may be piecewise smooth. However, only the active inequalities affect the local equilibrium sensitivity, so the problem can be reduced locally to a linear equality-constrained Stackelberg game. Appendix H formalizes this reduction. Thus, it suffices to consider the case where Y_i is equality-constrained.

Definition 5.4 (Functional constraints)**.** Each $e^{(i)} : X \times Y_i \to \mathbb{R}^{m_i}$ is a function denoting equality constraints for each player. Each $e^{(i)}$ is private to player i. Denote $m := \sum_{i=1}^N m_i$. Denote $\lambda := (\lambda_1, \cdots, \lambda_N) \in \mathbb{R}^m$ as the Lagrange multipliers for constraints of the players. Denote $\lambda^* := (\lambda_1^*, \cdots, \lambda_N^*) \in \mathbb{R}^m$ as the Lagrange multipliers satisfying the KKT condition of $\min_{y_i \in Y_i(x)} g_i(x, y_i, y_{\neg i}^*), \forall i \in [N]$.

In this setting, we use the same potential ϕ_ϵ as Definition 5.1 and define a constrained auxiliary game $\bar{\mathcal{G}}_\delta$ by $\bar{g}_i(x, y) = \phi_\epsilon(x, y) - g_i(x, y) - (\lambda^*)^\top By + \delta f(x, y)$. It follows that $\bar{\mathcal{G}}_\delta$ has the

same local strong monotonicity modulus $\bar{\mu}$ and local smoothness modulus $\bar{C}$ on the ball $\bar{B}_R(y^*)$ as the unconstrained case. Let $\bar{y}_\delta^*$ be the Nash equilibrium of $\bar{\mathcal{G}}_\delta$ and $\bar{\lambda}_\delta^*$ be the corresponding Lagrange multiplier. We modify the finite difference to address the constraints as: $v_\delta(x, \bar{y}_\delta^*, \bar{\lambda}_\delta^*) = \frac{1}{\delta}\sum_{i=1}^N [\nabla_x g_i(x, \bar{y}_{\delta,i}^*, y_{\neg i}^*) - \nabla_x g_i(x, y^*) + \frac{\partial}{\partial x} e^{(i)}(x, y_i^*)^\top (\bar{\lambda}_\delta^* - \lambda^*)]$

The constrained analysis follows the same argument as in the unconstrained case after replacing the game Jacobian by the corresponding KKT system. The transpose-aware construction in Section 5.1 therefore yields the same hypergradient approximation error order. The resulting auxiliary game remains locally strongly monotone, so the lower-level equilibrium can be computed to the required accuracy by first-order methods, and the approximate hypergradient can be used in inexact gradient descent. The full approximation and complexity guarantees are deferred to Appendices G and H.

Algorithm 1 First-Order Algorithm for Solving Stackelberg Game with Multiple Followers

notation: Let $S_{\mathcal{M}}(x, y) : X \times \mathbb{R}^n \to Y(x)$ be a solver that returns a Nash equilibrium of an arbitrary game $\mathcal{M}(x)$ given initial iterate y. Define $\mathcal{Q}_\delta := \mathcal{G}_\delta$ if $\frac{\partial}{\partial y} V(x, y^*(x))$ is symmetric, otherwise define $\mathcal{Q}_\delta := \bar{\mathcal{G}}_\delta$.
initialize: Initial upper-level variable $x_0 \in X$, initial lower-level variable $y_0 \in \mathbb{R}^n$, finite difference parameter $0 < \delta < 1$, potential function parameter $0 < \epsilon < 1$, learning rate $\eta > 0$.

1: **for** $t = 0, 1, 2, \ldots, T-1$ **do**
2: Compute Nash equilibrium $\hat{y} = S_{\mathcal{G}}(x_t, y_t)$ (and the dual solution $\hat{\lambda}$ if constrained)
3: Compute perturbed Nash equilibrium $\hat{y}_\delta = S_{\mathcal{Q}_\delta}(x_t, \hat{y})$ (and the dual solution $\hat{\lambda}_\delta$ if constrained).
4: Compute approximate hypergradient $\hat{\nabla} F(x_t) = \nabla_x f(x_t, \hat{y}) + \hat{v}_\delta(x, \hat{y}_\delta)$ (use $\hat{v}_\delta(x, \hat{y}_\delta, \hat{\lambda}_\delta)$ if constrained)
5: $y_{t+1} = \hat{y}$
6: $x_{t+1} = x_t - \eta \hat{\nabla} F(x_t)$
7: **end for**
8: return x_t, y_t

6 Learning Non-truthful Auctions

In this section, we instantiate the general first-order Stackelberg framework for learning non-truthful auction mechanisms, where the mechanism is parameterized by neural networks, and bidders respond strategically. Figure 1 and Algorithm 1 illustrate our framework.

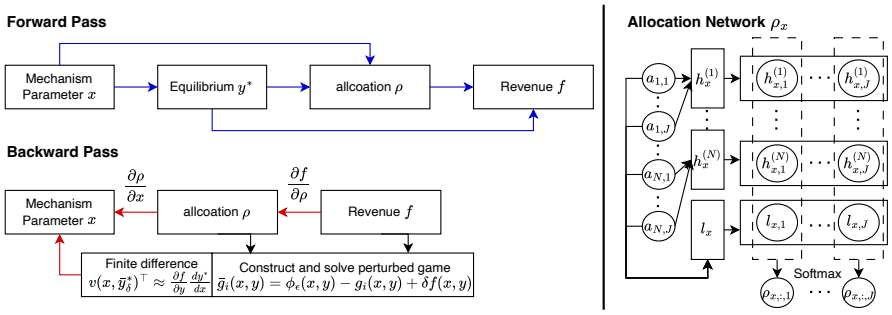


Figure 1: In forward pass, we compute the Nash equilibrium y^* of the original game and compute the revenue $f(x, y^*)$. $a_{i,j}$ is the bidder i's bid on item j. In the backward pass, we solve the perturbed game defined by $\bar{g}_i(x, y) = \phi_\epsilon(x, y) - g_i(x, y) + \delta f(x, y)$ and compute $v_\delta(x, \bar{y}_\delta^*) \approx \frac{d}{dx} y^*(x)^\top \nabla_y f(x, y^*)$.

6.1 Parameterization of Auction Mechanism

An N-bidder, J-item mechanism is given by an expected payment rule $P_x : \prod \mathcal{A}_i \to \mathbb{R}^N$ and a randomized allocation function $\rho_x : \prod \mathcal{A}_i \to \prod_{j=1}^J \Delta^{N+1}$. Without loss of generality, we assume

$0 \leq t_i \leq \omega_i, \forall i \in [N]$, for some $\omega_i \in \mathbb{R}_+^J$. Let $\rho_{x,i,j}$ denote the probability that player i receives item j. Define the additive expected utility function as $u_i(x, y, t_i) = \sum_{j=1}^{J} t_{i,j} \rho_{x,i,j}(\{\beta_{y_k}(t_k)\}_{k\in[N]}) - P_{x,i}(\{\beta_{y_k}(t_k)\}_{k\in[N]})$, with the bidding strategies $\{\beta_{y_k}\}_{k=1}^N$ being neural networks. We model the mechanism with temperature $\tau > 0$, neural networks $l_x : \prod \mathcal{A}_i \to \mathbb{R}^J$, and $h_x^{(i)} : \mathcal{A}_i \to \mathbb{R}^J$.

$$\rho_{x,i,j}(a) = \frac{\exp(h_{x,j}^{(i)}(a_i)/\tau)}{\sum_{k=1}^{N} \exp(h_{x,j}^{(k)}(a_k)/\tau) + \exp(l_{x,j}(a)/\tau)}, \quad P_{x,i}(a) = \sum_{j=1}^{J} \rho_{x,i,j}(a) a_{i,j}$$

Similar to a first-price auction, bidders pay their own bids if they receive an item. The function l_x controls the probability that no items are allocated.

6.2 Tracking a Local Equilibrium Branch

Algorithm 1 is used to find optimal non-truthful auctions. However, Nash equilibria may not be unique or even exist in auction games (Bichler et al., 2023b), in which case Problem P1 is ill-defined. When there exists multiple Nash equilibria, we select a differentiable branch of Nash equilibria $z^* : X \to Y(x)$ such that $z^*(x) \in \text{NE}(\mathcal{G}(x)), \forall x \in X$, and formulate the following Stackelberg game: $\min_{x\in X} F(x) := f(x, z^*(x)) := -\mathbb{E}_t \left[\sum_{i=1}^{N} P_{x,i}(\{\beta_{z_j^*(x)}(t_j)\}_{j=1}^N)\right]$. When Nash equilibrium does not exist, we may replace $\text{NE}(\mathcal{G})$ with $\text{LNE}(\mathcal{G})$. The problem is how we can select a differentiable branch of local Nash equilibria. We may assume local strong monotonicity around every local Nash equilibrium $y^*(x)$. Local Nash equilibria vary smoothly if the initial iterate is accurate and the mechanism parameters vary slightly (the hyperobjective's learning rate η is small). To ensure the initial game can be solved accurately, we pretrain the mechanism neural network to the allocation rule of a simultaneous first-price auction. We also initialize the lower-level solution as a parameter y that induces a truthful bidding strategy ($\beta_{y_i}(t_i) = t_i, \forall t_i \in \mathcal{T}_i$). The following theorem formalizes the warm-start strategy for auction design (see the proof in Appendix K).

Theorem 6.1. *Suppose $\mathcal{G}(x)$ is locally strongly monotone on a ball $\bar{B}_r(y^*(x))$ with radius $r > 0$ for all $y^*(x) \in$ LNE$(\mathcal{G}(x))$ and $x \in X$. Given initial iterate y, let $S(x, y, \hat{\delta}) : X \times \mathbb{R}^n \times \mathbb{R}_+ \to Y(x)$ be a solver can always return $\hat{y}$ such that $\|\hat{y} - y^*(x)\| \leq \hat{\delta}$ whenever $\|y - y^*(x)\| \leq r$ and $y^*(x) \in$ LNE$(\mathcal{G}(x))$.*

In Algorithm 1, suppose $\|y_0 - y^*(x_0)\| \leq \frac{r}{2}$ for some $y^*(x_0) \in$ LNE$(\mathcal{G}(x_0))$ and the learning rate η is sufficiently small, then $y_k = z^*(x_k)$, $\forall k \in [T]$ for some continuously differentiable function $z^* : X \to Y(x)$ such that $z^*(x) \in$ LNE$(\mathcal{G}(x))$.

7 Experiments

Experimental settings. We evaluate our method in three two-bidder auction settings. The first is a symmetric single-item auction, with bidders' values drawn independently and identically (iid) from Uniform$[0, 1]$. In this setting, a second-price auction with a reserve of 0.5 is revenue-optimal. A first-price auction with the same reserve achieves the same expected revenue at its symmetric Bayes-Nash equilibrium (Myerson, 1981; Krishna, 2009). The second setting is an asymmetric single-item auction, with values drawn independently from Uniform$[0, \frac{4}{3}]$ (strong bidder) and Uniform$[0, \frac{4}{5}]$ (weak bidder). Under the optimal auction (Myerson, 1981), the item is allocated to the bidder with the highest non-negative virtual value, where the strong and weak bidders' virtual values are $2t_1 - \frac{4}{3}$ and $2t_2 - \frac{4}{5}$, with corresponding reserve values of 0.667 and 0.4, respectively. The third setting has two items and additive bidders, with per-item valuations sampled iid from Uniform$[0, 1]$. We use GemNet (Wang et al., 2024) as a learned truthful benchmark; no closed-form revenue-optimal mechanism is known.

At the start of each training run, we sample $10,000$ random valuation profiles and reuse this fixed sample to estimate expected revenue throughout training. We replace softmax with max when reporting the expected revenue.

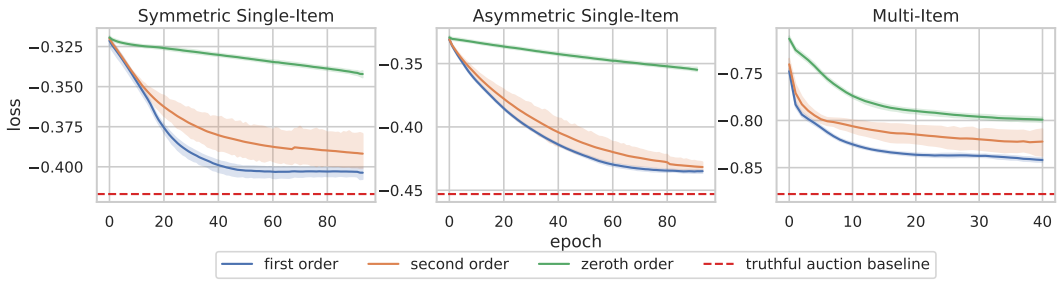


Figure 2: Convergence of negative expected revenue. Symmetric single-item auction (left), asymmetric single-item auction (middle), multi-item auction (right). The truthful multi-item auction baseline is GemNet.

Optimization performance. Figure 2 compares our first-order method with a second-order method (exact hypergradient) and a zeroth-order method (Kornowski & Shamir, 2024). Across all three settings, our method converges faster to a higher revenue. Moreover, the second-order method may be more sensitive to bias and noise based on its high variance. Figure 3 evaluates computational scaling by increasing the number of bidders in the symmetric single-item setting. Our first-order method has a significantly lower backward-pass runtime compared to the second-order implementation. This is consistent with the growing computational cost of the linear system required by the second-order method.

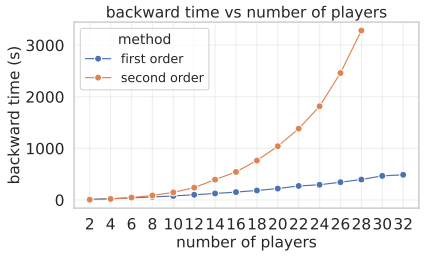


Figure 3: Backward time vs number of players in symmetric single-item auction

Comparison with truthful benchmarks. Our goal in these experiments is to assess whether optimizing pay-your-bid mechanisms can achieve revenues close to truthful benchmarks while retaining a simple, transparent payment rule—-not to establish a revenue advantage of non-truthfulness itself. Indeed, Table 1 demonstrates that the revenues of our non-truthful auction approximate those of the optimal auctions in single-item settings and GemNet in the multi-item setting. Appendix K and Figure 4 include allocation visualizations that provide qualitative comparisons with these benchmarks.

Table 1: Expected revenue comparison

Methods	Symmetric 2-bidder 1-item	Asymmetric 2-bidder 1-item	2-bidder 2-item
Truthful Baselines	0.417 (optimal)	0.453 (optimal)	0.878 (GemNet)
Ours	0.404	0.435	0.842

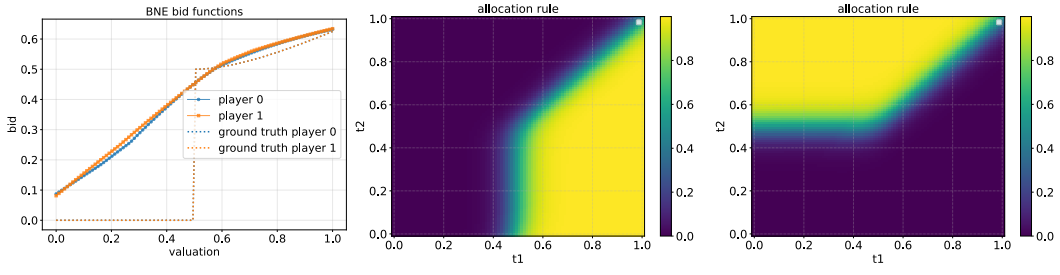


Figure 4: Equilibrium bidding strategy (left) and allocation rule (middle, right) for symmetric single-item auction. For allocation probability, x-axis and y-axis are the bid of bidder 1 and bidder 2 respectively. As shown in the leftmost plot, the learned auction is non-truthful, yet its revenue matches second-price auction with optimal reserve price of 0.5.

8 Conclusion

We formulate non-truthful auction design as a Stackelberg game and develop a novel fully first-order algorithm to solve it. Our key contribution is establishing non-asymptotic convergence guarantees even when the Jacobian is asymmetric. Our algorithm discovers non-truthful auction with revenue close to optimal truthful auctions even when the mechanism and bidders are neural networks.

9 Statement of LLM Use

In this work, we used generative AI tools for assistance with mathematical derivation and method implementation. We have not used generative AI tools for other tasks with required disclosure. Additionally, we used generative AI tools for assistance with debugging and copy-editing to improve readability. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

10 Ethics Statement

Our first order algorithm can be used to solve Stackelberg games, which is a widely used model in Economics. We recommend domain-specific validation when our algorithm is deployed in the real world.

11 Reproducibility Statement

For reproducibility, all assumptions and mathematical proofs are provided in the main text and Appendix with sufficient explanation. Experimental details are listed in Appendix K.

References

Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Batya Berzack, Rotem Oshman, and Inbal Talgam-Cohen. Optimal auctions for constrained buyers. *arXiv preprint arXiv:2606.30776*, 2026.

Martin Bichler, Maximilian Fichtl, Stefan Heidekrüger, Nils Kohring, and Paul Sutterer. Learning equilibria in symmetric auction games using artificial neural networks. *Nature machine intelligence*, 3(8):687–695, 2021.

Martin Bichler, Nils Kohring, and Stefan Heidekrüger. Learning equilibria in asymmetric auction games. *INFORMS Journal on Computing*, 35(3):523–542, 2023a.

Martin Bichler, Stephan B Lunowa, Matthias Oberlechner, Fabian R Pieroth, and Barbara Wohlmuth. On the convergence of learning algorithms in bayesian auction games. *arXiv preprint arXiv:2311.15398*, 2023b.

Jason Bigler. An update on first price auctions for google ad manager. *Retrieved October*, 22:2025, 2019.

Frank H Clarke. *Optimization and nonsmooth analysis*. SIAM, 1990.

Vincent Conitzer and Tuomas Sandholm. Complexity of mechanism design. *arXiv preprint cs/0205075*, 2002.

Vincent Conitzer and Tuomas Sandholm. Self-interested automated mechanism design and implications for optimal combinatorial auctions. In *Proceedings of the 5th ACM Conference on Electronic Commerce*, pp. 132–141, 2004.

Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *International Conference on Machine Learning*, pp. 6643–6670. PMLR, 2023.

Constantinos Daskalakis, Alan Deckelbaum, and Christos Tzamos. The complexity of optimal mechanism design. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pp. 1302–1318. SIAM, 2014.

Paul Dütting, Zhe Feng, Harikrishna Narasimhan, David Parkes, and Sai Srivatsa Ravindranath. Optimal auctions through deep learning. In *International Conference on Machine Learning*, pp. 1706–1715. PMLR, 2019.

Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

Zhe Feng, Harikrishna Narasimhan, and David C Parkes. Deep learning for revenue-optimal auctions with budgets. In *Proceedings of the 17th international conference on autonomous agents and multiagent systems*, pp. 354–362, 2018.

Zhe Feng, Sébastien Lahaie, Jon Schneider, and Jinchao Ye. Reserve price optimization for first price auctions in display advertising. In *International Conference on Machine Learning*, pp. 3230–3239. PMLR, 2021.

Rigel Galgana and Negin Golrezaei. Learning in repeated multiunit pay-as-bid auctions. *Manufacturing & Service Operations Management*, 27(1):200–229, 2025.

Shumpei Goke, Gabriel Y Weintraub, Ralph Mastromonaco, and Sam Seljan. Bidders' responses to auction format change in internet display advertising auctions. *arXiv preprint arXiv:2110.13814*, 2021.

Allen A Goldstein. Optimization of lipschitz continuous functions. *Mathematical Programming*, 13 (1):14–22, 1977.

Matt Harada. The ad exchange's place in a first-price world. *Marketing Technology Insights*, 2018. URL https://martechseries.com/mts-insights/guest-authors/ad-exchanges-place-first-price-world/.

Kexin Huang, Ziqian Chen, Xue Wang, Chongming Gao, Jinyang Gao, Bolin Ding, and Xiang Wang. Learning bayesian nash equilibrium in auction games via approximate best response. In *Forty-second International Conference on Machine Learning*, 2025.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *Advances in neural information processing systems*, 30, 2017.

Nils Kohring, Fabian Raoul Pieroth, and Martin Bichler. Enabling first-order gradient-based learning for equilibrium computation in markets. In *International Conference on Machine Learning*, pp. 17327–17342. PMLR, 2023.

Guy Kornowski and Ohad Shamir. An algorithm with optimal dimension-dependence for zero-order nonsmooth nonconvex stochastic optimization. *Journal of Machine Learning Research*, 25(122): 1–14, 2024.

Guy Kornowski, Swati Padmanabhan, Kai Wang, Zhe Zhang, and Suvrit Sra. First-order methods for linearly constrained bilevel optimization. *Advances in Neural Information Processing Systems*, 37: 141417–141460, 2024.

Vijay Krishna. *Auction theory*. Academic press, 2009.

Jeongyeol Kwon, Dohyun Kwon, Stephen Wright, and Robert D Nowak. A fully first-order method for stochastic bilevel optimization. In *International Conference on Machine Learning*, pp. 18083–18113. PMLR, 2023.

Roger B Myerson. Optimal auction design. *Mathematics of operations research*, 6(1):58–73, 1981.

April Niu, Kai Wang, and Juba Ziani. Finding a multiple follower stackelberg equilibrium: A fully first-order method. *arXiv preprint arXiv:2509.08161*, 2025.

Mai Pham, Vikrant Vaze, and Peter Chin. Advancing differentiable economics: A neural network framework for revenue-maximizing combinatorial auction mechanisms. *arXiv preprint arXiv:2501.19219*, 2025.

Jean-Charles Rochet and Philippe Choné. Ironing, sweeping, and multidimensional screening. *Econometrica*, pp. 783–826, 1998.

Vinzenz Thoma, Georgios Piliouras, and Luke Marris. Deep incentive design with differentiable equilibrium blocks. *arXiv preprint arXiv:2603.07705*, 2026.

Kai Wang, Lily Xu, Andrew Perrault, Michael K Reiter, and Milind Tambe. Coordinating followers to reach better equilibria: End-to-end gradient descent for stackelberg games. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 5219–5227, 2022.

Tonghan Wang, Yanchen Jiang, and David C Parkes. Gemnet: Menu-based, strategy-proof multi-bidder auctions through deep learning. *arXiv preprint arXiv:2406.07428*, 2024.

Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *International Conference on Machine Learning*, pp. 11173–11182. PMLR, 2020.

Zihao Zhao, Kai-Chia Mo, Shing-Hei Ho, Brandon Amos, and Kai Wang. A fully first-order layer for differentiable optimization. *arXiv preprint arXiv:2512.02494*, 2025.

APPENDIX

A AUXILIARY LEMMAS

Lemma A.1 (Variational inequality characterization of Nash Equilibrium)**.** *Suppose feasible set $Y_i(x)$ is convex, for all $i \in [N]$. If $(y_1^*, \cdots, y_N^*) \in$ LNE$(\mathcal{G}(x))$, then*

$$\forall y_i \in U_i : \langle \nabla_{y_i} g_i(x, y_i^* y_{\neg i}^*), y_i - y_i^* \rangle \geq 0$$

for some neighborhood $U_i \subseteq Y_i$ of y_i^*, for all $i \in [N]$. Under Assumption 3.4, each g_i is convex in y_i, the reverse direction also holds with $(y_1^*, \cdots, y_N^*) \in$ NE$(\mathcal{G}(x))$. Thus, local Nash equilibrium is a necessary and sufficient condition for Nash equilibrium in a strongly monotone and smooth game.

Proof. Let $U = \Pi_i U_i$ and $Y(x) = \Pi_i Y_i(x)$. Suppose $y^* = (y_1^*, \cdots, y_N^*) \in \text{LNE}(\mathcal{G}(x))$, then for all $i \in [N]$, $y_i^* \in \arg\min_{y_i \in U_i} g_i(x, y_i, y_{\neg i}^*)$ is a local minimum. Then, $\langle \nabla_{y_i} g_i(x, y_i^*, y_{\neg i}^*), d \rangle \geq 0$ for all feasible directions $d \in \mathbb{R}^{n_i}$ such that $y_i^* + td \in U_i$ for some $t > 0$. Since $U_i \subseteq Y_i(x)$ is convex, $\langle \nabla_{y_i} g_i(x, y_i^*, y_{\neg i}^*), y_i - y_i^* \rangle \geq 0$, for all $y_i \in U_i$. Then $\langle V(x, y^*), y - y^* \rangle \geq 0$, for all $y \in U$.

Suppose $\langle V(x, y^*), y - y^* \rangle \geq 0$, for all $y \in Y(x)$. Then, for all $i \in [N]$, $\langle \nabla_{y_i} g_i(x, y_i^*, y_{\neg i}^*), y_i - y_i^* \rangle \geq 0$. Hence, y_i^* is a stationary point. If V is monotone (which implies each g_i is convex in y_i), then $y_i^* \in \arg\min_{y_i \in Y_i(x)} g_i(x, y_i, y_{\neg i}^*)$ is global minimum, then $y^* \in \text{NE}(\mathcal{G}(x)) \subseteq \text{LNE}(\mathcal{G}(x))$. □

Since feasible set for each player is represented by linear constraints, we provide an alternative characterization of Nash equilibrium using KKT condition,

Corollary A.2 (KKT characterization of Nash Equilibrium)**.** *Under Assumption G.1, if $(y_1^*, \cdots, y_N^*) \in$ LNE$(\mathcal{G}(x))$, then there exists $\lambda^* := (\lambda_1^*, \cdots, \lambda_N) \in \mathbb{R}^m$ such that*

$$(y_i^*, \lambda_i^*) \text{ satisfies the KKT condition of } \min_{y_i \in Y_i(x)} g_i(x, y_i, y_{\neg i}^*), \forall i \in [N]$$

Under Assumption 3.4, each cost function g_i is convex in y_i. There exists optimal dual solution $\lambda^* := (\lambda_1^*, \cdots, \lambda_N) \in \mathbb{R}^m$ such that $(y_1^*, \cdots, y_N^*) \in$ NE$(\mathcal{G})$ if and only if

$$(y_i^*, \lambda_i^*) \text{ satisfies the KKT condition of } \min_{y_i \in Y_i(x)} g_i(x, y_i, y_{\neg i}^*), \forall i \in [N]$$

Lemma A.3 (Fixed point characterization of variational inequality)**.** *Let $Y \subseteq \mathbb{R}^n$ be a closed, convex set. Denote the Euclidean projection operator as $\Pi_Y(z) = \arg\min_{y \in Y} \frac{1}{2}(y - z)^\top (y - z)$. Let $V : \mathbb{R}^n \to \mathbb{R}^n$. Let $\eta > 0$ be an arbitrary number. $y^* \in Y$ satisfies $\langle V(y^*), y - y^* \rangle \geq 0$, $\forall y \in Y$ if and only if $y^* = \Pi_Y(y^* - \eta V(y^*))$*

Proof. By optimality condition of projection, $y^* = \Pi_Y(y^* - \eta V(y^*))$ if and only if $\forall y \in Y$,

$$\begin{aligned}\langle y^* - (y^* - \eta V(y^*)), y - y^* \rangle &\geq 0 \\ \langle V(y^*), y - y^* \rangle &\geq 0\end{aligned}$$

□

Lemma A.4. *Let $A \in \mathbb{R}^{m \times n}$. Let $\sigma_{\max}(A) = \sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_R(A) = \sigma_{\min}(A) > 0$ be non-zero singular values of A. Then,*

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sigma_{\max}(A)$$

If A is an invertible square matrix or A has full column rank, then

$$\min_{x \neq 0} \frac{\|Ax\|}{\|x\|} = \sigma_{\min}(A)$$

Proof. Let $A = U\Sigma V^\top$ be the singular value decomposition of A where $U \in \mathbb{R}^{m \times m}$, $V \in \mathbb{R}^{n \times n}$ and $\Sigma \in \mathbb{R}^{m \times n}$. U and V are orthogonal matrix with orthonormal columns. $\Sigma = \text{diag}(\sigma_1, \cdots, \sigma_R, 0, \cdots, 0)$. Let x be a unit vector,

$$\begin{aligned}\|Ax\|^2 &= x^\top A^\top A x \\ &= x^\top V \Sigma^\top U^\top U \Sigma V^\top x \\ &= x^\top V \Sigma^\top \Sigma V^\top x\end{aligned}$$

Let $y = V^\top x$, $\|y\| = \|x\| = 1$,

$$\begin{aligned}x^\top A^\top A x &= y^\top \Sigma^\top \Sigma y \\ &\leq \sigma_{\max}^2(A)\end{aligned}$$

which is an achievable upper bound if we choose x to be the corresponding right singular vector. If A is an invertible square matrix, then the diagonal Σ does not contain any zeros. If A has full column rank, then $\Sigma = \begin{bmatrix}\text{diag}(\sigma_1, \cdots, \sigma_R) \\ 0\end{bmatrix}$, and the diagonal of $\Sigma^\top \Sigma$ does not contain any zeros. We can minimize over x to get the minimum singular value. □

Lemma A.5. *If the function $V(y)$ is μ-strongly monotone with respect to $y \in \mathbb{R}^n$, then $\langle d, \frac{d}{dy} V(y) d \rangle \geq \mu \|d\|^2$, $\forall y \in Y$, $\forall d \in \mathbb{R}^n$. Furthermore, $\|\frac{d}{dy} V(y)\| \geq \sigma_{\min}(\frac{d}{dy} V(y)) \geq \mu$ and $\|\frac{d}{dy} V(y)^{-1}\| = \sigma_{\max}(\frac{d}{dy} V(y)^{-1}) \leq \frac{1}{\mu}$.*

Proof. Let $d \in \mathbb{R}^n$ be an arbitrary vector,

$$\begin{aligned}\langle V(y + td) - V(y), td \rangle &\geq \mu \|td\|^2 \\ \frac{\langle V(y + td) - V(y), d \rangle}{t} &\geq \mu \|d\|^2 \\ \langle \frac{d}{dy} V(y) d, d \rangle &\geq \mu \|d\|^2 \quad \text{(take } t \to 0\text{)}\end{aligned}$$

Thus, $\frac{d}{dy} V(y)$ is invertible and has full rank. The singular value decomposition of $\frac{d}{dy} V(y)$ is given by $U\Sigma V^\top$ where $U \in \mathbb{R}^{n \times n}$, $V \in \mathbb{R}^{n \times n}$ and $\Sigma \in \mathbb{R}^{n \times n}$. U and V are orthogonal matrix. $\Sigma = \text{diag}(\sigma_1, \cdots, \sigma_n)$. Following the proof in lemma A.4, we have

$$\min_{\|v\|=1} \|\frac{d}{dy} V(y) v\| = \sigma_{\min}(\frac{d}{dy} V(y))$$

Let $v \in \mathbb{R}^n$ be a unit vector with $\|v\| = 1$,

$$\begin{aligned}\|\frac{d}{dy}V(y)v\| &= \|\frac{d}{dy}V(y)v\|\|v\| \\ &\geq \langle \frac{d}{dy}V(y)v, v\rangle \\ &\geq \mu\|v\|^2 \\ &= \mu\end{aligned}$$

Taking minimum over v yields $\sigma_{\min}(\frac{d}{dy}V(y)) \geq \mu$. Inverting $\frac{d}{dy}V(y)$ using its singular value decomposition yields $\|\frac{d}{dy}V(y)^{-1}\| = \sigma_{\max}(\frac{d}{dy}V(y)^{-1}) \leq \frac{1}{\mu}$. □

Lemma A.6. *If the function $V(y)$ is C-Lipschitz with respect to $y \in \mathbb{R}^n$, then $\|\frac{d}{dy}V(y)\| = \sigma_{\max}(\frac{d}{dy}V(y)) \leq C$. If $\frac{d}{dy}V(y)^{-1}$ exists, then $\|\frac{d}{dy}V(y)^{-1}\| \geq \sigma_{\min}(\frac{d}{dy}V(y)^{-1}) \geq \frac{1}{C}$.*

Proof. By definition of C-Lipschitzness and using Lemma A.4, $\|\frac{d}{dy}V(y)\| = \sigma_{\max}(\frac{d}{dy}V(y)) \leq C$. If $\frac{d}{dy}V(y)^{-1}$ exists, using singular value decomposition of a full rank matrix, we get $\sigma_{\min}(\frac{d}{dy}V(y)^{-1}) \geq \frac{1}{C}$. □

Lemma A.7. *Let G be a square matrix such that $\langle d, Gd\rangle \geq \mu\|d\|^2$, $\forall d \in \mathbb{R}^n$ for some $\mu > 0$. Let B be a matrix with full row rank, then $BG^{-1}B^\top$ is invertible. If $\sigma_{\max}(G) \leq C$, then $\sigma_{\min}(BG^{-1}B^\top) \geq \frac{\mu}{C^2}\sigma_{\min}(B)^2$ and $\sigma_{\max}(BG^{-1}B^\top) \leq \frac{\sigma_{max}(B)^2}{\mu}$.*

Proof. G is invertible. Let $0 \neq v \in \mathbb{R}^n$, $d = G^{-1}v$

$$v^\top G^{-1}v = v^\top d = d^\top G^\top d = d^\top Gd > 0$$

Let $0 \neq x \in \mathbb{R}^n$, suppose

$$\begin{aligned}BG^{-1}B^\top x &= 0 \\ x^\top BG^{-1}B^\top x &= 0\end{aligned}$$

Since B has full row rank and $v^\top G^{-1}v > 0$ for any $v \neq 0$, this results in a contradiction. Hence, $BG^{-1}B^\top$ is invertible.

$$\begin{aligned}\sigma_{\max}(BG^{-1}B^\top) &\leq \|B\|^2\|G^{-1}\| \\ &\leq \frac{\|B\|^2}{\mu} \\ &= \frac{\sigma_{\max}(B)^2}{\mu}\end{aligned}$$

Let x be a unit vector. Let $v = B^\top x$ and $d = G^{-1}v$,

$$\begin{aligned}\|BG^{-1}B^\top x\| &= \|BG^{-1}B^\top x\|\|x\| \\ &\geq x^\top BG^{-1}B^\top x \\ &= d^\top Gd \\ &\geq \mu\|d\|^2 \\ &= \mu\|G^{-1}B^\top x\|^2 \\ &\geq \frac{\mu}{C^2}\|B^\top x\|^2 \\ &\geq \frac{\mu}{C^2}\sigma_{\min}(B)^2\end{aligned}$$

Minimizing over x, we get $\sigma_{\min}(BG^{-1}B^\top) \geq \frac{\mu}{C^2}\sigma_{\min}(B)^2$ □

Lemma A.8. *Given a block matrix* $H = \begin{bmatrix} G & B^\top \\ B & 0 \end{bmatrix}$. *If* B *has full row rank and* $(\forall d \neq 0$ *s.t. Bd=0*$)$ $(d^\top G d > 0)$*, then* H *is invertible.*

Proof. We want to show that $H \begin{bmatrix} a \\ b \end{bmatrix} = 0 \implies \begin{bmatrix} a \\ b \end{bmatrix} = 0$. Suppose $H \begin{bmatrix} a \\ b \end{bmatrix} = 0$. Suppose $Ba \neq 0$, then $H \begin{bmatrix} a \\ b \end{bmatrix} = 0$ is impossible by construction. Suppose $Ba = 0$. If $a \neq 0$, $a^\top G a > 0$, then $\begin{bmatrix} a & b \end{bmatrix} H \begin{bmatrix} a \\ b \end{bmatrix} = a^\top G a + 2b^\top B a > 0$, so $H \begin{bmatrix} a \\ b \end{bmatrix} \neq 0$ which violates the hypothesis. If $a = 0$, then $H \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} B^\top b \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. Since $B^\top$ has full column rank, $b = 0$. □

B Lipschitzness and Smoothness of Stackelberg Game with Linear Equality Constraints

Lemma B.1 (Implicit differentiation of KKT system)**.**

$$\begin{bmatrix} \frac{d}{dx} y^*(x) \\ \frac{d}{dx} \lambda^*(x) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial y} V(x, y^*) & B^\top \\ B & 0 \end{bmatrix}^{-1} \begin{bmatrix} -\frac{\partial}{\partial x} V(x, y^*) \\ -A \end{bmatrix} \tag{7}$$

Proof. Consider problem P1. By strong monotonicity assumption 3.4, each cost function g_i is convex in y_i. Hence, KKT condition is the necessary and sufficient condition for each y_i^* to be the Nash equilibrium. Denote $B := \operatorname{diag}((\frac{\partial}{\partial y} e^{(i)}(x, y_i^*))_{i \in [N]}) \in \mathbb{R}^{m \times n}$ as a block diagonal matrix, and $A := \begin{bmatrix} \frac{\partial}{\partial x} e^{(1)}(x, y_1^*) \\ \vdots \\ \frac{\partial}{\partial x} e^{(N)}(x, y_N^*) \end{bmatrix} \in \mathbb{R}^{m \times p}$. The KKT condition is

$$\Gamma(x, y^*, \lambda^*) = \begin{bmatrix} V(x, y^*) + B^\top \lambda^* \\ \begin{bmatrix} e^{(1)}(x, y_1^*) \\ \vdots \\ e^{(N)}(x, y_N^*) \end{bmatrix} \end{bmatrix} = 0$$

By Lemma A.8, the KKT matrix $\frac{\partial}{\partial (y^*, \lambda^*)} \Gamma(x, y^*, \lambda^*) = \begin{bmatrix} \frac{\partial}{\partial y} V(x, y^*) & B^\top \\ B & 0 \end{bmatrix}$ is invertible. Finally, by implicit function theorem, $y^*(x)$, $\lambda^*(x)$ are continuously differentiable functions of x and their derivatives match the statement of this Lemma. □

Corollary B.2 (Hypergradient)**.**

$$\nabla F(x) = \nabla_x f(x, y^*) + \begin{bmatrix} \frac{\partial}{\partial x} V(x, y^*)^\top & A^\top \end{bmatrix} \begin{bmatrix} \frac{\partial}{\partial y} V(x, y^*) & B^\top \\ B & 0 \end{bmatrix}^{-\top} \begin{bmatrix} -\nabla_y f(x, y^*) \\ 0 \end{bmatrix} \tag{8}$$

Proof.

$$\begin{aligned} & (\frac{d}{dx} y^*(x))^\top \nabla_y f(x, y^*(x)) \\ &= \begin{bmatrix} \frac{d}{dx} y^*(x)^\top & \frac{d}{dx} \lambda^*(x)^\top \end{bmatrix} \begin{bmatrix} \nabla_y f(x, y^*) \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial}{\partial x} V(x, y^*)^\top & A^\top \end{bmatrix} \begin{bmatrix} \frac{\partial}{\partial y} V(x, y^*) & B^\top \\ B & 0 \end{bmatrix}^{-\top} \begin{bmatrix} -\nabla_y f(x, y^*) \\ 0 \end{bmatrix} \end{aligned}$$

□

Lemma B.3. *Suppose*

$$\begin{bmatrix} G & B^\top \\ B & 0 \end{bmatrix} \begin{bmatrix} d_y \\ d_\lambda \end{bmatrix} = \begin{bmatrix} -G_{xy} \\ -A \end{bmatrix} \tag{9}$$

where G is a square matrix satisfying $\langle d, Gd\rangle \geq \mu\|d\|^2$, $\forall d \in \mathbb{R}^n$ for some $\mu > 0$, and B has full row rank, then

$$d_y = (I - P_B^G)B^\dagger(-A) + P_B^G(-G^{-1}G_{xy})$$

where $P_B^G := I - G^{-1}B^\top(BG^{-1}B^\top)^{-1}B$

Proof. Expanding the equations,

$$d_y = -G^{-1}B^\top d_\lambda - G^{-1}G_{xy}, \tag{10}$$
$$Bd_y = -A. \tag{11}$$

Solving for d_λ, we have

$$\begin{aligned} d_\lambda &= (BG^{-1}B^\top)^{-1}[A - BG^{-1}G_{xy}] \\ &= (BB^\top)^{-1}BG(I - P_B^G)(B^\dagger A - G^{-1}G_{xy}) \\ &= (B^\dagger)^\top G(I - P_B^G)(B^\dagger A - G^{-1}G_{xy}) \end{aligned}$$

Solving for d_y, we have

$$\begin{aligned} d_y &= -G^{-1}B^\top(BG^{-1}B^\top)^{-1}[A - BG^{-1}G_{xy}] - G^{-1}G_{xy} \\ &= (I - P_B^G)B^\dagger(-A) + P_B^G(-G^{-1}G_{xy}) \end{aligned}$$

□

Lemma B.4. *Let $P_B^G(x) = I - G^{-1}B^\top(BG^{-1}B^\top)^{-1}B$, where square matrix G satisfies $\langle d, Gd\rangle \geq \mu_g\|d\|^2$ for some $\mu_g > 0$ and matrix B has full row rank. Given $\sigma_{\max}(G) \leq C_g$ for some C_g, let $\kappa_g = \frac{C_g}{\mu_g}$, then $\|P_B^G\| \leq 1 + \frac{\sigma_{\max}(B)^2}{\sigma_{\min}(B)^2}\kappa_g^2 := M_P$.*

Proof.

$$\begin{aligned} \|P_B^G\| &\leq 1 + \frac{\|B\|^2}{\mu_g}\frac{1}{\sigma_{\min}(BG^{-1}B^\top)} \\ &\leq 1 + \frac{\|B\|^2}{\mu_g}\frac{C_g^2}{\mu_g\sigma_{\min}(B)^2} \\ &= 1 + \frac{\sigma_{\max}(B)^2}{\sigma_{\min}(B)^2}\kappa_g^2 \end{aligned}$$

The first inequality uses Lemma A.5 and Lemma A.4. The second inequality uses Lemma A.7. The last equality uses Lemma A.4. □

Lemma B.5. *Let $P_B^G := I - G^{-1}B^\top(BG^{-1}B^\top)^{-1}B$, where square matrix G satisfies $\langle d, Gd\rangle \geq \mu_g\|d\|^2$ for some $\mu_g > 0$ and matrix B has full row rank. Given $\sigma_{\max}(G) \leq C_g$ for some C_g. Suppose another square matrix $\hat{G}$ also satisfies the above properties with the same constants μ_g and C_g. Let $\kappa_g = \frac{C_g}{\mu_g}$, then $\|P_B^G - P_B^{\hat{G}}\| \leq L_P\|G - \hat{G}\|$ where $L_P := \frac{C_g^2\sigma_{\max}(B)^2}{\mu_g^3\sigma_{\min}(B)^2} + \frac{\sigma_{\max}(B)^2}{\mu_g}\frac{C_g^4\|B\|^2}{\mu_g^4\sigma_{\min}(B)^4}$.*

Proof. Denote $M = BG^{-1}B^\top$ and $\hat{M} = B\hat{G}^{-1}B^\top$,

$$\begin{aligned} &P_B^{\hat{G}} - P_B^G \\ &= G^{-1}B^\top M^{-1}B - \hat{G}^{-1}B^\top\hat{M}^{-1}B \\ &= (G^{-1}B^\top M^{-1}B - \hat{G}^{-1}B^\top M^{-1}B) + (\hat{G}^{-1}B^\top M^{-1}B - \hat{G}^{-1}B^\top\hat{M}^{-1}B) \\ &= (G^{-1} - \hat{G}^{-1})B^\top M^{-1}B + \hat{G}^{-1}B^\top(M^{-1} - \hat{M}^{-1})B \end{aligned}$$

Now we proceed to bound $\|G^{-1} - \hat{G}^{-1}\|$ and $\|M^{-1} - \hat{M}^{-1}\|$ using the identity $C^{-1} - D^{-1} = C^{-1}(D - C)D^{-1}$ for any invertible matrices C and D,

$$\begin{aligned}&\|G^{-1} - \hat{G}^{-1}\|\\ &\leq \frac{1}{\mu_g^2}\|G - \hat{G}\|\end{aligned}$$

Using Lemma A.7, we have $\sigma_{\min}(BG^{-1}B^\top) \geq \frac{\mu_g}{C_g^2}\sigma_{\min}(B)^2$ and $\sigma_{\max}(BG^{-1}B^\top) \leq \frac{\sigma_{max}(B)^2}{\mu_g}$.

$$\begin{aligned}&\|M^{-1} - \hat{M}^{-1}\|\\ &\leq \frac{1}{\sigma_{\min}(BG^{-1}B^\top)^2}\|M - \hat{M}\|\\ &\leq \frac{C_g^4\|B\|^2}{\mu_g^2\sigma_{\min}(B)^4}\|G^{-1} - \hat{G}^{-1}\|\\ &\leq \frac{C_g^4\|B\|^2}{\mu_g^4\sigma_{\min}(B)^4}\|G - \hat{G}\|\end{aligned}$$

Now we combine the results together,

$$\begin{aligned}&\|P_B^{\hat{G}} - P_B^G\|\\ &\leq \|G^{-1} - \hat{G}^{-1}\|\|B\|^2\|M^{-1}\| + \|\hat{G}^{-1}\|\|B\|^2\|M^{-1} - \hat{M}^{-1}\|\\ &\leq (\frac{C_g^2\sigma_{\max}(B)^2}{\mu_g^3\sigma_{\min}(B)^2} + \frac{\sigma_{\max}(B)^2}{\mu_g}\frac{C_g^4\|B\|^2}{\mu_g^4\sigma_{\min}(B)^4})\|G - \hat{G}\|\end{aligned}$$

□

Lemma B.6 (Lipshitzness and smoothness of lower-level solution mappings)**.** *Denote $\kappa = \frac{C}{\mu}$, $y^*(x)$ is L_y-Lipschitz and C_y-smooth in x for some constants L_y and C_y. Moreover, $\lambda^*(x)$ is L_λ-lipschitz and C_λ-smooth in x for some constants L_λ and C_λ.*

Proof. Denote $G = \frac{\partial}{\partial y}V(x, y^*(x))$, $G_{xy} = \frac{\partial}{\partial x}V(x, y^*(x))$. Directly applying Lemma B.3 and Lemma B.4, we can prove Lipschitzness of $y^*(x)$,

$$\begin{aligned}\|\frac{d}{dx}y^*(x)\| &= \|(I - P_B^G)B^\dagger(-A) + P_B^G(-G^{-1}G_{xy})\|\\ &\leq \|I - P_B^G\|\|B^\dagger A\| + \|P_B^G\|\|G^{-1}\|\|G_{xy}\|\\ &\leq (1 + M_P)\|B^\dagger A\| + M_P\frac{C}{\mu}\\ &:= L_y\end{aligned}$$

Denote $\hat{y} = y^*(\hat{x})$, $\hat{G} = \frac{\partial}{\partial y}V(\hat{x}, \hat{y})$ and $\hat{G}_{xy} = \frac{\partial}{\partial x}V(\hat{x}, \hat{y})$. We can also prove smoothness of $y^*(x)$,

$$\begin{aligned}&\|\frac{d}{dx}y^*(\hat{x}) - \frac{d}{dx}y^*(x)\|\\ &= \|(I - P_B^{\hat{G}})B^\dagger(-A) + P_B^{\hat{G}}(-\hat{G}^{-1}\hat{G}_{xy}) - (I - P_B^G)B^\dagger(-A) - P_B^G(-G^{-1}G_{xy})\|\\ &\leq \|(P_B^G - P_B^{\hat{G}})B^\dagger(-A)\| + \|(P_B^{\hat{G}} - P_B^G)(-\hat{G}^{-1}\hat{G}_{xy})\| + \|P_B^G(G^{-1}G_{xy} - \hat{G}^{-1}\hat{G}_{xy})\|\\ &\leq L_P\|G - \hat{G}\|\|B^\dagger A\| + L_P\|G - \hat{G}\|\frac{C}{\mu} + M_P\|G^{-1}G_{xy} - \hat{G}^{-1}\hat{G}_{xy}\|\end{aligned}$$

where Lemma B.5 and B.4 are applied in the last inequality. Now we can bound $\|G - \hat{G}\|$. Denote $L_G := S(1 + L_y)$, we have

$$\begin{aligned}\|G - \hat{G}\| &\leq \|\frac{\partial}{\partial y}V(x,y) - \frac{\partial}{\partial y}V(\hat{x},y)\| + \|\frac{\partial}{\partial y}V(\hat{x},y) - \frac{\partial}{\partial y}V(\hat{x},\hat{y})\| \\ &\leq S(1+L_y)\|x - \hat{x}\| \\ &= L_G\|x - \hat{x}\|\end{aligned}$$

and we can bound the difference between unconstrained implict gradients $\|G^{-1}G_{xy} - \hat{G}^{-1}\hat{G}_{xy}\|$, denote $L_u := \frac{S(1+L_y)}{\mu} + \frac{CS(1+L_y)}{\mu^2}$, we have

$$\begin{aligned}\|G^{-1}G_{xy} - \hat{G}^{-1}\hat{G}_{xy}\| &\leq \|G^{-1}G_{xy} - G^{-1}\hat{G}_{xy}\| + \|G^{-1}\hat{G}_{xy} - \hat{G}^{-1}\hat{G}_{xy}\| \\ &\leq \frac{1}{\mu}\|G_{xy} - \hat{G}_{xy}\| + \|\hat{G}_{xy}\|\|G^{-1}(\hat{G} - G)\hat{G}^{-1}\| \\ &\leq (\frac{S(1+L_y)}{\mu} + \frac{CS(1+L_y)}{\mu^2})\|x - \hat{x}\| \\ &= L_u\|x - \hat{x}\|\end{aligned}$$

We have obtained the smoothness bound for $y^*(x)$,

$$\begin{aligned}&\|\frac{d}{dx}y^*(\hat{x}) - \frac{d}{dx}y^*(x)\| \\ &\leq (L_P L_G\|B^\dagger A\| + \frac{CL_P L_G}{\mu} + M_P(L_u))\|x - \hat{x}\| \\ &:= C_y\|x - \hat{x}\|\end{aligned}$$

Directly applying Lemma B.3 and Lemma B.4, we can prove Lipschitzness of $\lambda^*(x)$,

$$\begin{aligned}\|\frac{d}{dx}\lambda^*(x)\| &= \|(B^\dagger)^\top G(I - P_B^G)(B^\dagger A - G^{-1}G_{xy})\| \\ &\leq \|B^\dagger\|\|G\|(1 + M_P)(\|B^\dagger A\| + \frac{C}{\mu}) \\ &:= L_\lambda\end{aligned}$$

The smoothness of $\lambda^*(x)$ is given by

$$\begin{aligned}&\|\frac{d}{dx}\lambda^*(x) - \frac{d}{dx}\lambda^*(\hat{x})\| \\ &= \|B^\dagger G(I - P_B^G)(B^\dagger A - G^{-1}G_{xy}) - B^\dagger\hat{G}(I - P_B^G)(B^\dagger A - G^{-1}G_{xy})\| \\ &+ \|B^\dagger\hat{G}(I - P_B^G)(B^\dagger A - G^{-1}G_{xy}) - B^\dagger\hat{G}(I - P_B^{\hat{G}})(B^\dagger A - G^{-1}G_{xy})\| \\ &+ \|B^\dagger\hat{G}(I - P_B^{\hat{G}})(B^\dagger A - G^{-1}G_{xy}) - B^\dagger\hat{G}(I - P_B^{\hat{G}})(B^\dagger A - \hat{G}^{-1}\hat{G}_{xy})\| \\ &\leq \|B^\dagger\|(1 + M_P)(\|B^\dagger A\| + \frac{C}{\mu})\|G - \hat{G}\| + C\|B^\dagger\|(\|B^\dagger A\| + \frac{C}{\mu})\|P_B^{\hat{G}} - P_B^G\| + C\|B^\dagger\|(1 + M_P)\|G^{-1}G_{xy} - \hat{G}^{-1}\hat{G}_{xy}\| \\ &\leq (L_G\|B^\dagger\|(1 + M_P)(\|B^\dagger A\| + \frac{C}{\mu}) + L_P C\|B^\dagger\|(\|B^\dagger A\| + \frac{C}{\mu}) + CL_u\|B^\dagger\|(1 + M_P))\|x - \hat{x}\| \\ &:= C_\lambda\|x - \hat{x}\|\end{aligned}$$

□

Theorem B.7 (Lipschitzness and smoothness of constrained hyper-objective F)**.** *The hyper-objective F is L_F-lipschitz and C_F-smooth in x, Where $L_F := L_f(1+L_y)$ and $C_F := (C_f(1+L_y) + L_f C_y + L_y C_f + L_y^2 C_f)$.*

Proof. Let $x, \hat{x} \in X$. Using Lemma B.6, Lipschitzness of F is given by

$$\begin{aligned}\|\nabla F(x)\| &\leq \|\nabla_x f(x, y^*(x))\| + \|(\frac{d}{dx}y^*(x))\|\|\nabla_y f(x, y^*(x))\| \\ &= L_f(1 + L_y) \\ &:= L_F\end{aligned}$$

Smoothness of F is given by

$$\begin{aligned}
&\|\nabla F(x) - \nabla F(\hat{x})\| \\
&\leq \|\nabla_x f(x, y^*(x)) - \nabla_x f(\hat{x}, y^*(\hat{x}))\| + \|\frac{d}{dx} y^*(x)^\top \nabla_y f(x, y^*(x)) - \frac{d}{dx} y^*(\hat{x})^\top \nabla_y f(\hat{x}, y^*(\hat{x}))\| \\
&\leq (C_f(1 + L_y) + L_f C_y + L_y C_f + L_y^2 C_f)\|x - \hat{x}\| \\
&:= C_F \|x - \hat{x}\|
\end{aligned}$$

□

Corollary B.8 (Smoothness of the unconstrained hyper-objective)**.** *Consider Problem P1 with unconstrained follower strategy sets $Y_i = \mathbb{R}^{n_i}$ for all $i \in [N]$. Suppose the equilibrium mapping $y^*(x)$ is L_y-Lipschitz and its Jacobian $\frac{d}{dx} y^*(x)$ is C_y-Lipschitz. Then the hyper-objective*

$$F(x) = f(x, y^*(x))$$

is L_F-Lipschitz and C_F-smooth, where

$$L_F := L_f(1 + L_y)$$

and

$$C_F := C_f(1 + L_y) + L_f C_y + L_y C_f + L_y^2 C_f.$$

Proof. In the unconstrained setting, the hypergradient is

$$\nabla F(x) = \nabla_x f(x, y^*(x)) + \left(\frac{d}{dx} y^*(x)\right)^\top \nabla_y f(x, y^*(x)).$$

The proof of Theorem B.7 only uses the L_y-Lipschitzness of $y^*(x)$, the C_y-Lipschitzness of $\frac{d}{dx} y^*(x)$, and the Lipschitzness and smoothness of f. Hence, after removing all constraint and dual-variable terms, the same argument gives

$$\|\nabla F(x)\| \leq L_f(1 + L_y) = L_F,$$

and, for any $x, \hat{x} \in X$,

$$\|\nabla F(x) - \nabla F(\hat{x})\| \leq \left(C_f(1 + L_y) + L_f C_y + L_y C_f + L_y^2 C_f\right) \|x - \hat{x}\|.$$

Therefore, F is L_F-Lipschitz and C_F-smooth. □

C Symmetric Jacobian Stackelberg Games: Unconstrained and Equality-Constrained Cases

We first provide auxiliary lemmas for section 4 with linear equality constraints. We then provide complete proof for theorem 4.1 and C.6 specialized to the unconstrained case.

C.1 Auxiliary Lemmas in constrained setting

Lemma C.1. *The perturbed game $\mathcal{G}_\delta := (x, [N], \{Y_i\}_{i\in[N]}, \{g_i + \delta f\}_{i\in[N]})$. $\mathcal{G}_\delta$ is a $C_\delta := (C + \delta C_f)$-smooth , $\mu_\delta := (\mu - \delta C_f)$-strongly monotone and $S_\delta := (S + \delta S_f)$-Hessian smooth. If we assume $\delta \leq \frac{\mu}{2C_f}$, then we set $\mu_\delta = \frac{\mu}{2}$, $C_\delta = C + \frac{\mu}{2}$ and $S_\delta = S + \frac{S_f \mu}{2C_f}$, so that μ_δ, C_δ and S_δ are independent of δ.*

Proof. Let $x, x' \in X$. $y, y' \in Y$, To prove strong monotonicity,

$$\begin{aligned}
&\langle V(x, y') + \delta \nabla_y f(x, y') - (V(x, y) + \delta \nabla_y f(x, y)), y' - y\rangle \\
&\geq \mu\|y' - y\|^2 + \langle \delta \nabla_y f(x, y') - \delta \nabla_y f(x, y), y' - y\rangle \\
&\geq \mu\|y' - y\|^2 - \delta\|\nabla_y f(x, y') - \nabla_y f(x, y)\|\|y' - y\| \\
&\geq (\mu - \delta C_f)\|y' - y\|^2
\end{aligned}$$

To prove smoothness,

$$\begin{aligned}
&\|V(x',y') + \delta\nabla_y f(x',y') - (V(x,y) + \delta\nabla_y f(x,y))\| \\
&\leq C\|y' - y\| + \delta\|\nabla_y f(x,y') - \nabla_y f(x,y)\| \\
&\leq (C + \delta C_f)\|y' - y\|
\end{aligned}$$

To prove Hessian-smoothness,

$$\begin{aligned}
&\frac{\partial}{\partial(x,y)}V(x',y') + \delta\frac{\partial}{\partial(x,y)}\nabla_y f(x',y') - \frac{\partial}{\partial(x,y)}V(x,y) - \delta\frac{\partial}{\partial(x,y)}\nabla_y f(x,y)\| \\
&\leq S\|(x',y') - (x,y)\| + \delta\|\frac{\partial}{\partial(x,y)}\nabla_y f(x',y') - \frac{\partial}{\partial(x,y)}\nabla_y f(x,y)\| \\
&\leq (S + \delta S_f)\|(x',y') - (x,y)\|
\end{aligned}$$

If we assume $\delta \leq \frac{\mu}{2C_f}$, then $\mu_\delta \geq \frac{\mu}{2}$, $C_\delta \leq C + \frac{\mu}{2}$ and $S_\delta \leq S + \frac{S_f\mu}{2C_f}$. □

Lemma C.2. *Let y_δ^* be the Nash equilibrium of the perturbed game $\mathcal{G}_\delta$. If $\delta \in [0, \Delta]$ where $\Delta \leq \frac{\mu}{2C_f}$, then*

$$\|y_\delta^* - y^*\| \leq \frac{2L_f}{\mu}\delta = \mathcal{O}(\delta)$$

Proof. By optimality of y_δ^* and y^* in their respective games,

$$\begin{aligned}
&\langle V(x,y_\delta^*) + \delta\nabla_y f(x,y_\delta^*), y^* - y_\delta^*\rangle \geq 0 \\
&\langle V(x,y^*), y_\delta^* - y^*\rangle \geq 0
\end{aligned}$$

By summing up the inequalities,

$$\langle V(x,y^*) - V(x,y_\delta^*) - \delta\nabla_y f(x,y_\delta^*), y_\delta^* - y^*\rangle \geq 0$$

By srtong monotonicity of $V + \delta\nabla_y f$,

$$\langle V(x,y_\delta^*) + \delta\nabla_y f(x,y_\delta^*) - V(x,y^*) - \delta\nabla_y f(x,y^*),\ y_\delta^* - y^*\rangle \geq (\mu - \delta C_f)\|y_\delta^* - y^*\|^2$$

We can bound the left-hand side by

$$\begin{aligned}
&\langle V(x,y_\delta^*) + \delta\nabla_y f(x,y_\delta^*) - V(x,y^*) - \delta\nabla_y f(x,y^*),\ y_\delta^* - y^*\rangle \\
&= \langle V(x,y_\delta^*) + \delta\nabla_y f(x,y_\delta^*) - V(x,y^*) + V(x,y^*) - (V(x,y^*) + \delta\nabla_y f(x,y^*)),\ y_\delta^* - y^*\rangle \\
&\leq \langle V(x,y^*) - (V(x,y^*) + \delta\nabla_y f(x,y^*)),\ y_\delta^* - y^*\rangle \\
&\leq \delta\|\nabla_y f(x,y^*))\|\|y_\delta^* - y^*\| \\
&\leq \delta L_f\|y_\delta^* - y^*\|
\end{aligned}$$

Hence, we get the desired inequality,

$$\begin{aligned}
\|y_\delta^* - y^*\| &\leq \frac{\delta}{\mu - \delta C_f}L_f \\
&\leq \frac{2L_f}{\mu}\delta
\end{aligned}$$

□

Lemma C.3 (Implicit differentiation of perturbed game)**.**

$$\begin{bmatrix} \frac{d}{d\delta}y_\delta^*|_{\delta=0} \\ \frac{d}{d\delta}\lambda_\delta^*|_{\delta=0} \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial y}V(x,y^*) & B^\top \\ B & 0 \end{bmatrix}^{-1} \begin{bmatrix} -\nabla_y f(x,y^*) \\ 0 \end{bmatrix}$$

Proof. By Lemma C.1, $\mathcal{G}_\delta$ is a $(\mu - \delta C_f)$-strongly monotone, $(C + \delta C_f)$-smooth and $(S + \delta S_f$-Hessian smooth. Let y_δ^*, λ_δ^* be the Nash equilibrium and lagrange multiplier for $\mathcal{G}_\delta$. We can again apply implicit function theorem and show that y_δ^* and λ_δ^* are continuously differentiable functions of δ and the derivatives of them are given by

$$\begin{bmatrix} \frac{\partial}{\partial y} V(x, y_\delta^*) + \delta \nabla_{yy}^2 f(x, y_\delta^*) & B^\top \\ B & 0 \end{bmatrix} \begin{bmatrix} \frac{d}{d\delta} y_\delta^* \\ \frac{d}{d\delta} \lambda_\delta^* \end{bmatrix} = \begin{bmatrix} -\nabla_y f(x, y_\delta^*) \\ 0 \end{bmatrix} \tag{12}$$

we can prove that $y^* = \lim_{\delta \to 0} y_\delta^*$ (Lemma C.2). Take $\delta \to 0$, we have the desired result. □

Lemma C.4. *Assume that $\frac{\partial}{\partial y} V(x, y^*)$ is symmetric. Denote y^*, λ^* as constants with respect to x, then the hypergradient $\nabla F(x)$ can be rewritten without second-order derivatives:*

$$v_\delta(x, y_\delta^*, \lambda_\delta^*) = \frac{1}{\delta} \sum_{i=1}^{N} \nabla_x [g_i(x, y_{\delta,i}^*, y_{\neg i}^*) + (\lambda_{\delta,i}^*)^\top e^{(i)}(x, y_i^*)] - \nabla_x [g_i(x, y_i^*, y_{\neg i}^*) + (\lambda_i^*)^\top e^{(i)}(x, y_i^*)]$$

$$(\frac{d}{dx} y^*(x))^\top \nabla_y f(x, y^*(x)) = \lim_{\delta \to 0} v_\delta(x, y_\delta^*, \lambda_\delta^*)$$

Proof.

$$\begin{aligned}
&(\frac{d}{dx} y^*(x))^\top \nabla_y f(x, y^*(x)) \\
&= \begin{bmatrix} \frac{d}{dx} y^*(x)^\top & \frac{d}{dx} \lambda^*(x)^\top \end{bmatrix} \begin{bmatrix} \nabla_y f(x, y^*) \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial}{\partial x} V(x, y^*)^\top & A^\top \end{bmatrix} \begin{bmatrix} \frac{\partial}{\partial y} V(x, y^*) & B^\top \\ B & 0 \end{bmatrix}^{-1} \begin{bmatrix} -\nabla_y f(x, y^*) \\ 0 \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial}{\partial x} V(x, y^*)^\top & A^\top \end{bmatrix} \begin{bmatrix} \frac{d}{d\delta} y_\delta^* \\ \frac{d}{d\delta} \lambda_\delta^* \end{bmatrix}_{\delta=0} \\
&= \begin{bmatrix} \frac{\partial}{\partial x_1} \nabla_{y_1} g_1(x, y^*) & \cdots & \frac{\partial}{\partial x_p} \nabla_{y_1} g_1(x, y^*) \\ \vdots & \cdots & \vdots \\ \frac{\partial}{\partial x_1} \nabla_{y_N} g_N(x, y^*) & \cdots & \frac{\partial}{\partial x_p} \nabla_{y_N} g_N(x, y^*) \end{bmatrix}^\top \frac{d}{d\delta} y_\delta^*|_{\delta=0} + \sum_{i=1}^{N} (\frac{\partial}{\partial x} e^{(i)}(x, y_i^*))^\top \frac{d}{d\delta} \lambda_{\delta,i}^*|_{\delta=0} \\
&= \sum_{i=1}^{N} \frac{\partial}{\partial y_i} \nabla_x g_i(x, y^*) \frac{d}{d\delta} y_{\delta,i}^*|_{\delta=0} + \frac{\partial}{\partial x} e^{(i)}(x, y_i^*))^\top \frac{d}{d\delta} \lambda_{\delta,i}^*|_{\delta=0} \\
&= \sum_{i=1}^{N} \lim_{\delta \to 0} \frac{1}{\delta} [\nabla_x g_i(x, y_{\delta,i}^*, y_{\neg i}^*) - \nabla_x g_i(x, y_i^*, y_{\neg i}^*)] + \frac{\partial}{\partial x} e^{(i)}(x, y_i^*))^\top \lim_{\delta \to 0} \frac{1}{\delta} [\lambda_{\delta,i}^* - \lambda_i^*] \\
&= \lim_{\delta \to 0} \frac{1}{\delta} \sum_{i=1}^{N} \nabla_x [g_i(x, y_{\delta,i}^*, y_{\neg i}^*) + (\lambda_{\delta,i}^*)^\top e^{(i)}(x, y_i^*)] - \nabla_x [g_i(x, y_i^*, y_{\neg i}^*) + (\lambda_i^*)^\top e^{(i)}(x, y_i^*)]
\end{aligned}$$

□

Lemma C.5 (Lipshitzness and smoothness of perturbed lower-level solution mappings)**.** *Denote $\kappa_\delta = \frac{C_\delta}{\mu_\delta}$ and suppose $\delta \le \Delta := \frac{\mu}{2C_f}$, y_δ^* is $L_{y,\delta}$-Lipschitz and $C_{y,\delta}$-smooth in $\delta \in \mathbb{R}_{>0}$. λ_δ^* is $L_{\lambda,\delta}$-lipschitz and $C_{\lambda,\delta}$-smooth in δ. Moreover, $L_{y,\delta}$, $C_{y,\delta}$, $L_{\lambda,\delta}$, $C_{\lambda,\delta}$ are independent of the value of δ, and can be treated as constants.*

Proof. $(\frac{d}{d\delta} y_\delta^*, \frac{d}{d\delta} \lambda_\delta^*)$ are the solutions to equation 12. Let $0 < \delta, \hat{\delta} \le \Delta$. Using Lemma B.6, substitute $G = \frac{\partial}{\partial y} V(x, y_\delta^*) + \delta \nabla_{yy}^2 f(x, y_\delta^*)$, $G_{xy} = \nabla_y f(x, y_\delta^*)$, $A = 0$, $\hat{y} = y_{\hat{\delta}}^*$, $\hat{G} = \frac{\partial}{\partial y} V(x, y_{\hat{\delta}}^*) + \hat{\delta} \nabla_{yy}^2 f(x, y_{\hat{\delta}}^*)$, $\hat{G}_{xy} = \nabla_y f(x, y_{\hat{\delta}}^*)$ $\kappa_\delta := \frac{C_\delta}{\mu_\delta}$. Denote $\|P_B^G - P_B^{\hat{G}}\| \le L_{P,\delta} \|G - \hat{G}\|$ and $\|P_B^G\| \le M_{P,\delta}$ according to Lemmas B.4 and B.5.

The Lipschitzness of y_δ^* is given by

$$\begin{aligned}\|\frac{d}{d\delta}y_\delta^*\| &= \|(I-P_B^G)B^\dagger(-A)+P_B^G(-G^{-1}G_{xy})\| \\ &\le \|I-P_B^G\|\|B^\dagger A\|+\|P_B^G\|\|G^{-1}\|\|G_{xy}\| \\ &\le (1+M_{P,\delta})(0+\frac{L_f}{\mu_\delta}) \\ &:= L_{y,\delta}\end{aligned}$$

Denote $L_{G,\delta} := (SL_{y,\delta}+C_f+\Delta S_f L_{y,\delta})$, we have

$$\begin{aligned}\|G-\hat{G}\| &= \|\frac{\partial}{\partial y}V(x,y_\delta^*)+\delta\nabla_{yy}^2 f(x,y_\delta^*)-(\frac{\partial}{\partial y}V(x,y_{\hat{\delta}}^*)+\hat{\delta}\nabla_{yy}^2 f(x,y_{\hat{\delta}}^*))\| \\ &\le SL_{y,\delta}\|\delta-\hat{\delta}\|+\|\delta\nabla_{yy}^2 f(x,y_\delta^*)-\hat{\delta}\nabla_{yy}^2 f(x,y_\delta^*)\|+\|\hat{\delta}\nabla_{yy}^2 f(x,y_\delta^*)-\hat{\delta}\nabla_{yy}^2 f(x,y_{\hat{\delta}}^*)\| \\ &\le (SL_{y,\delta}+C_f+\Delta S_f L_{y,\delta})\|\delta-\hat{\delta}\| \\ &:= L_{G,\delta}\|\delta-\hat{\delta}\|\end{aligned}$$

Denote $L_{u,\delta} = C_f L_{y,\delta}+\frac{L_f}{\mu_\delta^2}L_{G,\delta}$, we have

$$\begin{aligned}\|G^{-1}G_{xy}-\hat{G}^{-1}\hat{G}_{xy}\| &\le \|G^{-1}G_{xy}-G^{-1}\hat{G}_{xy}\|+\|G^{-1}\hat{G}_{xy}-\hat{G}^{-1}\hat{G}_{xy}\| \\ &\le \frac{1}{\mu_\delta}\|G_{xy}-\hat{G}_{xy}\|+\|\hat{G}_{xy}\|\|G^{-1}(\hat{G}-G)\hat{G}^{-1}\| \\ &\le (C_f L_{y,\delta}+\frac{L_f}{\mu_\delta^2}L_{G,\delta})\|\delta-\hat{\delta}\| \\ &:= L_{u,\delta}\|\delta-\hat{\delta}\|\end{aligned}$$

The smoothness of y_δ^* is given by

$$\begin{aligned}&\|\frac{d}{dx}y_\delta^*-\frac{d}{dx}y_\delta^*\| \\ &\le L_{P,\delta}\|G-\hat{G}\|0+L_{P,\delta}\|G-\hat{G}\|\frac{L_f}{\mu_\delta}+M_{P,\delta}\|G^{-1}G_{xy}-\hat{G}^{-1}\hat{G}_{xy}\| \\ &\le (L_{P,\delta}+L_{P,\delta}L_{G,\delta}\frac{L_f}{\mu_\delta}+M_{P,\delta}L_{u,\delta})\|\delta-\hat{\delta}\| \\ &:= C_{y,\delta}\|\delta-\hat{\delta}\|\end{aligned}$$

The Lipschitzness of λ_δ^* is given by

$$\begin{aligned}\|\frac{d}{d\delta}\lambda_\delta^*\| &= \|(B^\dagger)^\top G(I-P_B^G)(B^\dagger A-G^{-1}G_{xy})\| \\ &\le \|B^\dagger\|\|C_\delta\|(1+M_{P,\delta})(0+\frac{L_f}{\mu_\delta}) \\ &:= L_{\lambda,\delta}\end{aligned}$$

The smoothness of λ_δ^* is given by

$$
\begin{aligned}
&\|\frac{d}{d\delta}\lambda_\delta^* - \frac{d}{d\delta}\lambda_{\hat{\delta}}^*\| \\
&= \|B^\dagger G(I - P_B^G)(B^\dagger A - G^{-1}G_{xy}) - B^\dagger \hat{G}(I - P_B^G)(B^\dagger A - G^{-1}G_{xy})\| \\
&+ \|B^\dagger \hat{G}(I - P_B^G)(B^\dagger A - G^{-1}G_{xy}) - B^\dagger \hat{G}(I - P_B^{\hat{G}})(B^\dagger A - G^{-1}G_{xy})\| \\
&+ \|B^\dagger \hat{G}(I - P_B^{\hat{G}})(B^\dagger A - G^{-1}G_{xy}) - B^\dagger \hat{G}(I - P_B^{\hat{G}})(B^\dagger A - \hat{G}^{-1}\hat{G}_{xy})\| \\
&\le \|B^\dagger\|(1 + M_{P,\delta})(0 + \frac{L_f}{\mu_\delta})\|G - \hat{G}\| + C_\delta\|B^\dagger\|(0 + \frac{L_f}{\mu_\delta})\|P_B^{\hat{G}} - P_B^G\| + C_\delta\|B^\dagger\|(1 + M_{P,\delta})\|G^{-1}G_{xy} - \hat{G}^{-1}\hat{G}_{xy}\| \\
&\le (\|B^\dagger\|(1 + M_{P,\delta})\frac{L_f}{\mu_\delta}L_{G,\delta} + C_\delta\|B^\dagger\|\frac{L_f}{\mu_\delta}L_{P,\delta} + C_\delta\|B^\dagger\|(1 + M_{P,\delta})L_{u,\delta})\|\delta - \hat{\delta}\| \\
&:= C_{\lambda,\delta}\|\delta - \hat{\delta}\|
\end{aligned}
$$

□

C.2 Omitted proof in Section 4 in unconstrained games

Theorem 4.1. *If the Jacobian $J = \frac{\partial}{\partial y}V(x, y^*)$ is symmetric, then:*

$$\|v_\delta(x, y_\delta^*) + (\frac{\partial}{\partial x}V(x, y^*))^\top J(x, y^*)^{-\top}\nabla_y f(x, y^*)\| \le \mathcal{O}(N\delta). \tag{4}$$

Proof. Let y_t^* denote the Nash equilibrium of the perturbed game

$$G_t := (x, [N], \{\mathbb{R}^{n_i}\}_{i\in[N]}, \{g_i + tf\}_{i\in[N]}), \qquad t \in [0, \delta].$$

The equilibrium condition is

$$V(x, y_t^*) + t\nabla_y f(x, y_t^*) = 0.$$

Differentiating with respect to t gives

$$\left(J(x, y_t^*) + t\nabla_{yy}^2 f(x, y_t^*)\right)\frac{d}{dt}y_t^* = -\nabla_y f(x, y_t^*).$$

Taking $t = 0$,

$$\frac{d}{dt}y_t^*\Big|_{t=0} = J(x, y^*)^{-1}\left(-\nabla_y f(x, y^*)\right).$$

By implicit differentiation of the original equilibrium condition $V(x, y^*) = 0$,

$$\left(\frac{d}{dx}y^*(x)\right)^\top \nabla_y f(x, y^*) = \frac{\partial}{\partial x}V(x, y^*)^\top J(x, y^*)^{-\top}\left(-\nabla_y f(x, y^*)\right).$$

Since $J(x, y^*)$ is symmetric,

$$J(x, y^*)^{-\top} = J(x, y^*)^{-1}.$$

Hence,

$$\left(\frac{d}{dx}y^*(x)\right)^\top \nabla_y f(x, y^*) = \frac{\partial}{\partial x}V(x, y^*)^\top \frac{d}{dt}y_t^*\Big|_{t=0}.$$

By equality of mixed partial derivatives,

$$\frac{\partial}{\partial x}V(x, y^*)^\top \frac{d}{dt}y_t^*\Big|_{t=0} = \sum_{i=1}^N \frac{\partial}{\partial y_i}\nabla_x g_i(x, y^*)\frac{d}{dt}y_{t,i}^*\Big|_{t=0}.$$

For $0 \le t \le \delta \le \mu/(2C_f)$, the perturbed operator $V + t\nabla_y f$ is at least $\mu/2$-strongly monotone. Under Assumptions 3.4 and 3.5, there exist constants $L_{y,\delta}$ and $C_{y,\delta}$, independent of δ, such that

$$\|y_t^* - y^*\| \le L_{y,\delta}t, \qquad \left\|\frac{d}{dt}y_t^*\right\| \le L_{y,\delta},$$

and

$$\left\| \frac{d}{dt} y_t^* - \frac{d}{dt} y_t^* \Big|_{t=0} \right\| \leq C_{y,\delta} t.$$

By definition of finite difference in (3),

$$v_\delta(x, y_\delta^*) = \frac{1}{\delta} \sum_{i=1}^{N} \left[\nabla_x g_i(x, y_{\delta,i}^*, y_{-i}^*) - \nabla_x g_i(x, y_i^*, y_{-i}^*) \right].$$

For all $i \in [N]$, by the fundamental theorem of calculus,

$$\nabla_x g_i(x, y_{\delta,i}^*, y_{-i}^*) - \nabla_x g_i(x, y_i^*, y_{-i}^*) = \int_0^\delta \frac{\partial}{\partial y_i} \nabla_x g_i(x, y_{t,i}^*, y_{-i}^*) \frac{d}{dt} y_{t,i}^* \, dt.$$

We bound each term by

$$\begin{aligned}
&\left\| \frac{\nabla_x g_i(x, y_{\delta,i}^*, y_{-i}^*) - \nabla_x g_i(x, y_i^*, y_{-i}^*)}{\delta} - \frac{\partial}{\partial y_i} \nabla_x g_i(x, y^*) \, \frac{d}{dt} y_{t,i}^* \Big|_{t=0} \right\| \\
&\leq \frac{1}{\delta} \int_0^\delta \left\| \frac{\partial}{\partial y_i} \nabla_x g_i(x, y_{t,i}^*, y_{-i}^*) \frac{d}{dt} y_{t,i}^* - \frac{\partial}{\partial y_i} \nabla_x g_i(x, y^*) \, \frac{d}{dt} y_{t,i}^* \Big|_{t=0} \right\| dt \\
&\leq \frac{1}{\delta} \int_0^\delta \left(S L_{y,\delta}^2 t + C C_{y,\delta} t \right) dt \\
&= \frac{1}{2} \left(S L_{y,\delta}^2 + C C_{y,\delta} \right) \delta.
\end{aligned}$$

Summing over $i = 1, \ldots, N$, we obtain

$$\begin{aligned}
&\left\| v_\delta(x, y_\delta^*) - \left(\frac{d}{dx} y^*(x) \right)^\top \nabla_y f(x, y^*(x)) \right\| \\
&\leq \frac{N}{2} \left(S L_{y,\delta}^2 + C C_{y,\delta} \right) \delta = \mathcal{O}(N\delta).
\end{aligned}$$

□

Theorem C.6. *Consider problem P1 and assume that the Jacobian $J = \frac{\partial}{\partial y} V(x, y^*)$ is symmetric. Under assumptions 3.4 and 3.5, Algorithm 1 specialized to unconstrained case outputs an $\mathcal{O}(\epsilon)$-stationary point of the hyperobjective F with $\tilde{\mathcal{O}}(N\epsilon^{-2})$ oracle calls to f and g_i.*

Proof. By Assumptions 3.4 and 3.5 and the implicit function theorem, the lower-level solution mapping $y^*(x)$ is Lipschitz and smooth. By corollary B.8, the hyperobjective F is C_F-smooth for some constant $C_F > 0$.

Suppose we have access to approximate lower-level solutions $\hat{y}, \hat{y}_\delta$ such that

$$\|y^* - \hat{y}\| \leq \delta^2, \qquad \|y_\delta^* - \hat{y}_\delta\| \leq \delta^2.$$

We construct the inexact hypergradient

$$\nabla \hat{F}(x) = \nabla_x f(x, \hat{y}) + \hat{v}_\delta(x, \hat{y}_\delta),$$

where

$$\hat{v}_\delta(x, \hat{y}_\delta) = \frac{1}{\delta} \sum_{i=1}^{N} \left[\nabla_x g_i(x, \hat{y}_{\delta,i}, \hat{y}_{-i}) - \nabla_x g_i(x, \hat{y}_i, \hat{y}_{-i}) \right].$$

We first bound the error due to the approximate lower-level solutions:

$$\begin{aligned}
&\|\hat{v}_\delta(x, \hat{y}_\delta) - v_\delta(x, y_\delta^*)\| \\
&\leq \frac{1}{\delta}\sum_{i=1}^{N} \|\nabla_x g_i(x, \hat{y}_{\delta,i}, \hat{y}_{-i}) - \nabla_x g_i(x, y_{\delta,i}^*, y_{-i}^*)\| \\
&\quad + \frac{1}{\delta}\sum_{i=1}^{N} \|\nabla_x g_i(x, y_i^*, y_{-i}^*) - \nabla_x g_i(x, \hat{y}_i, \hat{y}_{-i})\| \\
&\leq \frac{1}{\delta}\sum_{i=1}^{N} \left(2C\delta^2 + C\delta^2\right) \\
&= 3CN\delta = \mathcal{O}(N\delta).
\end{aligned}$$

Also,

$$\|\nabla_x f(x, \hat{y}) - \nabla_x f(x, y^*)\| \leq C_f \|\hat{y} - y^*\| \leq C_f \delta^2.$$

Combining the above inequalities with Theorem 4.2,

$$\begin{aligned}
\|\nabla \hat{F}(x) - \nabla F(x)\| &\leq \|\nabla_x f(x, \hat{y}) - \nabla_x f(x, y^*)\| \\
&\quad + \|\hat{v}_\delta(x, \hat{y}_\delta) - v_\delta(x, y_\delta^*)\| \\
&\quad + \left\| v_\delta(x, y_\delta^*) - \left(\frac{d}{dx} y^*(x)\right)^\top \nabla_y f(x, y^*) \right\| \\
&\leq C_f \delta^2 + \mathcal{O}(N\delta) \\
&= \mathcal{O}(N\delta).
\end{aligned}$$

Set

$$\delta = \frac{\epsilon}{N}.$$

Then

$$\|\nabla \hat{F}(x) - \nabla F(x)\| \leq \mathcal{O}(\epsilon).$$

Using $\nabla \hat{F}(x)$ as the inexact gradient in Lemma D.2, an $\mathcal{O}(\epsilon)$-stationary point of F can be obtained with

$$\mathcal{O}(\epsilon^{-2})$$

calls to $\nabla \hat{F}$.

It remains to bound the oracle complexity for computing each $\nabla \hat{F}$. Since the lower-level game is unconstrained, PGD reduces to gradient descent,

$$y^{k+1} = y^k - \eta V(x, y^k).$$

By Assumption 3.4 and Lemma E.1, gradient descent converges linearly to y^*. Moreover, for $\delta \leq \mu/(2C_f)$, the perturbed operator

$$V_\delta(x, y) = V(x, y) + \delta \nabla_y f(x, y)$$

is at least $\mu/2$-strongly monotone and $(C + \delta C_f)$-smooth, so gradient descent also converges linearly to y_δ^*.

For each call to $\nabla \hat{F}(x)$,

$$\mathcal{O}\left(\log \frac{1}{\delta^2}\right) = \mathcal{O}\left(\log(N^2\epsilon^{-2})\right)$$

lower-level iterations are sufficient to obtain δ^2-accurate solutions $\hat{y}$ and $\hat{y}_\delta$. Each lower-level iteration requires one first-order oracle call to each of the N follower cost functions. Therefore, the total oracle complexity is

$$\begin{aligned}
&\mathcal{O}\left(N\epsilon^{-2}\log(N^2\epsilon^{-2})\right) \\
&= \widetilde{\mathcal{O}}(N\epsilon^{-2}).
\end{aligned}$$

□

D INEXACTNESS IN LOWER-LEVEL SOLUTIONS AND HYPERGRADIENT

Lemma D.1 (Inexactness in lower-level solutions)**.** *Assume we have access to approximate lower-level solutions $\hat{y}$, $\hat{\lambda}$, $\hat{y}_\delta$, $\hat{\lambda}_\delta$ such that for some $\hat{\delta} > 0$, $\|y^* - \hat{y}\| \leq \hat{\delta}$, $\|\lambda^* - \hat{\lambda}\| \leq \hat{\delta}$, $\|y_\delta^* - \hat{y}_\delta\| \leq \hat{\delta}$, $\|{\lambda_\delta}^* - \hat{\lambda}_\delta\| \leq \hat{\delta}$. If $\hat{\delta} = \delta^2$, then the approximate hypergradient $\hat{\nabla} F(x) := \nabla_x f(x, y^*(x)) + \hat{v}_\delta(x, \hat{y}_\delta, \hat{\lambda}_\delta)$, has an error of*

$$\|\hat{\nabla} F(x) - \nabla F(x)\| \leq \mathcal{O}(N\delta)$$

Proof.

$$v_\delta(x, y_\delta^*, \lambda_\delta^*) = \frac{1}{\delta} \sum_{i=1}^{N} \nabla_x [g_i(x, y_{\delta,i}^*, y_{\neg i}^*) + (\lambda_{\delta,i}^*)^\top e^{(i)}(x, y_i^*)] - \nabla_x [g_i(x, y_i^*, y_{\neg i}^*) + (\lambda_i^*)^\top e^{(i)}(x, y_i^*)]$$

$$\begin{aligned}
&\|\hat{v}_\delta(x, \hat{y}_\delta, \hat{\lambda}_\delta) - v_\delta(x, y_\delta^*, \lambda_\delta^*)\| \\
&\leq \frac{1}{\delta} \sum_{i=1}^{N} \|\nabla_x g_i(x, \hat{y}_{\delta,i}, \hat{y}_{\neg i}) - \nabla_x g_i(x, y_{\delta,i}^*, y_{\neg i}^*) + \nabla_x g_i(x, y_i^*, y_{\neg i}^*) - \nabla_x g_i(x, \hat{y}_i, \hat{y}_{\neg i})\| \\
&+ \frac{1}{\delta} \sum_{i=1}^{N} \|\nabla_x [(\hat{\lambda}_{\delta,i})^\top e^{(i)}(x, \hat{y}_i) - (\lambda_{\delta,i}^*)^\top e^{(i)}(x, y_i^*) + (\lambda_i^*)^\top e^{(i)}(x, y_i^*) - (\hat{\lambda}_i)^\top e^{(i)}(x, \hat{y}_i)]\| \\
&\leq \frac{1}{\delta} \sum_{i=1}^{N} (3C\hat{\delta} + 2\|A^{(i)}\|\hat{\delta}) \\
&= \mathcal{O}(N\delta)
\end{aligned}$$

$$\begin{aligned}
&\|\hat{\nabla} F(x) - \nabla F(x)\| \\
&\leq \|\nabla_x f(x, \hat{y}) - \nabla_x f(x, y^*)\| + \|\hat{v}_\delta(x, \hat{y}_\delta, \hat{\lambda}_\delta) - v_\delta(x, y_\delta^*, \lambda_\delta^*)\| + \|v_\delta(x, y_\delta^*, \lambda_\delta^*) - (\frac{d}{dx} y^*(x))^\top \nabla_y f(x, y^*)\| \\
&\leq C_f \hat{\delta} + \mathcal{O}(N\delta) \\
&= \mathcal{O}(N\delta)
\end{aligned}$$

□

Lemma D.2 (Inexact gradient descent on smooth function)**.** *Suppose the inexactness in the gradient of C_F-smooth function F is given by $\|\hat{\nabla} F(x) - \nabla F(x)\| \leq \alpha = \mathcal{O}(\epsilon)$. If learning rate $\eta = \frac{1}{C_F}$, then gradient descent using the inexact gradient $\hat{\nabla} F(x)$ requires $\mathcal{O}(\epsilon^{-2})$ to reach $\mathcal{O}(\epsilon)$-stationarity.*

Proof. Suppose $(\forall x \in X)\,(\hat{\nabla} F(x) = \nabla F(x) + \tau(x))$ where $\|\tau(x)\| \leq \mathcal{O}(\alpha)$. Denote $x^+ = x - \eta\hat{\nabla} F(x)$,

$$
\begin{aligned}
&F(x^+) \\
&\leq F(x) + \nabla F(x)^\top (x^+ - x) + \frac{C_F}{2}\|x^+ - x\|^2 \\
&= F(x) - \eta \nabla F(x)^\top \hat{\nabla} F(x) + \frac{C_F}{2}\|\eta\hat{\nabla} F(x)\|^2 \\
&= F(x) - \eta\|\nabla F(x)\|^2 + \frac{\eta^2 C_F}{2}\|\nabla F(x) + \tau(x)\|^2 - \eta \nabla F(x)^\top \tau(x) \\
&= F(x) - \eta\|\nabla F(x)\|^2 + \frac{\eta^2 C_F}{2}\|\nabla F(x)\|^2 + (\eta^2 C_F - \eta)\nabla F(x)^\top \tau(x) + \frac{\eta^2 C_F}{2}\|\tau(x)\|^2 \\
&= F(x) - \frac{2\eta - \eta^2 C_F}{2}\|\nabla F(x)\|^2 + (\eta^2 C_F - \eta)\nabla F(x)^\top \tau(x) + \frac{\eta^2 C_F}{2}\|\tau(x)\|^2 \\
&= F(x) - \frac{\eta}{2}\|\nabla F(x)\|^2 + 0 + \frac{\eta}{2}\|\tau(x)\|^2 \\
&\leq F(x) - \frac{\eta}{2}\|\nabla F(x)\|^2 + \frac{\eta}{2}\alpha^2
\end{aligned}
$$

Denote each step of inexact gradient descent by $x_k = x_{k-1} + \eta\hat{\nabla} F(x_{k-1})$ for $k = 0, 1, 2, \cdots, T-1$. Applying a telescope sum, we have

$$
\begin{aligned}
F(x_T) - F(x_0) &\leq -\frac{\eta}{2}\sum_{k=0}^{T-1}\|\nabla F(x_k)\|^2 + \frac{\eta}{2}\sum_{k=0}^{T-1}\alpha^2 \\
\frac{1}{T}\sum_{k=0}^{T-1}\|\nabla F(x_k)\|^2 &\leq \frac{2(F(x_0) - F(x_T))}{\eta T} + \alpha^2 \\
\min_{k=0,1,\cdots,T-1}\|\nabla F(x_k)\|^2 &\leq \frac{2(F(x_0) - F(x^*))}{\eta T} + \alpha^2 \quad (x^* \text{ is the minimizer of } F)
\end{aligned}
$$

When $T = \mathcal{O}(\epsilon^{-2})$ and $\alpha = \mathcal{O}(\epsilon)$, we have $\min_{k=0,1,\cdots,T-1}\|\nabla F(x_k)\| \leq \mathcal{O}(\epsilon)$. □

E Solving Smooth and Strongly Monotone Game

Smooth and Strongly monotone game can be solved using projected gradient descent (PGD) (Facchinei & Pang (2003)). Let $\eta > 0$ be the learning rate and let $\{y_k\}$ be the iterates that converges to the Nash equilibrium, the update rule is given by $y_{k+1} = \Pi_Y(y_k - \eta V(y_k))$. Here, we omit the upper-level variable for convenience. Following Facchinei & Pang (2003), we can show a linear convergence guarantee. We extend the guarantee to the case when we only have access to approximate projection and approximate gradient operator V in Lemma E.2.

Finding a Nash equilibrium is equivalent to solving a variational inequality in smooth and strongly monotone game (Lemma A.1). We consider solving for y^* in the following variational inequality

$$\forall y \in Y : \langle V(y^*), y - y^* \rangle \geq 0$$

where Y is a closed convex set and V is a μ-strongly monotone and C-Lipschitz vector-valued function. We also define Euclidean projection as $\Pi_Y(z) = \arg\min_{y \in Y}\|y - z\|^2$. Lemma A.3 shows that y^* satisfies the variational inequality above if and only $y^* = \Pi_Y(y^* - \eta V(y^*))$ for some $\eta > 0$.

Lemma E.1. *Let* $\gamma = 1 - 2\mu\eta + \eta^2 C^2$ $\forall x, y \in Y$,

$$\|\Pi_Y(x - \eta V(x)) - \Pi_Y(y - \eta V(y))\|^2 \leq \gamma\|y - x\|^2$$

If $0 < \eta < \frac{2\mu}{C^2}$, *then* $0 < \gamma < 1$ *and thus* $\Pi_Y(y - \eta V(y))$ *is a contraction mapping in* y.

Proof.

$$\begin{aligned}
&\|\Pi_Y(x-\eta V(x)) - \Pi_Y(y-\eta V(y))\|^2 \\
&\leq \|x-\eta V(x) - y + \eta V(y)\|^2 \\
&= \|x-y\|^2 - 2\eta\langle y-x, V(y)-V(x)\rangle + \eta^2\|V(y)-V(x)\|^2 \\
&\leq \|x-y\|^2 - 2\mu\eta\|x-y\|^2 + \eta^2 C^2\|y-x\|^2 \\
&= (1-2\mu\eta+\eta^2C^2)\|y-x\|^2
\end{aligned}$$

Solving $(1-2\mu\eta+\eta^2C^2) < 1$, we get the $0 < \eta < \frac{2\mu}{C^2}$. □

Lemma E.2 (PGD with inexact projection and gradient)**.** *Suppose for any vector z, we only have access to an approximate projection $\hat{\Pi}_Y(z)$ such that $\|\hat{\Pi}_Y(z) - \Pi_Y(z)\| \leq \epsilon_0$ for some ϵ_0. Let $\hat{V}(z)$ be an approximate gradient operator such that $\|\hat{V}(z) - V(z)\| \leq \delta_0$ for some $\delta_0 > 0$. Let $\gamma = 1-2\mu\eta+\eta^2C^2$ and $0 < \eta < \frac{2\mu}{C^2}$. Define an arbitrary initial iterate $y_0 \in Y$ and apply the following update rule recursively:*

$$y_{t+1} = \hat{\Pi}_Y(y_t - \eta\hat{V}(y_t))$$

If $\epsilon_0 = \frac{(1-\sqrt{\gamma})\epsilon}{4}$ and $\delta_0 = \frac{(1-\sqrt{\gamma})\epsilon}{4\eta}$, then $\|y_t - y^*\| \leq \epsilon$ when $T \geq \mathcal{O}(\log(\frac{2\|y_0-y^*\|}{\epsilon}))$

Proof. For every vector z, $\hat{\Pi}_Y(z) = \Pi_Y(z) + \tau_0$ for some vector τ_0 such that $\|\tau_0\| \leq \epsilon_0$,

$$\begin{aligned}
&\|y_{t+1} - y^*\| \\
&= \|\hat{\Pi}_Y(y_t - \eta\hat{V}(y_t)) - \Pi_Y(y_t - \eta V(y_t)) + \Pi_Y(y_t - \eta V(y_t)) - \Pi_Y(y^* - \eta V(y^*))\| \\
&\leq \|\Pi_Y(y_t - \eta\hat{V}(y_t)) - \Pi_Y(y_t - \eta V(y_t))\| + \|\Pi_Y(y_t - \eta V(y_t)) - \Pi_Y(y^* - \eta V(y^*))\| + \epsilon_0 \\
&\leq \|-\eta\hat{V}(y_t) + \eta V(y_t)\| + \|\Pi_Y(y_t - \eta V(y_t)) - \Pi_Y(y^* - \eta V(y^*))\| + \epsilon_0 \\
&\leq \sqrt{\gamma}\|y_t - y^*\| + \eta\delta_0 + \epsilon_0
\end{aligned}$$

$$\begin{aligned}
&\|y_{t+1} - y^*\| \\
&\leq (\sqrt{\gamma})^{t+1}\|y_0 - y^*\| + \sum_{k=0}^{t}(\sqrt{\gamma})^k\epsilon_0 + \eta\sum_{k=0}^{t}(\sqrt{\gamma})^k\delta_0 \\
&= (\sqrt{\gamma})^{t+1}\|y_0 - y^*\| + \frac{1-\sqrt{\gamma}^{t+1}}{1-\sqrt{\gamma}}\epsilon_0 + \frac{1-\sqrt{\gamma}^{t+1}}{1-\sqrt{\gamma}}\eta\delta_0 \\
&\leq (\sqrt{\gamma})^{t+1}\|y_0 - y^*\| + \frac{1}{1-\sqrt{\gamma}}(\epsilon_0 + \eta\delta_0) \\
&= (\sqrt{\gamma})^{t+1}\|y_0 - y^*\| + \frac{\epsilon}{2}
\end{aligned}$$

Solving for $(\sqrt{\gamma})^{t+1}\|y_0 - y^*\| \leq \frac{\epsilon}{2}$, we have

$$\begin{aligned}
(\sqrt{\gamma})^{t+1}\|y_0 - y^*\| &\leq \frac{\epsilon}{2} \\
(t+1)\log(\frac{1}{\sqrt{\gamma}}) + \log(\frac{1}{\|y_0 - y^*\|}) &\geq \log(\frac{2}{\epsilon}) \\
t+1 &\geq \log(\frac{2\|y_0 - y^*\|}{\epsilon})/\log(\frac{1}{\sqrt{\gamma}})
\end{aligned}$$

□

We leverage Lemma E.3 to verify convergence of PGD in practice.

Lemma E.3 (Sufficient condition for convergence to y^*)**.** *Let $r(y) := y - \Pi_Y(y - V(y))$ be the natural residual. Suppose $\|r(\hat{y})\| \leq \beta$, then $\|\hat{y} - y^*\| \leq \frac{C+1}{\mu}\beta$.*

Proof. Let $u := \hat{y} - V(\hat{y})$, by first-order optimality of Euclidean projection,

$$\begin{aligned}\langle \Pi_Y(u) - u, z - \Pi_Y(u)\rangle \geq 0 \quad &, \forall z \in Y \\ \langle V(\hat{y}) - r(\hat{y}), y^* - \hat{y} + r(\hat{y})\rangle \geq 0 \quad &, \text{substitute } z = y^*\end{aligned}$$

Since y^* solves the variational inequality $\langle V(y^*), y - y^*\rangle \geq 0$, $\forall y \in Y$, we have

$$\langle V(y^*), \hat{y} - r(\hat{y}) - y^*\rangle \geq 0$$

Summing the inequalities toegther, we have

$$\langle V(\hat{y}) - V(y^*) - r(\hat{y}), y^* - \hat{y} + r(\hat{y})\rangle \geq 0$$

Rearranging terms and applying Lipschitzness of V,

$$\begin{aligned}\langle V(\hat{y}) - V(y^*), \hat{y} - y^*\rangle &\leq \langle V(\hat{y}) - V(y^*), r(\hat{y})\rangle + \langle r(\hat{y}), \hat{y} - y^*\rangle \\ &\leq \|V(\hat{y}) - V(y^*)\|\|r(\hat{y})\| + \|r(\hat{y})\|\|\hat{y} - y^*\| \\ &\leq \beta(C+1)\|\hat{y} - y^*\|\end{aligned}$$

By strong monotonicity and rearranging terms, we get the desired bound,

$$\begin{aligned}\mu\|\hat{y} - y^*\|^2 &\leq \langle V(\hat{y}) - V(y^*), \hat{y} - y^*\rangle \\ &\leq \beta(C+1)\|\hat{y} - y^*\|\end{aligned}$$

□

Lemma E.4 (Inexact dual solution oracle)**.** *Consider a μ-strongly monotone and C-smooth game $\mathcal{G} := (x, [N], \{Y_i\}_{i\in[N]}, \{g_i\}_{i\in[N]})$. Given an approximate lower-level primal solution $\hat{y}$ such that $\|\hat{y} - y^*\| \leq \hat{\delta}$, we can obtain an approximate lower-level dual solution $\hat{\lambda}$ such that $\|\hat{\lambda} - \lambda^*\| \leq \hat{\delta}$ using the same method in Kornowski et al. (2024) within $\mathcal{O}(1)$ oracle calls.*

Proof. The KKT system for the Nash equilibrium of $\mathcal{G}$ is given by

$$\begin{bmatrix} V(x, y^*) + B^\top \lambda^* \\ \begin{bmatrix} e^{(1)}(x, y_1^*) \\ \vdots \\ e^{(N)}(x, y_N^*) \end{bmatrix} \end{bmatrix} = 0$$

$$\begin{bmatrix} V(x, y^*) + B^\top \lambda^* \\ By^* \end{bmatrix} = \begin{bmatrix} 0 \\ -Ax - b \end{bmatrix}$$

Denote $c_x = -Ax - b$, (y^*, λ^*) are given by

$$\begin{bmatrix} I & B^\top \\ B & 0 \end{bmatrix} \begin{bmatrix} y^* \\ \lambda^* \end{bmatrix} = \begin{bmatrix} -V(x, y^*) + y^* \\ c_x \end{bmatrix}$$

and the approximate lower-level solution $(\hat{y}, \hat{\lambda})$ are given by

$$\begin{bmatrix} I & B^\top \\ B & 0 \end{bmatrix} \begin{bmatrix} \hat{y} \\ \hat{\lambda} \end{bmatrix} = \begin{bmatrix} B^\top \hat{\lambda} + \hat{y} \\ c_x \end{bmatrix}$$

Given $\hat{y}$, we can obtain $\hat{\lambda}$ by solving

$$B^\top \hat{\lambda} = \underset{s \in \text{span}(B^\top)}{\arg\min} \ \| - V(x, \hat{y}) - s\|^2$$

From the KKT system, observe that

$$-V(x, y^*) = B^\top \lambda^* = \underset{s \in \text{span}(B^\top)}{\arg\min} \ \| - V(x, y^*) - s\|^2$$

$$\begin{aligned}\|B^\top\hat{\lambda} - (-V(x,y^*))\| &= \|B^\top\hat{\lambda} - B^\top\lambda^*\| \\ &\le \|V(x,y^*) - V(x,\hat{y})\| \\ &\le C\|\hat{y} - y^*\| \\ &\le C\hat{\delta}\end{aligned}$$

For some τ such that $\|\tau\| \le (1+C)\hat{\delta}$,

$$\begin{aligned}\| \begin{bmatrix}\hat{y}\\ \hat{\lambda}\end{bmatrix} - \begin{bmatrix}y^*\\ \lambda^*\end{bmatrix} \| &= \| \begin{bmatrix}I & B^\top\\ B & 0\end{bmatrix}^{-1} \| \| \begin{bmatrix}\tau\\ 0\end{bmatrix} \| \\ &\le \| \begin{bmatrix}I & B^\top\\ B & 0\end{bmatrix}^{-1} \| (1+C)\hat{\delta}\end{aligned}$$

□

F UNCONSTRAINED STACKELBERG GAME

The gradient of the potential function is

$$\begin{aligned}\nabla_{y_i}\phi_\epsilon(x,y) &= \frac{1}{\epsilon}[\nabla_{y_i}g_i(x, y_i + \epsilon(y_i - y_i^*), y_{\neg i})(1+\epsilon) - \nabla_{y_i}g_i(x,y)] \\ &+ \frac{1}{\epsilon}\sum_{j\neq i}[\nabla_{y_i}g_j(x, y_j + \epsilon(y_j - y_j^*), y_{\neg j}) - \nabla_{y_i}g_j(x,y)] \\ &= \frac{1}{\epsilon}\sum_{j=1}^{N}[\nabla_{y_i}g_j(x, y_j + \epsilon(y_j - y_j^*), y_{\neg j}) - \nabla_{y_i}g_j(x,y)] + g_i(x, y_i + \epsilon(y_i - y_i^*), y_{\neg i})\end{aligned}$$

Lemma F.1 (Local strong Monotonicity and smoothness of the perturbed game $\bar{\mathcal{G}}_\delta$)**.** *Consider a closed ball around y^*, $\bar{B}_R(y^*) = \{y \in \mathbb{R}^n : \|y - y^*\| \le R\}$. For $0 < \delta < \frac{\mu}{2C_f}$ and $0 < R < \frac{\mu - \delta C_f}{S(N+1)(2+\sqrt{N})}$ and $0 < \epsilon < 1$, the perturbed game $\bar{\mathcal{G}}_\delta$ is locally $\bar{\mu}$-strongly monotone:*

$$\langle \bar{V}(x,y) - \bar{V}(x,z), y - z\rangle \ge \bar{\mu}\|y - z\|^2,\ \forall y, z \in \bar{B}_R(y^*)$$

For any $\epsilon > 0$, $\delta > 0$ and $R > 0$, $\bar{\mathcal{G}}_\delta$ is locally $\bar{C}$-smooth:

$$\|\bar{V}(x,y) - \bar{V}(x,z)\| \le \bar{C}\|y - z\|,\ \forall y, z \in \bar{B}_R(y^*)$$

The local strong monotonicity modulus is $\bar{\mu} := \mu - \frac{SRN}{2}\epsilon - (SRN)(1 + \frac{\epsilon}{2} + \sqrt{N}) + \epsilon\mu - SR\epsilon(1 + \frac{\epsilon}{2} + \sqrt{N}) - \delta C_f > 0$. The local smoothness modulus is $\bar{C} := \sqrt{N}\{[C\sqrt{N} + SR(\sqrt{N}(1 + \frac{\epsilon}{2}) + N) + \delta C_f] + \epsilon[C + SR(2 + \frac{\epsilon}{2})]\}$

Proof. By the Fundamental Theorem of Calculus,

$$\nabla_{y_i}g_j(x, y_j + \epsilon(y_j - y_j^*), y_{\neg j}) - \nabla_{y_i}g_j(x,y) = \epsilon\int_0^1 \frac{\partial}{\partial y_j}\nabla_{y_i}g_j(x, y_j + t\epsilon(y_j - y_j^*), y_{\neg j})(y_j - y_j^*)dt$$

We can express the gradient of the perturbed cost function as

$$\begin{aligned}\nabla_{y_i}\bar{g}_i(x,y) &= \int_0^1 \sum_{j=1}^{N}\frac{\partial}{\partial y_j}\nabla_{y_i}g_j(x, y_j + t\epsilon(y_j - y_j^*), y_{\neg j})(y_j - y_j^*)dt \\ &+ \epsilon\int_0^1 \frac{\partial}{\partial y_i}\nabla_{y_i}g_i(x, y_i + t\epsilon(y_i - y_i^*), y_{\neg i})(y_i - y_i^*)dt + \delta\nabla_{y_i}f(x,y) \\ &= \int_0^1 \sum_{j=1}^{N} J_{ji}(x, y_j + t\epsilon(y_j - y_j^*), y_{\neg j})^\top(y_j - y_j^*)dt \\ &+ \epsilon\int_0^1 J_{ii}(x, y_i + t\epsilon(y_i - y_i^*), y_{\neg i})(y_i - y_i^*)dt + \delta\nabla_{y_i}f(x,y)\end{aligned}$$

We can further expand it according to the identity $A(y)(y-y^*)-A(z)(z-y^*)=A(y)(y-z)+[A(y)-A(z)](z-y^*)$ for arbitrary matrices $A(y)$ and $A(z)$.

$$
\begin{aligned}
&\langle \nabla_{y_i}\bar{g}_i(x,y)-\nabla_{y_i}\bar{g}_i(x,z), y_i-z_i\rangle\\
&=\underbrace{\int_0^1\sum_{j=1}^N\langle\, y_i-z_i,\, J_{ji}(x,y_j+t\epsilon(y_j-y_j^*),y_{\neg j})^\top(y_j-z_j)\,\rangle dt}_{\text{(a)}}\\
&+\int_0^1\sum_{j=1}^N\langle\, y_i-z_i,\, [J_{ji}(x,y_j+t\epsilon(y_j-y_j^*),y_{\neg j})^\top-J_{ji}(x,z_j+t\epsilon(z_j-y_j^*),z_{\neg j})^\top](z_j-y_j^*)\,\rangle dt\\
&+\epsilon\int_0^1\langle\, y_i-z_i,\, J_{ii}(x,y_i+t\epsilon(y_i-y_i^*),y_{\neg i})(y_i-z_i)\,\rangle dt\\
&+\epsilon\int_0^1\langle\, y_i-z_i,\, [J_{ii}(x,y_i+t\epsilon(y_i-y_i^*),y_{\neg i})-J_{ii}(x,z_i+t\epsilon(z_i-y_i^*),z_{\neg i})](z_i-y_i^*)\,\rangle dt\\
&+\delta\langle\, y_i-z_i,\nabla_{y_i}f(x,y)-\nabla_{y_i}f(x,z)\,\rangle
\end{aligned}
$$

We can further decompose term (a) according to the identity

$$J_{ji}(x,y_j+t\epsilon(y_j-y_j^*),y_{\neg j})^\top=J_{ji}(x,y)^\top+(J_{ji}(x,y_j+t\epsilon(y_j-y_j^*),y_{\neg j})^\top-J_{ji}(x,y)^\top)$$

For all $y,z\in B_R(y^*)$

$$
\begin{aligned}
&\langle\,\bar{V}(x,y)-\bar{V}(x,z),\,y-z\,\rangle\\
&=\sum_{i=1}^N\langle\,\nabla_{y_i}\bar{g}_i(x,y)-\nabla_{y_i}\bar{g}_i(x,z),\,y_i-z_i\,\rangle\\
&\geq\mu\|y-z\|^2-\underbrace{\int_0^1\sum_{i=1}^N\sum_{j=1}^N\|y_i-z_i\|\|y_j-z_j\|S\|t\epsilon(y_j-y_j^*)\|dt}_{\text{(a)}}\\
&\underbrace{-\int_0^1\sum_{i=1}^N\sum_{j=1}^N\|y_i-z_i\|S((1+t\epsilon)\|y_j-z_j\|+\|y_{\neg j}-z_{\neg j}\|)\|z_j-y_j^*\|dt}_{\text{(b)}}\\
&+\epsilon\mu\|y-z\|^2-\underbrace{\epsilon\int_0^1\sum_{i=1}^N\|y_i-z_i\|S((1+t\epsilon)\|y_i-z_i\|+\|y_{\neg i}-z_{\neg i}\|)\|z_i-y_i^*\|dt}_{\text{(c)}}\\
&-\delta C_f\|y-z\|^2
\end{aligned}
$$

$$
\begin{aligned}
\text{(a)}&=S\epsilon\int_0^1\sum_{i=1}^N t\|y_i-z_i\|(\sum_j\|y_j-z_j\|\|y_j-y_j^*\|)\\
&\leq S\sqrt{N}\epsilon\int_0^1 t\|y-z\|(\sum_j\|y_j-z_j\|\|y_j-y_j^*\|)\\
&\leq S\sqrt{N}\epsilon\int_0^1 t\|y-z\|(\sum_j\|y_j-z_j\|\|y-y^*\|)\\
&\leq\frac{SN}{2}\epsilon\|y-z\|^2\|y-y^*\|\\
&\leq\frac{SRN}{2}\epsilon\|y-z\|^2
\end{aligned}
$$

where the first and third inequalities are obtained using the identity $\frac{1}{\sqrt{N}}\|y\| = \sqrt{\frac{1}{N}\sum_i \|y_i\|^2} \geq \sum_i \frac{1}{N}\|y_i\|$ derived from Jensen's inequality. Using similar techniques, we can bound

$$\begin{aligned}\text{(b)} &\leq SN\|y-z\|^2\|z-y^*\| + \frac{SN}{2}\epsilon\|y-z\|^2\|z-y^*\| + SN\sqrt{N}\|y-z\|^2\|z-y^*\| \\ &\leq (SRN)(1+\frac{\epsilon}{2}+\sqrt{N})\|y-z\|^2\end{aligned}$$

$$\begin{aligned}\text{(c)} &\leq S\epsilon((1+\frac{\epsilon}{2})\|y-z\|^2\|z-y^*\| + \sqrt{N}\|y-z\|^2\|z-y^*\|) \\ &\leq SR\epsilon(1+\frac{\epsilon}{2}+\sqrt{N})\|y-z\|^2\end{aligned}$$

Putting these together, we have local strong monotonicity with modulus $\bar{\mu}$ for sufficiently small ϵ, δ and R

$$\begin{aligned}&\langle\, \bar{V}(x,y) - \bar{V}(x,z),\, y-z \,\rangle \\ &\geq [\mu - \frac{SRN}{2}\epsilon - (SRN)(1+\frac{\epsilon}{2}+\sqrt{N}) + \epsilon\mu - SR\epsilon(1+\frac{\epsilon}{2}+\sqrt{N}) - \delta C_f]\|y-z\|^2\end{aligned}$$

Rearranging terms, we get $\bar{\mu} > 0$ when

$$R < \frac{(1+\epsilon)\mu - \delta C_f}{SN(1+\epsilon+\sqrt{N}) + S\epsilon(1+\frac{\epsilon}{2}+\sqrt{N})}$$

$$\begin{aligned}\frac{(1+\epsilon)\mu - \delta C_f}{SN(1+\epsilon+\sqrt{N}) + S\epsilon(1+\frac{\epsilon}{2}+\sqrt{N})} &> \frac{\mu - \delta C_f}{SN(1+\epsilon+\sqrt{N}) + S\epsilon(1+\epsilon+\sqrt{N})} \\ &> \frac{\mu - \delta C_f}{S(N+1)(2+\sqrt{N})} \text{ (assume } 0<\epsilon<1)\end{aligned}$$

$0 < \delta < \frac{\mu}{2C_f}$ and $0 < R < \frac{\mu - \delta C_f}{S(N+1)(2+\sqrt{N})}$ is a sufficient condition for $\bar{\mu} > 0$.

To derive the smoothness bound,

$$\begin{aligned}\|\nabla_{y_i}\bar{g}_i(x,y) - \nabla_{y_i}\bar{g}(x,z)\| = \|&\int_0^1 \sum_{j=1}^N J_{ji}(x, y_j + t\epsilon(y_j - y_j^*), y_{\neg j})^\top (y_j - z_j)dt \\ &+ \int_0^1 \sum_{j=1}^N [J_{ji}(x, y_j + t\epsilon(y_j - y_j^*), y_{\neg j})^\top - J_{ji}(x, z_j + t\epsilon(z_j - y_j^*), z_{\neg j})^\top](z_j - y_j^*)dt \\ &+ \epsilon\int_0^1 J_{ii}(x, y_i + t\epsilon(y_i - y_i^*), y_{\neg i})(y_i - z_i)dt \\ &+ \epsilon\int_0^1 [J_{ii}(x, y_i + t\epsilon(y_i - y_i^*), y_{\neg i}) - J_{ii}(x, z_i + t\epsilon(z_i - y_i^*), z_{\neg i})](z_i - y_i^*)dt \\ &+ \delta\nabla_{y_i}f(x,y) - \nabla_{y_i}f(x,z)\| \\ &\leq C\sqrt{N}\|y-z\| + S\|z-y^*\|[\sqrt{N}(1+\frac{\epsilon}{2})\|y-z\| + N\|y-z\|] \\ &+ \epsilon C\|y_i - z_i\| + \epsilon S\|z-y^*\|[(1+\frac{\epsilon}{2})\|y_i - z_i\| + \|y-z\|] + \delta C_f\|y-z\| \\ &\leq [C\sqrt{N} + SR(\sqrt{N}(1+\frac{\epsilon}{2}) + N) + \delta C_f]\|y-z\| + \epsilon[C + SR(2+\frac{\epsilon}{2})]\|y-z\|\end{aligned}$$

Hence,

$$\begin{aligned}&\|\bar{V}(x,y) - \bar{V}(x,z)\| \\ &\leq \sqrt{N}\{[C\sqrt{N} + SR(\sqrt{N}(1+\frac{\epsilon}{2}) + N) + \delta C_f] + \epsilon[C + SR(2+\frac{\epsilon}{2})]\}\|y-z\|\end{aligned}$$

□

We need to guarantee the existence of Nash equilibrium $\bar{y}_\delta^*$ in the perturbed game $\bar{\mathcal{G}}_\delta$, and guarantee that $\|\bar{y}_\delta^* - y^*\|$ shrinks with decreasing $\delta > 0$.

Lemma F.2 (Existence of Nash equilibrium in the perturbed game)**.** *Let R be the radius of the closed ball $\bar{B}_R(y^*)$ on which $\bar{V}$ is $\bar{\mu}$-strongly-monotone and $\bar{C}$-Lipschitz. For $0 < \delta \leq \min\{\frac{\bar{\mu}R}{2L_f}, \frac{\mu}{2C_f}\}$, the perturbed game $\bar{\mathcal{G}}_\delta$ has a locally unique Nash equilibrium $\bar{y}_\delta^*$ in $\bar{B}_R(y^*)$, and*

$$\|\bar{y}_\delta^* - y^*\| \leq \frac{2L_f\delta}{\bar{\mu}}$$

Proof. Let $\gamma = 1 - 2\bar{\mu}\eta + \eta^2\bar{C}^2$. Denote $T(y) = y - \eta\bar{V}(x, y)$. From Lemm E.1, we have $\forall y, z \in \bar{B}_R(y^*)$,

$$\|T(y) - T(z)\|^2 \leq \gamma\|y - z\|^2$$

If $0 < \eta < \frac{2\bar{\mu}}{\bar{C}^2}$, then $0 < \gamma < 1$ and thus $T(y)$ is a contraction mapping in y. We want to prove that T maps from $\bar{B}_R(y^*)$ to $\bar{B}_R(y^*)$. Since T is a contraction, there exists a unique fixed point $\bar{y}_\delta^* = T(\bar{y}_\delta^*)$ by Banach fixed-point theorem. The fixed point is the Nash equilibrium by Lemma A.3.

Let $y \in \bar{B}_R(y^*)$,

$$\begin{aligned}\|T(y) - y^*\| &\leq \|T(y) - T(y^*)\| + \|T(y^*) - y^*\| \\ &\leq \sqrt{\gamma}\|y - y^*\| + \eta\|\bar{V}(y^*)\| \\ &\leq \sqrt{\gamma}R + \eta L_f\delta\end{aligned}$$

Set $\eta = \frac{\bar{\mu}}{\bar{C}^2}$, we have $1 - \sqrt{\gamma} = 1 - \sqrt{1 - \frac{\bar{\mu}^2}{\bar{C}^2}}$. We need

$$\delta \leq \frac{(1 - \sqrt{\gamma})R}{\eta L_f} = \frac{(1 - \sqrt{1 - \frac{\bar{\mu}^2}{\bar{C}^2}})R}{\frac{\bar{\mu}}{\bar{C}^2}L_f}$$

For any concave function $h : \mathbb{R} \to \mathbb{R}, \forall a, b \in \mathbb{R}, h(b) \leq h(a) + \nabla h(a)^\top(b - a)$. Set $h(b) = \sqrt{1 - b}$, $a = 0$, we have the identity $\sqrt{1 - b} \leq 1 - \frac{b}{2}$. Then

$$\frac{(1 - \sqrt{1 - \frac{\bar{\mu}^2}{\bar{C}^2}})R}{\frac{\bar{\mu}}{\bar{C}^2}L_f} \geq \frac{(1 - (1 - \frac{\bar{\mu}^2}{2\bar{C}^2}))R}{\frac{\bar{\mu}}{\bar{C}^2}L_f} = \frac{\bar{\mu}R}{2L_f}$$

A sufficient condition for the fixed point $\bar{y}_\delta^*$ to exist is $\delta \leq \frac{\bar{\mu}R}{2L_f}$. We can bound the distance to y^* in terms of δ,

$$\begin{aligned}\|\bar{y}_\delta^* - y^*\| &\leq \sqrt{\gamma}\|\bar{y}_\delta^* - y^*\| + \eta L_f\delta \\ \|\bar{y}_\delta^* - y^*\| &\leq \frac{\eta L_f\delta}{1 - \sqrt{\gamma}} \\ \|\bar{y}_\delta^* - y^*\| &\leq \frac{2L_f\delta}{\bar{\mu}}\end{aligned}$$

□

Lemma F.3.

$$\|\frac{\bar{y}_\delta^* - y^*}{\delta} - (J(x, y^*)^\top + \epsilon D(x, y^*))^{-1}(-\nabla_y f(x, y^*))\| \leq \frac{\sqrt{N}}{\bar{\mu}^2}\mathcal{O}(\delta)$$

Proof. Let $y = \bar{y}^*_\delta$, for all $i \in [N]$, we have

$$
\begin{aligned}
0 &= \nabla_{y_i} \bar{g}_i(x, y) \\
&= \int_0^1 \sum_{j=1}^N J_{ji}(x, y_j + t\epsilon(y_j - y_j^*), y_{\neg j})^\top (y_j - y_j^*) dt \\
&+ \epsilon \int_0^1 J_{ii}(x, y_i + t\epsilon(y_i - y_i^*), y_{\neg i})(y_i - y_i^*) dt + \delta \nabla_{y_i} f(x, y) \\
&= \underbrace{\int_0^1 \sum_{j=1}^N [J_{ji}(x, y_j + t\epsilon(y_j - y_j^*), y_{\neg j})^\top - J_{ji}(x, y_j^*, y_{\neg j})^\top](y_j - y_j^*) dt}_{\text{(a)}} \\
&+ \underbrace{\epsilon \int_0^1 [J_{ii}(x, y_i + t\epsilon(y_i - y_i^*), y_{\neg i}) - J_{ii}(x, y_i^*, y_{\neg i})](y_i - y_i^*) dt}_{\text{(b)}} \\
&+ \underbrace{\delta(\nabla_{y_i} f(x, y) - \nabla_{y_i} f(x, y^*))}_{\text{(c)}} \\
&+ \underbrace{\sum_{j=1}^N [J_{ji}(x, y_j^*, y_{\neg j})^\top - J_{ji}(x, y^*)^\top](y_j - y_j^*)}_{\text{(d)}} \\
&+ \underbrace{[J_{ii}(x, y_i^*, y_{\neg i})^\top - J_{ii}(x, y^*)^\top](y_i - y_i^*)}_{\text{(e)}} \\
&+ \sum_{j=1}^N J_{ji}(x, y^*)^\top (y_j - y_j^*) + \epsilon J_{ii}(x, y^*)(y_i - y_i^*) + \delta \nabla_{y_i} f(x, y^*)
\end{aligned}
$$

$$
\begin{aligned}
\|a\| &\le S(1 + \frac{\epsilon}{2})\|y - y^*\|^2 \\
\|b\| &\le S(1 + \frac{\epsilon}{2})\|y - y^*\|^2 \\
\|c\| &\le \delta L_f \|y - y^*\| \\
\|d\| &\le S\|y - y^*\|^2 \\
\|e\| &\le \epsilon S\|y - y^*\|^2
\end{aligned}
$$

Combining the equations for $i = 1, 2, \cdots, N$,

$$
J(x, y^*)^\top (y - y^*) + \epsilon D(x, y^*)(y - y^*) + \delta \nabla_y f(x, y^*) + \text{extra} = 0
$$

where $\|\text{extra}\| \le \sqrt{N} S(1 + \frac{\epsilon}{2})\|y - y^*\|^2 + S(1 + \frac{\epsilon}{2})\|y - y^*\|^2 + \delta L_f \|y - y^*\| + S\|y - y^*\|^2 + \epsilon S\|y - y^*\|^2$. Solving the equation, we have

$$
\frac{\bar{y}^*_\delta - y^*}{\delta} = (J(x, y^*)^\top + \epsilon D(x, y^*))^{-1}(-\nabla_y f(x, y^*) - \frac{\text{extra}}{\delta})
$$

Using Lemma F.2, we have

$$
\begin{aligned}
\|\frac{\bar{y}^*_\delta - y^*}{\delta} - (J(x, y^*)^\top + \epsilon D(x, y^*))^{-1}(-\nabla_y f(x, y^*))\| &\le \|(J(x, y^*)^\top + \epsilon D(x, y^*))^{-1}\| \|\frac{\text{extra}}{\delta}\| \\
&\le \frac{1}{\mu(1 + \epsilon)}[\mathcal{O}(\frac{\sqrt{N}\delta}{\bar{\mu}^2}) + \mathcal{O}(\frac{\sqrt{N}\delta\epsilon}{\bar{\mu}^2}))] \\
&\le \mathcal{O}(\frac{\sqrt{N}\delta}{\bar{\mu}^2}) \quad \text{(we ignore } \mathcal{O}(\delta\epsilon))
\end{aligned}
$$

□

Lemma F.4.

$$\|(J(x,y^*)^\top + \epsilon D(x,y^*))^{-1}(-\nabla_y f(x,y^*)) - J(x,y^*)^{-\top}(-\nabla_y f(x,y^*))\| \le \mathcal{O}(\epsilon)$$

Proof.

$$\begin{aligned}
&\|(J(x,y^*)^\top + \epsilon D(x,y^*))^{-1}(-\nabla_y f(x,y^*)) - J(x,y^*)^{-\top}(-\nabla_y f(x,y^*))\| \\
&= \|(J(x,y^*)^\top + \epsilon D(x,y^*))^{-1}[J(x,y^*)^\top - (J(x,y^*)^\top + \epsilon D(x,y^*))]J(x,y^*)^{-\top}(-\nabla_y f(x,y^*))\| \\
&\le \frac{\epsilon}{\mu^2(1+\epsilon)} L_f \\
&= \mathcal{O}(\epsilon)
\end{aligned}$$

The first equality uses the matrix identity $A^{-1} - B^{-1} = A^{-1}(B-A)B^{-1}$ for any invertible matrices A and B of the same dimensions. □

F.1 Proof of Theorem 5.2

Proof.

$$\begin{aligned}
v_\delta(x, \bar{y}_\delta^*) &= \frac{1}{\delta}\sum_{i=1}^N [\nabla_x g_i(x, \bar{y}_{\delta,i}^*, y_{\neg i}^*) - \nabla_x g_i(x, y^*)] \\
&= \frac{1}{\delta}\sum_{i=1}^N [\frac{\partial}{\partial y_i}\nabla_x g_i(x,y^*)(\bar{y}_{\delta,i}^* - y_i^*) + \text{H.O.T.}] \\
&= \frac{\partial}{\partial x} V(x,y^*)^\top \frac{\bar{y}_\delta^* - y^*}{\delta} + N\frac{\text{H.O.T.}}{\delta}]
\end{aligned}$$

where $\|\text{H.O.T.}\| = \mathcal{O}(\|\bar{y}_\delta^* - y^*\|^2) \le \mathcal{O}(\frac{\delta^2}{\bar{\mu}^2})$ by Lemma F.2. Using Lemma F.3 and F.4, we have

$$\begin{aligned}
&\|v_\delta(x, \bar{y}_\delta^*) - \frac{\partial}{\partial x} V(x,y^*)^\top \frac{\partial}{\partial y} V(x,y^*)^{-\top}(-\nabla_y f(x,y^*))\| \\
&\le \|v_\delta(x, \bar{y}_\delta^*) - \frac{\partial}{\partial x} V(x,y^*)^\top (J(x,y^*)^\top + \epsilon D(x,y^*))^{-1}(-\nabla_y f(x,y^*))\| \\
&+ \|\frac{\partial}{\partial x} V(x,y^*)^\top (J(x,y^*)^\top + \epsilon D(x,y^*))^{-1}(-\nabla_y f(x,y^*)) - \frac{\partial}{\partial x} V(x,y^*)^\top \frac{\partial}{\partial y} V(x,y^*)^{-\top}(-\nabla_y f(x,y^*))\| \\
&\le C(\mathcal{O}(\frac{\sqrt{N}\delta}{\bar{\mu}^2}) + \mathcal{O}(\frac{N\delta}{\bar{\mu}^2}) + \mathcal{O}(\epsilon)) \\
&= \mathcal{O}(\frac{N\delta}{\bar{\mu}^2} + \epsilon)
\end{aligned}$$

□

G Stackelberg Game with Linear Equality Constraints

By construction, $\bar{\mathcal{G}}_\delta$ has the same local strong monotonicity modulus $\bar{\mu}$ and local smoothness modulus $\bar{C}$ on the ball $\bar{B}_R(y^*)$ as the unconstrained case. We are making the following assumption.

Assumption G.1 (Constraint qualification (LICQ))**.** Assume the rows of the Jacobian of equality constraints and active inequality constraints are linearly independent for each player.

Lemma G.2 (Existence of Nash equilibrium in the perturbed game)**.** *Let R be the radius of the closed ball $\bar{B}_R(y^*)$ on which $\bar{V}$ is $\bar{\mu}$-strongly-monotone and $\bar{C}$-Lipschitz. For $0 < \delta \le \frac{\bar{\mu}R}{2L_f}$, the perturbed game $\bar{\mathcal{G}}_\delta$ has a locally unique Nash equilibrium $\bar{y}_\delta^*$ in $\bar{B}_R(y^*)$, and*

$$\|\bar{y}_\delta^* - y^*\| \le \frac{2L_f\delta}{\bar{\mu}}$$

$$\|\bar{\lambda}_\delta^* - \lambda^*\| \le \|B^\dagger\| L_f (1 + 2\frac{\bar{C}}{\bar{\mu}})\delta$$

Proof. Let $\gamma = 1 - 2\bar{\mu}\eta + \eta^2\bar{C}^2$. The feasible set is $Y = \{(y_1, \cdots, y_N) \in \mathbb{R}^n : e^{(i)}(x, y_i) = 0\,, \forall i \in [N]\}$. Denote $T(y) = \Pi_Y(y - \eta\bar{V}(x, y))$. From Lemma E.1, we have $\forall y, z \in \bar{B}_R(y^*)$,

$$\|T(y) - T(z)\|^2 \leq \gamma\|y - z\|^2$$

If $0 < \eta < \frac{2\bar{\mu}}{\bar{C}^2}$, then $0 < \gamma < 1$ and thus $T(y)$ is a contraction mapping in y. We want to prove that T maps from $\bar{B}_R(y^*)$ to $\bar{B}_R(y^*)$. Since T is a contraction, there exists a unique fixed point $\bar{y}_\delta^* = T(\bar{y}_\delta^*)$ by Banach fixed-point theorem. The fixed point is the Nash equilibrium by Lemma A.3.

Denote $H(y) = \Pi_Y(y - \eta V(x, y))$. We know that $y^* = H(y^*)$. KKT condition says that $V(x, y^*) + B^\top\lambda^* = 0$. Let $y \in \bar{B}_R(y^*)$,

$$\begin{aligned}
\|T(y) - y^*\| &\leq \|T(y) - T(y^*)\| + \|T(y^*) - H(y^*)\| + \|H(y^*) - y^*\| \\
&\leq \sqrt{\gamma}\|y - y^*\| + \eta\|-V(x, y^*) - B^\top\lambda^* - \delta\nabla_y f(x, y^*)\| + 0 \\
&\leq \sqrt{\gamma}\|y - y^*\| + \eta L_f\delta \\
&\leq \sqrt{\gamma}R + \eta\delta L_f
\end{aligned}$$

Thus, we can show the existence of the Nash equilibrium $\bar{y}_\delta^*$ and the desired bound on $\|\bar{y}_\delta^* - y^*\|$ using the same steps in Lemma F.2.

By KKT condition,

$$\begin{aligned}
\bar{V}(x, \bar{y}_\delta^*) + B^\top\bar{\lambda}_\delta^* &= 0 \\
V(x, y^*) + B^\top\lambda^* &= 0
\end{aligned}$$

subtracting these two equations and rearranging terms,

$$\begin{aligned}
\|\bar{\lambda}_\delta^* - \lambda^*\| &\leq \|B^\dagger\|\|\bar{V}(x, \bar{y}_\delta^*) - \bar{V}(x, y^*) + \bar{V}(x, y^*) - V(x, y^*)\| \\
&\leq \|B^\dagger\|(\bar{C}\|\bar{y}_\delta^* - y^*\| + \delta L_f) \\
&\leq \|B^\dagger\|L_f(1 + 2\frac{\bar{C}}{\bar{\mu}})\delta
\end{aligned}$$

□

Let $K(x, y) = \begin{bmatrix} \frac{\partial}{\partial y}V(x, y) & B^\top \\ B & 0 \end{bmatrix}$. Let $K_\epsilon(x, y) = \begin{bmatrix} \frac{\partial}{\partial y}V(x, y) + \epsilon D(x, y) & B^\top \\ B & 0 \end{bmatrix}$.

Lemma G.3. *Given a matrix $M = \begin{bmatrix} J & B^\top \\ B & 0 \end{bmatrix}$ where J is invertible. Then,*

$$\|M^{-1}\| \leq (1 + \|J^{-1}\|\|B\|)^2 \max(\|J^{-1}\|, \|(BJ^{-1}B^\top)^{-1}\|)$$

Proof. Let $S = -BJ^{-1}B^\top$. Let $L = \begin{bmatrix} I & 0 \\ -BJ^{-1} & I \end{bmatrix}$ and $U = \begin{bmatrix} I & -J^{-1}B^\top \\ 0 & I \end{bmatrix}$. Using L to eliminate rows of M and U to eliminates columns of M, we have

$$M = L^{-1}\begin{bmatrix} J & 0 \\ 0 & S \end{bmatrix}U^{-1}$$

Taking inverse,

$$\begin{aligned}
M^{-1} &= U\begin{bmatrix} J^{-1} & 0 \\ 0 & S^{-1} \end{bmatrix}L \\
&= \begin{bmatrix} I & -J^{-1}B^\top \\ 0 & I \end{bmatrix}\begin{bmatrix} J^{-1} & 0 \\ 0 & S^{-1} \end{bmatrix}\begin{bmatrix} I & 0 \\ -BJ^{-1} & I \end{bmatrix}
\end{aligned}$$

Then,

$$\|M^{-1}\| \leq (1 + \|J^{-1}\|\|B\|)^2 \max(\|J^{-1}\|, \|(BJ^{-1}B^\top)^{-1}\|)$$

□

Lemma G.4.

$$\| \begin{bmatrix} \frac{\bar{y}_\delta^* - y^*}{\delta} \\ \frac{\bar{\lambda}_\delta^* - \lambda^*}{\delta} \end{bmatrix} - (K_\epsilon(x, y^*))^{-\top} (- \begin{bmatrix} \nabla_y f(x, y^*) \\ 0 \end{bmatrix}) \| \leq \mathcal{O}(\frac{\sqrt{N}\delta}{\bar{\mu}^2})$$

Proof. Let $y = \bar{y}_\delta^*$ and $\lambda = \bar{\lambda}_\delta^*$. for all $i \in [N]$, we have

$$\begin{aligned}
0 &= \nabla_{y_i} \bar{g}_i(x, y) + (B^\top \lambda)_i \\
&= \int_0^1 \sum_{j=1}^N J_{ji}(x, y_j + t\epsilon(y_j - y_j^*), y_{\neg j})^\top (y_j - y_j^*) dt \\
&+ \epsilon \int_0^1 J_{ii}(x, y_i + t\epsilon(y_i - y_i^*), y_{\neg i})(y_i - y_i^*) dt + \delta \nabla_{y_i} f(x, y) + (B^\top(\lambda - \lambda^*))_i \\
&= \underbrace{\int_0^1 \sum_{j=1}^N [J_{ji}(x, y_j + t\epsilon(y_j - y_j^*), y_{\neg j})^\top - J_{ji}(x, y_j^*, y_{\neg j})^\top](y_j - y_j^*) dt}_{\text{(a)}} \\
&+ \underbrace{\epsilon \int_0^1 [J_{ii}(x, y_i + t\epsilon(y_i - y_i^*), y_{\neg i}) - J_{ii}(x, y_i^*, y_{\neg i})](y_i - y_i^*) dt}_{\text{(b)}} \\
&+ \underbrace{\delta(\nabla_{y_i} f(x, y) - \nabla_{y_i} f(x, y^*))}_{\text{(c)}} \\
&+ \underbrace{\sum_{j=1}^N [J_{ji}(x, y_j^*, y_{\neg j})^\top - J_{ji}(x, y^*)^\top](y_j - y_j^*)}_{\text{(d)}} \\
&+ \underbrace{[J_{ii}(x, y_i^*, y_{\neg i})^\top - J_{ii}(x, y^*)^\top](y_i - y_i^*)}_{\text{(e)}} \\
&+ \sum_{j=1}^N J_{ji}(x, y^*)^\top (y_j - y_j^*) + \epsilon J_{ii}(x, y^*)(y_i - y_i^*) + \delta \nabla_{y_i} f(x, y^*) + + (B^\top(\lambda - \lambda^*))_i
\end{aligned}$$

$$\begin{aligned}
\|a\| &\leq S(1 + \frac{\epsilon}{2}) \|y - y^*\|^2 \\
\|b\| &\leq S(1 + \frac{\epsilon}{2}) \|y - y^*\|^2 \\
\|c\| &\leq \delta L_f \|y - y^*\| \\
\|d\| &\leq S \|y - y^*\|^2 \\
\|e\| &\leq \epsilon S \|y - y^*\|^2
\end{aligned}$$

Combining the equations for $i = 1, 2, \cdots, N$,

$$J(x, y^*)^\top (y - y^*) + \epsilon D(x, y^*)(y - y^*) + \delta \nabla_y f(x, y^*) + B^\top(\lambda - \lambda^*) + \text{extra} = 0$$

where $\|\text{extra}\| \leq \sqrt{N}(S(1 + \frac{\epsilon}{2})\|y - y^*\|^2 + S(1 + \frac{\epsilon}{2})\|y - y^*\|^2) + \delta L_f \|y - y^*\| + S\|y - y^*\|^2 + \epsilon S \|y - y^*\|^2$ Solving the equation with the constraint that $B(y - y^*) = 0$, we have

$$\begin{bmatrix} \frac{\bar{y}_\delta^* - y^*}{\delta} \\ \frac{\bar{\lambda}_\delta^* - \lambda^*}{\delta} \end{bmatrix} = (K_\epsilon(x, y^*))^{-\top} (- \begin{bmatrix} \nabla_y f(x, y^*) + \frac{\text{extra}}{\delta} \\ 0 \end{bmatrix})$$

Using Lemma G.2, we have

$$
\begin{aligned}
&\| \begin{bmatrix} \frac{\bar{y}_\delta^* - y^*}{\delta} \\ \frac{\bar{\lambda}_\delta^* - \lambda^*}{\delta} \end{bmatrix} - (K_\epsilon(x, y^*))^{-\top} (- \begin{bmatrix} \nabla_y f(x, y^*) \\ 0 \end{bmatrix}) \| \\
&\leq \|K_\epsilon(x, y^*)^{-\top}\| \| \frac{\text{extra}}{\delta} \| \\
&\leq (1 + \frac{1}{\mu(1+\epsilon)} \|B\|)^2 \max(\frac{1}{\mu(1+\epsilon)}, \frac{C^2}{\mu(1+\epsilon)\sigma_{\min}(B)^2}) [\mathcal{O}(\frac{\sqrt{N}\delta}{\bar{\mu}^2}) + \mathcal{O}(\frac{\sqrt{N}\delta\epsilon}{\bar{\mu}^2})] \text{ (Lemma G.3)} \\
&\leq \mathcal{O}(\frac{\sqrt{N}\delta}{\bar{\mu}^2}) \quad \text{(we ignore } \mathcal{O}(\delta\epsilon)\text{)}
\end{aligned}
$$

Denote $G = J(x, y^*) + \epsilon D(x, y^*)$. The second inequality uses Lemma G.3 and the fact that $d^\top G d \geq (\mu + \epsilon\mu)\|d\|^2$. To bound $\|(BG^{-\top}B^\top)^{-1}\|$, we use Lemma A.7 and Lemma A.5:

$$
\begin{aligned}
\|(BG^{-\top}B^\top)^{-1}\| &\leq \frac{1}{\sigma_{\min}(BG^{-\top}B^\top)} \\
&\leq \frac{C^2}{\mu(1+\epsilon)\sigma_{\min}(B)^2}
\end{aligned}
$$

□

Lemma G.5.

$$
\|K_\epsilon(x, y^*)^{-\top}(-\nabla_y f(x, y^*)) - K(x, y^*)^{-\top}(-\nabla_y f(x, y^*))\| \leq \mathcal{O}(\epsilon)
$$

Proof.

$$
\begin{aligned}
&\|K_\epsilon(x, y^*)^{-\top} \begin{bmatrix} -\nabla_y f(x, y^*) \\ 0 \end{bmatrix} - K(x, y^*)^{-\top} \begin{bmatrix} -\nabla_y f(x, y^*) \\ 0 \end{bmatrix} \| \\
&= \|K_\epsilon(x, y^*)^{-\top} [K(x, y^*)^\top - K_\epsilon(x, y^*)^\top] K(x, y^*)^{-\top} (-\nabla_y f(x, y^*))\| \\
&\leq \mathcal{O}(L_f \epsilon) \\
&= \mathcal{O}(\epsilon)
\end{aligned}
$$

The first equality uses the matrix identity $A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}$ for any invertible matrices A and B of the same dimensions. The second inequality uses lemma G.3 which says $\|K_\epsilon(x, y^*)^{-1}\| = \mathcal{O}(1)$ and $\|K(x, y^*)^{-1}\| = \mathcal{O}(1)$ as ϵ decreases. □

G.1 Proof of Theorem G.7

Proof.

$$
\begin{aligned}
v_\delta(x, \bar{y}_\delta^*, \bar{\lambda}_\delta^*) &= \frac{1}{\delta} \sum_{i=1}^N [\nabla_x g_i(x, \bar{y}_{\delta,i}^*, y_{\neg i}^*) - \nabla_x g_i(x, y^*) + \frac{\partial}{\partial x} e^{(i)}(x, y_i^*)^\top (\bar{\lambda}_\delta^* - \lambda^*)] \\
&= \frac{1}{\delta} \sum_{i=1}^N [\frac{\partial}{\partial y_i} \nabla_x g_i(x, y^*)(\bar{y}_{\delta,i}^* - y_i^*) + \text{H.O.T.} + \frac{\partial}{\partial x} e^{(i)}(x, y_i^*)^\top (\bar{\lambda}_\delta^* - \lambda^*)] \\
&= \frac{\partial}{\partial x} V(x, y^*)^\top \frac{\bar{y}_\delta^* - y^*}{\delta} + N \frac{\text{H.O.T.}}{\delta} + A^\top \frac{\bar{\lambda}_\delta^* - \lambda^*}{\delta}]
\end{aligned}
$$

where $\|\text{H.O.T.}\| = \mathcal{O}(\|\bar{y}_\delta^* - y^*\|^2) \leq \mathcal{O}(\frac{\delta^2}{\bar{\mu}^2})$ by Lemma G.2. Let $v_\epsilon^* := K_\epsilon(x, y^*)^{-\top} \begin{bmatrix} -\nabla_y f(x, y^*) \\ 0 \end{bmatrix}$ and $v^* := K(x, y^*)^{-\top} \begin{bmatrix} -\nabla_y f(x, y^*) \\ 0 \end{bmatrix}$. Using Lemma G.4 and

G.5, we have

$$\begin{aligned}
&\|v_\delta(x,\bar{y}_\delta^*,\bar{\lambda}_\delta^*) - \begin{bmatrix}\frac{\partial}{\partial x}V(x,y^*)^\top & A^\top\end{bmatrix}\begin{bmatrix}\frac{\partial}{\partial y}V(x,y^*) & B^\top \\ B & 0\end{bmatrix}^{-\top}\begin{bmatrix}-\nabla_y f(x,y^*)\\ 0\end{bmatrix}\| \\
&\le \|v_\delta(x,\bar{y}_\delta^*,\bar{\lambda}_\delta^*) - \begin{bmatrix}\frac{\partial}{\partial x}V(x,y^*)^\top & A^\top\end{bmatrix} v_\epsilon^*\| \\
&+ \|\begin{bmatrix}\frac{\partial}{\partial x}V(x,y^*)^\top & A^\top\end{bmatrix}(v_\epsilon^* - v^*)\| \\
&\le (C+\|A\|)(\mathcal{O}(\frac{\sqrt{N}\delta}{\bar{\mu}^2}) + \mathcal{O}(\frac{N\delta}{\bar{\mu}^2}) + \mathcal{O}(\epsilon)) \\
&= \mathcal{O}(\frac{N\delta}{\bar{\mu}^2} + \epsilon)
\end{aligned}$$

□

Lemma G.6. *Let $\hat{y}$ and $\hat{\lambda}$ be approximate solutions of the original lower-level game such that $\|y^* - \hat{y}\| \le \hat{\delta}$ and $\|\lambda^* - \hat{\lambda}\| \le \hat{\delta}$. Define $\hat{\phi}_\epsilon(x,y) = \frac{1}{\epsilon}\sum_{j=1}^N g_j(x, y_j + \epsilon(y_j - \hat{y}_j, y_{\neg_j}) - g_j(x,y)$, $\hat{g}_i(x,y) = \hat{\phi}_\epsilon(x,y) - g_i(x,y) + \delta f(x,y) - (\hat{\lambda})^\top By$ and $\hat{V}(x,y) = (\nabla_{y_1}\hat{g}_1(x,y), \cdots, \nabla_{y_N}\hat{g}_N(x,y))$. Then*

$$\|\bar{V}(x,y) - \hat{V}(x,y)\| \le (CN + \|B\|)\hat{\delta}$$

Proof.

$$\|\nabla_{y_i}\hat{\phi}_\epsilon(x,y) - \nabla_{y_i}\phi_\epsilon(x,y)\| \le \frac{1}{\epsilon}\sum_j C\|\epsilon(y_j^* - \hat{y}_j)\| = C\sqrt{N}\|y^* - \hat{y}\|$$

$$\|V(x,y) - \hat{V}(x,y)\| \le CN\|y^* - \hat{y}\| + \|B\|\|\lambda^* - \hat{\lambda}\| \le (CN + \|B\|)\hat{\delta}$$

□

Theorem G.7. *For $0 < \delta < \frac{\mu}{2C_f}$ and $0 < R < \frac{\mu - \delta C_f}{S(N+1)(2+\sqrt{N})}$ and $0 < \epsilon < 1$, $\bar{V}$ is locally $\bar{\mu}$-strongly monotone and $\bar{C}$-smooth on $\bar{B}_R(y^*) = \{y \in \mathbb{R}^n : \|y - y^*\| \le R\}$ for some $\bar{\mu}$, $\bar{C} > 0$. Furthermore, if $\delta < \min(\frac{\mu}{2C_f}, \frac{\bar{\mu}R}{2L_f})$, then we have a fully first-order hypergradient approximation with non-asymptotic rate*

$$\|v_\delta(x,\bar{y}_\delta^*,\bar{\lambda}_\delta^*) - \begin{bmatrix}\frac{\partial}{\partial x}V(x,y^*)^\top & A^\top\end{bmatrix}\begin{bmatrix}\frac{\partial}{\partial y}V(x,y^*) & B^\top \\ B & 0\end{bmatrix}^{-\top}\begin{bmatrix}-\nabla_y f(x,y^*)\\ 0\end{bmatrix}\| \le \mathcal{O}(\frac{N\delta}{\bar{\mu}^2} + \epsilon)$$

Proof Sketch. Lemma F.1 proves local strong monoticity and smoothness. For $\delta < \min(\frac{\mu}{2C_f}, \frac{\bar{\mu}R}{2L_f})$, Lemma G.2 shows that $T(y) = \Pi_Y(y - \eta\bar{V}(x,y))$ is a contraction on $\bar{B}_R(y^*) \cap Y$. It proves the existence of Nash equilibrium $\bar{y}_\delta^*$ and $\|\bar{y}_\delta^* - y^*\| \le \frac{2L_f\delta}{\bar{\mu}}$. Let $v_\epsilon^* := K_\epsilon(x,y^*)^{-\top}\begin{bmatrix}-\nabla_y f(x,y^*)\\ 0\end{bmatrix}$ and $v^* := K(x,y^*)^{-\top}\begin{bmatrix}-\nabla_y f(x,y^*)\\ 0\end{bmatrix}$. Triangle inequality reduces the proof into bounding $\|v_\delta(x,\bar{y}_\delta^*,\bar{\lambda}_\delta^*) - \begin{bmatrix}\frac{\partial}{\partial x}V(x,y^*)^\top & A^\top\end{bmatrix} v_\epsilon^*\| + \|\begin{bmatrix}\frac{\partial}{\partial x}V(x,y^*)^\top & A^\top\end{bmatrix}(v_\epsilon^* - v^*)\|$. Combining the following equations yields the desired result,

$$\|v_\epsilon^* - v^*\| \le \mathcal{O}(\epsilon) \text{ (Lemma G.5)}$$

$$\|\begin{bmatrix}\frac{\bar{y}_\delta^* - y^*}{\delta}\\ \frac{\bar{\lambda}_\delta^* - \lambda^*}{\delta}\end{bmatrix} - v_\epsilon^*\| \le \frac{\sqrt{N}}{\bar{\mu}^2}\mathcal{O}(\delta) \text{ (Lemma G.4)}$$

$$v_\delta(x,\bar{y}_\delta^*,\bar{\lambda}_\delta^*) = \frac{\partial}{\partial x}V(x,y^*)^\top\frac{\bar{y}_\delta^* - y^*}{\delta} + N\frac{\text{H.O.T.}}{\delta} + A^\top\frac{\bar{\lambda}_\delta^* - \lambda^*}{\delta} \text{ (Taylor's expansion)}$$

where $\|\text{H.O.T.}\| = \mathcal{O}(\|\bar{y}_\delta^* - y^*\|^2) \le \mathcal{O}(\frac{\delta^2}{\bar{\mu}^2})$. □

Theorem G.8. *Consider problem P1. Under assumptions 3.4, 3.5 and G.1, inexact gradient descent (algorithm 1) outputs an $\mathcal{O}(\epsilon)$-stationary point of the hyperobjective F with $\tilde{\mathcal{O}}(N^2\epsilon^{-2})$ first-order oracle calls to f and g_i in both equality constrained and unconstrained settings.*

Proof. Following the argument of Theorem B.7, the hyperobjective F is smooth. Choose $\delta \leq \frac{\epsilon}{2N}$ and use the potential function $\phi_{\frac{\epsilon}{2}}$. By Lemma D.1, if the approximate lower-level solutions satisfy $\|y^* - \hat{y}\| \leq \delta^2$ and $\|\bar{y}^*_\delta - \hat{y}_\delta\| \leq \delta^2$, then the inexact hypergradient $\hat{\nabla}F(x) = \nabla_x f(x, \hat{y}) + \hat{v}_\delta(x, \hat{y}_\delta)$ satisfies $\|\hat{\nabla}F(x) - \nabla F(x)\| \leq \mathcal{O}(N\delta + \epsilon/2) \leq \mathcal{O}(\epsilon)$. Therefore, by Lemma D.2, inexact gradient descent requires $\mathcal{O}(\epsilon^{-2})$ hypergradient evaluations to reach an $\mathcal{O}(\epsilon)$-stationary point of F. It remains to bound the cost of each hypergradient evaluation. Since the potential is constructed using the approximate equilibrium $\hat{y}$, its gradient operator is also inexact, with an error of $\mathcal{O}(\epsilon)$ (Lemma G.6). Moreover, the auxiliary game is only locally strongly monotone on $\bar{B}_R(y^*)$. We therefore choose $\delta \leq \min\{\epsilon/(2N), \bar{\mu}R/(4L_f), \sqrt{R/2}\}$, which ensures $\|\bar{y}^*_\delta - \hat{y}\| \leq \|\bar{y}^*_\delta - y^*\| + \|y^* - \hat{y}\| \leq R$. Hence, gradient descent initialized at $\hat{y}$ remains in the local neighborhood and converges linearly to the auxiliary equilibrium. By Lemma E.2, $\mathcal{O}(\log(1/\delta^2))$ lower-level iterations are sufficient to obtain the required δ^2-accurate solutions. Each evaluation of the auxiliary game operator requires gradients of all N surrogate costs, and evaluating each surrogate cost requires $\mathcal{O}(N)$ calls to the original follower objectives g_i. Thus, each lower-level iteration uses $\mathcal{O}(N^2)$ first-order oracle calls. Combining this with $\mathcal{O}(\epsilon^{-2})$ outer iterations and $\mathcal{O}(\log(1/\delta^2))$ lower-level iterations gives a total oracle complexity of $\mathcal{O}(N^2\epsilon^{-2}\log(1/\delta^2)) = \widetilde{\mathcal{O}}(N^2\epsilon^{-2})$. □

H Extension to Stackelberg Game with Linear Inequality Constraints

Consider the following Stackelberg game with linear inequality constraints,

$$\begin{aligned}
&\min_{x\in X} F(x) := f(x, y^*(x)) \\
&\text{s.t. } g_i(x, y^*(x)) \leq g_i(x, y_i, y^*_{\neg i}(x)), \ \forall i \in [N], \ \forall y_i \in Y_i(x) \\
&\quad Y_i(x) = \left\{y_i \in \mathbb{R}^{n_i} : e^{(i)}(x, y_i) \leq 0\right\} \\
&\quad e^{(i)}(x, y_i) = A^{(i)}x + B^{(i)}y_i + b^{(i)}
\end{aligned} \tag{Q1}$$

In problem Q1, the hyper-objective F is Lipschitz continuous but nonsmooth because the active set may change discontinuously. (δ, ϵ)-stationarity is used to measure convergence because obtaining an ϵ-stationary point of a nonsmooth nonconvex function is intractable (Zhang et al. (2020)).

Definition H.1 ((δ, ϵ)-stationarity)**.** Let $F : \mathbb{R}^d \to \mathbb{R}$ be a locally Lipschitz continuous function. Let $x \in \mathbb{R}^d$. Let $B_\delta(x)$ be the Euclidean ball of radius $\delta > 0$ centered at x. The generalized gradient of F at x is the set $\partial F(x) = \text{conv}(\{g | g = \lim_{k\to\infty} \nabla f(x_k), \ x_k \to x\})$ (Clarke (1990)). Also, $\partial F(x) = \cap_{\delta>0} \cup_{y\in B_\delta(x)} \partial F(y)$. The Goldstein subdifferential of F at x is the set $\partial_\delta F(x) = \text{conv}(\cup_{y\in B_\delta(x)} \partial F(y))$ (Goldstein (1977)). x is called a (δ, ϵ)-stationary point of a locally Lipschitz continuous function F if $\text{dist}(0, \partial_\delta F(x)) \leq \epsilon$, where $\text{dist}(x, S) := \inf_{y\in S} \|x - y\|$.

Assumption H.2 (Active set identification)**.** For problem Q1, assume that the active constraints can be correctly identified.

By identifying the active set at the Nash equilibrium (similar to Zhao et al. (2025)), we can transform an arbitrary smooth and strongly monotone game $\mathcal{G} := (x, [N], \{Y_i\}_{i\in[N]}, \{g_i\}_{i\in[N]})$ with linear inequality constraints into an equivalent game with linear equality constraints. Then, we can directly apply the hypergradient approximation derived for Stackelberg game with linear equality constraints in Section 4 or G.

Given an arbitrary reference point $\bar{x} \in X$. For each player $i \in [N]$, we define the strict active set $\mathcal{I}_i := \{j \in [m_i] : (\lambda^*_i)_j > 0\}$, active dual variable $\lambda^*_{i,\mathcal{I}_i} = (\lambda^*_{i,j})_{j\in\mathcal{I}_i}$ and the active equality constraint function $e^{(i)}_{\mathcal{I}_i} = (e^{(i)}_j)_{j\in\mathcal{I}_i}$. We define the ghost cost function

$$\tilde{g}_i(x, y) = g_i(x, y) + (\lambda^*_i)^\top e^{(i)}(x, y_i)$$

and the ghost equality constraint $\tilde{Y}_i = \{y_i \in \mathbb{R}^{n_i} : \tilde{e^{(i)}}(x, y_i) = 0\}$, where

$$\tilde{e^{(i)}}(x, y_i) = \frac{\partial}{\partial x} e^{(i)}_{\mathcal{I}_i}(\bar{x}, y_i^*)(x - \bar{x}) + \frac{\partial}{\partial y} e^{(i)}_{\mathcal{I}_i}(\bar{x}, y_i^*)(y_i - y_i^*)$$

are the active constraints. We replace the original lower-level game $\mathcal{G} = (x, [N], \{Y_i\}_{i\in[N]}, \{g_i\}_{i\in[N]})$ with ghost lower-level game $\tilde{\mathcal{G}} := (x, [N], \{\tilde{Y}_i\}_{i\in[N]}, \{\tilde{g}_i\}_{i\in[N]})$. Denote $\tilde{y}^*$ as the Nash equilibrium of $\tilde{\mathcal{G}}$ and $\tilde{\lambda}^*$ as the optimal dual solution for $\tilde{\mathcal{G}}$. The ghost problem is defined as

$$\begin{aligned} &\min_{x\in X} \tilde{F}(x) := f(x, \tilde{y}^*(x)) \\ &\text{s.t. } \tilde{g}_i(x, \tilde{y}^*(x)) \le \tilde{g}_i(x, y_i, \tilde{y}^*_{\neg i}(x)) \,, \forall i \in [N] \,, \forall y_i \in \tilde{Y}_i(x) \\ &\quad \tilde{Y}_i(x) = \left\{ y_i \in \mathbb{R}^{n_i} : \frac{\partial}{\partial x} e^{(i)}_{\mathcal{I}_i}(\bar{x}, y_i^*)(x - \bar{x}) + \frac{\partial}{\partial y} e^{(i)}_{\mathcal{I}_i}(\bar{x}, y_i^*)(y_i - y_i^*) = 0 \right\} \end{aligned} \tag{Q2}$$

By comparing KKT conditions of the original and ghost lower-level problem (Lemma H.4), the ghost problem Q2 is in fact equivalent to the original Stackelberg game with inequality constraints Q1, such that $\tilde{y}^*(\bar{x}) = y^*(\bar{x})$ and $\frac{d}{dx}\tilde{y}^*(\bar{x}) = \frac{d}{dx}y^*(\bar{x})$. Hence, $\tilde{F}(\bar{x}) = F(\bar{x})$ and $\nabla\tilde{F}(\bar{x}) = \nabla F(\bar{x})$.

Lemma H.3. *The ghost lower-level game $\tilde{\mathcal{G}} := (x, [N], \{\tilde{Y}_i\}_{i\in[N]}, \{\tilde{g}_i\}_{i\in[N]})$ in problem Q2 is μ-strongly monotone, C-smooth and S-Hessian smooth.*

Proof. Define $\tilde{V}(x, y) := (\nabla_{y_1}\tilde{g}_1(x, y), \cdots, \nabla_{y_N}\tilde{g}_N(x, y)) \in \mathbb{R}^n$. Let $x, x' \in X$ and $y, y' \in \tilde{Y}$,

$$\begin{aligned} \tilde{V}(x', y') - \tilde{V}(x, y) &= V(x', y') - V(x, y) + B^\top\lambda^* - B^\top\lambda^* \\ &= V(x', y') - V(x, y) \end{aligned}$$

We can conclude the proof using the definitions of strong monotonicity, smoothness and Hessian smoothness in 3.4. □

Lemma H.4 (Equivalence of the ghost problem and the original problem)**.** *Let $\tilde{y}^*(x) \in NE(\tilde{\mathcal{G}})$ be the Nash equilibrium of the ghost lower-level game Q2 and let $y^*(x) \in NE(\mathcal{G})$ be the Nash equilibrium of the original lower-level game Q1. For all $\bar{x} \in X$ where active set is locally constant, $\tilde{y}^*(\bar{x}) = y^*(\bar{x})$, $\frac{d}{dx}\tilde{y}^*(\bar{x}) = \frac{d}{dx}y^*(\bar{x})$, $\tilde{\lambda}^*(\bar{x}) = 0$ where $\tilde{\lambda}^*$ is the optimal dual variable of $\tilde{\mathcal{G}}$*

Proof. Denote $B_{\mathcal{I}} := \operatorname{diag}((\frac{\partial}{\partial y} e^{(i)}_{\mathcal{I}_i}(\bar{x}, y_i^*))_{i\in[N]})$ as a block diagonal matrix corresponding to Jacobians of active inequality constraints, and $A_{\mathcal{I}} := \begin{bmatrix} \frac{\partial}{\partial x} e^{(1)}_{\mathcal{I}_1}(\bar{x}, y_1^*) \\ \vdots \\ \frac{\partial}{\partial x} e^{(N)}_{\mathcal{I}_N}(\bar{x}, y_N^*) \end{bmatrix}$. The KKT condition of the original problem Q1 is

$$\Gamma(\bar{x}, y^*, \lambda^*) = \begin{bmatrix} V(\bar{x}, y^*) + B_{\mathcal{I}}^\top\lambda^*_{\mathcal{I}} \\ \begin{bmatrix} \operatorname{diag}(\lambda^*_{1,\mathcal{I}_1}) e^{(1)}_{\mathcal{I}_1}(\bar{x}, y_1^*) \\ \vdots \\ \operatorname{diag}(\lambda^*_{N,\mathcal{I}_N}) e^{(N)}_{\mathcal{I}_N}(\bar{x}, y_N^*) \end{bmatrix} \end{bmatrix} = 0$$

The jacobian

$$\frac{\partial}{\partial(y^*, \lambda^*_{\mathcal{I}})}\Gamma(\bar{x}, y^*, \lambda^*_{\mathcal{I}}) = \begin{bmatrix} \frac{\partial}{\partial y}V(\bar{x}, y^*) & B_{\mathcal{I}}^\top \\ \operatorname{diag}(\lambda^*_{\mathcal{I}})B_{\mathcal{I}} & 0 \end{bmatrix}$$

is invertible following the same steps in Lemma A.8. Then by implicit function theorem, $y^*(\bar{x})$, $\lambda^*_{\mathcal{I}}(\bar{x})$ are continuously differentiable functions of x and their derivatives are given by

$$\begin{bmatrix} \frac{\partial}{\partial y}V(\bar{x}, y^*) & B_{\mathcal{I}}^\top \\ \operatorname{diag}(\lambda^*_{\mathcal{I}})B_{\mathcal{I}} & 0 \end{bmatrix} \begin{bmatrix} \frac{d}{dx}y^*(\bar{x}) \\ \frac{d}{dx}\lambda^*_{\mathcal{I}}(\bar{x}) \end{bmatrix} = \begin{bmatrix} -\frac{\partial}{\partial x}V(\bar{x}, y^*) \\ -\operatorname{diag}(\lambda^*_{\mathcal{I}})A_{\mathcal{I}} \end{bmatrix} \tag{13}$$

$$\begin{bmatrix} \frac{\partial}{\partial y}V(\bar{x}, y^*) & B_{\mathcal{I}}^\top \\ B_{\mathcal{I}} & 0 \end{bmatrix} \begin{bmatrix} \frac{d}{dx}y^*(\bar{x}) \\ \frac{d}{dx}\lambda^*_{\mathcal{I}}(\bar{x}) \end{bmatrix} = \begin{bmatrix} -\frac{\partial}{\partial x}V(\bar{x}, y^*) \\ -A_{\mathcal{I}} \end{bmatrix} \tag{14}$$

On the other hand, the KKT condition of the ghost problem Q2 is

$$\tilde{\Gamma}(\bar{x}, \tilde{y}^*, \tilde{\lambda}^*) = \begin{bmatrix} V(\bar{x}, \tilde{y}^*) + B_{\mathcal{I}}^\top \lambda_{\mathcal{I}}^* + B_{\mathcal{I}}^\top \tilde{\lambda}^* \\ \begin{bmatrix} \tilde{e}_{\mathcal{I}_1}^{(1)}(\bar{x}, \tilde{y}_1^*) \\ \vdots \\ \tilde{e}_{\mathcal{I}_N}^{(N)}(\bar{x}, \tilde{y}_N^*) \end{bmatrix} \end{bmatrix} = 0$$

Substituting $\tilde{y}^* = y^*$ and $\tilde{\lambda}^* = 0$, the KKT condition is satisfied. By LICQ and strong monotonicity (assumptions G.1, 3.4), the unique Nash equilibrium and optimal dual variable of the ghost lower-level game are $\tilde{y}^* = y^*$ and $\tilde{\lambda}^* = 0$ respectively. Using implicit function theorem, the derivatives of $\tilde{y}^*(\bar{x})$ and $\tilde{\lambda}^*(\bar{x})$ are given by

$$\begin{bmatrix} \frac{\partial}{\partial y} V(\bar{x}, \tilde{y}^*) & B_{\mathcal{I}}^\top \\ B_{\mathcal{I}} & 0 \end{bmatrix} \begin{bmatrix} \frac{d}{dx} \tilde{y}^*(\bar{x}) \\ \frac{d}{dx} \tilde{\lambda}_{\mathcal{I}}^*(\bar{x}) \end{bmatrix} = \begin{bmatrix} -\frac{\partial}{\partial x} V(\bar{x}, \tilde{y}^*) \\ -A_{\mathcal{I}} \end{bmatrix} \tag{15}$$

The derivatives of the KKT conditions of Q1 and Q2 are the same. Hence, $\frac{d}{dx}\tilde{y}^*(\bar{x}) = \frac{d}{dx} y^*(\bar{x})$. □

Assuming exact active set identification (assumption H.2), we can directly apply the finite difference hypergradient approximation in problem P1 and the gradient approximation guarantees follow.

However, the hyper-objective F is nonsmooth since the active set can change with x discontinuously. We can utilize Lemma H.5 (see Kornowski et al. (2024)) for minimizing Lipschitz nonsmooth nonconvex function F, and then we can conclude the convergence rate for problem Q1 with theorem H.6. Note that $\mathcal{O}(\hat{\delta}^{-1}\epsilon^{-3})$ for reaching a $(\hat{\delta}, \epsilon)$-stationary point is the optimal rate for nonsmooth nonconvex optimization Cutkosky et al. (2023).

Lemma H.5. *Suppose $F : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz and that $\|\hat{\nabla} F(\cdot) - \nabla F(\cdot)\| \leq \alpha$. Then there exists an algorithm which outputs a $(\hat{\delta}, \epsilon + \alpha)$-stationary point with $T = \mathcal{O}(\hat{\delta}^{-1}\epsilon^{-3})$ calls to $\hat{\nabla} F(\cdot)$ for arbitrary $\hat{\delta} > 0$.*

Theorem H.6. *Consider problem Q1. Under assumptions 3.4, 3.5, G.1 and H.2, there exists an algorithm that outputs an $(\hat{\delta}, 2\epsilon)$-stationary point of the hyper-objective F with $\tilde{\mathcal{O}}(\hat{\delta}^{-1}\epsilon^{-3})$ oracle calls to f and g_i for $\hat{\delta} > 0$.*

Proof. The proof is identical to that of theorem C.6 and G.8. Except that the complexity of the outer loop is $\mathcal{O}(\hat{\delta}^{-1}\epsilon^{-3})$ instead of $\mathcal{O}(\epsilon^{-2})$ □

I Discussion: Challenges of Fully First-Order Hypergradient Approximation in Asymmetric Game

Let $x \in \mathbb{R}$. Suppose the upper-level objective function is $f(x, y) = u^\top y$. Consider an unconstrained game $(x, [N], \{\mathbb{R}^{n_i}\}_{i\in[N]}, \{g_i\}_{i\in[N]})$, where the the cost function for each player $i \in [N]$ is

$$g_i(x, y) = \frac{1}{2} y_i^\top Q_{ii} y_i + y_i^\top \sum_{j \neq i} Q_{ij} y_j + x c_i^\top y_i$$

At equilibrium y^*, $\nabla_{y_i} g_i(x, y_i^*, y_{\neg i}^*) = 0, \forall i \in [N]$. Then,

$$V(x, y^*) = \begin{bmatrix} \nabla_{y_1} g_1(x, y^*) \\ \vdots \\ \nabla_{y_N} g_N(x, y^*) \end{bmatrix} = Qy + xc$$

Solving $V(x, y^*) = 0$, we have

$$y^* = -Q^{-1}(xc)$$

Using implict differentiation, the exact hypergradient term to be approximated is

$$\begin{aligned}(\frac{d}{dx}y^*(x))^\top \nabla_y f(x, y^*(x)) &= (-Q^{-1}c)^\top u \\ &= -c^\top Q^{-\top} u \end{aligned} \tag{16}$$

We may add perturbation $\delta f(x, y)$ to each player's cost function. The equilibrium y^*_δ of the perturbed game is given by

$$y^*_\delta = -Q^{-1}(xc + \delta u) = y^* - \delta Q^{-1} u$$

The first-order hypergradient approximation in Section 4 only matches equation 16 when Q is symmetric

$$\begin{aligned} v_\delta(x, y^*_\delta) &= \frac{1}{\delta}\sum_{i=1}^N \nabla_x g_i(x, y^*_{\delta,i}, y^*_{\neg i}) - \nabla_x g_i(x, y^*) \\ &= \frac{1}{\delta}\sum_{i=1}^N c_i^\top (y^*_{\delta,i} - y^*_i) \\ &= \frac{1}{\delta}\sum_{i=1}^N c_i^\top [\delta Q^{-1} u]_i \\ &= -c^\top Q^{-1} u \end{aligned}$$

Alternatively, using results in Section 5, we can construct a perturbed game with cost function $\bar{g}_i(x, y) = \phi_\epsilon(x, y) - g_i(x, y) + \delta f(x, y)$, where $\phi_\epsilon(x, y) = \frac{1}{\epsilon}\sum_{j=1}^N [g_j(x, y_j + \epsilon(y_j - y^*_j), y_{\neg j}) - g_j(x, y)]$ is a potential function. Its equilibrium $\bar{y}^*_\delta$ satisfies $\lim_{\delta\to 0} \bar{y}^*_\delta = y^*$ and $\frac{d\bar{y}^*_\delta}{d\delta}|_{\delta=0} = (Q + \epsilon D)^{-\top} u$, where $D = \text{diag}(Q_{11}, \cdots, Q_{NN})$ is a block-diagonal matrix. Then the first-order hypergradient approximation $v_\delta(x, \bar{y}^*_\delta)$ exactly matches the exact hypergradient in this toy example. The intuition of the construction is that the solution of $\bar{V}(x, y) = (\nabla_{y_1}\bar{g}_1(x, y), \cdots, \nabla_{y_N}\bar{g}_N(x, y)) = 0$ approximates the solution of $Q^\top(y - y^*) + \delta u = 0$ as $\epsilon \to 0$. Solving $\nabla_{y_i}\phi_\epsilon(x, \bar{y}^*_\delta) - \nabla_{y_i} g_i(x, \bar{y}^*_\delta) + \delta\nabla_{y_i} f(x, \bar{y}^*_\delta) = 0$, $\forall i \in [N]$:

$$\begin{aligned} &\nabla_{y_i}\phi_\epsilon(x, y) - \nabla_{y_i} g_i(x, y) + \delta\nabla_{y_i} f(x, y) \\ &= \frac{1}{\epsilon}\sum_{j=1}^N [\nabla_{y_i} g_j(x, y_j + \epsilon(y_j - y^*_j), y_{\neg j}) - \nabla_{y_i} g_j(x, y)] + \nabla_{y_i} g_i(x, y_i + \epsilon(y_i - y^*_i), y_{\neg i}) - \nabla_{y_i} g_i(x, y) + \delta\nabla_{y_i} f(x, y) \\ &= \frac{1}{\epsilon}[Q_{ii}(y_i + \epsilon(y_i - y^*_i)) + \sum_{j\neq i} Q_{ij} y_j - Q_{ii} y_i - \sum_{j\neq i} Q_{ij} y_j] \\ &+ \frac{1}{\epsilon}[\sum_{j\neq i} Q^\top_{ji}(y_j + \epsilon(y_j - y^*_j)) - \sum_{j\neq i} Q^\top_{ji} y_j] \\ &+ [Q_{ii}(y_i + \epsilon(y_i - y^*_i)) + \sum_{j\neq i} Q_{ij} y_j] \\ &- [Q_{ii} y_i + \sum_{j\neq i} Q_{ij} y_j] + \delta u_i \\ &= Q_{ii}(y_i - y^*_i) + \sum_{j\neq i} Q^\top_{ji}(y_j - y^*_j) + Q_{ii}\epsilon(y_i - y^*_i) = 0 \end{aligned}$$

Stacking these equations together, we have

$$\begin{aligned} Q^\top(y - y^*) + \epsilon D(y - y^*) + \delta u = 0 \\ y = y^* + (Q^\top + \epsilon D)^{-1}(-\delta u) \end{aligned}$$

We have achieved a desired formula for $\bar{y}^*_\delta$ in this toy example.

J Failure of Revelation Principle

To understand why non-truthful auction is important, we construct single-bidder and two-bidder settings where there exist non-truthful auctions with more revenue than any truthful auctions. Inspired by Example B.1. in Berzack et al. (2026), the key idea is to introduce valuation-dependent constraints over bids.

J.1 Single-Item Single-Bidder

Let the valuation be $t \sim \text{Uniform}[0,1]$. Fix $\alpha > 1$. $\{a \in \mathbb{R}_+ : a \le k(t)\}$ is the feasible set of the bid given valuation t, where

$$k(t) = \begin{cases} \alpha & \text{if } t < \frac{2}{3}, \\ 1 & \text{if } t \ge \frac{2}{3}. \end{cases}$$

Consider the following allocation $\rho : [0,1] \to \{0,1\}$ and payment rule $P : [0,1] \to \mathbb{R}$,

$$\rho(a) = \begin{cases} 1 & \text{if } a > 0, \\ 0 & \text{if } a = 0. \end{cases}$$

$$P(a) = \begin{cases} 0 & \text{if } a = 0 \\ \frac{2}{3} & \text{if } 0 < a \le 1, \\ \frac{1}{3} & \text{if } a > 1. \end{cases}$$

The utility of the bidder is $u(a,t) = t\rho(a) - P(a)$. We can now derive the best response of the bidder that maximizes the utility. When $t \le \frac{1}{3}$, bidding anything above zero would results in negative utility. The best response is to bid zero. If $\frac{1}{3} < t < \frac{2}{3}$, the constraint is $a \le \alpha$. The best response is to bid $a > 1$ and pay $\frac{1}{3}$. If $t \ge \frac{2}{3}$, the constraint is $a \le 1$. The best response is to bid any $0 < a \le 1$ and pay $\frac{2}{3}$. Hence, the expected revenue of this auction is

$$\begin{aligned} 0 + \mathbb{E}_{t \in [\frac{1}{3}, \frac{2}{3}]} \left[\frac{1}{3}\right] + \mathbb{E}_{t \ge \frac{2}{3}} \left[\frac{2}{3}\right] &= \frac{1}{3} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{2}{3} \\ &= \frac{1}{3} \end{aligned}$$

On the other hand, the optimal truthful auction is Myerson's optimal auction Myerson (1981) with an expected revenue of $\frac{1}{4}$. Hence, the non-truthful auction we propose in fact has a higher revenue than any truthful auction in this setting.

J.2 Single-Item Two-Bidder

Let the valuation of bidder i be $t_i \sim \text{Uniform}[0,1]$ sampled independently. Consider a singleton $B_1 = \{l_1 = 0\}$ and contiguous intervals $B_j = (l_j, m_j]$ for $j = 2,3,4$. $m_j = l_{j+1}$ for $j = 2,3$ and $m_j > l_j > l_1$ for $j = 2,3,4$. $\cup_{j=1}^4 B_j$ is the set of possible bids. Let $0 < c_1 < c_2, < c_3 < 1$. $\{a_i \in \mathbb{R}_+ : a_i \le k(t_i)\}$ is the feasible set of the ith bidder's bid, where

$$k(t) = \begin{cases} m_4 & \text{if } t < c_2, \\ m_3 & \text{if } c_2 \le t < c_3, \\ m_2 & \text{if } c_3 \le t \end{cases}$$

Consider the following allocation $\rho : [0,1] \times [0,1] \to \Delta^2$ and payment rule $P : [0,1] \times [0,1] \to \mathbb{R}^2$,

$$\rho_i(a_1, a_2) = \begin{cases} 1 & \text{if } a_i \ge a_{\neg i}, \\ 0 & \text{if } a_i < a_{\neg i} \text{ or } a_i = 0. \end{cases}$$

$$P_i(a_1, a_2) = \begin{cases} 0 & \text{if } a_i \in B_1 \\ c_3\rho_i(a) & \text{if } a_i \in B_2, \\ c_2\rho_i(a) & \text{if } a_i \in B_3, \\ c_1\rho_i(a) & \text{if } a_i \in B_4. \end{cases}$$

The utility of the bidder i given valuation t_i is $u_i(a, t_i) = t_i\rho_i(a) - P_i(a)$. Now we derive the Nash equilibrium. Consider player i. When $t_i < c_1$, The dominant strategy is to bid zero, because the utility will be negative even if bidder i wins. When $t_i \in [c_1, c_2)$, the constraint is $a_i \le m_4$, the dominant strategy is bidding $a_i \in B_4$ because it results in the least payment and maximizes the chance of winning. Following this logic, a Nash equilibrium bidding strategy is

$$\beta(t_i) = \begin{cases} 0 & \text{if } t < c_1 \\ \text{any } a_i \in B_4 & \text{if } t \in [c_1, c_2), \\ \text{any } a_i \in B_3 & \text{if } t \in [c_2, c_3), \\ \text{any } a_i \in B_2 & \text{if } t \in [c_3, 1). \end{cases}$$

Denote $\mathbb{P}_1 = \Pr(t_i < c_1)$, $\mathbb{P}_2 = \Pr(t_i \in [c_1, c_2))$, $\mathbb{P}_3 = \Pr(t_i \in [c_2, c_3))$, $\mathbb{P}_4 = \Pr(t_i \in [c_3, c_4))$. The expected revenue can be computed by the following Python program:

```
def two_bidder_four_interval(C):
    # C is a list containing c1, c2, c3
    C = numpy.array(C)
    P = (C[1:]-C[:-1])
    P = numpy.concatenate([ [C[0]], P, [(1-C[-1])] ])

    C_dim = len(C)
    P_dim = len(P)

    c1, c2, c3 = C[0], C[1], C[2]
    payment_matrix = np.array([
        [0, c1, c2, c3],
        [c1, c1, c1, c1],
        [c2, c1, c2, c2],
        [c3, c1, c2, c3]
    ])

    revenue = 0
    for i in range(P_dim):
        for j in range(P_dim):
            revenue += payment_matrix[i, j] * P[i] * P[j]

    return revenue
```

Substitute $c_1 = 0.4$, $c_2 = 0.6$, $c_3 = 0.8$, $B_2 = (0, 1]$, $B_3 = (1, 1.5]$, $B_4 = (1.5, 2]$. The expected revenue is 0.472. Myerson's optimal auction is the optimal truthful auction and has an expected revenue $\frac{5}{12} = 0.417$. The proposed non-truthful auction has a revenue higher than any truthful auction.

K Experiments

Theorem 6.1. *Suppose $\mathcal{G}(x)$ is locally strongly monotone on a ball $\bar{B}_r(y^*(x))$ with radius $r > 0$ for all $y^*(x) \in$ LNE$(\mathcal{G}(x))$ and $x \in X$. Given initial iterate y, let $S(x, y, \hat{\delta}) : X \times \mathbb{R}^n \times \mathbb{R}_+ \to Y(x)$ be a solver can always return $\hat{y}$ such that $\|\hat{y} - y^*(x)\| \le \hat{\delta}$ whenever $\|y - y^*(x)\| \le r$ and $y^*(x) \in$ LNE$(\mathcal{G}(x))$.*

In Algorithm 1, suppose $\|y_0 - y^*(x_0)\| \le \frac{r}{2}$ for some $y^*(x_0) \in$ LNE$(\mathcal{G}(x_0))$ and the learning rate η is sufficiently small, then $y_k = z^*(x_k)$, $\forall k \in [T]$ for some continuously differentiable function $z^* : X \to Y(x)$ such that $z^*(x) \in$ LNE$(\mathcal{G}(x))$.

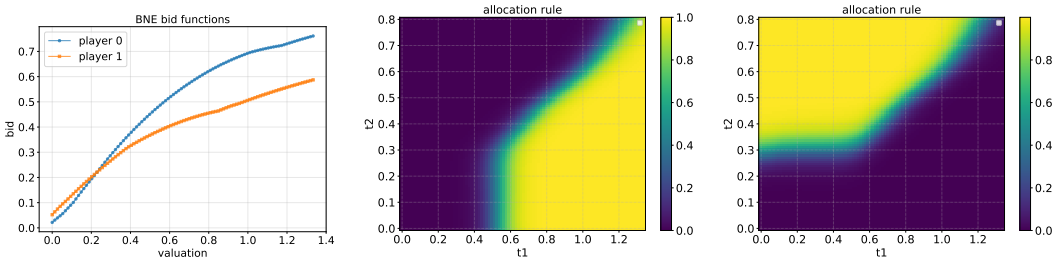
Proof. Let $y^*(x_k) \in \text{LNE}(\mathcal{G}(x_k))$. Since the KKT matrix $K(x_k, y^*(x_k))$ is invertible, by implicit function theorem, there exists a neighborhood $N_{x_k} \times N_{y^*(x_k)}$ of $(x_k, y^*(x_k))$ and a continuously differentiable and L_k-Lipschitz continuous function $z_k^* : N_{x_k} \to N_{y^*(x_k)}$ such that $z_k^*(x) \in \text{LNE}(\mathcal{G}(x))$, $\forall x \in N_{x_k}$, and $z_k^*(x_k) = y^*(x_k)$. Suppose $\|x_k - x_{k+1}\| \le \mathcal{O}(\eta)$ is sufficiently small such that $x_{k+1} \in N_{x_k}$ and $\|x_k - x_{k+1}\| \le \frac{r}{2L_k}$, then $\|z_k^*(x_k) - z_k^*(x_{k+1})\| \le \frac{r}{2}$.

Suppose $\|y_k - y^*(x_k)\| \leq \frac{r}{2}$, $\|y_k - z_k^*(x_{k+1})\| \leq r$ by triangle inequality. For arbitrary precision $\hat{\delta}$, the solver outputs $y_{k+1} = S(x_{k+1}, y_k, \hat{\delta})$ such that $\|y_{k+1} - z_k^*(x_{k+1})\| \leq \hat{\delta}$. By induction, the claim is proved. □

K.1 Experimental Details

Each bidding strategy β_{y_i} is two-layer fully connected neural network of width 10 with SeLU activation (Klambauer et al. (2017)) and ReLU activation (Agarap (2018)) for the output layer. The allocation rule is a two-layer fully connected neural network of width 10 with SeLU activation and Layer Normalization (Ba et al. (2016)). We use Adam (Kingma & Ba (2014)) to optimize the parameters of the auction mechanism. We choose the hyperobjective step size $\eta = 10^{-5}$ for single-item auction and $\eta = 10^{-4}$ for multi-item auction. For single-item auction, we run gradient descent with step size 10^{-4} for 30000 iterations to solve the original game, and with step size 10^{-4} for 1000 iterations to solve the perturbed game. For multi-item auction, we run gradient descent with step size 10^{-4} for 40000 iterations to solve the original game, and with step size 10^{-4} for 2000 iterations to solve the perturbed game.

We initialize the allocation rule of the mechanism to mimick first-price auction so that the initial lower-level game is easy to solve. Let $p \in \Delta^{N+1}$ be the output of the allocation rule of first-price auction. Let $q \in \Delta^{N+1}$ be the output of the allocation rule of the neural-network-parameterized allocation rule. Then, we run Adam with step size 10^{-4} for 10^6 iterations to minimize their KL divergence $D_{\mathrm{KL}}(p||q) \propto \sum_i p_i \log q_i$.



(a) bidding strategy (b) allocation for bidder 1 (c) allocation for bidder 2

Figure 5: Equilibrium bidding strategy and allocation rule for asymmetric single-item auction. For allocation probability, x-axis and y-axis are the bid of bidder 1 and bidder 2 respectively.

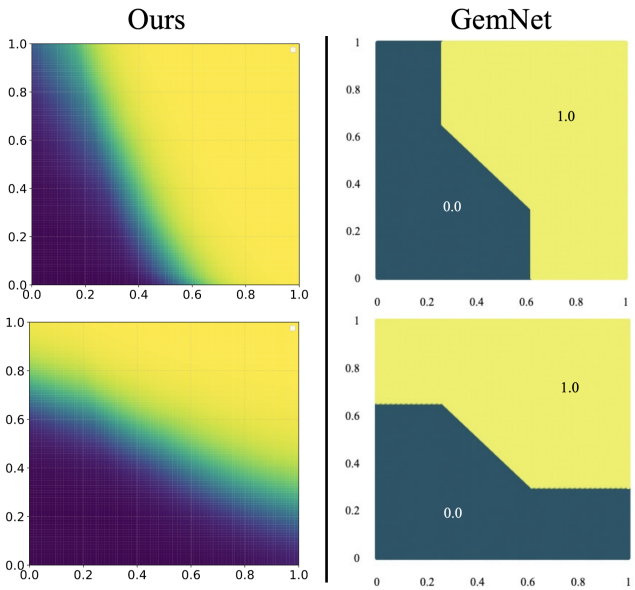


Figure 6: Comparison between non-truthful auction discovered by our method and truthful auction discovered by GemNet. For each plot, x-axis and y-axis is the bid of bidder 1 for item 1 and item 2 respectively. The top row and bottom row show the probability of item 1 and item 2 being allocated to bidder 1 respectively, given the bid of bidder 2 is (0,0.6).