A Omitted Proofs Section 2

Proposition A.1. *Given any convex-concave min-max game with dependent strategy sets* $(X, Y, f, \boldsymbol{g})$, *a Stackelberg equilibrium always exists.*

Proof of Proposition A.1. By Berge's maximum theorem [5], the outer player's value function $V(\boldsymbol{x}) = \max_{\boldsymbol{y} \in Y : \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}) \geq \boldsymbol{0}} f(\boldsymbol{x}, \boldsymbol{y})$ is continuous, and the inner solution correspondence $Y^*(\boldsymbol{x}) = \arg\max_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} f(\boldsymbol{x}, \boldsymbol{y})$ is non-empty, for all $\boldsymbol{x} \in \mathcal{X}$. Since V is continuous and X is compact and non-empty, by the extreme value theorem [50], there exists a minimizer $\boldsymbol{x}^*$ of V. Hence $(\boldsymbol{x}^*, \boldsymbol{y}^*(\boldsymbol{x}^*))$ where $\boldsymbol{y}^*(\boldsymbol{x}^*) \in \mathcal{Y}^*(\boldsymbol{x}^*)$ is well-defined and is a Stackelberg equilibrium of $(X, Y, f, \boldsymbol{g})$. □

B Envelope Theorem

Danskin's theorem [15] offers insights into optimization problems of the form:

$$\max_{\boldsymbol{y} \in Y} f(\boldsymbol{x}, \boldsymbol{y}) \ , \tag{4}$$

where $Y \subset \mathbb{R}^m$ is compact and non-empty. Among other things, Danskin's theorem allows us to compute the gradient of the objective function of this optimization problem with respect to $\boldsymbol{x}$.

Theorem B.1 (Danskin's Theorem)**.** *Consider Equation (4). Suppose that Y is convex and that f is concave in $\boldsymbol{y}$. Let $V(\boldsymbol{x}) = \max_{\boldsymbol{y} \in Y} f(\boldsymbol{x}, \boldsymbol{y})$ and $Y^*(\boldsymbol{x}) = \arg\max_{\boldsymbol{y} \in Y} f(\boldsymbol{x}, \boldsymbol{y})$. Then, V is differentiable at $\widehat{\boldsymbol{x}}$ if $Y^*(\widehat{\boldsymbol{x}})$ is a singleton. Additionally, the gradient at $\widehat{\boldsymbol{x}}$ is given by $V'(\widehat{\boldsymbol{x}}) = \nabla_{\boldsymbol{x}} f(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}}))$, where $\boldsymbol{y}^*(\widehat{\boldsymbol{x}}) \in Y^*(\widehat{\boldsymbol{x}})$.*

Unfortunately, Danskin's theorem does not hold when Y is replaced by even a non-empty compact-valued correspondence $\mathcal{Y} : X \rightrightarrows Y$, in which case the inner problem becomes $\max_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} f(\boldsymbol{x}, \boldsymbol{y})$.

Example B.2 (Danskin's theorem does not apply to min-max games with dependent strategy sets)**.** *Consider the optimization problem:*

$$\max_{y \in \mathbb{R} : y + x \geq 0} -y^2 + y + 2x + 2 \tag{5}$$

The solution function $y^*(\boldsymbol{x}) = \arg\max_{y \in \mathbb{R} : y + x \geq 0} -y^2 + y + 2x + 2$ for this problem is well defined since the solution is singleton-valued and is given by:

$$y^*(x) = \begin{cases} 1/2 & \text{if } x \geq -1/2 \\ -x & \text{if } x < -1/2 \end{cases} \tag{6}$$

The value function $V(x) = \max_{y \in \mathbb{R} : y + x \geq 0} -y^2 + y + 2x + 2$ is given by:

$$V(x) = f(x, y^*(x)) \tag{7}$$

$$= -y^*(x)^2 + y^*(x) + 2x + 2 \tag{8}$$

$$= \begin{cases} -1/4 + 1/2 + 2x + 2 & \text{if } x \geq -1/2 \\ -x^2 - x + 2x + 2 & \text{if } x < -1/2 \end{cases} \tag{9}$$

$$= \begin{cases} 9/4 + 2x & \text{if } x \geq -1/2 \\ -x^2 + x + 2 & \text{if } x < -1/2 \end{cases} \tag{10}$$

The derivative of the value function is given by:

$$\frac{\partial V}{\partial x} = \begin{cases} 2 & \text{if } x \geq -1/2 \\ 1 - 2x & \text{if } x < -1/2 \end{cases} \tag{11}$$

However, the derivative predicted by Danskin's theorem is 2, for all x. Hence, Danskin's theorem does not hold when the constraints are parameterized, i.e., when the problem is of the form $\min_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} f(\boldsymbol{x}, \boldsymbol{y})$ rather than $\min_{\boldsymbol{y} \in Y} f(\boldsymbol{x}, \boldsymbol{y})$ where $X \subset \mathbb{R}^n$, $Y \subset \mathbb{R}^m$, and $\mathcal{Y} : X \rightrightarrows Y$.

N.B. *For simplicity, we do not assume the constraint set is compact in this example; however, the conclusion still applies, since compactness of the constraint set is used to guarantee existence of a solution for all $\boldsymbol{x}$, but as a solution to this particular problem exists we can do away with the assumption.*

An answer to Dankin's theorem not holding when the constraints are parameterized can be found in the mathematical economics literature. In particular the following theorem due to Milgrom and Segal [41] generalizes Danskin's theorem (Theorem B.1).

Theorem B.3 (Envelope Theorem [41]). *Consider the maximization problem*

$$V(\boldsymbol{x}) = \max_{\boldsymbol{y} \in \mathbb{R}^m} f(\boldsymbol{x}, \boldsymbol{y}) \text{ subject to } g_k(\boldsymbol{x}, \boldsymbol{y}) \geq 0 \text{ for all } k = 1, \dots, K \ . \tag{12}$$

Define the solution correspondence $Y^*(\boldsymbol{x}) = \arg\max_{\boldsymbol{y} \in \mathbb{R}^m : \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y}) \geq \boldsymbol{0}} f(\boldsymbol{x}, \boldsymbol{y})$. Now suppose that Assumption 3.1 holds. Then, the value function V is absolutely continuous, and at any point $\widehat{\boldsymbol{x}}$ where it is differentiable:

$$\nabla_{\boldsymbol{x}} V(\widehat{\boldsymbol{x}}) = \nabla_{\boldsymbol{x}} L(\boldsymbol{y}^*(\widehat{\boldsymbol{x}}), \boldsymbol{\lambda}(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}}))), \widehat{\boldsymbol{x}}) = \nabla_{\boldsymbol{x}} f\left(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})\right) + \sum_{k=1}^{K} \lambda_k(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})) \nabla_{\boldsymbol{x}} g_k\left(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})\right) \ , \tag{13}$$

where $\boldsymbol{\lambda}(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})) = (\lambda_1(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})), \dots, \lambda_K(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})))^T \in \Lambda(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}}))$ are the Lagrange multipliers associated associated with $\boldsymbol{y}^*(\widehat{\boldsymbol{x}}) \in Y^*(\widehat{\boldsymbol{x}})$.

C Omitted Proofs Section 3

Proof of Theorem 3.2. Let $V(\boldsymbol{x}) = \max_{\boldsymbol{y} \in Y : g(\boldsymbol{x}, \boldsymbol{y}) \geq 0} f(\boldsymbol{x}, \boldsymbol{y})$. Reformulating the problem as a Lagrangian saddle point problem, for all, $\widehat{\boldsymbol{x}} \in X$, it holds that:

$$\begin{aligned} V(\widehat{\boldsymbol{x}}) &= \max_{\boldsymbol{y} \in Y : g(\widehat{\boldsymbol{x}}, \boldsymbol{y}) \geq 0} f(\widehat{\boldsymbol{x}}, \boldsymbol{y}) && (14) \\ &= \max_{\boldsymbol{y} \in Y} \min_{\boldsymbol{\lambda} \in \mathbb{R}^K_{++}} \left\{ f(\widehat{\boldsymbol{x}}, \boldsymbol{y}) + \sum_{k=1}^{K} \lambda_k g_k(\widehat{\boldsymbol{x}}, \boldsymbol{y}) \right\} && (15) \end{aligned}$$

Since an interior point exists by the assumptions, the Karush-Kuhn-Tucker Theorem [36] applies, so for all $\widehat{\boldsymbol{x}} \in X$, there exists $\boldsymbol{\lambda} \in \mathbb{R}^K$ that solves the above optimization problem Equation (15).

Let $Y^*(\boldsymbol{x}) = \arg\max_{\boldsymbol{y} \in Y : g(\boldsymbol{x}, \boldsymbol{y}) \geq 0} f(\boldsymbol{x}, \boldsymbol{y})$ and $\Lambda(\boldsymbol{x}, \boldsymbol{y}) = \arg\min_{\boldsymbol{\lambda} \in \mathbb{R}^K_{++}} \left\{ f(\boldsymbol{x}, \boldsymbol{y}) + \sum_{k=1}^{K} \lambda_k g_k(\boldsymbol{x}, \boldsymbol{y}) \right\}$. We can then re-express the value function as:

$$V(\widehat{\boldsymbol{x}}) = f(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})) + \sum_{k=1}^{K} \lambda_k(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})) g_k(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})), \quad \forall \boldsymbol{y}^*(\widehat{\boldsymbol{x}}) \in Y^*(\widehat{\boldsymbol{x}}), \lambda_k(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})) \in \Lambda(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})) \ .$$

Alternatively, we can take the maximum over $\boldsymbol{\lambda}$'s and $\boldsymbol{y}$'s to obtain:

$$V(\widehat{\boldsymbol{x}}) = \max_{\boldsymbol{y}^*(\widehat{\boldsymbol{x}}) \in Y^*(\widehat{\boldsymbol{x}})} \max_{\lambda_k(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})) \in \Lambda(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}}))} \left\{ f(\widehat{\boldsymbol{x}}, \boldsymbol{y}) + \sum_{k=1}^{K} \lambda_k(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})) g_k(\widehat{\boldsymbol{x}}, \boldsymbol{y}) \right\} \ .$$

Note that for fixed $\boldsymbol{y}^*(\widehat{\boldsymbol{x}}) \in Y^*(\widehat{\boldsymbol{x}})$ and corresponding $\lambda_k(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})) \in \Lambda(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}}))$, $f(\widehat{\boldsymbol{x}}, \boldsymbol{y}) + \sum_{k=1}^{K} \lambda_k(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})) g_k(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}}))$ is differentiable, since $f, g_1, \dots, g_K$ are differentiable.

Additionally, recall the pointwise maximum subdifferential property, i.e., if $f(\boldsymbol{x}) = \max_{\alpha \in \mathcal{A}} f_\alpha(\boldsymbol{x})$ for a family of functions $\{f_\alpha\}_{\alpha \in \mathcal{A}}$, then $\partial_{\boldsymbol{x}} f(\boldsymbol{a}) = \operatorname{conv}\left(\bigcup_{\alpha \in \mathcal{A}} \{\partial_{\boldsymbol{x}} f_{\alpha \in \mathcal{A}}(\boldsymbol{a}) \mid f_\alpha(\boldsymbol{a}) = f(\boldsymbol{x})\}\right)$

(see, for example, [7]), which then gives:

$$\partial_{\boldsymbol{x}} V(\widehat{\boldsymbol{x}}) = \partial_{\boldsymbol{x}} \left(\max_{\boldsymbol{y}^*(\widehat{\boldsymbol{x}}) \in Y^*(\widehat{\boldsymbol{x}})} \max_{\lambda_k(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})) \in \Lambda(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}}))} \left\{ f(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})) + \sum_{k=1}^{K} \lambda_k(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})) g_k(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})) \right\} \right) \tag{16}$$

$$= \operatorname{conv} \left(\bigcup_{\boldsymbol{y}^*(\widehat{\boldsymbol{x}}) \in Y^*(\widehat{\boldsymbol{x}})} \bigcup_{\lambda_k(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})) \in \Lambda(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}}))} \partial_{\boldsymbol{x}} \left\{ f(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})) + \sum_{k=1}^{K} \lambda_k(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})) g_k(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})) \right\} \right) \tag{17}$$

$$= \operatorname{conv} \left(\bigcup_{\boldsymbol{y}^*(\widehat{\boldsymbol{x}}) \in Y^*(\widehat{\boldsymbol{x}})} \bigcup_{\lambda_k(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})) \in \Lambda(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}}))} \left\{ \nabla_{\boldsymbol{x}} f\left(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})\right) + \sum_{k=1}^{K} \lambda_k(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})) \nabla_{\boldsymbol{x}} g_k\left(\widehat{\boldsymbol{x}}, \boldsymbol{y}^*(\widehat{\boldsymbol{x}})\right) \right\} \right) . \tag{18}$$

□

D Algorithms

The algorithms studied in our paper and described in Section 3 are presented below. We note that Π_Y is the projection operator on the set Y which is defined as $\Pi_Y(\boldsymbol{y}) = \arg\min_{\boldsymbol{z} \in Y} \|\boldsymbol{y} - \boldsymbol{z}\|_2$.

Algorithm 1 Max-Oracle Gradient Descent

Inputs: $X, Y, f, \boldsymbol{g}, \eta, T, \boldsymbol{x}^{(0)}$
Output: $(\boldsymbol{x}^\star, \boldsymbol{y}^\star)$

```
for t = 1, ..., T do
    Find ŷ ∈ Y such that f(x^(t−1), ŷ) ≥ max_{y∈Y: g(x^(t−1), y) ≥ 0} f(x^(t−1), y) − δ and g(x^(t−1), ŷ) ≥ 0
    Set y^(t−1) = ŷ
    Set λ^(t−1) = λ(x^(t−1), y^(t−1))
    Set x^(t) = Π_X ( x^(t−1) − η_t [ ∇_x f(x^(t−1), y^(t−1)) + Σ_{k=1}^K λ_k^(t−1) ∇_x g_k(x^(t−1), y^(t−1)) ] )
end for
Find ŷ ∈ Y such that f(x^(T), ŷ) ≥ max_{y∈Y: g(x^(T), y) ≥ 0} f(x^(T), y) − δ and g(x^(T), ŷ) ≥ 0
y^(T) = ŷ
return (x^(T), y^(T))
```

Algorithm 2 Nested Gradient Descent

Inputs: $X, Y, f, \boldsymbol{g}, \eta_{\boldsymbol{x}}, \eta_{\boldsymbol{y}}, T_{\boldsymbol{x}}, T_{\boldsymbol{y}}, \boldsymbol{x}^{(0)}, \boldsymbol{y}^{(0)}$
Output: $\boldsymbol{x}^\star, \boldsymbol{y}^\star$

```
for t = 1, ..., T_x do
    y^(t−1) = y^(0)
    for s = 1, ..., T_y do
        y^(t−1) = Π_{y∈Y: g(x^(t−1), y)} ( y^(t−1) + η_{s_y} ∇_y f(x^(t−1), y^(t−1))] )
    end for
    Set λ^(t−1) = λ(x^(t−1), y^(t−1))
    Set x^(t) = Π_X ( x^(t−1) − η_{t_x} [ ∇_x f(x^(t−1), y^(t−1)) + Σ_{k=1}^K λ_k^(t−1) ∇_x g_k(x^(t−1), y^(t−1)) ] )
end for
y^(T) = y^(0)
for s = 1, ..., T_y do
    y^(T) = Π_{y∈Y: g(x^(T), y)} ( y^(T) + η_{s_y} ∇_y f(x^(T), y^(T))] )
end for
return (x^(T), y^(T))
```

D.1 Omitted Proofs Section 3

Lemma D.1 (Lipschitz Objective, Lipschitz Value Function). *Let $f : X \times Y$ be a continuous function, where $X \subset \mathbb{R}^n$, $Y \subset \mathbb{R}^m$. Suppose that $\nabla_{\boldsymbol{x}} f$ is continuous in $(\boldsymbol{x}, \boldsymbol{y})$, X is compact and non-empty, and $\mathcal{Y} : X \rightrightarrows Y$ is nonempty-compact-valued correspondence, then $V(\boldsymbol{x}) = \max_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x})} f(\boldsymbol{x}, \boldsymbol{y})$ is ℓ_f-Lipschitz continuous, i.e., $\|V(\boldsymbol{x}_1) - V(\boldsymbol{x}_2)\| \leq \ell_f \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$, with $\ell_f = \max_{(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}) \in X \times Y} \|\nabla_{\boldsymbol{x}} f(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}})\|$.*

Proof of Lemma D.1. Let $\ell_f = \max_{(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}) \in X \times Y)} \|\nabla_{\boldsymbol{x}} f(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}})\|$. Clearly, we have $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in X, \boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}_1) \cap \mathcal{Y}(\boldsymbol{x}_2), \|f(\boldsymbol{x}_1, \boldsymbol{y}) - f(\boldsymbol{x}_2, \boldsymbol{y})\| \leq \ell_f \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|$.

Fix $\boldsymbol{x}_1, \boldsymbol{x}_2 \in X$. Then, for all $\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}_1) \cap \mathcal{Y}(\boldsymbol{x}_2)$, we have:

$$
\begin{aligned}
f(\boldsymbol{x}_1, \boldsymbol{y}) &\leq f(\boldsymbol{x}_1, \boldsymbol{y}) - f(\boldsymbol{x}_2, \boldsymbol{y}) + f(\boldsymbol{x}_2, \boldsymbol{y}) && (19) \\
&\leq \ell_f \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| + f(\boldsymbol{x}_2, \boldsymbol{y}) && (20)
\end{aligned}
$$

Taking the max over the $\boldsymbol{y}$'s on both sides (which is guaranteed to exist by the continuity of f, and compactness and non-emptyness of $\mathcal{Y}$), we obtain:

$$
\begin{aligned}
\max_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}_1)} f(\boldsymbol{x}_1, \boldsymbol{y}) &\leq \ell_f \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| + \max_{\boldsymbol{y} \in \mathcal{Y}(\boldsymbol{x}_2)} f(\boldsymbol{x}_2, \boldsymbol{y}) && (21) \\
V(\boldsymbol{x}_1) &\leq \ell_f \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| + V(\boldsymbol{x}_2) && (22) \\
V(\boldsymbol{x}_1) - V(\boldsymbol{x}_2) &\leq \ell_f \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| && (23)
\end{aligned}
$$

Since this inequality holds for arbitrary $\boldsymbol{x}_1, \boldsymbol{x}_2 \in X$, we also have:

$$
V(\boldsymbol{x}_2) - V(\boldsymbol{x}_1) \leq \ell_f \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \tag{24}
$$

Combining the two inequalities, we obtain

$$
\|V(\boldsymbol{x}_1) - V(\boldsymbol{x}_2)\| \leq \ell_f \|\boldsymbol{x}_1 - \boldsymbol{x}_2\| \tag{25}
$$

□

Proof of Theorem 3.3. Note that by Theorem 3.2, we have $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t-1)}) + \sum_{k=1}^{K} \lambda_k^{(t-1)} \nabla_{\boldsymbol{x}} g_k(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t-1)}) \in \partial_{\boldsymbol{x}} V(\boldsymbol{x}^{(t-1)}) = \partial_{\boldsymbol{x}} \max_{\boldsymbol{y} \in Y : \boldsymbol{g}(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}) \geq 0} f(\boldsymbol{x}^{(t-1)}, \boldsymbol{y})$. For notational clarity, let $\boldsymbol{g}(t-1) = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t-1)}) + \sum_{k=1}^{K} \lambda_k^{(t-1)} \nabla_{\boldsymbol{x}} g_k(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t-1)})$. Suppose that $\boldsymbol{x}^{\star} \in \arg\min_{\boldsymbol{x} \in X} \max_{\boldsymbol{y} \in Y : \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{y})} f(\boldsymbol{x}, \boldsymbol{y})$

$$
\begin{aligned}
\left\|\boldsymbol{x}^{(T)} - \boldsymbol{x}^*\right\|^2 &= \left\|\Pi_X\left(\boldsymbol{x}^{(T-1)} - \eta_T \boldsymbol{g}(T-1)\right) - \Pi_X\left(\boldsymbol{x}^*\right)\right\|^2 && (26) \\
&\leq \left\|\boldsymbol{x}^{(T-1)} - \eta_T \boldsymbol{g}(T-1) - \boldsymbol{x}^*\right\|^2 && (27) \\
&= \left\|\boldsymbol{x}^{(T-1)} - \boldsymbol{x}^*\right\|^2 - 2\eta_T \left\langle \boldsymbol{g}(T-1), \left(\boldsymbol{x}^{(T-1)} - \boldsymbol{x}^*\right)\right\rangle + \eta_T^2 \|\boldsymbol{g}(T-1)\|^2 && (28) \\
&\leq \left\|\boldsymbol{x}^{(T-1)} - \boldsymbol{x}^*\right\|^2 - 2\eta_T \left(f(\boldsymbol{x}^{(T-1)}, \boldsymbol{y}^{(T-1)}) - f(\boldsymbol{x}^*, \boldsymbol{y}^{(T-1)})\right) + \eta_T^2 \|\boldsymbol{g}(T-1)\|^2 && (29)
\end{aligned}
$$

where the first line follows from definitions, the second from the non-expansiveness of the projection operator, the third from algebra, the fourth from the definition of subgradients, i.e., $\boldsymbol{g}(t-1)^T \left(\boldsymbol{x}^{(t-1)} - \boldsymbol{x}^*\right) \geq f(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t-1)}) - f(\boldsymbol{x}^*, \boldsymbol{y}^{(t-1)})$. Applying the inequality above recursively, we obtain:

$$
\left\|\boldsymbol{x}^{(T)} - \boldsymbol{x}^*\right\|^2 \leq \left\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\right\|^2 - \sum_{t=1}^{T} 2\eta_t \left(f(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t-1)}) - f(\boldsymbol{x}^*, \boldsymbol{y}^{(t-1)})\right) + \sum_{t=1}^{T} \eta_t^2 \|\boldsymbol{g}(k-1)\|^2 \tag{30}
$$

Since $\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\| \geq 0$, we have:

$$2\sum_{t=1}^{T} \eta_t \left(f(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t-1)}) - f(\boldsymbol{x}^*, \boldsymbol{y}^{(t-1)}) \right) \leq \left\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\right\|^2 + \sum_{t=1}^{T} \eta_t^2 \left\|\boldsymbol{g}(t-1)\right\|^2 \tag{31}$$

Let $(\boldsymbol{x}^{(t)}_{\text{best}}, \boldsymbol{y}^{(t)}_{\text{best}}) = \arg\min_{(\boldsymbol{x}^{(k)}, \boldsymbol{y}^{(k)}):k\in[t]} f(\boldsymbol{x}^{(k)}, \boldsymbol{y}^{(k)})$, then we have we have:

$$\sum_{t=1}^{T} \eta_t \left(f(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t-1)}) - f(\boldsymbol{x}^*, \boldsymbol{y}^{(t-1)}) \right) \geq \sum_{t=1}^{T} \eta_t \left(f(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t-1)}) - \max_{\boldsymbol{y}\in Y:\boldsymbol{g}(\boldsymbol{x}^*,\boldsymbol{y})\geq \boldsymbol{0}} f(\boldsymbol{x}^*, \boldsymbol{y}) \right) \tag{32}$$

$$\geq \left(\sum_{t=1}^{T} \eta_t \right) \min_{t\in[T]} \left(f(\boldsymbol{x}^{(t-1)}, \boldsymbol{y}^{(t-1)}) - \max_{\boldsymbol{y}\in Y:\boldsymbol{g}(\boldsymbol{x}^*,\boldsymbol{y})\geq 0} f(\boldsymbol{x}^*, \boldsymbol{y}) \right) \tag{33}$$

$$= \left(\sum_{t=1}^{T} \eta_t \right) \left(f(\boldsymbol{x}^{(T)}_{\text{best}}, \boldsymbol{x}^{(T)}_{\text{best}}) - \max_{\boldsymbol{y}\in Y:\boldsymbol{g}(\boldsymbol{x}^*,\boldsymbol{y})\geq 0} f(\boldsymbol{x}^*, \boldsymbol{y}) \right) \tag{34}$$

Hence, we get the following bound:

$$f(\boldsymbol{x}^{(T)}_{\text{best}}, \boldsymbol{y}^{(T)}_{\text{best}}) - \max_{\boldsymbol{y}\in Y:\boldsymbol{g}(\boldsymbol{x}^*,\boldsymbol{y})\geq 0} f(\boldsymbol{x}^*, \boldsymbol{y}) \leq \frac{\left\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\right\|^2 + \sum_{t=1}^{T} \eta_t^2 ||\boldsymbol{g}(t-1)||^2}{2\left(\sum_{t=1}^{T} \eta_t\right)} \tag{35}$$

Since f is ℓ_f-Lipschitz with $\ell_f = \max_{(\widehat{\boldsymbol{x}},\widehat{\boldsymbol{y}})\in X\times Y} \|\nabla_{\boldsymbol{x}} f(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}})\|$, then for all $k \in \mathbb{N}$ we know that $\|\boldsymbol{g}(k-1)\| \leq \ell_f$.

$$f(\boldsymbol{x}^{(T)}_{\text{best}}, \boldsymbol{x}^{(T)}_{\text{best}}) - \max_{\boldsymbol{y}\in Y:\boldsymbol{g}(\boldsymbol{x}^*,\boldsymbol{y})\geq 0} f(\boldsymbol{x}^*, \boldsymbol{y}) \leq \frac{\left\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\right\|^2 + \ell_f^2 \sum_{t=1}^{T} \eta_t^2}{2\left(\sum_{t=1}^{T} \eta_t\right)} \tag{36}$$

$$f(\boldsymbol{x}^{(T)}_{\text{best}}, \boldsymbol{x}^{(T)}_{\text{best}}) - \min_{\boldsymbol{x}\in X} \max_{\boldsymbol{y}\in Y:\boldsymbol{g}(\boldsymbol{x},\boldsymbol{y})\geq 0} f(\boldsymbol{x}, \boldsymbol{y}) \leq \frac{\left\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\right\|^2 + \ell_f^2 \sum_{t=1}^{T} \eta_t^2}{2\left(\sum_{t=1}^{T} \eta_t\right)} \tag{37}$$

Under the assumption of the theorem:

$$\sum_{k=1}^{T} \eta_k^2 \leq \infty \qquad\qquad \sum_{k=1}^{T} \eta_k = \infty \tag{38}$$

as $t \to \infty$, $\lim_{k\to\infty} f(\boldsymbol{x}^{(k)}_{\text{best}}, \boldsymbol{y}^{(k)}) \leq \min_{\boldsymbol{x}\in X} \max_{\boldsymbol{y}\in Y:\boldsymbol{g}(\boldsymbol{x},\boldsymbol{y})\geq 0} f(\boldsymbol{x}, \boldsymbol{y})$, and since for all $k \in \mathbb{N}$, $\boldsymbol{y}^{(k)}_{\text{best}}$ satisfies $f(\boldsymbol{x}^{(k)}_{\text{best}}, \boldsymbol{y}^{(k)}_{\text{best}}) \geq \max_{\boldsymbol{y}\in Y:\boldsymbol{g}(\boldsymbol{x},\boldsymbol{y})\geq 0} f(\boldsymbol{x}^{(k)}_{\text{best}}, \boldsymbol{y}) - \delta$, as the number of iterations increases, the best iterate converges to a $(0, \delta)$-Stackelberg equilibrium. Additionally, setting $\eta_t = \frac{\left\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\right\|}{\ell_f \sqrt{T}}$ for all $t = 1, \ldots, T$, we get:

$$f(\boldsymbol{x}^{(T)}_{\text{best}}, \boldsymbol{x}^{(T)}_{\text{best}}) - \min_{\boldsymbol{x}\in X} \max_{\boldsymbol{y}\in Y:\boldsymbol{g}(\boldsymbol{x},\boldsymbol{y})\geq 0} f(\boldsymbol{x}^*, \boldsymbol{y}) \leq \frac{\ell_f \left\|\boldsymbol{x}^{(0)} - \boldsymbol{x}^*\right\|^2}{\sqrt{T}} \tag{39}$$

Hence, the best iterate converges to a (ε, δ)-Stackelberg equilibrium in $O(\varepsilon^{-2})$ iterations. □

Theorem D.2. *Suppose that Algorithm 1 is run on a convex-concave min-max game with dependent strategy sets given by $(X, Y, f, \boldsymbol{g})$ where X is convex. Suppose that Assumption 3.1 holds and that additionally f is μ-strongly convex in $\boldsymbol{x}$, i.e., $\forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in X, \boldsymbol{y} \in Y, f(\boldsymbol{x}_1, \boldsymbol{y}) \geq f(\boldsymbol{x}_2, \boldsymbol{y}) + \langle \boldsymbol{g}, (\boldsymbol{x}_1 - \boldsymbol{x}_2)\rangle + \frac{\mu}{2} \left\|\boldsymbol{x}_1 - \boldsymbol{x}_2\right\|^2$ where $\boldsymbol{g} \in \partial_{\boldsymbol{x}} f(\boldsymbol{x}_2, \boldsymbol{y})$. Then, if $(\boldsymbol{x}^{(t)}_{\text{best}}, \boldsymbol{y}^{(t)}_{\text{best}}) \in$*

$\arg\min_{(\boldsymbol{x}^{(k)},\boldsymbol{y}^{(k)}):k\in[t]} f(\boldsymbol{x}^{(k)},\boldsymbol{y}^{(k)})$, *for* $\varepsilon \in (0,1)$, *and* $\eta_t = \frac{2}{\mu(t+1)}$, *if we choose* T *large enough such that:*

$$T \geq N_T(\varepsilon) \doteq O(\varepsilon^{-1})$$

then there exists an iteration $T^\star \leq T$ *such that* $(\boldsymbol{x}_{\text{best}}^{(T^\star)}, \boldsymbol{y}_{\text{best}}^{(T^\star)})$ *is an* (ε,δ)*-Stackelberg equilibrium.*

Proof of Theorem D.2. Note that by Theorem 3.2, we have $\nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)}) + \sum_{k=1}^K \lambda_k^{(t-1)} \nabla_{\boldsymbol{x}} g_k(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)}) \in \partial_{\boldsymbol{x}} V(\boldsymbol{x}^{(t-1)}) = \partial_{\boldsymbol{x}} \max_{\boldsymbol{y}\in Y:\boldsymbol{g}(\boldsymbol{x}^{(t-1)},\boldsymbol{y})\geq 0} f(\boldsymbol{x}^{(t-1)},\boldsymbol{y})$. For notational clarity, let $\boldsymbol{g}(t-1) = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)}) + \sum_{k=1}^K \lambda_k^{(t-1)} \nabla_{\boldsymbol{x}} g_k(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)})$. Suppose that $\boldsymbol{x}^\star \in \arg\min_{\boldsymbol{x}\in X} \max_{\boldsymbol{y}\in Y:\boldsymbol{g}(\boldsymbol{x},\boldsymbol{y})} f(\boldsymbol{x},\boldsymbol{y})$. For any $t \in \mathbb{N}$ such that $t \geq 1$, we have:

$$\begin{aligned}
\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|^2 &= \left\|\Pi_X\left(\boldsymbol{x}^{(t-1)} - \eta_t \boldsymbol{g}(t-1)\right) - \Pi_X\left(\boldsymbol{x}^*\right)\right\|^2 && (40)\\
&\leq \left\|\boldsymbol{x}^{(t-1)} - \eta_t \boldsymbol{g}(t-1) - \boldsymbol{x}^*\right\|^2 && (41)\\
&= \left\|\boldsymbol{x}^{(t-1)} - \boldsymbol{x}^*\right\|^2 - 2\eta_t \left\langle \boldsymbol{g}(t-1), \left(\boldsymbol{x}^{(t-1)} - \boldsymbol{x}^*\right)\right\rangle + \eta_t^2 \left\|\boldsymbol{g}(t-1)\right\|^2 && (42)\\
&\leq \left\|\boldsymbol{x}^{(t-1)} - \boldsymbol{x}^*\right\|^2 - 2\eta_t \left[\frac{\mu}{2}\left\|\boldsymbol{x}^{(t-1)} - \boldsymbol{x}^*\right\|^2 + f(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t)}) - f(\boldsymbol{x}^*,\boldsymbol{y}^{(t)})\right] + \eta_t^2 \left\|\boldsymbol{g}(t-1)\right\|^2 && (43)\\
&= \left\|\boldsymbol{x}^{(t-1)} - \boldsymbol{x}^*\right\|^2 - \eta_t\mu \left\|\boldsymbol{x}^{(t-1)} - \boldsymbol{x}^*\right\|^2 - 2\eta_t \left(f(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)}) - f(\boldsymbol{x}^*,\boldsymbol{y}^{(t-1)})\right) + \eta_t^2 \left\|\boldsymbol{g}(t-1)\right\|^2 && (44)\\
&= (1-\eta_t\mu)\left\|\boldsymbol{x}^{(t-1)} - \boldsymbol{x}^*\right\|^2 - 2\eta_t \left(f(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)}) - f(\boldsymbol{x}^*,\boldsymbol{y}^{(t-1)})\right) + \eta_t^2 \left\|\boldsymbol{g}(t-1)\right\|^2 && (45)
\end{aligned}$$

where the first line follows from definitions, the second from the non-expansiveness of the projection operator, the third from algebra, the fourth from the definition of strong convexity, i.e., $\boldsymbol{g}(t-1)^T\left(\boldsymbol{x}^{(t-1)} - \boldsymbol{x}^*\right) \geq \frac{\mu}{2}\left\|\boldsymbol{x}^{(t-1)} - \boldsymbol{x}^*\right\|^2 + f(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)}) - f(\boldsymbol{x}^*,\boldsymbol{y}^{(t-1)})$.

Re-organizing expressions, we get:

$$f(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)}) - f(\boldsymbol{x}^*,\boldsymbol{y}^{(t-1)}) \leq \frac{1-\eta_t\mu}{2\eta_t}\left\|\boldsymbol{x}^{(t-1)} - \boldsymbol{x}^*\right\|^2 - \frac{1}{2\eta_t}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|^2 + \frac{\eta_t}{2}\left\|\boldsymbol{g}(t-1)\right\|^2 \quad (46)$$

Setting $\eta_t = \frac{2}{\mu(t+1)}$, we get:

$$\begin{aligned}
f(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)}) - f(\boldsymbol{x}^*,\boldsymbol{y}^{(t-1)}) &\leq \frac{\mu(t-1)}{4}\left\|\boldsymbol{x}^{(t-1)} - \boldsymbol{x}^*\right\|^2 - \frac{\mu(t+1)}{4}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|^2 + \frac{1}{\mu(t+1)}\left\|\boldsymbol{g}(t-1)\right\|^2 && (47)\\
t\left(f(\boldsymbol{x}^{(t-1)},\boldsymbol{y}^{(t-1)}) - f(\boldsymbol{x}^*,\boldsymbol{y}^{(t-1)})\right) &\leq \frac{\mu t(t-1)}{4}\left\|\boldsymbol{x}^{(t-1)} - \boldsymbol{x}^*\right\|^2 - \frac{\mu t(t+1)}{4}\left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^*\right\|^2 + \frac{1}{\mu}\left\|\boldsymbol{g}(t-1)\right\|^2 && (48)
\end{aligned}$$

where the last line was obtained by multiplying by t on both sides.

Summing up across all iterations on both sides:

$$\sum_{t=0}^{T} t\left(f(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) - f(\boldsymbol{x}^{*}, \boldsymbol{y}^{(t)})\right) \leq \sum_{t=0}^{T} \frac{\mu t(t-1)}{4} \left\|\boldsymbol{x}^{(t-1)} - \boldsymbol{x}^{*}\right\|^2 - \sum_{t=0}^{T} \frac{\mu t(t+1)}{4} \left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{*}\right\|^2 + \sum_{t=0}^{T} \frac{1}{\mu} \left\|\boldsymbol{g}(t-1)\right\|^2 \tag{49}$$

$$= \sum_{t=0}^{T} \frac{\mu t(t-1)}{4} \left\|\boldsymbol{x}^{(t-1)} - \boldsymbol{x}^{*}\right\|^2 - \sum_{t=1}^{T+1} \frac{\mu (t-1)t}{4} \left\|\boldsymbol{x}^{(t-1)} - \boldsymbol{x}^{*}\right\|^2 + \sum_{t=0}^{T} \frac{1}{\mu} \left\|\boldsymbol{g}(t-1)\right\|^2 \tag{50}$$

$$= -\frac{\mu t(t+1)}{4} \left\|\boldsymbol{x}^{(t)} - \boldsymbol{x}^{*}\right\|^2 + \sum_{t=0}^{T} \frac{1}{\mu} \left\|\boldsymbol{g}(t-1)\right\|^2 \tag{51}$$

$$\leq \sum_{t=0}^{T} \frac{1}{\mu} \left\|\boldsymbol{g}(t-1)\right\|^2 \tag{52}$$

$$\leq \frac{T}{\mu} \ell_f \tag{53}$$

where the last line was obtained by noticing that f is ℓ_f-Lipschitz with $\ell_f = \max_{(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}) \in X \times Y} \left\|\nabla_{\boldsymbol{x}} f(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}})\right\|$, which implies that for all $k \in \mathbb{N}$ we know that $\left\|\boldsymbol{g}(k-1)\right\| \leq \ell_f$.

Let $(\boldsymbol{x}_{\text{best}}^{(t)}, \boldsymbol{y}_{\text{best}}^{(t)}) = \arg\min_{(\boldsymbol{x}^{(k)}, \boldsymbol{y}^{(k)}): k \in [t]} f(\boldsymbol{x}^{(k)}, \boldsymbol{y}^{(k)})$:

$$\sum_{t=0}^{T} t\left(f(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) - f(\boldsymbol{x}^{*}, \boldsymbol{y}^{(t)})\right) \leq \frac{T}{\mu} \ell_f \tag{54}$$

$$\sum_{t=0}^{T} t\left(f(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) - \max_{\boldsymbol{y} \in Y : \boldsymbol{g}(\boldsymbol{x}^{*}, \boldsymbol{y}) \geq \boldsymbol{0}} f(\boldsymbol{x}^{*}, \boldsymbol{y})\right) \leq \frac{T}{\mu} \ell_f \tag{55}$$

$$\left(\sum_{t=0}^{T} t\right) \min_{t \in [T]} \left(f(\boldsymbol{x}^{(t)}, \boldsymbol{y}^{(t)}) - \max_{\boldsymbol{y} \in Y : \boldsymbol{g}(\boldsymbol{x}^{*}, \boldsymbol{y}) \geq \boldsymbol{0}} f(\boldsymbol{x}^{*}, \boldsymbol{y})\right) \leq \frac{T}{\mu} \ell_f \tag{56}$$

$$f(\boldsymbol{x}_{\text{best}}^{(T)}, \boldsymbol{y}_{\text{best}}^{(T)}) - \max_{\boldsymbol{y} \in Y : \boldsymbol{g}(\boldsymbol{x}^{*}, \boldsymbol{y}) \geq \boldsymbol{0}} f(\boldsymbol{x}^{*}, \boldsymbol{y}) \leq \frac{\ell_f}{\mu (T+1)} \tag{57}$$

$$\tag{58}$$

That is, as the number of iterations increases, the best iterate converges to a $(0, \delta)$-Stackelberg equilibirium. Additionally, the best iterate converges to a (ε, δ)-Stackelberg equilibrium in $O(\varepsilon^{-1})$ iterations. □

We present the following theorem which proves one of the cases given in Theorem 3.4. The proof for the other cases is the same as the proof below. We note that gradient ascent converges in $O(\varepsilon^{-1})$ iterations to a ε-maximum for a Lipschitz smooth objective, and in $O(\log(\varepsilon))$ iterations to a ε-maximum for a Lipschitz smooth and strongly concave objective [6].

Theorem D.3. *Suppose that Algorithm 2 is run on a convex-concave min-max game with dependent strategy sets given by* $(X, Y, f, \boldsymbol{g})$ *where* X, Y *are convex. Suppose that Assumption 3.1 holds and* f *is* $\ell_{\nabla f}$*-smooth, i.e.,* $\forall (\boldsymbol{x}_1, \boldsymbol{y}_1), (\boldsymbol{x}_2, \boldsymbol{y}_2) \in X \times Y, \left\|\nabla f(\boldsymbol{x}_1, \boldsymbol{y}_1) - \nabla f(\boldsymbol{x}_2, \boldsymbol{y}_2)\right\| \leq \ell_{\nabla f} \left\|(\boldsymbol{x}_1, \boldsymbol{y}_1) - (\boldsymbol{x}_2, \boldsymbol{y}_2)\right\|$.

Let $(\boldsymbol{x}_{\text{best}}^{(t)}, \boldsymbol{y}_{\text{best}}^{(t)}) \in \arg\min_{(\boldsymbol{x}^{(k)}, \boldsymbol{y}^{(k)}): k \in [t]} f(\boldsymbol{x}^{(k)}, \boldsymbol{y}^{(k)})$*. For* $\varepsilon \in (0, 1)$*, if we choose* $T_{\boldsymbol{x}}$ *and* $T_{\boldsymbol{y}}$ *large enough such that:*

$$T_{\boldsymbol{x}} \geq N_{T_{\boldsymbol{x}}}(\varepsilon) := O(\varepsilon^{-2}) \tag{59}$$

$$T_{\boldsymbol{x}} \geq N_{T_{\boldsymbol{y}}}(\varepsilon) := O(\varepsilon^{-1}) \tag{60}$$

then there exists an iteration $T^{\star} \leq T_{\boldsymbol{x}} T_{\boldsymbol{y}} = O(\varepsilon^{-3})$ *such that* $(\boldsymbol{x}_{\text{best}}^{(T^{\star})}, \boldsymbol{y}_{\text{best}}^{(T^{\star})})$ *is an* $(\varepsilon, \varepsilon)$*-Stackelberg equilibrium.*

Proof of Theorem 3.3. Since f is ℓ_f-smooth, it is well known that the inner gradient descent procedure will compute an ε-maximum of $f(\boldsymbol{x}^{(t)}, \cdot)$ for each iterate $\boldsymbol{x}^{(t)}$ in $O(\varepsilon^{-2})$ iterations [6]. Combining the iteration complexity of the outer and inner loops using this result and Theorem 3.3, we obtain an iteration complexity of $O(\varepsilon^{-2})O(\varepsilon^{-1}) = O(\varepsilon^{-3})$. □

E An Economic Application: Details

E.1 Experimental details

E.1.1 General Experiment Setup

Our experimental goals were two-folds. First, we wanted to understand the convergence complexity of our algorithms for different Fisher markets under which the objective function Equation (3) satisfies different smoothness properties. Secondly, we wanted to understand the approximate optimal $\boldsymbol{y}^{(t)}$ computed by the max-oracle in Algorithm 1 or by the inner loop in Algorithm 2 affected the preciseness of Stackelberg equilibrium outputed.

To answer these questions, we have collected data on the prices and allocations computed by Algorithm 1 with an exact max-oracle on each iteration and by Algorithm 2 on each iteration of the outer loop algorithms by running them on randomly initialized markets. We have initialized 500 different linear, Cobb-Douglas, Leontief Fisher markets with 5 buyers and 8 goods. For each of these markets, we have run Algorithm 1 and Algorithm 2 twice, one time with high starting prices and one time with low starting prices to understand the impact of initialization conditions on the algorithm. We have run Algorithm 1 and Algorithm 2's outer loop for 500, 300, and 700 iterations for linear, Cobb-Douglas, and Leontief Fisher markets respectively.

We have opted for a learning rate of 5 for both algorithms after manual hyper-parameter tuning and picked a decay rate of $t^{-1/2}$ for the learning rate based on our theory. For each run of the algorithm, we then computed the objective functions value for the iterates calculated by the algorithm throughout it to obtain Figure 1. Finally, to understand how much precision was lost in the accuracy of the Stackelberg equilibrium outputed by Algorithm 2 from not being able to compute a maximum of $f(\boldsymbol{x}, \cdot)$ for given $\boldsymbol{x} \in X$, we have run a first order James' test to see if the equilibrium strategies outputed by Algorithm 1 and Algorithm 2 were statistically distinguishable.

E.1.2 Computational Requirements, Packages, and Algorithmic Details

The experiments were run on MacOS machine with 8GB ram and an apple M1 chip and experiments took about 2 hours to run. Only CPU ressources were used.

We have run our experiments in Python 3.7 [64] and have used the NumPy [28], Pandas [60], and CVXPY [19]. The data from our experiments can be found on our code repository as well (`https://anonymous.4open.science/r/min-max-fisher-CEFA/`). Figure 1 was graphed via Matplotlib [31]. To run the first order James test, we transfer the data generated by our Python code to an R script [51], which we manipulate using the Tidyverse environment [67], and finally obtain the desired p-values via the STests package in R [30].

Licensing R as a package is licensed under GPL-2 | GPL-3. Python software and documentation are licensed under the PSF License Agreement. Numpy is distributed under a liberal BSD license. Pandas is distributed under a new BSD license. Matplotlib only uses BSD compatible code, and its license is based on the PSF license. CVXPY is licensed under an APACHE license. Tidyverse is distributed under an MIT license.

For our execution of algorithm Algorithm 1 for linear, Cobb-Douglas and Leontief Fisher markets, we used an exact Max-Oracle since the demand has a closed form solution for these markets [25]. As the computational overhead of the projection operation in the inner loop of Algorithm 2 can be high for most projection methods, we have opted to use CVXPY first order for the inner loop of Algorithm 2. In particular, we have opted for the ECOS solver and in case if any runtime exception occurred. Note that these solvers compute ε-optimal points as a result we believe that they present an accurate view of how Algorithm 2 would behave.

E.2 Fisher Market Algorithms

Algorithm 3 δ-Approximate Tâtonnement for Fisher Markets

Inputs: $U, \boldsymbol{b}, \eta, T, \boldsymbol{p}^{(0)}, \delta$
Output: $(\boldsymbol{X}^*, \boldsymbol{p}^*)$

```
for t = 1, ..., T do
    For all i ∈ [n], find x_i^(t) s.t. u_i(x_i^(t)) ≥ max_{x_i : x_i · p^(t−1) ≤ b_i} u_i(x_i) − δ and x_i^(t) · p^(t−1) ≤ b_i
    Set p^(t) = max{ p^(t−1) − η_t(1 − Σ_{i∈[n]} x_i^(t)), 0 }
end for
return (X^(T), p^(T))
```

For all $i \in [n]$, find $\boldsymbol{x}_i^{(t)}$ s.t. $u_i(\boldsymbol{x}_i^{(t)}) \geq \max_{\boldsymbol{x}_i : \boldsymbol{x}_i \cdot \boldsymbol{p}^{(t-1)} \leq b_i} u_i(\boldsymbol{x}_i) - \delta$ and $\boldsymbol{x}_i^{(t)} \cdot \boldsymbol{p}^{(t-1)} \leq b_i$; Set $\boldsymbol{p}^{(t)} = \max \left\{ \boldsymbol{p}^{(t-1)} - \eta_t (1 - \sum_{i \in [n]} \boldsymbol{x}_i^{(t)}), 0 \right\}$

Algorithm 4 δ-Approximate Nested Tâtonnement for Fisher Markets

Inputs: $U, \boldsymbol{b}, \eta, T_{\boldsymbol{p}}, T_{\boldsymbol{X}}, \boldsymbol{p}^{(0)}$
Output: $(\boldsymbol{X}^*, \boldsymbol{p}^*)$

```
for t = 1, ..., T_p do
    for s = 1, ..., T_X do
        For all i ∈ [n], x_i^(t) = Π_{x : x · p^(t−1) ≤ b_i} ( x_i^(t) + b_i / u_i(x_i^(t)) ∇_{x_i} u_i(x_i^(t)) )
    end for
    Set p^(t) = max{ p^(t−1) − η_t(1 − Σ_{i∈[n]} x_i^(t)), 0 }
end for
return (X^(T), p^(T))
```

For all $i \in [n]$, $\boldsymbol{x}_i^{(t)} = \Pi_{\left\{\boldsymbol{x} : \boldsymbol{x} \cdot \boldsymbol{p}^{(t-1)} \leq b_i\right\}} \left(\boldsymbol{x}_i^{(t)} + \frac{b_i}{u_i(\boldsymbol{x}_i^{(t)})} \nabla_{\boldsymbol{x}_i} u_i(\boldsymbol{x}_i^{(t)}) \right)$; Set $\boldsymbol{p}^{(t)} = \max \left\{ \boldsymbol{p}^{(t-1)} - \eta_t (1 - \sum_{i \in [n]} \boldsymbol{x}_i^{(t)}), 0 \right\}$

F Additional Related Work

Much progress has been made recently in solving min-max games with independent strategy sets, both in the convex-concave case and in non-convex-concave case. For the former case, when f is $\mu_{\boldsymbol{x}}$-strongly-convex-$\mu_{\boldsymbol{y}}$-strongly-concave, Tseng [63], Yurii Nesterov [69], and Gidel et al. [24] proposed variational inequality methods and Mokhtari, Ozdaglar, and Pattathil [42] gradient-descent-ascent (GDA)-based methods that compute a solution in $\tilde{O}(\mu_{\boldsymbol{y}} + \mu_{\boldsymbol{x}})$ iterations.. These upper bounds were recently complemented by the lower bound of $\tilde{\Omega}(\sqrt{\mu_{\boldsymbol{y}}\mu_{\boldsymbol{x}}})$, shown by Ibrahim et al. [32] and Zhang, Hong, and Zhang [70]. Subsequently, Lin, Jin, and Jordan [38] and Alkousa et al. [3] analyzed algorithms that converge in $\tilde{O}(\sqrt{\mu_{\boldsymbol{y}}\mu_{\boldsymbol{x}}})$ and $\tilde{O}(\min\left\{\mu_{\boldsymbol{x}}\sqrt{\mu_{\boldsymbol{y}}}, \mu_{\boldsymbol{y}}\sqrt{\mu_{\boldsymbol{x}}}\right\})$ iterations, respectively. For the special case where f is $\mu_{\boldsymbol{x}}$-strongly-convex-linear, Juditsky, Nemirovski, et al. [35], Hamedani and Aybat [27], and Zhao [72] all present methods that converge to an ε-approximate solution in $O(\sqrt{\mu_{\boldsymbol{x}}/\varepsilon})$. When assumptions on $f(\boldsymbol{x}, \cdot)$ are dropped and it is assumed to be $\mu_{\boldsymbol{x}}$-strongly-convex-concave, Thekumparampil et al. [61] provide an algorithm that converges to an approximate solution in $\tilde{O}(\mu_{\boldsymbol{x}}/\varepsilon)$, and Ouyang and Xu [49] provide a lower bound of $\tilde{\Omega}\left(\sqrt{\mu_{\boldsymbol{x}}/\varepsilon}\right)$. Lin, Jin, and Jordan then went on to develop a faster algorithm, with iteration complexity of $\tilde{O}\left(\sqrt{\mu_{\boldsymbol{x}}/\varepsilon}\right)$. When f is simply assumed to be convex-concave, Nemirovski [43], Nesterov [44], and Tseng [62] describe an algorithm with $\tilde{O}\left(\varepsilon^{-1}\right)$ and Ouyang and Xu [49] prove a lower bound of $\Omega(\varepsilon^{-1})$. We include a detailed summary table of these results in Table 4

When f is assumed to be non-convex-$\mu_{\boldsymbol{y}}$-strongly-concave, and the goal is to compute a first-order Nash or "local" Stackelberg equilibrium, Sanjabi et al. [54] provide an algorithm that converges to ε-an approximate solution in $O(\varepsilon^{-2})$ iterations. Jin, Netrapalli, and Jordan [34], Rafique et al. [52], Lin, Jin, and Jordan [37], and Lu, Tsaknakis, and Hong [39] provide algorithms that converge in $\tilde{O}\left(\mu_{\boldsymbol{y}}^2\varepsilon^{-2}\right)$, while Lin, Jin, and Jordan [38] provide an even faster algorithm, with an iteration complexity of $\tilde{O}\left(\sqrt{\mu_{\boldsymbol{y}}}\varepsilon^{-2}\right)$. When f is non-convex-non-concave and the goal to compute is an approximate first-order Nash equilibrium, Lu, Tsaknakis, and Hong [39] provide an algorithm with iteration complexity $\tilde{O}(\varepsilon^{-4})$, while Nouiehed et al. [47] provide an algorithm with iteration complexity $\tilde{O}(\varepsilon^{-3.5})$. More recently, Ostrovskii, Lowy, and Razaviyayn [48] and Lin, Jin, and Jordan [38] proposed an algorithm with iteration complexity $\tilde{O}\left(\varepsilon^{-2.5}\right)$. When f is non-convex-non-concave and the desired solution concept is a "local" Stackelberg equilibrium, Jin, Netrapalli, and Jordan [34], Rafique et al. [52], and Lin, Jin, and Jordan [37] provide algorithms with a $\tilde{O}\left(\varepsilon^{-6}\right)$ complexity. More recently, Thekumparampil et al. [61], Zhao [71], and Lin, Jin, and Jordan [38] have proposed algorithms that converge to an ε-approximate solution in $\tilde{O}\left(\varepsilon^{-3}\right)$ iterations. We include a detailed summary table of these results in Table 5

Table 4: Iteration complexities for min-max games with independent strategy sets in convex-concave settings. Note that these results assume that the objective function is Lipschitz-smooth.

Setting	Reference	Iteration Complexity
$\mu_{\boldsymbol{x}}$-Strongly-Convex-$\mu_{\boldsymbol{y}}$-Strongly-Concave	[63]	$\tilde{O}\left(\mu_{\boldsymbol{x}} + \mu_{\boldsymbol{y}}\right)$
	[69]	
	[24]	
	[42]	
	[3]	$\tilde{O}\left(\min\left\{\mu_{\boldsymbol{x}}\sqrt{\mu_{\boldsymbol{y}}}, \mu_{\boldsymbol{y}}\sqrt{\mu_{\boldsymbol{x}}}\right\}\right)$
	[38]	$\tilde{O}(\sqrt{\mu_{\boldsymbol{x}}\mu_{\boldsymbol{y}}})$
	[32]	$\tilde{\Omega}(\sqrt{\mu_{\boldsymbol{x}}\mu_{\boldsymbol{y}}})$
	[70]	
$\mu_{\boldsymbol{x}}$-Strongly-Convex-Linear	[35]	$O\left(\sqrt{\mu_{\boldsymbol{x}}/\varepsilon}\right)$
	[27]	
	[72]	
$\mu_{\boldsymbol{x}}$-Strongly-Convex-Concave	[61]	$\tilde{O}\left(\mu_{\boldsymbol{x}}/\sqrt{\varepsilon}\right)$
	[38]	$\tilde{O}(\sqrt{\mu_{\boldsymbol{x}}/\varepsilon})$
	[49]	$\tilde{\Omega}\left(\sqrt{\mu_{\boldsymbol{x}}/\varepsilon}\right)$
Convex-Concave	[43]	$O\left(\varepsilon^{-1}\right)$
	[44]	
	[62]	
	[38]	$\tilde{O}\left(\varepsilon^{-1}\right)$
	[49]	$\Omega(\varepsilon^{-1})$

Table 5: Iteration complexities for min-max games with independent strategy sets in non-convex-concave settings. Note that although all these results assume that the objective function is Lipschitz smooth, some authors make more assumptions, e.g., [47] prove their result for objective functions that satisfy the Lojasiwicz condition.

Setting	Reference	Iteration Complexity
Nonconvex-$\mu_{\boldsymbol{y}}$-Strongly-Concave, First Order Nash Equilibrium or Local Stackelberg Equilibrium	[34]	$\tilde{O}(\mu_{\boldsymbol{y}}^2 \varepsilon^{-2})$
	[52]	
	[37]	
	[39]	
	[38]	$\tilde{O}\left(\sqrt{\mu_{\boldsymbol{y}}} \varepsilon^{-2}\right)$
Nonconvex-Concave, First Order Nash Equilibrium	[39]	$\tilde{O}\left(\varepsilon^{-4}\right)$
	[47]	$\tilde{O}\left(\varepsilon^{-3.5}\right)$
	[48]	$\tilde{O}\left(\varepsilon^{-2.5}\right)$
	[38]	
Nonconvex-Concave Local Stackelberg Equilibrium	[34]	$\tilde{O}(\varepsilon^{-6})$
	[47]	
	[38]	
	[61]	$\tilde{O}(\varepsilon^{-3})$
	[71]	
	[38]	