Semi–Infinite Nonconvex Constrained Min–Max Optimization

Cody Melcher[1] **Zeinab Alizadeh**[2] **Lindsey Hiett**[2] **Afrooz Jalilzadeh**[2] **Erfan Yazdandoost Hamedani**[2]

[1]School of Mathematical Sciences, University of Arizona, Tucson, AZ, USA

[2]Department of Systems and Industrial Engineering, University of Arizona, Tucson, AZ, USA

Problem, Motivation, and Challenges

- Problem setup:

$$\min_{x \in \mathbb{R}^n} \max_{y \in Y} \phi(x, y) \quad \text{s.t.} \quad \psi(x, w) \leq 0, \ \forall w \in W$$

- Constrained min–max structure, infinite cardinality constraint set, non–convex in x
- Natural connection to robust and distributionally robust learning
- Classical methods only able to handle parts of this structure

Problem, Motivation, and Challenges

- Problem setup:

$$\min_{x\in\mathbb{R}^n} \max_{y\in Y} \phi(x,y) \quad \text{s.t.} \quad \psi(x,w) \le 0,\ \forall w \in W$$

- Constrained min–max structure, infinite cardinality constraint set, non–convex in x
- Natural connection to robust and distributionally robust learning
- Classical methods only able to handle parts of this structure

Challenges:

- Variable coupling due to min–max structure
- Feasibility enforcement over infinite W
- Finding feasible optimal solution (constrained non-convex)

Problem, Motivation, and Challenges

- Problem setup:

$$\min_{x\in\mathbb{R}^n} \max_{y\in Y} \phi(x,y) \quad \text{s.t.} \quad \psi(x,w) \le 0,\ \forall w \in W$$

- Constrained min–max structure, infinite cardinality constraint set, non–convex in x
- Natural connection to robust and distributionally robust learning
- Classical methods only able to handle parts of this structure

Challenges:

- Variable coupling due to min–max structure
- Feasibility enforcement over infinite W
- Finding feasible optimal solution (constrained non-convex)

✓ **This work:** first-order algorithm with non-asymptotic convergence guarantee to ε-KKT point

Application: Robust MTL

- **Formulation:**

$$\min_{x \in \mathbb{R}^m} \max_{y^{(1)} \in Y_1} \sum_{\xi_j \in D_1^{tr}} y_j^{(1)} \, \ell_1(x, \xi_j)$$

$$\text{s.t.} \quad \sum_{\xi_j \in D_i^{tr}} y_j^{(i)} \, \ell_i(x, \xi_j) \leq r_i, \quad \forall \, y^{(i)} \in Y_i, \; \forall \, i \in \{2, \ldots, T\}.$$

- Prioritize learning one task while ensure remaining tasks have learning loss no worse than target level r.

- Clear constrained min–max structure, task loss $\ell_i(x, \xi_j)$ may be nonconvex in x,

- Constraint must holds for all $y^{(i)} \in Y_i$, the ambiguity set which has infinite cardinality.

Assumptions: Structure and Regularity

- Semi-infinite min–max problem:

$$\min_{x \in \mathbb{R}^n} \max_{y \in Y} \phi(x, y) \quad \text{s.t.} \quad \psi(x, w) \le 0, \ \forall w \in W$$

Assumptions: Structure and Regularity

- Semi-infinite min–max problem:

$$\min_{x\in\mathbb{R}^n}\max_{y\in Y}\phi(x,y)\quad\text{s.t.}\quad\psi(x,w)\le 0,\ \forall w\in W$$

- Structural assumptions:
  - $Y\subseteq\mathbb{R}^m,\ W\subseteq\mathbb{R}^\ell$ nonempty, convex, closed.
  - For each x, $\phi(x,\cdot)$, $\psi(x,\cdot)$ are strongly concave or satisfy PL.
  - ϕ,ψ continuously differentiable, Lipschitz smooth, bounded ∇_x.

Assumptions: Structure and Regularity

- Semi-infinite min–max problem:

$$\min_{x\in\mathbb{R}^n}\max_{y\in Y}\phi(x,y) \quad \text{s.t.} \quad \psi(x,w)\le 0,\ \forall w\in W$$

- Structural assumptions:
  - $Y\subseteq\mathbb{R}^m,\ W\subseteq\mathbb{R}^\ell$ nonempty, convex, closed.
  - For each x, $\phi(x,\cdot)$, $\psi(x,\cdot)$ are strongly concave or satisfy PL.
  - ϕ,ψ continuously differentiable, Lipschitz smooth, bounded ∇_x.
- Regularity (Łojasiewicz-type):

$$[\psi(x,w)]_+^{2\theta}\le\mu\,\|\nabla_x\psi(x,w)[\psi(x,w)]_+\|, \quad \theta\in(0,1),\ \mu>0.$$

Implicit Problem and iDB–PD Updates

- Implicit constrained problem:

$$\min_x f(x) \text{ s.t. } g(x) \le 0, \qquad f(x) = \max_{y \in Y} \phi(x,y), \quad g(x) = \max_{w \in W} \psi(x,w).$$

Implicit Problem and iDB–PD Updates

- Implicit constrained problem:

$$\min_x f(x) \text{ s.t. } g(x) \le 0, \qquad f(x) = \max_{y\in Y} \phi(x,y), \quad g(x) = \max_{w\in W} \psi(x,w).$$

- One–step QP search direction:

$$\min_d \|\nabla f(x_k) + d\|^2 \quad \text{s.t.} \quad \nabla g(x_k)^\top d + \alpha_k \|\nabla g(x_k)\| \le 0$$

yields

$$\begin{aligned} d_k &= -\nabla_x \phi(x_k, y_k) - \lambda_k \nabla_x \psi(x_k, w_k), \\ \lambda_k &= \frac{\big[-\nabla_x \psi(x_k, w_k)^\top \nabla_x \phi(x_k, y_k) + \alpha_k \|\nabla_x \psi(x_k, w_k)\| \big]_+}{\|\nabla_x \psi(x_k, w_k)\|^2}. \end{aligned}$$

Issue: λ_k may blow up near feasible or critical points.

Implicit Problem and iDB–PD Updates

- Implicit constrained problem:

$$\min_x f(x) \text{ s.t. } g(x) \le 0, \qquad f(x) = \max_{y \in Y} \phi(x,y), \quad g(x) = \max_{w \in W} \psi(x,w).$$

- One–step QP search direction:

$$\min_d \|\nabla f(x_k) + d\|^2 \quad \text{s.t.} \quad \nabla g(x_k)^\top d + \alpha_k \|\nabla g(x_k)\| \le 0$$

yields

$$d_k = -\nabla_x \phi(x_k, y_k) - \lambda_k \nabla_x \psi(x_k, w_k),$$
$$\lambda_k = \frac{\big[-\nabla_x \psi(x_k,w_k)^\top \nabla_x \phi(x_k,y_k) + \alpha_k \|\nabla_x \psi(x_k,w_k)\| \big]_+}{\|\nabla_x \psi(x_k,w_k)\|^2}.$$

Issue: λ_k may blow up near feasible or critical points.

✓ Fix: use indicator $\zeta(x_k, w_k) = [\psi(x_k,w_k)]_+ \|\nabla_x \psi(x_k,w_k)\|$

Implicit Problem and iDB–PD Updates

- Implicit constrained problem:

$$\min_x f(x) \text{ s.t. } g(x) \le 0, \qquad f(x) = \max_{y \in Y} \phi(x,y), \quad g(x) = \max_{w \in W} \psi(x,w).$$

- One–step QP search direction:

$$\min_d \|\nabla f(x_k) + d\|^2 \quad \text{s.t.} \quad \nabla g(x_k)^\top d + \alpha_k \|\nabla g(x_k)\| \le 0$$

yields

$$d_k = -\nabla_x \phi(x_k, y_k) - \lambda_k \nabla_x \psi(x_k, w_k),$$
$$\lambda_k = \frac{\left[-\nabla_x \psi(x_k, w_k)^\top \nabla_x \phi(x_k, y_k) + \alpha_k \|\nabla_x \psi(x_k, w_k)\| \right]_+}{\|\nabla_x \psi(x_k, w_k)\|^2}.$$

Issue: λ_k may blow up near feasible or critical points.

✓ Fix: use indicator $\zeta(x_k, w_k) = [\psi(x_k, w_k)]_+ \|\nabla_x \psi(x_k, w_k)\|$

- Resulting iDB–PD updates:

$$x_{k+1} \leftarrow x_k + \gamma_k d_k, \quad y_{k+1} \approx \arg\max_{y \in Y} \phi(x_{k+1}, y), \quad w_{k+1} \approx \arg\max_{w \in W} \psi(x_{k+1}, w)$$

Convergence Analysis

Let $\{x_k, \lambda_k\}_{k=0}^{T-1}$ be the sequence generated by Algorithm 1. For any $k \geq 0$,

$$\alpha_k = \frac{T^{1/3}}{(k+2)^{1+\omega}}, \qquad \gamma_k = \gamma = \mathcal{O}\Big(\min\{T^{-1/3}, (L_f + L_{xy}^{\phi})^{-1}\}\Big),$$

$$N_k = \mathcal{O}(\log(k+1)),$$

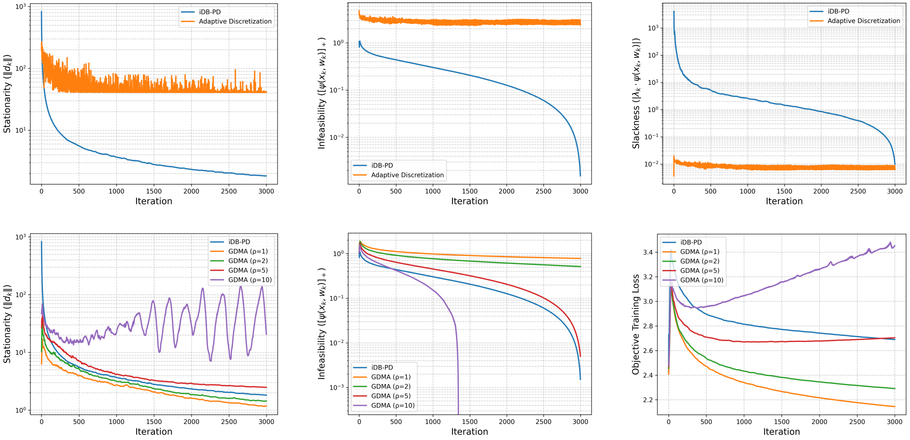
$$M_k = \begin{cases} \mathcal{O}(\max\{\max\{1, \frac{1}{2\theta}\}\log(T), \log(T[\psi(x_k, w_k)]_+^{4\theta-2})\}), & \zeta(x_k, w_k) > 0, \\ \mathcal{O}(\max\{1, \frac{1}{2\theta}\}\log(T)), & \text{otherwise.} \end{cases}$$

For any $\varepsilon > 0$, there exists $t \in \{0, \ldots, T-1\}$ such that

$$\begin{aligned} \|\nabla f(x_t) + \lambda_t \nabla g(x_t)\| \leq \varepsilon \quad &\text{in} \quad T = \mathcal{O}(\varepsilon^{-3}), \\ [g(x_t)]_+ \leq \varepsilon \quad &\text{in} \quad T = \mathcal{O}(\varepsilon^{-6\theta}), \\ |\lambda_t g(x_t)| \leq \varepsilon \quad &\text{in} \quad T = \mathcal{O}\big(\varepsilon^{-3\theta/(1-\theta)}\big). \end{aligned}$$

Experiment: Robust Multi-task Learning

- Two overlaid digits from MNIST with differing priority.
- Compare iDB–PD vs. AD (COOPER) and GDMA baselines.



- iDB–PD achieves a better joint trade-off between a reduction in stationarity and infeasibility than either AD or GDMA.