Recent advances in nonconvex-(non)concave min-max optimization

Dmitrii M. Ostrovskii
University of Southern California

John Hopkins University
AMS Seminar
February 9, 2022

Min-max optimization

Smooth min-max problem

$$\min_{x \in X} \max_{y \in Y} f(x, y)$$

X, Y are two convex sets, f is differentiable with λ-**Lipschitz gradient**.

- Can be seen as a 2-player **zero-sum game** with payoff fns $\pm f(x, y)$.

Min-max optimization

Smooth min-max problem

$$\min_{x \in X} \max_{y \in Y} f(x, y)$$

X, Y are two convex sets, f is differentiable with λ-**Lipschitz gradient**.

- Can be seen as a 2-player **zero-sum game** with payoff fns $\pm f(x, y)$.

Classical setup

f is **convex-concave**:

$f(\cdot, y)$ convex on X for any $y \in Y$; $\quad f(x, \cdot)$ concave on Y for any $x \in X$.

- Extensively studied: von Neumann & Nash ('30s-'40s) $\rightarrow$ Korpelevich (1977) $\rightarrow$ Nemirovski (2004) $\rightarrow$ modern Optim. & Machine Learning.

Min-max optimization

Smooth min-max problem

$$\min_{x \in X} \max_{y \in Y} f(x, y)$$

X, Y are two convex sets, f is differentiable with λ-**Lipschitz gradient**.

- Can be seen as a 2-player **zero-sum game** with payoff fns $\pm f(x, y)$.

Classical setup

f is **convex-concave**:

$f(\cdot, y)$ convex on X for any $y \in Y$; $\quad f(x, \cdot)$ concave on Y for any $x \in X$.

- Extensively studied: von Neumann & Nash ('30s-'40s) $\rightarrow$ Korpelevich (1977) $\rightarrow$ Nemirovski (2004) $\rightarrow$ modern Optim. & Machine Learning.

We focus on a more challenging setup, incorporating **nonconvexity** into f.

App. #1: Robust system design

x – controls; y – inputs/state.

Nominal-value design

$$\min_{x \in X} f(x, \hat{y})$$

$\mapsto$

Robust design

$$\min_{x \in X} \max_{\|y - \hat{y}\| \leqslant \delta} f(x, y)$$

App. #1: Robust system design

x – controls; y – inputs/state.

Nominal-value design

$$\min_{x \in X} f(x, \hat{y})$$

$\mapsto$

Robust design

$$\min_{x \in X} \max_{\|y - \hat{y}\| \leqslant \delta} f(x, y)$$

Concrete examples:

- designing a reliant transportation network with uncertain demand (Sharma et al., 2009; An and Lo, 2015);

App. #1: Robust system design

x – controls; y – inputs/state.

Nominal-value design

$$\min_{x \in X} f(x, \hat{y})$$

$\mapsto$

Robust design

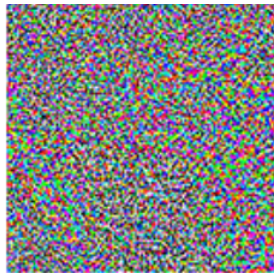
$$\min_{x \in X} \max_{\|y - \hat{y}\| \leqslant \delta} f(x, y)$$

Concrete examples:

- designing a reliant transportation network with uncertain demand (Sharma et al., 2009; An and Lo, 2015);
- adversarial attacks on neural networks (Goodfellow et al., 2015):

"panda"
57.7% confidence

$+ 0.007 \times$



$=$

"gibon"
99.3% confidence

$$\min_{\theta \in \mathbb{R}^d} \max_{\|\epsilon\| \leqslant \delta} \sum_{i=1}^{n} \ell(\theta, \zeta_i - \epsilon_i) + R(\theta).$$

App. #2: Minimization of a discrete maximum

$$\min_{x \in X} \quad \max_{j \in \{1,\ldots,K\}} f_k(x),$$

can be recast as a nonconvex-affine (hence **nonconvex-concave**) problem:

$$\min_{x \in X} \max_{y \in \Delta_K} \langle y, \vec{f}(x) \rangle$$

where Δ_K is the standard simplex and $\vec{f}(x) = [f_1(x); \ldots; f_K(x)] \in \mathbb{R}^K$.

App. #2: Minimization of a discrete maximum

$$\min_{x \in X} \quad \max_{j \in \{1,\ldots,K\}} \quad f_k(x),$$

can be recast as a nonconvex-affine (hence **nonconvex-concave**) problem:

$$\min_{x \in X} \max_{y \in \Delta_K} \langle y, \vec{f}(x) \rangle$$

where Δ_K is the standard simplex and $\vec{f}(x) = [f_1(x); \ldots; f_K(x)] \in \mathbb{R}^K$.

Example:

- max-min power control in MIMO for uniform QoS across users, (Nayebi et al., 2017); in particular using deep learning (D'Andrea et al., 2019).

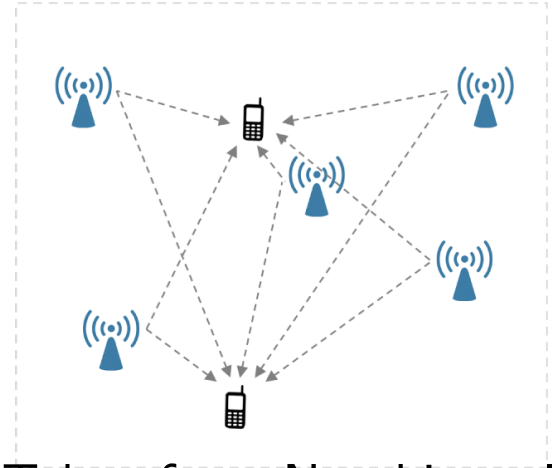


Figure: Taken from Nayebi et al. (2017).

Challenges of min-max optimization

①: Convergence cannot be taken for granted!

Smooth minimization (unconstrained)

$$\min_{x \in \mathcal{X}} f(x)$$

$\mathcal{X}$ is a Euclidean space, $f(\cdot)$ is differentiable with λ-Lipschitz gradient.

Gradient descent (GD) with constant stepsize:

$$x_{t+1} = x_t - \tfrac{1}{\lambda} \nabla f(x_t).$$

- decreases $f(x_t)$ and converges at $O(t^{-1/2})$ rate in terms of $\|\nabla f(x_t)\|$.
- **Convex case:** $O(t^{-1})$ convergence in $f(x_t) - \min_{x \in X} f(x)$ and $\|\nabla f(x_t)\|$.
- Faster convergence $O(t^{-2})$ by Nesterov's algorithm (add momentum).
- Similar results in the constrained setup (for projected GD & Nesterov).

Challenges of min-max optimization (cont'd)

①: Convergence cannot be taken for granted!

Smooth min-max optimization

$$\min_{x\in X}\max_{y\in Y} f(x,y). \qquad (\nabla f \text{ is } \lambda\text{-Lipschitz.})$$

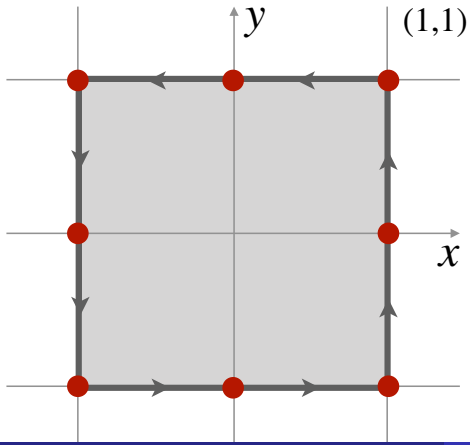
Projected **Gradient Descent-Ascent** (PGDA) with constant stepsizes:

$$\begin{bmatrix} x_{t+1} \\ y_{t+1} \end{bmatrix} = \begin{bmatrix} \Pi_X\left(x_t - \frac{1}{\lambda}\nabla_x f(x_t,y_t)\right) \\ \Pi_Y\left(y_t + \frac{1}{\lambda}\nabla_y f(x_t,y_t)\right) \end{bmatrix}.$$

- Convergence **not guaranteed** even in the convex-concave scenario:

$$\min_{|x|\leqslant 1}\ \max_{|y|\leqslant 1} xy$$

$\nabla f(x,y) = \begin{bmatrix} y \\ x \end{bmatrix}$ is 1-Lipschitz, but PGDA cycles on the boundary of the feasible set.



Challenges of min-max optimization (cont'd)

②: No minimax theorem beyond the convex-concave case!

- **Convex-concave** problems admit strong duality (a.k.a. minimax thm):

$$\min_{x\in X}\left\{\varphi(x) := \max_{y\in Y} f(x,y)\right\} = \max_{y\in Y}\left\{\psi(y) := \min_{x\in X} f(x,y)\right\} = f(x^*,y^*)$$

where (x^*, y^*) is a (global) **saddle point** or **Nash equilibrium**, that is

$$f(x^*,y) \leqslant f(x^*,y^*) \leqslant f(x,y^*), \quad \text{for any } (x,y) \in X \times Y.$$

Challenges of min-max optimization (cont'd)

②: No minimax theorem beyond the convex-concave case!

- **Convex-concave** problems admit strong duality (a.k.a. minimax thm):

$$\min_{x\in X}\left\{\varphi(x) := \max_{y\in Y} f(x,y)\right\} = \max_{y\in Y}\left\{\psi(y) := \min_{x\in X} f(x,y)\right\} = f(x^*,y^*)$$

  where (x^*,y^*) is a (global) **saddle point** or **Nash equilibrium**, that is

$$f(x^*,y) \leqslant f(x^*,y^*) \leqslant f(x,y^*), \quad \text{for any } (x,y)\in X\times Y.$$

- Common approach: $z^* = (x^*,y^*)$ satisfies the **variational inequality**

$$\langle F(z^*), z^* - z\rangle \leqslant 0 \quad \text{for all } z\in X\times Y,$$
$$\text{where } F(z) := [\nabla_x f(x,y); -\nabla_y f(x,y)].$$

- Moreover: if $\widehat{z}$ satisfies that $\langle F(\widehat{z}), \widehat{z} - z\rangle \leqslant \varepsilon$ for all $z\in X\times Y$, then

$$\varphi(\widehat{x}) - \min_{x\in X}\varphi(x) \leqslant \varphi(\widehat{x}) \cancel{-\varphi(x^*)} \cancel{+\psi(y^*)} - \psi(\widehat{y}) \quad \text{[minimax thm.]}$$
$$\leqslant \sup_{z\in Z}\langle F(\widehat{z}), \widehat{z} - z\rangle \leqslant \varepsilon. \quad \text{[convexity-concavity]}$$

- So, we can focus on (approximately) solving V.I.'s (Nemirovski, 2004).

Challenges of min-max optimization (cont'd)

②: No minimax theorem beyond the convex-concave case!

When we give up convexity of $f(\cdot, y)$, the minimax theorem does not apply:

$$\min_{x \in X} \{\varphi(x) := \max_{y \in Y} f(x, y)\} \neq \max_{y \in Y} \min_{x \in X} f(x, y).$$

- In fact, since $\varphi(\cdot)$ is nonconvex, we cannot hope to minimize it globally.

Challenges of min-max optimization (cont'd)

②: No minimax theorem beyond the convex-concave case!

When we give up convexity of $f(\cdot, y)$, the minimax theorem does not apply:

$$\min_{x \in X} \{\varphi(x) := \max_{y \in Y} f(x, y)\} \neq \max_{y \in Y} \min_{x \in X} f(x, y).$$

- In fact, since $\varphi(\cdot)$ is nonconvex, we cannot hope to minimize it globally.

Stackelberg or Nash?

S: Focus on **local minimizers** or **first-order stationary points** of φ.

N: Focus on f, i.e. on local / first-order Nash equilibria (solutions to VI).

- If $f(x, \cdot)$ is concave, these approaches are equivalent (Lin et al., 2019).
- But if $f(x, \cdot)$ is nonconcave, they may even lead to different solutions!

Challenges of min-max optimization (cont'd)

②: No minimax theorem beyond the convex-concave case!

When we give up convexity of $f(\cdot, y)$, the minimax theorem does not apply:

$$\min_{x \in X} \{\varphi(x) := \max_{y \in Y} f(x, y)\} \neq \max_{y \in Y} \min_{x \in X} f(x, y).$$

- In fact, since $\varphi(\cdot)$ is nonconvex, we cannot hope to minimize it globally.

Stackelberg or Nash?

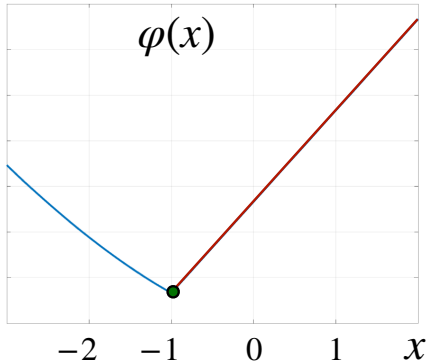
S: Focus on **local minimizers** or **first-order stationary points** of φ.

N: Focus on f, i.e. on local / first-order Nash equilibria (solutions to VI).

- If $f(x, \cdot)$ is concave, these approaches are equivalent (Lin et al., 2019).
- But if $f(x, \cdot)$ is nonconcave, they may even lead to different solutions!

$$\min_{x \in \mathbb{R}} \max_{|y| \leqslant 2} xy + \frac{y^3}{3}$$

$x^* = -1$ unique minimizer & FSP of $\varphi(x)$.
$(0, 0)$ – unique VI solution; **no local NEs**.



Challenge: nonsmoothness of $\varphi(x)$

We still can evaluate $\varphi(x)$ and its subgradient $\xi = \xi(x) \in \partial\varphi(x)$ at any x.

- One can escape spurious saddle points—those of $\varphi(\cdot)$—by augmenting first-order methods with random perturbations (Davis et al., 2021).
- So, let's focus on **first-order stationary points:** $x \in X : \partial\varphi(x) \ni 0$.

But what it means for x to be approximately stationary?

- The norm of $\xi \in \partial\varphi(x)$ is a poor criterion as it can be **discontinuous:**

$$\varphi(x) = |x|: \qquad \partial\varphi(0) \ni 0, \quad \text{but} \quad |\nabla\varphi(x)| = 1 \text{ for } x \neq 0.$$

Moreau envelope and ε-FSP

Definition (Moreau envelope)

The **Moreau envelope** for λ-weakly convex function φ is

$$\varphi_\lambda(x) := \min_{u \in X} \left\{ \varphi(u) + \lambda \|u - x\|^2 \right\}.$$

- Moreau envelope for an λ-weakly convex function φ is λ-smooth.

Definition (ε-approximate first-order stationary point)

If $f(\cdot, y)$ is λ-smooth $\forall y \in Y$, then $x \in X$: $\|\nabla \varphi_\lambda(x)\| \leqslant \varepsilon$ is called ε-FSP.

- **Motivation:**

Moreau envelope and ε-FSP

Definition (Moreau envelope)

The **Moreau envelope** for λ-weakly convex function φ is

$$\varphi_\lambda(x) := \min_{u \in X} \left\{ \varphi(u) + \lambda \|u - x\|^2 \right\}.$$

- Moreau envelope for an λ-weakly convex function φ is λ-smooth.

Definition (ε-approximate first-order stationary point)

If $f(\cdot, y)$ is λ-smooth $\forall y \in Y$, then $x \in X$: $\|\nabla \varphi_\lambda(x)\| \leqslant \varepsilon$ is called ε-FSP.

- **Motivation:**

Lemma (Lin et al., 2019)—mistake corrected in Ostrovskii et al. (2021b).

If $\|\nabla \varphi_\lambda(x)\| \leqslant \varepsilon$ for $x \in X$, then $x^+ := \operatorname*{argmin}_{u \in X} \{\varphi(u) + \lambda \|u - x\|^2\}$ satisfies

$$\|x^+ - x\| \leqslant \frac{\varepsilon}{2\lambda} \quad \text{and} \quad \lambda \|x^+ - \Pi_X[x^+ - \tfrac{1}{\lambda}\xi]\| \leqslant \varepsilon \text{ for some } \xi \in \partial \varphi(x^+).$$

Here $f(x, \cdot)$ may be **nonconcave** and $f(\cdot, y)$ may be **nonconvex**.

Problem formulation, finally

Given $\varepsilon > 0$ and smoothness parameter λ, find ε-FSP for the problem

$$\min_{x \in X} \max_{y \in Y} f(x, y).$$

Nonconvex-concave case

State of the art (early 2019)

There is an algorithm returning an ε-FSP in $O(\varepsilon^{-4})$ gradient computations.

$\downarrow$

Theorem (Ostrovskii et al., 2021b)

There is an algorithm returning an ε-FSP in $O(\varepsilon^{-3})$ gradient computations.

- Complexity bound improves to $O(\varepsilon^{-2})$ if $f(x, \cdot)$ is **strongly concave,** and this is optimal (Zhang et al., 2021). General optimality expected.
- Non-Euclidean projections; X, Y constrained; ε unknown in advance.
- Analysis exploits the known results on first-order methods with inexact oracle, and so is way easier than in concurrent work.

Nonconvex-concave case: insight

Key idea: emulate **proximal point method** for the primal problem:

$$x_{t+1} = \operatorname*{argmin}_x \left\{ \varphi_t(x) := \varphi(x) + \frac{\lambda}{2}\|x - x_t\|^2 \right\},$$

- With a **max-oracle** for $f(x, \cdot)$, we would be done in $\widetilde{O}(\varepsilon^{-2})$ iterations, always remaining stationary in y, by running Nesterov on $\varphi_t(x)$.
- Regularize $f(x, y)$ by $O(\varepsilon_y)$-term, preserving its y-gradient up to ε_y:

$$(x_{t+1}, y_{t+1}) = \operatorname*{argmin}_{x \in X} \left\{ \max_{y \in Y} \left[f(x, y) - \frac{\delta}{2D}\|y\|^2 \right] + \frac{\lambda}{2}\|x - x_t\|^2 \right\}$$

$$= \operatorname*{argmax}_{y \in Y} \left\{ \psi_t(y) := \min_{x \in X} \left[f(x, y) + \frac{\lambda}{2}\|x - x_t\|^2 \right] - \frac{\delta}{2D}\|y\|^2 \right\}$$

  Here $\psi_t(y)$ is λ-smooth and $O(\delta)$-strongly concave $\Rightarrow$ can be maximized in $\widetilde{O}(\delta^{-1/2})$ by Nesterov $\mapsto$ **inexact gradient** for $\varphi_t(x)$.
- Choosing $\delta \asymp \varepsilon^2$ gives the result.

Nonconvex-nonconcave case: new challenge

From now on, assume $\nabla_x f(\cdot)$ is Lipschitz: for any $x', x \in X$ and $y', y \in Y$:

$$\|\nabla_x f(x', y) - \nabla_x f(x, y)\| \leqslant \lambda \|x' - x\|,$$
$$\|\nabla_x f(x, y') - \nabla_x f(x, y)\| \leqslant \mu \|y' - y\|.$$

Thus, λ is the weak convexity modulus of φ, while μ is a coupling parameter.

Lyapunov analyses of first-order methods for minimizing max-functions (PGDA, subgradient, proximal-point method) need **full maximization**.

Key insight

No problem when Y is a singleton. Extend this to the case of a **small** Y?

Our strategy

Let $\hat{f}_k(x, y)$ be the k-order Taylor approximation of $f(x, \cdot)$ at some $\hat{y} \in Y$.

- $\hat{f}_k(x, \cdot)$ is a multivariate polynomial—**global** maximization for $k \leqslant 2$:
  - $\hat{f}_k(x, \cdot)$ is constant for $k = 0$ and affine for $k = 1$;
  - $\hat{f}_k(x, \cdot)$ is quadratic for $k = 2$, can be maximized on a ball via first-order methods (Carmon and Duchi, 2020).

Surrogate problem:

$$\min_{x \in X} \max_{y \in Y} \hat{f}_k(x, y).$$

Strategy

①: Prove that any ε-FSP of the surrogate problem remains $O(\varepsilon)$-FSP for the initial problem **when the diameter of Y is smaller than $\bar{\mathrm{D}} > 0$.**

We expect $\bar{\mathrm{D}} = O(\varepsilon^p)$ for some $p = p(k) > 0$.

②: Find some ε-FSP in the surrogate problem **by an efficient algorithm.**

Accuracy of Taylor approximation

- Assuming k^{th}-order regularity in y, i.e. that $\nabla^k_{y^k} f(x, \cdot)$ is ρ_k-Lipschitz

$$\|\nabla^k_{y^k} f(x, y') - \nabla^k_{y^k} f(x, y)\| \leqslant \rho_k \|y' - y\|,$$

yields

$$|\hat{f}_k(x, y) - f(x, y)| \leqslant \frac{\rho_k D^{k+1}}{(k+1)!}.$$

Accuracy of Taylor approximation

- Assuming k^{th}-order regularity in y, i.e. that $\nabla^k_{y^k} f(x,\cdot)$ is ρ_k-Lipschitz

$$\|\nabla^k_{y^k} f(x,y') - \nabla^k_{y^k} f(x,y)\| \leqslant \rho_k \|y' - y\|,$$

yields

$$|\hat{f}_k(x,y) - f(x,y)| \leqslant \frac{\rho_k D^{k+1}}{(k+1)!}.$$

- Similarly, assuming $\nabla^k_{y^k} f$ is Lipschitz in x ("higher-order interaction")

$$\|\nabla^k_{y^k} f(x',y) - \nabla^k_{y^k} f(x,y)\| \leqslant \sigma_k \|x' - x\|,$$

allows to control how well $\nabla_x \hat{f}_k(x,y)$ approximates $\nabla_x f(x,y)$.

Lemma (Approximation error for $\nabla_x f$.)

$$\|\nabla_x f(x,y) - \nabla_x \hat{f}_k(x,y)\| \leqslant \begin{cases} \dfrac{2\sigma_k D^k}{k!} & \text{for } k \geqslant 1, \\ \min\{\mu D, \sigma_0\} & \text{for } k = 0. \end{cases}$$

Accuracy of Taylor approximation (cont'd)

We have a problem:

- ε-FSP definition requires λ-weak convexity of $\varphi(x) = \max_{y \in Y} f(x, y)$.
- So to even *talk* about ε-FSP for the surrogate, we have to ensure that

$$\hat{\varphi}(x) := \max_{y \in Y} \hat{f}_k(x, y),$$

the surrogate primal function, *is also λ-weakly convex.*

- **Bilinear coupling (BC)**, i.e. $f(x, y) = g(x) + \langle Ax, y \rangle - h(y)$, ensures

$$\nabla^2_{xx} f(x, y) \, [= \nabla^2 g(x) = \nabla^2_{xx} f(x, \hat{y})] = \nabla^2_{xx} \hat{f}_k(x, y)$$

for all y, so in this case $\hat{f}_k(\cdot, y)$ is λ-smooth and $\hat{\varphi}$ is λ-weakly convex.
More generally, assuming $\|\nabla^{k+2}_{y^k x^2} f\| < \infty$ we have the following result:

Lemma (Weak convexity of $\hat{\varphi}$, simplified)

$\nabla_x \hat{f}_k(\cdot, y)$ is $\bar{\lambda}$-Lipschitz ($\hat{\varphi}$ is $\bar{\lambda}$-weakly convex) for $\bar{\lambda} = \lambda + O(D^k) \approx \lambda$.

Main result: critical diameter

Theorem (Ostrovskii et al., 2021a).

Given $k \geqslant 1$, let x^* be ε-FSP for the **surrogate problem**, and assume that

$$\min\left\{\sqrt{\frac{\lambda\rho_k \mathrm{D}^{k+1}}{(k+1)!}},\ \ \mu\mathrm{D} + \mathbb{1}\{k>0\}\frac{\sigma_k \mathrm{D}^k}{k!}\right\} \lesssim \varepsilon.$$

Then x^* is 6ε-FSP for the **initial problem**.

In other words, reduction to the surrogate problem works for $\mathrm{D} \lesssim \bar{\mathrm{D}}$ with

$$\bar{\mathrm{D}} := \max\left\{\frac{\varepsilon}{\mu},\ \ k\cdot\left(\frac{\varepsilon^2}{\lambda\rho_k}\right)^{\frac{1}{k+1}}\right\}.$$

- For $k=1$ we have $\bar{\mathrm{D}} = \frac{\varepsilon}{\min\{\mu,\sqrt{\lambda\rho_1}\}}$. Same rate as for $k=1$ except for a better constant factor in the strong coupling regime $\mu \geqslant \sqrt{\lambda\rho_1}$.
- For $k=2$ and in the nontrivial regime $\varepsilon \ll 1$, we have $\bar{\mathrm{D}} = \frac{\varepsilon^{2/3}}{(\lambda\rho_2)^{1/3}}$.
- Similar picture for $k>2$: coupling-independent $\bar{\mathrm{D}} = \bar{\mathrm{D}}(\varepsilon)$ when $\varepsilon \ll 1$.

Coupling-dependent bound

Proposition 1. For any $x^* \in X$ such that $\|\nabla\hat{\varphi}_{2\lambda}(x^*)\| \leqslant \varepsilon$, one has

$$\|\nabla\hat{\varphi}_{\lambda}(x^*) - \nabla\varphi_{\lambda}(x^*)\| \lesssim \begin{cases} \mu D + \dfrac{\sigma_k D^k}{k!} + \varepsilon & \text{for} \quad k \geqslant 1, \\ \min\{\mu D, \sigma_0\} + \varepsilon & \text{for} \quad k = 0. \end{cases}$$

Coupling-independent bound

Proposition 2. Moreau envelope gradients for φ and $\hat{\varphi}$ are *uniformly close*:

$$\|\nabla\hat{\varphi}_\lambda(x) - \nabla\varphi_\lambda(x)\| \lesssim \sqrt{\frac{\lambda\rho_k \mathrm{D}^{k+1}}{(k+1)!}} \quad \textit{for all } x \in X.$$

Coupling-independent bound: the proof

Proposition 2. Moreau envelope gradients for φ and $\hat{\varphi}$ are *uniformly close*:

$$\|\nabla\hat{\varphi}_\lambda(x) - \nabla\varphi_\lambda(x)\| \lesssim \sqrt{\frac{\lambda\rho_k \mathrm{D}^{k+1}}{(k+1)!}} \quad \textit{for all } x \in X.$$

Proof:

1^o. By the first-order optimality conditions for $\varphi_\lambda(x)$ and $\hat{\varphi}_\lambda(x)$ we have

$$\nabla\varphi_\lambda(x) = 2\lambda(x - x^+), \quad \nabla\hat{\varphi}_\lambda(x) = 2\lambda(x - \hat{x}^+),$$

where x^+ and $\hat{x}^+$ are the proximal-point mappings of x as per φ and $\hat{\varphi}$:

$$x^+ = \operatorname*{argmin}_{u\in X}\{\varphi(u) + \lambda\|u - x\|^2\}, \quad \hat{x}^+ = \operatorname*{argmin}_{u\in X}\{\hat{\varphi}(u) + \lambda\|u - x\|^2\}.$$

Thus $\|\nabla\varphi_\lambda(x) - \nabla\hat{\varphi}_\lambda(x)\| = 2\lambda\|\hat{x}^+ - x^+\|$. Let us bound $\|\hat{x}^+ - x^+\|$.

Coupling-independent bound: the proof (cont'd)

Proposition 2. Moreau envelope gradients for φ and $\hat{\varphi}$ are *uniformly close*:

$$\|\nabla\hat{\varphi}_\lambda(x) - \nabla\varphi_\lambda(x)\| \lesssim \sqrt{\frac{\lambda\rho_k \mathrm{D}^{k+1}}{(k+1)!}} \quad \textit{for all } x \in X.$$

Proof (cont'd):

2^o. Functions $\varphi(\cdot) + \lambda\|\cdot - x\|^2$ and $\hat{\varphi}(\cdot) + \lambda\|\cdot - x\|$ are λ-strongly convex and minimized at x^+ and $\hat{x}^+$ correspondingly, hence

$$\tfrac{1}{2}\lambda\|\hat{x}^+ - x^+\|^2 \leqslant \varphi(\hat{x}^+) + \lambda\|\hat{x}^+ - x\|^2 - \varphi(x^+) - \lambda\|x^+ - x\|^2,$$
$$\tfrac{1}{2}\lambda\|\hat{x}^+ - x^+\|^2 \leqslant \hat{\varphi}(x^+) + \lambda\|x^+ - x\|^2 - \hat{\varphi}(\hat{x}^+) - \lambda\|\hat{x}^+ - x\|^2.$$

Summing the two inequalities results in

$$\lambda\|\hat{x}^+ - x^+\|^2 \leqslant \hat{\varphi}(x^+) - \varphi(x^+) + \varphi(\hat{x}^+) - \hat{\varphi}(\hat{x}^+) \leqslant 2\sup_{x\in X}|\hat{\varphi}(x) - \varphi(x)|.$$

Coupling-independent bound: the proof (cont'd)

Proposition 2. Moreau envelope gradients for φ and $\hat{\varphi}$ are *uniformly close*:

$$\|\nabla\hat{\varphi}_\lambda(x) - \nabla\varphi_\lambda(x)\| \lesssim \sqrt{\frac{\lambda\rho_k \mathrm{D}^{k+1}}{(k+1)!}} \quad \textit{for all } x \in X.$$

Proof (cont'd):

2^o. Functions $\varphi(\cdot) + \lambda\|\cdot - x\|^2$ and $\hat{\varphi}(\cdot) + \lambda\|\cdot - x\|$ are λ-strongly convex and minimized at x^+ and $\hat{x}^+$ correspondingly, hence

$$\tfrac{1}{2}\lambda\|\hat{x}^+ - x^+\|^2 \leqslant \varphi(\hat{x}^+) + \lambda\|\hat{x}^+ - x\|^2 - \varphi(x^+) - \lambda\|x^+ - x\|^2,$$
$$\tfrac{1}{2}\lambda\|\hat{x}^+ - x^+\|^2 \leqslant \hat{\varphi}(x^+) + \lambda\|x^+ - x\|^2 - \hat{\varphi}(\hat{x}^+) - \lambda\|\hat{x}^+ - x\|^2.$$

Summing the two inequalities results in

$$\lambda\|\hat{x}^+ - x^+\|^2 \leqslant \hat{\varphi}(x^+) - \varphi(x^+) + \varphi(\hat{x}^+) - \hat{\varphi}(\hat{x}^+) \leqslant 2\sup_{x\in X}|\hat{\varphi}(x) - \varphi(x)|.$$

3^o. We arrive at $|\hat{\varphi}(x) - \varphi(x)| \leqslant \sup_{y\in Y}|\hat{f}_k(x,y) - f(x,y)| \leqslant \dfrac{\rho_k \mathrm{D}^{k+1}}{(k+1)!}$. ■

References I

Kun An and Hong K Lo. Robust transit network design with stochastic demand considering development density. *Transportation Research Part B: Methodological*, 81:737–754, 2015.

Y. Carmon and J. C. Duchi. First-order methods for nonconvex quadratic minimization. *SIAM Review*, 62(2):395–436, 2020.

Carmen D'Andrea, Alessio Zappone, Stefano Buzzi, and Merouane Debbah. Uplink power control in cell-free massive mimo via deep learning. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 554–558. IEEE, 2019.

Damek Davis, Mateo Díaz, and Dmitriy Drusvyatskiy. Escaping strict saddle points of the Moreau envelope in nonsmooth optimization. *arXiv preprint arXiv:2106.09815*, 2021.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2015.

G. M. Korpelevich. Extragradient method for finding saddle points and other problems. *Matekon*, 13(4):35–49, 1977.

References II

T. Lin, C. Jin, and M. I. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. *arXiv preprint arXiv:1906.00331*, 2019.

Elina Nayebi, Alexei Ashikhmin, Thomas L Marzetta, Hong Yang, and Bhaskar D Rao. Precoding and power optimization in cell-free massive mimo systems. *IEEE Transactions on Wireless Communications*, 16(7):4445–4459, 2017.

A. Nemirovski. Prox-method with rate of convergence $O(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1): 229–251, 2004.

Dmitrii M Ostrovskii, Babak Barazandeh, and Meisam Razaviyayn. Nonconvex-nonconcave min-max optimization with a small maximization domain. *arXiv preprint arXiv:2110.03950*, 2021a.

Dmitrii M Ostrovskii, Andrew Lowy, and Meisam Razaviyayn. Efficient search of first-order Nash equilibria in nonconvex-concave smooth min-max problems. *SIAM Journal on Optimization*, 31(4):2508–2538, 2021b.

Sushant Sharma, Satish V Ukkusuri, and Tom V Mathew. Pareto optimal multiobjective optimization for robust transportation network design problem. *Transportation Research Record*, 2090(1):95–104, 2009.

References III

S. Zhang, J. Yang, C. Guzmán, N. Kiyavash, and N. He. The complexity of nonconvex-strongly-concave minimax optimization. *arXiv preprint arXiv:2103.15888*, 2021.

Smooth minimization boot camp

$$\min_{x \in \mathcal{X}} f(x)$$

where $\mathcal{X}$ is a Euclidean space, $f(\cdot)$ is differentiable with **Lipschitz gradient:**

$$\|\nabla f(x) - \nabla f(x')\| \leqslant L\|x - x'\|, \quad \forall x, x' \in \mathcal{X}.$$

- $\nabla^2 f$ exists almost everywhere and bounded by $L \Rightarrow$ **Descent Lemma:**

$$f(x) \leqslant f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \tfrac{1}{2} L\|x - x_t\|^2.$$

- The RHS is a quadratic. Choosing x_{t+1} to minimize it we get

$$x_{t+1} = x_t - \tfrac{1}{L}\nabla f(x_t).$$

Gradient descent method generates a sequence $x_1, x_2, \ldots$ of such updates.

Smooth minimization boot camp (cont'd)

Gradient descent (GD)

$$x_{t+1} = x_t - \frac{1}{\lambda}\nabla f(x_t).$$

Descent Lemma

$$f(x) \leqslant f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2}\lambda \|x - x_t\|^2.$$

- Objective value decreases at each step:

$$f(x_{t+1}) \leqslant f(x_t) - \frac{1}{2L}\|\nabla f(x_t)\|^2$$

Smooth minimization boot camp (cont'd)

Gradient descent (GD)

$$x_{t+1} = x_t - \tfrac{1}{\lambda}\nabla f(x_t).$$

Descent Lemma

$$f(x) \leqslant f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \tfrac{1}{2}\lambda \|x - x_t\|^2.$$

- Objective value decreases at each step:

$$f(x_{t+1}) \leqslant f(x_t) - \tfrac{1}{2L}\|\nabla f(x_t)\|^2 \leqslant \quad \cdots \quad \leqslant f(x_1) - \tfrac{1}{2L}\sum_{\tau \leqslant t} \|\nabla f(x_\tau)\|^2$$

Smooth minimization boot camp (cont'd)

Gradient descent (GD)

$$x_{t+1} = x_t - \tfrac{1}{\lambda}\nabla f(x_t).$$

Descent Lemma

$$f(x) \leqslant f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \tfrac{1}{2}\lambda \|x - x_t\|^2.$$

- Objective value decreases at each step:

$$f(x_{t+1}) \leqslant f(x_t) - \tfrac{1}{2L}\|\nabla f(x_t)\|^2 \leqslant f(x_1) - \tfrac{t}{2L} \min_{\tau \leqslant t} \|\nabla f(x_\tau)\|^2.$$

$\Rightarrow$ $O(t^{-1/2})$ convergence in **gradient norm**–i.e. $O(\varepsilon^{-2})$ complexity.

Smooth minimization boot camp (cont'd)

Gradient descent (GD)

$$x_{t+1} = x_t - \tfrac{1}{\lambda}\nabla f(x_t).$$

Descent Lemma

$$f(x) \leqslant f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \tfrac{1}{2}\lambda \|x - x_t\|^2.$$

- Objective value decreases at each step:

$$f(x_{t+1}) \leqslant f(x_t) - \tfrac{1}{2L}\|\nabla f(x_t)\|^2 \leqslant f(x_1) - \tfrac{t}{2L} \min_{\tau \leqslant t} \|\nabla f(x_\tau)\|^2.$$

  $\Rightarrow O(t^{-1/2})$ convergence in **gradient norm**–i.e. $O(\varepsilon^{-2})$ complexity.

- Improves to $O(\varepsilon^{-1})$, in **obj. value** and **grad. norm**, if f is **convex**.

Smooth minimization boot camp (cont'd)

Gradient descent (GD)

$$x_{t+1} = x_t - \tfrac{1}{\lambda}\nabla f(x_t).$$

Descent Lemma

$$f(x) \leqslant f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \tfrac{1}{2}\lambda \|x - x_t\|^2.$$

- Objective value decreases at each step:

$$f(x_{t+1}) \leqslant f(x_t) - \tfrac{1}{2L}\|\nabla f(x_t)\|^2 \leqslant f(x_1) - \tfrac{t}{2L} \min_{\tau \leqslant t} \|\nabla f(x_\tau)\|^2.$$

  $\Rightarrow O(t^{-1/2})$ convergence in **gradient norm**–i.e. $O(\varepsilon^{-2})$ complexity.

- Improves to $O(\varepsilon^{-1})$, in **obj. value** and **grad. norm**, if f is **convex**.
- Can do better: $O(\varepsilon^{-1/2})$ via **Nesterov's method**, and this is optimal.

Convex-concave min-max optimization as V.I.

- **Convex-concave** problems admit strong duality (a.k.a. minimax thm):

$$\min_{x\in X}\left\{\varphi(x) := \max_{y\in Y} f(x,y)\right\} = \max_{y\in Y}\left\{\psi(y) := \min_{x\in X} f(x,y)\right\} = f(x^*,y^*)$$

where (x^*,y^*) is a (global) **saddle point** or **Nash equilibrium**, that is

$$f(x^*,y) \leqslant f(x^*,y^*) \leqslant f(x,y^*), \quad \text{for any } (x,y)\in X\times Y.$$

- Common approach: $z^* = (x^*,y^*)$ satisfies a **variational inequality**

$$\langle F(z^*), z^* - z\rangle \leqslant 0 \quad \text{for all } \ z\in X\times Y,$$
$$\text{where } \ F(z) := [\nabla_x f(x,y); -\nabla_y f(x,y)].$$

- Indeed: if $z_t = (x_t,y_t)$ satisfies $\langle F(z_t), z_t - z\rangle \leqslant \varepsilon \ \ \forall z\in X\times Y$, then

$$\begin{aligned}
\varphi(x_t) - \varphi(x^*) &\leqslant \varphi(x_t) \quad \cancel{-\varphi(x^*)} \qquad \cancel{+\psi(y^*)} \quad -\psi(y_t)\\
&= f(x_t,\bar{y}) - f(x_t,y_t) \qquad + f(x_t,y_t) - f(\bar{x},y_t)\\
\text{[convexity-concavity]} &\leqslant \langle -\nabla_y f(x_t,y_t), y_t - \bar{y}\rangle + \langle \nabla_x f(x_t,y_t), x_t - \bar{x}\rangle\\
&\leqslant \langle F(z_t), z_t - \bar{z}\rangle \leqslant \varepsilon.
\end{aligned}$$

Coupling-dependent bound

Proposition 1. For any $x^* \in X$ such that $\|\nabla\hat{\varphi}_{2\lambda}(x^*)\| \leqslant \varepsilon$, one has

$$\|\nabla\hat{\varphi}_{\lambda}(x^*) - \nabla\varphi_{\lambda}(x^*)\| \lesssim \begin{cases} \mu\mathrm{D} + \dfrac{\sigma_k \mathrm{D}^k}{k!} + \varepsilon & \text{for} \quad k \geqslant 1, \\ \min\{\mu\mathrm{D}, \sigma_0\} + \varepsilon & \text{for} \quad k = 0. \end{cases}$$

Coupling-dependent bound: the proof (cont'd)

Proposition 1. For any $x^* \in X$ such that $\|\nabla\hat{\varphi}_{2\lambda}(x^*)\| \leqslant \varepsilon$, one has

$$\|\nabla\hat{\varphi}_\lambda(x^*) - \nabla\varphi_\lambda(x^*)\| \lesssim \begin{cases} \mu\mathrm{D} + \dfrac{\sigma_k \mathrm{D}^k}{k!} + \varepsilon & \text{for} \quad k \geqslant 1, \\ \min\{\mu\mathrm{D}, \sigma_0\} + \varepsilon & \text{for} \quad k = 0. \end{cases}$$

Proof: (assuming $X = \mathcal{X}$ and $k \geqslant 1$ for simplicity)

1^o. Now let $x^+, \hat{x}^+$ be the proximal-point mappings of x^* as per $\varphi, \hat{\varphi}$:

$$\nabla\varphi_\lambda(x^*) = 2\lambda(x^* - x^+), \quad \nabla\hat{\varphi}_\lambda(x^*) = 2\lambda(x^* - \hat{x}^+),$$

Thus $\|\nabla\varphi_\lambda(x^*) - \nabla\hat{\varphi}_\lambda(x^*)\| = 2\lambda\|\hat{x}^+ - x^+\|$.

2^o. By the λ-strong convexity of $\varphi(\cdot) + \lambda\|\cdot - x^*\|^2$ and Cauchy-Schwarz:

$$\begin{aligned} \tfrac{1}{2}\lambda\|\hat{x}^+ - x^+\|^2 &\leqslant \lambda\|\hat{x}^+ - x^*\|^2 + \varphi(\hat{x}^+) - \varphi(x^+) - \lambda\|x^+ - x^*\|^2 \\ &\leqslant 4\lambda\|\hat{x}^+ - x^*\|^2 + \varphi(\hat{x}^+) - \varphi(x^+) - \tfrac{3}{4}\lambda\|\hat{x}^+ - x^+\|^2. \end{aligned}$$

Coupling-dependent bound: the proof (cont'd)

Rearranging, we get

$$(\lambda\|\hat{x}^+ - x^+\|)^2 \leqslant 8(\lambda\|\hat{x}^+ - x^*\|)^2 + 2\lambda\left[\varphi(\hat{x}^+) - \varphi(x^+) - \tfrac{3}{4}\lambda\|\hat{x}^+ - x^+\|^2\right].$$

3^o. Since x^* is an ε-FSP for $\hat{\varphi}_k$, the Moreau criterion characterization gives

$$\|\hat{x}^+ - x^*\| \leqslant \frac{\varepsilon}{2\lambda} \quad \text{and} \quad \|\hat{\xi}\| \leqslant \varepsilon \ \text{ for some } \hat{\xi} \in \partial\hat{\varphi}(\hat{x}^+).$$

Using the first inequality,

$$(\lambda\|\hat{x}^+ - x^+\|)^2 \leqslant 2\varepsilon^2 + 2\lambda\left[\varphi(\hat{x}^+) - \varphi(x^+) - \tfrac{3}{4}\lambda\|\hat{x}^+ - x^+\|^2\right].$$

4^o. By convexity of $\varphi(\cdot) + \frac{1}{2}\lambda\|\cdot - \hat{x}^+\|^2$, for **arbitrary** $\xi \in \partial\varphi(\hat{x}^+)$ we get

$$\varphi(\hat{x}^+) - \varphi(x^+) - \tfrac{\lambda}{2}\|\hat{x}^+ - x^+\|^2 \leqslant \langle \xi, \hat{x}^+ - x^+\rangle,$$

whence $(\lambda\|\hat{x}^+ - x^+\|)^2 \leqslant 2\varepsilon^2 + 2\lambda\left[\langle \xi, \hat{x}^+ - x^+\rangle - \frac{1}{4}\lambda\|\hat{x}^+ - x^+\|^2\right].$

Coupling-dependent bound: the proof (cont'd)

... whence $(\lambda\|\hat{x}^+ - x^+\|)^2 \leqslant 2\varepsilon^2 + 2\lambda\left[\langle\xi, \hat{x}^+ - x^+\rangle - \frac{1}{4}\lambda\|\hat{x}^+ - x^+\|^2\right]$.

5^o. Applying Cauchy-Schwarz twice we get

$$(\lambda\|\hat{x}^+ - x^+\|)^2 \leqslant 4\varepsilon^2 + 4\lambda\left[\langle\hat{\xi}, \hat{x}^+ - x^+\rangle - \tfrac{1}{4}\lambda\|\hat{x}^+ - x^+\|^2\right] + 4\|\hat{\xi} - \xi\|^2$$
$$\leqslant 4\varepsilon^2 + 4\|\hat{\xi}\|^2 + 4\|\hat{\xi} - \xi\|^2.$$

Recall that $\hat{\xi} \in \partial\hat{\varphi}(\hat{x}^+)$ was chosen to guarantee $\|\hat{\xi}\| \leqslant \varepsilon$. Thus we get

$$(\lambda\|\hat{x}^+ - x^+\|)^2 \leqslant 8\varepsilon^2 + 4\|\hat{\xi} - \xi\|^2.$$

6^o. It remains to bound $\|\hat{\xi} - \xi\|^2$. The "subgradient of max" rule implies:

$$\hat{\xi} \in \overline{\mathrm{conv}}\left(\left\{\nabla_x\hat{f}_k(\hat{x}^+, y),\ y \in \mathrm{Argmax}_{y\in Y}\,\hat{f}_k(\hat{x}^+, y)\right\}\right).$$

Besides, we can choose $\xi = \nabla_x f(\hat{x}^+, y^*)$ for $y^* \in \mathrm{Argmax}_{y\in Y} f(\hat{x}^+, y)$. Hence, choosing $\bar{y} \in \mathrm{Argmax}_{y\in Y}\|\nabla_x\hat{f}_k(\hat{x}^+, y) - \nabla_x f(\hat{x}^+, y^*)\|$ we get

$$\|\hat{\xi}_X^+ - \xi^+\| \leqslant \|\nabla_x\hat{f}_k(\hat{x}^+, \bar{y}) - \nabla_x f(\hat{x}^+, y^*)\|$$
$$\leqslant \underbrace{\|\nabla_x f(\hat{x}^+, \bar{y}) - \nabla_x f(\hat{x}^+, y^*)\|}_{\leqslant \mu D} + \underbrace{\|\nabla_x f(\hat{x}^+, y^*) - \nabla_x\hat{f}_k(\hat{x}^+, y^*)\|}_{\leqslant \frac{2}{k!}\sigma_k D^k}. \quad \blacksquare$$

Hessian approximation

Lemma (Weak convexity of $\hat{\varphi}$)

Assume $\|\nabla^{k+2}_{y^k x^2} f\| \leqslant \tau_k$. *Then* $\nabla_x \hat{f}_k(\cdot, y)$ *is* $\bar{\lambda}$*-Lipschitz with*

$$\bar{\lambda} := \lambda + \frac{2\tau_k D^k}{k!} \mathbb{1}\{k \geqslant 1\}.$$

In fact, under some mild measurability condition it suffices to assume that $\nabla^{k+1}_{y^k x} f(\cdot, y)$ is τ_k-Lipschitz for all $\forall y \in Y$, so we don't need $f \in C^{k+2}$.