Nonconvex-Nonconcave Min-Max Optimization with a Small Maximization Domain

arxiv.org/abs/2110.03950

Dmitrii M. Ostrovskii
Babak Barazandeh & Meisam Razaviyayn

Johns Hopkins U.
AMS Seminar
October 14, 2021

Outline

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

- **Background and challenges**.
- **Our approach**: restricting $\mathrm{diam}(Y)$.
- **Sharp bound** for the critical diameter.
- ~~**Algorithms** for finding stationary points.~~

Smooth min-max optimization

Given convex bodies X, Y in the corresponding Euclidean spaces $E_{\mathrm{x}}, E_{\mathrm{y}}$, find

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

assuming that f is smooth—has Lipschitz gradient $[\nabla_{\mathrm{x}} f(x, y);\ \nabla_{\mathrm{y}} f(x, y)]$.

- Full knowledge of X, Y: can compute proximal mappings.
- **Oracle** access to f: can query $f(x, y)$, $\nabla f(x, y), \ldots$ at $(x, y) \in X \times Y$.

- **Iterative methods:** form a sequence (x_t, y_t) such that $f(x_t, y_t) \to f^*$.
- **Complexity:** number of iterations T to guarantee a given accuracy.

Convex-concave setup

Classical setup: $f(\cdot, y)$ convex on X; $f(x, \cdot)$ concave on Y for all x, y.

- **Strong duality** (a.k.a. minimax theorem) under mild assumptions:

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

  (x^*, y^*) is a *saddle point*: $f(x^*, y) \leqslant f(x^*, y^*) \leqslant f(x, y^*)$ for all x, y

- Primal-dual algorithms minimize the **duality gap** (=primal+dual gap):

$$\varphi(x_t) \underbrace{- \varphi^* + \psi^*}_{= f^* - f^* = 0} - \psi(y_t) \leqslant \langle \nabla_x f(x_t, y_t), x_t - x^* \rangle + \langle \nabla_y f(x_t, y_t), y^* - y_t \rangle \, .$$

- Complexity $O(1/\epsilon)$ to reach ϵ duality gap is optimal without further assumptions—via extragradient-type algorithms (Nemirovski '2000).
- Well developed theory by now, although there is still ongoing work.

  (E.g. convergence of the last iterate vs. the averaged iterate.)

Nonconvex-concave setup

When $f(\cdot, y)$ is nonconvex, some of the nice structure is lost; in particular:

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

We still can evaluate $\varphi(x)$ and its subgradient $\xi = \xi(x) \in \partial\varphi(x)$ at any x. However, $\varphi(x)$ is nonconvex, so we lose all hope to minimize it globally.

Reasonable goal is to approximate a local minimizer or a **stationary point.**

- Under mild assumptions, we can escape "malignant" saddle points —those of $\varphi(\cdot)$—and focus on finding a **stationary point.**
  (Jin et al. '2017 for smooth minimization, Davis & Drusvyatskiy '2020).

But what it *means* for $x \in X$ to be ε-**stationary** when $\varphi(x)$ is nonsmooth?

It doesn't make sense to just use the norm of subgradients of φ. E.g., $\varphi(x) = |x|$: $x = 0$ is stationary ($\partial\varphi(0) \ni 0$), but $|\nabla\varphi(x)| \geqslant 1$ if $x \neq 0$.

Nash or Moreau?

But what it *means* for $x \in X$ to be ε-**stationary** when $\varphi(x)$ is nonsmooth?

- First-order Nash Equilibrium (ε-FNE): $\|\nabla_x f(x,y)\| + \|\nabla_y f(x,y)\| \leqslant \varepsilon$.
  Actually more complicated, taking into account the constraint sets...
  Stems from the primal-dual viewpoint: treats $f(\cdot, y), f(x, \cdot)$ equally.

- Or we can hold to the "primal-only" viewpoint if we make $\varphi(\cdot)$ smooth.
  It is possible since φ is λ-weakly convex (i.e., $\varphi(\cdot) + \frac{1}{2}\lambda\|\cdot\|$ is convex.)

Definition

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

is called the (standard) **Moreau envelope** of a λ-weakly convex function ϕ.

We have $\varphi(\cdot) = \max_{y \in Y} f(\cdot, y)$; each $f(\cdot, y)$ is λ-smooth $\Rightarrow$ λ-weakly convex.

- $\varphi_\lambda(\cdot)$ is differentiable and λ-smooth—same as each component $f(\cdot, y)$.

Moreau envelope criterion

Definition

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

is called the (standard) **Moreau envelope** of a λ-weakly convex function ϕ.

Proposition (Ostrovskii, Lowy, Razaviyayn '2020).

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

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

Here $f(x, \cdot)$ doesn't have to be concave. This motivates using $\|\nabla \varphi_\lambda(\cdot)\|$ as a measure of stationarity in the **general (nonconvex-nonconcave) setup.**

Definition (ε-first-order stationary point, or ε-FSP)

Let $f(\cdot, y)$ be λ-smooth $\forall y$. Then $x \in X$ is called ε-FSP if $\|\nabla \varphi_\lambda(x)\| \leqslant \varepsilon$.

Finding an ε-FSP: main challenge

From now on, we 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, and μ is the coupling parameter.

Problem of interest

Given a problem instance of the form $\min_{x\in X} \max_{y\in Y} f(x, y)$ and $\varepsilon > 0$, find a point x^* such that $\|\nabla \varphi_\lambda(x)\| \leqslant \varepsilon$, where φ_λ is the Moreau envelope.

Hard: Lyapunov-type analyses of local search methods (gradient descent-ascent, proximal-point method) rely on **full** maximization in y.

Key insight

Easy problem if Y is a singleton. Does this extend to the case of **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 $k = 2$, admits global maximization via first-order algorithms—see e.g. (Carmon and Duchi '2020).

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

Strategy

1^o. Prove that any ε-FSP of the surrogate problem remains $O(\varepsilon)$-FSP for the initial problem when $\mathrm{D} := \mathrm{diam}(Y)$ **is smaller than some** D^***.**

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

2^o. 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

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

- 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 $\nabla_x f$.)

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

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

Main result: critical diameter

Theorem

Given $k \geqslant 1$, let x^* be an ε-FSP in the **surrogate problem** (using $\bar{\lambda}_k$-weak convexity). Then x^* is also a 6ε-FSP for the **initial problem**, provided that

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

Moreover, for $k = 0$ it suffices that $\mu \mathrm{D} \lesssim \varepsilon$.

- In other words, for $k \geqslant 1$ the surrogate works as long as $\mathrm{D} \lesssim_k \bar{\mathrm{D}}$ with

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

- For $k = 0$ we have $\bar{\mathrm{D}} = \frac{\varepsilon}{\mu}$, same as for $k = 1$ except for a constant factor $\frac{1}{\mu} \leqslant \frac{1}{\min\{\mu, \sqrt{\lambda \rho_1}\}}$. Modest deterioration, and only if $\mu \geqslant \sqrt{\lambda \rho_1}$.
- For $k = 2$ we have $\bar{\mathrm{D}} = \frac{\varepsilon^{2/3}}{(\lambda \rho_k)^{1/3}}$, independent from μ whenever $\varepsilon \ll 1$.

Proof: μ-independent bound

Proposition 1. Moreau envelope gradients for φ and $\hat{\varphi}$ are *uniformly close*:

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

Proof:

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

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

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

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

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

Proof: μ-independent bound (cont'd)

Proposition 1. Moreau envelope gradients for φ and $\hat{\varphi}$ are *uniformly close*:

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

Proof:

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

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

Summing the two inequalities results in

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

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

Proof: μ-dependent bound

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

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

Proof: (assuming $X = E_x$ and $k \geqslant 1$ for simplicity)

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

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

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

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

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

Proof: μ-dependent bound (cont'd)

Rearranging, we get

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

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

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

Using the first inequality,

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

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

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

whence

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

Proof: μ-dependent bound (cont'd)

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

5*^o*. Applying Cauchy-Schwarz twice we get

$$\begin{aligned}(\bar\lambda_k\|\hat x^+ - x^+\|)^2 &\leqslant 4\varepsilon^2 + 4\bar\lambda_k\left[\langle \hat\xi, \hat x^+ - x^+\rangle - \tfrac{1}{4}\bar\lambda_k\|\hat x^+ - x^+\|^2\right] + 4\|\hat\xi - \xi\|^2\\ &\leqslant 4\varepsilon^2 + 4\|\hat\xi\|^2 + 4\|\hat\xi - \xi\|^2.\end{aligned}$$

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

$$(\bar\lambda_k\|\hat x^+ - x^+\|)^2 \leqslant 8\varepsilon^2 + 4\|\hat\xi - \xi\|^2,$$

6*^o*. It remains to bound $\|\hat\xi - \xi\|^2$. By the "subgradient of maximum" rule:

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

Also, we can choose $\xi = \nabla_x f(\hat x^+, y^*)$ for $y^* \in \text{Argmax}_{y\in Y}\, f(\hat x^+, y)$.
Whence by convexity of the norm:

$$\begin{aligned}&\|\hat\xi_X^+ - \xi^+\| \leqslant \max_{y\in Y}\|\nabla_x\hat f_k(\hat x^+, y) - \nabla_x f(\hat x^+, y^*)\|\\ &\leqslant \|\nabla_x f(\hat x^+, \bar y) - \nabla_x f(\hat x^+, y^*)\| + \|\nabla_x f(\hat x^+, y^*) - \nabla_x\hat f_k(\hat x^+, y^*)\|.\\ &\leqslant \mu D + \frac{2\sigma_k}{k!}. \qquad \blacksquare\end{aligned}$$

"Honest" Hessian approximation

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

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

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

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