Program:

- The notion of a min-max problem
- Game-theoretic interpretation
- Some applications

- Convex-concave case, saddle points, minimax thm & duality. duality gap.
  - Variational inequality viewpoint
  - "Baseline": gradient descent-ascent (GDA, a.k.a. Saddle-Point Mirror Descent).
    - Stochastic Oracles & ~~Non-Euclidean structure~~.
  - Extragradient method (a.k.a. Mirror Prox):
    - Motivation & Key Idea
    - $O(\frac{1}{T})$ convergence result
  - ~~Alternative approach: Nesterov's smoothing~~
  - Sampling for bilinear problems.

- Convex-strongly-concave case

- Noncouvex-[strongly-] concave case: ?
  - Algorithm for finding 1st-order stationary points
  - Convergence rates

Convexity

Let $X \subseteq \mathbb{R}^d$ be a convex body (set with nonempty interior)

Def.: $f: X \to \mathbb{R}$ is (closed) convex if $\forall x \in X$ one has:

$$f(y) \geq f(x) + \langle f'(x), y - x \rangle$$

for some vector $f'(x)$ called <u>subgradient</u> of $f(\cdot)$ at x, and any $y \in X$

Def.: <u>Subdifferential</u>: $\partial f(x) = \{ \text{all such } f'(x) \}$.

- If f is differentiable at x, then $\partial f(x) = \{ \nabla f(x) \}$
- If $f(x) = \max_{y \in Y} F(x,y)$ where $F(\cdot, y)$ is convex and differentiable for each $y \in Y$, then $\partial f(x)$ is the convex hull of the "active" gradients: $\partial f(x) = \operatorname{Conv}\left(\{ \nabla_x F(x, y^*) , y^* \in \operatorname{Argmax}_{y \in Y} F(x,y) \} \right)$

In particular, $\nabla f(x) = \nabla_x F(x, y^*(x))$ if $y^*(x)$ is a unique maximizer.

Black-Box model of convex optimization:

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

- $X \subseteq E_x$ is a known convex body $\Longrightarrow$ can project on it
- $f : X \to \mathbb{R}$ is an unknown function from certain class, available via a 1st-order oracle that returns $f'(x)$ at any $x \in X$.

Problem Classes:

- Convex and B-Lipshitz: $\|f'(x)\| \leq B$.
- Convex and L-smooth (differentiable with L-lipschitz gradient):

$$0 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2}\|y - x\|^2$$

- L-smooth and λ-strongly convex: $(\lambda > 0)$:

$$\frac{\lambda}{2}\|y - x\|^2 \leq f(y) - f(x) - \langle \nabla f(x), y - x \rangle \leq \frac{L}{2}\|y - x\|^2$$

$\boxed{\text{Black-Box model of convex optimization:}}$

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

- $X \subseteq E_x$ is a known convex body $\Longrightarrow$ can project on it
- $f : X \to \mathbb{R}$ is an unknown function from certain class, available via a 1st-order oracle that returns $f'(x)$ at any $x \in X$.

Accuracy Measures: $f(\hat{x}) - f^*$ where $f^* := \min_{x \in X} f(x)$ – objective error.

First-order algorithms construct a sequence of points $(\hat{x}_1, \dots, \hat{x}_T)$ such that $f(\hat{x}_T) - f^* \le \varepsilon(T)$ for $\varepsilon(T)$ depending on the problem parameters.

#1: $\varepsilon(T) = O(1)\dfrac{BD}{\sqrt{T}}$ for B-Lipschitz & convex and $\operatorname{diam}(X) = D$, with (projected) subgradient method.

#2: $\varepsilon(T) = O(1)\dfrac{LD^2}{T}$ for L-smooth & convex, with (projected) grad. descent.

Minimax Problems

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

- Game-theoretic interpretation: zero-sum game between Min and Max. $f(x,y)$ is the payoff of Max and the loss of Min when they choose a pair of strategies (x,y).

Who goes first, Min or Max?

- Primal & Dual problems:

$$(P) \quad \min_{x\in X} \left\{ \varphi(x) := \max_{y\in Y} f(x,y) \right\} \geq \max_{y\in Y} \left\{ \psi(y) := \min_{x\in X} f(x,y) \right\} \quad (D)$$

- Weak duality: $(P) \geq (D)$ "Going the second is better"

Proof: Let $x^* \in \operatorname{Argmin} \varphi(x)$, $y^* \in \operatorname{Argmax} \psi(y)$. Then

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

Examples

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

control parameter · nominal input · cost function

1° **Robust system design:**

E.g.: Adversarially robust training

2° **Linear Regression in ℓ_p norms:**

$$\min_{x\in X} \underbrace{\|Ax-b\|_p}_{\varphi(x)} = \min_{x\in X} \max_{\|y\|_q\le 1} \langle y, Ax-b\rangle$$

Bilinear SPP;

Y = the unit dual norm ball $\left(\frac{1}{p}+\frac{1}{q}=1\right)$

3° **Minimization of a maximum of m functions:**

$$\min_{x\in X} \max_{i\in\{1,..,m\}} g_i(x) = \min_{x\in X} \max_{y\in\Delta_m} \langle y, G(x)\rangle$$

where Δ_m is the standard simplex; $G(x)=\begin{bmatrix} g_1(x)\\ \vdots \\ g_m(x)\end{bmatrix}$.

Objective is linear ($\Rightarrow$ concave) in y;

- smooth if all $g_i(x)$ are smooth, and convex if $g_i(x)$ are so.

4° **Finite-sum minimization:** $\min_{x\in X} \sum_{i=1}^n \ell(x,z_i)$ where $\ell(x,z)\ge 0$

Then $\sum_{i=1}^n \ell(x,z_i) = \max_{y\in[0,1]^n} \langle y, L_n(x)\rangle$ where $L_n(x)=\begin{bmatrix} \ell(x,z_1)\\ \vdots \\ \ell(x,z_n)\end{bmatrix}$.

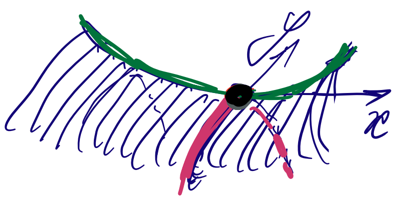
Convex - Concave Saddle-Point Problems

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

where $X \subseteq E_x$ and $Y \subseteq E_y$ are convex sets in their spaces;
$f(\cdot, y)$ is convex $\forall y \in Y$; $f(x, \cdot)$ is concave $\forall x \in X$.

⊙ Strong duality (a.k.a. Minimax Theorem; M. Sion 1958)

If f is convex-concave, and X or Y is compact, then $\varphi^* = \psi^*$
and there exists a saddle point — a point (x^*, y^*) for which



$$\varphi^* = \varphi(x^*) = f(x^*, y^*) = \psi(y^*) = \psi^*.$$

$$f(x^*, y) \le f(x^*, y^*) \le f(x, y^*) \quad \forall x \in X, \ \forall y \in Y$$

Also works with non-compact sets, but strong convexity or concavity

Convex-Concave Saddle-Point Problems:

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

$X \subseteq E_X$ and $Y \subseteq E_Y$ are convex sets in their spaces;
$f(\cdot, y)$ is convex $\forall y \in Y$; $f(x, \cdot)$ is concave $\forall x \in X$.

Task: Find a saddle point (x^*, y^*):

$$\boxed{\varphi^* = \varphi(x^*) = f(x^*, y^*) = \psi(y^*) = \psi^*.}$$

Duality Gap: How good is a candidate solution $(\hat{x}, \hat{y}) \in X \times Y$:

$$\mathrm{Gap}(\hat{x}, \hat{y}) := \varphi(\hat{x}) - \psi(\hat{y}) = \underbrace{\varphi(\hat{x}) - \cancel{\varphi^*}}_{\text{Primal gap} \geq 0} + \underbrace{\cancel{\psi^*} - \psi(\hat{y})}_{\text{Dual gap} \geq 0}$$

- Thus, $\mathrm{Gap}(\hat{x}, \hat{y}) \leq \varepsilon$ guarantees that $\hat{x}$ is ε-suboptimal for (P), and $\hat{y}$ is ε-suboptimal for (D).
- Typically we might upper-bound or even compute it – certificate.

Black-Box model for CCSPP:

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

- X and Y known (can project)
- $f : X \times Y \to \mathbb{R}$ is available via 1st-order oracle:

$$(x,y) \mapsto [f'_x(x,y); -f'_y(x,y)]$$

subgradient of $f(\cdot,y)$ at x supergradient of $f(x,\cdot)$ at y

Goal: In T queries, find $(\hat{x}_T, \hat{y}_T) \in X \times Y$ such that $\mathrm{Gap}(\hat{x}_T, \hat{y}_T) \le \varepsilon_T$.

Certificate:

$$\mathrm{Gap}(\hat{x},\hat{y}) = \varphi(\hat{x}) - \psi(\hat{y}) \le f(\hat{x},\hat{y}) - \min_{x \in X} f(x,\hat{y}) + \max_{y \in Y} f(\hat{x},y) - f(\hat{x},\hat{y})$$

$$\le \sup_{x \in X} \langle f'_x(\hat{x},\hat{y}), \hat{x} - x \rangle + \sup_{y \in Y} \langle -f'_y(\hat{x},\hat{y}), \hat{y} - y \rangle$$

The RHS can be bounded whenever we have linear maximization oracles (LMO) for X, Y, which is a very weak assumption.

CCSPP as a monotone variational inequality (MVI)

$\boxed{\min_{x \in X} \max_{y \in Y} f(x,y)}$ $\quad z = (x,y) \mapsto F(z) = [f'_x(x,y); -f'_y(x,y)]$

$z \in Z = X \times Y$ $\quad$ vector field

Optimality condition: $z^* = (x^*, y^*)$ is a SP iff $\langle F(z^*), z - z^* \rangle \geq 0 \ \forall z \in Z$.

Variational Inequality: for a vector field $F(\cdot)$ on Z, find z^* such that

$$\boxed{\langle F(z^*), z - z^* \rangle \geq 0 \quad \forall z \in Z}$$

- For any SPP (not necessarily a convex-concave one), a solution to the corresponding (strong) VI is a first-order Nash equilibrium ("SPP for the locally linearized objective")
- For CCSPPs, the operator $F(\cdot)$ is monotone: one has that

$$\langle F(z') - F(z), z' - z \rangle \geq 0 \qquad \forall z', z \in Z.$$

(Compare with the case of a subgradient of a convex function)

CCSPP as a monotone variational inequality (MVI)

$$\boxed{\min_{x \in X} \max_{y \in Y} f(x,y)} \qquad \begin{array}{c} z = (x,y) \\ \cap \\ Z = X \times Y \end{array} \longmapsto \underset{\text{vector field}}{F(z)} = \left[f'_x(x,y) \,;\, -f'_y(x,y) \right]$$

<u>Optimality condition:</u> $z^* = (x^*, y^*)$ is a SP iff $\langle F(z^*), z - z^* \rangle \geq 0 \;\; \forall z \in Z$.

<u>Variational Inequality:</u> for a vector field $F(\cdot)$ on Z, find z^* such that

$$\boxed{\langle F(z^*), z - z^* \rangle \geq 0 \quad \forall z \in Z}$$

- Recall that $\mathrm{Gap}(\hat{z}) = \varphi(\hat{x}) - \psi(\hat{y}) \leq \sup\limits_{z \in Z} \langle F(\hat{z}), \hat{z} - z \rangle$.
- ε-approximate VI solution, i.e. $\hat{z} \in Z$ such that

$$\boxed{\langle F(\hat{z}), \hat{z} - z \rangle \leq \varepsilon \quad \forall z \in Z}$$

  is also an ε-approximate saddle point, i.e. $\mathrm{Gap}(\hat{z}) \leq \varepsilon$.
- Hence, we can focus on solving MVIs (approximately).

CCSPP with a smooth objective (MVIs with continuous operator

- The common feature in all examples is that a non-smooth (primal) minimization problem $\min_{x \in X} \{\varphi(x) := \max_{y \in Y} f(x,y)\}$ translates to a smooth CCSPP, provided that we have a max-type representation for $\varphi(x)$.
- Recall that non-smooth (Lipshitz) convex minimization can be done in $O(\frac{1}{\sqrt{T}})$ in the black-box model — oracle $x \mapsto \varphi'(x)$ — via SGM. But here we "go out of the black box" in terms of φ: new oracle is $(x,y) \mapsto [\nabla_x f(x,y); -\nabla_y f(x,y)]$. Hopefully, we can do faster?
- For smooth convex minimization, projected gradient descent converges as $O(\frac{1}{T})$, and Nesterov as $O(\frac{1}{T^2})$ which is optimal.

▶ Can we generalize some of these results to MVIs w/ continuous operators?

- Yes, we can attain $\mathrm{Gap}(\bar{x}_T, \bar{y}_T) = O(\frac{1}{T})$ — and this is optimal already for bilinear CCSPPs.

Gradient Descent-Ascent (GDA) & the Looping Problem

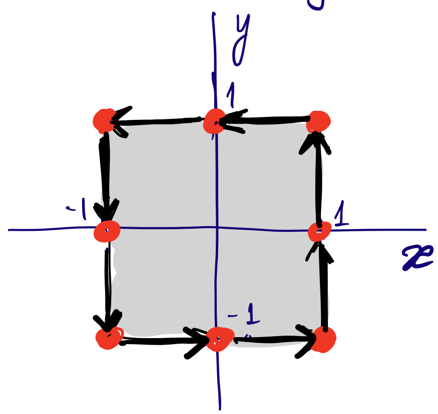
(Projected) GDA is a naive adaptation of (projected) gradient descent

Problem	Oracle	Method	
$\min_{x \in X} g(x)$, C^1, convex	$x \mapsto \nabla g(x)$	$x_{t+1} = \Pi_X\left[x_t - \frac{1}{L}\nabla g(x_t)\right]$	(PGD)
$\min_{x \in X} \max_{y \in Y} f(x,y)$ f is C^1 and convex-concave	$z=(x,y) \to F(z)$ where $F(z) = \begin{bmatrix} \nabla_x f(x,y) \\ -\nabla_y f(x,y) \end{bmatrix}$	$z_{t+1} = \Pi_Z\left[z_t - \frac{1}{L} F(z_t)\right]$ $\Longleftrightarrow$ $\begin{bmatrix} x_{t+1} \\ y_{t+1} \end{bmatrix} = \begin{bmatrix} \Pi_X\left[x_t - \frac{1}{L}\nabla_x f(x,y)\right] \\ \Pi_Y\left[y_t + \frac{1}{L}\nabla_y f(x,y)\right] \end{bmatrix}$	(PGDA)

Unfortunately, vanilla PGDA can "go in circles" even for bilinear SPPs

$$\min_{x \in [-1,1]} \max_{y \in [-1,1]} xy$$

Unique SP: $(x^*, y^*) = (0,0)$



$$F(x,y) = \begin{bmatrix} y \\ -x \end{bmatrix}$$

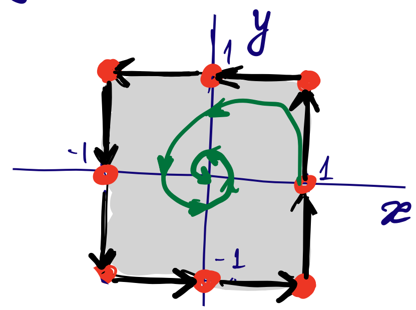
$$\begin{bmatrix} x_{t+1} \\ y_{t+1} \end{bmatrix} = \begin{bmatrix} \Pi_{[-1,1]}\left[x_t - y_t\right] \\ \Pi_{[-1,1]}\left[y_t + x_t\right] \end{bmatrix}$$

Ergodic Convergence of GDA

- We can enforce PGDA convergence via averaging & smaller steps.

$$\hat{z}_T = \frac{1}{T}\sum_{t=1}^{T} z_t$$

where $z_{t+1} = z_t - \gamma F(z_t)$



Theorem: Assume $\max\{\mathrm{diam}(X), \mathrm{diam}(Y)\} \le D$, and $\sup_{z \in Z} \|F(z)\| \le B$.

Then PGDA with $\gamma \simeq \frac{D}{B\sqrt{T}}$ ensures $\boxed{\mathrm{Gap}(\hat{z}_T) \lesssim \frac{BD}{\sqrt{T}}}$

- More precisely, if $\mathrm{diam}(X) \le D_x$, $\mathrm{diam}(Y) \le D_y$, and F is (B_x, B_y)-Lipschitz, we can attain $\mathrm{Gap}(\hat{z}_T) \le \frac{B_x D_x + B_y D_y}{\sqrt{T}}$

- In the C^1 case with $D_x f(z_x) = D_y f(z_y) = 0$ for some z_x, z_y:

$$\mathrm{Gap}(\hat{z}_T) \lesssim \frac{L_{xy} D_x D_y + L_{xx} D_x^2 + L_{yy} D_y^2}{\sqrt{T}}.$$

Note that $L_{xx} = L_{yy} = 0$ in the bilinear case.

GDA with a stochastic oracle

$$\left.\begin{aligned} \|F(x';y) - F(x,y)\| &\le L_{xx}\|x'-x\|; \\ \|F(x';y) - F(x,y)\| &\le L_{xy}\|y'-y\|; \\ \|F(x,y') - F(x,y)\| &\le L_{yy}\|y'-y\|. \end{aligned}\right] \text{for all } x', x \in X; \quad y', y \in Y.$$

$$\mathrm{Gap}(\bar{z}_T) \lesssim \frac{L_{xy} D_x D_y + L_{xx} D_x^2 + L_{yy} D_y^2}{\sqrt{T}},$$ with $L_{xx} = L_{yy} = 0$ in the bilinear case

- Stochastic oracle: $\tilde{F}(z) = F(z) + \xi(z)$ where $\xi(z) = \begin{bmatrix} \xi_x(z) \\ \xi_y(z) \end{bmatrix}$

- satisfies $\mathbb{E}[\xi(z)] = 0$ and $\mathbb{E}\begin{bmatrix} \|\xi_x(z)\|^2 \\ \|\xi_y(z)\|^2 \end{bmatrix} \le \begin{bmatrix} \sigma_x^2 \\ \sigma_y^2 \end{bmatrix}$ for all $z \in Z$

$$\mathrm{Gap}(\bar{z}_T) \lesssim [\ \dots\ ] + \frac{\sigma_x D_x + \sigma_y D_y}{\sqrt{T}}$$

where $[\dots]$ is the "deterministic" error (as if $\sigma_x = \sigma_y = 0$).

Extragradient method.

$\odot$ f is convex-concave and L-smooth (i.e. F is L-Lipschitz).

Then $\quad Gap(\bar{z}_t) \leq \dfrac{LD^2}{T} + \dfrac{\sigma_x D_x + \sigma_y D_y}{\sqrt{T}}$.

Key Ideas:

(1) Consider the following "conceptual method" (implicit update):

$$z_{t+1} = \Pi_Z \left[z_t - \frac{c}{L} F(z_{t+1}) \right]$$ with some $c \geq 0$ (to be chosen later).

Instead of mimicking PGD, we mimick the proximal point method (PPM).

$\odot$ It turns out that the analysis of PPM generalizes directly, and thus "conceptual method" converges as $\dfrac{LD^2}{T}$ when $c < 1$.

How to implement the conceptual method?

Extragradient method: basic version

[Understood since 1970s]

Conceptual update $z_{t+1} = \Pi_Z\left[z_t - \frac{c}{L} F(z_{t+1})\right]$ is a fixed point of the operator $P_{z_t}(z) = \Pi_Z\left[z_t - \frac{c}{L} F(z)\right]$. (By definition!)

Lemma. Assume $F(\cdot)$ is L-Lipschitz, and $c<1$. Then $P_{\bar z}(\cdot)$ is a contraction.

Proof: $\left\| P_{\bar z}(z') - P_{\bar z}(z) \right\| = \left\| \Pi_Z\left[\bar z - \frac{c}{L} F(z')\right] - \Pi_Z\left[\bar z - \frac{c}{L} F(z)\right] \right\|$

[by the non-expansiveness of projections] $\leq \left\| \cancel{\bar z} - \frac{c}{L} F(z') - \cancel{\bar z} + \frac{c}{L} F(z) \right\| \leq c \|z' - z\|$ ∎

Hence, given $\tilde z_t \approx z_t$ we can approximate z_{t+1} with K iterations of the form:

$$\tilde z_t \mapsto z_t^{(1)} = P_{z_t}(z_t) \mapsto z_t^{(2)} = P_{z_t}(z_t^{(1)}) \mapsto \dots \tilde z_{t+1} := z_t^{(K)} = P_{z_t}(z_t^{(K-1)})$$

with a linear (i.e. exponentially fast) convergence towards z_{t+1}.

This is "K-lookahead extragradient" update, and it gives the desired convergence rate.

Extragradient method with $K=2$

⊙ Note that $K=1$ gives PGDA. A. Nemirovski discovered that $K=2$ suffices!

Theorem: (Nemirovski, 2003) Let $c \leq \frac{1}{\sqrt{2}}$. Then the following method converges as $O(\frac{1}{T})$:

Inputs: z_t

$$z_{t+\frac{1}{2}} = \Pi_{\mathcal{Z}}\left[z_t - \frac{c}{L} F(z_t)\right]$$

Return: $z_{t+1} = \Pi_{\mathcal{Z}}\left[z_t - \frac{c}{L} F(z_{t+\frac{1}{2}})\right]$

(Convergence is for z_T, i.e. in terms of the last iterate.)

The proof is non-trivial, and heavily utilizes monotonicity and the algebraic structure of Bregman divergences.

BONUS: Weak/Strong MVIs, Certificates.

Monotone VIs, weak and strong solutions.

Strong solution: $\langle F(z^*), z - z^* \rangle \geq 0 \quad \forall z \in \mathcal{Z}.$

Monotonicity: $\langle F(z) - F(z^*), z - z^* \rangle \geq 0$

$\Rightarrow$ z^* is also a weak solution: $\langle F(z), z - z^* \rangle \geq 0 \quad \forall z \in \mathcal{Z}.$

- Moreover, in a monotone VI with a continuous operator, any weak solution is also a strong solution (Minty's lemma).

  Proof: Let $x_s = (1-s)x^* + s x$ for $s \in [0,1]$, where x^* is a weak solution.
  Then $0 \leq \langle F(x_s), x_s - x^* \rangle = s \langle F(x_s), x - x^* \rangle$, so $\langle F(x_s), x - x^* \rangle \geq 0$.
  Take $s \to 0$. ∎

- Hence, for CCSPPs with C_1-smooth objectives, SPs are weak VI solutions.

Certificates

- Certificate: let $\begin{bmatrix} \hat{x}_T \\ \hat{y}_T \end{bmatrix} = \sum_{t=1}^{T} w_t \begin{bmatrix} x_t \\ y_t \end{bmatrix}$ where $x_1, \dots, x_T$ are the iterates, $(w_1, \dots, w_T)$ are the convex weights.

$$\mathrm{Gap}(\hat{x}_T, \hat{y}_T) = \varphi(\hat{x}_T) - \psi(\hat{y}_T) \le \sum_{t=1}^{T} w_t \left[\varphi(x_t) - f(x_t, y_t) + f(x_t, y_t) - \psi(y_t) \right]$$

$$\le \sum_{t=1}^{T} w_t \left\{ \langle f'_y(x_t, y_t), y_t^{\varphi} - y_t \rangle + \langle f'_x(x_t, y_t), x_t - x_t^{\psi} \rangle \right.$$

Can upper-bound the RHS by maximizing linear functions on X and Y.