Optimization for Machine Learning

Lecture 15: Minimax problems: convex-concave

6.881: EECS, MIT

Suvrit Sra
Massachusetts Institute of Technology

13 Apr, 2021

$$\inf_x \sup_y \ \phi(x, y)$$

Minimax problems

- Minimax theory treats problems involving a combination of minimization and maximization

Minimax problems

- Minimax theory treats problems involving a combination of minimization and maximization
- Let $\mathcal{X}, \mathcal{Y}$ be nonempty sets; and $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$

Minimax problems

- Minimax theory treats problems involving a combination of minimization and maximization
- Let $\mathcal{X}, \mathcal{Y}$ be nonempty sets; and $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$
- First **inf** over $x \in \mathcal{X}$, then **sup** over $y \in \mathcal{Y}$:

$$\sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} \phi(x, y)$$

Minimax problems

- Minimax theory treats problems involving a combination of minimization and maximization
- Let $\mathcal{X}$, $\mathcal{Y}$ be nonempty sets; and $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$
- First **inf** over $x \in \mathcal{X}$, then **sup** over $y \in \mathcal{Y}$:

$$\sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} \phi(x, y)$$

- First **sup** over $y \in \mathcal{Y}$, then **inf** over $x \in \mathcal{X}$:

$$\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \phi(x, y)$$

When are "inf sup" and "sup inf" equal?

Weak minimax (*cf.* weak duality)

Theorem. Let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$. Then,

$$\sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} \phi(x, y) \quad \leq \quad \inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \phi(x, y)$$

Weak minimax (*cf.* weak duality)

Theorem. Let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$. Then,

$$\sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} \phi(x, y) \quad \leq \quad \inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \phi(x, y)$$

Proof:

$$x, y, \quad \inf_{x' \in \mathcal{X}} \phi(x', y) \quad \leq \quad \phi(x, y)$$

Weak minimax (*cf.* weak duality)

Theorem. Let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$. Then,

$$\sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} \phi(x, y) \quad \leq \quad \inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \phi(x, y)$$

Proof:

$$x, y, \quad \inf_{x' \in \mathcal{X}} \phi(x', y) \quad \leq \quad \phi(x, y)$$

$$x, y, \quad \inf_{x' \in \mathcal{X}} \phi(x', y) \quad \leq \quad \sup_{y' \in \mathcal{Y}} \phi(x, y')$$

Weak minimax (*cf.* weak duality)

Theorem. Let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$. Then,

$$\sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} \phi(x, y) \quad \leq \quad \inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \phi(x, y)$$

Proof:

$$x, y, \quad \inf_{x' \in \mathcal{X}} \phi(x', y) \quad \leq \quad \phi(x, y)$$

$$x, y, \quad \inf_{x' \in \mathcal{X}} \phi(x', y) \quad \leq \quad \sup_{y' \in \mathcal{Y}} \phi(x, y')$$

$$\forall x, \quad \sup_{y \in \mathcal{Y}} \inf_{x' \in \mathcal{X}} \phi(x', y) \quad \leq \quad \sup_{y' \in \mathcal{Y}} \phi(x, y')$$

Weak minimax (*cf.* weak duality)

Theorem. Let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$. Then,

$$\sup_{y\in\mathcal{Y}} \inf_{x\in\mathcal{X}} \phi(x,y) \quad \le \quad \inf_{x\in\mathcal{X}} \sup_{y\in\mathcal{Y}} \phi(x,y)$$

Proof:

$$\begin{aligned}
x, y, \quad & \inf_{x'\in\mathcal{X}} \phi(x',y) && \le && \phi(x,y) \\
x, y, \quad & \inf_{x'\in\mathcal{X}} \phi(x',y) && \le && \sup_{y'\in\mathcal{Y}} \phi(x,y') \\
\forall x, \quad & \sup_{y\in\mathcal{Y}} \inf_{x'\in\mathcal{X}} \phi(x',y) && \le && \sup_{y'\in\mathcal{Y}} \phi(x,y') \\
\implies \quad & \sup_{y\in\mathcal{Y}} \inf_{x'\in\mathcal{X}} \phi(x',y) && \le && \inf_{x\in\mathcal{X}} \sup_{y'\in\mathcal{Y}} \phi(x,y').
\end{aligned}$$

Weak minimax (*cf.* weak duality)

Theorem. Let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm\infty\}$. Then,

$$\sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} \phi(x, y) \quad \le \quad \inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \phi(x, y)$$

Proof:

$$\begin{aligned}
x, y, \quad & \inf_{x' \in \mathcal{X}} \phi(x', y) && \le \quad \phi(x, y) \\
x, y, \quad & \inf_{x' \in \mathcal{X}} \phi(x', y) && \le \quad \sup_{y' \in \mathcal{Y}} \phi(x, y') \\
\forall x, \quad & \sup_{y \in \mathcal{Y}} \inf_{x' \in \mathcal{X}} \phi(x', y) && \le \quad \sup_{y' \in \mathcal{Y}} \phi(x, y') \\
\Longrightarrow \quad & \sup_{y \in \mathcal{Y}} \inf_{x' \in \mathcal{X}} \phi(x', y) && \le \quad \inf_{x \in \mathcal{X}} \sup_{y' \in \mathcal{Y}} \phi(x, y').
\end{aligned}$$

Exercise: Show that weak duality is follows from above minimax inequality.

Hint: Use $\phi = \mathcal{L}$ (Lagrangian), and suitably choose y.

Saddle values, strong minimax

- If "inf sup" = "sup inf", common value **saddle-value**
- Value exists if there is a **saddle-point**, i.e., pair (x^*, y^*)

$$\phi(x, y^*) \geq \phi(x^*, y^*) \geq \phi(x^*, y) \quad \text{for all } x \in \mathcal{X}, y \in \mathcal{Y}.$$

Saddle values, strong minimax

- If "inf sup" = "sup inf", common value **saddle-value**
- Value exists if there is a **saddle-point**, i.e., pair (x^*, y^*)

$$\phi(x, y^*) \geq \phi(x^*, y^*) \geq \phi(x^*, y) \quad \text{for all } x \in \mathcal{X}, y \in \mathcal{Y}.$$

- Writing $f(x) := \sup_y \phi(x, y)$ and $g(y) := \inf_x \phi(x, y)$, we have

$$f(x^*) = \inf_{x \in \mathcal{X}} f(x) = \sup_{y \in \mathcal{Y}} g(y) = g(y^*)$$

- That is, ***strong minimax*** holds:

$$f(x^*) = \phi(x^*, y^*) = g(y^*).$$

Strong minimax

Def. Let ϕ be as before. Pair (x^*, y^*) is a saddle-point of ϕ **iff** the infimum in the expression

$$\inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \phi(x, y)$$

is **attained** at x^*, and the supremum in the expression

$$\sup_{y \in \mathcal{Y}} \inf_{x \in \mathcal{X}} \phi(x, y)$$

is **attained** at y^*, and these two extrema **are equal**.

Strong minimax

Def. Let ϕ be as before. Pair (x^*, y^*) is a saddle-point of ϕ **iff** the infimum in the expression

$$\inf_{x\in\mathcal{X}} \sup_{y\in\mathcal{Y}} \phi(x,y)$$

is **attained** at x^*, and the supremum in the expression

$$\sup_{y\in\mathcal{Y}} \inf_{x\in\mathcal{X}} \phi(x,y)$$

is **attained** at y^*, and these two extrema **are equal**.

$$x^* \in \operatorname*{argmin}_{x\in\mathcal{X}} \max_{y\in\mathcal{Y}} \phi(x,y), \qquad y^* \in \operatorname*{argmax}_{y\in\mathcal{Y}} \min_{x\in\mathcal{X}} \phi(x,y).$$

Strong minimax

♠ Classes of problems "dual" to each other can be generated by studying classes of functions ϕ

Strong minimax

- Classes of problems "dual" to each other can be generated by studying classes of functions ϕ
- **More interesting question:** Starting from the primal problem over $\mathcal{X}$, how to introduce a space $\mathcal{Y}$ and a "useful" function ϕ on $\mathcal{X} \times \mathcal{Y}$ so that we have a saddle-point?

Strong minimax

- ♠ Classes of problems "dual" to each other can be generated by studying classes of functions ϕ
- ♠ **More interesting question:** Starting from the primal problem over $\mathcal{X}$, how to introduce a space $\mathcal{Y}$ and a "useful" function ϕ on $\mathcal{X} \times \mathcal{Y}$ so that we have a saddle-point?

Sufficient conditions for saddle-point

- ▶ Function ϕ is continuous, and
- ▶ It is ***convex-concave***, i.e., $\phi(\cdot, y)$ convex for every $y \in \mathcal{Y}$, and $\phi(x, \cdot)$ concave for every $x \in \mathcal{X}$; and
- ▶ Both $\mathcal{X}$ and $\mathcal{Y}$ are convex; one of them is compact.

Strong minimax

- Classes of problems "dual" to each other can be generated by studying classes of functions ϕ
- **More interesting question:** Starting from the primal problem over $\mathcal{X}$, how to introduce a space $\mathcal{Y}$ and a "useful" function ϕ on $\mathcal{X} \times \mathcal{Y}$ so that we have a saddle-point?

Sufficient conditions for saddle-point

- Function ϕ is continuous, and
- It is ***convex-concave***, i.e., $\phi(\cdot, y)$ convex for every $y \in \mathcal{Y}$, and $\phi(x, \cdot)$ concave for every $x \in \mathcal{X}$; and
- Both $\mathcal{X}$ and $\mathcal{Y}$ are convex; one of them is compact.
- (More generally: ϕ is appropriately semicontinuous and quasiconvex-quasiconcave with convex $\mathcal{X}, \mathcal{Y}$)

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda\|x\|_1.$$

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda\|x\|_1.$$

$$\|x\|_1 = \max\{x^T v \mid \|v\|_\infty \leq 1\}$$
$$\|x\|_2 = \max\{x^T u \mid \|u\|_2 \leq 1\}.$$

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda\|x\|_1.$$

$$\|x\|_1 = \max\{x^T v \mid \|v\|_\infty \leq 1\}$$
$$\|x\|_2 = \max\{x^T u \mid \|u\|_2 \leq 1\}.$$

Saddle-point formulation

$$p^* \quad = \quad \min_x \max_{u,v} \left\{ u^T(b - Ax) + v^T x \mid \|u\|_2 \leq 1, \quad \|v\|_\infty \leq \lambda \right\}$$

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda\|x\|_1.$$

$$\|x\|_1 = \max\{x^T v \mid \|v\|_\infty \leq 1\}$$
$$\|x\|_2 = \max\{x^T u \mid \|u\|_2 \leq 1\}.$$

Saddle-point formulation

$$\begin{aligned} p^* &= \min_x \max_{u,v} \left\{ u^T(b - Ax) + v^T x \mid \|u\|_2 \leq 1, \;\; \|v\|_\infty \leq \lambda \right\} \\ &= \max_{u,v} \min_x \left\{ u^T(b - Ax) + x^T v \mid \|u\|_2 \leq 1, \;\; \|v\|_\infty \leq \lambda \right\} \end{aligned}$$

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda\|x\|_1.$$

$$\|x\|_1 = \max\{x^T v \mid \|v\|_\infty \leq 1\}$$
$$\|x\|_2 = \max\{x^T u \mid \|u\|_2 \leq 1\}.$$

Saddle-point formulation

$$\begin{aligned} p^* &= \min_x \max_{u,v} \left\{ u^T(b - Ax) + v^T x \mid \|u\|_2 \leq 1, \;\; \|v\|_\infty \leq \lambda \right\} \\ &= \max_{u,v} \min_x \left\{ u^T(b - Ax) + x^T v \mid \|u\|_2 \leq 1, \;\; \|v\|_\infty \leq \lambda \right\} \\ &= \max_{u,v} u^T b \qquad A^T u = v, \; \|u\|_2 \leq 1, \;\; \|v\|_\infty \leq \lambda \end{aligned}$$

Example: Lasso-like problem

$$p^* := \min_x \quad \|Ax - b\|_2 + \lambda\|x\|_1.$$

$$\|x\|_1 = \max\{x^T v \mid \|v\|_\infty \le 1\}$$
$$\|x\|_2 = \max\{x^T u \mid \|u\|_2 \le 1\}.$$

Saddle-point formulation

$$\begin{aligned}
p^* &= \min_x \max_{u,v} \left\{u^T(b - Ax) + v^T x \mid \|u\|_2 \le 1,\ \ \|v\|_\infty \le \lambda\right\} \\
&= \max_{u,v} \min_x \left\{u^T(b - Ax) + x^T v \mid \|u\|_2 \le 1,\ \ \|v\|_\infty \le \lambda\right\} \\
&= \max_{u,v} u^T b \qquad A^T u = v,\ \|u\|_2 \le 1,\ \ \|v\|_\infty \le \lambda \\
&= \max_u u^T b \qquad \|u\|_2 \le 1, \quad \|A^T v\|_\infty \le \lambda.
\end{aligned}$$

Theory & Algorithms

Convex-Concave SP problem

Convex-Concave Saddle Point Problem

$$\sigma^* := \inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \phi(x, y)$$

where $\phi(x, \cdot)$ is convex and $\phi(\cdot, y)$ is concave.

Convex-Concave SP problem

Convex-Concave Saddle Point Problem

$$\sigma^* := \inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \phi(x, y)$$

where $\phi(x, \cdot)$ is convex and $\phi(\cdot, y)$ is concave.

Primal-Dual pair of problems

$$\begin{aligned} \mathrm{Opt}(P) &:= \min_{x \in \mathcal{X}} f(x) = \sup_{y \in \mathcal{Y}} \phi(x, y), \\ \mathrm{Opt}(D) &:= \max_{y \in \mathcal{Y}} g(y) = \inf_{x \in \mathcal{X}} \phi(x, y). \end{aligned}$$

Convex-Concave SP problem

Convex-Concave Saddle Point Problem

$$\sigma^* := \inf_{x \in \mathcal{X}} \sup_{y \in \mathcal{Y}} \phi(x, y)$$

where $\phi(x, \cdot)$ is convex and $\phi(\cdot, y)$ is concave.

Primal-Dual pair of problems

$$\begin{aligned} \mathrm{Opt}(P) &:= \min_{x \in \mathcal{X}} f(x) = \sup_{y \in \mathcal{Y}} \phi(x, y), \\ \mathrm{Opt}(D) &:= \max_{y \in \mathcal{Y}} g(y) = \inf_{x \in \mathcal{X}} \phi(x, y). \end{aligned}$$

Assuming SP (x^*, y^*) exists, we have

$$\mathrm{Opt}(P) = \mathrm{Opt}(D) = \phi(x^*, y^*) = f(x^*) = g(y^*).$$

Judging solutions of the CCSP problem

Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Quantify accuracy of $z = (x, y)$ by the ***gap***

$$\epsilon_{\text{sp}}(z) := \sup_{q \in \mathcal{Y}} \phi(x, q) - \inf_{p \in \mathcal{X}} \phi(p, y) = f(x) - g(y).$$

Judging solutions of the CCSP problem

Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Quantify accuracy of $z = (x, y)$ by the ***gap***

$$\epsilon_{\text{sp}}(z) := \sup_{q \in \mathcal{Y}} \phi(x, q) - \inf_{p \in \mathcal{X}} \phi(p, y) = f(x) - g(y).$$

Let us rewrite this gap in a more revealing form

Judging solutions of the CCSP problem

Let $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Quantify accuracy of $z = (x, y)$ by the ***gap***

$$\epsilon_{\text{sp}}(z) := \sup_{q \in \mathcal{Y}} \phi(x, q) - \inf_{p \in \mathcal{X}} \phi(p, y) = f(x) - g(y).$$

Let us rewrite this gap in a more revealing form

$$\begin{aligned} f(x) - g(y) &= [f(x) - \text{Opt}(P)] + [\text{Opt}(D) - g(y)] \\ &= [f(x) - f(x^*)] + [g(y^*) - g(y)], \end{aligned}$$

i.e., sum of the primal and dual suboptimality.

Setting up Mirror-Descent for CC-SP

SP Operator: Let $\partial_x \phi(x, y)$ be subdifferential of $\phi(\cdot, y)$ at $x \in \mathcal{X}$.

Setting up Mirror-Descent for CC-SP

SP Operator: Let $\partial_x \phi(x, y)$ be subdifferential of $\phi(\cdot, y)$ at $x \in \mathcal{X}$.
Let $\partial_y[-\phi(x, y)]$ be subdiff of $-\phi(x, \cdot)$ at point $y \in \mathcal{Y}$.

Setting up Mirror-Descent for CC-SP

SP Operator: Let $\partial_x \phi(x, y)$ be subdifferential of $\phi(\cdot, y)$ at $x \in \mathcal{X}$. Let $\partial_y[-\phi(x, y)]$ be subdiff of $-\phi(x, \cdot)$ at point $y \in \mathcal{Y}$.

Subdiff: Let $\Phi(z) \equiv \Phi(x, y) = \partial_x \phi(x, y) \times \partial_y[-\phi(x, y)]$.
Exercise: Verify by definition that Φ is a monotone operator.

Setting up Mirror-Descent for CC-SP

SP Operator: Let $\partial_x \phi(x, y)$ be subdifferential of $\phi(\cdot, y)$ at $x \in \mathcal{X}$. Let $\partial_y[-\phi(x, y)]$ be subdiff of $-\phi(x, \cdot)$ at point $y \in \mathcal{Y}$.

Subdiff: Let $\Phi(z) \equiv \Phi(x, y) = \partial_x \phi(x, y) \times \partial_y[-\phi(x, y)]$.
Exercise: Verify by definition that Φ is a monotone operator.

Lemma O^*. A point z^* is an SP of ϕ **iff** for every selection $F(\cdot)$ of Φ (i.e., a vector field $F : \text{ri}(\mathcal{Z}) \to \mathbb{R}^d$ s.t., $F(z) \in \Phi(z)$ for every $z \in \text{ri}(\mathcal{Z})$) we have $\langle F(z), z - z^* \rangle \geq 0$ for all $z \in \text{ri}(\mathcal{Z})$.

Setting up Mirror-Descent for CC-SP

SP Operator: Let $\partial_x \phi(x, y)$ be subdifferential of $\phi(\cdot, y)$ at $x \in \mathcal{X}$. Let $\partial_y[-\phi(x, y)]$ be subdiff of $-\phi(x, \cdot)$ at point $y \in \mathcal{Y}$.

Subdiff: Let $\Phi(z) \equiv \Phi(x, y) = \partial_x \phi(x, y) \times \partial_y[-\phi(x, y)]$.
Exercise: Verify by definition that Φ is a monotone operator.

Lemma O^*. A point z^* is an SP of ϕ **iff** for every selection $F(\cdot)$ of Φ (i.e., a vector field $F : \mathrm{ri}(\mathcal{Z}) \to \mathbb{R}^d$ s.t., $F(z) \in \Phi(z)$ for every $z \in \mathrm{ri}(\mathcal{Z})$) we have $\langle F(z), z - z^* \rangle \geq 0$ for all $z \in \mathrm{ri}(\mathcal{Z})$.

Assumption: $\mathcal{Z}$ is bounded and ϕ is Lipschitz continuous on $\mathcal{Z}$ (in this case, $\mathrm{dom}\, \Phi = \mathcal{Z}$)

Mirror Descent Setup

Choose a norm $\|\cdot\|$ on $\mathcal{Z}$, and a ***Bregman divergence***

$$D_\omega(u,z) := \omega(u) - \omega(z) - \langle \omega'(z),\, u - z \rangle$$

that is strongly convex (in u) wrt the chosen norm.

Mirror Descent Setup

Mirror Descent Setup

Choose a norm $\|\cdot\|$ on $\mathcal{Z}$, and a ***Bregman divergence***

$$D_\omega(u,z) := \omega(u) - \omega(z) - \langle \omega'(z),\, u - z\rangle$$

that is strongly convex (in u) wrt the chosen norm.

(Bregman)-Prox-mapping

$$\mathrm{Prox}_z(\xi) := \operatorname*{argmin}_{u\in\mathcal{Z}}\; D_\omega(u,z) + \langle \xi,\, u\rangle$$

Mirror Descent Setup

Assumption: Subgradient-(selection) oracle: Given any $z = (x, y) \in \mathcal{Z}$, we can compute a vector $F(z) \in \Phi(x, y)$.

Mirror Descent Setup

Assumption: Subgradient-(selection) oracle: Given any $z = (x, y) \in \mathcal{Z}$, we can compute a vector $F(z) \in \Phi(x, y)$.

MD algorithm

1. Let $\gamma_t > 0$ for $t \geq 1$ be stepsizes

Mirror Descent Setup

Assumption: Subgradient-(selection) oracle: Given any $z = (x, y) \in \mathcal{Z}$, we can compute a vector $F(z) \in \Phi(x, y)$.

MD algorithm

1. Let $\gamma_t > 0$ for $t \geq 1$ be stepsizes
2. $z_1 = \mathrm{argmin}_{u \in \mathcal{Z}}\, \omega(u)$ *(initialization)*

Mirror Descent Setup

Assumption: Subgradient-(selection) oracle: Given any $z = (x, y) \in \mathcal{Z}$, we can compute a vector $F(z) \in \Phi(x, y)$.

MD algorithm

1. Let $\gamma_t > 0$ for $t \geq 1$ be stepsizes
2. $z_1 = \operatorname{argmin}_{u \in \mathcal{Z}} \omega(u)$ *(initialization)*
3. $z_{t+1} = \operatorname{Prox}_{z_t}(\gamma_t F(z_t))$ *(subgradient step)*

Mirror Descent Setup

Assumption: Subgradient-(selection) oracle: Given any $z = (x, y) \in \mathcal{Z}$, we can compute a vector $F(z) \in \Phi(x, y)$.

MD algorithm

1. Let $\gamma_t > 0$ for $t \geq 1$ be stepsizes
2. $z_1 = \operatorname{argmin}_{u \in \mathcal{Z}} \omega(u)$ *(initialization)*
3. $z_{t+1} = \operatorname{Prox}_{z_t}(\gamma_t F(z_t))$ *(subgradient step)*
4. $\bar{z}_t = \frac{\sum_{s=1}^{t} \gamma_s z_s}{\sum_{s=1}^{t} \gamma_s}$

Mirror Descent Setup

Assumption: Subgradient-(selection) oracle: Given any $z = (x, y) \in \mathcal{Z}$, we can compute a vector $F(z) \in \Phi(x, y)$.

MD algorithm

1. Let $\gamma_t > 0$ for $t \geq 1$ be stepsizes
2. $z_1 = \operatorname{argmin}_{u \in \mathcal{Z}} \omega(u)$ *(initialization)*
3. $z_{t+1} = \operatorname{Prox}_{z_t}(\gamma_t F(z_t))$ *(subgradient step)*
4. $\bar{z}_t = \frac{\sum_{s=1}^{t} \gamma_s z_s}{\sum_{s=1}^{t} \gamma_s}$ *(average iterate)*,

Recall: Mirror Descent Setups

- *Euclidean setup*: $\|\cdot\| = \|\cdot\|_2$, $\omega(x) = \frac{1}{2}x^T x$
- ℓ_1 *setup*: $\|\cdot\| = \|\cdot\|_1$, when $\mathcal{Z}$ a simplex, then $\omega(z) = \sum_i z_i \log z_i$
- ℓ_1 *setup*: $\|\cdot\| = \|\cdot\|_1$, when $\mathcal{Z}$ bounded (e.g., the unit ℓ_1-ball), one can set $\omega(z) = 2e \log n \sum_{i=1}^{n} |z_i|^{p(n)}$, where $p(n) = 1 + 1/2 \log n$.
- Many other examples,...

Take advantage of prob geometry; obtain faster FOMs

Convergence rate

Theorem. Assume $\|F(z)\|_* \le G$ for all $z \in \mathcal{Z}$. Then, $\forall t \ge 1$:

$$\epsilon_{\text{sp}}(\bar{z}_t) \le \left[\sum_{s=1}^t \gamma_s\right]^{-1} \left[\Omega + \frac{G^2}{2} \sum_{s=1}^t \gamma_s^2\right],$$

where $\Omega := \max_{u \in \mathcal{Z}} D_\omega(u, z_1) \le \max_{\mathcal{Z}} \omega(\cdot) - \min_{\mathcal{Z}} \omega(\cdot)$.

Convergence rate

Theorem. Assume $\|F(z)\|_* \leq G$ for all $z \in \mathcal{Z}$. Then, $\forall t \geq 1$:

$$\epsilon_{\text{sp}}(\bar{z}_t) \leq \left[\sum_{s=1}^t \gamma_s\right]^{-1} \left[\Omega + \frac{G^2}{2} \sum_{s=1}^t \gamma_s^2\right],$$

where $\Omega := \max_{u \in \mathcal{Z}} D_\omega(u, z_1) \leq \max_{\mathcal{Z}} \omega(\cdot) - \min_{\mathcal{Z}} \omega(\cdot)$.

Cor. Let $\gamma_t = \frac{\gamma}{G\sqrt{T}}$, for $t \in [T]$. Then, $\epsilon_{\text{sp}}(\bar{z}_T) \leq \frac{G}{\sqrt{T}}\left[\frac{\Omega}{\gamma} + \frac{G\gamma}{2}\right]$.

Exercise: Verify that for $\gamma_t = \frac{1}{G}\sqrt{\frac{2\Omega}{T}}$, $\epsilon_{\text{sp}}(\bar{z}_T) \leq G\sqrt{\frac{2\Omega}{T}}$.

Convergence rate

Theorem. Assume $\|F(z)\|_* \le G$ for all $z \in \mathcal{Z}$. Then, $\forall t \ge 1$:

$$\epsilon_{\mathrm{sp}}(\bar{z}_t) \le \left[\sum_{s=1}^t \gamma_s\right]^{-1}\left[\Omega + \frac{G^2}{2}\sum_{s=1}^t \gamma_s^2\right],$$

where $\Omega := \max_{u\in\mathcal{Z}} D_\omega(u, z_1) \le \max_{\mathcal{Z}} \omega(\cdot) - \min_{\mathcal{Z}} \omega(\cdot)$.

Cor. Let $\gamma_t = \frac{\gamma}{G\sqrt{T}}$, for $t \in [T]$. Then, $\epsilon_{\mathrm{sp}}(\bar{z}_T) \le \frac{G}{\sqrt{T}}\left[\frac{\Omega}{\gamma} + \frac{G\gamma}{2}\right]$.

Exercise: Verify that for $\gamma_t = \frac{1}{G}\sqrt{\frac{2\Omega}{T}}$, $\epsilon_{\mathrm{sp}}(\bar{z}_T) \le G\sqrt{\frac{2\Omega}{T}}$.

Essentially subgradient method style proof, except …

Convergence rate

Lemma (MD lemma). For any $u \in \mathcal{Z}$, we have

$$\gamma_t \langle F(z_t),\, z_t - u \rangle \leq D_\omega(u, z_t) - D_\omega(u, z_{t+1}) + \frac{\gamma_t^2}{2}\|F(z_t)\|_*^2.$$

Convergence rate

Lemma (MD lemma). For any $u \in \mathcal{Z}$, we have

$$\gamma_t \langle F(z_t), z_t - u \rangle \leq D_\omega(u, z_t) - D_\omega(u, z_{t+1}) + \frac{\gamma_t^2}{2} \|F(z_t)\|_*^2.$$

Why the above lemma?

Convergence rate

Lemma (MD lemma). For any $u \in \mathcal{Z}$, we have

$$\gamma_t \langle F(z_t), z_t - u \rangle \le D_\omega(u, z_t) - D_\omega(u, z_{t+1}) + \frac{\gamma_t^2}{2} \|F(z_t)\|_*^2.$$

Why the above lemma? Recall

Lemma O^*. A point z^* is an SP of ϕ **iff** for every selection $F(\cdot)$ of Φ (i.e., a vector field $F : \mathrm{ri}(\mathcal{Z}) \to \mathbb{R}^d$ s.t., $F(z) \in \Phi(z)$ for every $z \in \mathrm{ri}(\mathcal{Z})$) we have $\langle F(z), z - z^* \rangle \ge 0$ for all $z \in \mathrm{ri}(\mathcal{Z})$.

Convergence rate

Lemma (MD lemma). For any $u \in \mathcal{Z}$, we have

$$\gamma_t \langle F(z_t),\, z_t - u\rangle \le D_\omega(u, z_t) - D_\omega(u, z_{t+1}) + \frac{\gamma_t^2}{2}\|F(z_t)\|_*^2.$$

Why the above lemma? Recall

Lemma O^*. A point z^* is an SP of ϕ **iff** for every selection $F(\cdot)$ of Φ (i.e., a vector field $F : \mathrm{ri}(\mathcal{Z}) \to \mathbb{R}^d$ s.t., $F(z) \in \Phi(z)$ for every $z \in \mathrm{ri}(\mathcal{Z})$) we have $\langle F(z),\, z - z^*\rangle \ge 0$ for all $z \in \mathrm{ri}(\mathcal{Z})$.

Step 1. Summing up MD lemma for $s = 1, \ldots, t$, we get

$$\sum_{s=1}^{t} \gamma_s \langle F(z_s),\, z_s - u\rangle \quad \le$$

Convergence rate

Lemma (MD lemma). For any $u \in \mathcal{Z}$, we have

$$\gamma_t \langle F(z_t),\, z_t - u \rangle \le D_\omega(u, z_t) - D_\omega(u, z_{t+1}) + \tfrac{\gamma_t^2}{2}\|F(z_t)\|_*^2.$$

Why the above lemma? Recall

Lemma O^*. A point z^* is an SP of ϕ **iff** for every selection $F(\cdot)$ of Φ (i.e., a vector field $F : \mathrm{ri}(\mathcal{Z}) \to \mathbb{R}^d$ s.t., $F(z) \in \Phi(z)$ for every $z \in \mathrm{ri}(\mathcal{Z})$) we have $\langle F(z),\, z - z^* \rangle \ge 0$ for all $z \in \mathrm{ri}(\mathcal{Z})$.

Step 1. Summing up MD lemma for $s = 1, \ldots, t$, we get

$$\sum_{s=1}^{t} \gamma_s \langle F(z_s),\, z_s - u \rangle \quad \le \quad D_\omega(u, z_1) + \sum_{s=1}^{t} \frac{\gamma_s^2}{2}\|F(z_s)\|_*^2$$

Convergence rate

Lemma (MD lemma). For any $u \in \mathcal{Z}$, we have

$$\gamma_t \langle F(z_t),\, z_t - u\rangle \le D_\omega(u, z_t) - D_\omega(u, z_{t+1}) + \tfrac{\gamma_t^2}{2}\|F(z_t)\|_*^2.$$

Why the above lemma? Recall

Lemma O^*. A point z^* is an SP of ϕ **iff** for every selection $F(\cdot)$ of Φ (i.e., a vector field $F : \mathrm{ri}(\mathcal{Z}) \to \mathbb{R}^d$ s.t., $F(z) \in \Phi(z)$ for every $z \in \mathrm{ri}(\mathcal{Z})$) we have $\langle F(z),\, z - z^*\rangle \ge 0$ for all $z \in \mathrm{ri}(\mathcal{Z})$.

Step 1. Summing up MD lemma for $s = 1, \ldots, t$, we get

$$\begin{aligned}\sum\nolimits_{s=1}^{t} \gamma_s \langle F(z_s),\, z_s - u\rangle &\le D_\omega(u, z_1) + \sum\nolimits_{s=1}^{t} \frac{\gamma_s^2}{2}\|F(z_s)\|_*^2 \\ &\le \Omega + \tfrac{G^2}{2}\sum\nolimits_{s=1}^{t} \gamma_s^2.\end{aligned}$$

Convergence rate

Lemma (MD lemma). For any $u \in \mathcal{Z}$, we have

$$\gamma_t \langle F(z_t), z_t - u \rangle \leq D_\omega(u, z_t) - D_\omega(u, z_{t+1}) + \tfrac{\gamma_t^2}{2} \|F(z_t)\|_*^2.$$

Why the above lemma? Recall

Lemma O^*. A point z^* is an SP of ϕ **iff** for every selection $F(\cdot)$ of Φ (i.e., a vector field $F : \mathrm{ri}(\mathcal{Z}) \to \mathbb{R}^d$ s.t., $F(z) \in \Phi(z)$ for every $z \in \mathrm{ri}(\mathcal{Z})$) we have $\langle F(z), z - z^* \rangle \geq 0$ for all $z \in \mathrm{ri}(\mathcal{Z})$.

Step 1. Summing up MD lemma for $s = 1, \ldots, t$, we get

$$\begin{aligned} \sum_{s=1}^t \gamma_s \langle F(z_s), z_s - u \rangle &\leq D_\omega(u, z_1) + \sum_{s=1}^t \frac{\gamma_s^2}{2} \|F(z_s)\|_*^2 \\ &\leq \Omega + \tfrac{G^2}{2} \sum_{s=1}^t \gamma_s^2. \end{aligned}$$

Step 2. Show that $\phi(\bar{x}_t, y) - \phi(x, \bar{y}_t) \leq \sum_{s=1}^t \lambda_s \langle F(z_s), z_s - u \rangle$,

Convergence rate

Lemma (MD lemma). For any $u \in \mathcal{Z}$, we have

$$\gamma_t \langle F(z_t), z_t - u \rangle \leq D_\omega(u, z_t) - D_\omega(u, z_{t+1}) + \frac{\gamma_t^2}{2} \|F(z_t)\|_*^2.$$

Why the above lemma? Recall

Lemma O^*. A point z^* is an SP of ϕ **iff** for every selection $F(\cdot)$ of Φ (i.e., a vector field $F : \mathrm{ri}(\mathcal{Z}) \to \mathbb{R}^d$ s.t., $F(z) \in \Phi(z)$ for every $z \in \mathrm{ri}(\mathcal{Z})$) we have $\langle F(z), z - z^* \rangle \geq 0$ for all $z \in \mathrm{ri}(\mathcal{Z})$.

Step 1. Summing up MD lemma for $s = 1, \ldots, t$, we get

$$\begin{aligned} \sum_{s=1}^t \gamma_s \langle F(z_s), z_s - u \rangle &\leq D_\omega(u, z_1) + \sum_{s=1}^t \frac{\gamma_s^2}{2} \|F(z_s)\|_*^2 \\ &\leq \Omega + \frac{G^2}{2} \sum_{s=1}^t \gamma_s^2. \end{aligned}$$

Step 2. Show that $\phi(\bar{x}_t, y) - \phi(x, \bar{y}_t) \leq \sum_{s=1}^t \lambda_s \langle F(z_s), z_s - u \rangle$, then upon taking sup of (x, y) we arrive at $\epsilon_{\mathrm{sp}}(\bar{z}_t)$, as desired.

Proof of Step 2

Note $z_t = (x_t, y_t)$, and $\bar{z}_t = (\bar{x}_t, \bar{y}_t)$. Let $\lambda_t = \gamma_t / \sum_{s=1}^{t} \gamma_t$.

Proof of Step 2

Note $z_t = (x_t, y_t)$, and $\bar{z}_t = (\bar{x}_t, \bar{y}_t)$. Let $\lambda_t = \gamma_t / \sum_{s=1}^{t} \gamma_t$.

$$\sum_{s=1}^{t} \lambda_s \langle F(z_s),\, z_s - u\rangle = \sum_{s=1}^{t} \lambda_s \big[\langle \nabla_x \phi(x_s, y_s),\, x_t - x\rangle + \langle \nabla_y \phi(x_s, y_s),\, y - y_t\rangle\big]$$

Proof of Step 2

Note $z_t = (x_t, y_t)$, and $\bar{z}_t = (\bar{x}_t, \bar{y}_t)$. Let $\lambda_t = \gamma_t / \sum_{s=1}^{t} \gamma_t$.

$$\sum_{s=1}^{t} \lambda_s \langle F(z_s), z_s - u\rangle = \sum_{s=1}^{t} \lambda_s \left[\langle \nabla_x \phi(x_s, y_s), x_t - x\rangle + \langle \nabla_y \phi(x_s, y_s), y - y_t\rangle\right]$$

$$\geq \sum_{s=1}^{t} \lambda_s [\phi(x_s, y_s) - \phi(x, y_s) + \phi(x_s, y) - \phi(x_s, y_s)]$$

Proof of Step 2

Note $z_t = (x_t, y_t)$, and $\bar{z}_t = (\bar{x}_t, \bar{y}_t)$. Let $\lambda_t = \gamma_t / \sum_{s=1}^{t} \gamma_t$.

$$\sum_{s=1}^{t} \lambda_s \langle F(z_s),\, z_s - u \rangle = \sum_{s=1}^{t} \lambda_s \left[\langle \nabla_x \phi(x_s, y_s),\, x_t - x \rangle + \langle \nabla_y \phi(x_s, y_s),\, y - y_t \rangle \right]$$

$$\geq \sum_{s=1}^{t} \lambda_s [\phi(x_s, y_s) - \phi(x, y_s) + \phi(x_s, y) - \phi(x_s, y_s)]$$

$$= \sum_{s=1}^{t} \lambda_s [\phi(x_s, y) - \phi(x, y_s)]$$

Proof of Step 2

Note $z_t = (x_t, y_t)$, and $\bar{z}_t = (\bar{x}_t, \bar{y}_t)$. Let $\lambda_t = \gamma_t / \sum_{s=1}^t \gamma_t$.

$$\sum_{s=1}^{t} \lambda_s \langle F(z_s), z_s - u \rangle = \sum_{s=1}^{t} \lambda_s \left[\langle \nabla_x \phi(x_s, y_s), x_t - x \rangle + \langle \nabla_y \phi(x_s, y_s), y - y_t \rangle \right]$$

$$\geq \sum_{s=1}^{t} \lambda_s [\phi(x_s, y_s) - \phi(x, y_s) + \phi(x_s, y) - \phi(x_s, y_s)]$$

$$= \sum_{s=1}^{t} \lambda_s [\phi(x_s, y) - \phi(x, y_s)]$$

$$\geq \phi\left(\sum_{s=1}^{t} \lambda_s x_s, y\right) - \phi\left(x, \sum_{s=1}^{t} \lambda_s y_s\right)$$

Proof of Step 2

Note $z_t = (x_t, y_t)$, and $\bar{z}_t = (\bar{x}_t, \bar{y}_t)$. Let $\lambda_t = \gamma_t / \sum_{s=1}^t \gamma_t$.

$$
\begin{aligned}
\sum_{s=1}^{t} \lambda_s \langle F(z_s),\, z_s - u\rangle &= \sum_{s=1}^{t} \lambda_s \big[\langle \nabla_x \phi(x_s, y_s),\, x_t - x\rangle + \langle \nabla_y \phi(x_s, y_s),\, y - y_t\rangle\big] \\
&\geq \sum_{s=1}^{t} \lambda_s [\phi(x_s, y_s) - \phi(x, y_s) + \phi(x_s, y) - \phi(x_s, y_s)] \\
&= \sum_{s=1}^{t} \lambda_s [\phi(x_s, y) - \phi(x, y_s)] \\
&\geq \phi\left(\sum\nolimits_{s=1}^{t} \lambda_s x_s, y\right) - \phi\left(x, \sum\nolimits_{s=1}^{t} \lambda_s y_s\right) \\
&= \phi(\bar{x}_t, y) - \phi(x, \bar{y}_t).
\end{aligned}
$$

Proof of Step 2

Note $z_t = (x_t, y_t)$, and $\bar{z}_t = (\bar{x}_t, \bar{y}_t)$. Let $\lambda_t = \gamma_t / \sum_{s=1}^{t} \gamma_t$.

$$
\begin{aligned}
\sum_{s=1}^{t} \lambda_s \langle F(z_s), z_s - u \rangle &= \sum_{s=1}^{t} \lambda_s \left[\langle \nabla_x \phi(x_s, y_s), x_t - x \rangle + \langle \nabla_y \phi(x_s, y_s), y - y_t \rangle \right] \\
&\geq \sum_{s=1}^{t} \lambda_s [\phi(x_s, y_s) - \phi(x, y_s) + \phi(x_s, y) - \phi(x_s, y_s)] \\
&= \sum_{s=1}^{t} \lambda_s [\phi(x_s, y) - \phi(x, y_s)] \\
&\geq \phi\left(\sum\nolimits_{s=1}^{t} \lambda_s x_s, y \right) - \phi\left(x, \sum\nolimits_{s=1}^{t} \lambda_s y_s \right) \\
&= \phi(\bar{x}_t, y) - \phi(x, \bar{y}_t).
\end{aligned}
$$

Clearly, $\sup_{(x,y)} \phi(\bar{x}_t, y) - \phi(x, \bar{y}_t) \geq \epsilon_{\text{sp}}(\bar{z}_t)$.

Faster than MD

(Exploit structure)

Faster than MD: exploiting structure

- We saw MD yield $O(1/\sqrt{T})$ for the CCSP problem.

Faster than MD: exploiting structure

- We saw MD yield $O(1/\sqrt{T})$ for the CCSP problem.

Problems have structure that can be exploited.

Faster than MD: exploiting structure

- We saw MD yield $O(1/\sqrt{T})$ for the CCSP problem.

Problems have structure that can be exploited.

Nesterov (2005) introduced an *"excessive gap technique"*

1. use saddle point reformulation of (convex) $\min_{x \in \mathcal{X}} f(x)$
2. obtain thus a cheap **smooth** convex approximation f_{sm}

Faster than MD: exploiting structure

- We saw MD yield $O(1/\sqrt{T})$ for the CCSP problem.

Problems have structure that can be exploited.

Nesterov (2005) introduced an *"excessive gap technique"*

1. use saddle point reformulation of (convex) $\min_{x \in \mathcal{X}} f(x)$
2. obtain thus a cheap **smooth** convex approximation f_{sm}
3. minimize f_{sm} at a rate $O(1/T^2)$ using AGD

Faster than MD: exploiting structure

- We saw MD yield $O(1/\sqrt{T})$ for the CCSP problem.

Problems have structure that can be exploited.

Nesterov (2005) introduced an *"excessive gap technique"*
1. use saddle point reformulation of (convex) $\min_{x \in \mathcal{X}} f(x)$
2. obtain thus a cheap **smooth** convex approximation f_{sm}
3. minimize f_{sm} at a rate $O(1/T^2)$ using AGD
4. smoothness of f_{sm} deteriorates as $f_{\text{sm}} \to f$, final rate $O(1/T)$

We'll look at Mirror-Prox (Nemirovski 2004): simpler, more transparent, easier to extend, and delivers, $O(1/T)$ rate

Examples with structure

Ex. Let $f(x) = \max_{1 \le i \le m} f_i(x) = \max_{y \in \mathbb{R}^m_+, y^T 1 = 1} [\phi(x, y) := \sum_i y_i f_i(x)]$

Examples with structure

Ex. Let $f(x) = \max_{1 \le i \le m} f_i(x) = \max_{y \in \mathbb{R}^m_+, y^T 1 = 1} [\phi(x,y) := \sum_i y_i f_i(x)]$

Ex. Let $f(x) = \|Ax - b\|_p = \max_{\|y\|_q \le 1} y^T(Ax - b)$.

Exercise: What about $f(x) = \|[Ax - b]_+\|_p$?

Ex. Let $A(x) = A_0 + \sum_i x_i A_i$. Let $S_k(X) = \sum_{i=1}^k \lambda_i^{\downarrow}(X)$.
Then, $S_k(A(x)) = \max_{y \in \Sigma_n, y \preceq I/k} [\phi(x,y) := k\langle y, A(x) \rangle]$;
here Σ_n denotes the spectrahedron $\{X \mid X \succeq 0, \operatorname{Tr}(X) = 1\}$

Examples with structure

Ex. Let $f(x) = \max_{1 \le i \le m} f_i(x) = \max_{y \in \mathbb{R}^m_+, y^T 1 = 1} [\phi(x,y) := \sum_i y_i f_i(x)]$

Ex. Let $f(x) = \|Ax - b\|_p = \max_{\|y\|_q \le 1} y^T(Ax - b)$.

Exercise: What about $f(x) = \|[Ax - b]_+\|_p$?

Ex. Let $A(x) = A_0 + \sum_i x_i A_i$. Let $S_k(X) = \sum_{i=1}^k \lambda_i^{\downarrow}(X)$.
Then, $S_k(A(x)) = \max_{y \in \Sigma_n, y \preceq I/k} [\phi(x,y) := k\langle y, A(x)\rangle]$;
here Σ_n denotes the spectrahedron $\{X \mid X \succeq 0, \mathrm{Tr}(X) = 1\}$

Explore: Seek many other such SP examples

Exploitng structure via Mirror Prox

Assumption A: Let $\mathcal{X}, \mathcal{Y}$ be bounded
Assumption B: Let $\phi(x, y) \in C^1_L$

Exploitng structure via Mirror Prox

Assumption A: Let $\mathcal{X}, \mathcal{Y}$ be bounded
Assumption B: Let $\phi(x,y) \in C^1_L$

Then, we have $F(z) = [\nabla_x \phi(x,y), -\nabla_y \phi(x,y)] = [F_x(z), F_y(z)]$

Exploitng structure via Mirror Prox

Assumption A: Let $\mathcal{X}, \mathcal{Y}$ be bounded
Assumption B: Let $\phi(x, y) \in C_L^1$

Then, we have $F(z) = [\nabla_x \phi(x, y), -\nabla_y \phi(x, y)] = [F_x(z), F_y(z)]$

MD setup

Choose a norm $\|\cdot\|$ on $\mathcal{Z}$, and a ***Bregman divergence***

$$D_\omega(u, z) := \omega(u) - \omega(z) - \langle \omega'(z), u - z \rangle$$

that is strongly convex (in u) wrt the chosen norm.

(Bregman)-Prox-mapping

$$\text{Prox}_z(\xi) := \underset{u \in \mathcal{Z}}{\text{argmin}} \; D_\omega(u, z) + \langle \xi, u \rangle$$

Lipschitz gradient

$\|F(z) - F(z')\|_* \leq L\|z - z'\|$ for all $z, z' \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$

Mirror Prox

1. Let $\gamma_t > 0$ be stepsizes for $t \geq 1$
2. $z_1 = \operatorname{argmin}_{u \in \mathcal{Z}} \omega(u)$ *(initialization)*
3. $w_t = \operatorname{Prox}_{z_t}(\gamma_t F(z_t))$ *(gradient step)*
4. $z_{t+1} = \operatorname{Prox}_{z_t}(\gamma_t F(w_t))$ *(extra-gradient step)*
5. $\bar{z}_t = \frac{\sum_{s=1}^t \gamma_s w_s}{\sum_{s=1}^t \gamma_s}$ *(average iterate)*

Step 4 additional on top of MD; a bit mysterious (requires digression into why it helps). Roughly, the extra regularization allows us to exploit the smoothness of $\phi(x, y)$ to take longer steps, and thus converge faster.

Mirror Prox

1. Let $\gamma_t > 0$ be stepsizes for $t \geq 1$
2. $z_1 = \mathrm{argmin}_{u \in \mathcal{Z}}\, \omega(u)$ *(initialization)*
3. $w_t = \mathrm{Prox}_{z_t}(\gamma_t F(z_t))$ *(gradient step)*
4. $z_{t+1} = \mathrm{Prox}_{z_t}(\gamma_t F(w_t))$ *(extra-gradient step)*
5. $\bar{z}_t = \frac{\sum_{s=1}^{t} \gamma_s w_s}{\sum_{s=1}^{t} \gamma_s}$ *(average iterate)*

Step 4 additional on top of MD; a bit mysterious (requires digression into why it helps). Roughly, the extra regularization allows us to exploit the smoothness of $\phi(x, y)$ to take longer steps, and thus converge faster.

For the average iterate; *not possible* without averaging!

Convergence of MP

Theorem. Let $\delta_t := \gamma_t\langle F(w_t), w_t - z_{t+1}\rangle - D_\omega(z_{t+1}, z_t)$. For every $t \geq 1$, assuming bounded $\mathcal{X}, \mathcal{Y}$, $\phi \in C^1_L$, we have:

- $\epsilon_{\text{sp}}(\bar{z}_t) \leq \left[\sum_{s=1}^{t} \gamma_t\right]^{-1}\left[\Omega + \sum_{s=1}^{t} \delta_s\right]$
- If $\gamma_t \leq 1/L$ **and** $\delta_t \leq 0$, then $\forall t \geq 1 : \quad \epsilon_{\text{sp}}(\bar{z}_t) \leq \frac{\Omega L}{t}$

This is the $O(1/T)$ convergence rate for MP.

Convergence of MP

Theorem. Let $\delta_t := \gamma_t\langle F(w_t), w_t - z_{t+1}\rangle - D_\omega(z_{t+1}, z_t)$. For every $t \geq 1$, assuming bounded $\mathcal{X}, \mathcal{Y}$, $\phi \in C_L^1$, we have:

- $\epsilon_{\text{sp}}(\bar{z}_t) \leq \left[\sum_{s=1}^t \gamma_t\right]^{-1}\left[\Omega + \sum_{s=1}^t \delta_s\right]$
- If $\gamma_t \leq 1/L$ **and** $\delta_t \leq 0$, then $\forall t \geq 1 :\quad \epsilon_{\text{sp}}(\bar{z}_t) \leq \frac{\Omega L}{t}$

This is the $O(1/T)$ convergence rate for MP.

Proof: a small upgrade on top of the MD proof

Convergence of MP

Theorem. Let $\delta_t := \gamma_t \langle F(w_t), w_t - z_{t+1}\rangle - D_\omega(z_{t+1}, z_t)$. For every $t \geq 1$, assuming bounded $\mathcal{X}, \mathcal{Y}$, $\phi \in C_L^1$, we have:

- $\epsilon_{\text{sp}}(\bar{z}_t) \leq \left[\sum_{s=1}^t \gamma_t\right]^{-1} \left[\Omega + \sum_{s=1}^t \delta_s\right]$
- If $\gamma_t \leq 1/L$ **and** $\delta_t \leq 0$, then $\forall t \geq 1$: $\quad \epsilon_{\text{sp}}(\bar{z}_t) \leq \frac{\Omega L}{t}$

This is the $O(1/T)$ convergence rate for MP.

Proof: a small upgrade on top of the MD proof

Again recall Lemma O^*

Lemma O^*. A point z^* is an SP of ϕ **iff** for every selection $F(\cdot)$ of Φ such that $F(z) \in \Phi(z)$ we have $\langle F(z), z - z^*\rangle \geq 0$ for all $z \in \text{ri}(\mathcal{Z})$.

Convergence of MP

$$\mathrm{Prox}_{z}(\xi) := \operatorname*{argmin}_{u \in \mathcal{Z}} D_\omega(u, z) + \langle \xi, u \rangle$$

Recall: key MP update steps

$$w_t = \mathrm{Prox}_{z_t}(\gamma_t F(z_t)), \quad z_{t+1} = \mathrm{Prox}_{z_t}(\gamma_t F(w_t)), \quad \bar{z}_t = \sum_{s=1}^{t} \lambda_s w_s$$

Convergence of MP

$$\mathrm{Prox}_z(\xi) := \operatorname*{argmin}_{u \in \mathcal{Z}} \; D_\omega(u, z) + \langle \xi, u \rangle$$

Recall: key MP update steps

$$w_t = \mathrm{Prox}_{z_t}(\gamma_t F(z_t)), \quad z_{t+1} = \mathrm{Prox}_{z_t}(\gamma_t F(w_t)), \quad \bar{z}_t = \sum\nolimits_{s=1}^t \lambda_s w_s$$

Using Lemma O^*, we upper-bound $\sum_{s=1}^t \lambda_s \langle F(z_s), w_s - u \rangle$

Convergence of MP

$$\text{Prox}_z(\xi) := \underset{u \in \mathcal{Z}}{\text{argmin}}\; D_\omega(u, z) + \langle \xi,\, u\rangle$$

Recall: key MP update steps

$$w_t = \text{Prox}_{z_t}(\gamma_t F(z_t)), \quad z_{t+1} = \text{Prox}_{z_t}(\gamma_t F(w_t)), \quad \bar{z}_t = \sum_{s=1}^{t} \lambda_s w_s$$

Using Lemma O^*, we upper-bound $\sum_{s=1}^{t} \lambda_s \langle F(z_s),\, w_s - u\rangle$
Recall also that we previously proved for $\bar{z}_t = (\bar{x}_t, \bar{y}_t)$:

$$\sum_{s=1}^{t} \lambda_s \langle F(z_s),\, w_s - u\rangle \geq \phi(\bar{x}_t, y) - \phi(x, \bar{y}_t)$$

Convergence of MP

$$\mathrm{Prox}_z(\xi) := \operatorname*{argmin}_{u \in \mathcal{Z}} D_\omega(u, z) + \langle \xi, u \rangle$$

Recall: key MP update steps

$$w_t = \mathrm{Prox}_{z_t}(\gamma_t F(z_t)), \quad z_{t+1} = \mathrm{Prox}_{z_t}(\gamma_t F(w_t)), \quad \bar{z}_t = \sum\nolimits_{s=1}^{t} \lambda_s w_s$$

Using Lemma O^*, we upper-bound $\sum_{s=1}^{t} \lambda_s \langle F(z_s), w_s - u \rangle$
Recall also that we previously proved for $\bar{z}_t = (\bar{x}_t, \bar{y}_t)$:

$$\sum\nolimits_{s=1}^{t} \lambda_s \langle F(z_s), w_s - u \rangle \geq \phi(\bar{x}_t, y) - \phi(x, \bar{y}_t)$$

so that upon taking supremum over (x, y) we obtain

$$\sum\nolimits_{s=1}^{t} \lambda_s \langle F(z_s), w_s - u \rangle \geq \epsilon_{\mathrm{sp}}(\bar{z}_t).$$

Convergence of MP

$$\text{Prox}_{z}(\xi) := \underset{u \in \mathcal{Z}}{\text{argmin}}\; D_\omega(u, z) + \langle \xi,\, u\rangle$$

Recall: key MP update steps

$$w_t = \text{Prox}_{z_t}(\gamma_t F(z_t)), \quad z_{t+1} = \text{Prox}_{z_t}(\gamma_t F(w_t)), \quad \bar{z}_t = \sum\nolimits_{s=1}^{t} \lambda_s w_s$$

Using Lemma O^*, we upper-bound $\sum_{s=1}^{t} \lambda_s \langle F(z_s),\, w_s - u\rangle$
Recall also that we previously proved for $\bar{z}_t = (\bar{x}_t, \bar{y}_t)$:

$$\sum\nolimits_{s=1}^{t} \lambda_s \langle F(z_s),\, w_s - u\rangle \geq \phi(\bar{x}_t, y) - \phi(x, \bar{y}_t)$$

so that upon taking supremum over (x, y) we obtain

$$\sum\nolimits_{s=1}^{t} \lambda_s \langle F(z_s),\, w_s - u\rangle \geq \epsilon_{\text{sp}}(\bar{z}_t).$$

Remains to prove:

$$\sum\nolimits_{s=1}^{t} \lambda_s \langle F(z_s),\, w_s - u\rangle \leq O\Big([\sum\nolimits_s \gamma_s]^{-1} (\Omega + \sum\nolimits_s \delta_s) \Big)$$

Convergence of MP

Lemma (MD Lemma). Let $w = \mathrm{Prox}_z(\xi)$ and $z_+ = \mathrm{Prox}_z(\eta)$. Then, for all $u \in \mathcal{Z}$, we upper-bound $\langle \eta, w - u\rangle$ as follows:

$$
\begin{aligned}
&\le \quad D_\omega(u,z) - D_\omega(u,z_+) + \langle \eta, w - z_+\rangle - D_\omega(z_+,z)\\
&\le \quad D_\omega(u,z) - D_\omega(u,z_+) + \langle \eta - \xi, w - z_+\rangle - D_\omega(w,z) - D_\omega(z_+,w)\\
&\le \quad D_\omega(u,z) - D_\omega(u,z_+) + [\|\eta - \xi\|_* \|w - z_+\| - \tfrac{1}{2}\|z - w\|^2 - \tfrac{1}{2}\|z_+ - w\|^2]\\
&\le \quad D_\omega(u,z) - D_\omega(u,z_+) + \tfrac{1}{2}[\|\eta - \xi\|_*^2 - \|w - z\|^2].
\end{aligned}
$$

Convergence of MP

Lemma (MD Lemma). Let $w = \text{Prox}_z(\xi)$ and $z_+ = \text{Prox}_z(\eta)$. Then, for all $u \in \mathcal{Z}$, we upper-bound $\langle \eta, w - u \rangle$ as follows:

$$\begin{aligned}
&\leq D_\omega(u,z) - D_\omega(u,z_+) + \langle \eta, w - z_+ \rangle - D_\omega(z_+, z)\\
&\leq D_\omega(u,z) - D_\omega(u,z_+) + \langle \eta - \xi, w - z_+ \rangle - D_\omega(w,z) - D_\omega(z_+,w)\\
&\leq D_\omega(u,z) - D_\omega(u,z_+) + [\|\eta - \xi\|_* \|w - z_+\| - \tfrac{1}{2}\|z - w\|^2 - \tfrac{1}{2}\|z_+ - w\|^2]\\
&\leq D_\omega(u,z) - D_\omega(u,z_+) + \tfrac{1}{2}[\|\eta - \xi\|_*^2 - \|w - z\|^2].
\end{aligned}$$

Using this lemma with $z = z_t$, $\xi = \gamma_t F(z_t)$, $\eta = \gamma_t F(w_t)$, we get:

- $w = w_t$ and $z_+ = z_{t+1}$

Convergence of MP

Lemma (MD Lemma). Let $w = \mathrm{Prox}_z(\xi)$ and $z_+ = \mathrm{Prox}_z(\eta)$. Then, for all $u \in \mathcal{Z}$, we upper-bound $\langle \eta, w - u \rangle$ as follows:

$$\begin{aligned}
&\leq D_\omega(u,z) - D_\omega(u,z_+) + \langle \eta, w - z_+ \rangle - D_\omega(z_+,z) \\
&\leq D_\omega(u,z) - D_\omega(u,z_+) + \langle \eta - \xi, w - z_+ \rangle - D_\omega(w,z) - D_\omega(z_+,w) \\
&\leq D_\omega(u,z) - D_\omega(u,z_+) + [\|\eta - \xi\|_* \|w - z_+\| - \tfrac{1}{2}\|z - w\|^2 - \tfrac{1}{2}\|z_+ - w\|^2] \\
&\leq D_\omega(u,z) - D_\omega(u,z_+) + \tfrac{1}{2}[\|\eta - \xi\|_*^2 - \|w - z\|^2].
\end{aligned}$$

Using this lemma with $z = z_t$, $\xi = \gamma_t F(z_t)$, $\eta = \gamma_t F(w_t)$, we get:

- $w = w_t$ and $z_+ = z_{t+1}$
- $\gamma_t \langle F(w_t), w_t - u \rangle \leq D_\omega(u, z_t) - D_\omega(u, z_{t+1}) + \delta_t$

Convergence of MP

Lemma (MD Lemma). Let $w = \mathrm{Prox}_z(\xi)$ and $z_+ = \mathrm{Prox}_z(\eta)$. Then, for all $u \in \mathcal{Z}$, we upper-bound $\langle \eta,\, w - u\rangle$ as follows:

$$\begin{aligned}
&\leq D_\omega(u,z) - D_\omega(u,z_+) + \langle \eta,\, w - z_+\rangle - D_\omega(z_+,z)\\
&\leq D_\omega(u,z) - D_\omega(u,z_+) + \langle \eta - \xi,\, w - z_+\rangle - D_\omega(w,z) - D_\omega(z_+,w)\\
&\leq D_\omega(u,z) - D_\omega(u,z_+) + [\|\eta - \xi\|_*\|w - z_+\| - \tfrac{1}{2}\|z - w\|^2 - \tfrac{1}{2}\|z_+ - w\|^2]\\
&\leq D_\omega(u,z) - D_\omega(u,z_+) + \tfrac{1}{2}[\|\eta - \xi\|_*^2 - \|w - z\|^2].
\end{aligned}$$

Using this lemma with $z = z_t$, $\xi = \gamma_t F(z_t)$, $\eta = \gamma_t F(w_t)$, we get:

- $w = w_t$ and $z_+ = z_{t+1}$
- $\gamma_t \langle F(w_t),\, w_t - u\rangle \leq D_\omega(u,z_t) - D_\omega(u,z_{t+1}) + \delta_t$
- $\delta_t \leq \frac{1}{2}[\gamma_t^2\|F(w_t) - F(z_t)\|_*^2 - \|w_t - z_t\|^2]$

Convergence of MP

Lemma (MD Lemma). Let $w = \text{Prox}_z(\xi)$ and $z_+ = \text{Prox}_z(\eta)$. Then, for all $u \in \mathcal{Z}$, we upper-bound $\langle \eta,\, w - u\rangle$ as follows:

$$\begin{aligned}
&\leq \quad D_\omega(u,z) - D_\omega(u,z_+) + \langle \eta,\, w - z_+\rangle - D_\omega(z_+,z)\\
&\leq \quad D_\omega(u,z) - D_\omega(u,z_+) + \langle \eta - \xi,\, w - z_+\rangle - D_\omega(w,z) - D_\omega(z_+,w)\\
&\leq \quad D_\omega(u,z) - D_\omega(u,z_+) + [\|\eta - \xi\|_* \|w - z_+\| - \tfrac{1}{2}\|z - w\|^2 - \tfrac{1}{2}\|z_+ - w\|^2]\\
&\leq \quad D_\omega(u,z) - D_\omega(u,z_+) + \tfrac{1}{2}[\|\eta - \xi\|_*^2 - \|w - z\|^2].
\end{aligned}$$

Using this lemma with $z = z_t$, $\xi = \gamma_t F(z_t)$, $\eta = \gamma_t F(w_t)$, we get:

- $w = w_t$ and $z_+ = z_{t+1}$
- $\gamma_t \langle F(w_t),\, w_t - u\rangle \leq D_\omega(u, z_t) - D_\omega(u, z_{t+1}) + \delta_t$
- $\delta_t \leq \frac{1}{2}[\gamma_t^2 \|F(w_t) - F(z_t)\|_*^2 - \|w_t - z_t\|^2]$

Sum over $s \in [t]$, note $D_\omega(u, z_1) \leq \Omega$ and use $\lambda_s = \frac{\gamma_s}{\sum_{s'} \gamma_{s'}}$ to get

Convergence of MP

Lemma (MD Lemma). Let $w = \mathrm{Prox}_z(\xi)$ and $z_+ = \mathrm{Prox}_z(\eta)$. Then, for all $u \in \mathcal{Z}$, we upper-bound $\langle \eta,\, w - u\rangle$ as follows:

$$\begin{aligned}
&\le\ D_\omega(u,z) - D_\omega(u,z_+) + \langle \eta,\, w - z_+\rangle - D_\omega(z_+,z)\\
&\le\ D_\omega(u,z) - D_\omega(u,z_+) + \langle \eta - \xi,\, w - z_+\rangle - D_\omega(w,z) - D_\omega(z_+,w)\\
&\le\ D_\omega(u,z) - D_\omega(u,z_+) + [\|\eta - \xi\|_*\|w - z_+\| - \tfrac{1}{2}\|z - w\|^2 - \tfrac{1}{2}\|z_+ - w\|^2]\\
&\le\ D_\omega(u,z) - D_\omega(u,z_+) + \tfrac{1}{2}[\|\eta - \xi\|_*^2 - \|w - z\|^2].
\end{aligned}$$

Using this lemma with $z = z_t$, $\xi = \gamma_t F(z_t)$, $\eta = \gamma_t F(w_t)$, we get:

- $w = w_t$ and $z_+ = z_{t+1}$
- $\gamma_t\langle F(w_t),\, w_t - u\rangle \le D_\omega(u,z_t) - D_\omega(u,z_{t+1}) + \delta_t$
- $\delta_t \le \frac{1}{2}[\gamma_t^2\|F(w_t) - F(z_t)\|_*^2 - \|w_t - z_t\|^2]$

Sum over $s \in [t]$, note $D_\omega(u,z_1) \le \Omega$ and use $\lambda_s = \frac{\gamma_s}{\sum_{s'}\gamma_{s'}}$ to get

$$\sum_{s=1}^{t} \lambda_s\langle F(w_t),\, w_t - u\rangle \le \frac{\Omega + \sum_{s=1}^{t}\delta_s}{\sum_{s=1}^{t}\gamma_s}$$

Convergence of MP

Lemma (MD Lemma). Let $w = \mathrm{Prox}_z(\xi)$ and $z_+ = \mathrm{Prox}_z(\eta)$. Then, for all $u \in \mathcal{Z}$, we upper-bound $\langle \eta,\, w - u\rangle$ as follows:

$$\begin{aligned}
&\le D_\omega(u,z) - D_\omega(u,z_+) + \langle \eta,\, w - z_+\rangle - D_\omega(z_+,z)\\
&\le D_\omega(u,z) - D_\omega(u,z_+) + \langle \eta - \xi,\, w - z_+\rangle - D_\omega(w,z) - D_\omega(z_+,w)\\
&\le D_\omega(u,z) - D_\omega(u,z_+) + [\|\eta - \xi\|_*\|w - z_+\| - \tfrac{1}{2}\|z - w\|^2 - \tfrac{1}{2}\|z_+ - w\|^2]\\
&\le D_\omega(u,z) - D_\omega(u,z_+) + \tfrac{1}{2}[\|\eta - \xi\|_*^2 - \|w - z\|^2].
\end{aligned}$$

Using this lemma with $z = z_t$, $\xi = \gamma_t F(z_t)$, $\eta = \gamma_t F(w_t)$, we get:

- $w = w_t$ and $z_+ = z_{t+1}$
- $\gamma_t\langle F(w_t),\, w_t - u\rangle \le D_\omega(u,z_t) - D_\omega(u,z_{t+1}) + \delta_t$
- $\delta_t \le \frac{1}{2}[\gamma_t^2\|F(w_t) - F(z_t)\|_*^2 - \|w_t - z_t\|^2]$

Sum over $s \in [t]$, note $D_\omega(u,z_1) \le \Omega$ and use $\lambda_s = \frac{\gamma_s}{\sum_{s'}\gamma_{s'}}$ to get

$$\sum_{s=1}^{t} \lambda_s\langle F(w_t),\, w_t - u\rangle \le \frac{\Omega + \sum_{s=1}^{t}\delta_s}{\sum_{s=1}^{t}\gamma_s}$$

Using $\gamma_t \le 1/L$, we see that $\delta_t \le 0$, completing the argument.

Extensions

Mirror-Prox with Splitting

The $O(1/T)$ rate of MP assumes ϕ is smooth. If instead, it is nonsmooth but available in a composite form (i.e., the nonsmooth part is “simple” and can be handled via a suitable proximity operator), then one can extend MP to retain the $O(1/T)$ rate.

Mirror-Prox with Splitting

The $O(1/T)$ rate of MP assumes ϕ is smooth. If instead, it is nonsmooth but available in a composite form (i.e., the nonsmooth part is "simple" and can be handled via a suitable proximity operator), then one can extend MP to retain the $O(1/T)$ rate.

If $\phi(\cdot, y)$ is smooth and strongly concave, we can even accelerate to $O(1/T^2)$ rate.

This speedup also rediscovered in a recent paper: *"Efficient algorithms for smooth minimax optimization. In NeurIPS, pages 12659–12670, 2019"*

Other topics

What we did not cover

- Lower bounds
- Optimal methods (tight, essentially tight)
- Stochastic CCSP problems

What we did not cover

- Lower bounds
- Optimal methods (tight, essentially tight)
- Stochastic CCSP problems

Near-Optimal Algorithms for Minimax Optimization

Tianyi Lin DARREN_LIN@BERKELEY.EDU
University of California, Berkeley

Chi Jin CHIJ@PRINCETON.EDU
Princeton University

Michael. I. Jordan JORDAN@CS.BERKELEY.EDU
University of California, Berkeley

Settings	References	Gradient Complexity
Strongly-Convex-Strongly-Concave	Tseng (1995)	$\tilde{O}(\kappa_{\mathbf{x}} + \kappa_{\mathbf{y}})$
	Nesterov and Scrimali (2006)	
	Gidel et al. (2019)	
	Mokhtari et al. (2019b)	
	Alkousa et al. (2019)	$\tilde{O}(\min\{\kappa_{\mathbf{x}}\sqrt{\kappa_{\mathbf{y}}}, \kappa_{\mathbf{y}}\sqrt{\kappa_{\mathbf{x}}}\})$
	This paper (Theorem 9)	$\tilde{O}(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}})$
	Lower bound (Ibrahim et al., 2019)	$\tilde{\Omega}(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}})$
	Lower bound (Zhang et al., 2019)	$\tilde{\Omega}(\sqrt{\kappa_{\mathbf{x}}\kappa_{\mathbf{y}}})$
Strongly-Convex-Linear (special case of strongly-convex-concave)	Juditsky and Nemirovski (2011)	$O(\sqrt{\kappa_{\mathbf{x}}/\epsilon})$
	Hamedani and Aybat (2018)	
	Zhao (2019)	
Strongly-Convex-Concave	Thekumparampil et al. (2019)	$\tilde{O}(\kappa_{\mathbf{x}}/\sqrt{\epsilon})$
	This paper (Corollary 10)	$\tilde{O}(\sqrt{\kappa_{\mathbf{x}}/\epsilon})$
	Lower bound (Ouyang and Xu, 2019)	$\tilde{\Omega}(\sqrt{\kappa_{\mathbf{x}}/\epsilon})$
Convex-Concave	Nemirovski (2004)	$O(\epsilon^{-1})$
	Nesterov (2007)	
	Tseng (2008)	
	This paper (Corollary 11)	$\tilde{O}(\epsilon^{-1})$
	Lower bound (Ouyang and Xu, 2019)	$\Omega(\epsilon^{-1})$