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 $\phi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \cup \{\pm \infty\}$

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

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

**Exercise:** Show that weak duality follows from the above minimax inequality. Hint: Use $\phi = L$ (Lagrangian), and suitably choose $y$. 

Suvrit Sra (suvrit@mit.edu) 6.881 Optimization for Machine Learning (04/13/21; Lecture 15)
**Theorem.** Let \( \phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm \infty\} \). Then,

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

**Proof:**

\[
x, y, \inf_{x' \in \mathcal{X}} \phi(x', y) \leq \phi(x, y)
\]
**Theorem.** Let $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{\pm \infty\}$. Then,

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

**Proof:**

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

$$x, y, \quad \inf_{x' \in \mathcal{X}} \phi(x', y) \leq \sup_{y' \in \mathcal{Y}} \phi(x, y')$$
**Theorem.** Let \( \phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R} \cup \{ \pm \infty \} \). Then,

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

**Proof:**

\[
x, y, \inf_{x' \in \mathcal{X}} \phi(x', y) \leq \phi(x, y)
\]

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

\[
\forall x, \sup_{y \in \mathcal{Y}} \inf_{x' \in \mathcal{X}} \phi(x', y) \leq \sup_{y' \in \mathcal{Y}} \phi(x, y')
\]
Weak minimax (cf. weak duality)

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

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

**Proof:**

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

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

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

$$\implies \sup_{y \in \mathcal{Y}} \inf_{x' \in \mathcal{X}} \phi(x', y) \leq \inf_{x \in \mathcal{X}} \sup_{y' \in \mathcal{Y}} \phi(x, y').$$
Weak minimax (cf. weak duality)

**Theorem.** Let \( \phi : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbb{R} \cup \{\pm \infty\} \). Then,

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

**Proof:**

\[
x, y, \inf_{x' \in \mathcal{X}} \phi(x', y) \leq \phi(x, y)
\]

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

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

\[
\implies \sup_{y \in \mathcal{Y}} \inf_{x' \in \mathcal{X}} \phi(x', y) \leq \inf_{x \in \mathcal{X}} \sup_{y' \in \mathcal{Y}} \phi(x, y').
\]

**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 \((x^*, y^*)\)

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

\[
f(x^*) = \phi(x^*, y^*) = g(y^*).
\]
**Def.** Let $\phi$ be as before. Pair $(x^*, y^*)$ is a saddle-point of $\phi$ iff the infimum in the expression

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

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

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

is **attained** at $y^*$, and these two extrema are **equal**.
**Def.** Let $\phi$ be as before. Pair $(x^*, y^*)$ is a saddle-point of $\phi$ 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.

$$x^* \in \arg\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \phi(x, y), \quad y^* \in \arg\max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} \phi(x, y).$$
Strong minimax

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

More interesting question: Starting from the primal problem over X, how to introduce a space Y and a “useful” function φ on X×Y so that we have a saddle-point?

Sufficient conditions for saddle-point

▶ Function φ is continuous, and
▶ It is convex-concave, i.e., φ(·, y) convex for every y ∈ Y, and φ(x, ·) concave for every x ∈ X; and
▶ Both X and Y are convex; one of them is compact.

▶ (More generally: φ is appropriately semicontinuous and quasiconvex-quasiconcave with convex X, Y)
Strong minimax

♦ Classes of problems “dual” to each other can be generated by studying classes of functions $\phi$

♦ More interesting question: Starting from the primal problem over $\mathcal{X}$, how to introduce a space $\mathcal{Y}$ and a “useful” function $\phi$ on $\mathcal{X} \times \mathcal{Y}$ so that we have a saddle-point?

Sufficient conditions for saddle-point

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

(More generally: $\phi$ is appropriately semicontinuous and quasiconvex-quasiconcave with convex $\mathcal{X}$, $\mathcal{Y}$.)
Strong minimax

♠ Classes of problems “dual” to each other can be generated by studying classes of functions $\phi$

♠ More interesting question: Starting from the primal problem over $\mathcal{X}$, how to introduce a space $\mathcal{Y}$ and a “useful” function $\phi$ on $\mathcal{X} \times \mathcal{Y}$ so that we have a saddle-point?

**Sufficient conditions for saddle-point**

- Function $\phi$ 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 $\phi$

♠ **More interesting question:** Starting from the primal problem over $\mathcal{X}$, how to introduce a space $\mathcal{Y}$ and a “useful” function $\phi$ on $\mathcal{X} \times \mathcal{Y}$ so that we have a saddle-point?

**Sufficient conditions for saddle-point**

- Function $\phi$ 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: $\phi$ is appropriately semicontinuous and quasiconvex-quasiconcave with convex $\mathcal{X}, \mathcal{Y}$)
Example: Lasso-like problem

\[ p^* := \min_x \|Ax - b\|_2 + \lambda \|x\|_1. \]
Example: Lasso-like problem

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

\[ \|x\|_1 = \max \{x^Tv \mid \|v\|_\infty \leq 1 \} \]

\[ \|x\|_2 = \max \{x^Tu \mid \|u\|_2 \leq 1 \}. \]
Example: Lasso-like problem

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

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

**Saddle-point formulation**

\[ p^* = \min_x \max_{u,v} \left\{ u^T (b - Ax) + v^T x \mid \|u\|_2 \leq 1, \; \|v\|_\infty \leq \lambda \right\} \]
Example: Lasso-like problem

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

\[ \|x\|_1 = \max \{x^T v \mid \|v\|_\infty \leq 1\} \]

\[ \|x\|_2 = \max \{x^T u \mid \|u\|_2 \leq 1\}. \]

**Saddle-point formulation**

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

\[ = \max_{u,v} \min_x \left\{ u^T (b - Ax) + x^T v \mid \|u\|_2 \leq 1, \|v\|_\infty \leq \lambda \right\} \]
Example: Lasso-like problem

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

\[ \|x\|_1 = \max\{x^T v \mid \|v\|_\infty \leq 1\} \]

\[ \|x\|_2 = \max\{x^T u \mid \|u\|_2 \leq 1\}. \]

**Saddle-point formulation**

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

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

\[ = \max_{u,v} u^T b \quad A^T u = v, \|u\|_2 \leq 1, \|v\|_\infty \leq \lambda \]
Example: Lasso-like problem

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

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

Saddle-point formulation

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

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

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

Primal-Dual pair of problems

\[ \text{Opt}(P) := \min_{x \in X} f(x) = \sup_{y \in Y} \phi(x, y), \]
\[ \text{Opt}(D) := \max_{y \in Y} g(y) = \inf_{x \in X} \phi(x, y). \]
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

Opt(\(P\)) := \( \min_{x \in \mathcal{X}} f(x) = \sup_{y \in \mathcal{Y}} \phi(x, y) \),

Opt(\(D\)) := \( \max_{y \in \mathcal{Y}} g(y) = \inf_{x \in \mathcal{X}} \phi(x, y) \).

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

\[ \text{Opt}(P) = \text{Opt}(D) = \phi(x^*, y^*) = f(x^*) = g(y^*). \]
Judging solutions of the CCSP problem

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

$$
\epsilon_{sp}(z) := \sup_{q \in \mathcal{Y}} \phi(x, q) - \inf_{p \in \mathcal{X}} \phi(p, y) = f(x) - g(y).
$$
Judging solutions of the CCSP problem

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

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

Let us rewrite this gap in a more revealing form
Judging solutions of the CCSP problem

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

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

Let us rewrite this gap in a more revealing form

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

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 $x \in X$. 

**Lemma O**

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

**Assumption**: $Z$ is bounded and $\phi$ is Lipschitz continuous on $Z$ (in this case, $\text{dom} \Phi = Z$).
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 \( y \in \mathcal{Y} \).
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 $\Phi$ 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 $x \in X$. Let $\partial_y [-\phi(x, y)]$ be subdiff of $-\phi(x, \cdot)$ at point $y \in Y$.

**Subdiff:** Let $\Phi(z) \equiv \Phi(x, y) = \partial_x \phi(x, y) \times \partial_y [-\phi(x, y)]$.

**Exercise:** Verify by definition that $\Phi$ is a monotone operator.

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

![Lemma O*](image)

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

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

$$\text{Prox}_\mathcal{Z}(\xi) := \arg\min_{u \in \mathcal{Z}} D_\omega(u, z) + \langle \xi, u \rangle$$
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
**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. \( z_1 = \arg\min_{u \in \mathcal{Z}} \omega(u) \) \hspace{1cm} (initialization)
**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. \( z_1 = \arg\min_{u \in \mathcal{Z}} \omega(u) \) (initialization)
3. \( z_{t+1} = \text{Prox}_{z_t}(\gamma_tF(z_t)) \) (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. \( z_1 = \text{argmin}_{u \in \mathcal{Z}} \omega(u) \) (initialization)
3. \( z_{t+1} = \text{Prox}_{\gamma_t z_t}(\gamma_t F(z_t)) \) (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. $z_1 = \text{argmin}_{u \in \mathcal{Z}} \omega(u)$ (initialization)
3. $z_{t+1} = \text{Prox}_{z_t}(\gamma_t F(z_t))$ (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**: \[ \| \cdot \| = \| \cdot \|_2, \omega(x) = \frac{1}{2} x^T x \]

- \( \ell_1 \) setup: \[ \| \cdot \| = \| \cdot \|_1, \text{ when } Z \text{ a simplex, then} \]
  \[ \omega(z) = \sum_i z_i \log z_i \]

- \( \ell_1 \) setup: \[ \| \cdot \| = \| \cdot \|_1, \text{ when } Z \text{ bounded (e.g., the unit } \ell_1\text{-ball), one can set } \omega(z) = 2e \log n \sum_{i=1}^{n} |z_i|^{p(n)}, \text{ where} \]
  \[ p(n) = 1 + 1/2 \log n. \]

- Many other examples,...

Take advantage of prob geometry; obtain faster FOMs
Theorem. Assume $\|F(z)\|_* \leq G$ for all $z \in Z$. Then, $\forall t \geq 1$:

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

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

Cor. Let $\gamma_t = \gamma G \sqrt{T}$, for $t \in \lfloor T \rfloor$. Then,

$$
\epsilon_{sp}(\tilde{z}_T) \leq G \sqrt{T} \left[ \Omega \gamma + G \gamma^2 \right].
$$

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

Essentially subgradient method style proof, except . . .
Convergence rate

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

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

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, $\epsilon_{sp}(\bar{z}_T) \leq \frac{G}{\sqrt{T}} \left[ \frac{\Omega}{\gamma} + \frac{G \gamma}{2} \right]$.

**Exercise:** Verify that for $\gamma_t = \frac{1}{G} \sqrt{\frac{2 \Omega}{T}}$, $\epsilon_{sp}(\bar{z}_T) \leq G \sqrt{\frac{2 \Omega}{T}}$. 
**Theorem.** Assume $\|F(z)\|_* \leq G$ for all $z \in \mathcal{Z}$. Then, $\forall t \geq 1$:

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

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, $\epsilon_{sp}(\bar{z}_T) \leq \frac{G}{\sqrt{T}} \left[ \frac{\Omega}{\gamma} + \frac{G\gamma}{2} \right]$.

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

Essentially subgradient method style proof, except …
**Lemma** (MD lemma). For any $u \in \mathcal{Z}$, we have
\[
\gamma_t \langle F(z_t), z_t - u \rangle \leq D_\omega(u, z_t) - D_\omega(u, z_{t+1}) + \frac{\gamma_t^2}{2} \|F(z_t)\|_*^2.
\]
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
\[
\gamma_t \langle F(z_t), z_t - u \rangle \leq D_\omega(u, z_t) - D_\omega(u, z_{t+1}) + \frac{\gamma_t^2}{2} \|F(z_t)\|_*^2.
\]

Why the above lemma? Recall

**Lemma O*. A point \( z^* \) is an SP of \( \phi \) iff for every selection \( F(\cdot) \) of \( \Phi \) (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 \( z \in \text{ri}(\mathcal{Z}) \).
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 $\phi$ iff for every selection $F(\cdot)$ of $\Phi$ (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 $z \in \text{ri}(\mathcal{Z})$.

**Step 1.** Summing up MD lemma for $s = 1, \ldots, t$, we get
\[
\sum_{s=1}^{t} \gamma_s \langle F(z_s), z_s - u \rangle \leq
\]

Suvrit Sra (suvrit@mit.edu) 6.881 Optimization for Machine Learning (04/13/21; Lecture 15)
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 \( \phi \) iff for every selection \( F(\cdot) \) of \( \Phi \) (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 \( z \in \text{ri}(\mathcal{Z}) \).

**Step 1.** Summing up MD lemma for \( s = 1, \ldots, t \), we get
\[
\sum_{s=1}^{t} \gamma_s \langle F(z_s), z_s - u \rangle \leq D_\omega(u, z_1) + \sum_{s=1}^{t} \frac{\gamma_s^2}{2} \|F(z_s)\|_*^2
\]
Convergence rate

**Lemma** (MD lemma). For any \( u \in \mathcal{Z} \), we have
\[
\gamma_t \langle F(z_t), z_t - u \rangle \leq D_\omega(u, z_t) - D_\omega(u, z_{t+1}) + \frac{\gamma_t^2}{2} \|F(z_t)\|_*^2.
\]

Why the above lemma? Recall

**Lemma** \( O^* \). A point \( z^* \) is an SP of \( \phi \) iff for every selection \( F(\cdot) \) of \( \Phi \) (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 \( z \in \text{ri}(\mathcal{Z}) \).

**Step 1.** Summing up MD lemma for \( s = 1, \ldots, t \), we get
\[
\sum_{s=1}^t \gamma_s \langle F(z_s), z_s - u \rangle \leq D_\omega(u, z_1) + \sum_{s=1}^t \frac{\gamma_s^2}{2} \|F(z_s)\|_*^2
\]
\[
\leq \Omega + \frac{G^2}{2} \sum_{s=1}^t \gamma_s^2.
\]
**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** $O^\ast$. A point $z^\ast$ is an SP of $\phi$ **iff** for every selection $F(\cdot)$ of $\Phi$ (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 $\langle F(z), z - z^\ast \rangle \geq 0$ for all $z \in \text{ri}(\mathcal{Z})$.

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

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

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

Suvrit Sra (suvrit@mit.edu) 6.881 Optimization for Machine Learning (04/13/21; Lecture 15)
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 $\phi$ iff for every selection $F(\cdot)$ of $\Phi$ (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 $z \in \text{ri}(\mathcal{Z})$.

**Step 1.** Summing up MD lemma for $s = 1, \ldots, t$, we get
\[
\sum_{s=1}^t \gamma_s \langle F(z_s), z_s - u \rangle \leq D_\omega(u, z_1) + \sum_{s=1}^t \frac{\gamma_s^2}{2} \|F(z_s)\|_*^2 \\
\leq \Omega + \frac{G^2}{2} \sum_{s=1}^t \gamma_s^2.
\]

**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 $\epsilon_{sp}(\bar{z}_t)$, as desired.
Proof of Step 2

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

$$\sum_{s=1}^{t} \lambda_s \langle F(z_s), z_s - u \rangle = \sum_{s=1}^{t} \lambda_s \left[ \langle \nabla_x \phi(x_s, y_s), x_t - x \rangle + \langle \nabla_y \phi(x_s, y_s), y - y_t \rangle \right]$$
Proof of Step 2

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

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

\[
\geq \sum_{s=1}^{t} \lambda_s \left[ \phi(x_s, y_s) - \phi(x, y_s) + \phi(x_s, y) - \phi(x_s, y_s) \right]
\]
Note $z_t = (x_t, y_t)$, and $\bar{z}_t = (\bar{x}_t, \bar{y}_t)$. Let $\lambda_t = \gamma_t / \sum_{s=1}^{t} \gamma_t$.

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

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

$$
= \sum_{s=1}^{t} \lambda_s \left[ \phi(x_s, y) - \phi(x, y_s) \right]
$$
Proof of Step 2

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

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

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

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

\[
\geq \phi \left( \sum_{s=1}^t \lambda_s x_s, y \right) - \phi \left( x, \sum_{s=1}^t \lambda_s y_s \right)
\]
Proof of Step 2

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

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

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

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

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

$$
= \phi(\bar{x}_t, y) - \phi(x, \bar{y}_t).
$$
Proof of Step 2

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

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

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

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

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

\[ = \phi(\bar{x}_t, y) - \phi(x, \bar{y}_t). \]

Clearly, \( \sup_{(x,y)} \phi(\bar{x}_t, y) - \phi(x, \bar{y}_t) \geq \epsilon_{sp}(\bar{z}_t) \).
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.
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) $\min_{x \in X} f(x)$
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) $\min_{x \in X} f(x)$
2. obtain thus a cheap smooth convex approximation $f_{sm}$
3. minimize $f_{sm}$ at a rate $O(1/T^2)$ using AGD
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) $\min_{x \in X} f(x)$
2. obtain thus a cheap \textbf{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 $f_{sm} \to f$, 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 $f(x) = \max_{1 \leq i \leq m} f_i(x) = \max_{y \in \mathbb{R}_+^m, y^T 1 = 1} [\phi(x, y) := \sum_i y_i f_i(x)]$
Examples with structure

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

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

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

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

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

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

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

Ex. Let $A(x) = A_0 + \sum_i x_i A_i$. Let $S_k(X) = \sum_{i=1}^k \lambda_i^{1/k}(X)$. Then, $S_k(A(x)) = \max_{y \in \Sigma_n, y \preceq I/k} [\phi(x, y) := k \langle y, A(x) \rangle]$;

here $\Sigma_n$ denotes the spectrahedron $\{X \mid X \succeq 0, \text{Tr}(X) = 1\}$

Explore: Seek many other such SP examples
Exploiting structure via Mirror Prox

**Assumption A:** Let $\mathcal{X}, \mathcal{Y}$ be bounded

**Assumption B:** Let $\phi(x, y) \in C^1_L$
Exploiting 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 $F(z) = [\nabla_x \phi(x, y), -\nabla_y \phi(x, y)] = [F_x(z), F_y(z)]$
Exploiting 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 $F(z) = [\nabla_x \phi(x, y), -\nabla_y \phi(x, y)] = [F_x(z), F_y(z)]$

**MD setup**

| Choose a norm $\| \cdot \|$ on $Z$, 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**

$\text{Prox}_Z(\xi) := \arg\min_{u \in Z} D_\omega(u, z) + \langle \xi, u \rangle$

**Lipschitz gradient**

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

1. Let $\gamma_t > 0$ be stepsizes for $t \geq 1$
2. $z_1 = \arg\min_{u \in \mathcal{Z}} \omega(u)$ (initialization)
3. $w_t = \text{Prox}_{z_t}(\gamma_t F(z_t))$ (gradient step)
4. $z_{t+1} = \text{Prox}_{z_t}(\gamma_t F(w_t))$ (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.
Let $\gamma_t > 0$ be stepsizes for $t \geq 1$

1. $z_1 = \arg\min_{u \in \mathcal{Z}} \omega(u)$ (initialization)
2. $w_t = \text{Prox}_{z_t}(\gamma_t F(z_t))$ (gradient step)
3. $z_{t+1} = \text{Prox}_{z_t}(\gamma_t F(w_t))$ (extra-gradient step)
4. $\bar{z}_t = \frac{\sum_{s=1}^{t} \gamma_s w_s}{\sum_{s=1}^{t} \gamma_s}$ (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!
**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:

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

This is the $O(1/T)$ convergence rate for 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 \( X, Y, \phi \in C^1_L \), we have:

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

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

Proof: a small upgrade on top of the MD proof.
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:

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

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 $\phi$ iff for every selection $F(\cdot)$ of $\Phi$ 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

\[ \text{Prox}_z(\xi) := \operatorname{argmin}_{u \in Z} D_\omega(u, z) + \langle \xi, u \rangle \]

Recall: key MP update steps

\[ w_t = \text{Prox}_{z_t}(\gamma_tF(z_t)), \quad z_{t+1} = \text{Prox}_{z_t}(\gamma_tF(w_t)), \quad \bar{z}_t = \sum_{s=1}^{t} \lambda_s w_s \]
Convergence of MP

\[ \text{Prox}_z(\xi) := \arg\min_{u \in Z} D_\omega(u, z) + \langle \xi, u \rangle \]

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 \( \sum_{s=1}^{t} \lambda_s \langle F(z_s), w_s - u \rangle \)
Convergence of MP

\[ \text{Prox}_z(\xi) := \arg\min_{u \in Z} D_\omega(u, z) + \langle \xi, u \rangle \]

**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

\[ \text{Prox}_z(\xi) := \arg\min_{u \in \mathcal{Z}} D_\omega(u, z) + \langle \xi, u \rangle \]

**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 \( \bar{z}_t = (\bar{x}_t, \bar{y}_t) \):

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

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

\[ \sum_{s=1}^{t} \lambda_s \langle F(z_s), w_s - u \rangle \geq \epsilon_{sp}(\bar{z}_t). \]
Convergence of MP

Prox_z(\xi) := \arg\min_{u \in Z} D_\omega(u, z) + \langle \xi, u \rangle

Recall: key MP update steps

\omega_t = \text{Prox}_{z_t}(\gamma_t F(z_t)), \quad z_{t+1} = \text{Prox}_{z_t}(\gamma_t F(\omega_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)

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

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

Remains to prove:

\sum_{s=1}^{t} \lambda_s \langle F(z_s), w_s - u \rangle \leq O\left(\left[\sum_s \gamma_s\right]^{-1}(\Omega + \sum_s \delta_s)\right)
**Lemma** (MD Lemma). Let $w = \text{Prox}_z(\xi)$ and $z_+ = \text{Prox}_z(\eta)$. Then, for all $u \in Z$, we upper-bound $\langle \eta, w - u \rangle$ as follows:

\[
\begin{align*}
&\leq D_\omega(u, z) - D_\omega(u, z_+) + \langle \eta, w - z_+ \rangle - D_\omega(z_+, z) \\
&\leq D_\omega(u, z) - D_\omega(u, z_+) + \langle \eta - \xi, w - z_+ \rangle - D_\omega(w, z) - D_\omega(z_+, w) \\
&\leq D_\omega(u, z) - D_\omega(u, z_+) + [\|\eta - \xi\|_* \|w - z_+\| - \frac{1}{2} \|z - w\|^2 - \frac{1}{2} \|z_+ - w\|^2] \\
&\leq D_\omega(u, z) - D_\omega(u, z_+) + \frac{1}{2} [\|\eta - \xi\|_*^2 - \|w - z\|^2].
\end{align*}
\]
Lemma (MD Lemma). Let \( w = \text{Prox}_z(\xi) \) and \( z_+ = \text{Prox}_z(\eta) \). Then, for all \( u \in Z \), we upper-bound \( \langle \eta, w - u \rangle \) as follows:

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

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 \( z_+ = z_{t+1} \)
**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:

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

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 $z_+ = z_{t+1}$
- $\gamma_t \langle F(w_t), w_t - u \rangle \leq D_\omega(u, z_t) - D_\omega(u, z_{t+1}) + \delta_t$
**Lemma** (MD Lemma). Let $w = \text{Prox}_z(\xi)$ and $z_+ = \text{Prox}_z(\eta)$. Then, for all $u \in Z$, we upper-bound $\langle \eta, w - u \rangle$ as follows:

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

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 $z_+ = z_{t+1}$
- $\gamma_t \langle F(w_t), w_t - u \rangle \leq D_\omega(u, z_t) - D_\omega(u, z_{t+1}) + \delta_t$
- $\delta_t \leq \frac{1}{2} \left[ \gamma_t^2 \|F(w_t) - F(z_t)\|_*^2 - \|w_t - z_t\|^2 \right]$
**Lemma** (MD Lemma). Let \( w = \text{Prox}_z(\xi) \) and \( z_+ = \text{Prox}_z(\eta) \). Then, for all \( u \in Z \), we upper-bound \( \langle \eta, w - u \rangle \) as follows:

\[
\leq D_\omega(u, z) - D_\omega(u, z_+) + \langle \eta, w - z_+ \rangle - D_\omega(z_+, z)
\]

\[
\leq D_\omega(u, z) - D_\omega(u, z_+) + \langle \eta - \xi, w - z_+ \rangle - D_\omega(w, z) - D_\omega(z_+, w)
\]

\[
\leq D_\omega(u, z) - D_\omega(u, z_+) + \left[ \|\eta - \xi\| \|w - z_+\| - \frac{1}{2} \|z - w\|^2 - \frac{1}{2} \|z_+ - w\|^2 \right]
\]

\[
\leq D_\omega(u, z) - D_\omega(u, z_+) + \frac{1}{2} \left[ \|\eta - \xi\|^2 - \|w - z\|^2 \right].
\]

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

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
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:

\[
\leq D_\omega(u, z) - D_\omega(u, z_+) + \langle \eta - \xi, w - z_+ \rangle - D_\omega(w, z) - D_\omega(z_+, w)
\]

\[
\leq D_\omega(u, z) - D_\omega(u, z_+) + \frac{1}{2} \|\eta - \xi\| \|w - z_+\| - \frac{1}{2} \|z - w\|^2 - \frac{1}{2} \|z_+ - w\|^2
\]

\[
\leq D_\omega(u, z) - D_\omega(u, z_+) + \frac{1}{2} \|\eta - \xi\|_\star^2 - \|w - z\|^2.
\]

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

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

\[
\sum_{s=1}^{t} \lambda_s \langle F(w_t), w_t - u \rangle \leq \frac{\Omega + \sum_{s=1}^{t} \delta_s}{\sum_{s=1}^{t} \gamma_s}
\]
**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:

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

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 \( \lambda_s = \frac{\gamma_s}{\sum_{s'} \gamma_{s'}^s} \) to get

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

Using \( \gamma_t \leq 1/L \), we see that \( \delta_t \leq 0 \), completing the argument.
Extensions
The $O(1/T)$ rate of MP assumes $\phi$ 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.
The $O(1/T)$ rate of MP assumes $\phi$ 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  
University of California, Berkeley

Chi Jin  
Princeton University

Michael. I. Jordan  
University of California, Berkeley

<table>
<thead>
<tr>
<th>Settings</th>
<th>References</th>
<th>Gradient Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strongly-Convex-Strongly-Concave</td>
<td>Fuong (1995)</td>
<td>$O(n + s_g)$</td>
</tr>
<tr>
<td></td>
<td>Ne瘠ot and Scutari (2006)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gidel et al. (2019)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mohri et al. (2019)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Allena et al. (2019)</td>
<td>$O(\min{n\sqrt{s_g}, s_g \sqrt{n_g}})$</td>
</tr>
<tr>
<td>This paper (Theorem 9)</td>
<td></td>
<td>$O(\sqrt{s_g})$</td>
</tr>
<tr>
<td></td>
<td>Lower bound (Breath et al., 2019)</td>
<td>$\Omega(\sqrt{s_g})$</td>
</tr>
<tr>
<td></td>
<td>Lower bound (Zhang et al., 2019)</td>
<td>$\Omega(\sqrt{s_g})$</td>
</tr>
<tr>
<td>Strongly-Convex-Linear (special case of strongly convex-concave)</td>
<td>Jadranka and Nemirovski (2011)</td>
<td>$O(\sqrt{s_a/e})$</td>
</tr>
<tr>
<td></td>
<td>Hemelsens and Aytan (2018)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Zhao (2019)</td>
<td></td>
</tr>
<tr>
<td>Strongly-Convex-Concave</td>
<td>Thekunparampil et al. (2019)</td>
<td>$O(\sqrt{s_a})$</td>
</tr>
<tr>
<td>This paper (Corollary 16)</td>
<td></td>
<td>$O(\sqrt{s_a/e})$</td>
</tr>
<tr>
<td></td>
<td>Lower bound (Ouyang and Xa, 2019)</td>
<td>$\Omega(\sqrt{s_a/e})$</td>
</tr>
<tr>
<td>Convex-Concave</td>
<td>Nemirovski (2004)</td>
<td>$O(\epsilon^{-1})$</td>
</tr>
<tr>
<td></td>
<td>Nesterov (2007)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Teng (2008)</td>
<td></td>
</tr>
<tr>
<td>This paper (Corollary 13)</td>
<td></td>
<td>$O(\epsilon^{-1})$</td>
</tr>
<tr>
<td></td>
<td>Lower bound (Ouyang and Xa, 2019)</td>
<td>$\Omega(\epsilon^{-1})$</td>
</tr>
</tbody>
</table>

Suvrit Sra (suvrit@mit.edu)  6.881 Optimization for Machine Learning (04/13/21; Lecture 15)