Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 11: Primal-dual optimization I: Fundamentals of minimax problems

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

EE-556 (Fall 2021)
This work is released under a Creative Commons License with the following terms:

**Attribution**
- The licensor permits others to copy, distribute, display, and perform the work. In return, licensees must give the original authors credit.

**Non-Commercial**
- The licensor permits others to copy, distribute, display, and perform the work. In return, licensees may not use the work for commercial purposes – unless they get the licensor’s permission.

**Share Alike**
- The licensor permits others to distribute derivative works only under a license identical to the one that governs the licensor’s work.

**Full Text of the License**
Outline

▶ Today
  1. Min-max optimization (continued)

▶ Next week
  1. Algorithms for solving min-max optimization
A minimax optimization template

Minimax formulation

Consider the following problem that captures adversarial training, GANs, and robust reinforcement learning:

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

where $\Phi$ is differentiable and nonconvex in $x$ and nonconcave in $y$. (1)

Key questions:

1. Where do the algorithms converge?
2. When do the algorithm converge?
## A minimax optimization template

### Minimax formulation

Consider the following problem that captures adversarial training, GANs, and robust reinforcement learning:

\[
\min_{x \in X} \max_{y \in Y} \Phi(x, y),
\]

(1)

where \(\Phi\) is differentiable and nonconvex in \(x\) and nonconcave in \(y\).

- Key questions:
  1. Where do the algorithms converge?
  2. When do the algorithm converge?

### Recall: A buffet of negative results [5]

“Even when the objective is a Lipschitz and smooth differentiable function, deciding whether a min-max point exists, in fact even deciding whether an approximate min-max point exists, is NP-hard. More importantly, an approximate local min-max point of large enough approximation is guaranteed to exist, but finding one such point is PPAD-complete. The same is true of computing an approximate fixed point of the (Projected) Gradient Descent/Ascent update dynamics.”
The difficulty of the nonconvex-nonconcave setting

Minimax formulation
Consider the following problem that captures adversarial training, GANs, and robust reinforcement learning:

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

(2)

where $\Phi$ is differentiable and nonconvex in $x$ and nonconcave in $y$.

From minimax to minimization
Assume $\Phi(x, y) = f(x)$ for all $y$. The minimax optimization problem then seeks to find $x^*$ such that

$$f(x^*) \leq f(x), \forall x \in \mathbb{R}^p,$$

where $x^*$ is a global minimum of the nonconvex function $f$.

▶ Finding $x^*$ is NP-Hard even when $f$ is smooth! (see the complexity supplementary material)
▶ Finding solutions to a nonconvex-nonconvex min-max problem is harder in general.
Question 1 with a twist: Where do the algorithms want to converge?

Definition (Saddle points & Local Nash equilibria)

The point \((x^*, y^*)\) is called a saddle-point or a local Nash equilibrium (LNE) if it holds that

\[
\Phi(x^*, y) \leq \Phi(x^*, y^*) \leq \Phi(x, y^*)
\]

for all \(x\) and \(y\) within some neighborhood of \(x^*\) and \(y^*\), i.e., \(\|x - x^*\| \leq \delta\) and \(\|y - y^*\| \leq \delta\) for some \(\delta > 0\).

Necessary conditions

Through a Taylor expansion around \(x^*\) and \(y^*\) one can show that a LNE implies,

\[
\nabla_x \Phi(x, y), -\nabla_y \Phi(x, y) = 0 \\
\nabla_{xx} \Phi(x, y), -\nabla_{yy} \Phi(x, y) \succeq 0
\]

Figure: \(\Phi(x, y) = x^2 - y^2\)
Saddles of different shapes

Figure: The monkey saddle $\Phi(x, y) = x^3 - 3xy^2$ (left). The weird saddle $\Phi(x, y) = -x^2y^2 + xy$ (right) [17].
Question 2 with a twist: When do generalized Robbins-Monro schemes converge?

- Given \( \min_{x \in X} \max_{y \in Y} \Phi(x, y) \), define \( V(z) = [\nabla_x \Phi(x, y), -\nabla_y \Phi(x, y)] \) with \( z = [x, y]^T \).

- Given \( V(z) \), define stochastic estimates of \( V(z, \zeta) = V(z) + U(z, \zeta) \), where

  - \( U(z, \zeta) \) is a bias term
  - We often have unbiasedness: \( E U(z, \zeta) = 0 \)
  - The bias term can have bounded moments
  - We often have bounded variance: \( P(\|U(z, \zeta)\| \geq t) \leq 2 \exp\left(-\frac{t^2}{2\sigma^2}\right) \) for \( \sigma > 0 \).

- An abstract template for generalized Robbins-Monro schemes, dubbed as \( \mathcal{A} \):

\[
z^{k+1} = z^k - \alpha_k V(z^k, \zeta^k)
\]

The dessert section in the buffet of negative results: [12]

1. Bounded trajectories of \( \mathcal{A} \) always converge to an internally chain-transitive (ICT) set.

2. Trajectories of \( \mathcal{A} \) may converge with arbitrarily high probability to spurious attractors that contain no critical point of \( \Phi \).
Basic algorithms for minimax

- Given \( \min_{x \in X} \max_{y \in Y} \Phi(x, y) \), define \( V(z) = [\nabla_x \Phi(x, y), -\nabla_y \Phi(x, y)] \) with \( z = [x, y] \).

![Image of trajectory of different algorithms for a simple bilinear game](image_url)

**Figure**: Trajectory of different algorithms for a simple bilinear game \( \min_x \max_y xy \).

(In)Famous algorithms

- Gradient Descent Ascent (GDA)
- Proximal point method (PPM)
- Extra-gradient (EG)
- Optimistic Gradient Descent Ascent (OGDA)
- Reflected-Forward-Backward-Splitting (RFBS)

- EG and OGDA are approximations of the PPM
  - \( z^{k+1} = z^k - \alpha V(z^k) \).
  - \( z^{k+1} = z^k - \alpha V(z^{k+1}) \).
  - \( z^{k+1} = z^k - \alpha V(z^k - \alpha V(z^{k-1})) \).
  - \( z^{k+1} = z^k - \alpha [2V(z^k) - V(z^{k-1})] \).
  - \( z^{k+1} = z^k - \alpha (2z^k - z^{k-1}) \).
Minimax is more difficult than just optimization [11]

- Internally chain-transitive (ICT) sets characterize the convergence of dynamical systems [4].
  - For optimization, \( \{ \text{attracting ICT} \} \equiv \{ \text{solutions} \} \)
  - For minimax, \( \{ \text{attracting ICT} \} \equiv \{ \text{solutions} \} \cup \{ \text{spurious sets} \} \)

- “Almost” bilinear ≠ bilinear:
  \[
  \Phi(x, y) = xy + \epsilon \phi(x), \phi(x) = \frac{1}{2} x^2 - \frac{1}{4} x^4
  \]

- The “forsaken” solutions:
  \[
  \Phi(y, x) = y(x-0.5) + \phi(y) - \phi(x), \phi(u) = \frac{1}{4} u^2 - \frac{1}{2} u^4 + \frac{1}{6} u^6
  \]
A restricted minimax optimization template

A restricted minimax formulation

Consider the following problem

\[
\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \Phi(x, y),
\]

where \( \Phi \) is convex in \( x \) and concave in \( y \).

Key questions:

1. What problems does this template capture?
2. Where do the algorithms converge?
3. When do the algorithm converge?
General nonsmooth problems

- We will show that the restricted template captures the familiar composite minimization:

\[
\min_{x \in \mathbb{R}^p} f(x) + g(Ax).
\]

- \(f, g\) are convex, nonsmooth functions; and \(A\) is a linear operator.

**Examples**

- \(g(Ax) = \|Ax - b\|_1\) or \(g(Ax) = \|Ax - b\|_2^2\).

- \(g(Ax) = \delta_{\{b\}}(Ax)\), where \(\delta_{\{b\}}(Ax) = \begin{cases} 0, & \text{if } Ax = b, \\ +\infty, & \text{if } Ax \neq b. \end{cases}\)

**Observations:**

- The indicator example covers constrained problems, such as \(\min_{x \in X} \{f(x) : Ax = b\}\).

- We need a tool, called Fenchel conjugation, to reveal the underlying minimax problem.
Conjugation of functions

- Idea: Represent a convex function in max-form:

**Definition**

Let $Q$ be a Euclidean space and $Q^*$ be its dual space. Given a proper, closed and convex function $f : Q \to \mathbb{R} \cup \{+\infty\}$, the function $f^* : Q^* \to \mathbb{R} \cup \{+\infty\}$ such that

$$f^*(y) = \sup_{x \in \text{dom}(f)} \{ y^T x - f(x) \}$$

is called the Fenchel conjugate (or conjugate) of $f$.

**Observations:**
- $y$: slope of the hyperplane
- $-f^*(y)$: intercept of the hyperplane

*Figure:* The conjugate function $f^*(y)$ is the maximum gap between the linear function $x^T y$ (red line) and $f(x)$. 
Conjugation of functions

Definition
Given a proper, closed and convex function \( f : Q \rightarrow \mathbb{R} \cup \{+\infty\} \), the function \( f^* : Q^* \rightarrow \mathbb{R} \cup \{+\infty\} \) such that

\[
f^*(y) = \sup_{x \in \text{dom}(f)} \left\{ y^T x - f(x) \right\}
\]

is called the Fenchel conjugate (or conjugate) of \( f \).
Conjugation of functions

**Definition**
Given a proper, closed and convex function \( f : Q \rightarrow \mathbb{R} \cup \{+\infty\} \), the function \( f^* : Q^* \rightarrow \mathbb{R} \cup \{+\infty\} \) such that

\[
f^*(y) = \sup_{x \in \text{dom}(f)} \left\{ y^T x - f(x) \right\}
\]

is called the Fenchel conjugate (or conjugate) of \( f \).

**Properties**
- \( f^* \) is a convex and lower semicontinuous function by construction as the supremum of affine functions of \( y \).
- The conjugate of the conjugate of a convex function \( f \) is the same function \( f \); i.e., \( f^{**} = f \) for \( f \in \mathcal{F}(Q) \).
- The conjugate of the conjugate of a non-convex function \( f \) is its lower convex envelope when \( Q \) is compact:
  - \( f^{**}(x) = \sup \{ g(x) : g \text{ is convex and } g \leq f, \forall x \in Q \} \).
- For closed convex \( f \), \( \mu \)-strong convexity w.r.t. \( \| \cdot \| \) is equivalent to \( \frac{1}{\mu} \) smoothness of \( f^* \) w.r.t. \( \| \cdot \|^* \).
  - Recall dual norm: \( \| y \|^* = \sup_{x} \{ \langle x, y \rangle : \| x \| \leq 1 \} \).
  - See for example Theorem 3 in [16].
Examples

\( \ell_2 \)-norm-squared

\[ f(x) = \frac{1}{2} \|x\|^2 \Rightarrow f^*(y) = \max_x \langle y, x \rangle - \frac{1}{2} \|x\|^2. \]

○ Take the derivative and equate to 0: \( 0 = y - \lambda x \iff x = \frac{1}{\lambda} y \iff f^*(y) = \frac{1}{\lambda} \|y\|^2 - \frac{1}{2\lambda} \|y\|^2 = \frac{1}{2\lambda} \|y\|^2. \)

\( \ell_1 \)-norm

\[ f(x) = \lambda \|x\|_1 \Rightarrow f^*(y) = \max_x \langle y, x \rangle - \lambda \|x\|_1. \]

○ By definition of the \( \ell_1 \)-norm: \( f^*(y) = \max_x \sum_{i=1}^n y_i x_i - \lambda |x_i| = \max_x \sum_{i=1}^n y_i \text{sign}(x_i) |x_i| - \lambda |x_i|. \)

○ By inspection:

▶ If all \( |y_i| \leq \lambda \), then \( \forall i, (y_i \text{sign}(x_i) - \lambda) |x_i| \leq 0. \) Taking \( x = 0 \) gives the maximum value: \( f^*(y) = 0. \)

▶ If for at least one \( i, |y_i| > \lambda \), \( (y_i \text{sign}(x_i) - \lambda) |x_i| \rightarrow +\infty \) as \( |x_i| \rightarrow +\infty. \)

○ \( f^*(y) = \delta_{y:\|\cdot\|_\infty \leq \lambda}(y) = \begin{cases} 0, & \text{if } \|y\|_\infty \leq \lambda \\ +\infty, & \text{if } \|y\|_\infty > \lambda \end{cases} \)

Remark:

○ See advanced material at the end for non-convex examples, such as \( f(x) = \|x\|_0. \)
General nonsmooth problems

\[
\min_{x \in \mathbb{R}^p} f(x) + g(Ax)
\]

- By Fenchel-conjugation, we have \( g(Ax) = \max_y \langle Ax, y \rangle - g^*(y) \), where \( g^* \) is the conjugate of \( g \).

- Min-max formulation:

\[
\min_{x \in \mathbb{R}^p} f(x) + g(Ax) = \min_{x \in \mathbb{R}^p} \max_y \{ \Phi(x, y) := f(x) + \langle Ax, y \rangle - g^*(y) \}
\]

An example with linear constraints

- If \( g(Ax) = \delta_{\{b\}}(Ax) = \begin{cases} 0, & \text{if } Ax = b, \\ +\infty, & \text{if } Ax \neq b, \end{cases} \)

  \[
  \Rightarrow g^*(y) = \max_x \langle y, x \rangle - \delta_{\{b\}}(x) = \max_{x : x = b} \langle y, x \rangle = \langle y, b \rangle.
  \]

- We reach the minimax formulation (or the so-called “Lagrangian”) via conjugation:

\[
\min_x \{ f(x) : Ax = b \} = \min_x f(x) + g(Ax) = \min_x \max_y f(x) + \langle Ax - b, y \rangle.
\]
A special case in minimax optimization

Bilinear min-max template

$$\min_{x \in X} \max_{y \in Y} f(x) + \langle Ax, y \rangle - h(y),$$

where $X \subseteq \mathbb{R}^p$ and $Y \subseteq \mathbb{R}^n$.

- $f : X \to \mathbb{R}$ is convex.
- $h : Y \to \mathbb{R}$ is convex.
Example: Sparse recovery

An example from sparseland $b = Ax^\dagger + w$: constrained formulation

The basis pursuit denoising (BPDN) formulation is given by

$$x^* \in \arg \min_{x \in \mathbb{R}^p} \left\{ \|x\|_1 : \|Ax - b\|_2 \leq \|w\|_2, \|x\|_\infty \leq 1 \right\}.$$  \hspace{1cm} (BPDN)

A primal problem prototype

$$f^* := \min_{x \in \mathbb{R}^p} \left\{ f(x) : Ax - b \in \mathcal{K}, x \in \mathcal{X} \right\},$$

The above template captures BPDN formulation with

- $f(x) = \|x\|_1$.
- $\mathcal{K} = \{\|u\| \in \mathbb{R}^n : \|u\| \leq \|w\|_2\}$.
- $\mathcal{X} = \{x \in \mathbb{R}^p : \|x\|_\infty \leq 1\}$. 

Mathematics of Data | Prof. Volkan Cevher, volkan.cevher@epfl.ch

EPFL
An alternative formulation

**A primal problem** prototype

\[
f^* := \min_{x \in \mathbb{R}^p} \left\{ f(x) : Ax - b \in \mathcal{K}, \ x \in \mathcal{X} \right\},
\]

\[ (4) \]

- \( f \) is a proper, closed and convex function
- \( \mathcal{X} \) and \( \mathcal{K} \) are nonempty, closed convex sets
- \( A \in \mathbb{R}^{n \times p} \) and \( b \in \mathbb{R}^n \) are known
- An optimal solution \( x^* \) to (4) satisfies \( f(x^*) = f^* \), \( Ax^* - b \in \mathcal{K} \) and \( x^* \in \mathcal{X} \)

**A simplified template without loss of generality**

\[
f^* := \min_{x \in \mathbb{R}^p} \left\{ f(x) : Ax = b \right\},
\]

\[ (5) \]

- \( f \) is a proper, closed and convex function
- \( A \in \mathbb{R}^{n \times p} \) and \( b \in \mathbb{R}^n \) are known
- An optimal solution \( x^* \) to (5) satisfies \( f(x^*) = f^* \), \( Ax^* = b \)
Reformulation between templates

A primal problem template

$$\min_{x \in \mathbb{R}^p} \left\{ f(x) : Ax - b \in \mathcal{K}, x \in \mathcal{X} \right\}.$$ 

First step: Let $r_1 = Ax - b \in \mathbb{R}^n$ and $r_2 = x \in \mathbb{R}^p$.

$$\min_{x, r_1, r_2} \left\{ f(x) : r_1 \in \mathcal{K}, r_2 \in \mathcal{X}, Ax - b = r_1, x = r_2 \right\}.$$ 

Define $z = \begin{bmatrix} x \\ r_1 \\ r_2 \end{bmatrix} \in \mathbb{R}^{2p+n}$, $\bar{A} = \begin{bmatrix} A & -I_{n \times n} & 0_{n \times p} \\ I_{p \times p} & 0_{p \times n} & -I_{p \times p} \end{bmatrix}$, $\bar{b} = \begin{bmatrix} b \\ 0 \end{bmatrix}$, $\bar{f}(z) = f(x) + \delta_{\mathcal{K}}(r_1) + \delta_{\mathcal{X}}(r_2)$, where $\delta_{\mathcal{X}}(x) = 0$, if $x \in \mathcal{X}$, and $\delta_{\mathcal{X}}(x) = +\infty$, o/w.

The simplified template

$$\min_{z \in \mathbb{R}^{2p+n}} \left\{ \bar{f}(z) : \bar{A}z = \bar{b} \right\}.$$
From constrained formulation back to minimax

A general template

\[
\min_{x \in \mathbb{R}^p} \{ f(x) : Ax = b \}.
\]

Other examples:

- **Standard convex optimization** formulations: *linear programming*, *convex quadratic programming*, *second order cone programming*, *semidefinite programming* and *geometric programming*.
- **Reformulations** of existing unconstrained problems via **convex splitting**: *composite convex minimization*, *consensus optimization*, ... 

Formulating as min-max

\[
\max_{y \in \mathbb{R}^n} \langle y, Ax - b \rangle = \begin{cases} 
0, & \text{if } Ax = b, \\
+\infty, & \text{if } Ax \neq b.
\end{cases}
\]

\[
\min_{x \in \mathbb{R}^p} \{ f(x) : Ax = b \} = \min_{x \in \mathbb{R}^p} \max_{y \in \mathbb{R}^n} \{ \Phi(x, y) := f(x) + \langle y, Ax - b \rangle \}
\]
Dual problem

\[
\min_{x \in \mathbb{R}^p} \{ f(x) : Ax = b \} = \min_{x \in \mathbb{R}^p} \max_{y \in \mathbb{R}^n} \{ \Phi(x, y) := f(x) + \langle y, Ax - b \rangle \}
\]

- We define the dual problem

\[
\max d(y) := \max_{y \in \mathbb{R}^n} \{ \min_{x \in \mathbb{R}^p} f(x) + \langle y, Ax - b \rangle \}.
\]

Concavity of dual problem

Even if \( f(x) \) is not convex, \( d(y) \) is concave:

- For each \( x \), \( d(y) \) is linear; i.e., it is both convex and concave.
- Pointwise minimum of concave functions is still concave.

Remark:
- If we can exchange \( \min \) and \( \max \), we obtain a \textit{concave} maximization problem.
Example: Nonsmoothness of the dual function

- Consider a constrained convex problem:

\[
\begin{align*}
\min_{x \in \mathbb{R}^3} & \quad \left\{ f(x) := x_1^2 + 2x_2 \right\}, \\
\text{s.t.} & \quad 2x_3 - x_1 - x_2 = 1, \\
& \quad x \in \mathcal{X} := [-2, 2] \times [-2, 2] \times [0, 2].
\end{align*}
\]

- The dual function is concave and nonsmooth as written and then illustrated below.

\[
d(\lambda) := \min_{x \in \mathcal{X}} \left\{ x_1^2 + 2x_2 + \lambda(2x_3 - x_1 - x_2 - 1) \right\}
\]
Exchanging min and max: A dangerous proposal

- Weak duality:

\[
\max d(y) =: \max_{y \in \mathbb{R}^n} \min_{x \in \mathbb{R}^p} \Phi(x, y) \leq \min_{x \in \mathbb{R}^p} \max_{y \in \mathbb{R}^n} \Phi(x, y) = \min_{x \in \mathbb{R}^p} \{ f(x) : Ax = b \} = \begin{cases} f^*, & \text{if } Ax = b \\ +\infty, & \text{if } Ax \neq b \end{cases}
\]
A proof of weak duality

\[
f^* := \min_{x \in \mathbb{R}^p} \left\{ f(x) : Ax = b \right\} = \min_{x \in \mathbb{R}^p} \max_{y \in \mathbb{R}^n} \{ \Phi(x, y) := f(x) + \langle y, Ax - b \rangle \}
\]

○ Since \( Ax^* = b \), it holds for any \( y \)

\[
\Phi(x^*, y) = f^* = f(x^*) + \langle y, Ax^* - b \rangle \\
\geq \min_{x \in \mathbb{R}^p} \left\{ f(x) + \langle y, Ax - b \rangle \right\} \\
= \min_{x \in \mathbb{R}^p} \Phi(x, y).
\]

○ Take maximum of both sides in \( y \) and note that \( f^* \) is independent of \( y \):

\[
f^* = \min_{x \in \mathbb{R}^p} \max_{y \in \mathbb{R}^n} \Phi(x, y) \geq \max_{y \in \mathbb{R}^n} \min_{x \in \mathbb{R}^p} \Phi(x, y) =: \max_{y \in \mathbb{R}^n} d(y) = d^*.
\]
Strong duality and saddle points

**Strong duality**

\[
 f^* = f(x^*) = \min_{x \in \mathbb{R}^p} \max_{y \in \mathbb{R}^n} \Phi(x, y) = \max_{y \in \mathbb{R}^n} \min_{x \in \mathbb{R}^p} \Phi(x, y) =: \max_{y \in \mathbb{R}^n} d(y) = d^*. 
\]

Under strong duality and assuming existence of \(x^*\), \(\Phi(x, y)\) has a saddle point. We have primal and dual optimal values coincide, i.e., \(f^* = d^*\).
Strong duality and saddle points

**Strong duality**

\[
\begin{align*}
    f^* &= f(x^*) = \min_{x \in \mathbb{R}^p} \max_{y \in \mathbb{R}^n} \Phi(x, y) = \max_{y \in \mathbb{R}^n} \min_{x \in \mathbb{R}^p} \Phi(x, y) =: \max_{y \in \mathbb{R}^n} d(y) = d^*.
\end{align*}
\]

Under strong duality and assuming existence of \( x^* \), \( \Phi(x, y) \) has a saddle point. We have primal and dual optimal values coincide, i.e., \( f^* = d^* \).

**Recall saddle point / LNE**

A point \((x^*, y^*) \in \mathbb{R}^p \times \mathbb{R}^n\) is called a **saddle point** of \( \Phi \) if

\[
\Phi(x^*, y) \leq \Phi(x^*, y^*) \leq \Phi(x, y^*), \; \forall x \in \mathbb{R}^p, \; y \in \mathbb{R}^n.
\]
Toy example: Strong duality

**Primal problem**
- Consider the following primal minimization problem: \( \min_x P(x) := f(x) + g(x) := \frac{1}{2} \|x\|^2 + \|x\|_1 \)
- Using conjugation and strong duality
  \[
  P(x^*) = \min_x P(x) = \min_x \max_y f(x) + \langle x, y \rangle - g^*(y), \quad \text{by conjugation}
  = \max_y -g^*(y) + \min_x f(x) + \langle x, y \rangle, \quad \text{by changing min-max}
  = \max_y -g^*(y) - \max_x \langle x, -y \rangle - f(x), \quad \text{by } \min f = -\max -f
  = \max_y -g^*(y) - f^*(-y), \quad \text{by conjugation.}
  \]

**Dual problem**
- Dual problem: \( d^* = \max_y d(y) = -g^*(y) - f^*(-y) \)
- Recall \( f^*(-y) = \frac{1}{2} \|y\|^2 \) and \( g^*(y) = \delta_{y: \|y\|_\infty \leq 1}(y) \).
Toy example: Strong duality

Primal problem: \[ \min_x P(x) = \frac{1}{2} \|x\|^2 + \|x\|_1 \]

Dual problem: \[ \max_y -\frac{1}{2} \|y\|^2 - \delta_{\|y\|_\infty \leq 1}(y) \]

\[ d(y) = \begin{cases} -\frac{1}{2} \|y\|^2, & \text{if } \|y\|_\infty \leq 1 \\ -\infty, & \text{if } \|y\|_\infty > 1 \end{cases} \]
Back to convex-concave: Necessary and sufficient condition for strong duality

- Existence of a saddle point is not automatic even in convex-concave setting!
- Recall the minimax template:

$$\min_{x \in \mathbb{R}^p} \max_{y \in \mathbb{R}^n} \{ \Phi(x, y) := f(x) + \langle y, Ax - b \rangle \}$$

**Theorem (Necessary and sufficient optimality condition)**

*Under the Slater’s condition:* \( \text{relint}(\text{dom } f) \cap \{ x : Ax = b \} \neq \emptyset \), *strong duality holds,* where the primal and dual problems are given by

$$f^* := \begin{cases} \min_{x \in \mathbb{R}^p} f(x) \\
\text{s.t. } Ax = b, \end{cases} \quad \text{and} \quad d^* := \max_{y \in \mathbb{R}^n} d(y).$$

**Remarks:**

- By definition of \( f^* \) and \( d^* \), we always have \( d^* \leq f^* \) \((\text{weak duality})\).
- If a primal solution exists and the Slater’s condition holds, we have \( d^* = f^* \) \((\text{strong duality})\).
Slater’s qualification condition

- Denote $\text{relint}(\text{dom } f)$ the relative interior of the domain.
- The Slater condition requires
  \[ \text{relint}(\text{dom } f) \cap \{ x : Ax = b \} \neq \emptyset. \]  

(6)

Special cases

- If $\text{dom } f = \mathbb{R}^p$, then (6) $\iff \exists \bar{x} : A\bar{x} = b$.
- If $\text{dom } f = \mathbb{R}^p$ and instead of $Ax = b$, we have the feasible set $\{ x : h(x) \leq 0 \}$, where $h$ is $\mathbb{R}^p \to \mathbb{R}^q$ is convex, then
  \[ (6) \iff \exists \bar{x} : h(\bar{x}) < 0. \]
Example: Slater’s condition

Example

Let us consider solving $\min_{x \in D_\alpha} f(x)$ and so the feasible set is $D_\alpha := \mathcal{X} \cap A_\alpha$, where

$$\mathcal{X} := \{ x \in \mathbb{R}^2 : x_1^2 + x_2^2 \leq 1 \}, \quad A_\alpha := \{ x \in \mathbb{R}^2 : x_1 + x_2 = \alpha \},$$

where $\alpha \in \mathbb{R}$. 
Example: Slater’s condition

Example

Let us consider solving \( \min_{x \in \mathcal{D}_\alpha} f(x) \) and so the feasible set is \( \mathcal{D}_\alpha := \mathcal{X} \cap \mathcal{A}_\alpha \), where

\[ \mathcal{X} := \{ x \in \mathbb{R}^2 : x_1^2 + x_2^2 \leq 1 \}, \quad \mathcal{A}_\alpha := \{ x \in \mathbb{R}^2 : x_1 + x_2 = \alpha \}, \]

where \( \alpha \in \mathbb{R} \).

Two cases where Slater’s condition holds and does not hold

\( \mathcal{D}_{1/2} \) satisfies Slater’s condition – \( \mathcal{D}_{\sqrt{2}} \) does not satisfy Slater’s condition
Performance of optimization algorithms

\[ f^* := \min_{x \in \mathbb{R}^p} \left\{ f(x) : Ax = b, \right\}, \]

(Affine-Constrained)

**Exact vs. approximate solutions**

- Computing an **exact solution** \( x^* \) to (Affine-Constrained) is **impracticable**
- Algorithms seek \( x_\varepsilon^* \) that **approximates** \( x^* \) up to \( \varepsilon \) in some sense

**A performance metric: Time-to-reach \( \varepsilon \)**

\[ \text{time-to-reach } \varepsilon = \text{number of iterations to reach } \varepsilon \times \text{per iteration time} \]

**A key issue: Number of iterations to reach \( \varepsilon \)**

The notion of \( \varepsilon \)-accuracy is elusive in constrained optimization!
Numerical $\epsilon$-accuracy

- **Unconstrained case:** All iterates are feasible (no advantage from infeasibility)!
  \[ f(x^*_\epsilon) - f^* \leq \epsilon \]
  \[ f^* = \min_{x \in \mathbb{R}^p} f(x) \]

- **Constrained case:** We need to also measure the infeasibility of the iterates!
  \[ f^* - f(x^*_\epsilon) \leq \epsilon \]
  \[ f^* = \min_{x \in \mathbb{R}^p} \left\{ f(x) : Ax = b \right\} \] (7)

**Our definition of $\epsilon$-accurate solutions [22]**

Given a numerical tolerance $\epsilon \geq 0$, a point $x^*_\epsilon \in \mathbb{R}^p$ is called an $\epsilon$-solution of (7) if

\[
\begin{align*}
  f(x^*_\epsilon) - f^* & \leq \epsilon \text{ (objective residual)}, \\
  \|Ax^*_\epsilon - b\| & \leq \epsilon \text{ (feasibility gap)},
\end{align*}
\]

- When $x^*$ is unique, we can also obtain $\|x^*_\epsilon - x^*\| \leq \epsilon$ (iterate residual).
Numerical $\epsilon$-accuracy

Constrained problems

Given a numerical tolerance $\epsilon \geq 0$, a point $x_\epsilon^* \in \mathbb{R}^p$ is called an $\epsilon$-solution of (7) if

$$
\begin{aligned}
&f(x_\epsilon^*) - f^* \leq \epsilon \text{ (objective residual)}, \\
&\|Ax_\epsilon^* - b\| \leq \epsilon \text{ (feasibility gap)},
\end{aligned}
$$

- When $x^*$ is unique, we can also obtain $\|x_\epsilon^* - x^*\| \leq \epsilon$ (iterate residual).

General minimax problems

Since duality gap is 0 at the solution, we measure the primal-dual gap

$$
\text{Gap}(\bar{x}, \bar{y}) = \max_{y \in Y} \Phi(\bar{x}, y) - \min_{x \in X} \Phi(x, \bar{y}) \leq \epsilon. \quad (8)
$$

Remarks:

- $\epsilon$ can be different for the objective, feasibility gap, or the iterate residual.
- It is easy to show $\text{Gap}(x, y) \geq 0$ and $\text{Gap}(\bar{x}, \bar{y}) = 0$ iff $(\bar{x}, \bar{y})$ is a saddle point.
Primal-dual gap function for nonsmooth minimization

\[
\begin{align*}
\min_{x \in \mathcal{X}} f(x) + g(Ax) = \min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} & \left( f(x) + \langle Ax, y \rangle - g^*(y) \right) = \\
\max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} & \left( f(x) + \langle Ax, y \rangle - g^*(y) \right) = \Phi(x, y)
\end{align*}
\]

- Primal problem: \( \min_{x \in \mathcal{X}} P(x) \) where
  \[
  P(x) = \max_{y \in \mathcal{Y}} \Phi(x, y).
  \]

- Dual problem: \( \max_{y \in \mathcal{Y}} d(y) \) where
  \[
  d(y) = \min_{x \in \mathcal{X}} \Phi(x, y).
  \]

- The primal-dual gap, i.e., \(\text{Gap}(\bar{x}, \bar{y})\), is literally (primal value at \(\bar{x}\)) \(-\) (dual value at \(\bar{y}\)):
  \[
  \text{Gap}(\bar{x}, \bar{y}) = P(\bar{x}) - d(\bar{y}) = \max_{y \in \mathcal{Y}} \Phi(\bar{x}, y) - \min_{x \in \mathcal{X}} \Phi(x, \bar{y}).
  \]
Toy example for nonnegativity of gap

\( P(x) = \frac{1}{2} \| x \|^2 + \| x \|_1 \)

\( d(y) = -\frac{1}{2} \| y \|^2 - \delta_{y: \| y \|_\infty \leq 1}(y) \)

Recall the indicator function

\[ \delta_{y: \| y \|_\infty \leq 1}(y) = \begin{cases} 0, & \text{if } \| y \|_\infty \leq 1 \\ +\infty, & \text{if } \| y \|_\infty > 1 \end{cases} \]
Primal-dual gap function in the general case

\[
\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \Phi(x, y) = \max_{y \in \mathcal{Y}} \min_{x \in \mathcal{X}} \Phi(x, y)
\]

- Saddle point \((x^*, y^*)\) is such that \(\forall x \in \mathbb{R}^p, \forall y \in \mathbb{R}^n:\)
  \[
  \Phi(x^*, y) \leq \Phi(x^*, y^*) \leq \Phi(x, y^*).
  \]

- Nonnegativity of Gap:
  \[
  \text{Gap}(\bar{x}, \bar{y}) = \max_{y \in \mathcal{Y}} \Phi(\bar{x}, y) - \min_{x \in \mathcal{X}} \Phi(x, \bar{y}) \\
  \geq \Phi(\bar{x}, y^*) - \min_{x \in \mathcal{X}} \Phi(x, \bar{y}), \quad \text{by the definition of maximization} \\
  \geq \Phi(x^*, y^*) - \min_{x \in \mathcal{X}} \Phi(x, \bar{y}), \quad \text{by the inequality (**)} \\
  \geq \Phi(x^*, \bar{y}) - \min_{x \in \mathcal{X}} \Phi(x, \bar{y}), \quad \text{by the inequality (*)} \\
  \geq 0, \quad \text{by the definition of minimization.}
  \]

- If \((\bar{x}, \bar{y}) = (x^*, y^*)\), then all the inequalities will be equalities and \(\text{Gap}(\bar{x}, \bar{y}) = 0\).
Optimality conditions for minimax

**Saddle point**

We say \((x^*, y^*)\) is a primal-dual solution corresponding to primal and dual problems

\[
 f^* := \begin{cases} 
 \min_{x \in \mathbb{R}^p} \ f(x) & \text{and} \\
 \text{s.t.} \ Ax = b, 
\end{cases}
\]

and \(d^* := \max_{y \in \mathbb{R}^n} d(y) = \max_{y \in \mathbb{R}^n} \min_x \Phi(x, y).\)

if it is a saddle point of \(\Phi(x, y) = f(x) + \langle y, Ax - b \rangle:\)

\[
\Phi(x^*, y) \leq \Phi(x^*, y^*) \leq \Phi(x, y^*), \quad \forall x \in \mathbb{R}^p, \ y \in \mathbb{R}^n.
\]

**Karush-Khun-Tucker (KKT) conditions**

Under our assumptions, an equivalent characterization of \((x^*, y^*)\) is via the KKT conditions of the problem

\[
\min_{x \in \mathbb{R}^p} f(x) : Ax = b,
\]

which reads

\[
\begin{cases} 
 0 \in \partial_x \Phi(x^*, y^*) = A^T y^* + \partial f(x^*), \\
 0 = \nabla_y \Phi(x^*, \lambda^*) = Ax^* - b.
\end{cases}
\]
A naive proposal: Gradient descent-ascent (GDA)

Towards algorithms for minimax optimization

\[
\min_{x \in X} \max_{y \in Y} \Phi(x, y).
\]

We assume that

1. \(\Phi(\cdot, y)\) is convex,
2. \(\Phi(x, \cdot)\) is concave,
3. \(\Phi\) is smooth in the following sense:

\[
\left\| \begin{bmatrix} \nabla_x \Phi(x_1, y_1) \\ -\nabla_y \Phi(x_1, y_1) \end{bmatrix} - \begin{bmatrix} \nabla_x \Phi(x_2, y_2) \\ -\nabla_y \Phi(x_2, y_2) \end{bmatrix} \right\| \leq L \left\| \begin{bmatrix} x_1 - x_2 \\ y_1 - y_2 \end{bmatrix} \right\|. \tag{9}
\]

Let us try to use gradient descent for \(x\), gradient ascent for \(y\) to obtain a solution

**GDA**

1. Choose \(x^0, y^0\) and \(\tau\).
2. For \(k = 0, 1, \cdots\), perform:
   - \(x^{k+1} := x^k - \tau \nabla_x \Phi(x^k, y^k)\).
   - \(y^{k+1} := y^k + \tau \nabla_y \Phi(x^k, y^k)\).
GDA on a simple problem

Min-max problem

\[
\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \Phi(x, y).
\]

SimGDA

1. Choose \( x^0, y^0 \) and \( \tau \).
2. For \( k = 0, 1, \cdots \), perform:

\[
\begin{align*}
x^{k+1} &:= x^k - \tau \nabla_x \Phi(x^k, y^k), \\
y^{k+1} &:= y^k + \tau \nabla_y \Phi(x^k, y^k).
\end{align*}
\]

AltGDA

1. Choose \( x^0, y^0 \) and \( \tau \).
2. For \( k = 0, 1, \cdots \), perform:

\[
\begin{align*}
x^{k+1} &:= x^k - \tau \nabla_x \Phi(x^k, y^k), \\
y^{k+1} &:= y^k + \tau \nabla_y \Phi(x^{k+1}, y^k).
\end{align*}
\]

Example [9]

Let \( \Phi(x, y) = xy \), \( \mathcal{X} = \mathcal{Y} = \mathbb{R} \), then,

- for the iterates of SimGDA: \( x_{k+1}^2 + y_{k+1}^2 = (1 + \eta^2)(x_k^2 + y_k^2) \),
- for the iterates of AltGDA: \( x_{k+1}^2 + y_{k+1}^2 = C(x_0^2 + y_0^2) \).

\( \circ \) SimGDA diverges and AltGDA does not converge!
Practical performance

\[ \min_{x \in \mathbb{R}} \max_{y \in \mathbb{R}} xy \]

- Simultaneous GDA
- Alternating GDA
Between convex-concave and nonconvex-nonconcave

Nonconvex-concave problems

\[
\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \Phi(x, y)
\]

- \(\Phi(x, y)\) is nonconvex in \(x\), concave in \(y\), smooth in \(x\) and \(y\).

Recall

Define \(f(x) = \max_{y \in \mathcal{Y}} \Phi(x, y)\).

- Gradient descent applied to nonconvex \(f\) requires \(O(\epsilon^{-2})\) iterations to give an \(\epsilon\)-stationary point.
- (Sub)gradient of \(f\) can be computed using Danskin’s theorem:

\[
\nabla_x \Phi(\cdot, y^*(\cdot)) \in \partial f(\cdot), \text{ where } y^*(\cdot) \in \arg\max_{y \in \mathcal{Y}} \Phi(\cdot, y),
\]

which is tractable since \(\Phi\) is concave in \(y\) [19].

Remark:

- “Conceptually” much easier than nonconvex-nonconcave case.
### Epilogue

<table>
<thead>
<tr>
<th>Gradient complexity</th>
<th>Optimality measure</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>convex-concave</td>
<td>$O\left(\epsilon^{-1}\right)^1$</td>
<td>$\epsilon$ optimality w.r.t. duality gap</td>
</tr>
<tr>
<td>nonconvex-concave</td>
<td>$\tilde{O}\left(\epsilon^{-2.5}\right)^3$</td>
<td>$\epsilon$-stationarity w.r.t. gradient mapping norm</td>
</tr>
<tr>
<td>nonconvex-nonconcave</td>
<td>HARD</td>
<td>HARD</td>
</tr>
</tbody>
</table>

\(^1\)Rates are not directly comparable as duality gap and gradient mapping norm are not necessarily of the same order!


\(^3\)The rate is $\tilde{O}\left(\epsilon^{-2}\right)$ for strongly concave problems.


A new hope

\[ \min_{x \in \mathbb{R}} \max_{y \in \mathbb{R}} xy \]

- Next lecture: Some algorithms that actually **converge**!

- Convergence of the sequence:
  
  There exists \( z^* = (x^*, y^*) \), such that \( z_k \to z^* \).

- Convergence rate:

\[
\text{Gap} \left( \frac{1}{K} \sum_{k=1}^{K} x^k, \frac{1}{K} \sum_{k=1}^{K} y^k \right) \leq O \left( \frac{1}{K} \right).
\]
Wrap up!

- Try to finish Homework #2...
A **convex** proto-problem for **structured** sparsity

A combinatorial approach for estimating $x^\sharp$ from $b = Ax^\sharp + w$

We may consider the sparsest estimator or its surrogate with a valid sparsity pattern:

$$\hat{x} \in \arg\min_{x \in \mathbb{R}^p} \left\{ \|x\|_s : \|b - Ax\|_2 \leq \kappa, \|x\|_\infty \leq 1 \right\}$$  \hspace{1cm} (P_s)

with some $\kappa \geq 0$. If $\kappa = \|w\|_2$, then the structured sparse $x^\sharp$ is a feasible solution.

**Sparsity** and **structure** together [7]

Given some weights $d \in \mathbb{R}^d$, $e \in \mathbb{R}^p$ and an integer input $c \in \mathbb{Z}^l$, we define

$$\|x\|_s := \min_{\omega} \{d^T \omega + e^T s : M \begin{bmatrix} \omega \\ s \end{bmatrix} \leq c, \mathbb{1}_{\text{supp}(x)} = s, \omega \in \{0, 1\}^d\}$$

for all feasible $x$, $\infty$ otherwise. The parameter $\omega$ is useful for **latent** modeling.
A convex proto-problem for structured sparsity

A combinatorial approach for estimating $x^\sharp$ from $b = Ax^\sharp + w$

We may consider the sparsest estimator or its surrogate with a valid sparsity pattern:

$$\hat{x} \in \arg \min_{x \in \mathbb{R}^p} \left\{ \|x\|_s : \|b - Ax\|_2 \leq \kappa, \|x\|_\infty \leq 1 \right\}$$ (P_s)

with some $\kappa \geq 0$. If $\kappa = \|w\|_2$, then the structured sparse $x^\sharp$ is a feasible solution.

Sparsity and structure together [7]

Given some weights $d \in \mathbb{R}^d$, $e \in \mathbb{R}^p$ and an integer input $c \in \mathbb{Z}^l$, we define

$$\|x\|_s := \min_{\omega} \{d^T \omega + e^T s : M \omega \leq c, 1_{\text{supp}(x)} = s, \omega \in \{0, 1\}^d\}$$

for all feasible $x$, $\infty$ otherwise. The parameter $\omega$ is useful for latent modeling.

A convex candidate solution for $b = Ax^\sharp + w$

We use the convex estimator based on the tightest convex relaxation of $\|x\|_s$:

$$\hat{x} \in \arg \min_{x \in \text{dom}(\|\cdot\|_s)} \left\{ \|x\|_*^{**} : \|b - Ax\|_2 \leq \kappa \right\} \text{ with some } \kappa \geq 0, \text{ dom}(\|\cdot\|_s) := \{x : \|x\|_s < \infty\}.$$
Tractability & tightness of biconjugation

Proposition (Hardness of conjugation)

Let $F(s) : 2^\mathbb{R} \rightarrow \mathbb{R} \cup \{+\infty\}$ be a set function defined on the support $s = \text{supp}(x)$. Conjugate of $F$ over the unit infinity ball $\|x\|_\infty \leq 1$ is given by

$$g^*(y) = \sup_{s \in \{0,1\}^p} |y|^T s - F(s).$$

Observations:

- $F(s)$ is general set function
  
  Computation: NP-Hard

- $F(s) = \|x\|_s$
  
  Computation: Integer Linear Program (ILP) in general. However, if
  
  - $M$ is Totally Unimodular TU
  - $(M, c)$ is Total Dual Integral TDI
  
  then tight convex relaxations with a linear program (LP, which is “usually” tractable)

  Otherwise, relax to LP anyway!

- $F(s)$ is submodular
  
  Computation: Polynomial-time
Tree sparsity [15, 6, 3, 23]

**Structure:**  *We seek the sparsest signal with a rooted connected subtree support.*

**Linear description:** A valid support satisfy $s_{\text{parent}} \geq s_{\text{child}}$ over tree $\mathcal{T}$

$$T_{\text{supp}(x)} \triangleq Ts \geq 0$$

where $T$ is the directed edge-node incidence matrix, which is $TU$. 

Tree sparsity [15, 6, 3, 23]

**Structure:** We seek the sparsest signal with a rooted connected subtree support.

**Linear description:** A valid support satisfy \( s_{\text{parent}} \geq s_{\text{child}} \) over tree \( T \)

\[
T1_{\text{supp}(x)} := Ts \geq 0
\]

where \( T \) is the directed edge-node incidence matrix, which is \( TU \).

**Biconjugate:** \( \|x\|_* = \min_{s \in [0,1]^P} \{1^T s : Ts \geq 0, |x| \leq s \} \)

for \( x \in [-1, 1]^P \), \( \infty \) otherwise.
Tree sparsity [15, 6, 3, 23]

\( \mathcal{G}_H = \{\{1, 2, 3\}, \{2\}, \{3\}\} \)

valid selection of nodes

**Structure:** We seek the sparsest signal with a rooted connected subtree support.

**Linear description:** A valid support satisfy \( s_{\text{parent}} \geq s_{\text{child}} \) over tree \( T \)

\[
T 1_{\text{supp}(x)} := Ts \geq 0
\]

where \( T \) is the directed edge-node incidence matrix, which is \( TU \).

**Biconjugate:** \( \|x\|^{**} = \min_{s \in [0,1]^p} \left\{ T^Ts : Ts \geq 0, |x| \leq s \right\} = \sum_{G \in \mathcal{G}_H} \|x_G\|_{\infty} \) for \( x \in [-1,1]^p, \infty \) otherwise.

The set \( G \in \mathcal{G}_H \) are defined as each node and all its descendants.
Structure: We seek the sparsest signal with group allocation constraints.

Linear description: A valid support obeys budget constraints over $\mathcal{G}$

$\mathcal{B}^T s \leq c_u$

where $\mathcal{B}$ is the biadjacency matrix of $\mathcal{G}$, i.e., $\mathcal{B}_{ij} = 1$ iff $i$-th coefficient is in $\mathcal{G}_j$.

When $\mathcal{B}$ is an interval matrix or $\mathcal{G}$ has a loopless group intersection graph, it is TU.

Remark: We can also budget a lowerbound $c_\ell \leq \mathcal{B}^T s \leq c_u$. 
**Group knapsack sparsity [25, 10, 8]**

\[
\begin{bmatrix}
1 & 1 & \cdots & 1 & 1 & 1 & 0 & \cdots & 0 \\
0 & 1 & 1 & \cdots & 1 & 1 & 0 & \cdots & 0 \\
\vdots & & & & & & & & \vdots \\
0 & \cdots & 0 & 0 & 1 & 1 & \cdots & 1 & 1
\end{bmatrix}
\]

\((p - \Delta + 1) \times p\)

**Structure:** We seek the sparsest signal with group allocation constraints.

**Linear description:** A valid support obeys budget constraints over \(G\)

\[
\mathcal{B}^T s \leq c_u
\]

where \(\mathcal{B}\) is the biadjacency matrix of \(G\), i.e., \(\mathcal{B}_{ij} = 1\) iff \(i\)-th coefficient is in \(G_j\).

When \(\mathcal{B}\) is an interval matrix or \(G\) has a loopless group intersection graph, it is TU.

**Remark:** We can also budget a lowerbound \(c_\ell \leq \mathcal{B}^T s \leq c_u\).

**Biconjugate:** \[
\|x\|_{s}^{**} = \begin{cases} 
\|x\|_1 & \text{if } x \in [-1, 1]^p, \mathcal{B}^T|x| \leq c_u, \\
\infty & \text{otherwise}
\end{cases}
\]

For the neuronal spike example, we have \(c_u = 1\).
Group knapsack sparsity [25, 10, 8]

\[ \|x\|_{s}^{**} \leq 1 \text{ (middle)} \|x\|_{s}^{**} \leq 1.5 \text{ (right)} \|x\|_{s}^{**} \leq 2 \text{ for } \mathcal{G} = \{\{1, 2\}, \{2, 3\}\} \]

\textbf{Figure:} *

\begin{align*}
\text{(left) } & \|x\|_{s}^{**} \leq 1 \\
\text{(middle) } & \|x\|_{s}^{**} \leq 1.5 \\
\text{(right) } & \|x\|_{s}^{**} \leq 2
\end{align*}

\textbf{Structure:} We seek the sparsest signal with group allocation constraints.

\textbf{Linear description:} A valid support obeys budget constraints over $\mathcal{G}$

\[ B^T s \leq c_u \]

where $B$ is the biadjacency matrix of $\mathcal{G}$, i.e., $B_{ij} = 1$ iff $i$-th coefficient is in $G_j$.

When $B$ is an interval matrix or $\mathcal{G}$ has a loopless group intersection graph, it is TU.

\textbf{Remark:} We can also budget a lowerbound $c_{\ell} \leq B^T s \leq c_u$.

\textbf{Biconjugate:} $\|x\|_{s}^{**} = \begin{cases} 
\|x\|_1 & \text{if } x \in [-1, 1]^p, B^T|x| \leq c_u, \\
\infty & \text{otherwise}
\end{cases}$

For the neuronal spike example, we have $c_u = 1.$
Group knapsack sparsity example: A stylized spike train

- Basis pursuit (BP): $\|x\|_1$
- TU-relax (TU):

$$
\|x\|^*_{\mathcal{S}} = \begin{cases} 
\|x\|_1 & \text{if } x \in [-1, 1]^p, \mathcal{B}^T |x| \leq c_u, \\
\infty & \text{otherwise}
\end{cases}
$$

Figure: Recovery for $n = 0.18p$. 

relative errors: 

$$
\frac{\|x^\mathcal{G} - x^{\text{BP}}\|_2}{\|x^\mathcal{G}\|_2} = .200, \quad \frac{\|x^\mathcal{G} - x^{\text{TU}}\|_2}{\|x^\mathcal{G}\|_2} = .067
$$
Group knapsack sparsity: A simple variation

Structure: We seek the signal with the minimal overall group allocation.

Objective: \( \mathbf{1}^T \mathbf{s} \to \| \mathbf{x} \|_\omega = \begin{cases} \min_{\omega \in \mathbb{Z}_+} \omega & \text{if } \mathbf{x} \in [-1, 1]^p, \mathbf{B}^T \mathbf{s} \leq \omega \mathbf{1}, \\ \infty & \text{otherwise} \end{cases} \)

Linear description: A valid support obeys budget constraints over \( G \)

\[ \mathbf{B}^T \mathbf{s} \leq \omega \mathbf{1} \]

where \( \mathbf{B} \) is the biadjacency matrix of \( G \), i.e., \( \mathbf{B}_{ij} = 1 \) iff \( i \)-th coefficient is in \( G_j \).

When \( \mathbf{B} \) is an interval matrix or \( G \) has a loopless group intersection graph, it is TU.

Biconjugate: \( \| \mathbf{x} \|_{s^*} = \begin{cases} \max_{G \in G} \| \mathbf{x}^G \|_1 & \text{if } \mathbf{x} \in [-1, 1]^p, \\ \infty & \text{otherwise} \end{cases} \)

Remark: The regularizer is known as exclusive Lasso [25, 21].
**Group cover sparsity: Minimal group cover** [2, 20, 13]

Structure: We seek the signal covered by a minimal number of groups.

Objective: $1^T s \rightarrow d^T \omega$

Linear description: At least one group containing a sparse coefficient is selected

$\mathcal{B} \omega \geq s$

where $\mathcal{B}$ is the biadjacency matrix of $\mathcal{G}$, i.e., $\mathcal{B}_{ij} = 1$ iff $i$-th coefficient is in $\mathcal{G}_j$.

When $\mathcal{B}$ is an interval matrix, or $\mathcal{G}$ has a loopless group intersection graph it is TU.
**Group cover sparsity:** Minimal group cover [2, 20, 13]

![Diagram](image)

**Figure:** \( \mathcal{G} = \{\{1, 2\}, \{2, 3\}\} \), unit group weights \( \mathbf{d} = 1 \).

**Structure:** We seek the signal covered by a minimal number of groups.

**Objective:** \( \mathbf{1}^T \mathbf{s} \rightarrow \mathbf{d}^T \mathbf{\omega} \)

**Linear description:** At least one group containing a sparse coefficient is selected

\[
\mathbf{B}\mathbf{\omega} \geq \mathbf{s}
\]

where \( \mathbf{B} \) is the biadjacency matrix of \( \mathcal{G} \), i.e., \( \mathbf{B}_{ij} = 1 \) iff \( i \)-th coefficient is in \( \mathcal{G}_j \).

When \( \mathbf{B} \) is an interval matrix, or \( \mathcal{G} \) has a loopless group intersection graph it is TU.

**Biconjugate:** \( \|\mathbf{x}\|_{\omega}^{**} = \min_{\omega \in [0, 1]^M} \{\mathbf{d}^T \mathbf{\omega} : \mathbf{B}\mathbf{\omega} \geq |\mathbf{x}|\} \) for \( \mathbf{x} \in [-1, 1]^P, \infty \) otherwise
Group cover sparsity: **Minimal group cover** [2, 20, 13]

![Figure: $G = \{\{1, 2\}, \{2, 3\}\}$, unit group weights $d = 1$.](image)

**Structure:** We seek the signal covered by a minimal number of groups.

**Objective:** $1^T s \rightarrow d^T \omega$

**Linear description:** At least one group containing a sparse coefficient is selected

\[ \mathcal{B}\omega \geq s \]

where $\mathcal{B}$ is the biadjacency matrix of $G$, i.e., $\mathcal{B}_{ij} = 1$ iff $i$-th coefficient is in $G_j$.

When $\mathcal{B}$ is an interval matrix, or $G$ has a **loopless** group intersection graph it is **TU**.

**Biconjugate:** $\|x\|_{\omega^*} = \min_{\omega \in [0, 1]^M} \{d^T \omega : \mathcal{B}\omega \geq |x|\}$ for $x \in [-1, 1]^p$, $\infty$ otherwise

\[ \star \min_{\mathcal{V}_i \in \mathbb{R}^p} \{ \sum_{i=1}^M d_i \|v_i\|_{\infty} : x = \sum_{i=1}^M v_i, \forall \text{supp}(v_i) \subseteq G_i \}, \]
Group cover sparsity: *Minimal group cover* [2, 20, 13]

Figure: $\mathcal{G} = \{\{1, 2\}, \{2, 3\}\}$, unit group weights $d = 1$.

**Structure:** *We seek the signal covered by a minimal number of groups.*

**Objective:** $1^T s \rightarrow d^T \omega$

**Linear description:** At least one group containing a sparse coefficient is selected

$$\mathcal{B} \omega \geq s$$

where $\mathcal{B}$ is the biadjacency matrix of $\mathcal{G}$, i.e., $\mathcal{B}_{ij} = 1$ iff $i$-th coefficient is in $\mathcal{G}_j$.

When $\mathcal{B}$ is an interval matrix, or $\mathcal{G}$ has a *loopless* group intersection graph it is *TU*.

**Biconjugate:** $\|x\|^*_\omega = \min_{\omega \in [0,1]^M} \{d^T \omega : \mathcal{B} \omega \geq |x|\}$ for $x \in [-1,1]^p$, $\infty$ otherwise

$$\succeq \min_{v_i \in \mathbb{R}^p} \{\sum_{i=1}^M d_i\|v_i\|_\infty : x = \sum_{i=1}^M v_i, \forall \text{supp}(v_i) \subseteq \mathcal{G}_i\},$$

**Remark:** Weights $d$ can depend on the sparsity within each groups (not TU) [7].
**Budgeted group cover sparsity**

Structure: *We seek the sparsest signal covered by $G$ groups.*

**Objective:** $d^T \omega \rightarrow 1^T s$

**Linear description:** At least one of the $G$ selected groups cover each sparse coefficient.

$$\mathcal{B} \omega \geq s, 1^T \omega \leq G$$

where $\mathcal{B}$ is the biadjacency matrix of $\mathcal{G}$, i.e., $\mathcal{B}_{ij} = 1$ iff $i$-th coefficient is in $\mathcal{G}_j$.

When $\begin{bmatrix} \mathcal{B} \\ 1 \end{bmatrix}$ is an interval matrix, it is TU.
**Budgeted** group cover sparsity

**Structure:** We seek the sparsest signal covered by $G$ groups.

**Objective:** $d^T \omega \rightarrow 1^T s$

**Linear description:** At least one of the $G$ selected groups cover each sparse coefficient.

\[
\mathcal{B} \omega \geq s, 1^T \omega \leq G
\]

where $\mathcal{B}$ is the biadjacency matrix of $\mathcal{G}$, i.e., $\mathcal{B}_{ij} = 1$ iff $i$-th coefficient is in $\mathcal{G}_j$.

When \[
\begin{bmatrix}
\mathcal{B} \\
1
\end{bmatrix}
\]
is an interval matrix, it is TU.

**Biconjugate:** $\|x\|^{**} = \min_{\omega \in [0,1]^M} \{\|x\|_1 : \mathcal{B} \omega \geq |x|, 1^T \omega \leq G\}$

for $x \in [-1, 1]^P$, $\infty$ otherwise.
Budgeted group cover example: Interval overlapping groups

- Basis pursuit (BP): $\|x\|_1$
- Sparse group Lasso (SGL$_q$):
  \[
  (1 - \alpha) \sum_{G \in G} \sqrt{|G|} \|x^G\|_q + \alpha \|x^G\|_1
  \]
- TU-relax (TU):
  \[
  \|x\|^{**}_\omega = \min_{\omega \in [0,1]^M} \{\|x\|_1 : \exists \omega \geq |x|, 1^T \omega \leq G\}
  \]
  for $x \in [-1,1]^p$, $\infty$ otherwise.

**Figure:** Recovery for $n = 0.25p$, $s = 15$, $p = 200$, $G = 5$ out of $M = 29$ groups.
Group intersection sparsity [14, 24, 1]

Structure: We seek the signal intersecting with minimal number of groups.

Objective: $1^T s \rightarrow d^T \omega$

Linear description: All groups containing a sparse coefficient are selected

$$H_k s \leq \omega, \forall k \in \Psi$$

where $H_k(i, j) = \begin{cases} 1 & \text{if } j = k, j \in G_i \\ 0 & \text{otherwise} \end{cases}$, which is TU.
Group intersection sparsity $[14, 24, 1]$

$\mathcal{G} = \{\{1, 2\}, \{2, 3\}\}$, unit group weights $d = 1$

(left) intersection (right) cover.

**Structure:** We seek the signal intersecting with minimal number of groups.

**Objective:** $1^T s \rightarrow d^T \omega$

**Linear description:** All groups containing a sparse coefficient are selected

$$H_k s \leq \omega, \forall k \in \Psi$$

where $H_k(i, j) = \begin{cases} 
1 & \text{if } j = k, j \in G_i \\
0 & \text{otherwise}
\end{cases}$, which is TU.

**Biconjugate:** $\|x\|_{\omega}^{**} = \min_{\omega \in [0, 1]^M} \left\{ d^T \omega : H_k |x| \leq \omega, \forall k \in \Psi \right\}$

for $x \in [-1, 1]^p$, $\infty$ otherwise.
Group intersection sparsity [14, 24, 1]

$\mathcal{G} = \{\{1, 2\}, \{2, 3\}\}$, unit group weights $d = 1$
(left) intersection (right) cover.

**Structure:** We seek the signal intersecting with minimal number of groups.

**Objective:** $1^T s \rightarrow d^T \omega$ (submodular)

**Linear description:** All groups containing a sparse coefficient are selected

$$H_k s \leq \omega, \forall k \in \mathcal{G}$$

where $H_k(i, j) = \begin{cases} 1 & \text{if } j = k, j \in G_i \\ 0 & \text{otherwise} \end{cases}$, which is TU.

**Biconjugate:** $\|x\|^{**} = \min_{\omega \in [0, 1]^M} \{d^T \omega : H_k|x| \leq \omega, \forall k \in \mathcal{G}\} \equiv \sum_{G \in \mathcal{G}} \|x_G\|_{\infty}$

for $x \in [-1, 1]^p$, $\infty$ otherwise.
Group intersection sparsity [14, 24, 1]

\[ \mathcal{G} = \{\{1, 2, 3\}, \{2\}, \{3\}\} \], unit group weights \( d = 1 \).

**Structure:** We seek the signal intersecting with minimal number of groups.

**Objective:** \( 1^T s \rightarrow d^T \omega \) (submodular)

**Linear description:** All groups containing a sparse coefficient are selected

\[
H_k s \leq \omega, \forall k \in \Psi
\]

where \( H_k(i, j) = \begin{cases} 1 & \text{if } j = k, j \in G_i, \\
0 & \text{otherwise} \end{cases} \), which is TU.

**Biconjugate:** \( \|x\|^{**} = \min_{\omega \in [0,1]^M} \{d^T \omega : H_k |x| \leq \omega, \forall k \in \Psi\} = \sum_{G \in \mathcal{G}} \|x_G\|\infty \)

for \( x \in [-1,1]^p \), \( \infty \) otherwise.

**Remark:** For hierarchical \( \mathcal{G}_H \), group intersection and tree sparsity models coincide.
Beyond linear costs: Graph dispersiveness

Structure: We seek a signal dispersive over a given graph $G(\mathcal{P}, \mathcal{E})$

Objective: $\mathbf{1}^T s \rightarrow \sum_{(i,j) \in \mathcal{E}} s_i s_j$ (non-linear, supermodular function)

Linearization:

$$\|x\|_s = \min_{z \in \{0,1\}^|\mathcal{E}|} \left\{ \sum_{(i,j) \in \mathcal{E}} z_{ij} : z_{ij} \geq s_i + s_j - 1 \right\}$$

When edge-node incidence matrix of $G(\mathcal{P}, \mathcal{E})$ is TU (e.g., bipartite graphs), it is TU.
Beyond linear costs: Graph dispersiveness

Figure: (left) $\|x\|_{s}^{**} = 0$ (right) $\|x\|_{s}^{**} \leq 1$ for $E = \{\{1, 2\}, \{2, 3\}\}$ (chain graph)

**Structure:** We seek a signal dispersive over a given graph $G(\Psi, E)$

**Objective:** $1^T s \rightarrow \sum_{(i,j) \in E} s_i s_j$ (non-linear, supermodular function)

**Linearization:**

$$\|x\|_s = \min_{z \in \{0, 1\}^{|E|}} \{\sum_{(i,j) \in E} z_{ij} : z_{ij} \geq s_i + s_j - 1\}$$

When edge-node incidence matrix of $G(\Psi, E)$ is TU (e.g., bipartite graphs), it is TU.

**Biconjugate:** $\|x\|_{s}^{**} = \sum_{(i,j) \in E} (|x_i| + |x_j| - 1)^+ \text{ for } x \in [-1, 1]^p, \infty \text{ otherwise.}$
The difficulty of the nonconvex-nonconcave setting

**Definition (Local Nash Equilibrium)**

A pair of vectors \((x^*, y^*)\) with \(x^* \in A_x\) and \(y^* \in A_y\) is called \((\epsilon, \delta)\)-Local Nash Equilibrium if it holds that,

1. \(\Phi(x^*, y^*) \leq \Phi(x, y^*) + \epsilon\), for all \(x \in A_x\) with \(\|x - x^*\| \leq \delta\)
2. \(\Phi(x^*, y^*) \geq \Phi(x^*, y) - \epsilon\), for all \(y \in A_y\) with \(\|y - y^*\| \leq \delta\).

**Theorem [5]**

Deciding whether a function \(\Phi(x, y)\) admits an \((\epsilon, \delta)\)-Local Nash Equilibrium is \(NP\)-hard even for \((\epsilon, \delta) := (1/384, 1)\) and \((A_x, A_y) := ([0, 1]^{d_1}, [0, 1]^{d_2})\).
Reduction to 3-SAT(3)

**Definition (3-SAT(3))**

**Input**: A boolean CNF-formula \( \phi := (\phi_1, \ldots, \phi_m) \) with boolean variables \( x_1, \ldots, x_n \) such that every clause of \( \phi_j \) has at most 3 boolean variables and every boolean variable appears in at most 3 clauses.

**Output**: Return *Yes* if there exists an assignment of the boolean variables \( (x_1, \ldots, x_n) \) satisfying all clauses \( \{\phi_1, \ldots, \phi_m\} \) and *No* otherwise.

**Theorem [18]**

3−SAT(3) is NP−complete.
Reducing \((\epsilon, \delta)\)-LNE to 3-SAT(3)

Constructing the Function

Given an instance of 3-SAT(3) \(\phi := (\phi_1, \ldots, \phi_m)\), we construct a function \(\Phi(\cdot)\) as follows,

- For each boolean variable \(x_i\) (there are \(n\) boolean variables \(x_i\)) we correspond a respective real-valued variable \(x_i\).
- For each clause \(\phi_j\), we construct a polynomial \(P_j(x)\) as follows,
  - let \(\ell_i, \ell_k, \ell_m\) denote the literals participating in \(\phi_j\).
  - Consider the polynomial \(P_j(x) = P_{ji}(x) \cdot P_{jk}(x) \cdot P_{jm}(x)\) where

\[
P_{ji}(x) = \begin{cases} 
1 - x_i & \text{if } \ell_i = x_i \\
 0 & \text{if } \ell_i = \neg x_i
\end{cases}
\]

Example

For the clause \(\phi_j = x_1 \lor \neg x_2 \lor x_3\rightarrow P(x) := (1 - x_1) \cdot x_2 \cdot x_3\).
Reducing \((\epsilon, \delta)\)-LNE to 3-SAT(3)

**Constructing the Function**

Given an instance of 3-SAT(3) \(\phi := (\phi_1, \ldots, \phi_m)\), we construct a function \(\Phi(\cdot)\) as follows,

- For each boolean variable \(x_i\) (there are \(n\) boolean variables \(x_i\)) we correspond a respective real-valued variable \(x_i\)
- For each clause \(\phi_j\), we construct a polynomial \(P_j(x)\) as follows,
  - let \(\ell_i, \ell_k, \ell_m\) denote the literals participating in \(\phi_j\).
  - \(P_j(x) = P_{ji}(x) \cdot P_{jk}(x) \cdot P_{jm}(x)\) where
    
    \[
    P_{ij}(x) = \begin{cases} 
    1 - x_i & \text{if } \ell_i = x_i \\
    x_i & \text{if } \ell_i = \neg x_i 
    \end{cases}
    \]

The overall constructed function is

\[
\Phi(x, w, y) = \sum_{j=1}^{m} P_j(x) \cdot (w_j - y_j)^2
\]

where each \(w_j, y_j\) are additional variables associated with clause \(\phi_j\).
Reducing $(\epsilon, \delta)$-LNE to 3-SAT(3)

Lemma [5]

Let the minimizing player control $(x, w)$ and the maximizing player control $y$. A $(1/384, 1)$-Local Nash Equilibrium with $(x, w) \in [0, 1]^{n+m}$ and $y \in [0, 1]^m$ exists if and only if $\phi$ admits a satisfying assignment.
Proof of Lemma \(\rightarrow\)

Analysis

Let \(((x^*, w^*), y^*)\) an \((\epsilon, \delta)\)-Local NE for \(\epsilon = 1/384\) and \(\delta = 1\).

- \(P_j(x^*) \leq 16 \cdot \epsilon\) for all \(j = 1, \ldots, m\).

  - Let \(P_j(x^*) > 16 \cdot \epsilon\) for some \(j = 1, \ldots, m\)
    - If \(|w_j^* - y_j^*| \geq 1/4\) then the \textit{min player} can decrease \(\Phi(x, w, y)\) by at least \(\epsilon\) by setting \(w_j := y_j^*\).
    - If \(|w_j^* - y_j^*| \leq 1/4\) then the \textit{max player} can increase \(\Phi(x, w, y)\) by at least \(\epsilon\) by moving \(y_j\) to either 0 or 1.

- Randomly assign each boolean variable \(x_i\) to True or False with
  \[\Pr[x_i \text{ is set to True}] = x_i^*\]

- By the definition of \(P_j(x)\),
  \[\Pr[\phi_j \text{ is not satisfied}] = P_j(x^*) \leq 16 \cdot \epsilon = 1/24\]

- Since each boolean variable participates in at most 3 clauses. Each clause \(\phi_j\) \textit{shares boolean variables} with at most other 6 clauses. Since \(\Pr[\phi_j \text{ is not satisfied}] \leq 1/24\) by the \textit{Lovász Local Lemma},
  \[\Pr[\text{there exists an unsatisfied clause } \phi_j] < 1\]

Thus, there exists a satisfying assignment.
Proof of Lemma (←−)

Analysis

Let $x^* := (x_1^*, \ldots, x_n^*)$ be a satisfying boolean assignment for $\phi := (\phi_1, \ldots, \phi_m)$.

- If $x_i^* = \text{True}$ then we set the real-valued variable $x_i^*$ to 1.
- If $x_i^* = \text{False}$ then we set the real-valued variable $x_i^*$ to 0.
- Since each clause $\phi_j$ is satisfied then (by the definition of $P_j(x)$),
  \[ P_j(x^*) = 0 \quad \text{for all } j = 1, \ldots, m \]

Thus, all vectors $((x^*, w), y)$ are $(0, 1)$-Local Nash Equilibrium.
References I


Single-pixel imaging via compressive sampling. 

A totally unimodular view of structured sparsity.

*Spiking neuron models: Single neurons, populations, plasticity.*

A variational inequality perspective on generative adversarial networks.

Compressive sensing recovery of spike trains using a structured sparsity model.
References III

Convergence without convexity: Sampling, optimization, and games.  

The limits of min-max optimization algorithms: convergence to spurious non-critical sets.  

Learning with structured sparsity.  

Multi-scale mining of fmri data with hierarchical structured sparsity.  
In Pattern Recognition in NeuroImaging (PRNI), 2011.

Proximal methods for hierarchical sparse coding.  
References IV

[16] Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari.  
Regularization techniques for learning with matrices.  

Robust reinforcement learning via adversarial training with langevin dynamics.  

Reducibility among combinatorial problems.  

Near-optimal algorithms for minimax optimization.  

[20] G. Obozinski, L. Jacob, and J.P. Vert.  
Group lasso with overlaps: The latent group lasso approach.  
Joint covariate selection and joint subspace selection for multiple classification problems.  

[22] Q. Tran-Dinh and V. Cevher.  
Constrained convex minimization via model-based excessive gap.  
In *Proc. the Neural Information Processing Systems Foundation conference (NIPS2014)*, pages 1–9,  
Montreal, Canada, December 2014.

[23] Peng Zhao, Guilherme Rocha, and Bin Yu.  
Grouped and hierarchical model selection through composite absolute penalties.  

On model selection consistency of Lasso.  

Association screening of common and rare genetic variants by penalized regression.  