Mathematics of Data: From Theory to Computation

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 13: Primal-dual optimization I

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

EE-556 (Fall 2025)

License Information for Mathematics of Data Slides

- 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

General nonsmooth problems

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

$$\min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x}) + g(\mathbf{A}\mathbf{x}).$$

- f, g are convex, nonsmooth functions; and $\mathbf{A}$ is a linear operator.

Examples

- $g(\mathbf{A}\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_1$ or $g(\mathbf{A}\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2$.
- $g(\mathbf{A}\mathbf{x}) = \delta_{\{\mathbf{b}\}}(\mathbf{A}\mathbf{x})$, where $\delta_{\{\mathbf{b}\}}(\mathbf{A}\mathbf{x}) = \begin{cases} 0, & \text{if } \mathbf{A}\mathbf{x} = \mathbf{b}, \\ +\infty, & \text{if } \mathbf{A}\mathbf{x} \neq \mathbf{b}. \end{cases}$

Observations: ◦ The indicator example covers constrained problems, such as $\min_{\mathbf{x}\in\mathcal{X}}\{f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b}\}$.

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

$$f^*(\mathbf{y}) = \sup_{\mathbf{x} \in \mathrm{dom}(f)} \left\{ \mathbf{y}^T \mathbf{x} - f(\mathbf{x}) \right\}$$

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

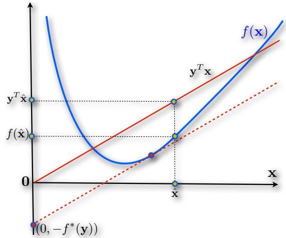


Figure: The conjugate function $f^*(\mathbf{y})$ is the maximum gap between the linear function $\mathbf{x}^T\mathbf{y}$ (red line) and $f(\mathbf{x})$.

Observations:
- ◦ $\mathbf{y}$: slope of the hyperplane
- ◦ $-f^*(\mathbf{y})$: intercept of the hyperplane

Conjugation of functions

Definition

Given a proper, closed and convex function $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$, the function $f^* : Q^* \to \mathbb{R} \cup \{+\infty\}$ such that

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

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

Conjugation of functions

Definition

Given a proper, closed and convex function $f : \mathcal{Q} \to \mathbb{R} \cup \{+\infty\}$, the function $f^* : Q^* \to \mathbb{R} \cup \{+\infty\}$ such that

$$f^*(\mathbf{y}) = \sup_{\mathbf{x} \in \mathrm{dom}(f)} \left\{ \mathbf{y}^T \mathbf{x} - f(\mathbf{x}) \right\}$$

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 $\mathbf{y}$.
- The conjugate of the conjugate of a convex function f is the same function f; i.e., $f^{**} = f$ for $f \in \mathcal{F}(\mathcal{Q})$.
- The conjugate of the conjugate of a non-convex function f is its lower convex envelope when $\mathcal{Q}$ is compact:
  - $f^{**}(\mathbf{x}) = \sup\{g(\mathbf{x}) : g$ is convex and $g \leq f, \forall \mathbf{x} \in \mathcal{Q}\}$.
- For closed convex f, μ-strong convexity w.r.t. $\|\cdot\|$ is equivalent to $\frac{1}{\mu}$ smoothness of f^* w.r.t. $\|\cdot\|_*$.
  - Recall dual norm: $\|\mathbf{y}\|_* = \sup_{\mathbf{x}}\{\langle \mathbf{x}, \mathbf{y}\rangle : \|\mathbf{x}\| \leq 1\}$.
  - See for example Theorem 3 in [12].

Examples

ℓ_2-norm-squared

$f(\mathbf{x}) = \frac{\lambda}{2}\|\mathbf{x}\|^2 \Rightarrow f^*(\mathbf{y}) = \max_{\mathbf{x}} \langle \mathbf{y}, \mathbf{x}\rangle - \frac{\lambda}{2}\|\mathbf{x}\|^2.$

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

ℓ_1-norm

$f(\mathbf{x}) = \lambda\|\mathbf{x}\|_1 \Rightarrow f^*(\mathbf{y}) = \max_{\mathbf{x}} \langle \mathbf{y}, \mathbf{x}\rangle - \lambda\|\mathbf{x}\|_1.$

∘ By definition of the ℓ_1-norm: $f^*(\mathbf{y}) = \max_{\mathbf{x}} \sum_{i=1}^n y_i x_i - \lambda|x_i| = \max_{\mathbf{x}} \sum_{i=1}^n y_i \mathrm{sign}(x_i)|x_i| - \lambda|x_i|.$

∘ By inspection:

- If all $|y_i| \leq \lambda$, then $\forall i, (y_i\mathrm{sign}(x_i) - \lambda)|x_i| \leq 0$. Taking $\mathbf{x} = 0$ gives the maximum value: $f^*(\mathbf{y}) = 0$.
- If for at least one $i, |y_i| > \lambda, (y_i\mathrm{sign}(x_i) - \lambda)|x_i| \to +\infty$ as $|x_i| \to +\infty$.

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

Remark: ∘ See advanced material at the end for non-convex examples, such as $f(\mathbf{x}) = \|\mathbf{x}\|_0$.

General nonsmooth problems

$$\min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x}) + g(\mathbf{A}\mathbf{x})$$

- By Fenchel-conjugation, we have $g(\mathbf{A}\mathbf{x}) = \max_{\mathbf{y}} \langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle - g^*(\mathbf{y})$, where g^* is the conjugate of g.
- Min-max formulation:

$$\min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x}) + g(\mathbf{A}\mathbf{x}) = \min_{\mathbf{x}\in\mathbb{R}^p} \max_{\mathbf{y}} \{\Phi(\mathbf{x},\mathbf{y}) := f(\mathbf{x}) + \langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle - g^*(\mathbf{y})\}$$

An example with linear constraints

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

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

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

$$\min_{\mathbf{x}} \{f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b}\} = \min_{\mathbf{x}} f(\mathbf{x}) + g(\mathbf{A}\mathbf{x}) = \min_{\mathbf{x}} \max_{\mathbf{y}} f(\mathbf{x}) + \langle \mathbf{A}\mathbf{x} - \mathbf{b}, \mathbf{y}\rangle.$$

A special case in minimax optimization

Bilinear min-max template

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

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

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

Example: Sparse recovery

An example from sparseland $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$: constrained formulation

The basis pursuit denoising (BPDN) formulation is given by

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \{\|\mathbf{x}\|_1 : \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \leq \|\mathbf{w}\|_2, \|\mathbf{x}\|_\infty \leq 1\}. \quad \text{(BPDN)}$$

A **primal problem** prototype

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

The above template captures BPDN formulation with

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

An alternative formulation

A **primal problem** prototype

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} - \mathbf{b} \in \mathcal{K},\ \mathbf{x} \in \mathcal{X} \right\}, \tag{1}$$

- f is a proper, closed and convex function
- $\mathcal{X}$ and $\mathcal{K}$ are nonempty, closed convex sets
- $\mathbf{A} \in \mathbb{R}^{n\times p}$ and $\mathbf{b} \in \mathbb{R}^n$ are known
- An optimal solution $\mathbf{x}^\star$ to (1) satisfies $f(\mathbf{x}^\star) = f^\star$, $\mathbf{A}\mathbf{x}^\star - \mathbf{b} \in \mathcal{K}$ and $\mathbf{x}^\star \in \mathcal{X}$

A simplified template without loss of generality

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b} \right\}, \tag{2}$$

- f is a proper, closed and convex function
- $\mathbf{A} \in \mathbb{R}^{n\times p}$ and $\mathbf{b} \in \mathbb{R}^n$ are known
- An optimal solution $\mathbf{x}^\star$ to (2) satisfies $f(\mathbf{x}^\star) = f^\star$, $\mathbf{A}\mathbf{x}^\star = \mathbf{b}$

Reformulation between templates

A primal problem template

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

First step: Let $\mathbf{r}_1 = \mathbf{A}\mathbf{x} - \mathbf{b} \in \mathbb{R}^n$ and $\mathbf{r}_2 = \mathbf{x} \in \mathbb{R}^p$.

$$\min_{\mathbf{x},\mathbf{r}_1,\mathbf{r}_2} \left\{ f(\mathbf{x}) : \mathbf{r}_1 \in \mathcal{K}, \mathbf{r}_2 \in \mathcal{X}, \mathbf{A}\mathbf{x} - \mathbf{b} = \mathbf{r}_1, \mathbf{x} = \mathbf{r}_2 \right\}.$$

- Define $\mathbf{z} = \begin{bmatrix}\mathbf{x}\\ \mathbf{r}_1 \\ \mathbf{r}_2\end{bmatrix} \in \mathbb{R}^{2p+n}$, $\bar{\mathbf{A}} = \begin{bmatrix}\mathbf{A} & -\mathbf{I}_{n\times n} & \mathbf{0}_{n\times p} \\ \mathbf{I}_{p\times p} & \mathbf{0}_{p\times n} & -\mathbf{I}_{p\times p}\end{bmatrix}$, $\bar{\mathbf{b}} = \begin{bmatrix}\mathbf{b}\\ \mathbf{0}\end{bmatrix}$, $\bar{f}(\mathbf{z}) = f(\mathbf{x}) + \delta_{\mathcal{K}}(\mathbf{r}_1) + \delta_{\mathcal{X}}(\mathbf{r}_2)$,
  where $\delta_{\mathcal{X}}(\mathbf{x}) = 0$, if $\mathbf{x} \in \mathcal{X}$, and $\delta_{\mathcal{X}}(\mathbf{x}) = +\infty$, o/w.

The simplified template

$$\min_{\mathbf{z}\in\mathbb{R}^{2p+n}} \left\{ \bar{f}(\mathbf{z}) : \bar{\mathbf{A}}\mathbf{z} = \bar{\mathbf{b}} \right\}.$$

From constrained formulation back to minimax

A general template

$$\min_{\mathbf{x}\in\mathbb{R}^p} \{f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b}\}.$$

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

$$\max_{\mathbf{y}\in\mathbb{R}^n} \langle \mathbf{y}, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle = \begin{cases} 0, & \text{if } \mathbf{A}\mathbf{x} = \mathbf{b}, \\ +\infty, & \text{if } \mathbf{A}\mathbf{x} \neq \mathbf{b}. \end{cases}$$

$$\boxed{\min_{\mathbf{x}\in\mathbb{R}^p} \left\{f(\mathbf{x})\colon \mathbf{A}\mathbf{x} = \mathbf{b}\right\} = \min_{\mathbf{x}\in\mathbb{R}^p} \max_{\mathbf{y}\in\mathbb{R}^n} \left\{\Phi(\mathbf{x},\mathbf{y}) := f(\mathbf{x}) + \langle \mathbf{y}, \mathbf{A}\mathbf{x} - \mathbf{b}\rangle\right\}}$$

Dual problem

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

◦ We define the dual problem

$$\max_{\mathbf{y}\in\mathbb{R}^n} d(\mathbf{y}) := \max_{\mathbf{y}\in\mathbb{R}^n}\{\underbrace{\min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x}) + \langle \mathbf{y}, \mathbf{A}\mathbf{x}-\mathbf{b}\rangle}_{d(\mathbf{y})}\}.$$

Concavity of dual problem

Even if $f(\mathbf{x})$ is not convex, $d(\mathbf{y})$ is concave:

- For each $\mathbf{x}$, $d(\mathbf{y})$ 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 **concave** maximization problem.

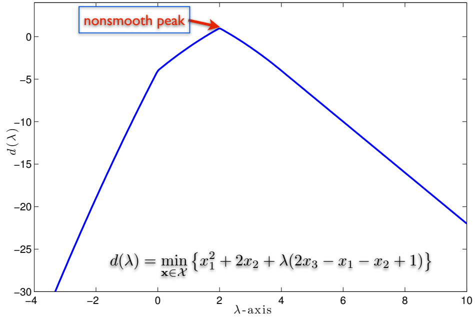
Example: Nonsmoothness of the dual function

- Consider a constrained convex problem:

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

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

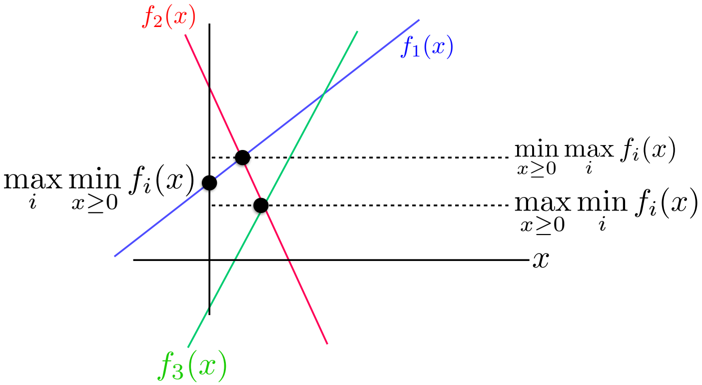
$$d(\lambda) := \min_{\mathbf{x}\in\mathcal{X}} \left\{ x_1^2 + 2x_2 + \lambda(2x_3 - x_1 - x_2 - 1) \right\}$$



Exchanging $\min$ and $\max$: A dangerous proposal

◦ Weak duality:

$$\underbrace{\max_{\mathbf{y}\in\mathbb{R}^n} d(\mathbf{y})}_{\text{Dual problem}} =: \boxed{\max_{\mathbf{y}\in\mathbb{R}^n}\min_{\mathbf{x}\in\mathbb{R}^p} \Phi(\mathbf{x},\mathbf{y}) \le \min_{\mathbf{x}\in\mathbb{R}^p}\max_{\mathbf{y}\in\mathbb{R}^n} \Phi(\mathbf{x},\mathbf{y})} = \underbrace{\min_{\mathbf{x}\in\mathbb{R}^p}\left\{f(\mathbf{x}) : \mathbf{A}\mathbf{x}=\mathbf{b}\right\}}_{\text{Primal problem}} = \begin{cases} f^\star, \text{ if } \mathbf{A}\mathbf{x}=\mathbf{b} \\ +\infty, \text{ if } \mathbf{A}\mathbf{x}\neq\mathbf{b} \end{cases}$$



A proof of weak duality

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

◦ Since $\mathbf{A}\mathbf{x}^\star = \mathbf{b}$, it holds for any $\mathbf{y}$

$$\begin{aligned}
\Phi(\mathbf{x}^\star,\mathbf{y}) = f^\star &= f(\mathbf{x}^\star) + \langle \mathbf{y}, \mathbf{A}\mathbf{x}^\star - \mathbf{b} \rangle \\
&\geq \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) + \langle \mathbf{y}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle \right\} \\
&= \min_{\mathbf{x}\in\mathbb{R}^p} \Phi(\mathbf{x},\mathbf{y}).
\end{aligned}$$

◦ Take maximum of both sides in $\mathbf{y}$ and note that $f^\star$ is independent of $\mathbf{y}$:

$$f^\star = \min_{\mathbf{x}\in\mathbb{R}^p} \max_{\mathbf{y}\in\mathbb{R}^n} \Phi(\mathbf{x},\mathbf{y}) \geq \max_{\mathbf{y}\in\mathbb{R}^n} \min_{\mathbf{x}\in\mathbb{R}^p} \Phi(\mathbf{x},\mathbf{y}) =: \max_{\mathbf{y}\in\mathbb{R}^n} d(\mathbf{y}) = d^\star.$$

Strong duality and saddle points

Strong duality

$$f^\star = f(\mathbf{x}^\star) = \min_{\mathbf{x}\in\mathbb{R}^p} \max_{\mathbf{y}\in\mathbb{R}^n} \Phi(\mathbf{x},\mathbf{y}) = \max_{\mathbf{y}\in\mathbb{R}^n} \min_{\mathbf{x}\in\mathbb{R}^p} \Phi(\mathbf{x},\mathbf{y}) =: \max_{\mathbf{y}\in\mathbb{R}^n} d(\mathbf{y}) = d^\star.$$

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

Strong duality and saddle points

Strong duality

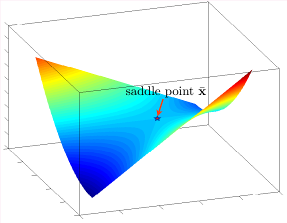
$$f^\star = f(\mathbf{x}^\star) = \min_{\mathbf{x}\in\mathbb{R}^p} \max_{\mathbf{y}\in\mathbb{R}^n} \Phi(\mathbf{x},\mathbf{y}) = \max_{\mathbf{y}\in\mathbb{R}^n} \min_{\mathbf{x}\in\mathbb{R}^p} \Phi(\mathbf{x},\mathbf{y}) =: \max_{\mathbf{y}\in\mathbb{R}^n} d(\mathbf{y}) = d^\star.$$

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

Recall saddle point / LNE

A point $(\mathbf{x}^\star, \mathbf{y}^\star) \in \mathbb{R}^p \times \mathbb{R}^n$ is called a saddle point of Φ if

$$\Phi(\mathbf{x}^\star,\mathbf{y}) \leq \Phi(\mathbf{x}^\star,\mathbf{y}^\star) \leq \Phi(\mathbf{x},\mathbf{y}^\star), \ \forall \mathbf{x} \in \mathbb{R}^p, \ \mathbf{y} \in \mathbb{R}^n.$$



Toy example: Strong duality

Primal problem

- Consider the following primal minimization problem: $\min_{\mathbf{x}} P(\mathbf{x}) := f(\mathbf{x}) + g(\mathbf{x}) := \frac{1}{2}\|\mathbf{x}\|^2 + \|\mathbf{x}\|_1$
- Using conjugation and strong duality

$$\begin{aligned}
P(\mathbf{x}^\star) = \min_{\mathbf{x}} P(\mathbf{x}) &= \min_{\mathbf{x}} \max_{\mathbf{y}} f(\mathbf{x}) + \langle \mathbf{x}, \mathbf{y} \rangle - g^*(\mathbf{y}), && \text{by conjugation} \\
&= \max_{\mathbf{y}} -g^*(\mathbf{y}) + \min_{\mathbf{x}} f(\mathbf{x}) + \langle \mathbf{x}, \mathbf{y} \rangle, && \text{by changing min-max} \\
&= \max_{\mathbf{y}} -g^*(\mathbf{y}) - \max_{\mathbf{x}} \langle \mathbf{x}, -\mathbf{y} \rangle - f(\mathbf{x}), && \text{by } \min f = -\max -f \\
&= \max_{\mathbf{y}} -g^*(\mathbf{y}) - f^*(-\mathbf{y}), && \text{by conjugation.}
\end{aligned}$$

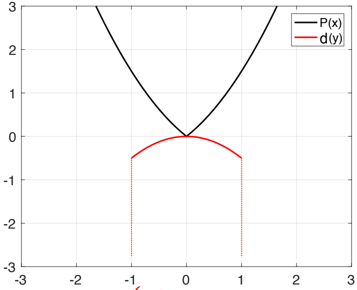
Dual problem

- Dual problem: $d^\star = \max_{\mathbf{y}} d(\mathbf{y}) = -g^*(\mathbf{y}) - f^*(-\mathbf{y})$
- Recall $f^*(-\mathbf{y}) = \frac{1}{2}\|\mathbf{y}\|^2$ and $g^*(\mathbf{y}) = \delta_{\mathbf{y}:\|\mathbf{y}\|_\infty \leq 1}(\mathbf{y})$.

Toy example: Strong duality

Primal problem: $\min_{\mathbf{x}} P(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2 + \|\mathbf{x}\|_1$

Dual problem: $\max_{\mathbf{y}} -\frac{1}{2}\|\mathbf{y}\|^2 - \delta_{\mathbf{y}:\|\mathbf{y}\|_\infty \leq 1}(\mathbf{y})$



$$d(\mathbf{y}) = \begin{cases} -\frac{1}{2}\|\mathbf{y}\|^2, & \text{if } \|\mathbf{y}\|_\infty \leq 1 \\ -\infty, & \text{if } \|\mathbf{y}\|_\infty > 1 \end{cases}$$

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:

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

Theorem (Necessary and sufficient optimality condition)

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

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

Remarks:

◦ By definition of $f^\star$ and $d^\star$, we always have $\boxed{d^\star \leq f^\star}$ (**weak duality**).

◦ If a primal solution exists and the Slater's condition holds, we have $\boxed{d^\star = f^\star}$ (**strong duality**).

Slater's qualification condition

- Denote $\text{relint}(\text{dom } f)$ the relative interior of the domain.
- The Slater condition requires

$$\boxed{\text{relint}(\text{dom } f) \cap \{\mathbf{x} \; : \; \mathbf{A}\mathbf{x} = \mathbf{b}\} \neq \emptyset.} \tag{3}$$

Special cases

- If $\text{dom } f = \mathbb{R}^p$, then (3) $\Leftrightarrow$ $\boxed{\exists \bar{\mathbf{x}} \; : \; \mathbf{A}\bar{\mathbf{x}} = \mathbf{b}}$.
- If $\text{dom } f = \mathbb{R}^p$ and instead of $\mathbf{A}\mathbf{x} = \mathbf{b}$, we have the feasible set $\{\mathbf{x} : h(\mathbf{x}) \leq 0\}$, where h is $\mathbb{R}^p \to R^q$ is convex, then

$$(3) \Leftrightarrow \boxed{\exists \bar{\mathbf{x}} \; : \; h(\bar{\mathbf{x}}) < 0.}$$

Example: Slater's condition

Example

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

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

where $\alpha \in \mathbb{R}$.

Example: Slater's condition

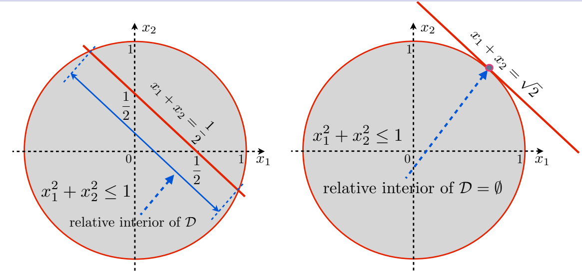
Example

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

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

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

$$f^\star := \min_{\mathbf{x}\in\mathbb{R}^p} \Big\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b}, \Big\}, \qquad \text{(Affine-Constrained)}$$

Exact vs. approximate solutions

- Computing an **exact solution** $\mathbf{x}^\star$ to (Affine-Constrained) is **impracticable**
- Algorithms seek $\mathbf{x}^\star_\epsilon$ that approximates $\mathbf{x}^\star$ up to ϵ in some sense

A performance metric: Time-to-reach ϵ

time-to-reach ϵ = number of iterations to reach ϵ $\times$ per iteration time

A key issue: Number of iterations to reach ϵ

The notion of ϵ-accuracy is elusive in constrained optimization!

Numerical ϵ-accuracy

- **Unconstrained case:** All iterates are feasible (no advantage from infeasibility)!

$$f(\mathbf{x}^\star_\epsilon) - f^\star \le \epsilon$$

$$\boxed{f^\star = \min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x})}$$

- **Constrained case:** We need to also measure the infeasibility of the iterates!

$$f^\star - f(\mathbf{x}^\star_\epsilon) \le \epsilon \quad !!!$$

$$\boxed{f^\star = \min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b} \right\}} \tag{4}$$

Our definition of ϵ-accurate solutions [16]

Given a numerical tolerance $\epsilon \ge 0$, a point $\mathbf{x}^\star_\epsilon \in \mathbb{R}^p$ is called an ϵ-solution of (4) if

$$\begin{cases} f(\mathbf{x}^\star_\epsilon) - f^\star & \le \epsilon \text{ (objective residual)}, \\ \|\mathbf{A}\mathbf{x}^\star_\epsilon - \mathbf{b}\| & \le \epsilon \text{ (feasibility gap)}, \end{cases}$$

- When $\mathbf{x}^\star$ is unique, we can also obtain $\|\mathbf{x}^\star_\epsilon - \mathbf{x}^\star\| \le \epsilon$ (iterate residual).

Numerical ϵ-accuracy

Constrained problems

Given a numerical tolerance $\epsilon \geq 0$, a point $\mathbf{x}^\star_\epsilon \in \mathbb{R}^p$ is called an ϵ-solution of (4) if

$$\begin{cases} f(\mathbf{x}^\star_\epsilon) - f^\star & \leq \epsilon \text{ (objective residual)}, \\ \|\mathbf{A}\mathbf{x}^\star_\epsilon - \mathbf{b}\| & \leq \epsilon \text{ (feasibility gap)}, \end{cases}$$

- When $\mathbf{x}^\star$ is unique, we can also obtain $\|\mathbf{x}^\star_\epsilon - \mathbf{x}^\star\| \leq \epsilon$ (iterate residual).

General minimax problems

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

$$\mathrm{Gap}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = \max_{\mathbf{y} \in \mathcal{Y}} \Phi(\bar{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x} \in \mathcal{X}} \Phi(\mathbf{x}, \bar{\mathbf{y}}) \leq \epsilon. \tag{5}$$

Remarks:

- ϵ can be different for the objective, feasibility gap, or the iterate residual.
- It is easy to show $\mathrm{Gap}(\mathbf{x}, \mathbf{y}) \geq 0$ and $\mathrm{Gap}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = 0$ iff $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ is a saddle point.

Primal-dual gap function for nonsmooth minimization

$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) + g(\mathbf{A}\mathbf{x}) = \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \underbrace{f(\mathbf{x}) + \langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle - g^*(\mathbf{y})}_{\Phi(\mathbf{x},\mathbf{y})} = \max_{\mathbf{y}\in\mathcal{Y}} \min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) + \langle \mathbf{A}\mathbf{x}, \mathbf{y}\rangle - g^*(\mathbf{y})$$

- Primal problem: $\min_{\mathbf{x}\in\mathcal{X}} P(\mathbf{x})$ where

$$P(\mathbf{x}) = \max_{\mathbf{y}\in\mathcal{Y}} \Phi(\mathbf{x},\mathbf{y}).$$

- Dual problem: $\max_{\mathbf{y}\in\mathcal{Y}} d(\mathbf{y})$ where

$$d(\mathbf{y}) = \min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x},\mathbf{y}).$$

- The primal-dual gap, i.e., $\mathrm{Gap}(\bar{\mathbf{x}},\bar{\mathbf{y}})$, is literally $(\text{primal value at } \bar{\mathbf{x}}) - (\text{dual value at } \bar{\mathbf{y}})$:

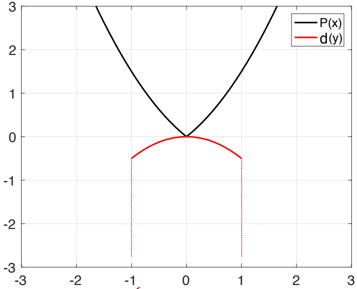
$$\mathrm{Gap}(\bar{\mathbf{x}},\bar{\mathbf{y}}) = P(\bar{\mathbf{x}}) - d(\bar{\mathbf{y}}) = \max_{\mathbf{y}\in\mathcal{Y}} \Phi(\bar{\mathbf{x}},\mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x},\bar{\mathbf{y}}).$$

Toy example for nonnegativity of gap

- $P(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2 + \|\mathbf{x}\|_1$
- $d(\mathbf{y}) = -\frac{1}{2}\|\mathbf{y}\|^2 - \delta_{\mathbf{y}:\|\mathbf{y}\|_\infty \le 1}(\mathbf{y})$

Recall the indicator function

$$\delta_{\mathbf{y}:\|\mathbf{y}\|_\infty \le 1}(\mathbf{y}) = \begin{cases} 0, \text{ if } \|\mathbf{y}\|_\infty \le 1 \\ +\infty, \text{ if } \|\mathbf{y}\|_\infty > 1 \end{cases}$$



$$d(\mathbf{y}) = \begin{cases} -\frac{1}{2}\|\mathbf{y}\|^2, & \text{if } \|\mathbf{y}\|_\infty \le 1 \\ -\infty, & \text{if } \|\mathbf{y}\|_\infty > 1 \end{cases}$$

Primal-dual gap function in the general case

$$\boxed{\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \Phi(\mathbf{x},\mathbf{y}) = \max_{\mathbf{y}\in\mathcal{Y}} \min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x},\mathbf{y})}$$

◦ Saddle point $(\mathbf{x}^\star, \mathbf{y}^\star)$ is such that $\forall \mathbf{x} \in \mathbb{R}^p$, $\forall \mathbf{y} \in \mathbb{R}^n$:

$$\Phi(\mathbf{x}^\star, \mathbf{y}) \overset{(*)}{\leq} \Phi(\mathbf{x}^\star, \mathbf{y}^\star) \overset{(**)}{\leq} \Phi(\mathbf{x}, \mathbf{y}^\star).$$

◦ Nonnegativity of Gap:

$$\begin{aligned}
\mathrm{Gap}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) &= \max_{\mathbf{y}\in\mathcal{X}} \Phi(\bar{\mathbf{x}}, \mathbf{y}) - \min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x}, \bar{\mathbf{y}}) \\
&\geq \Phi(\bar{\mathbf{x}}, \mathbf{y}^\star) - \min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x}, \bar{\mathbf{y}}), && \text{by the definition of maximization} \\
&\geq \Phi(\mathbf{x}^\star, \mathbf{y}^\star) - \min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x}, \bar{\mathbf{y}}), && \text{by the inequality } (**) \\
&\geq \Phi(\mathbf{x}^\star, \bar{\mathbf{y}}) - \min_{\mathbf{x}\in\mathcal{X}} \Phi(\mathbf{x}, \bar{\mathbf{y}}), && \text{by the inequality } (*) \\
&\geq 0, && \text{by the definition of minimization.}
\end{aligned}$$

◦ If $(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = (\mathbf{x}^\star, \mathbf{y}^\star)$, then all the inequalities will be equalities and $\mathrm{Gap}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = 0$.

Optimality conditions for minimax

Saddle point

We say $(\mathbf{x}^\star, \mathbf{y}^\star)$ is a primal-dual solution corresponding to primal and dual problems

$$f^\star := \begin{cases} \min\limits_{\mathbf{x}\in\mathbb{R}^p} & f(\mathbf{x}) \\ \text{s.t.} & \mathbf{A}\mathbf{x} = \mathbf{b}, \end{cases} \quad \text{and} \quad d^\star := \max_{\mathbf{y}\in\mathbb{R}^n} d(\mathbf{y}) = \max_{\mathbf{y}\in\mathbb{R}^n} \min_{\mathbf{x}} \Phi(\mathbf{x}, \mathbf{y}).$$

if it is a saddle point of $\Phi(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) + \langle \mathbf{y}, \mathbf{A}\mathbf{x} - \mathbf{b} \rangle$:

$$\Phi(\mathbf{x}^\star, \mathbf{y}) \leq \Phi(\mathbf{x}^\star, \mathbf{y}^\star) \leq \Phi(\mathbf{x}, \mathbf{y}^\star), \ \forall \mathbf{x} \in \mathbb{R}^p, \ \mathbf{y} \in \mathbb{R}^n.$$

Karush-Khun-Tucker (KKT) conditions

Under our assumptions, an equivalent characterization of $(\mathbf{x}^\star, \mathbf{y}^\star)$ is via the KKT conditions of the problem

$$\min_{\mathbf{x}\in\mathbb{R}^p} f(\mathbf{x}) : \mathbf{A}\mathbf{x} = \mathbf{b},$$

which reads

$$\begin{cases} 0 & \in \partial_{\mathbf{x}} \Phi(\mathbf{x}^\star, \mathbf{y}^\star) & = \mathbf{A}^T \mathbf{y}^\star + \partial f(\mathbf{x}^\star), \\ 0 & = \nabla_{\mathbf{y}} \Phi(\mathbf{x}^\star, \lambda^\star) & = \mathbf{A}\mathbf{x}^\star - \mathbf{b}. \end{cases}$$

Primal approach: The Penalty Method

$$\boxed{\min_{\mathbf{x}\in\mathbb{R}^p}\left\{f(\mathbf{x}) : \mathbf{A}\mathbf{x}=\mathbf{b}\right\}}$$

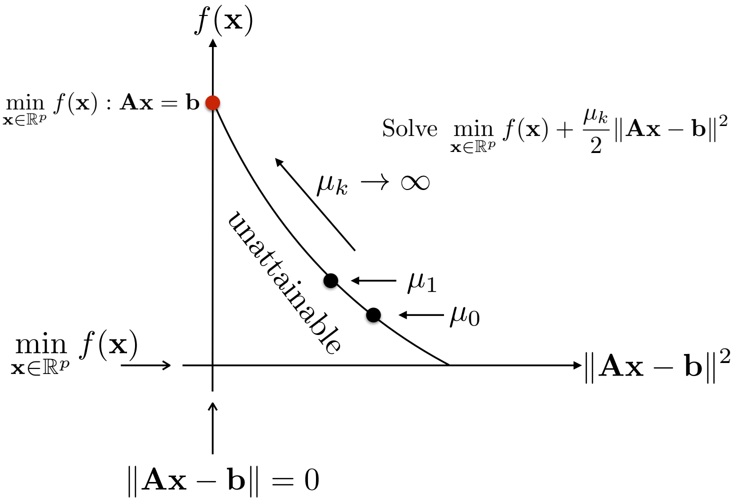
Penalty methods

- Convert constrained problem (difficult) to unconstrained (easy).
- Penalized function with penalty parameter $\mu > 0$:

$$F_\mu(\mathbf{x}) := \left\{f(\mathbf{x}) + \frac{\mu}{2}\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2\right\} \quad \overset{\mu\to\infty}{\Longleftrightarrow} \quad \min_{\mathbf{x}\in\mathbb{R}^p}\left\{f(\mathbf{x}) : \mathbf{A}\mathbf{x}=\mathbf{b}\right\}.$$

- **Observations**:
  - Minimize a weighted combination of $f(\mathbf{x})$ and $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2$ at the same time.
  - μ determines the weight of $\|\mathbf{A}\mathbf{x}-\mathbf{b}\|^2$.
  - As $\mu\to\infty$, we enforce $\mathbf{A}\mathbf{x}=\mathbf{b}$.
  - Other functions than the quadratic $\frac{1}{2}\|\cdot\|^2$ are also possible e.g., exact nonsmooth penalty functions:
    - $\mu\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2$ or $\mu\|\mathbf{A}\mathbf{x}-\mathbf{b}\|_1$
    - They work with finite μ, but they are difficult to solve [13, Section 17.2], [4]

Quadratic penalty: Intuition



Quadratic penalty: Conceptual algorithm

Quadratic penalty method (QP):
1. Choose $\mathbf{x}_0 \in \mathbb{R}^p$ and $\mu_0 > 0$.
2. For $k = 0, 1, \cdots$, perform:
2.a. $\mathbf{x}_k := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) + \frac{\mu_k}{2}\|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 \right\}$.
2.b. Update $\mu_{k+1} > \mu_k$.

Theorem [13, Theorem 17.1]

Assume that f is smooth and $\mu_k \to \infty$. Then, every limit point $\bar{\mathbf{x}}$ of the sequence $\{\mathbf{x}_k\}$ is a solution of the constrained problem

$$\mathbf{x}^\star \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) \colon \mathbf{A}\mathbf{x} = \mathbf{b} \right\}.$$

Limitation

- The minimization problems of step 2.a. of the algorithm become ill-conditioned as $\mu_k \to \infty$.
- Common improvements:
  - Solve the subproblem inexactly, *i.e.*, up to ϵ accuracy.
  - Linearization to simplify subproblems (up next).

Quadratic penalty: Linearization

Generalized quadratic penalty method:
1. Choose $\mathbf{x}_0 \in \mathbb{R}^p$, $\mu_0 > 0$ and positive semidefinite matrix $\mathbf{Q}_k$.
2. For $k = 0, 1, \cdots,$ perform:
2.a. $\mathbf{x}_k := \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ f(\mathbf{x}) + \frac{\mu_k}{2}\lVert \mathbf{A}\mathbf{x} - \mathbf{b}\rVert^2 + \frac{1}{2}\lVert \mathbf{x} - \mathbf{x}_{k-1}\rVert^2_{\mathbf{Q}_k} \right\}$.
2.b. Update $\mu_{k+1} > \mu_k$.

Ideas

- Minimize a majorizer of $F_\mu(\mathbf{x})$, parametrized by $\mathbf{Q}_k$ in step 2.a..
- $\mathbf{Q}_k = \mathbf{0}$ gives the standard QP; $\mathbf{Q}_k = \mathbf{I}$ gives strongly convex subproblems.
- $\mathbf{Q}_k = \alpha_k \mathbf{I} - \mu_k \mathbf{A}^\top \mathbf{A}$, with $\alpha_k \geq \mu_k \lVert \mathbf{A}\rVert^2$ gives

$$\mathbf{x}_k = \mathrm{prox}_{\frac{1}{\alpha_k} f}\left(\mathbf{x}_{k-1} - \frac{\mu_k}{\alpha_k}\mathbf{A}^\top(\mathbf{A}\mathbf{x}_{k-1} - \mathbf{b})\right)$$

Only one proximal operator!

and picking $\alpha_k = \mu_k \lVert \mathbf{A}\rVert^2$ gives

$$\mathbf{x}_k = \mathrm{prox}_{\frac{1}{\mu_k \lVert \mathbf{A}\rVert^2} f}\left(\mathbf{x}_{k-1} - \frac{1}{\lVert \mathbf{A}\rVert^2}\mathbf{A}^\top(\mathbf{A}\mathbf{x}_{k-1} - \mathbf{b})\right).$$

Per-iteration time: The key role of the prox-operator

Recall: Prox-operator

$$\mathrm{prox}_f(\mathbf{x}) := \arg\min_{\mathbf{z}\in\mathbb{R}^p} \left\{ f(\mathbf{z}) + \frac{1}{2}\|\mathbf{z}-\mathbf{x}\|^2 \right\}.$$

Key properties:

- single valued & non-expansive since f is a proper convex function.
- distributes when the primal problem has decomposable structure:

$$f(\mathbf{x}) := \sum_{i=1}^{m} f_i(\mathbf{x}_i), \quad \text{and} \quad \mathcal{X} := \mathcal{X}_1 \times \cdots \times \mathcal{X}_m.$$

  where $m \geq 1$ is the number of components.
- often efficient & has closed form expression. For instance, if $f(\mathbf{z}) = \|\mathbf{z}\|_1$, then the prox-operator performs coordinate-wise soft-thresholding by 1.

Quadratic penalty: Linearized methods

Linearized QP method (LQP)	Accelerated linearized QP method (ALQP)
1. Choose $\mathbf{x}_0 \in \mathbb{R}^p$, $\sigma_0 = 1$, $\mu_0 > 0$.	**1.** Choose $\mathbf{x}_0, \mathbf{y}_0 \in \mathbb{R}^p$, $\tau_0 = 1$, $\mu_0 > 0$.
2. For $k = 0, 1, \cdots$:	**2.** For $k = 0, 1, \cdots$:
2.a. $\mathbf{x}_{k+1} := \mathrm{prox}_{\frac{1}{\mu_k \Vert\mathbf{A}\Vert^2} f}\left(\mathbf{x}_k - \frac{1}{\Vert\mathbf{A}\Vert^2}\mathbf{A}^\top(\mathbf{A}\mathbf{x}_k - \mathbf{b})\right)$.	**2.a.** $\mathbf{x}_{k+1} := \mathrm{prox}_{\frac{1}{\mu_k \Vert\mathbf{A}\Vert^2} f}\left(\mathbf{y}_k - \frac{1}{\Vert\mathbf{A}\Vert^2}\mathbf{A}^\top(\mathbf{A}\mathbf{y}_k - \mathbf{b})\right)$.
2.b. Update σ_{k+1} s.t. $\frac{(1-\sigma_{k+1})^2}{\sigma_{k+1}} = \frac{1}{\sigma_k}$.	**2.b.** $\mathbf{y}_{k+1} := \mathbf{x}_{k+1} + \frac{\tau_{k+1}(1-\tau_k)}{\tau_k}(\mathbf{x}_{k+1} - \mathbf{x}_k)$.
2.c. Update $\mu_{k+1} = \sqrt{\sigma_{k+1}}$.	**2.c.** Update $\mu_{k+1} = \mu_k(1 + \tau_{k+1})$.
	2.d. Update $\tau_{k+1} \in (0, 1)$ as the unique positive root of $\tau^3 + \tau^2 + \tau_k^2\tau - \tau_k^2 = 0$.

Theorem (Convergence [17])

- ***LQP***:

$$|f(\mathbf{x}_k) - f(\mathbf{x}^\star)| \leq \mathcal{O}\left(\mu_0 k^{-1/2} + \mu_0^{-1} k^{-1/2}\right)$$

$$\Vert\mathbf{A}\mathbf{x}_k - \mathbf{b}\Vert \leq \mathcal{O}\left(\mu_0^{-1} k^{-1/2}\right)$$

- ***ALQP***:

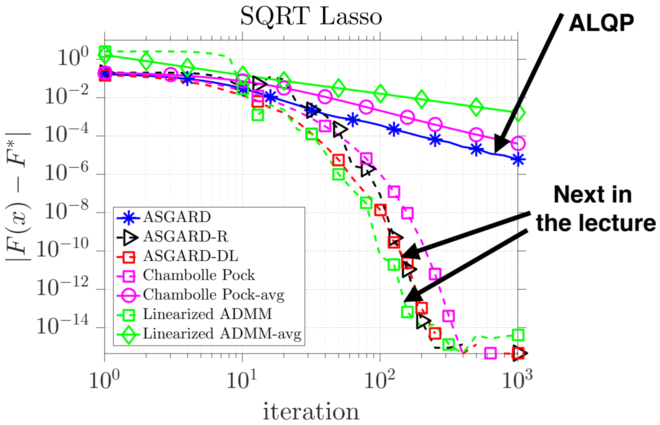
$$|f(\mathbf{x}_k) - f(\mathbf{x}^\star)| \leq \mathcal{O}\left(\mu_0 k^{-1} + \mu_0^{-1} k^{-1}\right)$$

$$\Vert\mathbf{A}\mathbf{x}_k - \mathbf{b}\Vert \leq \mathcal{O}\left(\mu_0^{-1} k^{-1}\right)$$

In practice: poor (worst case) performance

◦ A nonsmooth problem: SQRT Lasso

$$\min_{\mathbf{x}\in\mathbb{R}^p} \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2 + \lambda\|\mathbf{x}\|_1.$$



Wrap up!

- Try to finish Homework #2...

A *convex* proto-problem for *structured* sparsity

A combinatorial approach for estimating $\mathbf{x}^\natural$ from $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$

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

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \{\|\mathbf{x}\|_{\boldsymbol{s}} : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \kappa, \|\mathbf{x}\|_\infty \leq 1\} \quad (\mathcal{P}_s)$$

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

Sparsity and structure together [6]

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

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

for all feasible $\mathbf{x}$, ∞ otherwise. The parameter $\boldsymbol{\omega}$ is useful for latent modeling.

A *convex* proto-problem for *structured* sparsity

A combinatorial approach for estimating $\mathbf{x}^\natural$ from $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$

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

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{\|\,\mathbf{x}\,\|_{\boldsymbol{s}} : \|\,\mathbf{b} - \mathbf{A}\mathbf{x}\,\|_2 \leq \kappa, \|\mathbf{x}\|_\infty \leq 1\right\} \qquad (\mathcal{P}_s)$$

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

Sparsity and structure together [6]

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

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

for all feasible $\mathbf{x}$, ∞ otherwise. The parameter $\boldsymbol{\omega}$ is useful for latent modeling.

A convex candidate solution for $\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w}$

We use the convex estimator based on the tightest convex relaxation of $\|\,\mathbf{x}\,\|_{\boldsymbol{s}}$:
$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathrm{dom}(\|\cdot\|_{\boldsymbol{s}})} \left\{\|\,\mathbf{x}\,\|_{\boldsymbol{s}}^{**} : \|\,\mathbf{b} - \mathbf{A}\mathbf{x}\,\|_2 \leq \kappa\right\}$ with some $\kappa \geq 0$, $\mathrm{dom}(\|\cdot\|_{\boldsymbol{s}}) := \{\mathbf{x} : \|\,\mathbf{x}\,\|_{\boldsymbol{s}} < \infty\}$.

Tractability & tightness of biconjugation

Proposition (Hardness of conjugation)

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

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

Observations:

- $F(\boldsymbol{s})$ is general set function

  Computation: NP-Hard

- $F(\boldsymbol{s}) = \|\mathbf{x}\|_{\boldsymbol{s}}$

  Computation: Integer Linear Program (ILP) in general. However, if
  - $\boldsymbol{M}$ is Totally Unimodular TU
  - $(\boldsymbol{M}, \boldsymbol{c})$ is Total Dual Integral TDI

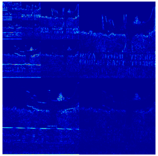
  then tight convex relaxations with a linear program (LP, which is "usually" tractable)

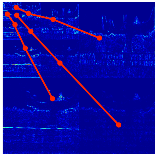
  Otherwise, relax to LP anyway!

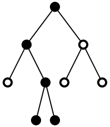
- $F(\boldsymbol{s})$ is submodular

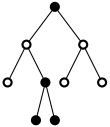
  Computation: Polynomial-time

Tree sparsity [11, 5, 3, 18]


Wavelet coefficients


Wavelet tree


Valid selection of nodes


Invalid selection of nodes

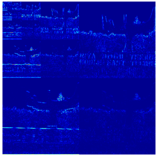
Structure: *We seek the sparsest signal with a rooted connected subtree support.*

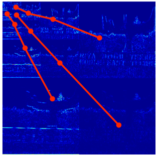
Linear description: A valid support satisfy $s_{\text{parent}} \geq s_{\text{child}}$ over tree $\mathcal{T}$

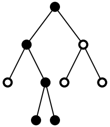
$$\boldsymbol{T}\mathbb{1}_{\text{supp}(\mathbf{x})} := \boldsymbol{T}\boldsymbol{s} \geq 0$$

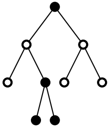
where $\boldsymbol{T}$ is the directed edge-node incidence matrix, which is TU.

Tree sparsity [11, 5, 3, 18]


Wavelet coefficients


Wavelet tree


Valid selection of nodes


Invalid selection of nodes

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

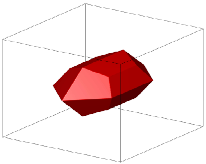
Linear description: A valid support satisfy $s_{\text{parent}} \geq s_{\text{child}}$ over tree $\mathcal{T}$

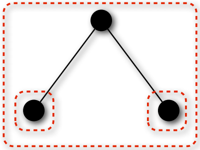
$$\boldsymbol{T}\mathbb{1}_{\text{supp}(\mathbf{x})} := \boldsymbol{T}\boldsymbol{s} \geq 0$$

where $\boldsymbol{T}$ is the directed edge-node incidence matrix, which is TU.

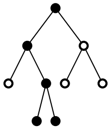
Biconjugate: $\|\mathbf{x}\|_{\boldsymbol{s}}^{**} = \min_{\boldsymbol{s}\in[0,1]^p}\{\mathbb{1}^T\boldsymbol{s} : \boldsymbol{T}\boldsymbol{s} \geq 0, |\mathbf{x}| \leq \boldsymbol{s}\}$
for $\mathbf{x} \in [-1,1]^p$, ∞ otherwise.

Tree sparsity [11, 5, 3, 18]





$\mathfrak{G}_H = \{\{1,2,3\},\{2\},\{3\}\}$



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 $\mathcal{T}$

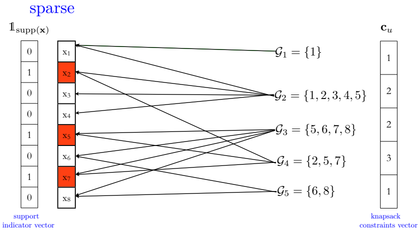
$$\boxed{\boldsymbol{T}\mathbb{1}_{\text{supp}(\mathbf{x})} := \boldsymbol{T}\boldsymbol{s} \geq 0}$$

where $\boldsymbol{T}$ is the directed edge-node incidence matrix, which is TU.

Biconjugate: $\|\mathbf{x}\|_{\boldsymbol{s}}^{**} = \min_{\boldsymbol{s}\in[0,1]^p}\{\mathbb{1}^T\boldsymbol{s} : \boldsymbol{T}\boldsymbol{s} \geq 0, |\mathbf{x}| \leq \boldsymbol{s}\} \overset{\star}{=} \sum_{\mathcal{G}\in\mathfrak{G}_H} \|x_{\mathcal{G}}\|_\infty$
for $\mathbf{x} \in [-1,1]^p$, ∞ otherwise.

The set $\mathcal{G} \in \mathfrak{G}_H$ are defined as each node and all its descendants.

Group knapsack sparsity [20, 8, 7]



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

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

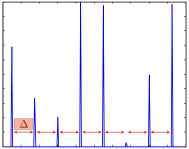
$$\mathfrak{B}^T \boldsymbol{s} \leq \boldsymbol{c}_u$$

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

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

Remark: We can also budget a lowerbound $\boldsymbol{c}_\ell \leq \mathfrak{B}^T \boldsymbol{s} \leq \boldsymbol{c}_u$.

Group knapsack sparsity [20, 8, 7]



$$\mathfrak{B}^T = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 1 & \cdots & 1 & 1 & 0 & \cdots & 0 \\ & & & & \ddots & & & & \\ 0 & \cdots & 0 & 0 & 1 & 1 & \cdots & 1 & 1 \end{bmatrix}_{(p-\Delta+1)\times p}$$

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

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

$$\boxed{\mathfrak{B}^T \boldsymbol{s} \leq \boldsymbol{c}_u}$$

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

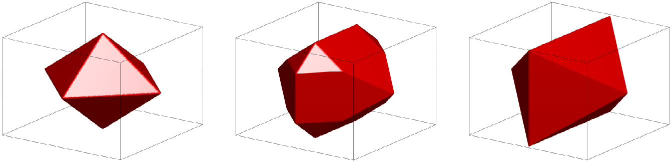
When $\mathfrak{B}$ is an interval matrix or $\mathfrak{G}$ has a *loopless* group intersection graph, it is TU.

Remark: We can also budget a lowerbound $\boldsymbol{c}_\ell \leq \mathfrak{B}^T \boldsymbol{s} \leq \boldsymbol{c}_u$.

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

For the neuronal spike example, we have $\boldsymbol{c}_u = \mathbb{1}$.

Group knapsack sparsity [20, 8, 7]



(left) $\|\mathbf{x}\|_s^{**} \leq 1$ (middle) $\|\mathbf{x}\|_s^{**} \leq 1.5$ (right) $\|\mathbf{x}\|_s^{**} \leq 2$ for $\mathfrak{G} = \{\{1,2\},\{2,3\}\}$

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

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

$$\boxed{\mathfrak{B}^T \boldsymbol{s} \leq \boldsymbol{c}_u}$$

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

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

Remark: We can also budget a lowerbound $\boldsymbol{c}_\ell \leq \mathfrak{B}^T \boldsymbol{s} \leq \boldsymbol{c}_u$.

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

For the neuronal spike example, we have $\boldsymbol{c}_u = \mathbb{1}$.

Group knapsack sparsity example: A stylized spike train

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

$$\|\mathbf{x}\|_{\boldsymbol{s}}^{**} = \begin{cases} \|\mathbf{x}\|_1 & \text{if } \mathbf{x} \in [-1,1]^p, \mathfrak{B}^T|\mathbf{x}| \leq \boldsymbol{c}_u, \\ \infty & \text{otherwise} \end{cases}$$

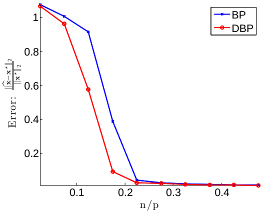
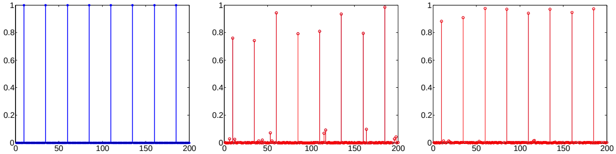


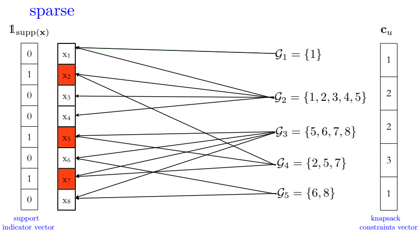
Figure: *Recovery for $n = 0.18p$.*



$\mathbf{x}^\natural$ | $\mathbf{x}^{\text{BP}}$ solution | $\mathbf{x}^{\text{TU}}$ solution

relative errors: $\frac{\|\mathbf{x}^\natural - \mathbf{x}^{\text{BP}}\|_2}{\|\mathbf{x}^\natural\|_2} = .200$ $\frac{\|\mathbf{x}^\natural - \mathbf{x}^{\text{TU}}\|_2}{\|\mathbf{x}^\natural\|_2} = .067$

Group knapsack sparsity: A simple variation



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

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

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

$$\boxed{\mathfrak{B}^T \boldsymbol{s} \leq \omega \mathbb{1}}$$

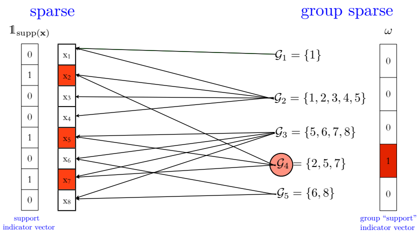
where $\mathfrak{B}$ is the biadjacency matrix of $\mathfrak{G}$, i.e., $\mathfrak{B}_{ij} = 1$ iff i-th coefficient is in $\mathcal{G}_j$.

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

Biconjugate: $$\|\mathbf{x}\|_s^{**} = \begin{cases} \max_{\mathcal{G} \in \mathfrak{G}} \|\mathbf{x}^{\mathcal{G}}\|_1 & \text{if } \mathbf{x} \in [-1,1]^p, \\ \infty & \text{otherwise} \end{cases}$$

Remark: The regularizer is known as *exclusive Lasso* [20, 15].

Group cover sparsity: Minimal group cover [2, 14, 9]



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

$$\text{Objective: } \mathbb{1}^T \boldsymbol{s} \rightarrow \boldsymbol{d}^T \boldsymbol{\omega}$$

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

$$\boxed{\mathfrak{B}\boldsymbol{\omega} \geq \boldsymbol{s}}$$

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

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

Group cover sparsity: Minimal group cover [2, 14, 9]

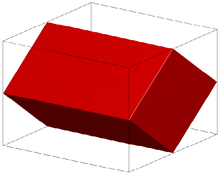


Figure: $\mathfrak{G} = \{\{1,2\},\{2,3\}\}$, unit group weights $\boldsymbol{d} = \mathbb{1}$.

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

Objective: $\mathbb{1}^T \boldsymbol{s} \to \boldsymbol{d}^T \boldsymbol{\omega}$

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

$$\boxed{\mathfrak{B}\boldsymbol{\omega} \geq \boldsymbol{s}}$$

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

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

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

Group cover sparsity: Minimal group cover [2, 14, 9]

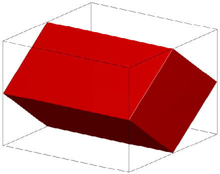


Figure: $\mathfrak{G} = \{\{1,2\},\{2,3\}\}$, unit group weights $\boldsymbol{d} = \mathbb{1}$.

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

Objective: $\mathbb{1}^T \boldsymbol{s} \rightarrow \boldsymbol{d}^T \boldsymbol{\omega}$

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

$$\boxed{\mathfrak{B}\boldsymbol{\omega} \geq \boldsymbol{s}}$$

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

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

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

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

Group cover sparsity: Minimal group cover [2, 14, 9]

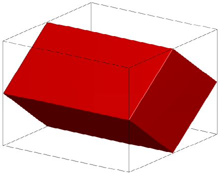


Figure: $\mathfrak{G} = \{\{1,2\},\{2,3\}\}$, unit group weights $\boldsymbol{d} = \mathbb{1}$.

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

Objective: $\mathbb{1}^T \boldsymbol{s} \to \boldsymbol{d}^T \boldsymbol{\omega}$

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

$$\boxed{\mathfrak{B}\boldsymbol{\omega} \geq \boldsymbol{s}}$$

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

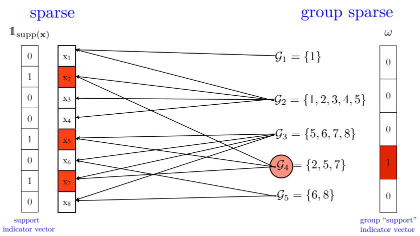
When $\mathfrak{B}$ is an interval matrix, or $\mathfrak{G}$ has a *loopless* group intersection graph it is TU.

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

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

Remark: Weights $\boldsymbol{d}$ can depend on the sparsity within each groups (not TU) [6].

Budgeted group cover sparsity



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

$$\text{Objective: } \boldsymbol{d}^T\boldsymbol{\omega} \to \mathbb{1}^T\boldsymbol{s}$$

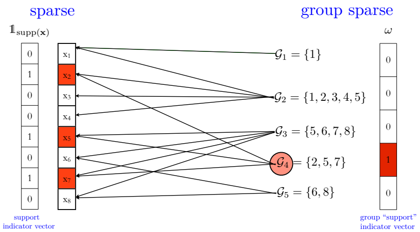
Linear description: At least one of the G selected groups cover each sparse coefficient.

$$\boxed{\mathfrak{B}\boldsymbol{\omega} \geq \boldsymbol{s}, \mathbb{1}^T\boldsymbol{\omega} \leq G}$$

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

When $\begin{bmatrix}\mathfrak{B}\\ \mathbb{1}\end{bmatrix}$ is an interval matrix, it is TU.

Budgeted group cover sparsity



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

Objective: $\boldsymbol{d}^T\boldsymbol{\omega} \to \mathbb{1}^T\boldsymbol{s}$

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

$$\mathfrak{B}\boldsymbol{\omega} \geq \boldsymbol{s}, \mathbb{1}^T\boldsymbol{\omega} \leq G$$

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

When $\begin{bmatrix}\mathfrak{B}\\ \mathbb{1}\end{bmatrix}$ is an interval matrix, it is TU.

Biconjugate: $\|\mathbf{x}\|_{\boldsymbol{\omega}}^{**} = \min_{\boldsymbol{\omega}\in[0,1]^M}\{\|\mathbf{x}\|_1 : \mathfrak{B}\boldsymbol{\omega} \geq |\mathbf{x}|, \mathbb{1}^T\boldsymbol{\omega} \leq G\}$
for $\mathbf{x} \in [-1,1]^p$, ∞ otherwise.

Budgeted group cover example: Interval overlapping groups

- Basis pursuit (BP): $\|\mathbf{x}\|_1$
- Sparse group Lasso (SGL$_q$):

$$(1-\alpha)\sum_{\mathcal{G}\in\mathfrak{G}}\sqrt{|\mathcal{G}|}\|\mathbf{x}^{\mathcal{G}}\|_q + \alpha\|\mathbf{x}^{\mathcal{G}}\|_1$$

- TU-relax (TU):

$$\|\mathbf{x}\|_{\boldsymbol{\omega}}^{**} = \min_{\boldsymbol{\omega}\in[0,1]^M}\{\|\mathbf{x}\|_1 : \mathfrak{B}\boldsymbol{\omega} \geq |\mathbf{x}|, \mathbb{1}^T\boldsymbol{\omega} \leq G\}$$

for $\mathbf{x} \in [-1,1]^p, \infty$ otherwise.

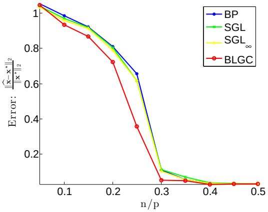
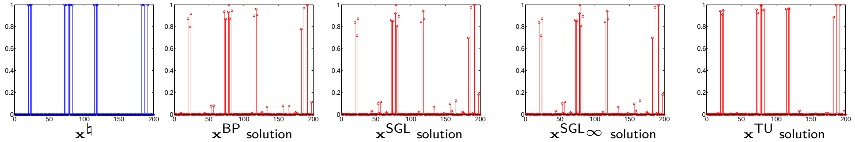
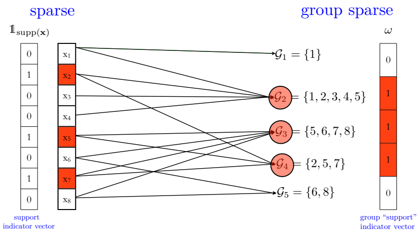


Figure: *Recovery for $n = 0.25p, s = 15, p = 200, G = 5$ out of $M = 29$ groups.*



relative errors: $\frac{\|\mathbf{x}^\natural - \mathbf{x}^{\text{BP}}\|_2}{\|\mathbf{x}^\natural\|_2} = .128$ $\frac{\|\mathbf{x}^\natural - \mathbf{x}^{\text{SGL}}\|_2}{\|\mathbf{x}^\natural\|_2} = .181$ $\frac{\|\mathbf{x}^\natural - \mathbf{x}^{\text{SGL}\infty}\|_2}{\|\mathbf{x}^\natural\|_2} = .085$ $\frac{\|\mathbf{x}^\natural - \mathbf{x}^{\text{TU}}\|_2}{\|\mathbf{x}^\natural\|_2} = .058$

Group intersection sparsity [10, 19, 1]



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

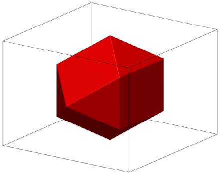
Objective: $\mathbb{1}^T \boldsymbol{s} \to \boldsymbol{d}^T \boldsymbol{\omega}$

Linear description: All groups containing a sparse coefficient are selected

$$\boxed{\boldsymbol{H}_k \boldsymbol{s} \leq \boldsymbol{\omega}, \forall k \in \mathfrak{P}}$$

where $\boldsymbol{H}_k(i,j) = \begin{cases} 1 & \text{if } j = k, j \in \mathcal{G}_i \\ 0 & \text{otherwise} \end{cases}$, which is TU.

Group intersection sparsity [10, 19, 1]



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

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

Objective: $\mathbb{1}^T \boldsymbol{s} \to \boldsymbol{d}^T \boldsymbol{\omega}$

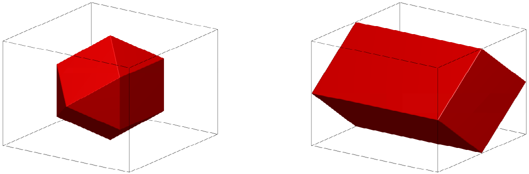
Linear description: All groups containing a sparse coefficient are selected

$$\boldsymbol{H}_k \boldsymbol{s} \leq \boldsymbol{\omega}, \forall k \in \mathfrak{P}$$

where $\boldsymbol{H}_k(i,j) = \begin{cases} 1 & \text{if } j = k, j \in \mathcal{G}_i \\ 0 & \text{otherwise} \end{cases}$, which is TU.

Biconjugate: $\|\mathbf{x}\|_{\boldsymbol{\omega}}^{**} = \min_{\boldsymbol{\omega} \in [0,1]^M} \{\boldsymbol{d}^T \boldsymbol{\omega} : \boldsymbol{H}_k |\mathbf{x}| \leq \boldsymbol{\omega}, \forall k \in \mathfrak{P}\}$
for $\mathbf{x} \in [-1,1]^p$, ∞ otherwise.

Group intersection sparsity [10, 19, 1]



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

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

Objective: $\mathbb{1}^T \boldsymbol{s} \to \boldsymbol{d}^T \boldsymbol{\omega}$ (*submodular*)

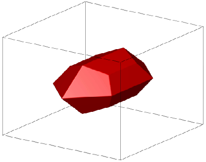
Linear description: All groups containing a sparse coefficient are selected

$$\boldsymbol{H}_k \boldsymbol{s} \leq \boldsymbol{\omega}, \forall k \in \mathfrak{P}$$

where $\boldsymbol{H}_k(i,j) = \begin{cases} 1 & \text{if } j = k, j \in \mathcal{G}_i \\ 0 & \text{otherwise} \end{cases}$, which is TU.

Biconjugate: $\|\mathbf{x}\|_{\boldsymbol{\omega}}^{**} = \min_{\boldsymbol{\omega} \in [0,1]^M} \{\boldsymbol{d}^T \boldsymbol{\omega} : \boldsymbol{H}_k |\mathbf{x}| \leq \boldsymbol{\omega}, \forall k \in \mathfrak{P}\} \overset{\star}{=} \sum_{\mathcal{G} \in \mathfrak{G}} \|x_{\mathcal{G}}\|_\infty$
for $\mathbf{x} \in [-1,1]^p$, ∞ otherwise.

Group intersection sparsity [10, 19, 1]



$\mathfrak{G} = \{\{1,2,3\},\{2\},\{3\}\}$, unit group weights $\boldsymbol{d} = \mathbb{1}$.

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

Objective: $\mathbb{1}^T \boldsymbol{s} \rightarrow \boldsymbol{d}^T \boldsymbol{\omega}$ (*submodular*)

Linear description: All groups containing a sparse coefficient are selected

$$\boldsymbol{H}_k \boldsymbol{s} \leq \boldsymbol{\omega}, \forall k \in \mathfrak{P}$$

where $\boldsymbol{H}_k(i,j) = \begin{cases} 1 & \text{if } j = k, j \in \mathcal{G}_i \\ 0 & \text{otherwise} \end{cases}$, which is TU.

Biconjugate: $\|\mathbf{x}\|_{\boldsymbol{\omega}}^{**} = \min_{\boldsymbol{\omega} \in [0,1]^M} \{\boldsymbol{d}^T \boldsymbol{\omega} : \boldsymbol{H}_k |\mathbf{x}| \leq \boldsymbol{\omega}, \forall k \in \mathfrak{P}\} \overset{\star}{=} \sum_{\mathcal{G} \in \mathfrak{G}} \|x_{\mathcal{G}}\|_\infty$ for $\mathbf{x} \in [-1,1]^p$, ∞ otherwise.

Remark: For hierarchical $\mathfrak{G}_H$, group intersection and tree sparsity models coincide.

Beyond linear costs: Graph dispersiveness

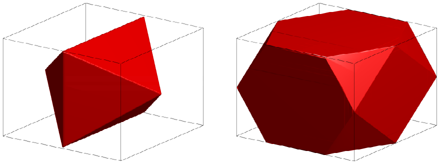


Figure: (left) $\|\mathbf{x}\|_s^{**} = 0$ (right) $\|\mathbf{x}\|_s^{**} \leq 1$ for $\mathcal{E} = \{\{1,2\},\{2,3\}\}$ (chain graph)

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

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

Linearization:

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

When edge-node incidence matrix of $\mathcal{G}(\mathfrak{P}, \mathcal{E})$ is TU (e.g., bipartite graphs), it is TU.

Beyond linear costs: Graph dispersiveness

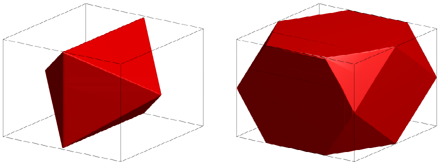


Figure: (left) $\|\mathbf{x}\|_s^{**} = 0$ (right) $\|\mathbf{x}\|_s^{**} \leq 1$ for $\mathcal{E} = \{\{1,2\},\{2,3\}\}$ (chain graph)

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

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

Linearization:

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

When edge-node incidence matrix of $\mathcal{G}(\mathfrak{P}, \mathcal{E})$ is TU (e.g., bipartite graphs), it is TU.

Biconjugate: $\|\mathbf{x}\|_{\boldsymbol{s}}^{**} = \sum_{(i,j)\in\mathcal{E}} (|x_i| + |x_j| - 1)_+$ for $\mathbf{x} \in [-1,1]^p, \infty$ otherwise.

References I

[1] Francis Bach.
Structured sparsity-inducing norms through submodular functions.
Adv. Neur. Inf. Proc. Sys. (NIPS), pages 118–126, 2010.
(Cited on pages 59, 60, 61, and 62.)

[2] L. Baldassarre, N. Bhan, V. Cevher, and A. Kyrillidis.
Group-sparse model selection: Hardness and relaxations.
arXiv preprint arXiv:1303.3207, 2013.
(Cited on pages 52, 53, 54, and 55.)

[3] R.G. Baraniuk, V. Cevher, M.F. Duarte, and C. Hegde.
Model-based compressive sensing.
IEEE Trans. Inf. Theory, 56(4):1982–2001, April 2010.
(Cited on pages 44, 45, and 46.)

[4] Dimitri P Bertsekas.
Necessary and sufficient conditions for a penalty method to be exact.
Mathematical programming, 9(1):87–99, 1975.
(Cited on page 33.)

References II

[5] Marco F. Duarte, Dharmpal Davenport, Mark A. adn Takhar, Jason N. Laska, Ting Sun, Kevin F. Kelly, and Richard G. Baraniuk.
Single-pixel imaging via compressive sampling.
IEEE Sig. Proc. Mag., 25(2):83–91, March 2008.
(Cited on pages 44, 45, and 46.)

[6] Marwa El Halabi and Volkan Cevher.
A totally unimodular view of structured sparsity.
preprint, 2014.
arXiv:1411.1990v1 [cs.LG].
(Cited on pages 41, 42, 52, 53, 54, and 55.)

[7] W Gerstner and W. Kistler.
Spiking neuron models: Single neurons, populations, plasticity.
Cambridge university press, 2002.
(Cited on pages 47, 48, and 49.)

[8] C. Hegde, M. Duarte, and V. Cevher.
Compressive sensing recovery of spike trains using a structured sparsity model.
In *Sig. Proc. with Adapative Sparse Struct. Rep. (SPARS)*, 2009.
(Cited on pages 47, 48, and 49.)

References III

[9] J. Huang, T. Zhang, and D. Metaxas.
Learning with structured sparsity.
J. Mach. Learn. Res., 12:3371–3412, 2011.
(Cited on pages 52, 53, 54, and 55.)

[10] R. Jenatton, A. Gramfort, V. Michel, G. Obozinski, F. Bach, and B. Thirion.
Multi-scale mining of fmri data with hierarchical structured sparsity.
In *Pattern Recognition in NeuroImaging (PRNI)*, 2011.
(Cited on pages 59, 60, 61, and 62.)

[11] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach.
Proximal methods for hierarchical sparse coding.
J. Mach. Learn. Res., 12:2297–2334, 2011.
(Cited on pages 44, 45, and 46.)

[12] Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari.
Regularization techniques for learning with matrices.
Journal of Machine Learning Research, 13(59):1865–1890, 2012.
(Cited on pages 5 and 6.)

References IV

[13] J. Nocedal and S.J. Wright.
Numerical Optimization.
Springer Series in Oper. Res. and Financial Engineering. Springer, 2 edition, 2006.
(Cited on pages 33 and 35.)

[14] G. Obozinski, L. Jacob, and J.P. Vert.
Group lasso with overlaps: The latent group lasso approach.
arXiv preprint arXiv:1110.0413, 2011.
(Cited on pages 52, 53, 54, and 55.)

[15] G. Obozinski, B. Taskar, and M.I. Jordan.
Joint covariate selection and joint subspace selection for multiple classification problems.
Statistics and Computing, 20(2):231–252, 2010.
(Cited on page 51.)

[16] Quoc Tran-Dinh and Volkan Cevher.
Constrained convex minimization via model-based excessive gap.
In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'14, 2014.
(Cited on page 27.)

References V

[17] Quoc Tran-Dinh, Olivier Fercoq, and Volkan Cevher.
A smooth primal-dual optimization framework for nonsmooth composite convex minimization.
SIAM Journal on Optimization, 28(1):96–134, 2018.
(Cited on page 38.)

[18] Peng Zhao, Guilherme Rocha, and Bin Yu.
Grouped and hierarchical model selection through composite absolute penalties.
Department of Statistics, UC Berkeley, Tech. Rep, 703, 2006.
(Cited on pages 44, 45, and 46.)

[19] Peng Zhao and Bin Yu.
On model selection consistency of Lasso.
J. Mach. Learn. Res., 7:2541–2563, 2006.
(Cited on pages 59, 60, 61, and 62.)

[20] H. Zhou, M.E. Sehl, J.S. Sinsheimer, and K. Lange.
Association screening of common and rare genetic variants by penalized regression.
Bioinformatics, 26(19):2375, 2010.
(Cited on pages 47, 48, 49, and 51.)