Advanced Topics in Data Sciences

Prof. Volkan Cevher
volkan.cevher@epfl.ch

Lecture 3: Structured sparsity

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

EE-731 (Spring 2016)

lions@epfl

MARIE CURIE ACTIONS

FNSNF
FONDS NATIONAL SUISSE
SCHWEIZERISCHER NATIONALFONDS
FONDO NAZIONALE SVIZZERO
SWISS NATIONAL SCIENCE FOUNDATION

License Information for Mathematics of Data Slides

Outline

- This lecture
  1. Review of compressive sensing problem
  2. Overview of structured sparsity
  3. Convex relaxation
  4. Fenchel conjugate
  5. Examples

Recommended Reading

- *Structured sparsity-inducing norms through submodular functions*, Bach, 2010.
- *A totally unimodular view of structured sparsity*, El Halabi and Cevher, 2015.

Signal recovery from linear measurements

Problem statement

Recover an accurate estimate $\hat{\mathbf{x}}$ of a signal $\mathbf{x}^\natural \in \mathbb{C}^p$, in the sense $\|\hat{\mathbf{x}} - \mathbf{x}^\natural\| \leq \epsilon$, from a set of linear measurements

$$\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w},$$

where $\mathbf{A} \in \mathbb{C}^{n \times p}$ is a *known* measurement matrix, and $\mathbf{w} \in \mathbb{C}^{n \times 1}$ an *unknown* noise.

Signal recovery from linear measurements

Problem statement

Recover an accurate estimate $\hat{\mathbf{x}}$ of a signal $\mathbf{x}^\natural \in \mathbb{C}^p$, in the sense $\|\hat{\mathbf{x}} - \mathbf{x}^\natural\| \leq \epsilon$, from a set of linear measurements

$$\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w},$$

where $\mathbf{A} \in \mathbb{C}^{n \times p}$ is a *known* measurement matrix, and $\mathbf{w} \in \mathbb{C}^{n \times 1}$ an *unknown* noise.

The following problem is fundamental in signal processing, machine learning, and many other areas.

- Image compression
- Medical resonance imaging (MRI)
- Communications
- Linear regression

Signal recovery from linear measurements

Problem statement

Recover an accurate estimate $\hat{\mathbf{x}}$ of a signal $\mathbf{x}^\natural \in \mathbb{C}^p$, in the sense $\|\hat{\mathbf{x}} - \mathbf{x}^\natural\| \leq \epsilon$, from a set of linear measurements

$$\mathbf{b} = \mathbf{A}\mathbf{x}^\natural + \mathbf{w},$$

where $\mathbf{A} \in \mathbb{C}^{n \times p}$ is a *known* measurement matrix, and $\mathbf{w} \in \mathbb{C}^{n \times 1}$ an *unknown* noise.

The following problem is fundamental in signal processing, machine learning, and many other areas.

- Image compression
- Medical resonance imaging (MRI)
- Communications
- Linear regression

Two regimes of interest:

- $n < p$ (*underdetermined*): Infinitely many solutions; impossible in general
- $n > p$ (*overdetermined*): Solvable using classical techniques such as least squares

Least-squares estimation in the linear model

Recall the least-squares (LS) estimator.

LS estimation in the linear model

The LS estimator for $\mathbf{x}^\natural$ given $\mathbf{A}$ and $\mathbf{b}$ is defined as

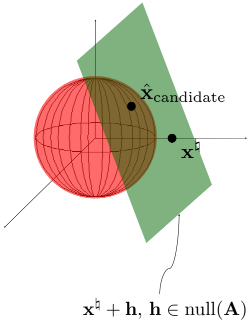
$$\hat{\mathbf{x}}_{\mathsf{LS}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2 \right\}.$$

- If $\mathbf{A}$ has full column rank, $\hat{\mathbf{x}}_{\mathsf{LS}} = \mathbf{A}^\dagger\mathbf{b}$ is uniquely defined.
- *In the case that $n < p$*, $\mathbf{A}$ cannot have full column rank, and we can only conclude that $\hat{\mathbf{x}}_{\mathsf{LS}} \in \left\{ \mathbf{A}^\dagger\mathbf{b} + \mathbf{h} : \mathbf{h} \in \mathrm{null}\left(\mathbf{A}\right) \right\}$.

Observation: The estimation error $\left\| \hat{\mathbf{x}}_{\mathsf{LS}} - \mathbf{x}^\natural \right\|_2$ can be *arbitrarily large*!

A candidate solution

- There are infinitely many solutions $\mathbf{x}$ such that $\mathbf{b} = \mathbf{A}\mathbf{x}$
- Suppose that $\mathbf{w} = 0$ (i.e. no noise). Should we just choose the one $\hat{\mathbf{x}}_{\text{candidate}}$ with the smallest norm $\|\mathbf{x}\|_2$?



Unfortunately, *this still fails when* $n < p$

A candidate solution contd.

Proposition ([5])

Suppose that $\mathbf{A} \in \mathbb{R}^{n \times p}$ *is a matrix of i.i.d. standard Gaussian random variables, and* $\mathbf{w} = \mathbf{0}$. *We have*

$$(1-\epsilon)\left(1 - \frac{n}{p}\right)\left\|\mathbf{x}^\natural\right\|_2^2 \leq \left\|\hat{\mathbf{x}}_{\text{candidate}} - \mathbf{x}^\natural\right\|_2^2 \leq (1-\epsilon)^{-1}\left(1 - \frac{n}{p}\right)\left\|\mathbf{x}^\natural\right\|_2^2$$

with probability at least $1 - 2\exp\left[-(1/4)(p-n)\epsilon^2\right] - 2\exp\left[-(1/4)p\epsilon^2\right]$, *for all* $\epsilon > 0$ *and* $\mathbf{x}^\natural \in \mathbb{R}^p$.

Observation: The estimation error may *not* diminish unless n is very close to p.

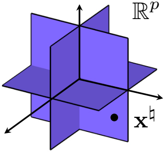
Impact: It is impossible to estimate $\mathbf{x}^\natural$ accurately using $\hat{\mathbf{x}}_{\text{candidate}}$ when $n \ll p$ even if $\mathbf{w} = \mathbf{0}$.

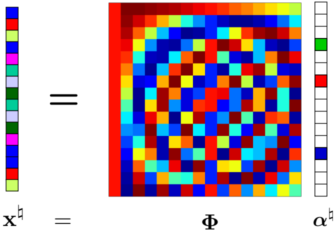
- The statistical error $\left\|\hat{\mathbf{x}}_{\text{candidate}} - \mathbf{x}^\natural\right\|_2^2$ can also be arbitrarily large when $\mathbf{w} \neq \mathbf{0}$. Hence, the solution is also not robust.
- **We need additional information on $\mathbf{x}^\natural$!**

A natural signal model

Definition (s-sparse vector)

A vector $\boldsymbol{\alpha} \in \mathbb{R}^p$ is s-sparse if it has at most s non-zero entries.

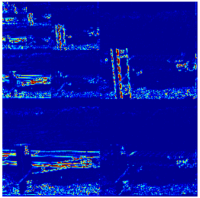




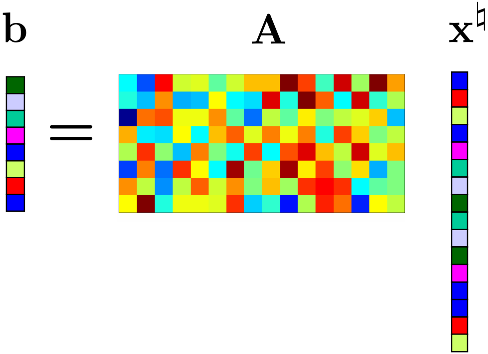
Sparse representations

$\boldsymbol{\alpha}^\natural$: *sparse* transform coefficients

- Basis representations $\boldsymbol{\Phi} \in \mathbb{R}^{p \times p}$
  - *Wavelets*, DCT, ...
- Frame representations $\boldsymbol{\Phi} \in \mathbb{R}^{m \times p}$, $m > p$
  - Gabor, curvelets, shearlets, ...
- Other *dictionary* representations...

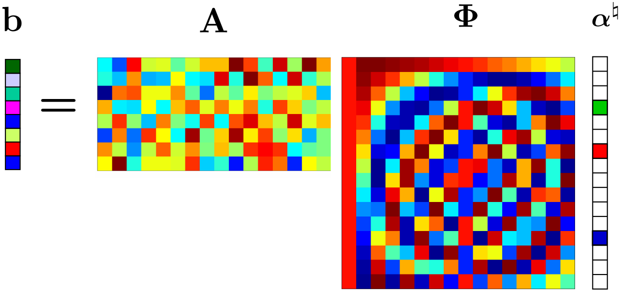


Sparse representations strike back!



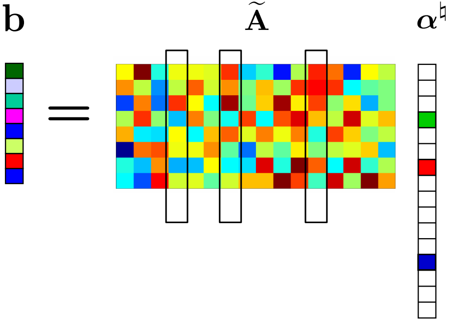
- $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n \times p}$, and $n < p$

Sparse representations strike back!



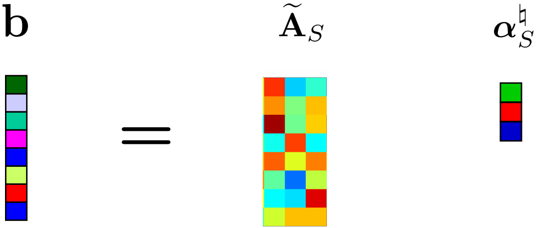
- $\mathbf{b} \in \mathbb{R}^n$, $\mathbf{A} \in \mathbb{R}^{n \times p}$, and $n < p$
- $\mathbf{\Psi} \in \mathbb{R}^{p \times p}$, $\boldsymbol{\alpha}^\natural \in \mathbb{R}^p$, and $\|\boldsymbol{\alpha}^\natural\|_0 \leq s < n$

Sparse representations strike back!



- $\mathbf{b} \in \mathbb{R}^n$, $\tilde{\mathbf{A}} \in \mathbb{R}^{n \times p}$, and $\boldsymbol{\alpha}^\natural \in \mathbb{R}^p$, and $\|\boldsymbol{\alpha}^\natural\|_0 \leq s < n < p$

Sparse representations strike back!



A fundamental impact:

The matrix $\tilde{\mathbf{A}}$ effectively becomes *overcomplete*.

We could easily solve for $\boldsymbol{\alpha}^\natural$ (and hence $\mathbf{x}^\natural$) if we knew *the location of the non-zero entries of* $\mathbf{x}^\natural$.

Sparse recovery

Sparse estimators

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_0 : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{w}\|_2 \right\} \qquad (\mathcal{P}_0)$$

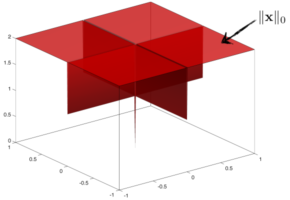
$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 + \rho \|\mathbf{x}\|_0 \qquad (\mathcal{P}_0')$$

where $\|\mathbf{x}\|_0 := \mathbb{1}^T \boldsymbol{s}$, $\boldsymbol{s} = \mathbb{1}_{\mathrm{supp}(\mathbf{x})}$, $\mathrm{supp}(\mathbf{x}) = \{i | x_i \neq 0\}$.

Sparse estimators characteristics:

- Sample complexity: $\mathcal{O}(s)$
- Computational effort: *NP-Hard*
- *Not robust* to noise.

$\|\mathbf{x}\|_0$ over the unit ℓ_∞-ball



Convex relaxation of sparse recovery

Convex sparse estimators

Basis pursuit (BP):

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_1 : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{w}\|_2 \right\} \qquad \text{(BP)}$$

Least absolute shrinkage and selection operator (Lasso):

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 + \rho \|\mathbf{x}\|_1 \qquad \text{(LASSO)}$$

where $\|\mathbf{x}\|_1 := \mathbb{1}^T|\mathbf{x}|$.

Convex estimators characteristics [7]:

- Sample complexity: $\mathcal{O}(s\log(\frac{p}{s}))$
- Computational effort: *Polynomial*
- *Robust* to noise.

Why is $\|\mathbf{x}\|_1$ a good convex surrogate for $\|\mathbf{x}\|_0$?

Convex relaxation of sparse recovery

Convex sparse estimators

Basis pursuit (BP):

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ \|\mathbf{x}\|_1 : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{w}\|_2 \right\} \tag{BP}$$

Least absolute shrinkage and selection operator (Lasso):

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 + \rho \|\mathbf{x}\|_1 \tag{LASSO}$$

where $\|\mathbf{x}\|_1 := \mathbb{1}^T|\mathbf{x}|$.

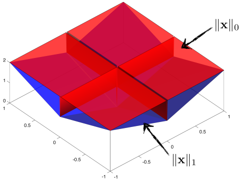
Convex estimators characteristics [7]:

- Sample complexity: $\mathcal{O}(s\log(\frac{p}{s}))$
- Computational effort: *Polynomial*
- *Robust* to noise.

Convex relaxation:

Convex envelope is the *largest* convex lower bound.

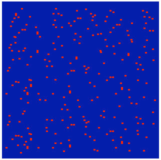
$\|\mathbf{x}\|_1$ is the *convex envelope* of $\|\mathbf{x}\|_0$



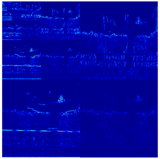
A technicality: Restrict $\mathbf{x}^\natural \in [-1, 1]^p$.

Beyond sparsity towards model-based or *structured* sparsity

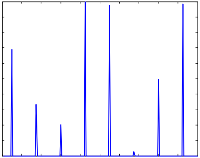
- The following signals can look the **same** from a **sparsity** perspective!



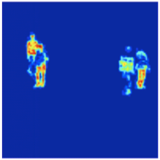
Sparse image



Wavelet coefficients of a natural image

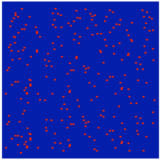


Spike train

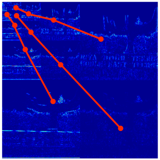


Background substracted image

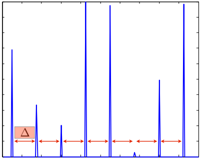
- In reality, these signals have additional structures beyond the simple sparsity



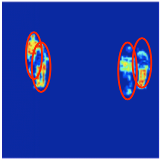
Sparse image



Wavelet coefficients of a natural image



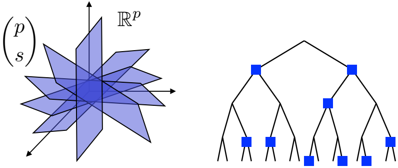
Spike train



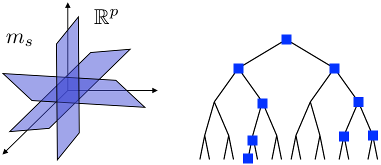
Background substracted image

Beyond sparsity towards model-based or ***structured*** sparsity

Sparsity model: Union of *all* s-dimensional canonical subspaces.



Structured sparsity model: A *particular* union of m_s s-dimensional canonical subspaces.



Three upshots of structured sparsity: [3]

1. Reduced sample complexity: e.g., $\mathcal{O}(s\log(\frac{p}{s})) \to \mathcal{O}(s)$ for tree-sparse signals [1]
2. Better noise robustness
3. Better interpretability

[1] this was proved for a greedy method (CoSaMP). Convex methods in practice require similar number of samples.

Structured sparse recovery

We encode the structure over the support by $g(\mathbf{x}) = F(\mathrm{supp}(\mathbf{x}))$

Structured sparsity estimators

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \left\{ F(\mathrm{supp}(\mathbf{x})) : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{w}\|_2 \right\}$$

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 + \rho F(\mathrm{supp}(\mathbf{x}))$$

where $F(\boldsymbol{s}) : \{0,1\}^p \to \mathbb{R} \cup \{+\infty\}, \mathrm{supp}(\mathbf{x}) = \{i | x_i \neq 0\}$.

Tractable & stable recovery:
How to choose a good convex surrogate of g?

1. Case by case heuristics
2. *Convex envelope:* given by the *biconjugate* of g, i.e., the fenchel conjugate of the fenchel conjugate of g.

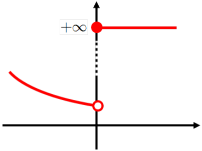
Lower semi-continuity

Definition

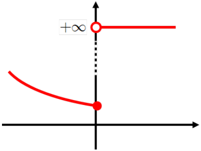
A function $f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is lower semi-continuous (l.s.c.), also called closed, if

$$\liminf_{\mathbf{x} \to \mathbf{y}} f(\mathbf{x}) \geq f(\mathbf{y}), \quad \text{for any } \mathbf{y} \in \mathrm{dom}(f).$$

$$f(x) = \begin{cases} e^{-x}, & \text{if } x < 0 \\ +\infty, & \text{if } x \geq 0 \end{cases}$$



$$f(x) = \begin{cases} e^{-x}, & \text{if } x \leq 0 \\ +\infty, & \text{if } x > 0 \end{cases}$$



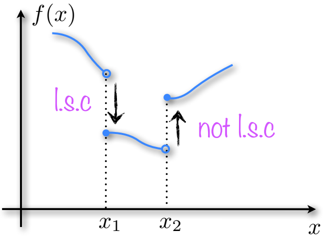
Lower semi-continuity

Definition

A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is lower semi-continuous (l.s.c.) if

$$\liminf_{\mathbf{x}\to\mathbf{y}} f(\mathbf{x}) \geq f(\mathbf{y}), \quad \text{for any } \mathbf{y} \in \mathrm{dom}(f).$$

- **Rule of thumb:** a lower semi-continuous function *only jumps down.*



- f is l.s.c iff its epigraph $\mathrm{epi} f = \{(\mathbf{x}, \alpha) : \mathbf{x} \in \mathbb{R}^p, \alpha \in \mathbb{R}, f(\mathbf{x}) \leq \alpha\}$ is a *closed* set.
- f is l.s.c iff all its sublevel sets $\{\mathbf{x} \in \mathbb{R}^p : f(\mathbf{x}) \leq \alpha\}$ are *closed*.

Fenchel conjugate

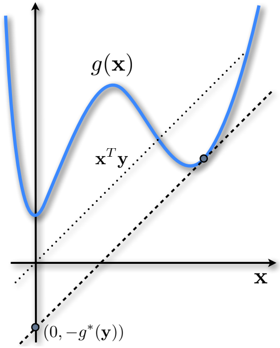
Definition (Fenchel conjugate)

Let $g : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ be a *proper* function (non-empty domain), its fenchel convex conjugate is defined as follows:

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

where the domain of g is defined as $\mathrm{dom}(g) = \{\mathbf{x} \in \mathbb{R}^p : g(\mathbf{x}) \neq +\infty\}$.

Fenchel conjugate



- For a given direction $\mathbf{y} \in \mathbb{R}^p$, $g^*(\mathbf{y})$ is the maximum gap between the linear function $\mathbf{x}^T\mathbf{y}$ (dotted line) and $g(\mathbf{x})$.
- Given $x^\star \in \arg\max_{\mathbf{x}\in\mathrm{dom}(g)} \left\{\mathbf{y}^T\mathbf{x} - g(\mathbf{x})\right\}$, $x^\star$ will lie on the convex envelope.
- g^* may be seen as minus the intercept of the tangent to the graph of g with slope $\mathbf{y}$; i.e., the line $\mathbf{x}^T\mathbf{y} - g^*(\mathbf{y})$.
- By definition of conjugation, g is always above the lines $\mathbf{x}^T\mathbf{y} - g^*(\mathbf{y}), \forall \mathbf{y} \in \mathbb{R}^p$.

Fenchel conjugate

Definition (Fenchel conjugate)

Let $g : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ be a *proper* function (non-empty domain), its fenchel convex conjugate is defined as follows:

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

where the domain of g is defined as $\mathrm{dom}(g) = \{\mathbf{x} \in \mathbb{R}^p : g(\mathbf{x}) \neq +\infty\}$.

Properties of conjugation [8]:

- As a pointwise supremum of linear functions, g^* is always *convex* and *l.s.c*, even if g is not.
- If g is convex and l.s.c itself, then its biconjugate is equal to g; $g^{**} = g$.
- The biconjugate g^{**} is the *l.s.c convex envelope* of g; i.e., the largest l.s.c convex lower bound on g.

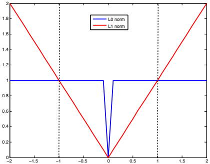
Example: Tight convex relaxation of sparsity

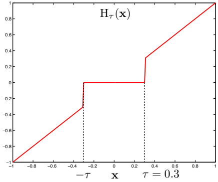
The convex envelope of the ℓ_0-"norm", over the unit ℓ_∞-ball, is the ℓ_1-norm.

Proof:

1. Compute the conjugate $\|\cdot\|_0^*$ of the ℓ_0-"norm", for all $\mathbf{y} \in \mathbb{R}^p$:

$$\begin{aligned}
\|\mathbf{y}\|_0^* &= \sup_{\|\mathbf{x}\|_\infty \leq 1} \mathbf{x}^T\mathbf{y} - \|\mathbf{y}\|_0 \\
&= \sup_{\boldsymbol{s} \in \{0,1\}^p} \sup_{\substack{\|\mathbf{x}\|_\infty \leq 1 \\ \mathbb{1}_{\mathrm{supp}(\mathbf{x})} = \boldsymbol{s}}} \mathbf{x}^T\mathbf{y} - \mathbb{1}^T\boldsymbol{s} \\
&= \max_{\boldsymbol{s} \in \{0,1\}^p} |\mathbf{y}|^T\boldsymbol{s} - \mathbb{1}^T\boldsymbol{s} \\
&= \sum_{|y_i|>1} |y_i|
\end{aligned}$$





Hard Thresholding

Example: Tight convex relaxation of sparsity

The convex envelope of the ℓ_0-"norm", over the unit ℓ_∞-ball, is the ℓ_1-norm.
Proof:

1. $\|\mathbf{y}\|_0^* = \sum_{|y_i|>1} |y_i|$.
2. Compute the conjugate $\|\cdot\|_0^{**}$ of $\|\cdot\|_0^*$, for all $\mathbf{x} \in \mathbb{R}^p$ such that $\|\mathbf{x}\|_\infty \leq 1$:

$$\begin{aligned}
\|\mathbf{x}\|_0^{**} &= \sup_{\mathbf{y}\in\mathbb{R}^p} \mathbf{x}^T\mathbf{y} - \|\mathbf{y}\|_0^* \\
&= \sup_{\mathbf{y}\in\mathbb{R}^p} \mathbf{x}^T\mathbf{y} - \sum_{|y_i|>1} |y_i| \\
&= \sum_{i=1}^p |x_i| = \|\mathbf{x}\|_1
\end{aligned}$$

How do we compute the biconjugate of structured sparsity models in general?

- Computing the conjugate of $g(\mathbf{x}) = F(\mathrm{supp}(\mathbf{x}))$ is *NP-Hard*.
- Computing both the conjugate and the biconjugate of g becomes *tractable*, if F is *submodular*, or *linear* over an *integral polytope* domain.

Fenchel conjugate of structured sparsity models

Let $F(\boldsymbol{s}) : \{0,1\}^p \to \mathbb{R} \cup \{+\infty\}$ be any set function, then

$$\begin{aligned} g^*(\mathbf{y}) &= \sup_{\|\mathbf{x}\|_\infty \leq 1} \mathbf{x}^T\mathbf{y} - F(\operatorname{supp}(\mathbf{x})) \\ &= \sup_{\boldsymbol{s}\in\{0,1\}^p} \sup_{\substack{\|\mathbf{x}\|_\infty \leq 1 \\ \mathbb{1}_{\operatorname{supp}(\mathbf{x})} = \boldsymbol{s}}} \mathbf{x}^T\mathbf{y} - F(\boldsymbol{s}) \\ &= \max_{\boldsymbol{s}\in\{0,1\}^p} |\mathbf{y}|^T\boldsymbol{s} - F(\boldsymbol{s}) \end{aligned}$$

The Fenchel conjugate of general structured sparsity models is a discrete optimization problem which, in general, is *NP-Hard*.

Fenchel conjugate of submodular structured sparsity models

Let $F(\boldsymbol{s}) : \{0,1\}^p \to \mathbb{R} \cup \{+\infty\}$ be a submodular function,

$$\begin{aligned} g^*(\mathbf{y}) &= \sup_{\|\mathbf{x}\|_\infty \leq 1} \mathbf{x}^T \mathbf{y} - F(\mathrm{supp}(\mathbf{x})) \\ &= \max_{\boldsymbol{s} \in \{0,1\}^p} |\mathbf{y}|^T \boldsymbol{s} - F(\boldsymbol{s}) \\ &= \min_{\boldsymbol{s} \in \{0,1\}^p} -|\mathbf{y}|^T \boldsymbol{s} + F(\boldsymbol{s}) \end{aligned}$$

The Fenchel conjugate of submodular structured sparsity models is a *submodular minimization* problem, and hence is tractable.

Biconjugate of submodular structured sparsity models

Let $F(\boldsymbol{s}) : \{0,1\}^p \to \mathbb{R} \cup \{+\infty\}$ be a submodular function,

$$\begin{aligned} g^*(\mathbf{y}) &= \max_{\boldsymbol{s} \in \{0,1\}^p} |\mathbf{y}|^T \boldsymbol{s} - F(\boldsymbol{s}) \\ &= \max_{\boldsymbol{s} \in [0,1]^p} |\mathbf{y}|^T \boldsymbol{s} - F_L(\boldsymbol{s}) \end{aligned}$$

where F_L is the Lovász extension of F.

Recall from Lecture 2:

- The Lovász extension of $F(s) = -|\mathbf{y}|^T \boldsymbol{s}, \forall s \in \{0,1\}^p$ is $F_L(s) = -|\mathbf{y}|^T \boldsymbol{s}, \forall s \in [0,1]^p$.
- The Lovász extension of $\tilde{F}(\boldsymbol{s}) = -|\mathbf{y}|^T \boldsymbol{s} + F(\boldsymbol{s})$ is $\tilde{F}_L(\boldsymbol{s}) = -|\mathbf{y}|^T \boldsymbol{s} + F_L(\boldsymbol{s})$.
- $\min_{\boldsymbol{s} \in \{0,1\}^p} \tilde{F}(\boldsymbol{s}) = \min_{\boldsymbol{s} \in [0,1]^p} \tilde{F}_L(\boldsymbol{s})$

Biconjugate of submodular structured sparsity models

Let $F(\boldsymbol{s}) : \{0,1\}^p \to \mathbb{R} \cup \{+\infty\}$ be a submodular function,

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

$$= \max_{\boldsymbol{s} \in [0,1]^p} |\mathbf{y}|^T \boldsymbol{s} - F_L(\boldsymbol{s})$$

where F_L is the Lovász extension of F.

Theorem ([2])

Given a monotone submodular function F, the biconjugate of $g(\mathbf{x}) = F(\mathrm{supp}(\mathbf{x}))$ is given by $F_L(|\mathbf{x}|), \forall \mathbf{x} \in [-1,1]^p$.

Example (Sparsity)

Given the modular function $F(s) = \mathbb{1}^T \boldsymbol{s}$, the biconjugate of $g(\mathbf{x}) = \|\mathbf{x}\|_0$ is $F_L(|\mathbf{x}|)$.
Recall from lecture 2 that $F_L(s) = \mathbb{1}^T \boldsymbol{s}$, thus $g^{**}(x) = \mathbb{1}^T |\mathbf{x}| = \|\mathbf{x}\|_1$.

Fenchel conjugate of TU structured sparsity models

Let $F(\boldsymbol{s}) : \{0,1\}^p \to \mathbb{R} \cup \{+\infty\}$ be a *linear* function over an *integral* polytope.

In particular, let $F(\boldsymbol{s}) = \boldsymbol{e}^T\boldsymbol{s} + \iota_{\{\boldsymbol{M}\boldsymbol{s}\leq\mathbf{c}\}}(\boldsymbol{s})$ where $\boldsymbol{e} \in \mathbb{R}^p, \mathbf{c} \in \mathbb{Z}^\ell$ and $\boldsymbol{M} \in \mathbb{R}^{\ell\times p}$ is a *totally unimodular (TU)* matrix.

$$
\begin{aligned}
g^*(\mathbf{y}) &= \sup_{\|\mathbf{x}\|_\infty \leq 1} \mathbf{x}^T\mathbf{y} - F(\mathrm{supp}(\mathbf{x})) \\
&= \max_{\boldsymbol{s}\in\{0,1\}^p} |\mathbf{y}|^T\boldsymbol{s} - F(\boldsymbol{s}) \\
&= \max_{\boldsymbol{s}\in\{0,1\}^p} \{|\mathbf{y}|^T\boldsymbol{s} - \boldsymbol{e}^T\boldsymbol{s} : \boldsymbol{M}\boldsymbol{s} \leq \mathbf{c}\} \\
&= \max_{\boldsymbol{s}\in[0,1]^p} \{|\mathbf{y}|^T\boldsymbol{s} - \boldsymbol{e}^T\boldsymbol{s} : \boldsymbol{M}\boldsymbol{s} \leq \mathbf{c}\} \qquad \text{(cf., Lecture 2)}
\end{aligned}
$$

The Fenchel conjugate of TU structured sparsity models is a *linear program*, and hence is tractable.

Biconjugate of TU structured sparsity models

Let $F(\boldsymbol{s}) = \boldsymbol{e}^T\boldsymbol{s} + \iota_{\{\boldsymbol{M}\boldsymbol{s}\leq \mathbf{c}\}}(\boldsymbol{s})$ where $\boldsymbol{e} \in \mathbb{R}^p, \mathbf{c} \in \mathbb{Z}^\ell$ and $\boldsymbol{M} \in \mathbb{R}^{\ell\times p}$ is a *TU* matrix.

$$g^*(\mathbf{y}) = \max_{s\in[0,1]^p} \{|\mathbf{y}|^T\boldsymbol{s} - \boldsymbol{e}^T\boldsymbol{s} : \boldsymbol{M}\boldsymbol{s} \leq \mathbf{c}\}$$

Theorem ([4])

Given $F(\boldsymbol{s}) = \boldsymbol{e}^T\boldsymbol{s} + \iota_{\{\boldsymbol{M}\boldsymbol{s}\leq \mathbf{c}\}}(\boldsymbol{s})$*, the biconjugate of* $g(\mathbf{x}) = F(\mathrm{supp}(\mathbf{x}))$*,*
$\forall \mathbf{x} \in [-1,1]^p$ *is given by:*

$$g^{**}(\mathbf{x}) = \min_{s\in[0,1]^p} \{\boldsymbol{e}^T\boldsymbol{s} : \boldsymbol{M}\boldsymbol{s} \leq \mathbf{c}, \boldsymbol{s} \geq |\mathbf{x}|\}$$

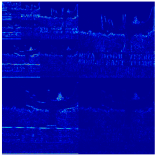
if $\exists \boldsymbol{s} \in [0,1]^p$ *such that* $\boldsymbol{M}\boldsymbol{s} \leq \mathbf{c}, \boldsymbol{s} \geq |\mathbf{x}|$*, and infinity otherwise.*

Example (Sparsity)

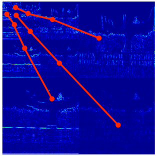
Given the function $F(s) = \mathbb{1}^T\boldsymbol{s}$, the biconjugate of $g(\mathbf{x}) = \|\mathbf{x}\|_0$ is given by:

$$g^{**}(\mathbf{x}) = \min_{s\in[0,1]^p} \{\mathbb{1}^T\boldsymbol{s} : \boldsymbol{s} \geq |\mathbf{x}|\} = \mathbb{1}^T|\mathbf{x}| = \|\mathbf{x}\|_1$$

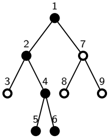
Example of TU structure: Tree sparsity



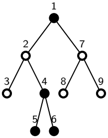
Wavelet coefficients



Wavelet tree



Valid selection of nodes



Invalid selection of nodes

We seek the sparsest signal with a rooted connected tree support. [3]

Objective: $\|\mathbf{x}\|_0 \equiv \mathbb{1}^T \boldsymbol{s}$ s.t. $\mathbb{1}_{\mathrm{supp}(\mathbf{x})} = \boldsymbol{s}$

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

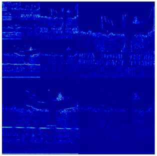
$$\boxed{\boldsymbol{T}\mathbb{1}_{\mathrm{supp}(\mathbf{x})} := \boldsymbol{T}\boldsymbol{s} \geq 0}$$

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

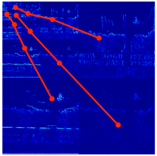
Recall that *any* directed edge-node incidence matrix is *TU*.

$$\boldsymbol{T} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix}$$

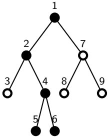
Example of TU structure: Tree sparsity



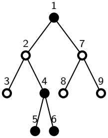
Wavelet coefficients



Wavelet tree



Valid selection of nodes



Invalid selection of nodes

We seek the sparsest signal with a rooted connected tree support. [3]

Objective: $\|\mathbf{x}\|_0 \equiv \mathbb{1}^T \boldsymbol{s}$ s.t. $\mathbb{1}_{\mathrm{supp}(\mathbf{x})} = \boldsymbol{s}$

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

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

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

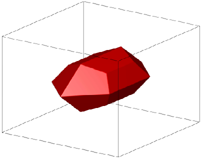
Recall that *any* directed edge-node incidence matrix is *TU*.

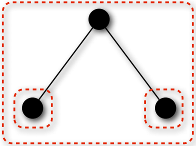
$$\boldsymbol{T} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix}$$

Biconjugate: Tractable! $\sum_{\mathcal{G} \in \mathfrak{G}_H} \|x_{\mathcal{G}}\|_\infty$

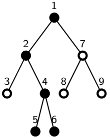
This is known as the hierarchical group lasso [10, 6].

Example of TU structure: Tree sparsity





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



valid selection of nodes

We seek the sparsest signal with a rooted connected tree support. [3]

Objective: $\|\mathbf{x}\|_0 \equiv \mathbb{1}^T \boldsymbol{s}$ s.t. $\mathbb{1}_{\mathrm{supp}(\mathbf{x})} = \boldsymbol{s}$

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

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

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

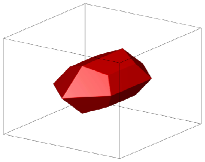
Recall that *any* directed edge-node incidence matrix is *TU*.

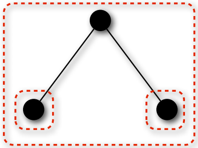
$$\boldsymbol{T} = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix}$$

Biconjugate: Tractable! $\sum_{\mathcal{G}\in\mathfrak{G}_H} \|x_{\mathcal{G}}\|_\infty$

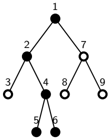
This is known as the hierarchical group lasso [10, 6].

Example of submodular structure: Tree sparsity





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



valid selection of nodes

Tree sparsity can be enforced by a submodular function too:

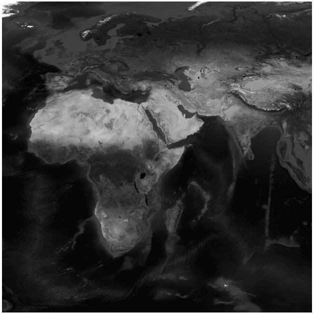
$$F(S) = \sum_{G \in \mathfrak{G}_H} \mathbb{1}_{G \cap S \neq \emptyset}(S)$$

Recall that F is submodular, and its Lovász extension $F_L(s) = \sum_{G \in \mathfrak{G}_H} \max_{k \in G} s_k$.

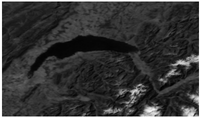
$$g^{**}(\mathbf{x}) = F_L(|\mathbf{x}|) = \sum_{G \in \mathfrak{G}_H} \max_{k \in G} |x_k| = \sum_{\mathcal{G} \in \mathfrak{G}_H} \|x_{\mathcal{G}}\|_\infty$$

Tree sparsity example: 1:100-compressive sensing [9, 1]

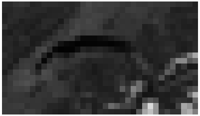
World [1Gpix]



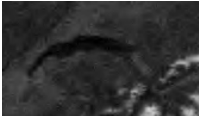
Lac Léman



World [10Mpix]

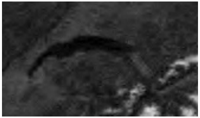


sparse



PNSR $= 31.83$db

tree-sparse



PNSR $= 32.48$db

PSNR: Peak signal-to-noise ratio

References I

[1] Ben Adcock, Anders C. Hansen, Clarice Poon, and Bogdan Roman.
Breaking the coherence barrier: A new theory for compressed sensing.
http://arxiv.org/abs/1302.0561, Feb. 2013.

[2] F. Bach.
Structured sparsity-inducing norms through submodular functions.
In *NIPS*, pages 118–126, 2010.

[3] R.G. Baraniuk, V. Cevher, M.F. Duarte, and C. Hegde.
Model-based compressive sensing.
Information Theory, IEEE Transactions on, 56(4):1982–2001, 2010.

[4] Marwa El Halabi and Volkan Cevher.
A totally unimodular view of structured sparsity.
In *18th Int. Conf. Artificial Intelligence and Statistics*, 2015.

[5] Rémi Gribonval, Volkan Cevher, and Mike E. Davies.
Compressible distributions for high-dimensional statistics.
IEEE Trans. Inf. Theory, 58(8):5016–5034, 2012.

[6] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach.
Proximal methods for hierarchical sparse coding.
J. Mach. Learn. Res., 12:2297–2334, 2011.

References II

[7] Samet Oymak, Christos Thrampoulidis, and Babak Hassibi.
Simple bounds for noisy linear inverse problems with exact side information.
2013.
arXiv:1312.0641v2 [cs.IT].

[8] R.T. Rockafellar.
Convex analysis.
Princeton University Press (Princeton, NJ), 1970.

[9] Quoc Tran-Dinh and Volkan Cevher.
Constrained convex minimization via model-based excessive gap.
In *Advances in Neural Information Processing Systems*, pages 721–729, 2014.

[10] 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.