lions@epfl

EE-731: ADVANCED TOPICS IN DATA SCIENCES
LABORATORY FOR INFORMATION AND INFERENCE SYSTEMS
SPRING 2016

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

INSTRUCTOR: VOLKAN CEVHER

SCRIBERS: SOROOSH SHAFIEEZADEH-ABADEH, MICHAËL DEFFERRARD

STRUCTURED SPARSITY

Many problems in signal processing, image compression, and communications can be cast as a linear regression problem where an unknown signal $\mathbf{x}^\natural \in \mathbb{C}^p$ is related to its observations $\mathbf{b} \in \mathbb{C}^p$ via

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

where $\mathbf{A} \in \mathbb{C}^{n \times p}$ is a known measurement matrix and $\mathbf{w} \in \mathbb{C}^n$ is an unknown noise. The goal is to recover an accurate estimate $\hat{\mathbf{x}}$ of the signal $\mathbf{x}^\natural$ such that $\|\hat{\mathbf{x}} - \mathbf{x}^\natural\| \leq \epsilon$.

There exists three regimes to that problem:

- Critical ($n = p$ if the rows of $\mathbf{A}$ are linearly independent): for every variable giving a degree of freedom, there exists a corresponding constraint. The problem has therefore a unique solution (which is not $\mathbf{x}^\natural$ in general because of the noise).
- Overdetermined: there is more observations than the dimension of the unknown signal ($n > p$). There is thus no solution and an approximation is needed, e.g. least squares if $\mathbf{w}$ is assumed to be Gaussian noise.
- Underdetermined: there is less observations than the dimension of the unknown signal ($n < p$). There exists infinitely many solutions and some prior knowledge is needed to hope for a correct recovery.

These notes will discuss such a prior, known as structured sparsity.

1 Background on Sparse Signals

In statistics and mathematics, linear least squares is an approach to fit a mathematical or statistical model to data in cases where the idealized value provided by the model for any data point is expressed linearly in terms of the unknown parameters of the model. The resulting fitted model can be used to summarize the data, predict unobserved values from the same system or understand the mechanisms that may underlie it.

The method of least squares (LS) in linear model is a standard approach in statistics to approximate solution of overdetermined systems. The LS estimator for $\mathbf{x}^\natural$ given $\mathbf{A}$ and $\mathbf{b}$ is defined as

$$\hat{\mathbf{x}}_{\text{LS}} \in \arg\min_{\mathbf{x} \in \mathbb{C}^p} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2^2.$$

If the measurement matrix $\mathbf{A}$ is full column rank, one can show that the LS solution can be uniquely defined as $\hat{\mathbf{x}}_{\text{LS}} = (\mathbf{A}^\top\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{b} = \mathbf{A}^\dagger\mathbf{b}$. However, in the underdetermined case, $\mathbf{A}$ is not full column rank. Hence, one can only conclude that $\mathbf{x}_{\text{LS}} \in \{\mathbf{A}^\dagger\mathbf{b} + \mathbf{h} : \mathbf{h} \in \text{null}(\mathbf{A})\}$ which means we cannot recover the unknown signal uniquely. The downside is that the estimation error $\|\hat{\mathbf{x}}_{\text{LS}} - \mathbf{x}^\natural\|_2$ can be arbitrarily large. Figure 1 illustrates this phenomenon, where we are interested in solving a linear regression model in a noiseless setting, i.e. $\mathbf{A}\mathbf{x} = \mathbf{b}$. In this case, choosing a candidate solution with the least amount of energy (i.e. smallest norm $\|\mathbf{x}\|_2$) leads to a significant estimation error. The following proposition asserts that the estimation error may not diminish unless n is very close to p.

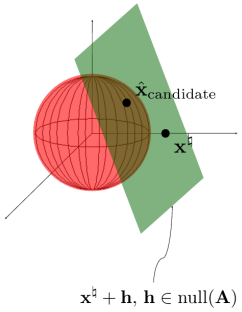


Figure 1: The estimation error for a candidate solution given by LS method.

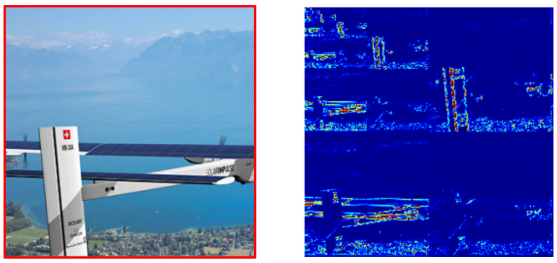


Figure 2: The image $\mathbf{x}$ and its sparse representation $\boldsymbol{\alpha}$ in the wavelet domain $\boldsymbol{\Psi}$.

Proposition 1.1 ([4]). *Suppose that $\mathbf{A} \in \mathbb{R}^{n\times p}$ is a matrix of i.i.d. standard Gaussian random variables, and $\mathbf{w} = 0$. Then, we have*

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

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

The following comments are in order:

1. It is impossible to estimate $\mathbf{x}^\sharp$ accurately using $\hat{\mathbf{x}}_{\text{candidate}}$ when $n \ll p$ even if $\mathbf{w} = 0$.
2. The statistical error $\left\|\hat{\mathbf{x}}_{\text{candidate}} - \mathbf{x}^\sharp\right\|_2^2$ can also be arbitrarily large when $\mathbf{w} \neq 0$. The solution is therefore not robust.
3. We need additional information on $\mathbf{x}^\sharp$ in order to have an accurate estimation. This knowledge is often a constraint on the complexity of $\mathbf{x}^\sharp$, e.g. that it admits an accurate approximation by a sparse set of coefficients.

Any signal $\mathbf{x} \in \mathbb{R}^p$ can be represented in term of coefficients $\boldsymbol{\alpha} \in \mathbb{R}^p$ in an orthogonal basis $\boldsymbol{\Psi} \in \mathbb{R}^{p\times p}$ via $\mathbf{x} = \boldsymbol{\Psi}\boldsymbol{\alpha}$. Signal $\mathbf{x}$ has a sparse representation if only $s \ll p$ entries of $\boldsymbol{\alpha}$ are nonzero. Indeed, sparse representations frequently appear in signal and image processing applications. For instance, Figure 2 shows the sparse structure of an image represented in the wavelet basis. However, we do not know the location of the nonzero entries of $\boldsymbol{\alpha}^\sharp$.

To account for sparse signals in an appropriate basis, one should modify the linear regression problem as $\mathbf{b} = \mathbf{A}\mathbf{x}^\sharp + \mathbf{w} = \mathbf{A}\boldsymbol{\Psi}\boldsymbol{\alpha}^\sharp + \mathbf{w}$. Hence, given the measurement matrix $\mathbf{A}$, the basis matrix $\boldsymbol{\Psi}$, and the observations $\mathbf{b}$, one needs to select the s columns of matrix $\tilde{\mathbf{A}} = \mathbf{A}\boldsymbol{\Psi}$ which corresponds to nonzero entries of $\boldsymbol{\alpha}^\sharp$ and solve the least square problem. When s is strictly smaller than the number of observation n, the least squares solution is accurate. Given the $s-$sparse least squares solution $\boldsymbol{\alpha}_{\text{LS}}$, we can recover the signal as $\mathbf{x} = \boldsymbol{\Psi}\boldsymbol{\alpha}_{\text{LS}}$. Figure 3 is an illustration of sparse recovery for a 3-sparse vector. In practice, however, we do not know the location of the nonzero entries of $\boldsymbol{\alpha}^\sharp$. Instead, we are interested in solving an optimization problem which induces sparsity to the solution.

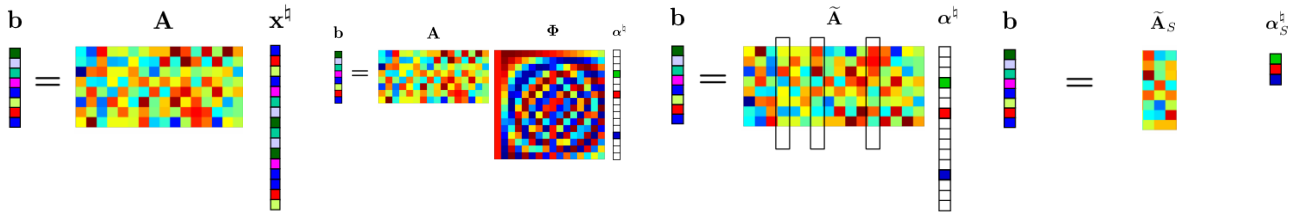
In the sequel, we will assume without loss of generality that the signal $\mathbf{x}^\sharp$ is sparse or compressible in the canonical domain so that the sparsity basis $\boldsymbol{\Psi}$ is the identity and $\boldsymbol{\alpha}^\sharp = \mathbf{x}^\sharp$.

The most natural way to induce sparsity is to consider the sparse estimator

$$\hat{\mathbf{x}} \in \arg\min_{\mathbf{x}\in\mathbb{R}^p} \{\|\mathbf{x}\|_0 : \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 \leq \|\mathbf{w}\|_2\}, \tag{$\mathcal{P}_0$}$$

where $\|\mathbf{x}\|_0 = \mathbf{1}^\top\mathbf{s}, \mathbf{s} = \mathbb{1}_{\text{supp}(\mathbf{x})}$, and $\text{supp}(\mathbf{x}) = \{i : x_i \neq 0\}$. The sparse estimator corresponding to $(\mathcal{P}_0)$ is the solution of a constrained optimization problem. The straightforward method to represent $(\mathcal{P}_0)$ as an unconstrained optimization problem is to consider the sparse estimator

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



(a) Relation between unknown signal and observations.

(b) Sparse representation of unknown signal through basis $\boldsymbol{\Psi}$.

(c) Select collumns which correspond to nonzero entries.

(d) Solve the LS problem to find the values of the nonzero entries of $\boldsymbol{\alpha}^\sharp$

Figure 3: Sparse recovery of a 3-sparse vector.

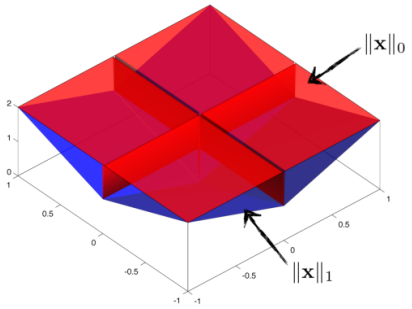


Figure 4: The relation between ℓ_0 and ℓ_1 norms over the region $\mathcal{X} = \{\mathbf{x} : \mathbf{x} \in [-1, 1]^2\}$.

where ρ is the regularization factor which reflects the trade-off between data fidelity and sparsity. These natural sparse estimators, which are based on ℓ_0-norm[1], have a sample complexity of $O(s)$. In other words, a $s-$sparse signal of length p can be recovered using only $n = O(s)$ linear measurements. Nevertheless, these sparse estimators are known to be computationally hard (NP-Hard) and not robust to noise.

A standard convex relaxation of the problem is to replace the ℓ_0-norm (cardinality of the support) by the ℓ_1-norm. Estimators may then be obtained as solutions of convex programs. Hence, we consider the following constrained convex optimization problem:

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

where $\|\mathbf{x}\|_1 := \mathbf{1}^\top|\mathbf{x}| = \sum_i |x_i|$. This problem, known as Basis Pursuit (BP), can similarly be represented in the unconstrained form

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

known as the Least Absolute Shrinkage and Selection Operator (LASSO). In addition to a polynomial computational effort, convex sparse estimators (BP) and (LASSO) are robust to noise and have sample complexity of order $O(s\log(\frac{p}{s}))$ [6], which is a key result of the compressed sensing theory.

Why is the ℓ_1-norm a good proxy for ℓ_0? Intuitively, Figure 4 shows that the ℓ_1-norm is the largest possible convex function which lower bounds the ℓ_0-norm over the region $[-1, 1]^2$. In general, we aim to show that the ℓ_1-norm is the convex envelope of $\|\cdot\|_0$ over the region $\mathcal{X} = \{\mathbf{x} : \mathbf{x} \in [-1, 1]^p\}$.

Definition 1 (Convex envelope)**.** *The convex envelope of a (non-convex) function f over some region $\mathcal{X}$ is the largest possible convex lower bound of f over the region $\mathcal{X}$, i.e.,*

$$\text{conv}_{f,\mathcal{X}}(\boldsymbol{x}) = \sup\{c(\boldsymbol{x}) : c(\boldsymbol{x}') \leq f(\boldsymbol{x}')\ \forall \boldsymbol{x}' \in \mathcal{X}, c \text{ is convex}\}.$$

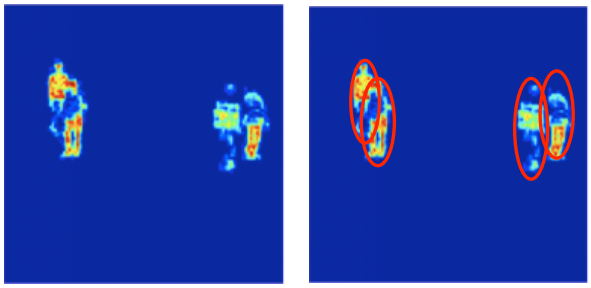
2 From sparsity to structured sparsity

While many natural and man-made signals and images can be described to the first-order as sparse or compressible, their sparse supports (sets of nonzero coefficients) often have an underlying structure. Figure 5 illustrates a real-world example of structured sparse support in image processing. Figure 5(a) represents a background subtracted image, which is sparse; while Figure 5(b) illustrates their clustered structure. Another well-known example is the tree-like structure of wavelet representations. Nevertheless, these structural properties are not exploited by the standard sparse recovery algorithms presented above.

Heuristically, a general $s-$sparse vector can be specified with $2s$ numbers: s for the values of the nonzero coefficients and another s for the locations of these coefficients. If the s significant coefficients have some special patterns, e.g. in Figure 5(b), then the indexing cost is significantly reduced and hence the total description is reduced. In fact, exploiting this embedded structure along with the sparse representation would potentially lead to:

1. reconstruction using fewer measurements (i.e., reduced sample complexity),
2. better reconstructed signals (i.e., better robustness to noise),
3. better interpretability,
4. although at the cost of usually slower reconstruction algorithms.

[1] ℓ_0-norm is not really a norm since it does not satisfy the triangle inequality.



(a) Background subtracted image. (b) Exisisting clusters in the foreground image.

Figure 5: Structured sparsity in background subtracted images.

One way to encode structure and sparsity is to define a combinatorial set function F on the support of the unknown parameter. Hence, the structured sparse estimator over the function $F(\text{supp}(\mathbf{x}))$ can be

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

which is the solution of a constrained optimization problem. Again, the constraint can be integrated in the objective, which gives the estimator

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

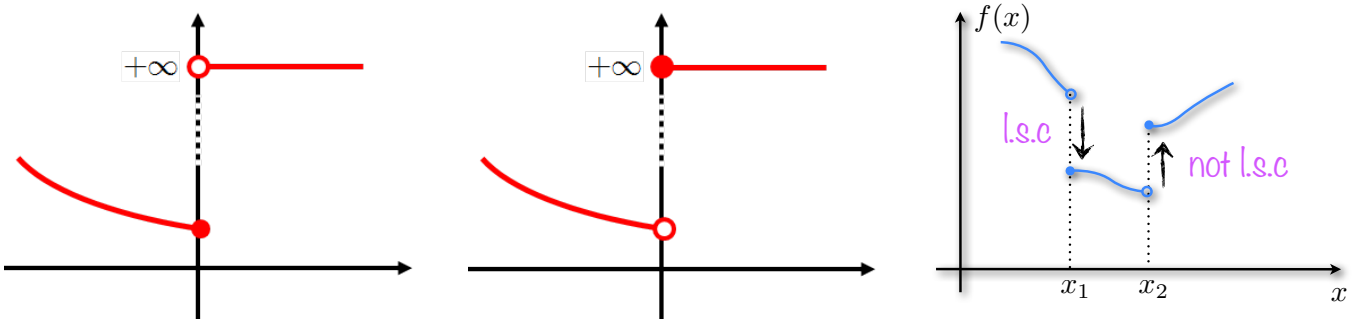
Determining a good convex surrogate of $F(\text{supp}(\mathbf{x}))$ can be done with case by case heuristics or by determining its convex envelope, which is mathematically given by its biconjugate, i.e., Fenchel conjugate. Let's introduce some definitions of the Fenchel conjugate that will be used to derive the convex envelope.

Definition 2 (Lower semi-continuity)**.** *A function $f : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ is lower semi-continuous if*

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

Note that a lower semi-continuous function is also called a closed function as its epigraph and sublevel set are closed. The epigraph, denoted epi f, describes the set of input-output pairs that f can achieve, as well as anything above, i.e., epi $f = \{(\mathbf{x}, t) : \mathbf{x} \in \mathbb{R}^p, t \in \mathbb{R}, f(\mathbf{x}) \leq t\}$. The sublevel set is a set of all points that achieve at most a certain value t for f; i.e., $\{\mathbf{x} \in \mathbb{R}^p : f(\mathbf{x}) \leq t\}$.

Figure 6 shows different types of functions. For instance, Figure 6(a) is a closed and lower semi-continuous function while Figure 6(b) is not (it is upper semi-continuous). Notice that the lower semi-continuous condition is important in convex optimization problems as it is essential for solving minimization problems in order to obtain its infimum[2]. Intuitively, a function is lower semi-continuous if it is continuous or, if not, it only jumps down. For instance, Figure 6(c) is not a lower semi-continuous function as it jumps up at x_2.



(a) A lower semi-continuous function.

(b) A function which is not a lower semi-continuous function.

(c) As a rule of thumb, a lower semi-continuous function *only jumps down*.

Figure 6: Lower semi-continuous functions.

[2] A function which may not be necessarily lower semi-continuous can still obtain its infimum.

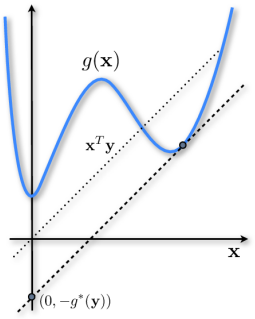


Figure 7: The geometric interpretation of the Fenchel conjugate.

Definition 3 (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:*

$$g^*(\boldsymbol{y}) = \sup_{\boldsymbol{x}\in\mathrm{dom}(g)} \left\{\boldsymbol{y}^\top\boldsymbol{x} - g(\boldsymbol{x})\right\},$$

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

For a given direction $\mathbf{y} \in \mathbb{R}^p$, the Fenchel conjugate $g^*(\mathbf{y})$ is the maximum gap between the linear function $\mathbf{x}^\top\mathbf{y}$ and $g(\mathbf{x})$. For instance in Figure 7, for a given $\mathbf{y} \in \mathbb{R}$, the Fenchel conjugate is the gap between the dotted and dashed lines. The following comments on the properties of the Fenchel conjugate are in order:

1. Given $\mathbf{x}^\star \in \arg\max_{\mathbf{x}\in\mathrm{dom}(g)}\{\mathbf{y}^\top\mathbf{x} - g(\mathbf{x})\}$, $\mathbf{x}^\star$ will lie on the convex envelope.
2. $g^\star$ may be seen as minus the intercept of the tangent to the graph of g with slope $\mathbf{y}$; i.e., the line $\mathbf{x}^\top\mathbf{y} - g^\star(\mathbf{y})$.
3. By definition of conjugation, g is always above the line $\mathbf{x}^\top\mathbf{y} - g^\star(\mathbf{y}), \forall\mathbf{y} \in \mathbb{R}^p$.
4. As a pointwise supremum of linear functions, g^* is always convex and lower semi-continuous, even if g is not.
5. If g is convex and lower semi-continuous itself, then its biconjugate coincides with g, i.e., $g^{**} = g$.
6. The biconjugate g^{**} is the lower semi-continuous convex envelope of g; i.e., the largest lower semi-continuous convex lower bound on g.

We now give the proof of why the ℓ_1-norm is the "best" convex relaxation of the ℓ_0-norm (two other proofs are presented in Corollary 2.3 and 3.2).

Proposition 2.1. *The lower semi-continuous convex envelope of the ℓ_0-norm, over the unit $\ell_\infty-$ball is the ℓ_1-norm.*

Proof. We start the proof by computing the conjugate of the ℓ_0-norm, for all $\mathbf{y} \in \mathbb{R}^p$:

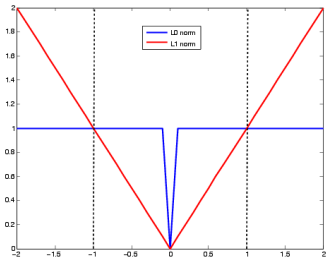
$$\begin{aligned}
\|\mathbf{y}\|_0^* &= \sup_{\|\mathbf{x}\|_\infty\leq 1} \mathbf{y}^\top\mathbf{x} - \|\mathbf{x}\|_0 \\
&= \sup_{\mathbf{s}\in\{0,1\}^p} \sup_{\substack{\|\mathbf{x}\|_\infty\leq 1 \\ \mathbb{1}_{\mathrm{supp}(\mathbf{x})}=s}} \mathbf{y}^\top\mathbf{x} - \mathbf{1}^\top\mathbf{s} \\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

where the second equality follows from the definition of the ℓ_0-norm (i.e., $\|x\|_0 = \mathbf{1}^\top\mathbb{1}_{\mathrm{supp}(\mathbf{x})}$), the third equality follows from the explicit maximization over $\mathbf{x}$ (x_i would be 1 or -1 depending on the sign of y_i), and the last equality holds by explicitly maximizing over $\mathbf{s}$. Figure 8 illustrates the relation between the ℓ_0 and ℓ_1 norms and depicts the Fenchel conjugate of the ℓ_0-norm.

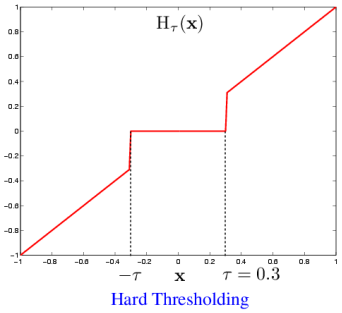
We then compute the biconjugate of the ℓ_0-norm for all $\mathbf{x} \in \mathbb{R}^p$ such that $\|\mathbf{x}\|_\infty \leq 1$:

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

where the last equality holds by explicitly maximizing over $\mathbf{y}$ (y_i is 1 or -1 depending on the sign of s_i). Thus, the claim is proved. □



(a) The realation between the ℓ_0 and ℓ_1 norms.



(b) The Fenchel conjugate of the ℓ_0-norm.

Figure 8: Convex relaxation of the ℓ_0-norm.

Now the question is how to compute the biconjugate of structured sparsity models. Note that computing the conjugate of $g(\mathbf{x}) = F(\text{supp}(\mathbf{x}))$ is NP-Hard in general. However, from previous lecture, we know that computing both conjugate and biconjugate of g becomes tractable if the set function F is submodular, or linear over an integral polytope domain. Formally speaking, let $F(\mathbf{s}) = \{0,1\}^p \to \mathbb{R} \cup \{+\infty\}$ be any set function, then

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

Hence, computing the Fenchel conjugate for a general structured sparse model reduces to solving a discrete optimization problem which is known to be NP-Hard in general. However, the Fenchel conjugate of a submodular structured sparse model can be written as the submodular minimization problem

$$g^*(\mathbf{y}) = \min_{\mathbf{s}\in\{0,1\}^p} -|\mathbf{y}|^\top \mathbf{s} + F(\mathbf{s}),$$

which is tractable. Recall from previous lecture that we can relax the constraints in submodular minimization problems using the Lovász extension. In the following, we denote the Lovász extension of the set function F by F_L. For the set function $F(s) = -|\mathbf{y}|^\top \mathbf{s}, \forall \mathbf{s} \in \{0,1\}^p$, the Lovász extension is $F_L(\mathbf{s}) = -|\mathbf{y}|^\top s, \forall \mathbf{s} \in [0,1]^p$. Hence, computing the Fenchel conjugate of a submodular structured sparse model can be reformulated as

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

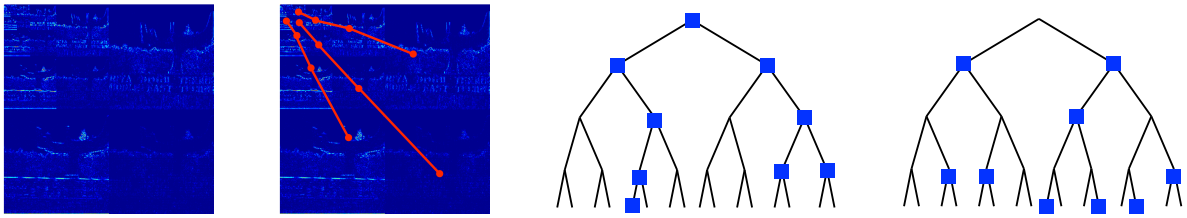
Theorem 2.2 ([1]). *Given a monotone submodular function F, the biconjugate of $g(\boldsymbol{x}) = F(\text{supp}(\boldsymbol{x}))$ is given by $F_L(|\boldsymbol{x}|), \forall \boldsymbol{x} \in [-1,1]^p$.*

Proof. For all $\mathbf{x}$ such that $\|\mathbf{x}\|_\infty \leq 1$, we have

$$\begin{aligned} g^{**}(\mathbf{x}) = & \max_{\mathbf{y}} \mathbf{x}^\top \mathbf{y} - g^*(\mathbf{y}) \\ = & \max_{\mathbf{y}} \min_{\mathbf{s}\in[0,1]^p} \mathbf{x}^\top \mathbf{y} - |\mathbf{y}|^\top \mathbf{s} + F_L(\mathbf{s}) \\ = & \min_{\mathbf{s}\in[0,1]^p} \max_{\mathbf{y}} \mathbf{x}^\top \mathbf{y} - |\mathbf{y}|^\top \mathbf{s} + F_L(\mathbf{s}) \\ = & \min_{\substack{\mathbf{s}\in[0,1]^p \\ \mathbf{s}\geq|\mathbf{x}|}} F_L(\mathbf{s}) \\ = & F_L(|\mathbf{x}|), \end{aligned}$$

where third equality holds by strong duality due to the Slater's condition [3], and the last equality follows as the set function F is monotone implying that F_L is non-decreasing with respect to all of its components. □

Corollary 2.3. *Given the ℓ_0-norm and its corresponding submodular function $F(\mathbf{s}) = \mathbf{1}^\top \mathbf{s}$, the biconjugate of $g(\boldsymbol{x}) = \|\boldsymbol{x}\|_0$ is $F_L(|\boldsymbol{x}|)$. Recall from lecture 2 that $F_L(s) = \mathbf{1}^\top \mathbf{s}$, thus $g^{**}(\boldsymbol{x}) = \mathbf{1}^\top |\boldsymbol{x}| = \|\boldsymbol{x}\|_1$.*



(a) Wavelet coefficients (b) Wavelet tree (c) Valid selection of nodes (d) Invalid selection of nodes

Figure 9: Tree sparsity as an example of a TU structure.

3 Totally unimodular structured sparsity models

We now show that many structured sparsity models can be naturally represented by linear matrix inequalities on the support of the unknown parameters, where the constraint matrix has a totally unimodular (TU) structure. For such structured models, it is known that tight convex relaxations can be obtained in polynomial time via linear programming.

Let $F(\mathbf{s}) : \{0,1\}^p \to \mathbb{R} \cup \{+\infty\}$ be a linear function over an integral polytope, i.e., $F(\mathbf{s}) = \mathbf{e}^\top \mathbf{s} + \iota_{\{\mathbf{M}\mathbf{s} \le \mathbf{c}\}}(\mathbf{s})$ where $\mathbf{e} \in \mathbb{R}^p, \mathbf{c} \in \mathbb{Z}^\ell$ and $\mathbf{M} \in \mathbb{R}^{\ell \times p}$ is a totally unimodular (TU) matrix. Hence

$$\begin{aligned} g^*(\mathbf{y}) = & \sup_{\|\mathbf{x}\|_\infty \le 1} \mathbf{y}^\top \mathbf{x} - F(\mathrm{supp}(\mathbf{x})) \\ = & \max_{\mathbf{s} \in \{0,1\}^p} |\mathbf{y}|^\top \mathbf{s} - F(\mathbf{s}) \\ = & \max_{\mathbf{s} \in \{0,1\}^p} \{|\mathbf{y}|^\top \mathbf{s} - \mathbf{e}^\top \mathbf{s} : \mathbf{M}\mathbf{s} \le \mathbf{c}\} \\ = & \max_{\mathbf{s} \in \{0,1\}^p} \{|\mathbf{y}|^\top \mathbf{s} - \mathbf{e}^\top \mathbf{s} : \mathbf{M}\mathbf{s} \le \mathbf{c}\}. \end{aligned}$$

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

Theorem 3.1 ([5]). *Given $F(\mathbf{s}) = \mathbf{e}^\top \mathbf{s} + \iota_{\{\mathbf{M}\mathbf{s} \le \mathbf{c}\}}(\mathbf{s})$, the biconjugate of $g(\boldsymbol{x}) = F(\mathrm{supp}(\boldsymbol{x})), \forall \boldsymbol{x} \in [-1,1]^p$ is given by*

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

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

Corollary 3.2. *Given the ℓ_0-norm and its corresponding TU submodular function $F(\mathbf{s}) = \mathbf{1}^\top \mathbf{s}$ (no constraint), the biconjugate of $g(\boldsymbol{x}) = \|\boldsymbol{x}\|_0$ is given by*

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

In this lecture, we only focus on the tree-structured sparse representation as additional prior information, which emerges from the wavelet transform of natural images (see Figures 9(a) and 9(b)). As mentioned before, considering structured sparse models lead to sample complexity reduction. For instance, it was proved in [2] that the greedy method CoSaMP can recover an $s-$sparse signal using tree-structured sparsity with a sample complexity of $O(s)$, to be compared to complexity $O(s \log(\frac{p}{s}))$ of LASSO. In practice, convex methods dealing with tree-sparse signals have similar sample complexity but it has not been proved yet.

Let first describe tree sparsity in the context of sparse wavelet decompositions. Of particular interest to us is the following problem: given an arbitrary signal $\mathbf{x} \in \mathbb{R}^p$, find the $s-$sparse tree signal that minimizes the fitting error. For instance, Figure 9(c) is a valid or feasible solution for a 9−sparse tree signal while Figure 9(d) is not a valid one and should be discarded from the search space. In fact, the tree sparsity model comprises the set of $s-$sparse signals whose nonzero coefficients form a rooted, connected subtree. It was shown in [2] that the size of this model is upper bounded by $(2e)^s/(s+1)$ which is far less than $\binom{p}{s}$ considered by LASSO for large p. This will help us be more robust to noise and improve the model performance.

The tree sparsity model consists of a collection of all groups $\mathfrak{G}_H = \{\mathcal{G}_1, \ldots, \mathcal{G}_M\}$, where each group corresponds to a node and all its descendants. See Figure 10 for an example. Tree sparse structure can also be encoded by a set of linear constrains which correspond to the relation between parent and child nodes. This structure can be formulated as $s_{\text{parent}} \ge s_{\text{child}}$ over the tree $\mathcal{T}$. Therefore, we can aggregate all the information into the so called directed edge-node incident matrix denoted by $\mathbf{T}$ with the constraint $\mathbf{T}\mathbb{1}_{\mathrm{supp}(\mathbf{x})} := \mathbf{T}\mathbf{s} \ge \mathbf{0}$. Notice that any directed edge-node incident matrix is TU. Hence, we can apply Theorem 3.1 to extract tractable reformulation for tree-sparse signals.

Corollary 3.3 (Tree sparsity using TU structure)**.** *Tree sparsity can be enforced by the function $F(\mathbf{s}) = \mathbf{1}^\top \mathbf{s} + \iota_{\mathbf{T}\mathbf{s} \ge \mathbf{0}}(\mathbf{s})$, the biconjugate of $g(\boldsymbol{x}) = F(\mathrm{supp}(\boldsymbol{x})), \forall \boldsymbol{x} \in [-1,1]^p$ is given by*

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

which coincides with $g^{**}(\boldsymbol{x}) = F_L(|\boldsymbol{x}|) = \sum_{G \in \mathfrak{G}_H} \|x_{\mathcal{G}}\|_\infty$. (See [5, Proposition 2].)

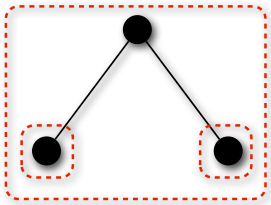


Figure 10: The collection of all groups for this simple tree is $\mathfrak{G}_H = \{\{1,2,3\},\{2\},\{3\}\}$.

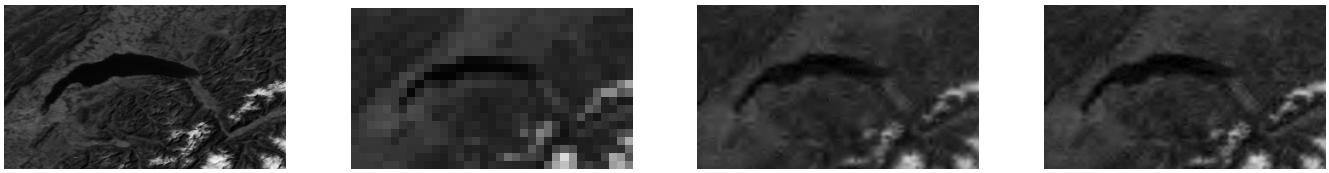
Corollary 3.4 (Tree sparsity using submodular structure). *Tree sparsity can be enforced by the submodular function*

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

Recall that F is submodular and its Lovász extension is $F_L = \sum_{G \in \mathfrak{G}_H} \max_{k \in G} s_k$. The biconjugate of $g(\boldsymbol{x}) = F(\mathrm{supp}(\boldsymbol{x})), \forall \boldsymbol{x} \in [-1,1]^p$ is given by

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

Figure 11 shows a comparison between the tree-sparse reconstruction and the sparse construction (LASSO) in an imaging experiment, where the peak signal-to-noise ratio (PNSR) is used as a quality measure. Figures 11(a) and 11(b) show the Lac Léman area of a 1Gpix and 10Mpix image, respectively. The reconstructed image from the 10Mpix image using sparse and tree-sparse models are shown in Figure 11(c) and 11(d), respectively. The PNSR for the tree-sparse model is 32.48 db while it is 31.83 db for the regular sparse model.



(a) Lac Léman 1Gpix (b) Lac Léman 10Mpix (c) Sparse model: PNSR = 31.83 db (d) Tree-sparse model: PNSR = 32.48 db

Figure 11: 1:100 compressive sensing

References

[1] F. R. Bach. Structured sparsity-inducing norms through submodular functions. In *Advances in Neural Information Processing Systems*, pages 118–126, 2010.

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

[3] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[4] R. Gribonval, V. Cevher, and M. E. Davies. Compressible distributions for high-dimensional statistics. *IEEE Transactions on Information Theory*, 58(8):5016–5034, 2012.

[5] M. E. Halabi and V. Cevher. A totally unimodular view of structured sparsity. *arXiv preprint arXiv:1411.1990*, 2014.

[6] S. Oymak, C. Thrampoulidis, and B. Hassibi. Simple bounds for noisy linear inverse problems with exact side information. *arXiv preprint arXiv:1312.0641*, 2013.