6.883 Learning with Combinatorial Structure
Note for Lecture 12
Author: Chiyuan Zhang

1 Sparsity and ℓ_1 relaxation

Last time we talked about sparsity and characterized when an ℓ_1 relaxation could recover the original sparse solution of the ℓ_0 problem. Today we look at a specific setting, in which the behavior could be characterized more easily.

Specifically, instead of reconstructing the signal using a sparse linear combination of atoms from a overcomplete dictionary, we assume the signal is sparse in the original Euclidean basis. We get a noisy observation z of the signal, and would like to recover the original sparse signal by

$$\min_w \frac{1}{2}\|z-w\|^2, \quad s.t. \|w\|_1 \leq R$$

Figure 1 gives a nice geometric intuition of why the ℓ_1 constrained problem recovers sparse solutions, while ℓ_2 constrained problem does not. Under this setting, this phenomenon could also be explained algebraically. In particular, we consider an equivalent formulation with the constraints replaced by a regularizer penalty with appropriately chosen regularization coefficient λ:

$$\min_w \frac{1}{2}\|z-w\|^2 + \lambda\|w\|_1 \tag{1}$$

Since this is a convex optimization problem, from our previous lectures, we can write the optimality condition as

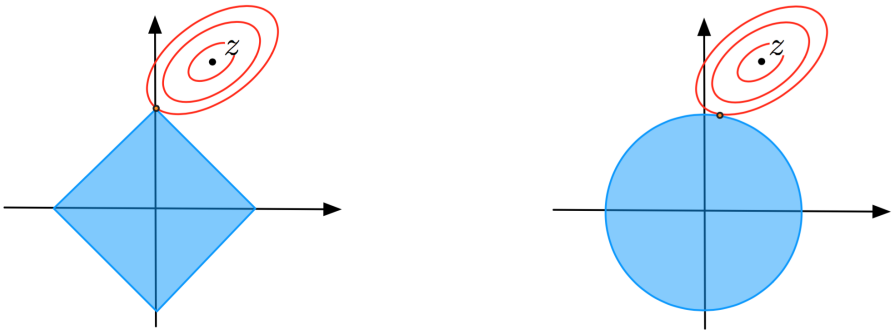
$$0 \in \partial_w \left(\frac{1}{2}\|z-w\|^2 + \lambda\|w\|_1 \right)$$

In other words, $\exists g \in \partial_w \|w\|_1$ such that $(w-z) + \lambda g = 0$. In this case, the optimal solution is $w^\star = z - \lambda g$. To get the subdifferential of the ℓ_1 norm, notice that

$$\|w\|_1 = \max_{s \in [-1,1]^d} s^\top w \quad \Rightarrow \quad \operatorname{sign}(w) \in \partial_w \|w\|_1$$

Where $\operatorname{sign}(\cdot)$ is the component-wise sign function, and when $w_i = 0$, $\operatorname{sign}(w_i)$ could take any value in $[-1,1]$. To make sure $\operatorname{sign}(w^\star)$ is consistent with g, we get the following rules

$$\begin{aligned} z_i > \lambda &\Rightarrow w_i^\star = z - \lambda,\ g_i = +1 \\ z_i < -\lambda &\Rightarrow w_i^\star = z + \lambda,\ g_i = -1 \\ |z_i| \leq \lambda &\Rightarrow w_i^\star = 0,\ g_i = z_i/\lambda \end{aligned} \tag{2}$$



$$\min_w \ \tfrac{1}{2}\|z-w\|^2 \quad \text{s.t.} \ \|w\|_1 \le R \qquad \min_w \ \tfrac{1}{2}\|z-w\|^2 \quad \text{s.t.} \ \|w\|_2 \le R$$

Figure 1: Geometric intuition of why ℓ_1 constraints give sparse solution, while ℓ_2 constraints do not.

In summary, we can see that the optimal solution $w^\star$ is given by shrinking the magnitude of each component z_i by λ, and then thresholding at 0. If $|z_i| \le \lambda$ for some i, the corresponding $w_i^\star$ will be 0. Therefore, ℓ_1 regularized problem leads to sparse solutions.

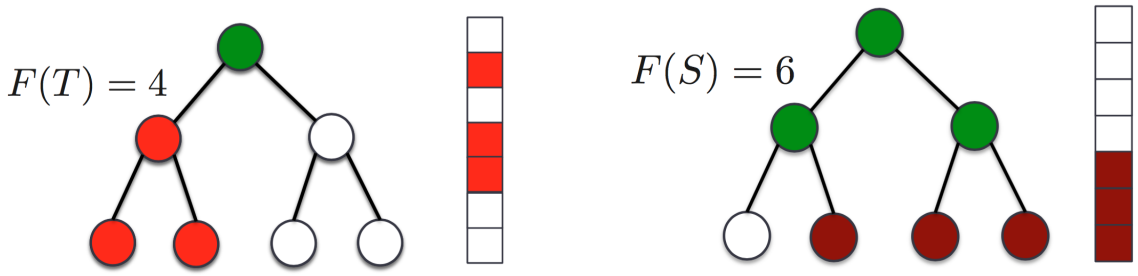
2 Structural Sparsity and Convex Relaxation

2.1 Formulation of Structural Sparsity

In the last lecture, we also mentioned that sometimes we need to enforce structures in the sparsity patterns of our signals. For examples, the coefficients under a wavelet basis of a natural image generally follows a hierarchical structural pattern. In this section, we will talk about how to generalize the technique of relaxation to a general class of structural sparsity problems.

Recall that the original sparsity constraint is formulated with $\|w\|_0 = |\operatorname{supp}(w)|$, which is actually a set function on $\operatorname{supp}(w)$. If we replace the set function with any non-decreasing submodular function $F(\cdot)$, a lot of useful structural sparsity constraints can actually be represented. For example, let

$$F(S) = \sum_{i=1}^{k} \min\{|S \cap G_i|, 1\}$$



(a) Preferred sparsity pattern.

(b) Not-so-preferred sparsity pattern.

Figure 2: Illustration of hierarchical sparsity, by using the submodular function $F(\cdot)$ defined in (3).

then we get the group sparsity, according to partition of variables into $\{G_i\}_{i=1}^k$. Note the cost is zero for group G_i if all the variables in that group is zero, but otherwise is 1 no matter how many variables in G_i comes into play.

Another example is to prefer hierarchical (tree-like) sparsity patterns, by using a set function that satisfies

$$F(T) < F(S), \text{ if } T \text{ is a tree and } S \text{ is not, while } |S| = |T|$$

In particular, we can use

$$F(S) = \left| \bigcup_{s \in S} \text{ancestors}(s) \right| \tag{3}$$

An illustration of this hierarchical sparsity is shown in Figure 2.

2.2 Convex Relaxation of Structural Sparsity

The reason that we chose to use a submodular function $F(\cdot)$ in the formulation of the structural sparsity problems is that when we develop the convex relaxation, the tools we developed in the previous lectures of this class will naturally come into play and help solving the problems.

Recall in the case of ℓ_0 sparsity, we relax the ℓ_0-norm with the ℓ_1-norm, which is its *convex envelope* on $[-1, 1]^d$. See Figure 3 for an illustration.

Definition 1 (Convex Envelope)**.** *A convex envelope of a function $g(x) : \mathcal{D} \to \mathbb{R}$ is the largest convex function $h(x) : \mathcal{D} \to \mathbb{R}$ such that $h(x) \leq g(x)\ \forall x \in \mathcal{D}$.*

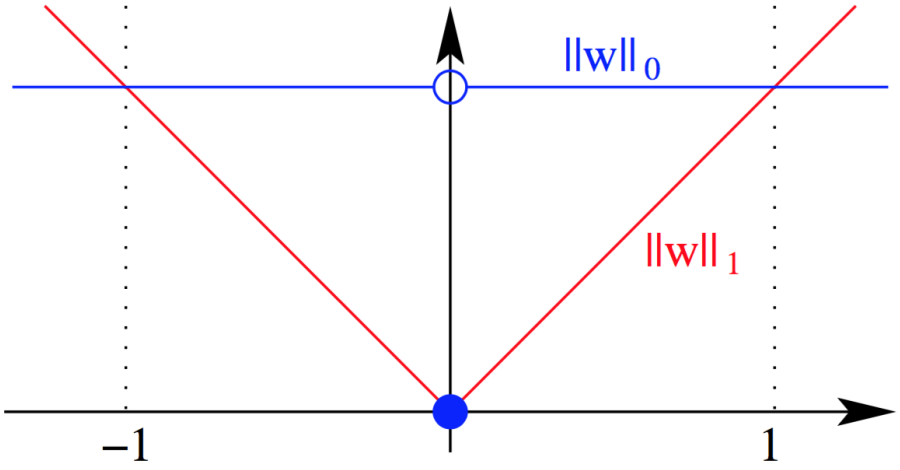


Figure 3: ℓ_1-norm is the convex envelope of the ℓ_0-norm on $[-1,1]^d$.

Note that talking about the convex envelope of ℓ_0-norm on the whole $\mathbb{R}^d$ is not very useful, because it will be the constant-0 function. The choice of the set $[-1,1]^d$ (the unit ℓ_∞ ball) is somewhat arbitrary, and choosing other sets will lead to different convex envelopes. But we will see this choice interplay nicely with submodular function related tools.

The convex envelope of a function $g(x)$ can be computed by taking the conjugate function

$$g^*(y) = \sup_{x \in \mathcal{D}} x^\top y - g(x)$$

and then compute the g^{**}, the conjugate of the conjugate. Note when g itself is convex, $g^{**} = g$.

Proposition 1. *Let F be a non-decreasing submodular function. The convex envelope of $g(w) = F(\text{supp}(w))$ on $w \in [-1,1]^d$ is*

$$\Omega(w) = f(|w|) \tag{4}$$

where $f(\cdot)$ is the Lovász extension of $F(\cdot)$, and $|w|$ is the component-wise absolute value of the vector w.

Proof. The proof is done by directly calculating the conjugate of the conjugate. First of all,

$$g^*(s) = \sup_{w \in [-1,1]^d} s^\top w - g(w) = \sup_{w \in [-1,1]^d} s^\top w - F(\text{supp}(w))$$

Since F only depends on the support of w, we can separate the sup into two parts, one part is about the support, and the other about the magnitude of each component on the

support:

$$g^*(s) = \sup_{\delta\in\{0,1\}^d} \sup_{w\in([-1,1]\setminus\{0\})^d} (\delta \circ w)^\top s - F(\text{supp}(\delta)) = \sup_{\delta\in\{0,1\}^d} \delta^\top |s| - F(\text{supp}(\delta))$$

here $\delta \circ w$ means elementwise multiplication of two vectors. Note

$$\sup_{\delta\in\{0,1\}^d} \delta^\top |s| - F(\text{supp}(\delta)) = \inf_{T\subset\{1,\dots,d\}} \left\{ \tilde{F}(T) = F(T) - \sum_{i\in T} |s_i| \right\} \tag{5}$$

where $\tilde{F}$ is a submodular function because F is submodular and the red part is modular. From the homework, we know that minimizing a submodular function is equivalent to minimizing its Lovász extension on $[0,1]^d$. Meanwhile, it is easy to show that the Lovász extension of the sum of two submodular functions is equal to the sum of their Lovász extensions. And the Lovász extension of a modular function $H(T) = \sum_{i\in T} H_i$ is equal to $h(x) = \sum_i x_i H_i$ for $x \in [0,1]^d$.

Therefore, we can replace the submodular minimization problem in (5) by minimizing the Lovász extension of $\tilde{F}$:

$$g^*(s) = \sup_{\delta\in[0,1]^d} \delta^\top |s| - f(\delta)$$

where $f(\cdot)$ is the Lovász extension of $F(\cdot)$. Now let us compute the conjugate of g^*:

$$g^{**}(w) = \sup_s s^\top w - g^*(s) = \sup_s \inf_{\delta\in[0,1]^d} s^\top w - \delta^\top |s| + f(\delta)$$

Remember $f(\cdot)$ is convex because $F(\cdot)$ is submodular. By verifying the conditions for the saddle point theorem, we can switch the inf and sup:

$$g^{**}(w) = \inf_{\delta\in[0,1]^d} f(\delta) + \sup_s s^\top w - \delta^\top |s|$$

Note that

$$\sup_s s^\top w - \delta^\top |s| = \begin{cases} 0 & \forall i:\ |w_i| \le \delta_i \\ \infty & o.w. \end{cases}$$

Therefore,

$$g^{**}(w) = \inf_{\delta\in[0,1]^d, \delta\ge|w|} f(\delta) = f(|w|)$$

where the last equality comes from the following fact: the Lovász extension is

$$f(x) = \sum_{i=1}^{d} x_{(i)}(F(S_i) - F(S_{i-1}))$$

where $x_{(i)}$ are sorted coordinates such that $x_{(1)} \ge \dots \ge x_{(d)}$, and S_i are corresponding level sets. Note $\emptyset = S_0 \subset S_1 \subset S_2 \subset \cdots \subset S_d = \{1,\dots,d\}$. By assumption, $F(S)$ is

a submodular function defined on the cardinality $|S|$ and is non-decreasing. So $F(S_i) - F(S_{i-1}) \geq 0$. Therefore, the minimizer of $f(x)$ is achieved by taking the minimal possible coordinate for each component x_i. □

Proposition 2. *$\Omega(w) = f(|w|)$ is a norm if $F(\{a\}) > 0$ for all $a \in \mathcal{V}$, where f is the Lovász extension of F.*

Some common examples of this formulation include

- $F(S) = |S|$: $f(|w|) = \|w\|_1$
- $F(S) = \min |S|, 1$: $f(|w|) = \|w\|_\infty$
- $F(S) = \sum_{i=1}^k \min\{|S \cap G_i|, 1\}$: $f(|w|) = \sum_{i=1}^k \|w_{G_i}\|_\infty$

2.3 Geometry of Convex Relaxation

By the definition Lovász extension,

$$\Omega(w) = f(|w|) = \max_{s \in \mathcal{P}_F} s^\top |w| = \max_{|s| \in \mathcal{P}_F} s^\top w \tag{6}$$

where the last equality is because when $F(S) = F(|S|)$ is non-decreasing, all the vertices of $\mathcal{P}_F$ is in the positive orthant. Recall the dual norm is defined by

$$\|u\|_* = \sup_{\|x\| \leq 1} u^\top x$$

So $\Omega(\cdot)$ is actually the dual norm of a norm Ω_* whose unit norm ball is defined by

$$\mathbb{B}_{\Omega_*} = \{s : \Omega_*(s) \leq 1\} = \{s : |s| \in \mathcal{P}_F\} = \{s : \|s_A\|_1 \leq F(A), \forall A \subset \mathcal{V}\} = \left\{s : \max_{A \subset \mathcal{V}, A \neq \emptyset} \frac{\|s_A\|_1}{F(A)} \leq 1\right\}$$

Comparing the left hand side to the right hand side, we get

$$\Omega_*(s) = \max_{A \subset \mathcal{V}, A \neq \emptyset} \frac{\|s_A\|_1}{F(A)} \tag{7}$$

From the unit norm ball of Ω_*, we can also get the unit norm ball of Ω by the fact that the unit dual norm ball is the polar set of the unit norm ball:

$$\mathbb{B}_\Omega = \{y \in \mathbb{R}^d : s^\top y \leq 1, \forall s \in \mathbb{B}_{\Omega_*}\}$$

For our specific case,

$$\mathbb{B}_\Omega = \{w : f(|w|) \leq 1\} = \operatorname{conv}\left\{\frac{1}{F(\operatorname{supp}(w))} w : w \in \{\pm 1, 0\}^d\right\} \tag{8}$$

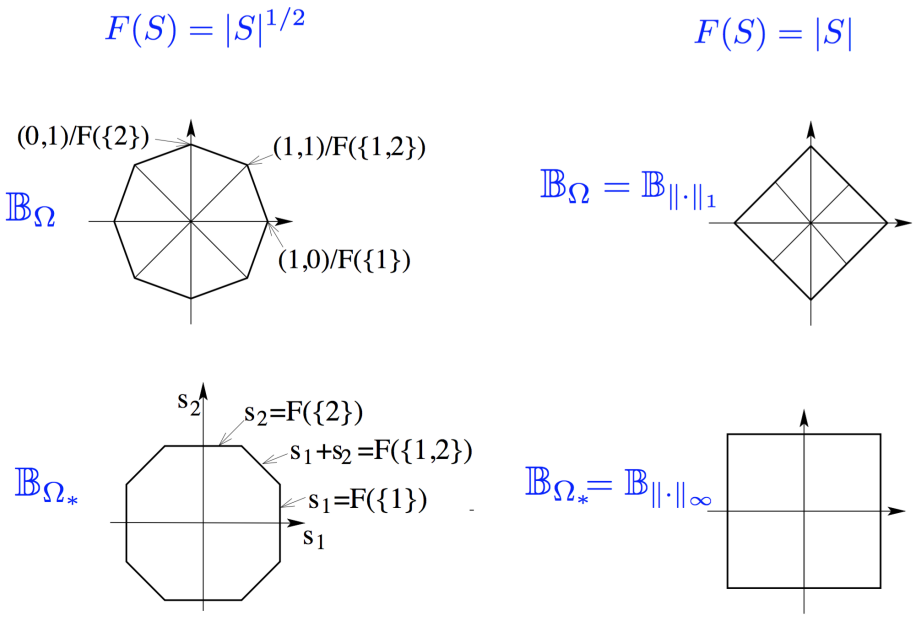


Figure 4: Examples of unit norm balls and unit dual norm balls.

Figure 4 shows some examples of unit norm balls and unit dual norm balls associated with different $F(\cdot)$. Please refer to the slides for more examples.

The sparsity patterns that we could get from those norms can be characterized by the notion of *stable set*.

Definition 2 (Stable Set). *A set $S \subset \mathcal{V}$ is* stable *if $\forall e \notin S$, $F(S \cup \{e\}) > F(S)$. In other words, adding any element to the set will increase the "cost".*

For example, for $F(S) = |S|$, then every set is stable; for $F(S) = \min\{1, |S|\}$, then only $\emptyset$ is stable.

Proposition 3 (Bach, Prop. 5.3). *Assume y has an absolute continuous density w.r.t. Lebesgue measure, and $X^\top X$ is invertible. Then the minimizer*

$$\hat{w} \in \arg\min_w \frac{1}{2}\|Xw - y\|^2 + \lambda f(|w|) \tag{9}$$

is unique and with probability one its support is a stable set.

The key for generalization to high dimensions is decomposability. Please refer to *Bach 2010* for details of results on support recovery and errors.

3 Proximal Gradient Descent

Our formulation of convex relaxed problem can be written in the regularization form as

$$\min_w L(w) + \lambda f(|w|) \tag{10}$$

where $L(w)$ is the loss function, which is usually smooth and convex (e.g. the square loss), while $f(\cdot)$ is the Lovász extension of some submodular function $F(\cdot)$. So $f(|w|)$ is convex, but usually not smooth.

In order to solve this problem, we can use subgradient descent, which gives us an approximation error of $O(1/\sqrt{t})$ for t iterations. Gradient descent gives faster rate, but it cannot be applied here because $f(|w|)$ is non-differentiable. But by using the fact that $L(w)$ is still smooth, we can potentially use the *proximal gradient method* to achieve $1/t$ convergence rate.

The proximal gradient method is a generalization of the projected gradient method. Recall that the projected gradient method is developed to solve a constrained optimization problem. By using the indicator function of the constraint set C, we can write the problem as

$$\min_w L(w) + \delta_C(w) \tag{11}$$

which could be solved by the following iterations:

$$w^{t+1} \leftarrow \Pi_C(w^t - \eta_t \nabla L(w^t)) \tag{12}$$

Now if we replace $\delta_C(w)$ by a general convex function $h(w)$ in (11), we can solve the new problem by modified iterations:

$$w^{t+1} = \mathrm{prox}_{\alpha,h}(w^t - \eta_t \nabla L(w^t)) \tag{13}$$

where prox is the Euclidean proximity operator, defined by

$$\mathrm{prox}_{\alpha,h}(z) = \arg\min_w h(w) + \frac{1}{2\alpha}\|w - z\|^2 \tag{14}$$

Note that if $h(w) = \delta_C(w)$, then the proximity operator is equivalent to the projection operator $\Pi_C(\cdot)$. If the proximity operator is easy to compute for our $h(w)$, then we get a practical algorithm to solve (10). Some familiar special cases include

- Lasso: when $h(w) = \lambda\|w\|_1$, the proximity operator has a closed form solution, because the operator is exactly computing (1), whose solution is completely characterized by (2). The resulting algorithm also has a specific name: ISTA (Iterative Shrinkage-Thresholding Algorithm).

- Submodular penalty: when $h(w) = \lambda f(|w|)$ for f being the Lovász extension of some submodular function. Then computing the proximity operator is equivalent to solving a minimum-norm point problem, similar to the problems in Homework 2 (see also *Bach 2010*).

If we expand the definition of the proximal operator, and let $\alpha = \eta_t$

$$\begin{aligned} w^{t+1} &= \text{prox}_{\eta_t, h}(w^t - \eta_t \nabla L(w^t)) \\ &= \arg\min_w h(w) + \frac{1}{2\eta_t}\|w - (w^t - \eta_t \nabla L(w^t))\|^2 \\ &= \arg\min_w h(w) + L(w^t) + \nabla L(w^t)^\top (w - w^t) + \frac{1}{2\eta_t}\|w - w^t\|^2 \end{aligned}$$

where in the last equality we removed the constant $\|\nabla L(w^t)\|^2$ and add the constant $L(w^t)$ since those constants does not change the optimum.

We can interpret this as solving an approximation to the original problem by locally approximating the part $L(w)$ by a linear function. Note only $L(w)$ is approximated, while $h(w)$ is untouched, intuitively because we know $L(w)$ is smooth, so the local linear approximation is "nicer". The extra quadratic term $\|w - w^t\|^2$ is added to prevent us from going too far away from w^t, which is the starting point of the local linear approximation.

Note w^* is a stationary point of the proximal iteration if and only if

$$w^* = \text{prox}_{\eta_t, h}(w^* - \eta_t \nabla L(w^*))$$

which is equivalent to

$$0 \in \partial h(w^*) + \frac{1}{2\eta_t} \nabla \left(\|w - (w^* - \eta_t \nabla L(w^*))\|^2 \right) |_{w=w^*}$$

Plugging in the gradient, we get

$$0 \in \partial h(w^*) + \nabla L(w^*)$$

which is exactly the optimality condition of the original problem (10). The convergence of the proximal gradient descent algorithm will be covered in the next lecture.