6.883 Learning with Combinatorial Structure
Notes for Lecture 13
Author: Tuhin Sarkar

1 Convergence of Proximal Gradient Descent

In the last class, we talked about the Proximal Gradient descent method used to minimize the following regularized function

$$L(w) + \lambda\, h(w) \tag{1}$$

where $L(w)$ is the loss function, which is smooth and convex (*e.g.*: squared loss function), while $h(\cdot)$ is a convex function that may not be smooth (*e.g.*: Lovász extension of a submodular function $F(\cdot)$). In this section we will prove the convergence of the proximal gradient algorithm.

1.1 Proof Idea

To prove convergence, we basically modify the proof of the Gradient descent algorithm. First, let $\phi(w) = L(w) \; + \; h(w)$. The proof proceeds in three steps:

- We begin by bounding the progress in one iteration: $L(w^{t+1}) \leq L(w^t) + \ldots$
- Using convexity, we show that

$$\phi(w^{t+1}) - \phi^* \leq \frac{1}{2\alpha}\big(||w^t - w^*||^2 - ||w^{t+1} - w^*||^2\big)$$

- Finally, summing the above over all t (telescoping sum), we can show that

$$\big(T+1\big)\big[\phi(w^{T+1}) - \phi(w^*)\big] \leq \frac{1}{2\alpha}\big(||w^1 - w^*||^2\big)$$

1.2 Convergence Analysis

Recall that in proximal gradient descent, we have

$$w^{t+1} = \text{prox}_{\alpha,h}\big(w^t - \alpha_t \nabla L(w^t)\big) \tag{2}$$

Next, we define $G_\alpha(w)$, which is a gradient like object at iteration t,

$$G_\alpha(w) \triangleq \frac{1}{\alpha}\big(w - \text{prox}_{\alpha,h}\big(w - \alpha \nabla L(w)\big)\big)$$

At every iteration, $w^{t+1} = w^t - \alpha_t G_{\alpha_t}(w^t)$. Further at the optimum, $G_\alpha(w^t) = 0$.

We additionally require that the gradient of $L(\cdot)$ is Lipschitz continuous, *i.e.*, $||\nabla L(y) - \nabla L(x)|| \leq M||y - x|| \; \forall \; x, y$.

Lemma 1 (Descent Lemma)**.**

$$L(y) \leq L(w) + \big\langle \nabla L(w), y - w \big\rangle + \frac{M}{2}||y - w||^2 \tag{3}$$

Lemma 1 $\implies$ $L(w^{t+1}) \leq L(w^t) + \big\langle \nabla L(w^t), w^{t+1} - w^t \big\rangle + \frac{M}{2}||w^{t+1} - w^t||^2$.

Lemma 2 (Central Lemma)**.** *Let $\phi(w) = L(w) + h(w), 0 \leq \alpha \leq 1/M, w^{t+1} = w^t - \alpha G_\alpha(w^t)$. Then, for any z:*

$$\phi(w^{t+1}) \leq \phi(z) + \big\langle G_\alpha(w^t), w^t - z \big\rangle - \frac{\alpha}{2}||G_\alpha(w)||^2 \tag{4}$$

Proof. By convexity of $L(\cdot)$ we have, $L(z) \geq L(w) + \big\langle \nabla L(w), z - w \big\rangle$. For simpicity, $w = w^t, w' = w^{t+1}$. Combining this with Descent Lemma we get

$$L(w') \leq L(z) - \big\langle \nabla L(w), z - w \big\rangle + \big\langle \nabla L(w), w' - w \big\rangle + \frac{M}{2}||w' - w||^2 \tag{5}$$

Since $h(\cdot)$ is a convex function we have the following inequality for any $g \in \partial h(w')$

$$h(z) \geq h(w') + \langle g, z - w' \rangle$$

Recall that $w^{t+1} = \text{prox}_{\alpha,h}\big(w^t - \alpha_t \nabla L(w^t)\big) = \arg\min_w h(w) + \frac{1}{2\alpha_t}||w - (w^t - \alpha_t \nabla L(w^t))||^2$. Then, by optimality conditions we have

$$\begin{aligned} 0 &\in \partial h(w') + \frac{1}{\alpha_t}\big(w' - w + \alpha_t \nabla L(w)\big) \\ &\implies \frac{1}{\alpha_t}\big(w - w' - \alpha_t \nabla L(w)\big) \in \partial h(w') \\ &\implies G_\alpha(w) - \nabla L(w) = g \in \partial h(w') \end{aligned}$$

here $\alpha_t = \alpha$ (constant step size). Then we have from Eqn. (5)

$$L(w') + h(w') \leq L(z) + h(z) + \langle \nabla L(w) + G_\alpha(w) - \nabla L(w), w' - z \rangle + \frac{M}{2}||w' - w||^2 \quad (6)$$

Setting $w' = w - \alpha G_\alpha(w)$ in Eqn. (6) and noting that $M \leq 1/\alpha$, we have

$$\phi(w') \leq \phi(z) + \langle G_\alpha(w), w - \alpha G_\alpha(w) - z \rangle + \frac{M}{2}||w' - w||^2$$

$$\phi(w') \leq \phi(z) + \langle G_\alpha(w), w - z \rangle - \frac{1}{2\alpha}||w' - w||^2 \quad (7)$$

Since $\|w' - w\| = \|\alpha G_\alpha(w)\|$, this proves the lemma. □

Next, we use Lemma 2. By setting $z = w$, we get $\phi(w') \leq \phi(w) - \frac{1}{2\alpha}||w - w'||^2$. This means we make progress in each iteration.

Next we set $z = w_{\text{opt}}$, then Eqn. (7) gives us

$$\begin{aligned}
\phi(w^{t+1}) - \phi(w_{\text{opt}}) &\leq \langle G_\alpha(w^t), w^t - w_{\text{opt}} \rangle - \frac{1}{2\alpha}||w^{t+1} - w^t||^2 \\
&= \frac{1}{2\alpha}\big(\langle 2\alpha G_\alpha(w^t), w^t - w_{\text{opt}} \rangle - \|\alpha G_\alpha\|^2\big) \\
&= \frac{1}{2\alpha}\big(||w^t - w_{\text{opt}}||^2 - \|\alpha G_\alpha - w^t + w_{\text{opt}}\|^2\big) \\
&= \frac{1}{2\alpha}\big(||w^t - w_{\text{opt}}||^2 - ||w^{t+1} - w_{\text{opt}}||^2\big) \quad (8)
\end{aligned}$$

Since $\phi(w^{t+1}) \leq \phi(w^t)$, summing up Eqn. (8) $\forall\ t = 1, 2, \ldots, T$ we get the following relation

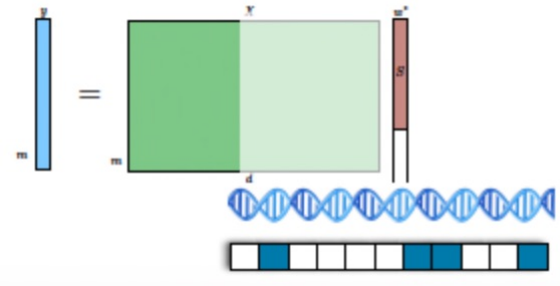
$$\begin{aligned}
(T+1)(\phi(w^{T+1}) - \phi(w_{\text{opt}})) &\leq \frac{1}{2\alpha}\big(||w^0 - w_{\text{opt}}||^2 - ||w^{T+1} - w_{\text{opt}}||^2\big) \\
&\leq \frac{1}{2\alpha}||w^0 - w_{\text{opt}}||^2 \quad (9)
\end{aligned}$$

Eqn. (9) shows that the convergence is $O(1/T)$. We can improve convergence rates to $O(1/T^2)$ by averaging cleverly: set $z^1 = w^0$, and iterate

$$\begin{aligned}
w^t &= \text{prox}_{\alpha_t, h}\big(z^t - \alpha_t \nabla L(z^t)\big) \\
\beta_{t+1} &= \tfrac{1}{2}(1 + \sqrt{1 + 4\beta_t^2}) \\
z^{t+1} &= w^t + \frac{\beta_t - 1}{\beta_{t+1}}(w^t - w^{t-1})
\end{aligned}$$

Then we have $\phi(w^T) - \phi_{\text{opt}} \leq \frac{2M}{(T+1)^2}||w^1 - w_{\text{opt}}||^2$. (See Nesterov 1983, 2004; Beck & Teboulle 2009, Tseng 2008 for details).

2 Structured Sparsity



Previously, we had used regularization to enforce sparsity constraints on our optimization problems.

$$\min_{w} L_N(w) + \lambda\Omega(w)$$

$\Omega(w)$ could be the $||w||_0, ||w||_1$ functions. However, we can further generalize this notion of regularization by

- Penalty on support via submodular functions: $F\big(\text{supp}(w)\big)$ relaxed to $f(|w|)$
- assuming sparse combinations of atoms

In the second scenario, w_{opt} is a combination of few atoms from a collection $\mathcal{A}$, *i.e.*, $w_{\text{opt}} = \sum_{i=1}^{k} \gamma_i a_i$ here $\gamma_i \geq 0$ and $a_i \in \mathcal{A}$. Some examples are

- Sparse vectors: $\mathcal{A} = \{\pm e_i\}$ = set of canonical basis vectors $\implies$ w_{opt} k-sparse.
- Low rank matrices: $\mathcal{A} = \{uv^T \,|\, ||u||_2 \leq 1, ||v||_2 \leq 1\}$ $\implies$ w^* rank k.

2.1 Atomic Norms

In the example of sparse vectors, the ℓ_1 ball is the convex hull conv($\mathcal{A}$). Using this analogy, we generalize to *atomic norms* based on the atomic sets. We will need the following assumptions:

- No element $a \in \mathcal{A}$ lies in the convex hull of another element in the atomic set $\mathcal{A}$ (every a is an extreme point);
- $\mathcal{A}$ is centrally symmetric about the origin.

We now define the **gauge function** as

$$\begin{aligned} ||w||_{\mathcal{A}} &= \inf\{t > 0 | w \in t\,\text{conv}(\mathcal{A})\} \\ &= \inf\left\{ \sum_{a\in\mathcal{A}} \gamma_a | w = \sum_{a\in\mathcal{A}} \gamma_a a, \gamma_a \geq 0 \,\forall\, a \in \mathcal{A} \right\} \end{aligned}$$

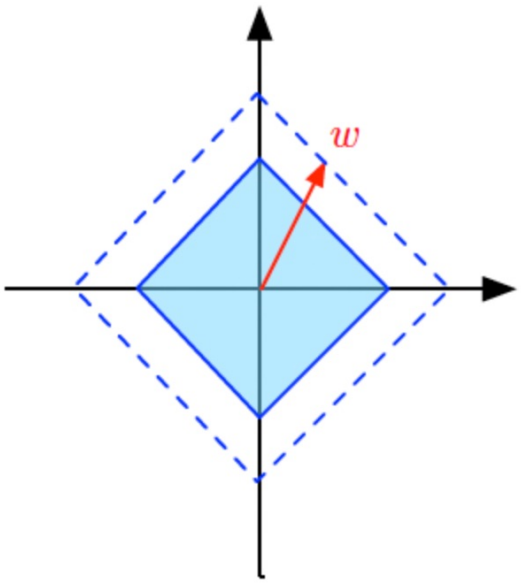


Figure 1: Gauge function

Under the above assumptions, the gauge function is a norm and is here called the atomic norm. Examples: $\mathcal{A} = \{\pm e_i\} \implies ||w||_{\mathcal{A}} = ||w||_1$ and $\mathcal{A} = \{-1, 1\}^d \implies ||w||_{\mathcal{A}} = ||w||_\infty$ because here $\text{conv}(\mathcal{A}) = \{a \,|\, ||a||_\infty \leq 1\}$. We will now give examples of some atomic norms.

- $\mathcal{A} = \{uv^T \,|\, ||u||_2 \leq 1, ||v||_2 \leq 1\}$, then $||w||_{\mathcal{A}} = \inf\left\{ \sum_{a\in\mathcal{A}} \gamma_a | w = \sum_{a\in\mathcal{A}} \gamma_a a, \gamma_a \geq 0 \,\forall\, a \in \mathcal{A} \right\} = \sum_i |\sigma_i(W)|$ (follows from the existence of SVD for any matrix W). $||w||_{\mathcal{A}}$ is also the nuclear norm. Such norm constraints are used to find low rank solutions, and have applications in, *e.g.*, predicting user rating of different movies given a set of preferences.
- Recovering a ranking of m elements: Here, we take $\mathcal{A}$ to be the collection of all $m \times m$ permutation matrices (rankings can be expressed as permutations). Then $\text{conv}(\mathcal{A})$ is the *Birkhoff polytope*, the set of all doubly stochastic matrices (see also Homework 1, Problem 1).

1	5	0	*
*	*	4	5
5	2	*	4
1	*	5	3

Figure 2: User Rating in Netflix

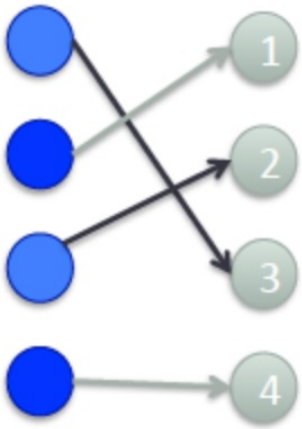


Figure 3: Ranking

2.2 Recovery in the noiseless case

Let $\Delta = \widehat{w} - w^*$, and w^* is the true underlying signal ($y = Xw^*$.). Recall the restricted nullspace condition for the ℓ_1-norm. We have

$$\text{null}(X) = \{\Delta \,|\, X\Delta = 0\}$$
$$\text{cone of descent directions of } \|\cdot\|_1 \text{ at } w^* = \{\Delta \,|\, ||\widehat{w}||_1 \leq ||w^*||_1\}$$

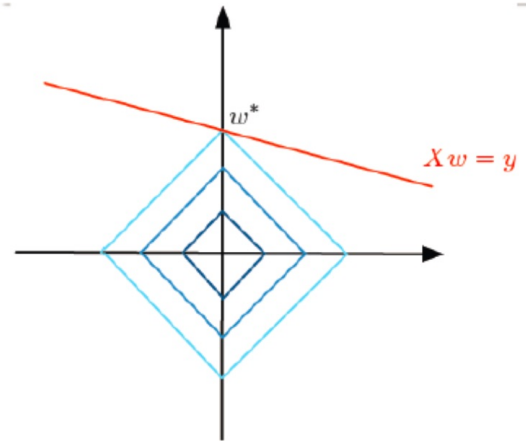


Figure 4: Restricted nullspace condition. The descent directions are the directions into the polytope (starting from w^*).

The restricted nullspace condition demands that null$(X) \cap$ cone of descent directions $= \{0\}$. This ensures that w^* is the point with minimum $\|w\|_1$ that satisfies $Xw = y$, any point in a descent direction would not satisfy the constraint. In other words, w^* is the optimal solution to the optimization problem of minimizing $\|w\|_1$ such that $Xw = y$.

We generalize this to atomic norms by using the cone of descent directions for atomic norms:

$$\mathcal{T}_{\mathcal{A}}(w^*) = \{z - w^* \,|\, ||z||_{\mathcal{A}} \leq ||w^*||_{\mathcal{A}}\}$$

$\mathcal{T}_{\mathcal{A}}(w^*)$ is the *tangent cone* of the convex hull $\text{conv}(\mathcal{A})$ at w^*, and is the set of descent directions. Our restricted null space condition is then $\mathcal{T}_{\mathcal{A}}(w^*) \cap \text{null}(X) = \{0\}$.

2.3 Recovery in noisy case

Assume that we have

1. $y = Xw + \epsilon$
2. $||\epsilon|| \leq \delta$
3. $||X\nu|| \geq \alpha||\nu|| \quad \forall\, \nu \in \mathcal{T}_{\mathcal{A}}(w^*)$

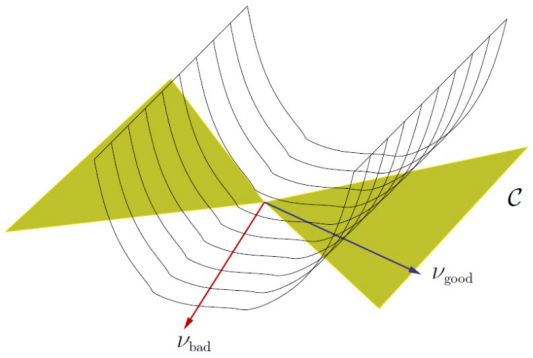


Figure 5: The loss function (squared error) does not have curvature in all directions. The redularization helps rule out the "bad" (flat) directions.

Then we solve for

$$\widehat{w} = \arg\min_{w} ||w||_{\mathcal{A}} \quad \text{s.t. } ||Xw - y|| \leq \delta. \tag{10}$$

Under the assumptions above we have

$$\begin{aligned} ||X(\widehat{w} - w^*)|| &\leq ||X(\widehat{w} - y)|| + ||X(w^* - y)|| \\ &\leq \delta + \delta = 2\,\delta \end{aligned}$$

Further, by assumption item 3, we have $||X(\widehat{w} - w^*)|| \geq \alpha\,||\widehat{w} - w^*||$, then we get $||w^* - \widehat{w}|| \leq 2\,\delta/\alpha$. The problem here is to ensure Assumption 3 holds.

To that effect we will study how this condition can be met in certain random matrices. Assume $X \in \mathbb{R}^{d\times N}$ has iid Gaussian entries, $x_{ij} \sim \mathcal{N}(0, 1/N)$. We will need the set $\mathcal{S} = \mathcal{T}_{\mathcal{A}}(w^*) \cap \mathbb{S}^{d-1}$. Next we define the **gaussian width** of a set $S \subset \mathbb{R}^d$.

Gaussian Width: The gaussian width of some set $S \subset \mathbb{R}^d$ is defined as

$$w(S) = \mathbb{E}_{\mathbf{g}\sim\mathcal{N}(0,I)}\left[\sup_{z\in S} \mathbf{g}^T z\right]$$

The Gaussian width is a measure of the width of the cone of descent directions; the larger the width, larger the cone.

The following results are drawn from (Chandrasekaran et al, *The Convex Geometry of Linear Inverse Problems*, 2012).

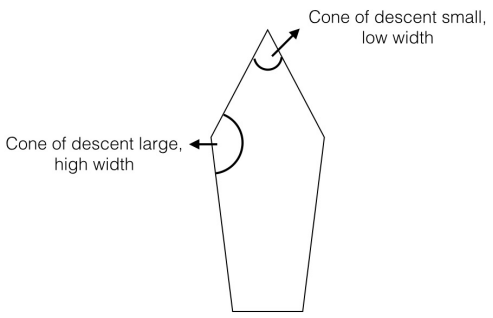


Figure 6: Gaussian Width for a convex polytope

Theorem 1. $\mathbb{E}\left[\sqrt{N}\,\min_{\nu\in\mathcal{S}}||X\nu||\right] \geq \lambda_N - w(\mathcal{S})$. *Here λ_N is the expected length of a $N-$ dimensional Gaussian vector.*

Corollary 1. *If*

$$N \geq \frac{w(\mathcal{S})^2 + 3/2}{(1-\alpha)^2}$$

Then we have

$$\mathbb{P}\big(||X\nu|| \geq \alpha||\nu||,\, \nu \in \mathcal{S}\big) \geq 1 - \exp\left(-\frac{1}{2}\left[\lambda_N - w(\mathcal{S}) - \sqrt{N}\alpha\right]\right)$$

The results show that the wider the tangent cone (cone of descent direction), the more samples (larger N) we need to guarantee a small error with a certain confidence. In a sharp cone, there are only very few descent directions, and the nullspace condition is less strict and hence, we can do with fewer samples or measurements.

3 Atomic Norm and Lovász extension

We talked about two generalizations for structured sparsity: (1) via submodular functions and the Lovász extension of the absolute value, $f(|w|) = \sup\{\langle w, y\rangle \,|\, |y| \in \mathcal{P}_F\}$, and (2) via sets of atoms and atomic norms $||w||_{\mathcal{A}} = \inf\{t > 0 \,|\, w \in t\,\mathrm{conv}(\mathcal{A})\}$. Are these related?

The relation is via dual norms. While the Lovász extension version creates a norm via a support function, the atomic norm is a gauge function. Support functions and gauge functions are related via dual norms: Let $||\cdot||_*$ be the dual norm of $||\cdot||_{\mathcal{A}}$, then

$$\begin{aligned} ||u||_* &= \sup\{\langle w, u\rangle \,|\, ||w||_{\mathcal{A}} \leq 1\} \\ &= \sup\{\langle w, u\rangle \,|\, w \in \mathrm{conv}(\mathcal{A})\}. \end{aligned}$$

The last line is the support function of $\text{conv}(\mathcal{A})$.

(In the special case that $\text{conv}(\mathcal{A})$ is a symmetrized version of the submodular polytope $\mathcal{P}_F$, then the Lovász norm is the dual norm of the atomic norm.)

4 Optimization

To use atmomic norms in practice, we will need to solve

$$\min_w \frac{1}{2}||Xw - y||^2 \ + \ \lambda \, ||w||_{\mathcal{A}}.$$

One idea is to use the proximal gradient method to solve this, where we need to be able to compute the proximity operator. This operator corresponds to solving

$$\text{prox}_{\|\cdot\|_{\mathcal{A}}}(z) = \arg\min_w \frac{1}{2\alpha_t}||w - z||^2 \ + \ \lambda||w||_{\mathcal{A}}.$$

This is not always simple, except for special cases. The paper by (Chandrasekaran et al) discusses this in a bit more detail.

However, there is another way of viewing this problem

$$\min_w \frac{1}{2\alpha_t}||w - z||^2 \quad ||w||_{\mathcal{A}} \leq R \tag{11}$$

Eqn. (11) is equivalent to minimizing a differentiable convex function over a polytope. We will talk about Conditional Gradient/Frank-Wolfe algorithm to solve such problems in the next lecture.