Lecture 12

Last time

- ▷ Forward - Backward method
- ▷ Examples
- ▷ Constraints via proximal operator
- ▷ Analysis

Today

- ▷ Finish analysis
- ▷ Guarantees for strongly convex
- ▷ Accelerated Forward Backward Method.
- ▷ More proximal methods
- ▷ Alternating Projections

Theorem For any convex, L-smooth f and convex h such that $x^* \in \operatorname{argmin}(f+h)(x)$. Then, the iterates of FBM with $\alpha_k = 1/L$ satisfies

$$(f+h)(x_{k+1}) - \min(f+h) \leq \frac{L\|x_0 - x^*\|^2}{2k}.$$

Proof: We start by proving

$$(\ast\ast) \quad 0 \leq (f+h)(x_{k+1}) - \min(f+h) \leq \frac{L}{2}\left(\|x_k - x^*\|^2 - \|x_{k+1} - x^*\|^2\right)$$

By definition x_{k+1} minimizes

$$\varphi_k(x) = \underbrace{f(x_k) + \nabla f(x_k)^T(x - x_k) + h(x)}_{\text{0-strongly convex}} + \underbrace{\frac{L}{2}\|x - x_k\|^2}_{\text{L-strongly convex!}}$$

By HW 2 P2:

$$(1) \quad \varphi_k(x_{k+1}) + \frac{L}{2}\|x^* - x_{k+1}\|^2 \leq \varphi_k(x^*)$$

Using the characterization of L-smooth convex functions

(2) $$(f+h)(x_{k+1}) \leq \varphi_k(x_{k+1})$$

Using the convexity f

(3) $$\varphi_k(x^*) \leq \underbrace{f(x^*) + h(x^*)}_{\min(f+h)} + \frac{L}{2}\|x^* - x_k\|^2$$

Then

$$\begin{aligned}(f+h)(x_{k+1}) - \min(f+h) &\overset{(2)}{\leq} \varphi_k(x_{k+1}) - \min(f+h) \\ &\overset{(1)}{\leq} \varphi_k(x^*) - \frac{L}{2}\|x^* - x_{k+1}\|^2 \\ &\quad - \min(f+h) \\ &\overset{(3)}{\leq} \frac{L}{2}\left(\|x^* - x_k\|^2 - \|x^* - x_{k+1}\|^2\right),\end{aligned}$$

which establishes $(\underline{**})$.

$\square$

Again convergence should speed up under quadratic growth.

Theorem If in addition, we suppose that $f+h$ is μ strongly convex. Then

$$(f+h)(x_{k+1}) - \min f+h \leq \frac{1}{2}\left((f+h)(x_0) - \min f+h\right)$$

for $k > \lceil 2L/\mu \rceil$.

Proof: By the previous theorem

$$(f+h)(x_{k+1}) - \min f+h \le \frac{L}{2k}\|x_0 - x^*\|^2$$

HW 2 quadratic growth $\rightarrow$ $$\le \frac{L}{\mu k}\left((f+h)(x_0) - \min f+h\right)$$

$k > \frac{2L}{\mu}$ $\rightarrow$ $$\le \frac{1}{2}\left((f+h)(x_0) - \min f+h\right).$$

$\dashv$

Thus we achieve accuracy ε after

$$\log_2\left(\frac{f+h(x_0) - \min f+h}{\varepsilon}\right) \text{ iterations.}$$

Accelaration

We consider the algorithm that starts at $y_0 = x_0$ an $\lambda_0 = 0$, and updates

$$y_{k+1} = \text{prox}_{\alpha h}(x_k - \alpha\nabla f(x_k))$$

$$x_{k+1} = y_{k+1} + \frac{(\lambda_k - 1)}{\lambda_{k+1}}(y_{k+1} - y_k)$$

$$\lambda_{k+1} = \frac{1 + \sqrt{1+4\lambda_k^2}}{2}.$$

This algorithm goes by different names:

- Accelerated/Fast Proximal/Projected Gradient Method
- FISTA.

Just as before it exhibits faster convergence.

Theorem: For any convex f with L-Lipschitz gradient and convex h. The iterates of AFBM with $\alpha = 1/L$ satisfy

$$(f+h)(y_k) - \min f+h \leq \frac{2L \, \|x_0 - x^*\|^2}{(k+1)^2}.$$

Proof: Details are very similar to the proof for AGD (see Beck's book Theorem 10.34). $\square$

More proximal methods

A natural question is what happens when we have

$$\min \; f(x) + g(x)$$

$f, g : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$, and none of the two is smooth. Maybe the proximal operator of both f and g is easy to compute

Examples

- Intersection of two sets

Find $x \in C_1 \cap C_2$ $\equiv$ $\min \iota_{C_1}(x) + \iota_{C_2}(x)$.

convex and closed

Then the two prox are projection

- Compressed sensing

Prox is projection

$$\min \|x\|_1 \quad \text{s.t.} \quad Ax = b \quad \equiv \quad \min \|x\|_1 + \iota_{\{Ax=b\}}(x)$$

Prox is easy

There are a number of methods to tackle these problems:

- Alternating Projections (Example 1)
- Alternating Direction Method of Multipliers (ADMM)
- Primal - Dual Hybrid Gradient (PDHG)

Example 2.

In order to understand the ideas behind

ADMM and PDLP we need more convex analysis, so these algorithms will be covered in Nonlinear 2.

Alternating Projections

Assume we want to solve

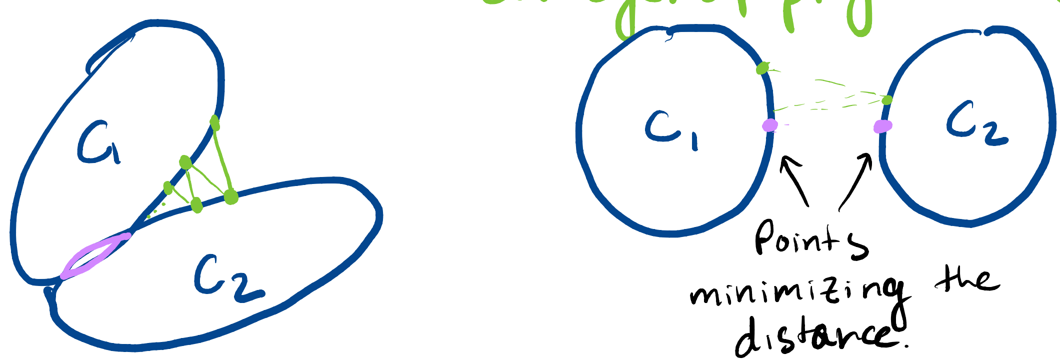
$$\min \|x-y\| \text{ s.t. } x \in C_1,\ y \in C_2.$$

The alternating projections method was originally proposed by John von Neumann. It updates as follows

$$x_{k+1} \leftarrow P_{C_1} P_{C_2}(x_k)$$

orthogonal projection.

Intuition



Another perspective to analyze iterated algorithms based on proximal mappings is through the lens of a fixed-point iteration.

Def: Given an operator $F: \mathbb{R}^d \to \mathbb{R}^d$, a fixed-point iteration updates

$$x_{k+1} \leftarrow F(x_k).$$

The goal of this iteration is to find a fixed point $x^* = F(x^*)$.

Proposition: The following two are equivalent

- x^* is a fixed point of $\underbrace{P_{C_1} P_{C_2}}_{F}$.
- $(x^*, P_{C_1} x^*)$ is a solution of

$$\min_{\substack{x \in C_2 \\ y \in C_1}} \frac{1}{2}\|x - y\|^2.$$

Proof Consider

$$\min \underbrace{\frac{1}{2}\|x-y\|^2 + \iota_{C_1}(x) + \iota_{C_2}(y)}_{f(x,y)}$$

Then (x^*, y^*) is a solution iff

$$\begin{bmatrix} 0 \\ 0 \end{bmatrix} \in \begin{bmatrix} x^* - y^* \\ y^* - x^* \end{bmatrix} + \begin{bmatrix} \partial \iota_{C_1}(x^*) \\ \partial \iota_{C_2}(y^*) \end{bmatrix}$$

Thus $y^* - x^* \in \partial \iota_{C_1}(x^*) \Leftrightarrow \mathrm{proj}_{C_1}(y^*) = x^*$

$x^* - y^* \in \partial \iota_{C_2}(y^*) \Leftrightarrow \mathrm{proj}_{C_2}(x^*) = y^*$. $\square$

Fact 1: $R = 2F - I$ is 1-Lipschitz. $\dashv$

Check!

Fact 2: For all $a, b \in \mathbb{R}^d$

$$\|\tfrac{1}{2} a + \tfrac{1}{2} b\|^2 = \tfrac{1}{2}\|a\|^2 + \tfrac{1}{2}\|b\|^2 - \tfrac{1}{4}\|a-b\|^2. \dashv$$

Theorem The iterates of AP satisfy

$$\frac{1}{T}\sum_{k=0}^{T-1} \|x_k - F(x_k)\|^2 \leq \frac{\|x_0 - x^*\|^2}{T}. \dashv$$

Proof: Rewrite $F = \frac{R}{2} + \frac{I}{2}$, then

$$\|x_{k+1} - x^*\|^2 = \|\tfrac{1}{2}(x_k - x^*) + \tfrac{1}{2}(R(x_k) - R(x^*))\|^2$$

(*)
$$= \frac{1}{2}\|x_k - x^*\|^2 + \frac{1}{2}\|R(x_k) - R(x^*)\|^2 - \frac{1}{4}\|x_k - x^* - R(x_k) + R(x^*)\|^2$$

$$\leq \|x_k - x^*\|^2 - \frac{1}{4}\|x_k - R(x_k)\|^2.$$

Reordering and summing up first $T-1$ ineq

$$\frac{1}{4T}\sum \underbrace{\|x_k - R(x_k)\|^2}_{= \|2(x_k - F(x_k))\|^2} \leq \frac{1}{T}\left(\|x_0 - x^*\|^2 - \|x_T - x^*\|^2\right)$$

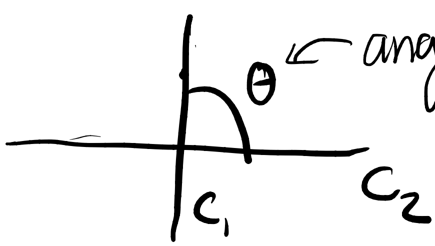
$$= \frac{1}{T}\|x_0 - x^*\|^2$$

$\square$

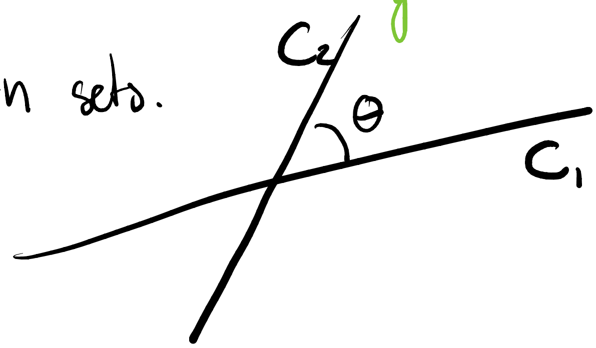
Corollary: The iterates converge to a fixed point x^*.

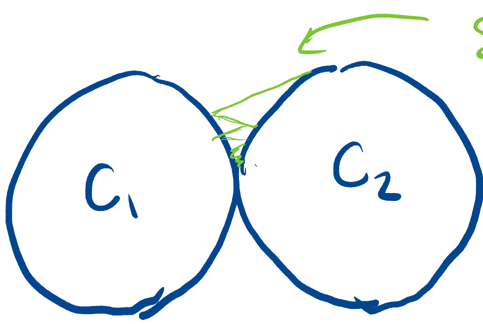
Proof: Let $S = \{x \mid F(x) = x\}$. By the previous Theorem, the x_k's are bounded. Thus, there is some accumulation point x^*. By the previous Theorem $\|x_k - F(x_k)\| \to 0$, by continuity $x^* = F(x^*)$. Moreover by $(*)$ $x_k \to x^*$. $\square$

More generally one can prove that the convergence depends on transversality.



← angle between sets.





Slow convergence due to lack of transversality.