Convex Optimisation

Matthew Tam
(University of Melbourne)
matthew.tam@unimelb.edu.au

ACE Network
S1 2021

---

* Assignment 4
  - online after lecture today
  - due June 3rd @ 5pm

Last few weeks

Frank-Wolfe

differentiable

$$\min_{x \in C} f(x)$$

C ← compact

Proximal gradient descent

proper, lsc, convex

$$\min_{x \in \mathbb{R}^n} f(x) + g(x)$$

f: differentiable

- FISTA = accelerated version
- activated f via Df
- activated g via $\text{prox}_{\lambda g}$

$$x_{k+1} = \text{prox}_{\lambda g}(x_k - \lambda Df(x_k)) \quad \forall k \in \mathbb{N}.$$

3.4 Subgradient method

In this section, we consider the problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

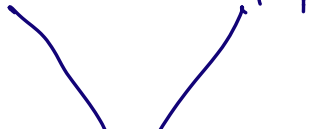
where $f: \mathbb{R}^n \to \mathbb{R}$ is a convex function, but not necessarily differentiable.

If f were differentiable, then we could apply the method from Section 3.2 with $g = 0$ to obtain algorithm

$$x_{k+1} = x_k - \lambda \nabla f(x_k).$$

replace this with $\partial f(x_k)$.

Since our function is not differentiable, we will replace the gradient with subgradients in the above iteration.



Example 3.4.1

Consider $\min_{x \in \mathbb{R}} f(x)$ where $f(x) = |x|$ which has solution $x^* = 0$. Then

$$\partial f(x) = \begin{cases} \{+1\} & x > 0 \\ \{-1\} & x < 0 \\ [-1, +1] & x = 0. \end{cases}$$

Let $\lambda > 0$ and $x_0 \in \left(\frac{1}{3}\lambda, \frac{2}{3}\lambda\right)$. Consider the iteration

$$x_{k+1} = x_k - \lambda \phi_k \quad \text{where} \quad \phi_k \in \partial f(x_k)$$

Then

$$x_1 = \begin{cases} x_0 - \lambda(+1) < \frac{2}{3}\lambda - \lambda = -\frac{1}{3}\lambda \\ x_0 - \lambda(+1) > \frac{1}{3}\lambda - \lambda = -\frac{2}{3}\lambda \end{cases}$$

$$\Rightarrow \quad x_1 \in \left(-\frac{2}{3}\lambda, -\frac{1}{3}\lambda\right).$$

By symmetry, it follows that $x_2 \in \left(\frac{1}{3}\lambda, \frac{2}{3}\lambda\right)$.

Thus, $|x_k| > \frac{1}{3}\lambda \quad \forall k \in \mathbb{N}$. and

So (x_k) cannot converge to $x^* = 0$. □

Given an initial $x_0 \in \mathbb{R}^n$, we consider the subgradient method which computes (x_k) according to

$$x_{k+1} = x_k - \lambda_k \cdot \frac{\phi_k}{\|\phi_k\|} \qquad (3.18)$$

where $\phi_k \in \partial f(x_k)$.

where $(\lambda_k) \subseteq \mathbb{R}_{++}$ is a specially chosen sequence of stepsizes. In fact, we will require that

$$\lim_{k\to\infty} \lambda_k = 0 \quad \text{and} \quad \sum_{k=0}^{\infty} \lambda_k = +\infty. \qquad (3.19)$$

This is satisfied, for instance, by

$$\lambda_k = 1/\sqrt{k} \quad \forall k \in \mathbb{N}.$$

actually not

actually not needed.

Theorem 3.4.2

Let $f: \mathbb{R}^n \to \mathbb{R}$ be a continuous convex function, and let $x^* \in \operatorname{argmin}_{x \in \mathbb{R}^n} f(x)$. Let (x_k) be the sequence given by (3.18) with stepsizes given by (3.19). Then

$$\min_{0 \le i \le k} f(x_i) \to f(x^*) \quad \text{as } k \to \infty.$$

Proof

Suppose by way of a contradiction, that there exists $\tilde{f} \in \mathbb{R}$ such that

$$f(x_k) \geq \tilde{f} > f(x^*) \qquad \forall k \in \mathbb{N}.$$

By continuity of f, there exists $\rho > 0$ such that $f(x) \geq \tilde{f}$ for all x such that $\|x - x^*\| \leq \rho$. In particular, for

$$y_k := x^* + \rho \frac{\phi_k}{\|\phi_k\|},$$

we have $f(y_k) \le \tilde{f}$ since

$$\|y_k - x^*\| = \left\| \rho \cdot \frac{\phi_k}{\|\phi_k\|} \right\| = \rho.$$

Since $\phi_k \in \partial f(x_k)$, we have.

$$\begin{aligned} f(y_k) &\ge f(x_k) + \langle \phi_k, y_k - x_k \rangle \\ &\ge \tilde{f} + \langle \phi_k, x^* - x_k + y_k - x^* \rangle \\ &= \tilde{f} + \langle \phi_k, x^* - x_k \rangle + \langle \phi_k, y_k - x^* \rangle \\ &= \tilde{f} + \langle \phi_k, x^* - x_k \rangle + \langle \phi_k, \rho \cdot \frac{\phi_k}{\|\phi_k\|} \rangle \\ &= \tilde{f} + \langle \phi_k, x^* - x_k \rangle + \rho \|\phi_k\|. \end{aligned}$$

This implies

$$\begin{aligned} \langle \phi_k, x_k - x^* \rangle &\ge \rho \|\phi_k\| + \underbrace{\tilde{f} - f(y_k)}_{\to 0}. \\ &\ge \rho \|\phi_k\| \end{aligned}$$

$$\Rightarrow \frac{\langle \phi_k, x_k - x^* \rangle}{\|\phi_k\|} \nrightarrow \rho.$$

Note that $\phi_k \neq 0$ because $x_k \notin \operatorname{argmin}_{x \in \mathbb{R}^n} f(x)$, so division by $\|\phi_k\|$ is valid.

Consequently, we have

$$\begin{aligned}
&\|x_{k+1} - x^*\|^2 \\
&= \left\| (x_k - x^*) - \lambda_k \frac{\phi_k}{\|\phi_k\|} \right\|^2 \\
&= \|x_k - x^*\|^2 - 2\lambda_k \left\langle \frac{\phi_k}{\|\phi_k\|}, x_k - x^* \right\rangle + \left\| \lambda_k \frac{\phi_k}{\|\phi_k\|} \right\|^2 \\
&= \|x_k - x^*\|^2 - 2\lambda_k \frac{\langle \phi_k, x_k - x^* \rangle}{\|\phi_k\|} + \lambda_k^2 \\
&\leq \|x_k - x^*\|^2 - 2\lambda_k \rho + \lambda_k^2 \leq \lambda_k \rho
\end{aligned}$$

Since $\lambda_k \to 0$, there is a sufficiently large $k_0 \in \mathbb{N}$ such that

$$\lambda_k \leq \rho \quad \forall k \geq k_0.$$

Hence, for $k \geq k_0$, we have

$$\begin{aligned} &\|x_{k+1} - x^*\|^2 \\ &\leq \|x_k - x^*\|^2 - 2\lambda_k \rho + \lambda_k \rho \\ &= \|x_k - x^*\|^2 - \lambda_k \rho. \end{aligned}$$

Equivalently,

$$\|x_{k+1} - x^*\|^2 + \lambda_k \rho \leq \|x_k - x^*\|^2. \quad \forall k \geq k_0.$$

Telescoping this inequality gives

$$\|x_{k+1} - x^*\|^2 + \rho \sum_{i=k_0}^{k} \lambda_i \leq \|x_{k_0} - x^*\|^2 \qquad \forall k \geq k_0$$

This implies

$$\rho \sum_{i=k_0}^{k} \lambda_i \leq \|x_{k_0} - x^*\|^2 \qquad \forall k \geq k_0$$

$\to +\infty$

Taking the limit as $k \to \infty$, gives

$$\rho \sum_{i=k_0}^{\infty} \lambda_i \geq \|x_{k_0} - x^*\|^2 < \infty,$$

which contradicts (3.19). ∎

Example 3.4.3

Consider the nonsmooth minimisation problem

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{where } f(x) = \|Ax - b\|.$$

where $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. This is a bad way to formulate this problem as it is equivalent to the smooth minimisation problem considered last week:

$$\min_{x \in \mathbb{R}^n} \frac{1}{2}\|Ax - b\|^2.$$

Nevertheless, we will use it to illustrate the importance of problem formulation, and the difference between GD and sub-gradient method.

In order to apply the subgradient method, we must compute the subdifferential at a point x such that $Ax \neq b$. To do this, note that

$$f(x) = \left(\|Ax-b\|^2\right)^{\frac{1}{2}}$$

Then using the chain-rule for differentiation, we have.

$$\nabla f(x) = \frac{1}{2}\left(\|Ax-b\|^2\right)^{-\frac{1}{2}} \cdot 2A^T(Ax-b)$$

$$= \frac{A^T(Ax-b)}{\|Ax-b\|}$$

By Prop. 2.5.3, we have

$$\partial f(x) = \left\{ \frac{A^T(Ax-b)}{\|Ax-b\|} \right\} \quad \text{if } Ax \neq b.$$