OP202 HW3

Yanyu ZHOU

February 21, 2023

1 Q1

Consider

$$\min_{x \in X} f(x) \tag{PC}$$

$$\Updownarrow$$

$$\min_{x \in \mathbf{R}^n} f(x) + \iota_X(x) \tag{PS$'$}$$

where $\iota_X(x)$ is a indicator function. When $x \in X$, $\iota_X(x) = 1$, else $\iota_X(x) = \infty$.

Proof:$\partial \iota_X(x) \equiv N_X(x)$

For (PC), the necessary and sufficient conditions for optimality is

$$\partial f(x^*) + N_X(x^*) \ni \mathbf{0}$$

where $N_X(x)$ is the normal cone operator of X at x. By definition, the *normal cone* of a set $\mathbf{X}$ at a boundary point x_0 is the set of all vectors y such that $y^T(x - x_0) \geq 0$ for all $x \in \mathbf{X}$.

By definition, $\partial \iota_X(x) = \{g \in \mathbf{R}^n : g^T x \geq g^T y \quad \forall y \in \mathbf{X}\}$. Therefore, $\partial \iota_X(x) \equiv N_X(x)$.

2 Q2

Work out one of the proofs of Theorem 17 (the simplest one is $f_1 \in \mathbf{S}^{1,1}_{m,L}$). Here we prove that Theorem 12 in class 2 is also valid for the proximal gradient method. Referrence: CMU 10-725 Lecture 8.

For $f(x) = f_1(x) + f_2(x)$, we assume that

- f_1 is convex and at least $f_1 \in \mathbf{S}^{1,1}_{0,L}$
- f_2 is convex, $\operatorname{prox}_{\alpha h}(x) = \arg\min_z \left\{ \frac{\|x-z\|^2}{2\alpha} + h(z) \right\}$ can be evaluated.

Notation: The update step of proximal gradient method can be written as

$$x_k = x_{k-1} - t_k G_{t_k}(x_{k-1})$$

where $G_t(x)$ is the generalized gradient and is given by

$$G_t(x) = \frac{x - \text{prox}_t(x - t\nabla f_1(x))}{t}.$$

Then the following holds:
Theorem: The proximal gradient method with constant step size $\alpha \leq \frac{2}{L}$ has an ergodic(global) convergence certificate of

$$f\left(\bar{x}_t\right) - f\left(x^*\right) \leq \frac{C}{t+1}\left\|x_0 - x^*\right\|^2$$

for $\bar{x}_t = \frac{1}{t+1}\sum_{k=0}^{t} x_k,$ and $C > 0$.
Proof: We begin by showing that

$$f(y) \leq f_1(x) + \nabla g(x)^T(y-x) + \frac{L}{2}\left\|y-x\right\|^2 + f_2(y) \quad \forall x, y.$$

Since ∇f_1 is Lipschitz with constant L, $\nabla^2 f_1$
Remark:

- This has the same convergence rate $O(\frac{1}{t})$ as that of gradient descent.

3 Q2

Depending on f, you can derive conditions for which the dual ascent is converging depending on α and f you can have linear convergence, or $O(1/t), O\left(1/t^2\right), O(1/\sqrt{t})\ldots$

4 Q3

Homework. Can you think of a dual Newton's method? Can you think of a dual proximal method? When would you apply these methods and with which guarantees?