Sub gradient Descent Method

Saturday, 30 September 2023 8:50 PM

A function $f: \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is said to be proper if $f(x) > -\infty \ \forall x$ and $f(x) < \infty$ for some $x \in \mathbb{R}^n$.

Effective domain of is

$$\mathrm{dom}(f) = \{ x \mid f(x) < \infty \}$$

For $x \in \mathrm{dom}(f)$, $g \in \mathbb{R}^n$ is said to be a subgradient of f at x if $f(y) \geq f(x) + g^T(y-x) \quad \forall y \in \mathrm{dom}(f)$

Collection of all subgradients is called sub-differential of f at x and denoted as $\partial f(x)$.

$x^* \in \mathrm{dom}(f)$ will be optimal solution of f iff $0 \in \partial f(x^*)$.

g is a sub differential of f at x, then $-g$ is not necessarily a descent direction.

<u>for example:-</u>

Consider $f(x) = |x_1| + 2|x_2|$

Let $x = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$, $\partial f(x) = \begin{pmatrix} 1 \\ [-2, 2] \end{pmatrix}$

Clearly $g = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \in \partial f(x)$

Now for $\alpha > 0$,

$$\begin{aligned} f(x - \alpha g) &= f\left(\begin{pmatrix} 1-\alpha \\ -2\alpha \end{pmatrix}\right) \\ &= |1-\alpha| + 2|-2\alpha| \\ &\geq 1 - \alpha + 4\alpha \\ &= 1 + 3\alpha \\ &> 1 = f\begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{for any } \alpha > 0. \end{aligned}$$

Hence $d = -g = \begin{pmatrix} -1 \\ -2 \end{pmatrix}$ is not a descent direction

Suppose $x \in \mathrm{dom}(f)$ and $0 \notin \partial f(x)$ then there exists some $\xi \in \partial f(x)$ such that $-\xi$ is a descent direction of f at x.

define $\xi_{min} = \underset{\xi \in \partial f(x)}{\arg\min} \|\xi\|_2$

then $-\xi_{min}$ is a descent direction of f at x.

However finding ξ_{min} is not easy.

<u>Subgradient descent method</u>

<u>step0</u>:- Input f, x^0, set $k := 0$

<u>step1</u>:- Consider $d^k = -\xi^k$ where $\xi^k \in \partial f(x^k)$ is choosen randomly.

<u>step2</u>: Update $x^{k+1} = x^k - \alpha_k \xi^k$

<u>step3</u>: Define $f_{best}^{k+1} = \mathrm{argmin}\{ f_{best}^k, f(x^{k+1}) \}$

<u>Note</u>:- In gradient descent method, $f(x^{k+1}) < f(x^k)$ holds for all k, So in step 3, $f_{best}^{k+1} = x^{k+1}$ holds for all k.

<u>Step length selection</u>

Following techniques are used to find step length:-

1. Constant step size, $\alpha_k = h \ \forall k$
2. <u>Square summable but not summable</u>:- Choose $\{\alpha_k\}$ s.t.

$$\sum \alpha_k = \infty \quad \text{but} \quad \sum \alpha_k^2 < \infty$$

- One typical example

$$\alpha_k = \frac{a}{b+k} \qquad a > 0, b \geq 0$$

3. <u>Non-summable diminishing</u>:-

The step size satisfy

$$\lim_{k \to \infty} \alpha_k = 0, \quad \sum \alpha_k = \infty.$$

Typical example:

$$\alpha_k = \frac{a}{\sqrt{k}} \qquad a > 0$$

<u>Assumption</u>:- f satisfies Lipschitz condition i.e

$$|f(x) - f(y)| \leq L\|x - y\| \quad \forall x, y \in \mathrm{dom}(f).$$

If f is Lipschitz then $\|\xi\| \leq L \quad \forall \xi \in \partial f(x)$ for any $x \in \mathrm{dom}(f)$.

[Try to prove this].

Suppose $x^* = \mathrm{argmin}\, f(x)$

We have

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &= \|x^k - \alpha_k \xi^k - x^*\|^2 = \|x^k - x^*\|^2 - 2\alpha_k \xi^{kT}(x^k - x^*) + \alpha_k^2 \|\xi_k\|^2 \\ &\leq \|x^k - x^*\|^2 - 2\alpha_k (f(x^k) - f(x^*)) + \alpha_k^2 \|\xi_k\|^2 \qquad \text{— (1)} \end{aligned}$$

last inequality holds since

$$f(x^*) \geq f(x^k) + \xi^{kT}(x^* - x^k)$$

Applying (1) recurcively,

$$\|x^{k+1} - x^*\|^2 \leq \|x^0 - x^*\|^2 - 2\sum_{i=0}^{k} \alpha_i (f(x^i) - f^*) + \sum_{i=0}^{k} \alpha_i^2 \|\xi^i\|^2$$

This implies,

$$\begin{aligned} \|x^0 - x^*\|^2 + \sum_{i=0}^{k} \alpha_i^2 \|\xi^i\|^2 &\geq 2\sum_{i=0}^{k} \alpha_i (f(x^i) - f^*) + \|x^{k+1} - x^*\| \\ &\geq 2\sum_{i=0}^{k} \alpha_i (f(x^i) - f^*) \end{aligned}$$

Combining this with

$$\sum_{i=0}^{k} \alpha_i (f(x^i) - f^*) \geq \min_{i=0,1,\dots,k} (f(x^i) - f^*) \sum_{i=0}^{k} \alpha_i$$

we have

$$f_{best}^k - f^* = \left(\min_{i=0,1,\dots,k} f(x^i) - f^*\right) \leq \frac{\|x^0 - x^*\|^2 + \sum_{i=0}^{k} \alpha_i^2 \|\xi^i\|^2}{2\sum_{i=0}^{k} \alpha_i} \qquad \text{— (2)}$$

<u>Constant step size</u>,

Let $\alpha_k = h \ \forall k$, Then from (2)

$$f_{best}^k - f^* \leq \frac{\|x^0 - x^*\|^2 + L^2 h^2 k}{2hk}$$

RHS $\to \frac{L^2 h}{2}$ as $k \to \infty$.

Hence Subgradien method converges within $Gh/2$ of the optimal soln.

<u>Square summable but not summable</u>.

Suppose $\sum_{i=0}^{\infty} \alpha_i^2 = \gamma < \infty$ and

$$\sum_{i=0}^{\infty} \alpha_i = \infty$$

Then from (2)

$$f_{best}^k - f^* \leq \frac{\|x^0 - x^*\|^2 + L^2 \sum \alpha_i^2}{2\sum_{i=0}^{k} \alpha_i}$$

Clearly RHS converges to 0 as $k \to \infty$.

i.e subgradient descent converges in the sense $f_{best}^k \to f^*$.

Similar result holds for third case.

<u>Practice problems</u>:-

1. Find $\xi \in \partial f(x^0)$ s.t $-\xi$ is a descent direction of f at x^0,

(a) $f(x) = 3|x_1| + 5|x_2|$

$x^0 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$

(b) $f(x) = \max_i \{ |x_i| \quad i = 1, 2 \}$

$x^0 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$

(c) $f(x) = \max\{ 3x_1 - 2x_2 + 1, -5x_1 + 3x_2 - 4 \}$

$x^0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$

(d) $f(x) = \max\{ |x_1 - 2|, |x_2 + 1| \}$

$x^0 = \begin{pmatrix} -1 \\ 2 \end{pmatrix}$