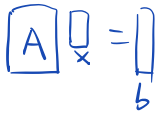
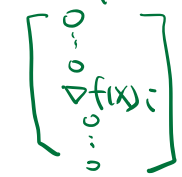
Part III : Randomized/Stochastic Methods

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



$\nabla f(x) = A^T(Ax-b)$

$\nabla f(x)^{(i)}$ denotes i-th entry of $\nabla f(x)$



① Grdient Descent is $x_{k+1} = x_k - \eta \nabla f(x_k)$

② Coordinate Descent $x_{k+1} = x_k - \eta \nabla f(x_k)^{i(k)}$

$i(1)=1$
$i(2)=2$
$\vdots$

only $i(k)$-th entry of x_k is updated

$\nabla f(x)^{(1)} =$ 

③ Randomized Coordinate Descent

$$x_{k+1} = x_k - \eta \nabla f(x_k)^{i(k)}$$

$i(k) \sim$ i.i.d. uniform distribution in $\{1,2,\cdots,n\}$

identical
independent
distributed

Consider an opertator $T: \mathbb{R}^n \to \mathbb{R}^n$

$x \mapsto T(x)$

and a fixed point iteration

$$x_{k+1} = T(x_k)$$

① An operator S is nonexpansive if

$$\|S(x) - S(y)\| \leq \|x-y\|$$

Example : If $\nabla f(x)$ is L-Cont. and $f(x)$ is convex

then $S(x) = [I - \frac{2}{L}\nabla f](x)$ is nonexpansive

$$\|S(x) - S(y)\|^2 = \|-\tfrac{2}{L}[\nabla f(x) - \nabla f(y)] + (x-y)\|^2$$

$$= \|x-y\|^2 + \tfrac{4}{L^2}\|\nabla f(x) - \nabla f(x)\|^2 - \tfrac{4}{L}\langle x-y, \nabla f(x) - \nabla f(y)\rangle$$

$$\leq \|x-y\|^2$$

2.2.3 Convergence for convex functions

Theorem 2.8. *Assume $\nabla f(\mathbf{x})$ is Lipschitz-continuous with Lipschitz constant L and $f(\mathbf{x}) : \mathbb{R}^n \longrightarrow \mathbb{R}$ is convex. Then for any $\mathbf{x}, \mathbf{y}$:*

1. $f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle + \frac{1}{2L}\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2$
2. $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\|^2 \leq L\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle$.

② $T = (1-\theta)I + \theta S$ with $\theta \in (0,1)$

is called θ-averaged if S is nonexpansive

Example : $S = [I - \frac{2}{L}\nabla f]$

$$T = I - \eta\nabla f = (1-\theta)I + \theta(I - \tfrac{2}{L}\nabla f)$$

$$\theta = \frac{\eta L}{2} \in (0,1) \iff 0 < \eta < \frac{2}{L}$$

③ Recall we did the following on Mar 3 :

Theorem (Browder-Göhde-Kirk)

$S : \mathbb{R}^n \to \mathbb{R}^n$ is nonexpansive $\Rightarrow$ S has at least one fixed point.

$S(x_*) = x_*$.

$x_{k+1} = S(x_k)$ may not converge to x_*

Example: $S(x) = -x$ $\quad x_* = 0$

Theorem If $S: \mathbb{R}^n \to \mathbb{R}^n$ is nonexpansive, then

$$x_{k+1} = \theta x_k + (1-\theta) S(x_k), \quad 0 < \theta < 1$$

converges to one fixed point of $S(x)$.

Example: This implies GD converges if $\eta < \frac{2}{L}$.

④ $T(x) = \begin{bmatrix} [T(x)]_1 \\ [T(x)]_2 \\ \vdots \\ [T(x)]_n \end{bmatrix}$ $\quad T_i(x) = \begin{bmatrix} x_1 \\ \vdots \\ x_{i-1} \\ [T(x)]_i \\ x_{i+1} \\ \vdots \\ x_n \end{bmatrix}$

If $x_{k+1} = x_k - \eta \nabla f(x_k) = T(x_k)$ is GD

$x_{k+1} = x_k - \eta \nabla f(x)^{i(k)} \iff x_{k+1} = T_{i(k)}(x_k)$

Theorem Assume

① T is θ-averaged ($\Leftrightarrow \eta < \frac{2}{L}$ in GD)

② $i(k) \in \{1, \dots, n\}$ is i.i.d. with uniform probability.

Then $x_{k+1} = T_{i(k)}(x_k)$

converges to one fixed point of $T(x)$ with probability 1.

Proof: Define R, R_i by $\begin{cases} T = I - \theta R \\ T_i = I - \theta R_i \end{cases}$

$$\Rightarrow R_i(x) = \begin{bmatrix} 0 \\ \vdots \\ [R(x)]_i \\ \vdots \\ 0 \end{bmatrix} \qquad I - R = I - \frac{I-T}{\theta} = (1-\tfrac{1}{\theta})I + \tfrac{1}{\theta}T = S$$

$$x_{k+1} = T_{i(k)}(x_k) \Leftrightarrow x_{k+1} = x_k - \theta R_{i(k)}[x_k]$$

T is θ-averaged $\Leftrightarrow T = (1-\theta)I + \theta S$ with S nonexpansive

$\Leftrightarrow \frac{1}{\theta}T - (\frac{1}{\theta}-1)I$ is nonexpansive

$\Leftrightarrow I - R$ is nonexpansive

$\Leftrightarrow \|x - Rx - y + Ry\|^2 \le \|x-y\|^2$

$\Leftrightarrow \frac{1}{2}\|Rx - Ry\|^2 \le \langle x-y, Rx-Ry \rangle$

$T(x_*) = x_* \Leftrightarrow R(x_*) = 0$

$y = x_* \Rightarrow \frac{1}{2}\|Rx\|^2 \le \langle Rx, x - x_* \rangle$

$\mathbb{E}$ denotes expectation w.r.t. random variables: $i(0), i(1), \dots$

Example: ① X is a random variable taking values in $\{0,1\}$ with equal probability

$$\mathbb{E}(X) = 0\cdot\frac{1}{2} + 1\cdot\frac{1}{2} = \frac{1}{2} \qquad P(X=0) = \frac{1}{2}$$
$$\mathbb{E}(X^2) = 0^2\cdot\frac{1}{2} + 1^2\cdot\frac{1}{2} = \frac{1}{2} \qquad P(X=1) = \frac{1}{2}$$
$$\mathbb{E}(f(X)) = f(0)\cdot\frac{1}{2} + f(1)\cdot\frac{1}{2} = \frac{1}{2}f(1)$$

② X is a random variable taking values in $\{x_1, x_2, \dots, x_k\}$ with probability $p_1, p_2, \dots, p_k$ $\quad \sum_{i=1}^{k} p_i = 1,\ p_i \geq 0$

$$\mathbb{E}(X) = x_1\cdot p_1 + x_2\cdot p_2 + \dots + x_k p_k \rightarrow \text{convex combination}$$
$$\mathbb{E}(f(X)) = \sum_{i=1}^{k} f(x_i) p_i \qquad \text{Prob}(X = x_i) = p_i$$

$f(\mathbb{E}(X)) \leq \mathbb{E}(f(X))$ if $f(x)$ is convex

③ X, Y are i.i.d. random variable taking values in $\{x_1, x_2, \dots, x_k\}$ with probability $p_1, p_2, \dots, p_k$

$$\sum_{i=1}^{k} p_i = 1,\ p_i \geq 0$$

Joint probability

independence $\Rightarrow P(X=x_i, Y=x_j) = P(X=x_i)\,P(Y=x_j)$

$$\mathbb{E}[f(X,Y)] = \sum_{i=1}^{k}\sum_{j=1}^{k} f(x_i, x_j) p_i p_j$$

In $\begin{cases} X_{k+1} = X_k - \eta \nabla f(X_k)^{i(k)} \\ X_{k+1} = T_{i(k)}(X_k) \end{cases}$,

X_0 is deterministic and $i(0), i(1), \cdots, i(N)$ are random

X_N is a function of $N+1$ random variables

$E(X_N)$ denotes expectation w.r.t. $N+1$ random variables

④ Conditional Probability & Expectation

X is a random variable taking values in $\{x_1, x_2, \cdots, x_k\}$ with probability $p_1, p_2, \cdots, p_k$

Y is a random variable taking values in $\{y_1, y_2, \cdots, y_\ell\}$ with probability $q_1, q_2, \cdots, q_\ell$

$$P(X = x_i \mid Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}$$ ← joint prob.

$$E(X \mid Y = y_j) = \sum_{i=1}^{k} x_i P(X = x_i \mid Y = y_j)$$

$$= \sum_{i=1}^{k} x_i \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}$$

$$E(f(x) \mid Y = y_j) = \sum_{i=1}^{k} f(x_i) \cdot P(X = x_i \mid Y = y_j)$$

$$= \sum_{i=1}^{k} f(x_i) \cdot \frac{P(X = x_i, Y = y_j)}{}$$

$P(Y=y_j)$

E_k denotes the conditional expectation w.r.t. $i(k)$

Conditioned on the past random variables

$i(k-1), i(k-2), \cdots, i(0)$

Then ① $E_k(X_k) = X_k$ because X_k does NOT depend on $i(k)$

② Let X be something depending on all $i(k)$ for $k=1,\cdots,N$

$$E[E_k(X)] = E[X]$$

Law of Total Expectation

$E(X|Y)$ is a function of Y

$$E(X|Y) = \sum_{i=1}^{k} x_i P(X=x_i|Y)$$

$$E[E(X|Y)] = E\left[\sum_{i=1}^{k} x_i P(X=x_i|Y)\right]$$

$$= \sum_{j=1}^{l}\left[\sum_{i=1}^{k} x_i P(X=x_i|Y=y_j)\right] q_j$$

$$= \sum_{j=1}^{l}\sum_{i=1}^{k} x_i P(X=x_i|Y=y_j)\cdot P(Y=y_j)$$

$$= \sum_{j=1}^{l}\sum_{i=1}^{k} x_i P(X=x_i, Y=y_j)$$

$$= \sum_{i=1}^{k} x_i \left[\sum_{j=1}^{l} P(X=x_i, Y=y_j)\right]$$

$$= \sum_{i=1}^{k} x_i \, P(X = x_i)$$

$$= E(X)$$