

Part III : Randomized/Stochastic Methods

$$\boxed{A} \begin{bmatrix} x \\ \end{bmatrix} = \begin{bmatrix} b \end{bmatrix}$$

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 $\nabla f(x)$

$$\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \nabla f(x)_i \\ 0 \\ \vdots \end{bmatrix}$$

① Gradient 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

$$\nabla f(x)^{(i)} = \begin{bmatrix} \text{zeros} \\ \text{row } i \text{ of } A^T(Ax - b) \end{bmatrix}$$

③ Randomized Coordinate Descent

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

$i(k) \sim \text{i.i.d. uniform distribution in } \{1, 2, \dots, n\}$
identical
independent
distributed

Consider an operator $T: \mathbb{R}^n \rightarrow \mathbb{R}^n$
 $x \mapsto T(x)$

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

$$\begin{aligned} \|S(x) - S(y)\|^2 &= \left\| -\frac{2}{L} [\nabla f(x) - \nabla f(y)] + (x - y) \right\|^2 \\ &= \|x - y\|^2 + \frac{4}{L^2} \|\nabla f(x) - \nabla f(y)\|^2 - \frac{4}{L} \langle x - y, \nabla f(x) - \nabla f(y) \rangle \\ &\leq \|x - y\|^2 \end{aligned}$$

2.2.3 Convergence for convex functions

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

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

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

is called θ -averaged if S is nonexpansive

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

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

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

③ Assume S is nonexpansive

and S has at least one fixed point

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

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

Theorem If $S: \mathbb{R}^n \rightarrow \mathbb{R}^n$ is nonexpansive,
and there is at least one fixed point,

then $x_{k+1} = \theta x_k + (1-\theta)S(x_k)$, $0 < \theta < 1$
converges to one fixed point of $S(x)$.

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

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

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

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

$\downarrow$
 $\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \nabla f(x)_i \\ 0 \\ \vdots \\ 0 \end{bmatrix}$

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
$$\begin{cases} T = I - \theta R \\ T_i = I - \theta R_i \end{cases}$$

$$\Rightarrow R_i(x) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ [R(x)]_i \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

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

T is θ -averaged $\Leftrightarrow T = (1-\theta)I + \theta S$ $\begin{cases} T = I - \theta R \\ T = (1-\theta)I + \theta S \end{cases}$
with S nonexpansive

$$\begin{aligned} I - R &= I - \frac{I - T}{\theta} \\ &= (1 - \frac{1}{\theta})I + \frac{1}{\theta}T \\ &= S \end{aligned}$$

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

$$\Leftrightarrow I - R \text{ is nonexpansive}$$

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

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

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

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

$$T = I - \eta \nabla f$$

$$R = \frac{\eta}{\theta} \nabla f = \frac{\theta}{L} \nabla$$

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

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

$$E(X) = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}$$

$$P(X=0) = \frac{1}{2}$$

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

$$P(X=1) = \frac{1}{2}$$

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

② X is a random variable taking values

in $\{x_1, x_2, \dots, x_k\}$

with probability $p_1, p_2, \dots, p_k$

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

$$E(X) = x_1 \cdot p_1 + x_2 \cdot p_2 + \dots + x_k \cdot p_k \rightarrow \text{Convex combination}$$

$$E(f(X)) = \sum_{i=1}^k f(x_i) p_i$$

$$\text{Prob}(X=x_i) = p_i$$

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

② X, Y are i.i.d. random variable taking values

$\{x_1, x_2, \dots, x_k\}$

with probability $p_1, p_2, \dots, p_k$

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

Joint probability

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

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

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

X_0 is deterministic and $i(0), i(1), \dots, i(N)$ 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 $\{x_1, x_2, \dots, x_k\}$

with probability $p_1, p_2, \dots, p_k$

Y is a random variable taking values

in $\{y_1, y_2, \dots, y_\ell\}$

with probability $q_1, q_2, \dots, q_\ell$

$$P(X=x_i | Y=y_j) = \frac{P(X=x_i, Y=y_j)}{P(Y=y_j)} \leftarrow \text{joint prob.}$$

$$\begin{aligned} E(X | Y=y_j) &= \sum_{i=1}^k x_i P(X=x_i | Y=y_j) \\ &= \sum_{i=1}^k x_i \frac{P(X=x_i, Y=y_j)}{P(Y=y_j)} \end{aligned}$$

$$\begin{aligned} E(f(x) | Y=y_j) &= \sum_{i=1}^k f(x_i) \cdot P(X=x_i | Y=y_j) \\ &= \sum_{i=1}^k f(x_i) \cdot \frac{P(X=x_i, Y=y_j)}{P(Y=y_j)} \end{aligned}$$

$$P(Y=y_j)$$

E_k denotes the conditional expectation w.r.t. $i(k)$
Conditioned on the past random variables

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

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 $k=1, \dots, N$

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

Law of Total Expectation

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

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

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

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

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

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

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

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

$$= E(X)$$