Randomized Coordinate Descent $\quad \min_{x \in \mathbb{R}^n} f(x)$

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

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

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

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

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

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

$\text{Prob}(i(1)=1) = \frac{1}{n}$

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

$$x \longmapsto T(x)$$

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

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

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

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

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

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

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

③ Recall we did the following on Mar 3 :

<u>Theorem</u> (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

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

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

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

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

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 ($\Rightarrow T$ has at least one fixed point)

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.

Example: $S = I - \frac{2}{L}\nabla f$ is nonexpansive if $f(x)$ is convex

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

$= (1-\theta) I + \theta S$ is θ-averaged

if $\theta = \frac{\eta L}{2} < 1 \iff \eta < \frac{2}{L}$

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

$\Rightarrow \quad R_i(x) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ [R(x)]_i \\ 0 \\ \vdots \\ 0 \end{bmatrix}$ $\quad R = \frac{\eta}{\theta}\nabla f = \frac{2}{L}\nabla f$

$T = I - \theta \cdot \frac{2}{L}\nabla f$

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

T is θ-averaged $\iff T = (1-\theta)I + \theta S$

with S nonexpansive

$I - R = I - \frac{I-T}{\theta}$

$= (1-\frac{1}{\theta})I + \frac{1}{\theta}T$

$= S$

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

$\iff I - R$ is nonexpansive

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

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

$T = I - \theta R$

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

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

$\frac{1}{2}\|\frac{2}{L}\nabla f(x)\|^2 \leq \langle \frac{2}{L}\nabla f(x), x - x_*\rangle$

$\iff \|\nabla f(x)\|^2 \leq L\langle \nabla f(x), x - x_*\rangle$

Definition of Expectation & more :

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

definition
$$\begin{cases} E(X) = 0\cdot\frac{1}{2} + 1\cdot\frac{1}{2} = \frac{1}{2} \\ E(X^2) = 0^2\cdot\frac{1}{2} + 1^2\cdot\frac{1}{2} = \frac{1}{2} \\ E(f(X)) = f(0)\cdot\frac{1}{2} + f(1)\cdot\frac{1}{2} = \frac{1}{2}f(1) \end{cases}$$

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

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

E denotes expectation w.r.t. random variable X

Expectation for disrete random variables are defined as above

② X is a random variable taking values in $\{x_1, x_2, \cdots, x_k\}$

with probability $P_1, P_2, \cdots, P_k$ $\quad \sum_{i=1}^{k} P_i = 1$, $P_i \geq 0$

definition
$$\begin{cases} E(X) = x_1\cdot P_1 + x_2\cdot P_2 + \cdots + x_k P_k \rightarrow \text{convex combination} \\ E(f(X)) = \sum_{i=1}^{k} f(x_i) P_i \end{cases}$$

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

Jensen's: $f(E(X)) \leq E(f(X))$ if $f(x)$ is convex

Example: $f(x) = x^2 \Rightarrow [E(X)]^2 \leq E(X^2)$

③ X, Y are i.i.d. random variable taking values $\{a_1, a_2, \cdots, a_k\}$

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

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

Joint probability $P(X=a_i, Y=a_j)$

X & Y are independent $\Rightarrow$

$$P(X=a_i, Y=a_j) = P(X=a_i) P(Y=a_j)$$

$$E[f(X,Y)] = \sum_{i=1}^{k} \sum_{j=1}^{k} f(a_i, a_j) P(X=a_i, Y=a_j)$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{k} f(a_i, a_j) p_i p_j$$

In $\quad x_{k+1} = T_{i(k)}(x_k)$

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

x_0 is deterministic and $i(0), i(1), \cdots, i(N)$ are random

x_N is a function of these $N+1$ random variables

$E(x_N)$ denotes expectation w.r.t. $N+1$ i.i.d. 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 $q_1, q_2, \cdots, q_\ell$

Probability of event $X = x_i$ given the knowledge $Y = y_j$

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

The conditional expectation w.r.t. X given $Y = y_j$

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

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

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

is $\begin{cases} \text{a function of } y \\ \text{a random variable} \end{cases}$

We can also write it as $g(Y) = E(X \mid Y)$

So $E(X \mid Y)$ is a random variable.

$$E_k = E[i(k) \mid i(k-1), i(k-2), \cdots, i(0)]$$

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

Conditioned on the past random variables

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

$$X_{k+1} = T_{i(k)}(X_k)$$

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

Then ① $\underbrace{E_k(X_k)}_{\text{random variable}} = X_k$ because X_k does NOT depend on $i(k)$

② Let X be something depending on all $i(k)$
$i(k-1), i(k-2), \ldots, i(0)$

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

Law of Total Expectation : $E[E(X|Y)] = E(X)$

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

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

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

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

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

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

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

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

$$= E(X)$$

③ $$E_k\left[R_{i(k)}(X_k) \right] = \frac{1}{n} R(X_k)$$

$$R = \frac{2}{L} \nabla f \qquad R_i(X) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \nabla f(X)_i \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

E_k is expectation w.r.t. $i(k)$

$$\begin{cases} P(i(k) = 1) = \frac{1}{n} \\ P(i(k) = 2) = \frac{1}{n} \\ \quad \vdots \\ P(i(k) = n) = \frac{1}{n} \end{cases}$$

$$\Rightarrow E_k\left[R_{i(k)}(X_k) \right] = \sum_{j=1}^{n} P\left[i(k) = j \right] \cdot R_j(X_k)$$

$$= \frac{1}{n} \sum_{j=1}^{n} R_j(X_k) = \frac{1}{n} R(X_k)$$

$$E_k \left\| R_{i(k)}(X_k) \right\|^2 = \frac{1}{n} \left\| R(X_k) \right\|^2$$

$$\left\| R_{i(k)}(X_k) \right\|^2 = \left(\left[R(X_k) \right]_{i(k)} \right)^2$$

$$E_k \left(\left[R(X_k) \right]_{i(k)} \right)^2 = \sum_{j=1}^{n} P\left[i(k) = j \right] \cdot \left(R_j(X_k) \right)^2$$

$$= \frac{1}{n}\sum_{j=1}^{n}\left(R_j(x_k)\right)^2 = \frac{1}{n}\|R(x_k)\|^2$$

④ Let x_* be a fixed point to $x_* = T(x_*)$

$$\|x_{k+1} - x_*\|^2 = \|x_k - \theta R_{i(k)}(x_k) - x_*\|^2$$
$$= \|x_k - x_*\|^2 - 2\theta \langle R_{i(k)} x_k, x_k - x_* \rangle$$
$$+ \theta^2 \|R_{i(k)}(x_k)\|^2$$

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

Expectation is linear

$$E_k \|x_{k+1} - x_*\|^2 = E_k\|x_k - x_*\|^2 - 2\theta E_k \langle R_{i(k)} x_k, x_k - x_* \rangle$$
$$+ \theta^2 E_k \|R_{i(k)}(x_k)\|^2$$
$$= \|x_k - x_*\|^2 - 2\theta \langle E_k[R_{i(k)}(x_k)], x_k - x_* \rangle$$
$$+ \theta^2 \cdot \frac{1}{n}\|R(x_k)\|^2$$

$$= \|x_k - x_*\|^2 - \frac{2\theta}{n} \langle R(x_k), x_k - x_* \rangle$$
$$+ \theta^2 \cdot \frac{1}{n}\|R(x_k)\|^2$$
$$\leq \|x_k - x_*\|^2 - \frac{\theta}{n}\|R(x_k)\|^2 + \frac{\theta^2}{n}\|R(x_k)\|^2$$

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

Take E for both sides

$$\Rightarrow \quad E\|x_{k+1} - x_*\|^2 \leq E\|x_k - x_*\|^2 - (1-\theta)\frac{\theta}{n} E\|R(x_k)\|^2$$

$\Rightarrow \quad \{ E\|X_k - X_*\| \} \searrow$

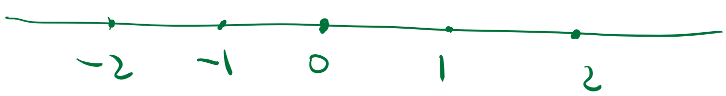
⑤ Discrete Martingale

is a sequence of random variable $X_1, X_2, X_3, \cdots$ satisfying

① $E(|X_n|) < +\infty$

② $E(X_{n+1} \mid X_1, \cdots, x_n) = X_n$

Example : positions of 1D random walk



Integer grid ; $X_0 = 0$; $X_{k+1} = \begin{cases} X_k + 1 & \text{with prob. } \frac{1}{2} \\ X_k - 1 & \text{with prob. } \frac{1}{2} \end{cases}$

$E(|X_n|) \leq n$

Given $x_1, \cdots, x_n$, expected value of X_{n+1} is X_n

$E(X_{n+1} \mid x_1, \cdots, X_n) = x_n$

Supermartingale if $E(X_{n+1} \mid X_1, \cdots, x_n) \leq X_n$

submartingale if $E(X_{n+1} \mid X_1, \cdots, x_n) \geq X_n$

$\boxed{\text{Supermartingale Convergence Theorem}}$

Let X_k, Y_k be random variables satisfying

① $X_k \geq 0$, $Y_k \geq 0$ almost surely

② $E[X_{k+1} | x_1, \cdots, x_k] \leq X_k - Y_k$

Then almost surely (with probility 1)

1) X_k converges to X_∞, $k \to \infty$

2) $\sum_{k=0}^{\infty} Y_k < +\infty$

Remark: X_∞ is a random variable

$$\boxed{E_k \|X_{k+1} - X_*\|^2 \leq \|X_k - X_*\|^2 - (1-\theta)\frac{\theta}{n} \|R(X_k)\|^2}$$

$$\Rightarrow \begin{cases} \sum_{k=0}^{\infty} \|R(X_k)\|^2 < \infty \Rightarrow \|R(X_k)\| \to 0 \\ \lim_{k\to\infty} \|X_k - X_*\|^2 \text{ exists} \end{cases} \quad \text{with probability } 1$$

Let Fix(T) be the set of all fixed points of T

$\forall x_* \in \mathrm{Fix}(T)$, $[\lim_{k\to\infty} \|X_k - X_*\| \text{ exists with prob. } 1]$

$\Rightarrow$ With prob. 1 $[\forall x_* \in \mathrm{Fix}(T), \lim_{k\to\infty} \|X_k - X_*\| \text{ exists}]$

nontrivial, skipped, see last reference book.

⑥ With prob. 1, we have

1) $\{x_k\}$ is a bounded sequence

thus a convergent subsequence $\{X_{k_j}\} \to Z$

2) $\|R(x_k)\| \to 0 \Rightarrow \|\frac{1}{\theta}(I-T)(x_k)\| \to 0$

$\Rightarrow \|(I-T)x_k\| \to 0$

T is θ averaged

$\Rightarrow \|T(x)-T(y)\| \leq \|x-y\|$

$\Rightarrow$ T is continuous

$\Rightarrow$ $I-T$ is continuous

$\Rightarrow \|(I-T)x_{k_j}\| \to 0$

$\Rightarrow \|z-T(z)\| = 0$

$\Rightarrow z - T(z) = 0$

$\Rightarrow z \in Fix(T)$

$$\boxed{\|x_{k+1}-x_*\|^2 = \|x_k - x_*\|^2 - 2\theta\langle R_{i(k)}x_k, x_k - x_*\rangle + \theta^2\|R_{i(k)}(x_k)\|^2}$$

$\boxed{\|R(x_k)\| \to 0}$

set $x_* = z \to \{x_k\}$ has the same limit as x_{k_j}