Today's Plan
- I. Review of convergence of GD
- II. Conditional Expectation & Martingale
- III. Convergence of Randomized Coordinate Descent
- IV. Randomized Coordinate of fixed point iteration

I. Review: Convergence of $x_{k+1} = x_k - \eta \nabla f(x_k)$ under the assumptions of
- ① ∇f is L-cont.
- ② f is convex
- ③ there is at least one minimizer

Let x_* be any minimizer of $f(x)$

Step 1 ∇f is L-cont., f is convex $\Rightarrow \|\nabla f(x) - \nabla f(y)\|^2 \le L \langle \nabla f(x) - \nabla f(y), x - y \rangle$

↓ Theorem 2.9 in Sec 2.3 in Typed Notes

$y = x_* \Rightarrow -L \langle \nabla f(x) - \nabla f(x_*), x - x_* \rangle \le -\|\nabla f(x) - \nabla f(x_*)\|^2$

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

Step 2

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

$$\|x_{k+1} - x_*\|^2 = \|x_k - x_* - \eta \nabla f(x_k)\|^2$$

$$= \|x_k - x_*\|^2 - 2\eta \langle x_k - x_*, \nabla f(x_k) \rangle + \eta^2 \|\nabla f(x_k)\|^2$$

$$\le \|x_k - x_*\|^2 - 2\eta \cdot \frac{1}{L} \cdot \|\nabla f(x_k)\|^2 + \eta^2 \|\nabla f(x_k)\|^2$$

$$= \|x_k - x_*\|^2 + \eta\left(\eta - \frac{2}{L}\right)\|\nabla f(x_k)\|^2$$

$\propto \eta < \frac{2}{L} \Rightarrow \|x_{k+1} - x_*\|$ ↘

Step 3 Let $\begin{cases} \omega = -\eta\left(\eta - \frac{2}{L}\right) > 0 \\ r_k^2 = \|x_k - x_*\|^2 \end{cases}$, then

$$r_{k+1}^2 \le r_k^2 - \omega\|\nabla f(x_k)\|^2 \Rightarrow \sum_{k=0}^{\infty}\|\nabla f(x_k)\|^2 \le \frac{1}{\omega} r_0^2$$

$$\Rightarrow \|\nabla f(x_k)\|^2 \to 0 \Rightarrow \nabla f(x_k) \to 0 \quad (*)$$

Step 4 $\|x_{k+1} - x_*\|$ non-increasing $\Rightarrow \{x_k\}$ is bounded

$\Rightarrow$ a convergent subsequence $x_{k_j} \to \bar{x}$

Continuity of $\nabla f \Rightarrow \nabla f(x_{k_j}) \to \nabla f(\bar{x})$ $(**)$

$\left.\begin{matrix}(*)\\(**)\end{matrix}\right\} \Rightarrow \nabla f(\bar{x}) = 0 \Rightarrow \bar{x}$ is one minimizer

Step 5 $\forall\ k > k_j,\ 0 \le \|x_k - \bar{x}\| \le \|x_{k_j} - \bar{x}\|$

$\Rightarrow x_k \to \bar{x}$

II. Discrete Martingale

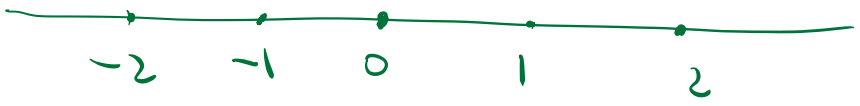
is a sequence of random variable $X_1, X_2, X_3, \ldots$

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

Supermartingale Convergence Theorem

Let X_k, Y_k be random variables satisfying

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

② $E[X_{k+1} \mid X_1, \cdots, X_k] \leq X_k - Y_k$

Then almost surely (with probility 1)

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

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

III. Randomized Coordinate Descent

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

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

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

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

Definition of Expectation & more:

① 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
$$\begin{cases} E(X) = x_1 \cdot p_1 + x_2 \cdot p_2 + \cdots + x_k p_k \to \text{convex combination} \\ E(f(x)) = \sum_{i=1}^{k} f(x_i) p_i \end{cases}$$

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

② 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 $X_{k+1} = T_{i(k)}(X_k)$

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

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

X_{N+1} is a function of these $N+1$ random variables

$E(X_{N+1})$ denotes expectation w.r.t. $N+1$ i.i.d. random variables

③ 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 $Y = y_j$

$$P(X = x_i \mid Y = y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}$$ ← 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 { a function of y / a random variable

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

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

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

For $f(x, y)$, $E(X \mid Y) f \triangleq \sum_{i=1}^{k} f(x_i, y) P(X = x_i \mid Y = y)$

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

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

Conditioned on the past random variables

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

$$X_{k+1} = X_k - \eta\, \nabla f(X_k)^{i(k)}$$ X_k depends on $i(0),\cdots,i(k-1)$ but NOT $i(k)$

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

($E_k(X_k)$: random variable; X_k)

② Let X be something depending on all $i(k)$

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

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

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

③ $$E_k\left[\nabla f^{i(k)}(X_k)\right] = \frac{1}{n}\nabla f(X_k)$$

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

$$\nabla f^{i(k)} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \nabla f(X_k)_{i(k)} \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

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

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

$$= \frac{1}{n}\sum_{j=1}^{n} \nabla f^j(X_k) = \frac{1}{n}\nabla f(X_k)$$

$$\|\nabla f^{i(k)}(x_k)\|^2 = \left([\nabla f(x_k)]_{i(k)}\right)^2$$

$$E_k \left\|[\nabla f(x_k)]^{i(k)}\right\|^2 = \sum_{j=1}^{n} P[i(k)=j] \cdot \left\|\nabla f^j(x_k)\right\|^2$$

$$= \frac{1}{n}\sum_{j=1}^{n}\left(\nabla f_j(x_k)\right)^2 = \frac{1}{n}\|\nabla f(x_k)\|^2$$

So we have $\begin{cases} E_k[\nabla f^{i(k)}(x_k)] = \frac{1}{n}\nabla f(x_k) \\ E_k\|\nabla f^{i(k)}(x_k)\|^2 = \frac{1}{n}\|\nabla f(x_k)\|^2 \end{cases}$

IV. Convergence of Randomized Coordinate Descent

Assume
① ∇f is L-cont.
② f is convex
③ there is at least one minimizer

Step 1 ∇f is L-cont., f is convex $\Rightarrow \|\nabla f(x)-\nabla f(y)\|^2 \le L\langle \nabla f(x)-\nabla f(y), x-y\rangle$

↓ Theorem 2.9 in Sec 2.3 in Typed Notes

$y = x_* \Rightarrow -L\langle \nabla f(x)-\nabla f(x_*), x-x_*\rangle \le -\|\nabla f(x)-\nabla f(x_*)\|^2$

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

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

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

$$= \|x_k - x_*\|^2 - 2\eta \langle x_k - x_*, \nabla f^{i(k)}(x_k)\rangle + \eta^2 \|\nabla f^{i(k)}(x_k)\|^2$$

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

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

$$E_k\|x_{k+1}-x_*\|^2 = E_k\|x_k - x_*\|^2 - 2\eta E_k\langle x_k - x_*, \nabla f^{i(k)}(x_k)\rangle + \eta^2 E_k\|\nabla f^{i(k)}(x_k)\|^2$$

For E_k, x_k is constant

$$= \|x_k - x_*\|^2 - 2\eta \langle x_k - x_*, E_k \nabla f^{i(k)}(x_k)\rangle + \eta^2 E_k\|\nabla f^{i(k)}(x_k)\|^2$$

$$= \|x_k - x_*\|^2 - 2\eta \langle x_k - x_*, \frac{1}{n}\nabla f(x_k)\rangle + \eta^2 \frac{1}{n}\|\nabla f(x_k)\|^2$$

$$\le \|x_k - x_*\|^2 - 2\eta \frac{1}{n}\cdot\frac{1}{L}\cdot\|\nabla f(x_k)\|^2 + \eta^2 \frac{1}{n}\|\nabla f(x_k)\|^2$$

$$= \|x_k - x_*\|^2 + \frac{1}{n}\eta\left(\eta - \frac{2}{L}\right)\|\nabla f(x_k)\|^2$$

Take E for both sides (Law of Total Expectation)

$$E\left[E_k\|x_{k+1}-x_*\|^2\right] = E\left[\|x_{k+1}-x_*\|^2\right]$$

$$\Rightarrow E\|x_{k+1}-x_*\|^2 \le E\|x_k - x_*\|^2 + \frac{1}{n}\eta\left(\eta - \frac{2}{L}\right)E\|\nabla f(x_k)\|^2$$

$\propto \eta < \frac{2}{L} \Rightarrow E\|x_{k+1} - x_*\| \searrow$

Step 3 Let $\begin{cases} \omega = -\eta\left(\eta - \frac{2}{L}\right) > 0, \text{ then} \\ r_k^2 = \|x_k - x_*\|^2 \end{cases}$

$$E_k r_{k+1}^2 \le r_k^2 - \frac{1}{n} w \|\nabla f(x_k)\|^2$$

Supermartingale Convergence Theorem

Let X_k, Y_k be random variables satisfying

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

② $E[X_{k+1} \mid x_1, \dots, x_k] \le X_k - Y_k$

Then almost surely (with probility 1)

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

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

$\Rightarrow$ With probability 1

$\sum_{k=0}^{\infty} \|\nabla f(x_k)\|^2 < \infty \Rightarrow \|\nabla f(x_k)\| \to 0$ (*)

$\lim_{k\to\infty} \|x_k - x_*\|^2$ exists

with probability 1

for any minimizer x_*, $\left[\lim_{k\to\infty} \|x_k - x_*\| \text{ exists with prob. } 1\right]$

$\Rightarrow$ With prob. 1 [for any x_*, $\lim_{k\to\infty} \|x_k - x_*\|$ exists]

→ nontrivial, skipped, see book BS

Step 4 With probability 1

$\|x_{k+1} - x_*\|$ non-increasing $\Rightarrow \{x_k\}$ is bounded

$\Rightarrow$ a convergent subsequence $x_{k_j} \to \bar{x}$

Continuity of $\nabla f \Rightarrow \nabla f(x_{k_j}) \to \nabla f(\bar{x})$

(*) $\Rightarrow \nabla f(\bar{x}) = 0 \Rightarrow \bar{x}$ is one minimizer

Step 5 With probability 1

$$\forall\ k > k_j,\ 0 \le \|x_k - \bar{x}\| \le \|x_{k_j} - \bar{x}\|$$

$\Rightarrow x_k \to \bar{x}$

V. Convergence of Fixed Point Iteration

Lecture 21 on Oct 14

Theorem Assume

① T is θ-averaged with at least one fixed point

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

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

converges to one fixed point of $T(x)$

Same Proof !

Example: $S = I - \frac{2}{L}\nabla f$ is nonexpansive if $\begin{cases} f(x) \text{ is convex} \\ \nabla f \text{ is } L\text{-cont.} \end{cases}$

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

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

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