Markov Chain V.S. Martingale

- Martingale is a sequence $\{X_k\}_{k=1}^{\infty}$ satisfying
  ① X_k is a R.V. (random variable)
  ② $E(|X_k|) < +\infty$, $\forall k$
  ③ $E(X_{k+1} \mid X_k, \underbrace{X_{k-1}, \cdots, X_1}_{\text{forgetting history}}) = X_k$

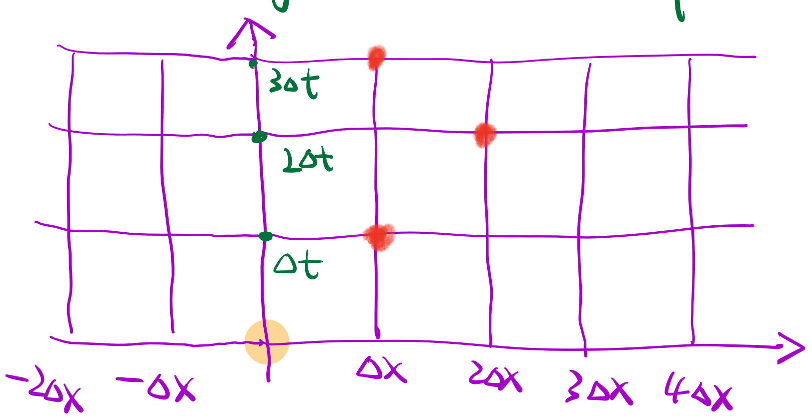
- Markov Chain is a sequence $\{X_k\}_{k=1}^{\infty}$ satisfying:
  ① X_k is a R.V. (random variable)
  ② $\mathrm{Prob}(X_{k+1} = x \mid X_k = x_k, \underbrace{X_{k-1} = x_{k-1}, \cdots, X_1 = x_1}_{\text{forgetting history}})$

$$= \mathrm{Prob}(X_{k+1} = x \mid X_k = x_k)$$ if both are well defined.

Random walk is both a martingale and a markov chain

Example: Random Walk

2D rectangular lattice defined by $(m\Delta x, n\Delta t)$



1) A particle starts at $x=0$ at time $t=0$

2) At time step n, the position is X_n

$$X_{n+1} = \begin{cases} X_n - \Delta x & , \text{ with probability } \frac{1}{2} \\ X_n + \Delta x & , \text{ with prob. } \frac{1}{2} \end{cases}$$

Example: Assume $Y_0, Y_1, Y_2, Y_3, \cdots$ are i.i.d. Gaussian $N(0,1)$

$$E(Y_i) = 0$$

Define $\{X_n\}_{n\geq 0}$ by

$X_{k+1} = X_k + Y_{k+1} \cdot X_0$ with X_0 independent of Y_i

Then $E(X_{k+1} \mid X_k, X_{k-1}, \cdots, X_0)$

$= E(X_k + Y_{k+1} X_0 \mid X_k, X_{k-1}, \cdots, X_0)$

$= X_k + X_0 E(Y_{k+1}) = X_k$

So $\{X_n\}_{n\geq 0}$ is a martingale

But $\{X_n\}_{n\geq 0}$ is not a Markov chain

because $P(X_{k+1} = x \mid X_k = x_k)$

$\neq P(X_{k+1} = x \mid X_k = x_k, \cdots, X_0 = x_0\}$

Stochastic Gradient Descent for $\min_{x \in \mathbb{R}^n} \frac{1}{N} \sum_{i=1}^{N} f_i(x)$

Example: Linear Regression for given data $(x_i, y_i)_{i=1}^{N}$

$\phi_j(x)$ $j = 1, 2, \cdots, n$ are some model basis

Want to solve the eqn in least square sense.

$$\underbrace{c_1 \phi_1(x_i) + c_2 \phi_2(x_i) + \cdots + c_n \phi_n(x_i)}_{g(x_i, c)} = y_i \quad , \quad i = 1, \cdots, N$$

$$\min_{c \in \mathbb{R}^n} \frac{1}{N} \sum_{i=1}^{N} \| g(x_i, c) - y_i \|^2$$

(Full batch) Gradient Descent:

$$x_{k+1} = x_k - \eta \left[\frac{1}{N} \sum_{i=1}^{N} \nabla f_i(x) \right]$$

Stochastic Gradient Descent:

$$x_{k+1} = x_k - \eta_k \nabla f_{i(k)}(x_k)$$

$i(k) \in \{1, \cdots, N\}$ are i.i.d. random variables with uniform distribution

$$f(x) = \sum_{j=1}^{n} f_j(x) = \sum_{i=1}^{N} F_i(x) \qquad \frac{n}{N} = m$$

$$F_i = \sum_{j = m \cdot (i-1)+1}^{m \cdot i} f_j(x)$$

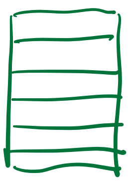
Two Randomized / Stochastic Methods

I. Randomized Coordinate Descent

II. Stochastic Gradient Descent

Example: $\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{2}\|Ax-b\|^2$ $\quad A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$

a_i^T is i-th row of A $\Rightarrow f(x) = \frac{1}{2}\|Ax-b\|^2$

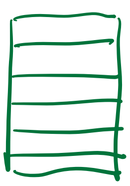
A= 

$$= \frac{1}{2}\sum_{i=1}^{m} |\langle a_i, x\rangle - b_i|^2$$

$$= \frac{1}{m}\left(\frac{1}{2}m\sum_{i=1}^{m} |\langle a_i, x\rangle - b_i|^2\right)$$

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

a_i^T is i-th row of A

A= 

$$= x_k - \eta A^T(Ax-b)$$

$= x_k - \eta$ 

$$= x_k - \eta \underbrace{\sum_{i=1}^{m} a_i(\langle a_i, x_k\rangle - b^i)}_{\nabla f(x)}$$

$$= \sum_{i=1}^{m} a_i(a_i^T x - b^i)$$

$$\nabla f = \begin{pmatrix} \nabla f^1 \\ \nabla f^2 \\ \vdots \\ \nabla f^n \end{pmatrix} \qquad \nabla f^{(j)} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \nabla f^j \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\nabla f(x) = \boxed{A^T}\left(\boxed{A}\,\underset{x}{\boxed{\,}} - \underset{b}{\boxed{\,}}\right)$$

j-th row of A^T

$$\nabla f(x)^{(j)} = \boxed{\begin{matrix} \text{———} \\ \text{———} \\ \text{———} \\ \text{———} \end{matrix}}\left(\boxed{A}\,\underset{x}{\boxed{\,}} - \underset{b}{\boxed{\,}}\right)$$

0 row

$$\sum_{i=1}^{m} a_i(a_i^T x_k - b^i)$$

Gradient Descent : $x_{k+1} = x_k - \eta \underbrace{\nabla f(x_k)}$

Coordinate Descent : $x_{k+1} = x_k - \eta \underbrace{\nabla f(x_k)^{(j)}}$

a sparse vector with only one nonzero entry

j-th entry of $\sum_{i=1}^{m} a_i(a_i^T x_k - b^i)$

SGD : $f = \frac{1}{m}\sum_{i=1}^{m} f_i$ $\quad f(x) = \frac{1}{m}\left(\frac{1}{2}m\sum_{i=1}^{m}|\langle a_i, x\rangle - b_i|^2\right)$

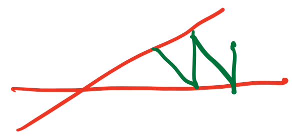
$$x_{k+1} = x_k - \eta \underbrace{\nabla f_i(x_k)}_{m\,a_i(a_i^T x_k - b^i)}$$

i can be taken sequentially or randomly or in a mini-batch

$$x_{k+1} = x_k - \eta \sum_{i \in S} \nabla f_i(x_k)$$

Kaczmarz Method for solving a least square
(1937)

$$\min_x \frac{1}{2}\|Ax-b\|^2 :$$



If we take some special η, SGD becomes

$$a_i^T x - b^i = 0, \quad i=1,\dots,m$$

$$x_{k+1} = x_k - \frac{(a_i^T x_k - b^i)}{\|a_i\|_2^2} a_i ,$$

which is the Kaczmarz Method.

a good reading choice

The normalization factor $\frac{1}{\|a_i\|_2^2}$ corresponds to Importance Sampling in SGD.

a good reading choice

Code demo for ① Coordinate Descent for $\begin{cases} \text{GD} \\ \text{Proximal Grad} \end{cases}$

② TV minimization