Lecture - 6

Def $\{S_n\}_{n\geq 1}$ is a martingale if

(i) $E|S_n| < \infty$

(ii) $E(S_{n+1} | S_1 \dots S_n) = S_n$

Ex 1: (Random Walk)

$X_i \to$ iid zero mean

$S_n = X_1 + X_2 + \dots + X_n$

$$E(S_{n+1} | S_1 \dots S_n) = E(X_{n+1} + S_n | S_1 \dots S_n)$$
$$= E(X_{n+1}) + S_n = S_n$$

Ex 2: $\{X_i\}_{i\geq 1} \to$ dependent r.v.s $\quad E(X_{i+1} | X_1 \dots X_i) = 0$

$S_n = X_1 + \dots + X_n$

$$E(S_{n+1} | S_1 \dots S_n) = E(X_{n+1} + S_n | S_1 \dots S_n)$$
$$= E(X_{n+1} | S_1 \dots S_n) + E(S_n | S_1 \dots S_n)$$
$$= E(X_{n+1} | X_1 \dots X_n) + S_n$$
$$= S_n$$

Ex 3: $S_n = X_1 \dots X_n$, $\{X_i\}$ iid, mean one

$$E(S_{n+1} | S_1 \dots S_n) = E(X_{n+1} S_n | S_1 \dots S_n)$$
$$= E(X_{n+1}) S_n = S_n$$

Question: ① $\{S_n\}$ martingale

$$E(S_{n+2} \mid S_1 \ldots S_n) = S_n$$

② $\{X_i\}$ iid

Is $\{X_i\}$ a martingale?

NO

Ex4: A stochastic gradient algorithm

$$x_{k+1} = x_k - a_k \big(\underbrace{\hat{\nabla} f(x_k)}_{\text{gradient estimate}} \big)$$

$$\boxed{x_{k+1} = x_k - a_k \big(\nabla f(x_k) + \underbrace{w_{k+1}}_{\text{noise}} \big)}$$

$$w_{k+1} = \hat{\nabla} f(x_k) - \nabla f(x_k)$$

Suppose $\boxed{E(\hat{\nabla} f(x_k) \mid x_1 \ldots x_k) = \nabla f(x_k)}$ unbiased (conditionally) gradient estimate

Then,

$$E(w_{k+1} \mid x_1 \ldots x_k)$$

$$= E(\hat{\nabla} f(x_k) \mid x_1, \ldots, x_k) - E[\nabla f(x_k) \mid x_1 \ldots x_k]$$

$$= \nabla f(x_k) - \nabla f(x_k) = 0$$

$\{w_k\} \to$ martingale difference sequence.

Side note:
$E[w_{k+1}] = 0$
In general, $E[w_{k+1} \mid x_k] = 0$ $\Rightarrow$ $E[w_{k+1}] = 0$
converse does not hold

In general,

$\{S_n\}_{n \geq 1}$ is a martingale difference sequence if

(i) $E|S_n| < \infty$

(ii) $E(S_{n+1} \mid S_0, S_1, \ldots S_n) = 0$

Example: $\{Z_n\}$ is a martingale.

Let $S_n = Z_n - Z_{n-1}$

Then, $\{S_n\}$ is a martingale difference sequence.

An application: Mean estimation

Consider a random variable (r.v.) Y with mean μ & finite variance, say σ^2.

Suppose we are given iid samples $Y_1, \ldots, Y_n$

Let x_n be the estimate of μ.

$$x_n = \frac{1}{n} \sum_{k=1}^{n} Y_k$$

$$x_{n+1} = \frac{1}{n+1} \sum_{k=1}^{n+1} Y_k$$

$$= \frac{n}{n+1} \left(\frac{1}{n} \sum_{k=1}^{n} Y_k \right) + \frac{1}{n+1} Y_{n+1}$$

$$x_{n+1} = \frac{n}{n+1} x_n + \frac{1}{n+1} Y_{n+1}$$ ← iterative scheme for updating sample mean

$$\boxed{x_{n+1} = x_n + \frac{1}{n+1} \left(Y_{n+1} - x_n \right)}$$

$$x_{n+1} = x_n + a_n \left(Y_{n+1} - x_n \right) \quad \text{— } (**)$$

step-size (aka learning rate)

where $a_n = \frac{1}{n+1}$

$(**)$ is a stochastic approximation algorithm (a more detailed intro later)

What does Strong Law of large numbers say about x_n?

$$x_n \longrightarrow \mu \text{ a.s. as } n \to \infty$$

with step size $a_n = \frac{1}{n+1}$

$$x_{n+1} = x_n + a_n (Y_{n+1} - x_n)$$

$$= x_n + a_n \left((\mu - x_n) + \underbrace{(Y_{n+1} - \mu)}_{W_{n+1}} \right)$$

$$E(W_{n+1} \mid x_1, \dots, x_n)$$

$$= E(W_{n+1} \mid Y_1, \dots, Y_n)$$

$$= E(Y_{n+1} \mid Y_1, \dots, Y_n) - \mu$$

$$= E[Y_{n+1}] - \mu = \mu - \mu = 0$$

Application: "Urn model"

Initially empty urn

Add red/blue ball randomly each time.

$$Y_{n+1} = \begin{cases} 1 & \text{if } (n+1)\text{th ball is red} \\ 0 & \text{else} \end{cases}$$

$$S_n = \sum_{k=1}^{n} Y_k \rightarrow \text{total \# of red balls}$$

$$x_n = \frac{S_n}{n} \rightarrow \text{fraction of red balls}$$

$$x_{n+1} = \frac{1}{n+1} \sum_{i=1}^{n+1} Y_k$$

$$= \left(1 - \frac{1}{n+1}\right) x_n + \frac{1}{n+1} Y_{n+1}$$

$$= x_n + \underbrace{a_n}_{= \frac{1}{n+1}} \left(Y_{n+1} - x_n \right)$$

Suppose the conditional probability that the "next" ball added at $(n+1)$, given the past, depends only on x_n, i.e.,

$$P(Y_{n+1}=1 \mid x_1, \dots, x_n) = p(x_n)$$

Then,

$$x_{n+1} = x_n + a_n\Big((p(x_n) - x_n) + \underbrace{(Y_{n+1} - p(x_n))}\Big)$$

$$\boxed{x_{n+1} = x_n + a_n\big((p(x_n) - x_n) + \underbrace{w_{n+1}}_{\text{"noise"}}\big)}$$

$$E(w_{n+1} \mid x_1, \dots, x_n)$$

$$= E(Y_{n+1} \mid x_1, \dots, x_n) - p(x_n)$$

$$= P(Y_{n+1}=1 \mid x_1, \dots, x_n) - p(x_n) = p(x_n) - p(x_n)$$

$$= 0$$

$\{w_{n+1}\} \to$ martingale difference sequence

Def $\{S_n\}_{n\geq 1}$ is a Sub-martingale if

(i) $E|S_n| < \infty$

(ii) $E(S_{n+1} \mid S_1, \dots, S_n) \geq S_n$

Def $\{S_n\}_{n\geq 1}$ is a Supermartingale if

(i) $E|S_n| < \infty$

(ii) $E(S_{n+1} \mid S_1, \dots, S_n) \leq S_n$

Ex: $\{X_i\}$ iid

$E X_i \geq 0 \Rightarrow \{X_i\}$ sub-martingale

$E X_i \leq 0 \Rightarrow \{X_i\}$ super-martingale

A popular ML algorithm for "training"

Want to solve $\min_x f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x)$

Common assumption: f_i smooth & f is convex/strongly-convex

Batch GD:

$$x_{n+1} = x_n - a_n \Bigg(\underbrace{\frac{1}{m}\sum_{i=1}^{m} \nabla f_i(x_n)}_{\text{computationally expensive for large } m} \Bigg) \leftarrow \text{Noiseless algorithm}$$

computationally expensive for large m (e.g. m → # training examples in a large dataset)

Alternative: SGD

Pick "i_n" uniformly at random in $\{1 \ldots m\}$

$$i_n = \begin{cases} 1 & \text{w.p. } 1/m \\ \vdots & \\ m & \text{— " — } 1/m \end{cases}$$

$$x_{n+1} = x_n - a_n \nabla f_{i_n}(x_n) \longrightarrow \text{One sample gradient used for update}$$

Rewriting the update rule,

$$x_{n+1} = x_n - a_n\left(\frac{1}{m}\sum_{i=1}^{n} \nabla f_i(x_n)\right) - a_n\left(\nabla f_{i_n}(x_n) - \frac{1}{m}\sum_{i=1} \nabla f_i(x_n)\right)$$

$$\boxed{x_{n+1} = x_n - a_n\left(\frac{1}{m}\sum_{i=1}^{n} \nabla f_i(x_n) + w_{n+1}\right)}$$

$\longrightarrow$ A stochastic approximation algorithm

Note,

$$\boxed{E\left(\nabla f_{i_n}(x_n) \,\middle|\, x_1 \ldots x_n\right) = \frac{1}{m}\sum_{i=1}^{m} \nabla f_i(x_n)}$$

Thus, $E\left(w_{n+1} \,\middle|\, x_1 \ldots x_n\right) = 0$

or, $\{w_n\}$ is a martingale difference sequence

Bottomline! Convergence of stochastic approximation algorithms is tied to whether the effect of underlying noise (martingale diff) can be ignored in the long run.

Convergence of martingales

A few useful facts:

(I) $A_1, A_2, \ldots$ increasing sequence of events

$$A_1 \subseteq A_2 \subseteq A_3 \subseteq \ldots$$

Let $A = \bigcup_{i=1}^{\infty} A_i = \lim_{i\to\infty} A_i$

Then, $\boxed{P(A) = \lim_{i\to\infty} P(A_i)}$

Proof:-

$$A = \underbrace{A_1 \cup (A_2 \setminus A_1) \cup (A_3 \setminus A_2) \ldots}_{\text{disjoint union}}$$

$$P(A) = P(A_1) + \sum_{i=1}^{\infty} P(A_{i+1} \setminus A_i)$$

$$= P(A_1) + \lim_{n\to\infty} \sum_{i=1}^{n-1} \left(P(A_{i+1}) - P(A_i)\right)$$

$$= \lim_{n\to\infty} P(A_n)$$

$\blacksquare$

(II) If $B_1, B_2, \ldots$ decreasing sequence of events, i.e.,

$$B_1 \supseteq B_2 \supseteq B_3 \supseteq \cdots$$

and $B = \bigcap_{i=1}^{\infty} B_i = \lim_{i \to \infty} B_i$

Then, $P(B) = \lim_{i \to \infty} P(B_i)$

Hint: Take complements & use the claim above for unions.

(III) Cauchy convergence

Real case: $x_1, x_2, x_3, \ldots$ Cauchy if
$\forall \epsilon > 0, \ \exists N > 0$ s.t. $|x_n - x_m| < \epsilon \quad \forall m, n > N$

Fact: Every Cauchy sequence converges to a limit, i.e.,
$x_1, x_2, x_3 \ldots$ Cauchy $\Rightarrow \exists x^*$ s.t. $x_n \to x^*$ as $n \to \infty$
(or)
$\forall \epsilon > 0, \exists N$ s.t. $|x_n - x^*| < \epsilon \ \forall n > N$.

Extending Cauchy convergence to "probabilistic" case!

$\{X_n\}_{n\geq 1}$ r.v.s is Cauchy convergent if

the set of sample points ω for which $\{X_n(\omega)\}_{n\geq 1}$ is

Cauchy convergent has probability one

(or)

$$P(\{\omega \in \Omega \mid X_m(\omega) - X_n(\omega) \to 0 \text{ as } m,n\to\infty\}) = 1$$

Fact: $\{X_n(\omega)\}$ converges $\iff$ $\{X_n(\omega)\}$ Cauchy convergent

$\{X_n\}_{n\geq 1}$ converges almost surely $\iff$ $\{X_n\}_{n\geq 1}$ Cauchy convergent

(iv) "Uncorrelated with the past"

$\{S_n\}_{n\geq 1}$ martingale i.e., $E|S_n| < \infty$, $E(S_{n+1}|S_1,\dots,S_n) = S_n$

Claim: $E(S_m(S_{m+n} - S_m)) = 0 \quad \forall\, m,n \geq 1$

Why? $E(S_m(S_{m+n} - S_m))$

$= E(E(S_m(S_{m+n} - S_m) | S_1,\dots,S_m))$

$= E(S_m\, E(S_{m+n} - S_m | S_1,\dots,S_m))$

$= 0$ since $E(S_{m+n}|S_1,\dots,S_m) = S_m$

Lecture - 8

Book: Grimmett & Stirzaker "Prob & RP".

Martingale convergence theorem

$\{S_n\}$ martingale satisfying $ES_n^2 < M < \infty$ for some M & $\forall n$

Then, $\exists$ a r.v. S s.t.

(i) $S_n \xrightarrow[a.s]{} S$ as $n \to \infty$.

(ii) $S_n \xrightarrow[L^2]{} S$ as $n \to \infty$ (mean-squared sense)

Before proving this, we establish a useful inequality.

Doob-Kolmogorov inequality:

$\{S_n\}$ martingale & $ES_n^2 < \infty$ $\forall n$. Then,

$$\boxed{P\left(\max_{i=1\ldots n} |S_i| \geq \epsilon\right) \leq \frac{ES_n^2}{\epsilon^2}}$$ for any $\epsilon > 0$.

Pf:

$A_k = \{|S_1| < \epsilon, \ldots, |S_k| < \epsilon\}$, $A_0 = \Omega$

$B_1 = \{|S_1| \geq \epsilon\} = A_0 \cap \{|S_1| \geq \epsilon\}$

$B_2 = \{|S_1| < \epsilon, |S_2| \geq \epsilon\} = A_1 \cap \{|S_2| \geq \epsilon\}$

$\vdots$

$B_k = A_{k-1} \cap \{|S_k| \geq \epsilon\}$

$$\underbrace{A_n \cup \left(\bigcup_{i=1}^{n} B_i \right)}_{\text{disjoint union}} = \Omega$$

$$E S_n^2 = \sum_{i=1}^{n} E\left(S_n^2 \, I(B_i)\right) + E\left(S_n^2 \, I(A_n)\right)$$

indicator

$$I(A) = \begin{cases} 1 & \text{if } A \text{ happens} \\ 0 & \text{else} \end{cases}$$

$$E S_n^2 \geq \sum_{i=1}^{n} E\left(S_n^2 \, I(B_i)\right)$$

$$E\left(S_n^2 \, I(B_i)\right) = E\left((S_n - S_i + S_i)^2 \, I(B_i) \right)$$

$$= E\left((S_n - S_i)^2 \, I(B_i) \right) + 2 E\left((S_n - S_i) S_i \, I(B_i) \right) + E\left(S_i^2 \, I(B_i) \right)$$

$$\overset{(*)}{\geq} \underbrace{2 E\left((S_n - S_i) S_i \, I(B_i) \right)}_{0} + E\left(S_i^2 \, I(B_i) \right)$$

$$\geq \epsilon^2 \, P(B_i) \qquad \text{since} \quad |S_i| \geq \epsilon \text{ on } B_i$$

$$E S_n^2 \geq \sum_{i=1}^{n} E\left(S_n^2 \, I(B_i)\right)$$

$$\geq \sum_{i=1}^{n} \epsilon^2 \, P(B_i)$$

why?!

$$\geq \epsilon^2 \, P\left(\max_{i=1 \dots n} |S_i| \geq \epsilon \right) \quad \blacksquare$$

Justification for (*):

$$E\left(\left(S_n - S_i\right) S_i \, I(B_i) \mid S_1, \ldots, S_i\right)$$

$$= I(B_i) \, E\left(\left(S_n - S_i\right) \mid S_1, \ldots, S_i\right) S_i$$

$$= 0$$

Theorem:

$\{S_n\}$ martingale satisfying $E S_n^2 < M < \infty$ for some M & $\forall n$

Then, $\exists$ a r.v. S s.t.

(i) $S_n \xrightarrow[a.s]{} S$ as $n \to \infty$.

(ii) $S_n \xrightarrow[L^2]{} S$ as $n \to \infty$ (mean-squared sense)

Pf: S_m and $(S_{m+n} - S_m)$ are uncorrelated $m, n \geq 1$

Since $E\left(S_m \left(S_{m+n} - S_m\right)\right) = 0$ $\leftarrow$ fact (IV)

$$E\left(S_{m+n}^2\right) = E S_m^2 + E\left(\left(S_{m+n} - S_m\right)^2\right)$$

$$\geq E S_m^2$$

$\{E S_n^2\}$ non-decreasing, bounded above (by assumption)

Choosing M s.t.

$$E S_n^2 \uparrow M \quad \text{as} \quad n \to \infty \quad \text{——} \quad (*)$$

Enough to show:

$\{S_n(\omega)\}_{n \geq 1}$ is Cauchy convergent

as it would imply a.s. convergence.

Let $C = \{\omega \mid \{S_n(\omega)\} \text{ is Cauchy convergent}\}$

Set C can be written as

$$C = \{\omega \mid \forall \epsilon > 0, \exists m \text{ s.t. } |S_{m+i}(\omega) - S_{m+j}(\omega)| < \epsilon \quad \forall i, j \geq 1\}$$

If $|S_{m+i} - S_m| < \epsilon$ & $|S_{m+j} - S_m| < \epsilon$

then $|S_{m+i} - S_{m+j}| < 2\epsilon$ triangle Inequality

So,

$$C = \{\omega \mid \forall \epsilon > 0, \exists m \text{ s.t. } |S_{m+i}(\omega) - S_m(\omega)| < \epsilon \quad \forall i \geq 1\}$$

$$= \bigcap_{\substack{\epsilon > 0 \\ (\text{rational})}} \bigcup_{m} \{|S_{m+i} - S_m| < \epsilon, \forall i \geq 1\}$$

$$C^c = \bigcup_{\epsilon > 0} \bigcap_{m \geq 1} \{|S_{m+i} - S_m| \geq \epsilon, \text{ for some } i \geq 1\}$$

Let $A_m(\epsilon) = \{ |S_{m+i} - S_m| \geq \epsilon \text{ for some } i \geq 1 \}$

Then, $C^c = \bigcup_{\epsilon > 0} \bigcap_{m \geq 1} A_m(\epsilon)$

If $\epsilon \geq \epsilon'$, $A_m(\epsilon) \subseteq A_m(\epsilon')$

Want $P(C^c) = 0$ Note: $0 \leq \lim_{\epsilon \downarrow 0} P\left(\bigcap_{m \geq 1} A_m(\epsilon)\right)$

$$\leq \lim_{\epsilon \downarrow 0} \lim_{m \to \infty} P(A_m(\epsilon))$$

So, if $\lim_{m \to \infty} P(A_m(\epsilon)) = 0$ for any $\epsilon > 0$, then $P(C^c) = 0$

Let $Y_n = S_{m+n} - S_m$, m fixed.

Is $\{Y_n\}$ a martingale?

Please check $E[Y_{n+1} | Y_1, \ldots, Y_n] = Y_n$

Doob Kolmogorov ineq for $\{Y_i\}$

$$P(|Y_i| \geq \epsilon \text{ for some } 1 \leq i \leq n) \leq \frac{E Y_n^2}{\epsilon^2}$$

$$P(|S_{m+i} - S_m| \geq \epsilon \text{ for some } i) \leq \frac{E(S_{m+n} - S_m)^2}{\epsilon^2}$$

$$0 \leq P(A_m(\epsilon)) \leq \frac{E S_{m+n}^2 + E S_m^2 - 2E(S_{m+n} S_m)}{\epsilon^2}$$

$$E(S_{m+n} S_m) = E\left(E\left[S_{m+n} S_m \mid S_1, \ldots, S_m\right]\right)$$

$$= E S_m^2$$

$$0 \le P(A_m(\epsilon)) \le \frac{E S_{m+n}^2 - E S_m^2}{\epsilon^2}$$

$$\xrightarrow{n \to \infty} \frac{M - E S_m^2}{\epsilon^2} \quad (\text{why}(?))$$

As $m \to \infty$

$$\lim_{m \to \infty} P(A_m(\epsilon)) = 0 \quad \text{since } E S_m^2 \uparrow M$$

So, $P(C^c) = 0 \quad P(C) = 1$

$\{S_n\}$ is Cauchy convergent

$\Rightarrow \exists$ r.v. S s.t. $S_n \longrightarrow S$ a.s. as $n \to \infty$

Mean-square convergence:

Fatou's lemma: If $\{X_n\}$ s.t. $X_n \geq 0, \forall n$, then

$$E\left(\liminf_{n\to\infty} X_n\right) \leq \liminf_{n\to\infty} E X_n$$

Background

$\{x_n\}_{n\geq 1}$

$$\liminf_{n\to\infty} x_n = \lim_{n\to\infty}\left(\inf_{m\geq n} x_m\right)$$

$$\limsup_{n\to\infty} x_n = \lim_{n\to\infty}\left(\sup_{m\geq n} x_m\right)$$

e.g. $x_n = (-1)^n$

$\liminf x_n = -1$

$\limsup x_n = +1$

If limit exists, then $\liminf = \limsup$

Back to martingale conv. in mean square

Want $S_n \xrightarrow{L^2} S$ or $E(S_n - S)^2 \xrightarrow[n\to\infty]{} 0$

$$E\left((S_n - S)^2\right)$$

$$= E\left(\liminf_{m\to\infty} (S_n - S_m)^2\right) \quad \text{— ①}$$

$$\leq \liminf_{m\to\infty} E\left((S_n - S_m)^2\right) \quad \text{— ②}$$

$$= M - E S_n^2 \xrightarrow{n\to\infty} 0$$

L^2 conv used "a.s. convergence" + Fatou's lemma

$\Rightarrow$ $E(S_n - S)^2 \xrightarrow{n\to\infty} 0$ or $S_n \xrightarrow{L^2} S$ $\blacksquare$

Justification for ①:

"n is fixed"

$$E\left(\lim_{m\to\infty} (S_n - S_m)^2\right)$$

$$= E\left(\lim_{m\to\infty} \left(S_n^2 + S_m^2 - 2S_m S_n\right)\right)$$

$$= E\left(S_n^2 + S^2 - 2S_n S\right)$$

$$= E\left((S_n - S)^2\right)$$

Justification for ②:

$$\lim_{m\to\infty} E\left((S_n - S_m)^2\right) = \lim_{m\to\infty} \left(E S_n^2 + E S_m^2 - 2E(S_n S_m)\right)$$

$$= \lim_{m\to\infty} \left[E S_m^2 - E S_n^2\right]$$

[Check $E(S_n S_m) = E S_n^2$]

$$= M - E S_n^2$$