Lecture-10

"Kushner-Clark lemma & applications"

(1) $x_{n+1} = x_n + a(n)\left(h(x_n) + M_{n+1}\right)$

$a(n)$: stepsize h: want to find $h(x)=0$ M_{n+1}: noise

(2) $x_{n+1} = x_n + a(n)\left(h(x_n) + M_{n+1} + \beta_n\right)$

β_n: additional error

Assumptions:-

ODE $\dot{x}(t) = h(x(t))$

(A1) $h : \mathbb{R}^d \to \mathbb{R}$ Lipschitz

(A2) $\{\beta_n\}$ bounded (random) seq s.t. $\beta_n \to 0$

(e.g. $x_{n+1} = x_n + a(n)\left(-\nabla f(x_n) + M_{n+1} + C_1 \delta_n^2\right)$)

(A3) $a(n) \to 0$ and $\sum_n a(n) = \infty$

(A4) $\{M_{n+1}\}$ seq s.t. $\forall \epsilon > 0$

$$\boxed{\lim_{n\to\infty} P\left(\sup_{m \geq n} \left\| \sum_{i=n}^{m} a(i) M_{i+1} \right\| > \epsilon\right) = 0}$$

Note: $\{M_{n+1}\}$ is not assumed to be a martingale diff.

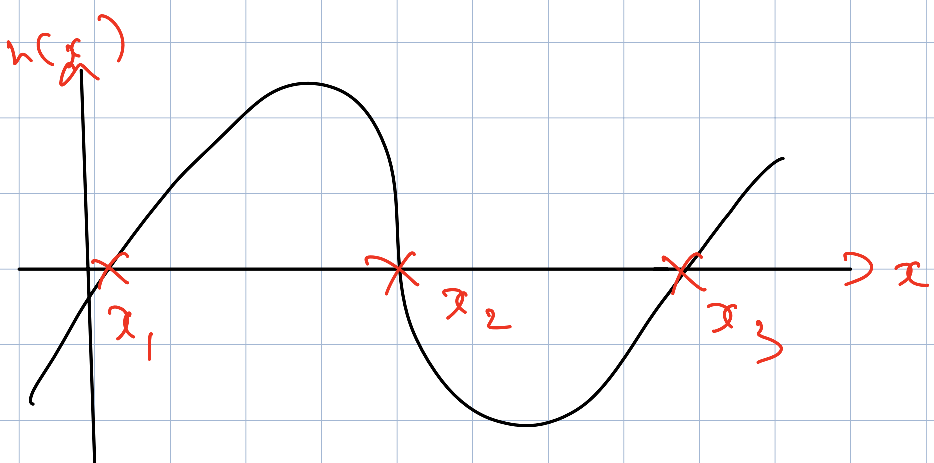
(A5) Bounded iterates: $\sup_n \|x_n\| < \infty$

Kushner-Clark lemma

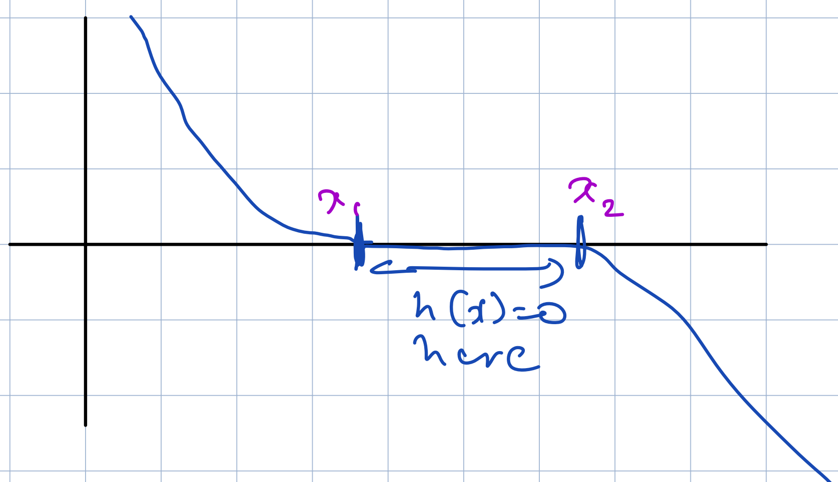
Under (A1) - (A5),

$$x_n \longrightarrow \{x^* \mid h(x^*)=0\} \text{ a.s. as } n \to \infty$$

Convergence to a set



K-C lemma says conv to one of these points $\{x_1, x_2, x_3\}$



K-C lemma says conv to $[x_1, x_2]$

Case of a MDS $\{M_{n+1}\}$: $\mathcal{F}_n = \sigma(x_1, \dots x_n)$

$$\boxed{\begin{aligned} &\text{A6} \quad E[M_{n+1} \mid \mathcal{F}_n) = 0 \\ &\qquad E[\|M_{n+1}\|^2 \mid \mathcal{F}_n] \leq C(1 + \|x_n\|^2) \end{aligned}}$$

$$\boxed{(\text{A3}') \qquad \sum a(n) = \infty , \quad \sum a(n)^2 < \infty}$$

K-C lemma variant:

Under A1, A2, A3', A5, A6,

$$x_n \to \{x^* \mid h(x^*) = 0\} \text{ a.s. as } n \to \infty$$

Background "Doob's martingale inequality"

$\{W_m\}$ martingale

$$(*) \quad P\left(\sup_{m \geq 0} \|W_m\| \geq \epsilon\right) \leq \frac{1}{\epsilon^2} \lim_{n \to \infty} E\|W_m\|^2$$

$$\sum_{n=k}^{i} a(n) M_{n+1} \quad \leftarrow \text{martingale}$$

Apply Doob's ineq. to

$$\lim_{k\to\infty} P\left(\sup_{l\geq k}\left\|\sum_{n=k}^{l} a(n) M_{n+1}\right\| \geq \epsilon\right)$$

$$\leq \frac{1}{\epsilon^2}\lim_{k\to\infty}\sum_{n=k}^{\infty} a(n)^2\, E\|M_{n+1}\|^2$$

$$\leq \frac{\text{const}}{\epsilon^2}\underbrace{\lim_{k\to\infty}\sum_{n=k}^{\infty} a(n)^2}_{\to 0 \text{ since } \sum a(n)^2<\infty}$$

$$= 0$$

So, (A4) of K-C lemma is satisfied

The claim follows

Coming next!

Stochastic gradient algorithm

$$x_{n+1} = x_n - a(n)\,\hat{\nabla} f(x_n)$$

$$E(\hat{\nabla} f(x_n) \mid \mathcal{F}_n) = \nabla f(x_n)$$

"unbiased gradient oracle"

$$x \longrightarrow \boxed{} \longrightarrow \nabla f(x) + \underset{\text{noise}}{\underset{\uparrow}{\epsilon}}$$

$$E(\hat{\nabla} f(x_n) \mid \mathcal{F}_n) \neq \nabla f(x_n)$$

$$E(\hat{\nabla} f(x_n) \mid \mathcal{F}_n) - \nabla f(x_n) \leq C\delta_n^2$$

"Zeroth-order optimization"

Biased gradient oracle

$$x, \delta \longrightarrow \boxed{} \longrightarrow \nabla f(x) + \epsilon_{\text{noise}} + \underset{\text{bias}}{\underset{\downarrow}{\delta}}$$

Recall:

Simultaneous perturbation gradient estimation

$$\hat{\nabla}_i f(x_n) = \frac{f(x_n + \delta_n \Delta_n) - f(x_n - \delta_n \Delta_n)}{2\delta_n \Delta_n^i}$$

$\Delta_n^i \rightarrow$ ith coord

$$\Delta_n = (\Delta_n^1, \ldots, \Delta_n^d) \qquad \Delta_n^i \rightarrow \text{Rademacher}$$

$$E\left(\hat{\nabla} f(x_n) \,\middle|\, \mathcal{F}_n\right) - \nabla f(x_n) \le C\delta_n^2$$

(assuming f is smooth, etc)

$$\hat{\nabla}_i f(x_n) = \frac{f(x_n+\delta_n\Delta_n)+\xi_n^+ - \left(f(x_n-\delta_n\Delta_n)+\xi_n^-\right)}{2\delta_n\Delta_n^i}$$

$\Delta_n^i \rightarrow$ ith coord

Assume $E\left(\xi_n^+ - \xi_n^- \,\middle|\, \mathcal{F}_n\right) = 0$

$$x,\delta \longrightarrow \boxed{\text{Biased gradient oracle}} \longrightarrow \hat{\nabla} f(x)$$

$$E\left[\hat{\nabla} f(x) \,\middle|\, x\right] - \nabla f(x) \le C_1\delta^2$$

$$E\left[\left\|\hat{\nabla} f(x) - E\left(\hat{\nabla} f(x) \,\middle|\, x\right)\right\|^2\right] \le \frac{C_2}{\delta^2}$$

K-C lemma variant (MDS Core)

$$x_{n+1} = x_n + a(n)\left(h(x_n) + M_{n+1} + \beta_n\right)$$

(B1) $h : \mathbb{R}^d \to \mathbb{R}$ Lipschitz

(B2) $\{\beta_n\}$ bounded (random) seq s.t. $\beta_n \to 0$

(B3) $E[M_{n+1} \mid \mathcal{F}_n] = 0$

$$E[\|M_{n+1}\|^2 \mid \mathcal{F}_n] \le C(1 + \|x_n\|^2)$$

(B4) $\sum a(n) = \infty$, $\sum a(n)^2 < \infty$

(B5) Bounded iterates: $\sup_n \|x_n\| < \infty$

Kushner-Clark lemma

Under (B1) - (B5),

$$x_n \longrightarrow \{x^* \mid h(x^*) = 0\} \text{ a.s. as } n \to \infty$$

Sto-grad algo with unbiased gradient information!

$$x_{n+1} = x_n + a(n)\left(-\hat{\nabla} f(x_n)\right)$$
$$= x_n + a(n)\left(-\nabla f(x_n) + M_{n+1}\right)$$

(C1) f is L-smooth ($\Leftrightarrow$ ∇f is L-Lipschitz)

(C2) $E(M_{n+1} \mid \mathcal{F}_n) = 0$ (or $E(\hat{\nabla} f(x_n) \mid \mathcal{F}_n) = \nabla f(x_n)$)

$$E\left(\left\| \hat{\nabla} f(x_n) - E(\hat{\nabla} f(x_n) \mid \mathcal{F}_n) \right\|^2\right) \leq \sigma^2$$

(C3) $\sum a(n) = \infty$, $\sum a(n)^2 < \infty$

(C4) $\sup_n \|x_n\| < \infty$ a.s.

Claim: Under (C1)-(C4),

$$x_n \to \{x^* \mid \nabla f(x^*) = 0\} \quad \text{a.s as } n \to \infty$$

Proof:-

1. (C1 $\Rightarrow$ B1) "∇f Lipschitz"
2. $\beta_n = 0$ (B2) trivial
3. (C2) $\Rightarrow$ (B3) $\{M_{n+1}\}$
4. (C3) $\Rightarrow$ (B4) stepsizes
5. (C4) $\Rightarrow$ (B5) Bounded iterates

Hence proved. $\blacksquare$

H.W. Use K-C lemma to establish asymp. convergence of a stochastic fixed point iteration

$$x_{n+1} = x_n + a(n)\left(f(x_n) - x_n + M_{n+1} \right)$$

"Make reasonable assumptions" for K-C lemma invocation.

e.g. $\quad \|f(x) - f(y)\| \leq \alpha \|x - y\|$

$$0 < \alpha < 1$$

Sto-grad algo with biased gradient info:

$$x_{n+1} = x_n + a(n)\left(-\hat{\nabla} f(x_n)\right) \quad \text{—} (**)$$

$$x_n, \delta_n \longrightarrow \boxed{\text{BGO}} \longrightarrow \hat{\nabla} f(x_n)$$

(D1) $\quad \left\| E\left(\hat{\nabla} f(x_n) \mid \mathcal{F}_n\right) - \nabla f(x_n) \right\| \leq C_1 \delta_n^2$

$$E \left\| \hat{\nabla} f(x_n) - E\left(\hat{\nabla} f(x_n) \mid \mathcal{F}_n\right) \right\|^2 \leq \frac{C_2}{\delta_n^2}$$

bias variance tradeoff

for some constants C_1, C_2; $\forall n \geq 1$

Rewriting (**)

$$x_{n+1} = x_n + a(n)\left(-\hat{\nabla} f(x_n)\right)$$

$$x_{n+1} = x_n + a(n)\left(-\nabla f(x_n) + M_{n+1} + \beta_n\right)$$

So that $E(M_{n+1} \mid \mathcal{F}_n) = 0$ &

$$\beta_n \le C_1 \delta_n^2$$

Side-note: Unbiased Case

$$x_{n+1} = x_n + a(n)\left(-\hat{\nabla} f(x_n)\right)$$

$$= x_n + a(n)\left(-\nabla f(x_n) - \underbrace{\left(\hat{\nabla} f(x_n) - \nabla f(x_n)\right)}_{M_{n+1}}\right)$$

$$E(M_{n+1} \mid \mathcal{F}_n) = 0$$

BGO Core:

$$x_{n+1} = x_n - a(n)\left(\hat{\nabla} f(x_n)\right)$$

$$x_{n+1} = x_n - a(n)\left[\hat{\nabla} f(x_n) - E\left(\hat{\nabla} f(x_n) \mid \mathcal{F}_n\right) + E\left(\hat{\nabla} f(x_n) \mid \mathcal{F}_n\right) - \nabla f(x_n) + \nabla f(x_n) \right]$$

$$= x_n - a(n)\left(\nabla f(x_n) + M_{n+1} + B_n\right)$$

Lecture-11

Today → ① Asymptotic analysis of 2DO-SGD

② Variants of zeroth-order gradient estimation

$$x_{n+1} = x_n - a(n)\left[\hat{\nabla} f(x_n) - E(\hat{\nabla} f(x_n) | \mathcal{F}_n) + E(\hat{\nabla} f(x_n) | \mathcal{F}_n) - \nabla f(x_n) + \nabla f(x_n)\right]$$

$$= x_n - a(n)\left(\nabla f(x_n) + M_{n+1} + B_n\right)$$

(D1) + the following assumptions:

(D2) $a(n), \delta_n \to 0$ as $n \to \infty$

$$\sum_n a(n) = \infty$$

$$\sum_n \left(\frac{a(n)}{\delta_n}\right)^2 < \infty$$

H.W.
Suppose
$a(n) = \frac{1}{n^b}$
$\delta_n = \frac{1}{n^c}$
Under what choice for b, c is (D2) satisfied?

(D3) f is L-smooth

(D4) $E\|M_{n+1}\|^2 \le \frac{C}{\delta_n^2}, \forall n$

Claim: Under some assumptions,

$$x_n \longrightarrow \{ x^* \mid \nabla f(x^*) = 0 \} \quad \text{a.s as } n \to \infty$$

Pf:

① f is L-smooth ⇒ (B1) of K-C lemma satisfied

② $\beta_n = O(\delta_n^2)$ & by assumption (D2), $\delta_n \to 0$. So, $\beta_n \to 0$ ⇒ (B2) of K-C lemma satisfied.

③ K-C lemma requires

$$\boxed{\lim_{n\to\infty} P\Big(\sup_{m \geq n} \Big\| \underbrace{\sum_{i=n}^{m} a(i) M_{i+1}}_{\text{denote this as } W_m} \Big\| \geq \epsilon \Big) = 0} \quad (**)$$

$\{W_m\}$ is a martingale

$\{W_m\}$ martingale

(*) $$P\left(\sup_{m \geq 0} \|W_m\| \geq \epsilon\right) \leq \frac{1}{\epsilon^2} \lim_{n \to \infty} E\|W_m\|^2$$

Apply this to $W_m = \sum_{i=n}^{m} a(i) M_{i+1}$

$$P\left(\sup_{m \geq n} \left\|\sum_{i=n}^{m} a(i) M_{i+1}\right\| \geq \epsilon\right)$$

$$\leq \frac{1}{\epsilon^2} E\left\|\sum_{i=n}^{\infty} a(i) M_{i+1}\right\|^2$$

$$\leq \frac{1}{\epsilon^2} \sum_{i=n}^{\infty} a(i)^2 \, E\|M_{i+1}\|^2$$

(Used $E[M_m M_n]$ for $m<n$

$= E[M_m E(M_n | \mathcal{F}_m)]$

$= 0$)

using (D4)

$$\leq \frac{1}{\epsilon^2} \sum_{i=n}^{\infty} \frac{a(i)^2}{\delta_i^2}$$

Since $\sum_{i} \frac{a(i)^2}{\delta_i^2} < \infty$

$$\lim_{n\to\infty} \sum_{i=n}^{\infty} \frac{a(i)^2}{\delta_i^2} = 0$$

So, K-C lemma condition $(\cancel{\ast\ast})$ holds.

(T4) Since $\sup_n \|x_n\| < \infty$,

last condition of K-C lemma holds.

Hence, the claim follows. ∎

Why is assumption (D4) reasonable:

For $i = 1, \ldots d$,

$$\hat{\nabla}_i f(x_n) = \frac{\left(f(x_n + \delta_n \Delta_n) + \xi^+\right) - \left(f(x_n - \delta \Delta_n) + \xi^-\right)}{2\delta_n \Delta_n^i}$$

SPSA gradient estimate

$\Delta_n =$ vector of Rademacher r.v.s.

$$E \, \|\hat{\nabla} f(x_n) - E(\hat{\nabla} f(x_n) \mid \mathcal{F}_n)\|^2$$

$$\leq E \, \|\hat{\nabla} f(x_n)\|^2 \quad \text{———} \quad (\cancel{***})$$

$$\hat{\nabla}_i f(x_n)^2$$

$$= \left[\frac{(f(x^+) + \xi^+) - (f(x^-) + \xi^-)}{2 \delta \Delta_i} \right]^2$$

If $\quad f(x^+)^2, f(x^-)^2 \leq C_2$

$$E(\xi^+)^2 \leq C_3, \quad E(\xi^-)^2 \leq C_4$$

$$E \, \nabla_i f(x)^2 \leq \frac{\text{Const}}{\delta_n^2}$$

So, avg$(\cancel{x}\ x\ x)$ depends on d

$$\mathbb{E}\,\|\hat{\nabla} f(x_n)\|^2 \le \frac{\text{Const}}{\delta_n^2}$$