Lecture - 8 (contd)

# INTRODUCTION TO STOCHASTIC APPROXIMATION

e.g. 1 find maximizer

$\max_x f(x)$

$\nabla f(x) = 0$ or $h = \nabla f$

2 $f(x) = x \leftarrow$ fixed point

$h(x) = f(x) - x$

Problem: Solve $\boxed{h(x) = 0}$ given noisy measurements of $h$, i.e., given access to a black box that, on input $x \in \mathcal{R}^d$, gives as output $h(x)+$ noise.

$x \to \boxed{\text{Box}} \to$ "$h(x)$+noise"

**Robbins-Monro algorithm:** Starting with $x_0 \in \mathcal{R}^d$, do:

Stochastic root finding

$$x(n+1) = x(n) + \underbrace{a(n)}_{\text{stepsize}}[h(x(n)) + M(n+1)], \; n \geq 0.$$

Here the stepsize sequence (or 'learning parameter') $\{a(n)\}$ satisfies: $a(n) \geq 0$ and

$$\sum_n a(n) = \infty, \; \sum_{\cancel{N}\, n} a(n)^2 < \infty.$$

($\Longrightarrow$ slow decrease to zero, e.g., $\frac{1}{n}$, $\frac{1}{n \log n}$, $\frac{1}{n^{2/3}}$ etc.).

(A 0) $\sum a(n) = \infty$, $\sum a(n)^2 < \infty$

(A 1) $h : \mathcal{R}^d \mapsto \mathcal{R}^d$ is Lipschitz: $\|h(x) - h(y)\| \leq L\|x - y\|$ for $x, y \in \mathcal{R}^d$.

Euclidean norm

(A 2) $\{M(n)\}$ a square-integrable martingale difference sequence, i.e., for

$$\mathcal{F}_n := \overbrace{\sigma(x_0, M_m, m \leq n)}^{\text{information up to time "}n\text{"}}, n \geq 0,$$

we have

$$E\left[\|M(n)\|^2\right] < \infty$$

and in addition, it is 'uncorrelated with past', i.e., $M_i(\cdot)$ is a martingale difference

$$E[M_i(n+1)|\mathcal{F}_n] = 0 \ \forall \ i.$$

. (Equivalently,

$$E[M_i(n+1)|x_0, M_m, m \le n] = 0 \ \forall \ i.)$$

. Thus

$$\begin{aligned} & E[M_i(n+1)f(x_0, M_1, \cdots, M_n)] \\ &= E[E[M_i(n+1)f(x_0, M_1, \cdots, M_n)|x_0, M_m, m \le n]] \\ &= E[E[M_i(n+1)|x_0, M_m, m \le n]f(x_0, M_1, \cdots, M_n)] \\ &= 0. \end{aligned}$$

Hence ‘uncorrelated with past’.

Furthermore, we assume that for some $K > 0$, Part of

$$\boxed{E\left[\|M(n+1)\|^2|\mathcal{F}_n\right] \leq K\left(1+\|x(n)\|^2\right)} \ \forall\ n \geq 0.$$

Assumption (A2)

(equivalently,

$$E\left[\|M(n+1)\|^2|x_0, M_m, m \leq n\right] \leq K\left(1+\|x(n)\|^2\right)\ \forall\ n \geq 0.)$$

In particular, if

(A3)

$$\boxed{\sup_m \|x(n)\| < \infty \text{ a.s.},}$$

"Iterate is bounded"

we have

$$\sup_n E\left[\|M(n+1)\|^2|\mathcal{F}_n\right] < \infty \text{ a.s.}$$

# Convergence claim: A rough introduction

$x \in \mathbb{R}^d$

(1) $\rightarrow$ $x(n+1) = x(n) + a(n)\big(\underbrace{h(x(n)) + M_{n+1}}_{(*)}\big)$

$x(n) \xrightarrow[a.s]{} x^*$ s.t. $h(x^*) = 0$ [Assume $x^*$ is unique]

If (i) $h$ is Lipschitz

(ii) $\sum a(n) = \infty$ $\quad \sum a(n)^2 < \infty$

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

$E(\|M_{n+1}\|^2 \mid \mathcal{F}_n) \leq K(1 + \|x(n)\|^2)$

(iv) $\sup_n \|x(n)\| < \infty$

---

Lecture-9

Tip:

To infer the limit of (1), take the conditional expectation of (*) w.r.t. $\mathcal{F}_n$ & equate it to zero

e.g.
$$x(n+1) = x(n) - a(n)\big(\hat{\nabla} f(x(n))\big)$$
$$= x(n) - a(n)\big(\nabla f(x(n)) + M_{n+1}\big)$$
$$M_{n+1} = \hat{\nabla}(f(x(n)) - \nabla f(x(n)))$$

Under suitable assumptions, $x(n) \xrightarrow[a.s]{} x^*$ as $n \to \infty$, where $\nabla f(x^*) = 0$

This is more general than it appears. Suppose the algorithm is

noise is a parameter

$$x(n+1) = x(n) + a(n) f(x(n), \xi(n+1)), \; n \geq 0,$$

$= x(n) + a(n)\big( E(f(x(n), \xi(n+1) | \mathcal{F}_n) + M(n+1)\big)$

where $\{\xi(n)\}$ are IID. This is often how many recursive algorithms are stated.

This can be put in the above form by letting

$$h(x) = E[f(x, \xi(n)] = E\left[f(x(n), \xi(n+1)) | x(n) = x\right]$$

$$= E\left[f(x(n), \xi(n+1) | \mathcal{F}_n\right],$$

$$M(n+1) = f(x(n), \xi(n+1)) - h(x(n)).$$

$E( M(n+1) | \mathcal{F}_n ) = 0$

Examples: Stochastic Gradient Descent ($h = -\nabla f$),

reinforcement learning algorithms (more later)

Highlights:

1. Typically small amount of computation and memory requirements per iterate (Contrast: batch-mode algorithms)

2. Incremental: makes a small change in the current iterate at each step

$a(n) = \frac{1}{n}$

3. Slowly decreasing stepsize captures 'exploration' (≈ large steps initially) vs 'exploitation' (≈ small steps later) trade-off

① $x(n) \xrightarrow[\text{a.s.}]{} x^*$ as $n \to \infty$ "SLLN spirit"

② $\sqrt{n}(x(n) - x^*) \xrightarrow[d]{} N(0, \Sigma)$ "CLT spirit"

4. Averages out the noise (can be thought of as a generalization of the Strong Law of Large Numbers)

③ $P(\|x(n) - x^*\| > \epsilon) \leq \exp(-cn\epsilon^2)$ "Concentration bound"

1.-3. typical of adaptive behavior $\Longrightarrow$ extremely well suited for adaptive algorithms or models of adaptation

One of the two main workhorses of statistical computation, MCMC being the other.

"Markov Chain Monte Carlo"

Hoeffding

$P(\hat{\mu}_n - \mu > \epsilon) \leq \exp(-cn\epsilon^2)$

Applications:

statistics, signal processing, machine learning, adaptive control and communications

Also for models of learning, bounded rationality, herding behavior, etc.

Classical approach for analysis: uses 'almost supermartingales' etc. (Robbins-Siegmund, $\cdots$)

Alternative approach: ODE (Ordinary Differential Equations) approach (Meerkov ’72, Derevetskii-Fradkov ‘74, Ljung ‘77)

ODE approach: Treat the iterates as a noisy discretization of the ODE

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

Recall the Euler scheme for this ODE:

$$x(n+1) = x(n) + ah(x(n)),\ n \geq 0,$$

where $a > 0$ is a small discrete time step.

$$x(n+1) = x(n) + a(n)\Big(h(x(n)) + \underbrace{M_{n+1}}_{\text{noise}}\Big)$$

Then SA can be viewed as an Euler scheme to approximate the ODE with slowly decreasing time steps $\{a(n)\}$ and measurement noise.

Robbins-Monro conditions:

$\sum_n a(n) = \infty \implies$ the entire time axis is covered. This is essential because we want to track the asymptotic (as $t \uparrow \infty$) behavior of the ODE.

$\sum_n a(n)^2 < \infty \implies$ the approximation of the ODE gets better with time:

$a(n) \to 0$ ensures that errors due to discretization are asymptotically zero

$\sum_n a(n)^2 < \infty$ ensures that errors due to the martingale difference noise are asymptotically zero, a.s. (multiplication by $a(n)$ reduces the (conditional) variance of noise)

Advantages:

1. Once you have mastered the approach, you can often write the limiting ODE by inspection and analyze it.

2. Designing algorithms: any convergent ODE is a template for an algorithm.

3. Finer dynamic phenomena lead to useful results, e.g., avoidance of unstable equilibria a.s. $\Longrightarrow$ avoidance of 'traps' (undesirable equilibria)

Analogy with SLLN suggests related results for fluctuations, e.g., central limit theorem, law of iterated logarithms, concentration inequalities

Further issues and variations:

stability tests, multiple timescales, distributed and asynchronous implementations, differential inclusion limits, constant stepsizes, other noise models, etc.

"Stochastic fixed point iterations

Covered on blackboard"

---

## An application: Mean estimation

Consider a random variable (r.v.) $X$ with mean $\mu$
& finite variance, say $\sigma^2$.

Suppose we are given iid samples $X_1, \ldots, X_m$

Let $r_m$ be the estimate of $\mu$.

Sample mean $\rightarrow$ $r_m = \frac{1}{m} \sum_{k=1}^{m} X_k$

$$r_{m+1} = \frac{1}{m+1} \sum_{k=1}^{m+1} X_k$$

$$= \frac{m}{m+1} \left( \frac{1}{m} \sum_{k=1}^{m} X_n \right) + \frac{1}{m+1} X_{m+1}$$

$$r_{m+1} = \frac{m}{m+1} r_m + \frac{1}{m+1} X_{m+1}$$ $\leftarrow$ iterative scheme for updating sample mean

new sample, prev. average

$$r_{m+1} = r_m + \frac{1}{m+1} \left( X_{m+1} - r_m \right)$$

iterate, step size, new sample

$$r_{m+1} = r_m + \beta_m \underbrace{\left( X_{m+1} - r_m \right)}_{Hr_m + w_m} \quad - \quad (**)$$

where $\beta_m = \frac{1}{m+1}$

$(**)$ resembles the sto. iter. algo $(*)$

What does Strong Law of Large numbers say about $r_m$?

$$r_m \longrightarrow \mu \text{ a.s. as } m \to \infty$$

with step size $\beta_m = \frac{1}{m+1}$

$$r_{m+1} = r_m + \beta_m \left( X_{m+1} - r_m \right)$$

$$= r_m + \beta_m \left( (\mu - r_m) + (X_{m+1} - \mu) \right)$$

$$r_{m+1} = (1-\beta_m)\, r_m + \beta_m \Big( \underbrace{\mu}_{Hr_m} + \underbrace{(X_{m+1} - \mu)}_{w_m} \Big)$$

$$E w_m = 0, \quad E w_m^2 < \infty, \quad \cancel{Hr^* = r^*} \Leftrightarrow \cancel{r^* = \mu}.$$

So, $(**)$ is really a sto. iterative algorithm.

And, we get $r_m \to r^* \ (= \mu)$ a.s. as $m \to \infty$.

The theory of sto. iter. algo that we shall develop ensures $r_m \to \mu$ a.s. as $m \to \infty$ for more general step sizes that satisfy

$$\sum_m \beta_m = \infty, \quad \sum_m \beta_m^2 < \infty \qquad \text{e.g. } \beta_m = \frac{1}{m}$$

Question: Why do we need these conditions?

noise

$r \to \boxed{\text{Box}} \to Hr + w$

# On step-size requirements:

$$r_{m+1} = (1-\beta_m) r_m + \beta_m (H r_m + w_m)$$

$$r_{m+1} = r_m + \beta_m \big( \underbrace{H r_m + w_m - r_m}_{\text{increment}} \big)$$

Suppose $\{w_i\}$ is independent of $r_m, \forall m$

& has variance $\sigma^2$, eg. $w \sim N(0, \sigma^2 I)$

$Var(X+Y) = Var X + Var Y$ if $X, Y$ independent.

$Var(cX) = c^2 Var X$

Variance of $r_{m+1}$

$$= Var(r_{m+1})$$

$$= Var\left[(1-\beta_m) r_m + \beta_m H r_m\right] + \beta_m^2 Var(w_m)$$

$$= Var\left[(1-\beta_m) r_m + \beta_m H r_m\right] + \beta_m^2 \sigma^2$$

$$\geq \beta_m^2 \sigma^2$$

$\beta_m = \beta$ "constant" stepsize

$\beta_m = \frac{1}{m}$ "diminishing" stepsize

Now, if $\beta_m = \beta \ \forall m$, then

Want $r_m \to r^*$ s.t. $H r^* = r^*$

$$Var(r_{m+1}) \geq \beta^2 \sigma^2$$

So, fixing $\beta > 0$, $r_m \not\to r^*$ for any $r^*$.

Aside: constant step size sto-approx algos → convergence guarantees are in "distribution" & not "a.s."

So, the step-size has to vanish asymptotically. i.e., $\beta_m \to 0$ as $m \to \infty$.

But, $\beta_m$ cannot go down too fast.

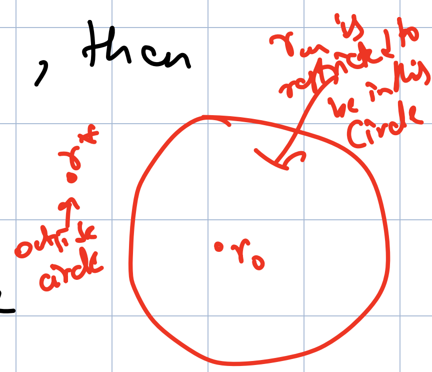
$$r_{m+1} = r_m + \beta_m \underbrace{\left( Hr_m + w_m - r_m \right)}_{\text{increment}}$$

$$|r_m - r_0| \overset{\text{triangle inequality}}{\leq} \sum_{\tau=0}^{m-1} \underbrace{\beta_\tau \left| Hr_\tau - r_\tau + w_\tau \right|}_{\text{sum of increments}}$$

So, if $\left| Hr_\tau - r_\tau + w_\tau \right| \leq C_1$ and

if $\displaystyle\sum_{\tau=0}^{\infty} \beta_\tau \leq C_2 < \infty$, then

$|r_m - r_0|$ is bounded above $\forall m$




$(\Rightarrow)$ $r_m$ is within a certain radius of $r_0$

problematic if $r^*$ lies outside the radius.

So, we need $\displaystyle\sum_\tau \beta_\tau = \infty$

$$\boxed{\text{"} \sum \beta_t = \infty \ \& \ \beta_t \to 0 \text{ as } t \to \infty \text{"}}$$

$\beta_t = \dfrac{1}{t^{1/4}}$