Text :- ↳ V.S Borkar, Stochastic Approximation : A dynamical systems viewpoint.
(2008)

2nd Edition (2022)

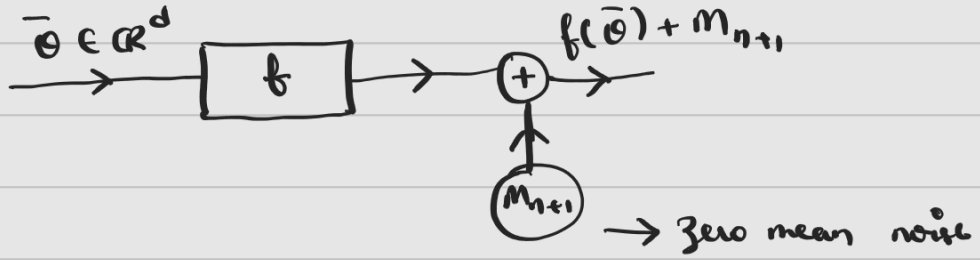
↳ Bruce Hajek, Random processes for Engineers
Cambridge Univ Press, 2015

↳ V.S. Borkar, Probability theory : Advanced Course, Springer 1995.

Evaluation :- 2 Mid-terms (20% + 20%)
1 Project (20%)
1 Final exam (40%)

Paper :- K.L. Robbins and H. Monro,
Annals of Math. Stat, 1951

Suppose $f : \mathbb{R}^d \to \mathbb{R}^d$ and we don't know the functional form of f.



Q) Can we find a θ^* for which $f(\theta^*) = 0$

Algorithm :- (Robbins - Monro Algorithm)

$$\theta_{n+1} = \theta_n + a(n)\left(f(\theta_n) + M_{n+1} \right)$$

Starting from some $\theta_0 \in \mathbb{R}^d$

$a(n)$: learning rate parameter or stepsize.

Typically $a(n) = 1/n$, $n \geq 1$

↳ As $n \to \infty$, under some conditions, $\theta_n \to \theta^*$ s.t. $f(\theta^*) = 0$.

Examples where R-M algorithm is useful

↳ Finding fixed points of a function given noisy observations.

Finds θ^* s.t. $\theta^* = f(\theta^*)$

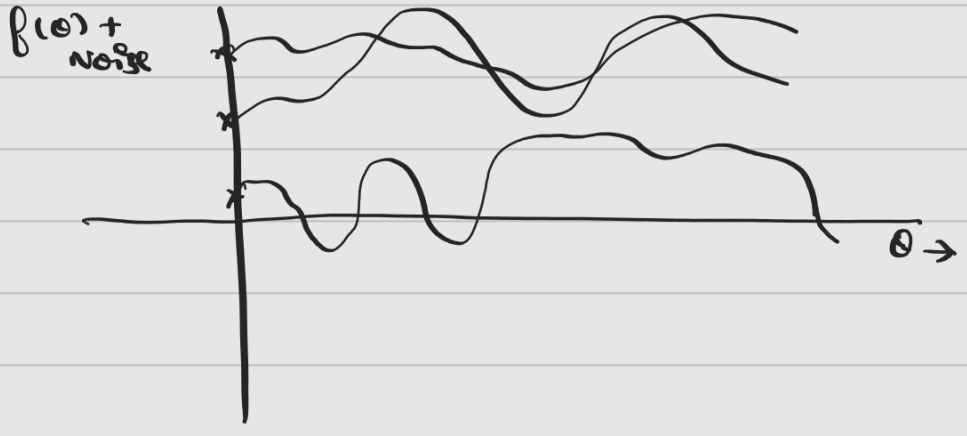
let $g(\theta) = f(\theta) - \theta$

(zeros of g are fixed points of f)

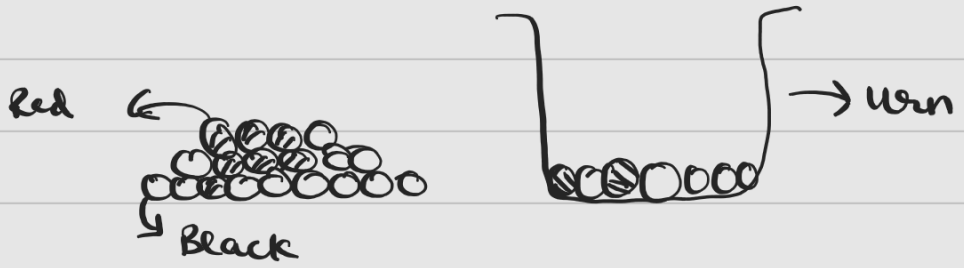
↳ Finding local minima of a function $g: \mathbb{R}^d \to \mathbb{R}$

let $f = \nabla g$

The zeros of f will be stationary points of g.



Chapter - 1



At any instant, we randomly pick a ball and drop it in urn.
By time n, we drop n balls in the urn.

Objective :- Find the long-run function of red balls in the urn.

Let y_n = # of red balls in urn by time n
x_n = fraction of red balls in urn at time n.

$$\therefore \quad x_n = \frac{y_n}{n}$$

Let ξ_{n+1} = $I_{\{n+1^{st} \text{ ball is red in colour}\}}$ $= \begin{cases} 1 & \text{if } n+1^{th} \text{ ball is red} \\ 0 & \text{o.w.} \end{cases}$

Then $\quad y_{n+1} = y_n + \xi_{n+1}$

Assumption :- Probability that $n+1^{st}$ ball dropped is red in colour given the entire history of red and black balls dropped until time n depends only on $p(x_n)$, where

$$p : [0,1] \to [0,1]$$

$$P(\xi_{n+1} = 1 \mid \xi_n, \xi_{n-1}, \dots, \xi_1) = p(x_n) \text{ a.s.}$$

Note :- $\quad x_n = \frac{y_n}{n}$

$$\therefore x_{n+1} = \frac{y_{n+1}}{(n+1)} = \frac{(y_n + \xi_{n+1})}{(n+1)}$$

$$= \frac{n}{n+1} \cdot \frac{y_n}{n} + \frac{1}{n+1} \xi_{n+1}$$

$$= \frac{n}{n+1} x_n + \frac{\xi_{n+1}}{n+1}$$

$$x_{n+1} = x_n + \frac{1}{n+1} (\xi_{n+1} - x_n)$$

$$= x_n + \frac{1}{n+1} (p(x_n) - x_n) + \frac{1}{n+1} (\xi_{n+1} - p(x_n))$$

Let $M_{n+1} = \xi_{n+1} - p(x_n)$, $n \geq 0$

$E[M_{n+1} | \xi_1, \xi_2, \cdots, \xi_n]$

$$= E[\xi_{n+1} | \xi_1, \xi_2, \cdots, \xi_n] - E[p(x_n) | \xi_1, \cdots, \xi_n]$$

$$= P(\xi_{n+1} = 1 | \xi_1, \cdots, \xi_n) - E[p(x_n) | \xi_1, \cdots, \xi_n]$$

$$= p(x_n) - p(x_n)$$

$$= 0$$

$$E[M_{n+1}] = E[E[M_{n+1} | \xi_1, \cdots, \xi_n]] = 0$$

Algorithm :- $x_{n+1} = x_n + \frac{1}{n+1} (p(x_n) - x_n) + \frac{1}{n+1} M_{n+1}$, $n \geq 0$

The ODE approach [1977 - Ljung; 1977-78 - Dvoretsky; 1978 - Kushner and Clark]

Under some conditions we can approximate the algorithm via an ODE $\dot{x}(t) = p(x(t)) - x(t)$ (*) → Numerical schemes for solving ODE

Euler discretization of the ODE

From (*), we have

$$x(t) = x_0 + \int_0^t p(x(\tau)) - x(\tau)\, d\tau$$

Suppose $h > 0$ is a small time element.

Then $$\int_0^t p(x(\tau)) - x(\tau)\, d\tau = \sum_{m=0}^{n} (p(x(mh)) - x(mh))\, h$$

$t \approx nh$

$$x(nh) = x_0 + \sum_{m=0}^{n-1} p(x(mh)) - x(mh))\, h$$

$$x((n+1)h) = x_0 + \sum_{m=0}^{n} p(x(mh)) - x(mh))\, h$$

Then $x((n+1)h) = x(nh) + h(p(x(nh)) - x(nh))$ (Euler discretization) ↳ (2)

□

Comparing (1) and (2)

↳ $h = 1/n+1$

↳ In (1), additional term = $\frac{M_{n+1}}{n+1}$

↳

$$x_{n+1} = x_0 + \sum_{m=0}^{n} \frac{1}{m+1} (p(x_m) - x_m) + \underbrace{\sum_{m=0}^{n} \frac{1}{m+1} M_{m+1}}$$

$$\downarrow$$

As $n \to \infty$ $\sum_{m=0}^{n} \frac{M_{m+1}}{m+1} < \infty$ w.p. 1

Let's consider the ODE

$$\dot{x}(t) = p(x(t)) - x(t)$$

Assume $p : [0,1] \to [0,1]$ is lipschitz continuous

$\exists L > 0, \forall x, y \in [0,1]$ $|p(x) - p(y)| \leq L|x - y|$

We can show that the ODE is <u>well-posed.</u>

Suppose, we start ODE (*) s.t.

$$x_0 \in [0,1]$$

Note :- when $x(t)=0$, $\dot{x}(t) \geq 0$

when $x(t)=1$, $\dot{x}(t) \leq 0$

Suppose $\dot{x}(t) > 0$ when $x(t)=0$

⇒ The ODE will move in the right direction.

However $\dot{x}(t) \leq 0$ when $x(t) = 1$

↳ Note that at some point $x(t) \in [x(0), 1]$, $\dot{x}(t) = 0$ by Continuity of trajectory

↳ Let $H = \{x \mid p(x) = x\} \equiv$ set of equilibria of the ODE (*).

↳ This also means that H is nonempty.

↳ Note :- $x(t) \longrightarrow H$ as $t \to \infty$.

We then argue under some conditions that algorithm (1) satisfies

$$x_n \to H \text{ as } n \to \infty \quad \text{w.p. 1.}$$

Assumptions

1) $p : [0,1] \to [0,1]$ is lipschitz.
2) $a(n) > 0$ $\forall n$ [$a(n) = \frac{1}{n+1}$ here]

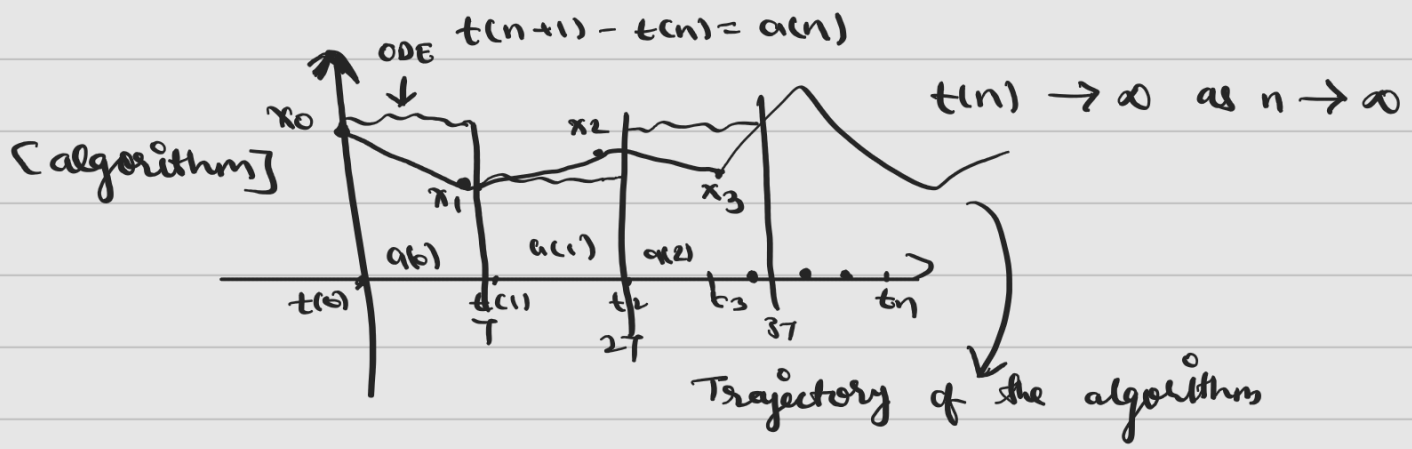
$$\sum_n a(n) = \infty \quad , \quad \sum_n (a(n))^2 < \infty$$

Example :- $a(n) = \frac{1}{n+1}$, $\frac{1}{(n+1)^\alpha}$, $\alpha \in (0.5, 1]$

$$\frac{\log(n+1)}{(n+1)} \quad , \quad \frac{1}{(n+1)\log(n+1)} \quad , \quad n \geq 1$$

Suppose $t(n) = \sum_{i=0}^{n-1} a(i)$, $n \geq 0$

with $t(0) = 0$



Let $\gamma > 0$ be fixed.

Second condition :- $\sum_n a(n^2) < \infty$

↓

Error due to noise & discretization asymptotically vanish.

generally $a(m)$

$$z_n = \sum_{m=0}^{n} \left(\frac{1}{m+1} * M_{m+1} \right), \quad n \geq 0$$

↓

Martingale sequence.

↳ Due to the fact that $\sum_n a(n)^2 < \infty$.

$z_n \xrightarrow{a.s.} z_\infty$ as $n \to \infty$. $\Rightarrow \sum_{m=n}^{\infty} a(m) M_{m+1} \to 0$ as $n \to \infty$ a.s.

3) $E[M_{n+1} \mid x_m, M_m, m \leq n] = 0 \quad \forall n$

$E[\|M_{n+1}\|^2 \mid x_m, M_m, m \leq n] \leq K(1 + \|x_n\|^2)$

4) Crucial requirement [nontrivial]

$\sup_n \|x_n\| < \infty \quad$ w.p. 1

$\leq C(\omega) < \infty$

Does not mean $\sup_\omega C(\omega) < \infty$.

2nd setting :- Finding a local minimum.

↳ Consider a repeated experiment with i/p - o/p pairs (x_n, y_n) with $x_n \in \mathbb{R}^m$, $y_n \in \mathbb{R}^k$. (i/p, o/p)

↳ Assume (x_n, y_n), $n \geq 0$ are i.i.d.

Goal :- Find a best fit

$$y_n = f_w(x_n) + \epsilon_n$$

Here we are given a parameterized family of functions

$$f_w : \mathbb{R}^m \to \mathbb{R}^k$$

with ϵ_n as the error.

↳ let $g(w) = \frac{1}{2} E[\|\epsilon_n\|^2]$

$$g(w) = \frac{1}{2} E[\|y_n - f_w(x_n)\|^2]$$

Goal :- Find a w^* s.t. $g(w)$ is minimized at w^*.

Let $\|\cdot\|$ be the eucledian norm.

Thus, $g(w) = \frac{1}{2} E[(y_n - f_w(x_n))^T (y_n - f_w(x_n))]$

Assuming that the gradient can be taken inside the expectation, we have

$$\nabla_w g(w) = \frac{1}{2} \nabla_w E[(y_n - f_w(x_n))^T (y_n - f_w(x_n))]$$

$$= \frac{1}{2} E[\nabla_w ((y_n - f_w(x_n))^T (y_n - f_w(x_n))]$$

$$= \frac{1}{2} E[-2 \nabla f_w(x_n)^T (y_n - f_w(x_n))]$$

$= -E[\nabla f_w(x_n)^T (y_n - f_w(x_n))]$

A gradient descent scheme could be :-

$$w_{n+1} = w_n + \alpha E[(y_n - f_w(x_n))^T \nabla f_w(x_n)]$$

Problem :- We don't know the distribution of (x_n, y_n) pairs.

Alternative :- $w_{n+1} = w_n + a(n)((y_n - f_w(x_n))^T \nabla f_w(x_n))$

$w_{n+1} = w_n + a(n)[E[\nabla f_w(x_n)^T (y_n - f_w(x_n)] + M_{n+1}]$

$M_{n+1} = (y_n - f_w(x_n))^T \nabla f_w(x_n)$
$- E[(y_n - f_w(x_n))^T) \nabla f_w(x_n)]$

$= w_n + a(n)[-\nabla_w g(w_n) + M_{n+1}]$

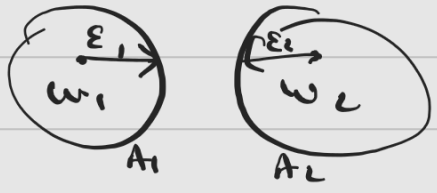
SGD (Stochastic gradient descent scheme)

$$\dot{w}(t) = -\nabla_w g(w(t)) \longrightarrow (+)$$

Let $H = \{w \mid \nabla_w g(w) = 0\}$

↓

Stationary points of the ODE (+)



$A_1 \cap A_2 = \phi$

Note :- Stationary points are those for which gradient is zero.

⇒ local maxima, local minima, saddle points

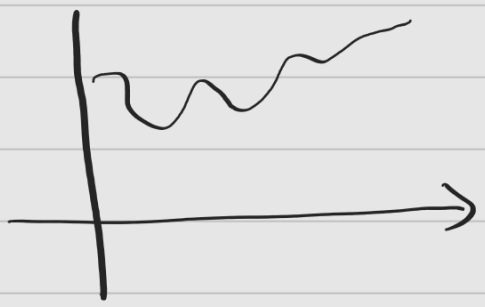
Algorithm :- $x_{n+1} = x_n + a(n)(h(x_n) + M_{n+1}), \; n \geq 0 \quad -(4)$

↳ Analyze convergence of (4) assuming stability (Chapter 2)

($\sup_k \|x_k\| < \infty$ w.p. 1)

↳ Provide sufficient conditions for stability of (4) (Chapter 3)

↳ Lock-in-probability of SA. (Chapter 4)



↳ Consider recursions such as

$x_{n+1} = x_n + a(n)(f(x_n, \xi_n) + M_{n+1}), \; n \geq 0$

Here $\{\xi_n\}$ is parameter dependent Markov process.

↳ $P(\xi_{n+1} = y' \mid \xi_n = y, x_n = x) = p_x(y, y')$ Under ergodicity of (Chapter 7) Markov process, $h(x) = \int f(x,y)\, \nu_x(dy)$

$\dot{x}(t) = h(x)$ ← Stationary distribution of $\{\xi_n\}$ for a fixed x.

↳ Algorithm with set-valued maps. (Stochastic recursive inclusions)

Chapter 5-6

$h(x_n) \subset \mathbb{R}^d$

↳ Differential inclusion $\boxed{\dot{x}(t) \in h(x(t))}$

↳ If process is not ergodic ⇒ stationary distribution is not unique.

$$H(x) = \left\{ \int f(x,y)\, \nu_x(dy) \right\}$$

DI : $\dot{x}(t) \in H(x(t)), \quad t \geq 0$

Some basics on martingales

[Chapter 10 of Bruce Hajek's book]

Suppose X and Y be two discrete r.v.s.

Then

$$g(i) \triangleq E[X|Y=i] = \sum_j u_j P(X=u_j | Y=i) \quad \text{— (1)}$$

Here $X \in \{u_1, u_2, \ldots\}$

Note :- (1) is well defined if $P(Y=i) > 0$

Also, either the sum restricted $\{j \mid u_j > 0\}$ or $\{j \mid u_j < 0\}$ is convergent.

⇒ Define r.v. $E[X|Y] = g(Y)$ where $g(i) = E[X|Y=i]$

Similarly, if X and Y are continuous r.v.s having a joint PDF.

$$E[X|Y=y] = \int x f_{X|Y}(x|y)\, dx = g(y)$$

and $E[X|Y] = g(Y)$.

⇒ Given a set Ω, a set of subsets $\mathcal{D}$ of Ω is called a σ-algebra if

(i) $\Omega \in \mathcal{D}$

(ii) If $A \in \mathcal{D} \Rightarrow A^c \in \mathcal{D}$

(iii) If $A_1, A_2, \ldots \in \mathcal{D} \Rightarrow \bigcup A_i \in \mathcal{D}$

⇒ In particular, we require that the set of events $\mathcal{F}$ in a probability space $(\Omega, \mathcal{F}, P)$ to be a σ-algebra.

⇒ We say that $\mathcal{D}$ is a sub σ-algebra of $\mathcal{F}$ if $\mathcal{D}$ is a σ-algebra and $\mathcal{D} \subset \mathcal{F}$.

⇒ We say that a r.v. Z is $\mathcal{D}$ measurable if

$\{Z \le c\} \in \mathcal{D} \quad \forall c \in \mathbb{R}$.

Definition:- σ-algebra generated by a set of r.v.s $(Y_i, i \in I)$ denoted by $\sigma(Y_i : i \in I)$ is the smallest σ-algebra containing all sets of the form $\{Y_i \le c\}$, $i \in I$, $c \in \mathbb{R}$.

Fact:- The smallest σ-algebra exists and equals the intersection of σ-algebras $\sigma(Y_i)$, $i \in I$.

↳ σ-algebra generated by a random variable Y will be denoted by $\sigma(Y)$ or $\mathcal{F}^Y$.

↳ **Borel - measurable functions**

Let $g : \mathbb{R} \to \mathbb{R}$. We say that g is Borel measurable if $\{x \mid g(x) \le c\} \in \mathcal{B}$ $\forall c \in \mathbb{R}$.

Here $\mathcal{B}$ is the σ-algebra of open subsets of $\mathbb{R}$.

$$g^{-1}((-\infty, c]) \in \mathcal{B} \quad \forall c \in \mathbb{R}$$