Last time:

Stochastic Approximation Iteration:

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

We came to this via: Urn model $\left(h(x) = p(x) - x\right)$

We saw the need for two conditions on step-sizes:

① $\sum_n a(n) = \infty$

② $\sum_n a^2(n) < \infty$

Today: Example 2: Regression.

Preliminaries: $(X, Y) \sim P_{XY}$, $X \in \mathbb{R}^n$, $Y \in \mathbb{R}^k$

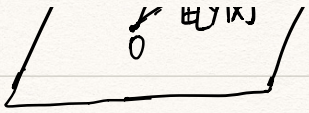
Problem: Observe X, estimate Y.

to $\min_{f(\cdot)} \frac{1}{2}\mathbb{E}\left[\|Y - f(X)\|^2\right]$

The answer to this is: $f^*(X) = \mathbb{E}[Y|X]$.

Proof (Exercise): Add & sub $\mathbb{E}[Y|X]$ & use the orthogonality property:

$$\mathbb{E}\left[\left(Y - \mathbb{E}[Y|X]\right)^T f(X)\right] = 0$$



Modified Problem: Given a family of parameterized functions $\{f_w : \mathbb{R}^m \to \mathbb{R}^k, w \in \mathbb{R}^d\}$, find the best fit w.

$$w^* = \operatorname*{argmin}_w g(w), \quad g(w) = \frac{1}{2}\mathbb{E}[\|Y - f_w(x)\|^2].$$

eg: Linear fns, frequencies of sines/cosines

$$g(w) = \frac{1}{2}\mathbb{E}[Y^TY - 2f_w(x)^TY + f_w(x)^Tf_w(x)]$$

$$\nabla g(w) = -\mathbb{E}\begin{bmatrix} \vdots \\ \frac{\partial}{\partial w_i}\sum_{j=1}^{k} f_{w,j}(x)Y_j \\ \vdots \end{bmatrix} + \mathbb{E}\begin{bmatrix} \vdots \\ \frac{1}{2}\sum_{j=1}^{k} f_{w,j}^2(x) \\ \vdots \end{bmatrix}$$

(Assuming we can push $\frac{\partial}{\partial w_i}$ inside $\mathbb{E}[\cdot]$)

$$= -\mathbb{E}\begin{bmatrix} \vdots \\ \sum_{j=1}^{k}\frac{\partial}{\partial w_i} f_{w,j}(x)\left(Y_j - f_{w,j}(x)\right) \\ \vdots \end{bmatrix}$$

$$= -\mathbb{E}\left[\begin{bmatrix} \frac{\partial}{\partial w_1}f_{w,1}(x) & \frac{\partial}{\partial w_1}f_{w,2}(x) & \cdots & \frac{\partial}{\partial w_1}f_{w,k}(x) \\ \vdots & & \ddots & \vdots \end{bmatrix}\begin{bmatrix} Y - f_w(x) \end{bmatrix}\right]$$

(*)

$$\left[\left[\frac{\partial}{\partial w_d} f_{w,1}(x) \quad \cdots \quad \frac{\partial}{\partial w_d} f_{w,K}(x)\right]\right]$$

As an aside, Jacobian: Given $f_w(x) = \begin{bmatrix} f_{w,1}(x) \\ \vdots \\ f_{w,K}(x) \end{bmatrix}$,

$$\text{Jacobian} = \begin{bmatrix} \frac{\partial}{\partial w_1} f_{w,1}(x) & \cdots & \frac{\partial}{\partial w_d} f_{w,1}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial w_1} f_{w,K}(x) & \cdots & \frac{\partial}{\partial w_d} f_{w,K}(x) \end{bmatrix} =: D_w f_w(x)$$

So $(*) = (D_w f_w(x))^T$

We thus have $\nabla g(w) = -\mathbb{E}\left[(D_w f_w(x))^T (Y - f_w(x))\right]$

So the idea is that: $\nabla g(w)$ gives direction of steepest ascent. So take a step in opp dir^n.

Try $w_{n+1} = w_n - \nabla g(w)\Big|_{w=w_n}$

$$\Rightarrow w_{n+1} = w_n + \mathbb{E}\left[(D_w f_w(x))^T (Y - f_w(x))\right]$$

Now, we don't know $P_{X,Y}$ & hence it's difficult for us to evaluate the expectation.

S'pose we have access to samples from the joint distⁿ

$$(X_n, Y_n)_{n\geq 1} \overset{iid}{\sim} P_{XY}$$

Try: "Plug-in estimate" for the gradient:

$$w_{n+1} = w_n + D_w f_w(X_n)^T (Y_n - f_w(X_n))\Big|_{w=w_n}$$

Problem: Like a random walk, can wander all over.

Idea: Smoothen the dynamics.

Smoothed dynamics:

$$w_{n+1} = (1-a(n))\, w_n + a(n)\left[w_n + D_w f_w(X_n)^T (Y - f_w(X_n))\Big|_{w=w_n}\right]$$

$$= w_n + a(n)\left[D_w f_w(X_n)^T (Y_n - f_w(X_n))\Big|_{w=w_n}\right]$$

As time progresses, $a(n) \to 0$

Add & sub. mean.

$$\Rightarrow w_{n+1} = w_n + a(n)\Big[-\nabla g(w_n) + D_w f_w(X_n)^T (Y - f_w(X_n))\Big|_{w=w_n} - (-\nabla g(w_n))\Big]$$

$$\underbrace{\qquad\qquad}_{M_{n+1},\ \text{noise}}$$

$$\Rightarrow w_{n+1} = w_n + a(n)\left[-\nabla g(w_n) + M_{n+1}\right] \quad \text{(Stoch. grad. descent)}$$

This is a noisy discretization of

$$\dot{w}(t) = -\nabla g(w(t)) \quad \longrightarrow \text{Gradient dynamics}$$

We anticipate that $w_n \xrightarrow[n\to\infty]{} H = \{w : \nabla g(w) = 0\}$.

If "noise is rich enough", we can show convergence to minima, possibly local.

$$\sum_{n=0}^{\infty} a(n) = \infty, \quad \sum_{n=0}^{\infty} a^2(n) < \infty$$

Chapter 2: Preliminaries: ODEs.

$$\dot{x}(t) = h(x(t)), \quad x(0) = \bar{x} \in \mathbb{R}^d.$$

Defn: The ode is "well-posed" if for any $\bar{x} \in \mathbb{R}^d$, the solution is unique, $x(\cdot) \in C([0,\infty); \mathbb{R}^d)$,

(C: space of continuous fns; cont $\forall t \in [0,\infty)$; $x(\cdot) \in \mathbb{R}^d$)

and further the mapping $\bar{x} \in \mathbb{R}^d \longmapsto x(\cdot) \in C([0,\infty); \mathbb{R}^d)$

is continuous.

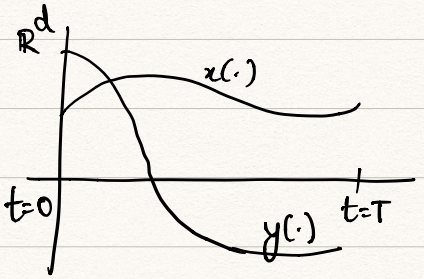
Remarks:

- $C([0,\infty); \mathbb{R}^d)$: space of continuous fns for all positive time,

taking values in $\mathbb{R}^d$.

$x(\cdot) \in C([0,\infty); \mathbb{R}^d)$ is a function $x: [0,\infty) \to \mathbb{R}^d$.

- $C([0,\infty); \mathbb{R}^d)$ is a metric space $\rho(x(\cdot), y(\cdot))$
- This induces a topology $\rightarrow$ collection of open sets generated by open balls.
- A mapping is continuous, if inverse images of open sets are open.
- To understand ρ, let us consider first $C([0,T]; \mathbb{R}^d)$, $T>0$, finite.



$$\|x(\cdot) - y(\cdot)\|_T = \sup_{t\in[0,T]} \|x(t) - y(t)\|$$

= width of the smallest tube around $x(\cdot)$ that swallows $y(\cdot)$

(Exercise: Verify that this is a metric).

- (Back to ρ):

$$\rho(x(\cdot), y(\cdot)) = \sum_{T=1}^{\infty} \left(\left(\|x(\cdot)|_T - y(\cdot)|_T\|_T \right) \wedge 1 \right) 2^{-T}$$