Chapter 6.

一、Motivating example: mean estimation.

1. expectation: $E[X] \approx \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$

$\bar{x} \to E[X]$ as $N \to \infty$

2. Calculate:

directly: have to wait for a long time.

iteration: $w_{k+1} = \frac{1}{k}\sum_{i=1}^{k} x_i \quad k = 1, 2, 3, \cdots$

$w_k = \frac{1}{k-1}\sum_{i=1}^{k-1} x_i \quad k = 1, 2, 3, \cdots$

$$w_{k+1} = \frac{1}{k}\sum_{i=1}^{k} x_i = \frac{1}{k}\left(\sum_{i=1}^{k-1} x_i + x_k\right)$$

$$= \frac{1}{k}\left((k-1)w_k + x_k\right)$$

$$= w_k - \frac{1}{k}(w_k - x_k)$$

Iteration algorithm: $w_{k+1} = w_k - \frac{1}{k}(w_k - x_k)$ (*)

Furthermore, general expression:

$$w_{k+1} = w_k - a_k(w_k - x_k)$$

where $a_k > 0$ and satisfy some mild condition

= Robbins - Monro algorithm

1. Stochastic appoximation (SA)

① SA refers to a broad of stochastic iterative algorithm solving root finding or optimization problem.

② SA does not require to know the expresstion of the objective function

③ Stochastic gradient descent (SGD) is a special form of RM algorithm

2. find root of $g(w)=0$

example: $g(w) = \nabla_w J(w) = 0$

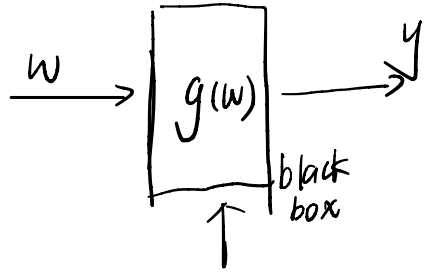
or $g(w) = c \Rightarrow g(w) - c = 0$

$g'(w)$

If $g(w)$ is known: use numerical algorithm

If $g(w)$ is unknown?

example: such as a artificial network



$y = g(w)$

when $y=0$ $w=?$

The Robbins-Monro algorithm can solve this problem: $w_{k+1} = w_k - a_k \bar{g}(w_k, \eta_k)$ $k = 1, 2, 3 \cdots$

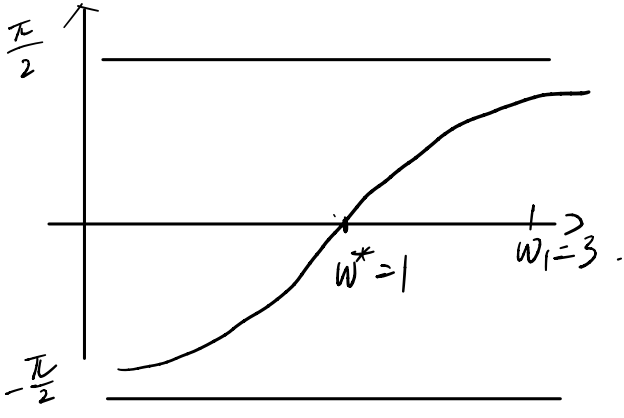
$\bar{g}(w_k, \eta_k) = g(w_k) + \eta_k$ noisy observation

need data

3. Convergence properties.

an illustrative example:

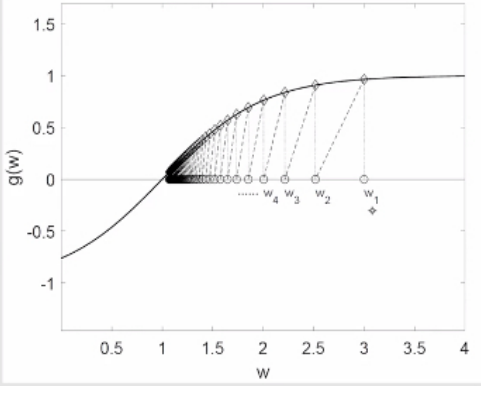
· $g(w) = \tanh(w-1)$



$w_1 = 3 \quad a_k = \frac{1}{k} \quad \eta_k \equiv 0$ (no noise)

$w_{k+1} = w_k - a_k g(w_k)$

sult: w_k converges to the true root $w^* = 1$.



不动点迭代
牛顿迭代

Intuition: w_{k+1} is closer to w^* than w_k.

- When $w_k > w^*$, we have $g(w_k) > 0$. Then, $w_{k+1} = w_k - a_k g(w_k) < w_k$ and hence w_{k+1} is closer to w^* than w_k.
- When $w_k < w^*$, we have $g(w_k) < 0$. Then, $w_{k+1} = w_k - a_k g(w_k) > w_k$ and w_{k+1} is closer to w^* than w_k.

严格数学推导：

Theorem (Robbins-Monro Theorem)

1) $0 < c_1 \le \nabla_w g(w) \le c_2$ for all w;

2) $\sum_{k=1}^{\infty} a_k = \infty$ and $\sum_{k=1}^{\infty} a_k^2 < \infty$;

3) $E[\eta_k | \mathcal{H}_k] = 0$ and $E[\eta_k^2 | \mathcal{H}_k] < \infty$;

where $\mathcal{H}_k = \{w_k, w_{k-1}, \dots\}$ then w_k converge with probability 1 to the root w^* satisfying $g(w^*) = 0$

依概率收敛？

Explanation:

$\nabla g(w)$ 有界且大于0，$g(w) = 0$ exist and unique.

The gradient is bounded from the above.

$$\min J(w)$$

$$g(w) = \nabla_w J(w) = 0$$

$$0 < c_1 \le \nabla_w g(w) \le c_2 \text{ for all } w.$$

$J(w)$ 是一个凸函数 convex.

$\sum_{k=1}^{\infty} a_k = \infty$ and $\sum_{k=1}^{\infty} a_k^2 < \infty$

The condition of $\sum_{k=1}^{\infty} a_k^2 < \infty$ ensure that a_k converges to zero as $k \to \infty$

The condition of $\sum_{k=1}^{\infty} a_k = \infty$ ensure that a_k do not converge to zero too fast.

$E[\eta_k | \mathcal{H}_k] = 0$ and $E[\eta_k^2 | \mathcal{H}_k] < \infty$

A special yet common case is that $\{\eta_k\}$ is an iid stochastic sequence satisfying $E[\eta_k] = 0$ and $E[\eta_k^2] < \infty$

The observation error η_k is not required to be Gaussian.

① $\sum_{k=1}^{\infty} a_k^2 < \infty \Rightarrow a_k \to 0$ as $k \to \infty$

$w_{k+1} - w_k = -a_k \tilde{g}(w_k, \eta_k)$ $k \to \infty$ $w_{k+1} - w_k \to 0$

② $\sum_{k=1}^{\infty} a_k = \infty$

$w_2 = w_1 - a_1 \tilde{g}(w_1, \eta_1)$

$w_3 = w_2 - a_2 \tilde{g}(w_2, \eta_2)$

$\vdots$

$w_{k+1} = w_k - a_k \tilde{g}(w_k, \eta_k)$

累加之, $w_1 - w_\infty = \sum_{k=1}^{\infty} a_k \tilde{g}(w_k, \eta_k)$

$w_\infty = w_1 - \sum_{k=1}^{\infty} a_k \tilde{g}(w_k, \eta_k)$

$w_\infty \to w^*$

$\sum_{k=1}^{\infty} a_k < \infty$ then $\sum_{k=1}^{\infty} a_k \tilde{g}(w_k, \eta_k)$ bounded.

One typical sequence: $a_k = \frac{1}{k}$

It hold that: $\lim_{n\to\infty} \left(\sum_{k=1}^{n} \frac{1}{k} - \ln n \right) = \text{常数}$

Euler-Mascheroni constant.

It is notable that:

$$\sum_{k=1}^{\infty} \frac{1}{k^2} = \frac{\pi}{6} < \infty$$

Consider a Function:

$$g(w) = w - E(X)$$

The observation:

$$\tilde{g}(w, x) = w - x$$

$$\hat{g}(w, \eta) = w - x = w - x + E[X] - E[X]$$
$$= \underbrace{(w - E[X])}_{g(w)} + \underbrace{(E[X] - x)}_{\eta}$$
$$= g(w) + \eta$$

The RM for $g(x) = 0$

$$w_{k+1} = w_k - \alpha_k \hat{g}(w_k, \eta_k) = w_k - \alpha_k (w_k - x_k)$$

三、SGD (Stochastic gradient descent)
随机梯度下降.

1、Algorithm description:

solve the following optimization problem:

$$\min_w J(w) = E[f(w, X)]$$

w is parameter to be optimized

X is random variable.

① method 1: gradient descent (GD)

$$w_{k+1} = w_k - \alpha_k \nabla_w E[f(w_k, X)] = w_k - \alpha_k E[\nabla_w f(w_k, X)]$$

draw-back: the expected value is difficult to obtain.

② method 2: batch gradient descent (BGD)

$$E[\nabla_w f(w_k, X)] \approx \frac{1}{n}\sum_{i=1}^{n} \nabla_w f(w_k, x_i)$$

$$w_{k+1} = w_k - \alpha_k \frac{1}{n}\sum_{i=1}^{n} \nabla_w f(w_k, x_i)$$

draw-back: requires many samples in each iteration.

③ SGD

$$w_{k+1} = w_k - \alpha_k \underbrace{\nabla_w f(w_k, x_k)}_{\text{stochastic gradient}}$$

BGD let $n=1$ = SGD

2. Example:

$$\min_w J(w) = E[f(w,X)] = E\left[\frac{1}{2}\|w-X\|^2\right]$$

$$f(w,X) = \|w-X\|^2/2 \qquad \nabla_w f(w,X) = w-X$$

optimal solution: $\nabla_w J(w) = 0 \Rightarrow E[\nabla_w f(w,X)] = 0$

$$\Rightarrow E[w-X] = 0 \Rightarrow w = E(X)$$

① GD: $w_{k+1} = w_k - \alpha_k \nabla_w J(w_k) = w_k - \alpha_k E[\nabla_w f(w_k, X)]$

$$= w_k - \alpha_k E[w_k - X]$$

② SGD: $w_{k+1} = w_k - \alpha_k \nabla_w f(w_k, x_k) = w_k - \alpha_k (w_k - x_k)$

stochastic gradiant = true gradient

3. SGD properties

$$\delta_k \le \frac{\nabla_w f(w_k, x_k) - E[\nabla_w f(w_k, X)]}{c\,|w_k - w^*|}$$

w_k 离 w^* 很远时 SGD ≈ GD