Chapter 6 Stochastic Approximation

Stochastic Approximation, Robbins-Monro Algorithm, Stochastic Gradient Descent

Stochastic Approximation

1. What is Stochastic Approximation?

Robbins-Monro Algorithm

2. Explain the idea and formula of Robbins-Monro algorithm.
3. Explain the convergence conditions of RM and their meanings.
4. How to use RM for mean estimation?

Stochastic Gradient Descent

5. Explain the derivation of SGD.
6. How to use SGD for mean estimation?
7. What is BGD and MBGD?
8. What is the convergence conditions and their meanings of SGD?
9. Explain the converge pattern of SGD.

1. Stochastic approximation is a broad class of stochastic iterative algorithms for root-finding or optimization problems. This is not a reinforcement learning algorithm, but is the basic for temporal-difference learning algorithms.
2. Robbins-Monro algorithm is an incremental method used for solving root-finding problems like g(w) = 0. It doesn't need the expression or derivative of the function. We just need noisy outputs $g\tilde{}(w_k,\eta_k)$, where η is the noise (error term).

$$w_{k+1} = w_k - a_k \tilde{g}(w_k, \eta_k), \qquad k = 1, 2, 3, \ldots$$

   where $a_k > 0$.

3.

Theorem 6.1 (Robbins-Monro theorem). *In the Robbins-Monro algorithm in* (6.5), *if*

(a) $0 < c_1 \leq \nabla_w g(w) \leq c_2$ *for all* w*;*

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

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

where $\mathcal{H}_k = \{w_k, w_{k-1}, \ldots\}$, *then* w_k *almost surely converges to the root* w^* *satisfying* $g(w^*) = 0$.

(a) makes sure g(w) is monotonically increasing and can avoid g(w) is too large which will lead to the divergence of the algorithm.

(b) makes sure that the learning rate is decreasing but should not be that fast.

(c) makes sure that the existed error terms {η}'s expectation is always around 0, and every η_k should not be so far away from 0.

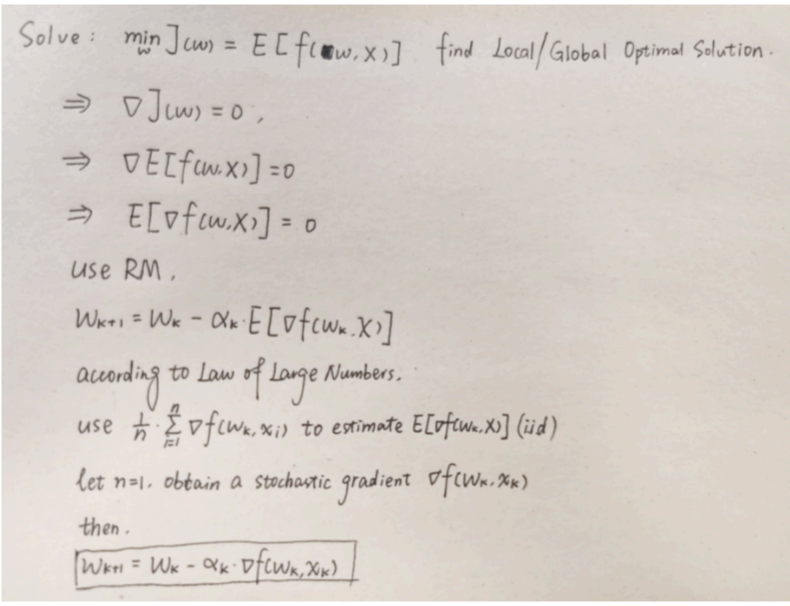
4. To solve w = E[x], we define g(w) = w - E[x], then the prob is converted to solve g(w) = 0.

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

where (w_k - x_k) is g˜(w_k,η_k), which is:

$$\begin{aligned}\tilde{g}(w, \eta) &= w - x \\ &= w - x + \mathbb{E}[X] - \mathbb{E}[X] \\ &= (w - \mathbb{E}[X]) + (\mathbb{E}[X] - x) \doteq g(w) + \eta\end{aligned}$$

5. SGD is used for solving optimization problem.



Solve: $\min_w J(w) = E[f(w,X)]$ find Local/Global Optimal Solution.

$\Rightarrow \nabla J(w) = 0$,

$\Rightarrow \nabla E[f(w,X)] = 0$

$\Rightarrow E[\nabla f(w,X)] = 0$

use RM,

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

according to Law of Large Numbers,

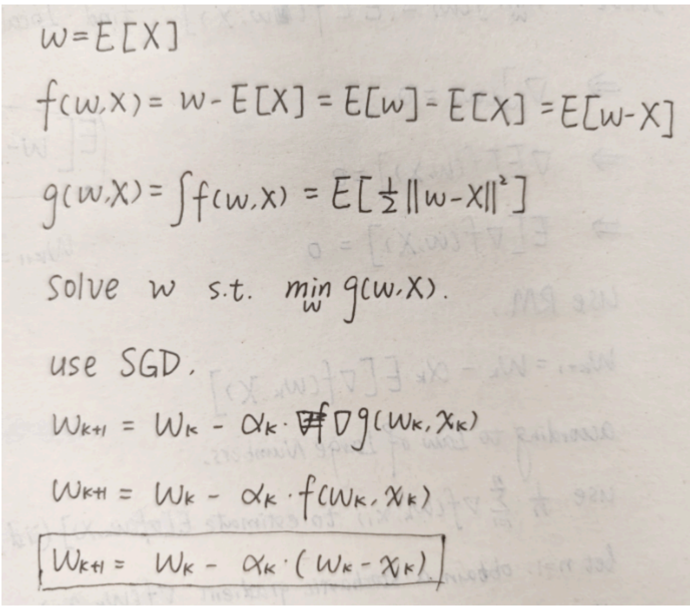
use $\frac{1}{n} \cdot \sum_{i=1}^{n} \nabla f(w_k, x_i)$ to estimate $E[\nabla f(w_k, X)]$ (iid)

let n=1. obtain a stochastic gradient $\nabla f(w_k, x_k)$

then.

$\boxed{w_{k+1} = w_k - \alpha_k \cdot \nabla f(w_k, x_k)}$

6.



$w = E[X]$

$f(w,X) = w - E[X] = E[w] - E[X] = E[w-X]$

$g(w,X) = \int f(w,X) = E[\frac{1}{2}\|w-X\|^2]$

Solve w s.t. $\min_w g(w,X)$.

use SGD,

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

$w_{k+1} = w_k - \alpha_k \cdot f(w_k, x_k)$

$\boxed{w_{k+1} = w_k - \alpha_k \cdot (w_k - x_k)}$

7. If X has n values, 1 < m < n, then

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

$$w_{k+1} = w_k - \alpha_k \frac{1}{m} \sum_{j \in \mathcal{I}_k} \nabla_w f(w_k, x_j), \qquad \text{(MBGD)}$$

$$w_{k+1} = w_k - \alpha_k \nabla_w f(w_k, x_k). \qquad \text{(SGD)}$$

8.

$$
\begin{aligned}
&(a) \quad 0 < c_1 \leq \nabla_w^2 f(w, X) \leq c_2; \\
&(b) \quad \textstyle\sum_{k=1}^{\infty} a_k = \infty \ and \ \sum_{k=1}^{\infty} a_k^2 < \infty; \\
&(c) \quad \{x_k\}_{k=1}^{\infty} \ are \ i.i.d.
\end{aligned}
$$

(a) makes sure that f(w,X) is convex function and ∇f(w,X) is not that large to avoid gradient explosion.

(b) makes sure that the learning rate is decreasing but should not be that fast.

(c) is a general condition.

9. If the estimate w_k is far from the solution w^*, then it converges fast. By contrast, it behaves more randomly and slow when they are close.