Paper Summary[A stochastic approximation method (1951)]

Paper Summary[A stochastic approximation method (1951)]
- Problems (WHAT & WHY)
- Idea (HOW)
- Details (HOW)
- Application (WHERE)
- Summary
- Reference

Problems (WHAT & WHY)

Sometimes in practice (bioassay, sensitivity data), the original data $M(x)$ cannot be observed, instead, researchers are given a data Y which has the property:

$$E(Y|X=x)=M(x) \tag{1}$$

With data Y known and $M(x)$ unknown, we want to find the root of an equation

$$M(x)=\alpha \tag{2}$$

Idea (HOW)

Suppose θ is the only root for equation (2), authors came to an idea that approaching θ step by step. Formally, they wanted to construct a sequence $(x_1, x_2, \dots, x_n)$ such that:

$$\lim_{n\to\infty} x_n = \theta \tag{3}$$

Details (HOW)

First, initiate x_1 randomly, and then set $\{x_n\}$ as:

$$x_{n+1}-x_n = a_n\left(\alpha - y_n\right) \tag{4}$$

where α_n is the learning rate.

Notice that the stochastics method is an online learning way: it uses new-received information $(x_1, \ldots, x_{n-1};\ M(x_1), \ldots, M(x_{n-1});\ M'(x_1), \ldots, M'(x_{n-1}))$ when updating x_n's value.

To make sure that $\{x_n\}$ goes to θ, consider the "distance" between x_n and θ:

$$b_n = \mathbb{E}\left[(x_n - \theta)^2\right] \tag{5}$$

If we can show that $\lim_{n\to\infty} b_n = 0$, then x_n can be proved to converge to θ in probability, which is what we want. To achieve this, the authors listed some conditions:

$$\sum_{n=2}^{\infty} \frac{a_n}{a_1 + \cdots + a_{n-1}} = \infty, \sum_{n=1}^{n\to\infty} a_n^2 < \infty \tag{6}$$

$$\exists C > 0, s.t. \Pr[|Y(x)| \le C] = 1 \tag{7}$$

$$M(x)\ is\ nondecreasing \tag{8}$$

$$M'(\theta) > 0 \tag{9}$$

Condition (6) is interesting, it shows that the learning rate $\{\alpha_n\}$ should be divergent sequence like $\frac{1}{n}$.

The detailed proof is omitted here.

Application (WHERE)

In machine learning, $M(x)$ is often the derivative of the loss function and α is set to 0 to search for the optimal point. This is the famous **stochastic gradient descent** (SGD) method.

Besides, when $M(x)$ is a distribution function $F(x)$, we are indeed finding the α-quantile by the stochastic method.

Summary

The authors proposed a stochastic method for solving a very important problem: how to find a function's root without direct information of the function. Nowadays with the development of machine learning, the goal function is more and more complex. Both researchers and engineers are more and more likely to use SGD to solve equations.

The paper's proof part is jumped here but it is actually full of wisdom, all proving ideas make sense, and the design of a sequence $\{k_n\}$ is intuitive.