第六课 stochastic approximation

1.stochastic approximation

Stochastic approximation (SA):

- SA refers to a broad class of stochastic iterative algorithms solving root finding or optimization problems.
- Compared to many other root-finding algorithms such as gradient-based methods, SA is powerful in the sense that it does *not* require to know the expression of the objective function nor its derivative.

2.Robbins-Monro algorithm

RM是SA领域里的开创性算法

目标：求解g(x)=0

The Robbins-Monro (RM) algorithm can solve this problem:

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

where

- w_k is the kth estimate of the root
- $\tilde{g}(w_k, \eta_k) = g(w_k) + \eta_k$ is the kth noisy observation
- a_k is a positive coefficient.

The function $g(w)$ is a black box! This algorithm relies on data:

- Input sequence: $\{w_k\}$
- Noisy output sequence: $\{\tilde{g}(w_k, \eta_k)\}$

Philosophy: without model, we need data!

- Here, the model refers to the expression of the function.

收敛条件:

Theorem (Robbins-Monro Theorem)

In the Robbins-Monro algorithm, if

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

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

条件一要求函数单增

条件二要求ak收敛到0，且不能收敛的太快

例，ak=1/k满足条件二。实际中常会将ak设为一个非常小的正数。

3.Stochastic gradient descent(SGD)

可以证明，SGD算法就是特殊的RM算法。

注，对于损失函数loss，我们的目标是求其导数为0的点，因此使用RM算法的条件是loss的二阶段大于0，也就是凸函数。

SGD性质：当所求参数离真实点很远时，SGD中使用真实样本数据代替期望值的误差很小；而只有当所求参数就在真实点附近时，才会有明显的震荡。

有两种形式的SGD：

带随机变量的SGD

Suppose we aim to solve the following optimization problem:

$$\min_{w} \quad J(w) = \mathbb{E}[f(w, X)]$$

- w is the parameter to be optimized.
- X is a random variable. The expectation is with respect to X.
- w and X can be either scalars or vectors. The function $f(\cdot)$ is a scalar.

和更常见的不带随机变量的SGD

- The formulation of SGD we introduced above involves random variables and expectation.
- One may often encounter a deterministic formulation of SGD without involving any random variables.

Consider the optimization problem:

$$\min_{w} \quad J(w) = \frac{1}{n}\sum_{i=1}^{n} f(w, x_i),$$

- $f(w, x_i)$ is a parameterized function.
- w is the parameter to be optimized.
- a set of real numbers $\{x_i\}_{i=1}^{n}$, where x_i does not have to be a sample of any random variable.

两者可以通过如下推理统一起来

A quick answer to the above questions is that we can introduce a random variable manually and convert the *deterministic formulation* to the *stochastic formulation* of SGD.
In particular, suppose X is a random variable defined on the set $\{x_i\}_{i=1}^{n}$. Suppose its probability distribution is uniform such that

$$p(X = x_i) = 1/n$$

Then, the deterministic optimization problem becomes a stochastic one:

$$\min_{w} \quad J(w) = \frac{1}{n}\sum_{i=1}^{n} f(w, x_i) = \mathbb{E}[f(w, X)].$$

$\{x_i\}$
$\downarrow$
x_k

- The last equality in the above equation is strict instead of approximate. Therefore, the algorithm is SGD.
- The estimate converges if x_k is *uniformly* and independently sampled from $\{x_i\}_{i=1}^{n}$. x_k may repeatedly take the same number in $\{x_i\}_{i=1}^{n}$ since it is sampled randomly.