Mean estimation

Suppose

$$w_{k+1} = \frac{1}{k}\sum_{i=1}^{k} x_i, \quad k = 1, 2, \ldots$$

Then, w_{k+1} can be expressed in terms of w_k as

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

> Consider an algorithm a more general expression
>
> $$w_{k+1} = w_k - a_k(w_k - x_k)$$
>
> in future discussions.

Robbins-Monro algorithm

Stochastic approximation (SA):

- SA refers to a broad class of stochastic iterative algorithms solving root finding or optimization problems.
- SA does not require to know the expression of the objective function nor its derivative.

Problem statement:

Suppose we would like to find root of equation

$$g(w) = 0$$

> Many problem can be converted to this problem. Like $J(w)$ is to be minimized. Then we can use
>
> $$g(w) = \nabla_w J(w) = 0$$

How to calculate the equation when expression of the g is **unknown**?

Robbins-Monro algorithm - The algorithm

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

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

- $a_k > 0$

The algorithm relies on data:

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

Robbins-Monro algorithm - Convergence properties

Theorem (Robbins-Monro theorem):

> If
>
> 1. $0 < c_1 \leq \nabla_w g(w) \leq c_2$ 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 $\mathcal{H}_k = \{w_k, w_{k-1}, \ldots\}$, then w_k converges with probability 1 (w.p.1) to the root w^* satisfying $g(w^*) = 0$.

- $0 < c_1 \leq \nabla_w g(w) \leq c_2$ for all w

  This condition indicates

  - g to be monotonically increasing. So root of $g(w) = 0$ exists and is unique.
  - The gradient is bounded.
- $\sum_{k=1}^{\infty} a_k = \infty$ and $\sum_{k=1}^{\infty} a_k^2 < \infty$

  Ensures that a_k converges to zero as $k \to \infty$ and not converge to fast.

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

    Indicates that $a_k \to 0$ and $k \to \infty$.

    Since

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

    And if $a_k \to 0$, $w_{k+1} - w_k \to 0$, which grants that w_k converges.

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

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

    If $\sum_{k=1}^{\infty} a_k^2 < \infty$, then $-\sum_{k=1}^{\infty} a_k \tilde{g}(w_k, \eta_k)$ may be bounded.
- $\mathbb{E}[\eta_k|\mathcal{H}_k] = 0$ and $\mathbb{E}[\eta_k^2|\mathcal{H}_k] < \infty$

  $\{\eta_k\}$ is an i.i.d. (independent and identically distributed) stochastic sequence $\{\eta_k\}$ satisfying $\mathbb{E}[\eta_k] = 0$ and $\mathbb{E}[\eta_k^2] < \infty$

Robbins-Monro algorithm - Apply to mean estimation

Consider a function:

$$g(w) = w - \mathbb{E}[X]$$

Our aim is to solve $g(w) = 0$.

Note that

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

The form satisfies RM algorithm. So we can solve $g(w) = 0$ by using RM algorithm

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

Stochastic gradient descent (SGD)

Suppose we aim to solve following optimization problem:

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

- 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.

SGD:

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

Example

Now consider:

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

where

$$f(w, X) = \frac{1}{2}||w - X||^2 \quad \nabla_w f(w, X) = w - X$$

- The GD algorithm for solving this problem:

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

- The SGD algorithm for solving this problem:

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

Convergence

From GD to SGD:

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

$\nabla_w f\left(w_k, x_k\right)$ can be viewed as a noisy measurement of $\mathbb{E}\left[\nabla_w f\left(w_k, X\right)\right]$:

$$\nabla_w f(w_k, x_k) = \mathbb{E}[\nabla_w f(w, X)] + \underbrace{\nabla_w f(w_k, x_k) - \mathbb{E}[\nabla_w f(w, X)]}_{\eta}.$$

Now we only need to show that **SGD is a special RM algorithm**.

The aim of SGD is minimize

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

And we can convert it to find the root of $g(w) = 0$, where

$$g(w) = \nabla J_w(w)$$

What we can measure is

$$\tilde{g}(w, \eta) = \nabla_w f(w, x) = \underbrace{\mathbb{E}[\nabla_w f(w, X)]}_{g(w)} + \underbrace{\nabla_w f(w, x) - \mathbb{E}[\nabla_w f(w, X)]}_{\eta}.$$

Then, the RM algorithm for solving $g(w) = 0$ is

$$w_{k+1} = w_k - a_k \tilde{g}(w_k, \eta_k) = w_k - \alpha_k \nabla_w f\left(w_k, x_k\right)$$

Theorem (Convergence of SGD):

If

1. $0 < c_1 \leq \nabla_w^2 f(w, X) \leq c_2$;
2. $\sum_{k=1}^{\infty} a_k = \infty$ and $\sum_{k=1}^{\infty} a_k^2 < \infty$;
3. $\{x_k\}_{k=1}^{\infty}$ is i.i.d.;

then w_k converges to the root of $\nabla_w \mathbb{E}[f(w, X)] = 0$ w.p.1.

> Here we assume that w is scalar. If w is vector then $\nabla_w^2 f(w, X)$ is the well-known **Hessian matrix**.

Proof:

- $0 < c_1 \leq \nabla_w g(w) \leq c_2$;

  Since $0 < c_1 \leq \nabla_w^2 f(w, X) \leq c_2$, so $\nabla_w g(w) = \nabla_w \mathbb{E}[\nabla_w f(w, X)] = \mathbb{E}[\nabla_w^2 f(w, X)]$ satisfies $0 < c_1 \leq \nabla_w g(w) \leq c_2$
- $\sum_{k=1}^{\infty} a_k = \infty$ and $\sum_{k=1}^{\infty} a_k^2 < \infty$;

  Same as Robbins-Monro theorem.
- $\mathbb{E}[\eta_k | \mathcal{H}_k] = 0$ and $\mathbb{E}[\eta_k^2 | \mathcal{H}_k] < \infty$;

  Since $\{x_k\}$ is i.i.d., $\mathbb{E}_{x_k}[\nabla_w f(w, x_k)] = \mathbb{E}[\nabla_w f(w, X)]$ holds for all k. Therefore,

  $$\mathbb{E}[\eta_k | \mathcal{H}_k] = \mathbb{E}[\nabla_w f(w_k, x_k) - \mathbb{E}[\nabla_w f(w_k, X)] | \mathcal{H}_k]$$

  Since $\mathcal{H} = \{w_k, w_{k-1}, \ldots\}$ and x_k is independent of $\mathcal{H}_k$, $\mathbb{E}[\nabla_w f(w_k, x_k) | \mathcal{H}_k] = \mathbb{E}_{x_k}[\nabla_w f(w_k, x_k)]$ and $\mathbb{E}[\mathbb{E}[\nabla_w f(w_k, X)] | \mathcal{H}_k] = \mathbb{E}[\nabla_w f(w_k, X)]$. Therefore

$$\mathbb{E}[\eta_k|\mathcal{H}_k] = \mathbb{E}_{x_k}[\nabla_w f(w_k, x_k)] - \mathbb{E}[\nabla_w f(w_k, X)] = 0$$

Similarly, it can be proven that $\mathbb{E}[\eta_k^2|\mathcal{H}_k] < \infty$.

Convergence pattern

Convergence is not slow:

Consider the relative error:

$$\delta_k = \frac{|\nabla_w f(w_k, x_k) - \mathbb{E}[\nabla_w f(w_k, X)]|}{|\mathbb{E}[\nabla_w f(w_k, X)]|}$$

Since $\mathbb{E}[\nabla_w f(w^*, X)] = 0$, we further have

$$\delta_k = \frac{|\nabla_w f(w_k, x_k) - \mathbb{E}[\nabla_w f(w_k, X)]|}{|\mathbb{E}[\nabla_w f(w_k, X)] - \mathbb{E}[\nabla_w f(w^*, X)]|} = \frac{|\nabla_w f(w_k, x_k) - \mathbb{E}[\nabla_w f(w_k, X)]|}{|\mathbb{E}[\nabla_w^2 f(\widetilde{w}_k, X)(w_k - w^*)]|}$$

where use mean value theorem and $\widetilde{w}_k \in [w_k, w^*]$

Suppose f is strictly convex such that

$$\nabla_w^2 f \geq c > 0$$

for all w, X.

Then, the denominator of δ_k becomes

$$|\mathbb{E}[\nabla_w^2 f(\widetilde{w}_k, X)(w_k - w^*)]| = |\mathbb{E}[\nabla_w^2 f(\widetilde{w}_k, X)||(w_k - w^*)]| \geq c|w_k - w^*|$$

So

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

Which implies that when $|w_k - w^*|$ is large, δ_k is small and SGD behaves like GD.

A deterministic formulation:

> Example:
>
> $$\min_w J(w) = \frac{1}{n}\sum_{i=1}^{n} f(w, x_i)$$
>
> - x_i does not have to be a sample of any **random variable**.

The gradient descent algorithm for solving this problem is:

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

Suppose the set $\{x_i\}$ is large and we can only fetch a single number every time. In this case, we can use

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

Suppose the probability distribution

$$p(X = x_i) = \frac{1}{n}$$

Then, the deterministic optimization problem becomes a stochastic one:

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

BGD, MBGD, and SGD

The algorithms solve the problem that minimizing $J(w) = E[f(w, X)]$ (where X consists of a set of random samples $\{x_i\}_{i=1}^n$) are:

$$\begin{aligned}
w_{k+1} &= w_k - \alpha_k \frac{1}{n}\sum_{i=1}^{n} \nabla_w f(w_k, x_i), && \text{(BGD)} \\
w_{k+1} &= w_k - \alpha_k \frac{1}{m}\sum_{j\in\mathcal{I}_k} \nabla_w f(w_k, x_j), && \text{(MBGD)} \\
w_{k+1} &= w_k - \alpha_k \nabla_w f(w_k, x_k). && \text{(SGD)}
\end{aligned}$$

- BGD uses all samples per iteration.
- MBGD uses i.d.d. samplings m times.
- SGD randomly samples one time.