Supplementary Material

ESC 2024 Winter Session 2nd Week

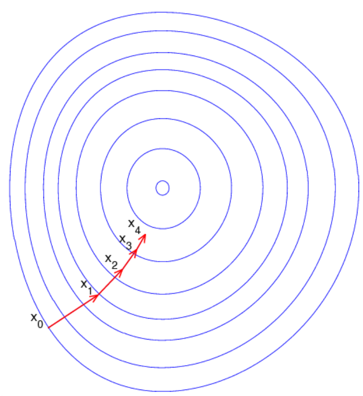
전인태

1. Gradient-Descent Method

Let $f(\mathbf{x})$ be a function that we want to optimize and set $\mathbf{x}_0$ be an initial point. Given $\mathbf{x}_i$, next point $\mathbf{x}_{i=1}$ is computed by

$$\mathbf{x}_{i+1} = \mathbf{x}_i - \eta_i \nabla f(\mathbf{x}_i)$$

where η_i is a hyperparameter meaning the learning rate, which is the distance to move at each iteration. If η_i is too huge, then it may not converge to some point ; if η_i is too small, it may get stuck in some local optimal points.



2. Robbins-Monro Algorithm

Consider θ and z governed by $p(z, \theta)$ and define the regression function

$$f(\theta) = E[z|\theta] = \int zp(z|\theta)dz.$$

We'd like to seek θ^* such that $f(\theta^*) = 0$.
Assume we are given samples from $p(z, \theta)$, one at a time. Successive estimates of θ^* are then given by

$$\theta^{(N)} = \theta^{(N-1)} - a_{N-1} z(\theta^{(N-1)}).$$

Conditions on a_N for convergence :

$$\lim_{N\to\infty} a_N = 0, \ \sum_{N=1}^{\infty} a_N = \infty, \ \sum_{N=1}^{\infty} a_N^2 < \infty$$

Regarding

$$-\lim_{N\to\infty}\frac{1}{N}\sum_{n=1}^{N}\frac{\partial}{\partial\theta}\ln p(x_n|\theta) = E\left[-\frac{\partial}{\partial\theta}\ln p(x|\theta)\right]$$

as a regression function, finding its root is equivalent to finding the maximum likelihood solution θ_{ML}. Thus

$$\theta^{(N)} = \theta^{(N-1)} - a_{N-1}\frac{\partial}{\partial\theta^{(N-1)}}\left[-\ln p(x_N|\theta^{(N-1)})\right].$$

Example : Updating μ_{ML} of the Gaussian Distribution
Suppose $X \sim N(\mu, \sigma^2)$ and we are going to be provided one sample data at a time. The probability density function $f(x)$ is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

Fix σ for convenience. Since $\log f(x)$ partially differentiated with respect to μ is

$$\frac{x-\mu}{\sigma^2}.$$

If we substitute $\ln p(x_N|\theta^{(N-1)}$ with $\partial f(x)/\partial\mu$, we have that

$$\mu^{(N)} = \mu^{(N-1)} + a_{N-1}\frac{x-\mu^{(N-1)}}{\sigma^2}.$$

Note that the parameter we'd like to estimate here is μ, then $\theta := \mu$.

3. Newton-Rhaphson Method

The Newton's method is an idea to approximate the root of a real-valued function. If the tangent line to the curve $f(x)$ at $x = x_n$ intercepts the x-axis at x_{n+1}, then the slope is

$$f'(x_n) = \frac{f(x_n) - 0}{x_n - x_{n+1}}.$$

Solving for x_{n+1} gives

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}.$$