MAT 653: Statistical Simulation

Instructor: Dr. Wei Li Scribe: Yuan Liu

Sep 23th , 2021

Necessary Conditions

(a) First-Order necessary condition

Fact: Let f be continuously differentiable, if x^* is a minimizer of f, then $\nabla f(x^*) = 0$.

(b) Second-order necessary condition

Fact: let f be continuously twice differentiable, if x^* is a minimizer of f, then $\nabla f(x^*) = 0$, and $\nabla^2 f(x^*) \geq 0$.

Sufficient Conditions(2nd order condition)

Fact: suppose $\nabla^2 f(x)$ is continuous and at some point x^*, if $\nabla^2 f(x^*) > 0$, and $\nabla f(x^*) = 0 \Rightarrow$ then x^* is a minimizer(strict local).

Convex function and global minimizer

Fact: when f is convex on S, then any local minimizer x^* is a global minimizer.

If in addition, f is differentiable, then any stationary point is a global minimizer.

f: $\mathbb{R}^n \to \mathbb{R}$ Hessian matrix at x^*, such that $\nabla f(x^*) = 0$

1 If $\nabla^2 f(x^*) > 0$, then any local minimizer x^* is a strict local minimizer.

2 If $\nabla^2 f(x^*) < 0$, then local maximizer x^* is a strict local maximizer.

3 If $\nabla^2 f(x^*)$ is indefinite(i.e. neither positive semi-definite nor negative semi-definite), then the local minimizer x^* is a saddle point.

4 If those cases not listed above, then the test is inconclusive.

Numerical method for optimization

$minf(x) \to x^*$

Two methods:

1. Gradient method

2. Newton's method

Both guarantee that the stationary points fo f can be found (i.e. $\nabla f(x^*) = 0$).

Steps

1 start the process with some initial point x_0;

2 then iterate steps denoted by $x_k \to x_{k+1}$ going downhill toward a stationary point of f.

Repeat **2** until the sequance of points converge to a stationary point.

For a general(non-convex) function, we run the procedure several times with diffrent initial values x_0.

Gradient descent

Choose a direction P_k and search along the direction from the current iterate x_k for a new iterate with lower function value.

$f(x_{k+1}) < f(x_k)$

$x_k + \alpha P_k$ (α = "step length", "learning rate", $\alpha > 0$ scalar)

Fix α : Taylor approximation of f at x_k

$f(x_k + \alpha p) \approx f(x_k) + \alpha p^T \nabla f(x_k)$

$\min_{||p||=1} P^T \nabla f(x_k)$: solution gives us the unit direction that is most rapid decrease.

$p^T \nabla f(x_k) = ||p|| \cdot ||\nabla f(x_k)|| cos(\theta)$, $0 \leq \theta \leq \pi$

$\Rightarrow cos(\theta) = -1$

$\Rightarrow$ p is the exact opposite direction of $\nabla f(x_k)$, $p = \frac{-\nabla f(x_k)}{||\nabla f(x_k)||}$

Any P_k such that $P_k^T \nabla f(x_k) < 0$ would work. It's called "descent direction".

Algorithm

Gradient descent (fix α), set $k = 0$, given x_0

Repeat $x_{k+1} \leftarrow x_k - \alpha \nabla f(x_k)$, k $\leftarrow$ k+1

Until stopping condition is met ($||\nabla f(x_k)|| = 0$).