Class #6

Friday, February 7, 2025 10:20 AM

Stopping Criterion

- Objective function stops changing significantly

$$\left| f(x^{(k)}) - f(x^{(k-1)}) \right| \le \epsilon \quad \text{for } \epsilon \text{ small}$$

e.g., $\epsilon = 10^{-8}$

- Iterates stop changing significantly

$$\| x^{(k)} - x^{(k-1)} \| \le \epsilon \quad \text{for } \epsilon \text{ small}$$

- Function gradient evaluated at the iterates becomes very small

$$\| \nabla f(x^{(k)}) \| \le \epsilon \quad \text{for } \epsilon \text{ small}$$

Descent Optimization Methods

An optimization method is termed a 'descent method' if

$$f(x^{(k+1)}) < f(x^{(k)}) \quad \forall\, k, \text{ except when}$$

$x^{(k)} = x^*$.

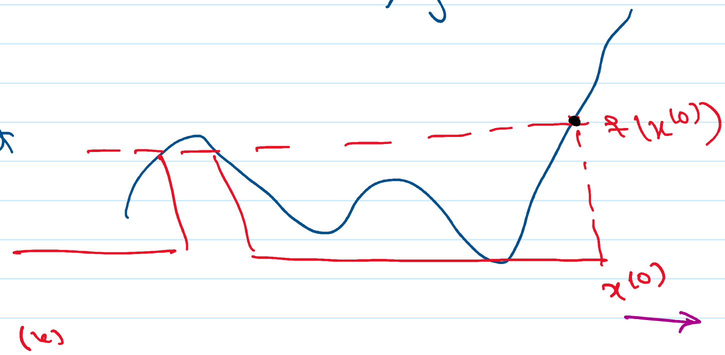
strict inequality

When we initialize at $x^{(0)}$

All iterates $x^{(k)}$ for a descent method stay with the set the set

$$S = \{x : f(x) \le f(x^{(0)})\}$$

Sublevel set of $f(x)$ at $f(x) = f(x^{(0)})$



$$x^{(k+1)} \leftarrow x^{(k)} + t^{(k)} \Delta x^{(k)}$$

In descent methods, $\Delta x^{(k)}$ is called descent direction.

What direction is a descent direction?

We will focus on all continuously differentiable functions and answer this question.

$$C^1 = \{f(x) : f(x) \text{ has continuous derivatives at all } x \in \operatorname{dom} f\}$$

① Make use of Taylor's theorem

Since $f \in C^1$, we have that

$$f(z) = f(x) + \nabla f(x)^T (z-x) + h(x)\|z-x\|_2$$

where $h(x) \to 0$ as $z \to x$ faster than $\|z-x\|_2$

$$f(z) = f(x) + \nabla f(x)^T (z-x) + o(\|z-x\|_2)$$

e.g. $g(n) = n^{-(2+\varepsilon)}$

$$\frac{h(x)\,\|z-x\|_2}{\|z-x\|_2} \to 0$$

e.g.; $g(n) = n^{-(2+\varepsilon)}$

$g(n) = o(n^{-2})$

Take

$x = x^{(k)}$

$z = x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x^{(k)}$

$$f(x^{(k+1)}) = f(x^{(k)}) + \nabla f(x^{(k)})^T \left(t^{(k)} \Delta x^{(k)}\right) + o\left(\|t^{(k)} \Delta x^{(k)}\|_2\right)$$

$$= f(x^{(k)}) + t^{(k)} \nabla f(x^{(k)})^T \Delta x^{(k)} + \underbrace{o\left(t^{(k)} \|\Delta x^{(k)}\|_2\right)}_{\approx 0}$$

Let $t^{(k)} \to 0$ (very small)

$$f(x^{(k+1)}) = f(x^{(k)}) + t^{(k)} \nabla f(x^{(k)})^T \Delta x^{(k)}$$

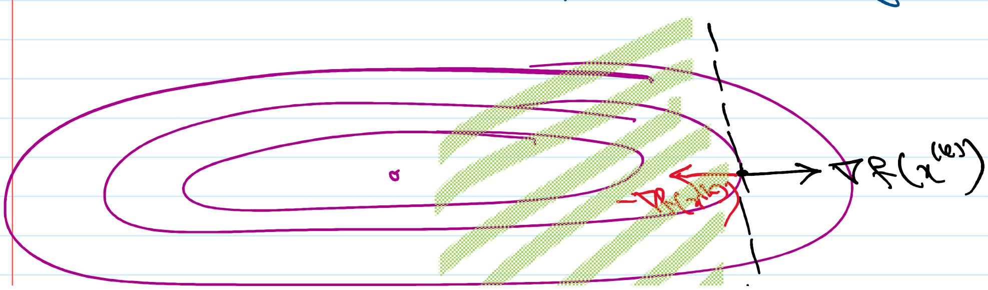
$$f(x^{(k+1)}) < f(x^{(k)})$$

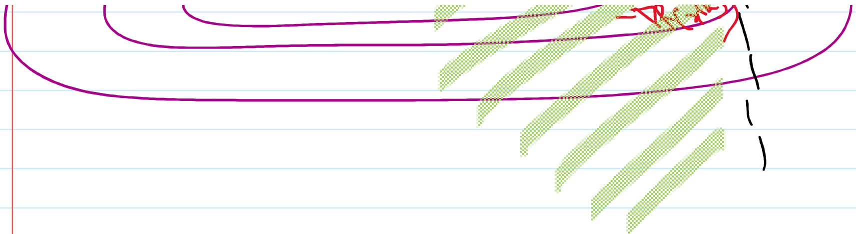
$$\Longleftrightarrow \nabla f(x^{(k)})^T \Delta x^{(k)} < 0$$

$$\Longleftrightarrow \left[-\nabla f(x^{(k)})\right]^T \Delta x^{(k)} > 0$$

$\Delta x^{(k)}$ is a descent if and only if

It makes an acute angle with $-\nabla f(x^{(k)})$





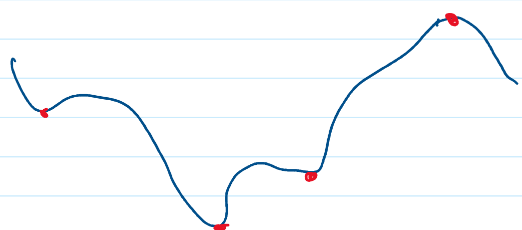
In particular, $\Delta x^{(k)} = -\nabla f(x^{(k)})$ is a descent direction

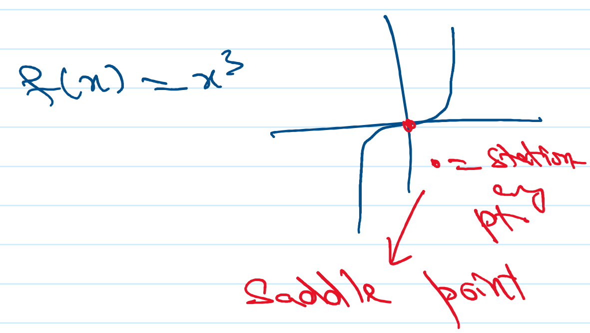
In a descent method, the optimization method 'might' not be able to reduce function value further when

$$\nabla f(x^{(k)}) = 0$$

Strict use of a descent method means that any x for which $\nabla f(x^{(k)}) = 0$ is a fixed point of the method.

Any $x \in \operatorname{dom} f$ for which $\nabla f(x) = 0$ is called a stationary point of f.





General form of descent direction

$\Delta x^{(k)}$ is a descent direction when

$$\boxed{\Delta x^{(k)} = -B^{(k)} \nabla f(x^{(k)})} \longrightarrow (*)$$

where $B^{(k)}$ is a positive definite matrix; $\in S^n_{++}$

Descent methods $\Rightarrow$ $x^{(k+1)} = x^{(k)} - t^{(k)} B^{(k)} \nabla f(x^{(k)})$

$$\left[-\nabla f(x^{(k)})\right]^T \Delta x^{(k)}$$

$$= -\nabla f(x^{(k)})^T \left(-B^{(k)} \nabla f(x^{(k)})\right)$$

$$= \nabla f(x^{(k)})^T B^{(k)} \nabla f(x^{(k)})$$

> 0 b/c $B^{(k)}$ is PD.

Based on the choice of $B^{(k)}$, we have different names for descent methods.

① Gradient descent : $B^{(k)} = I$

$$\Leftrightarrow \Delta x^{(k)} = -\nabla f(x^{(k)})$$

$$x^{(k+1)} = x^{(k)} - t^{(k)} \nabla f(x^{(k)})$$

② Newton's method

$$B^{(k)} = \text{Hessian matrix at } x^{(k)}$$

$$= \nabla^2 f(x^{(k)})$$

Hessian matrix $\Rightarrow$ matrix of second-order partial derivatives.

Another type is called Quasi-Newton method, in which $B^{(k)}$ is built from $\nabla f(x^{(k)})$, but is meant in which $B^{(k)}$ is built from $\nabla f(x^{(k)})$, but is meant to approximate $\nabla^2 f(x^{(k)})$

③ Steepest descent

$B^{(k)}$ is chosen based on the geometry of the function.

Issue: The previous analysis guarantees a descent direction, but only when $t^{(k)}$ is very small.

Can we use descent methods with a larger step size?

↳ we can, but we need to assume additional regularity on the function.

$C_L^1(\mathbb{R}^n) = \{ f(x) :$ f has continuous derivatives that are L-Lipschitz continuous $\}$ → L-Smooth functions

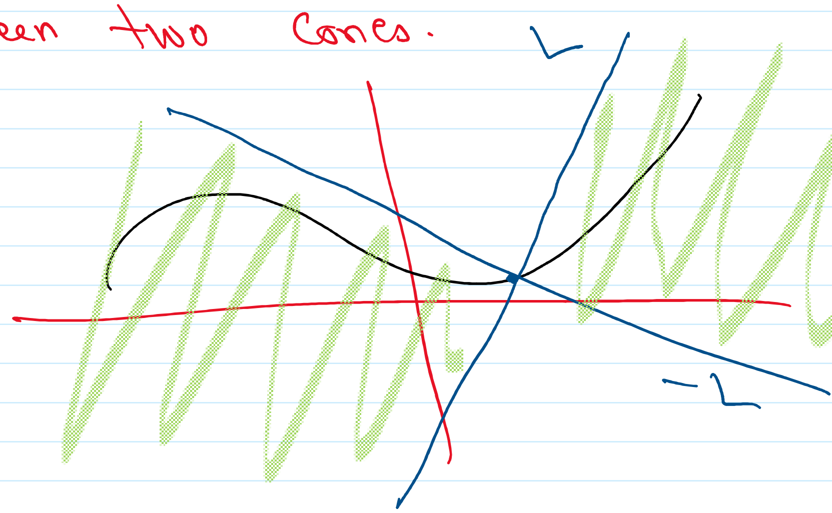
$$C_L^1 \subseteq C^1$$

The gradients $\nabla f(x)$ of $f(x)$ are called L-Lipschitz continous if and only if

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L \|x-y\|_2, \forall x,y$$

L: Lipschitz constant

Lipschitz Continity is a Stronger form of Continuity, Which says that the function value always lies between two Cones.



e.g.; ① $f(x) = x^2$

$$f'(x) = 2x$$

$$|f'(x) - f'(y)| = |2x - 2y| \le 2|x-y|$$

$L = 2$

② $f(x) = x_1^2 + x_2^2 \longrightarrow x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$

$$\nabla f(x) = 2x = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix}$$

$$\|\nabla f(x) - \nabla f(y)\|_2 = \|2x - 2y\|_2 \le 2\|x-y\|_2$$

$L = 2$