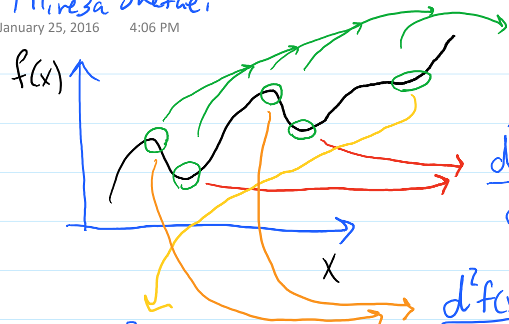
Alireza Shafaei

January 25, 2016 4:06 PM



$\frac{d f(x)}{d x} = 0$, the first derivative is zero

$\frac{d^2 f(x)}{d x^2} > 0 \Rightarrow$ local minima (locally convex)

$\frac{d^2 f(x)}{d x^2} < 0 \Rightarrow$ local maxima (locally concave)

$\frac{d^2 f(x)}{d x^2} = 0$ ↳ Saddle point, change of direction
think of $f(x) = x^3$ at $x = 0$

The second derivative

In multivariate case we similarly use the Hessian to understand the curvature

$\nabla^2 f(x)$ is a squre matrix with eigenvalues λ_i at the point x.

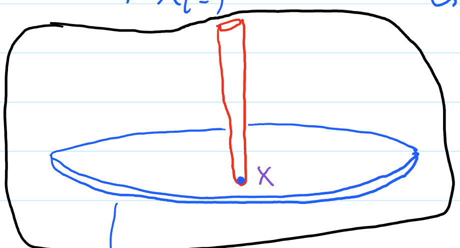
① if $\lambda_i \geq 0$ for all i $\Rightarrow$ locally convex (local minima) (strong if strict)

② if $\lambda_i \leq 0$ for all i $\Rightarrow$ locally concave (local maxima) (strong if strict)

Otherwise we will have some sort of saddle point.

$$\mu I = \begin{bmatrix} \mu & & 0 \\ & \mu & \\ 0 & & \mu \end{bmatrix} \qquad L I = \begin{bmatrix} L & & 0 \\ & L & \\ 0 & & L \end{bmatrix} \qquad \mu I \preceq \nabla^2 f(x) \preceq L I$$

↳ $\lambda_i = \mu$ ↳ $\lambda_i = L$ ↳ Comparison of the eigenvalues.



if true at all x, then we are certain our function does not have a flat surface ($\lambda_i \geq \mu > 0$) hence strongly convex
when upperbounded by L, it means it can't change too quickly hence strongly smooth.

an alternative definition is to say $\mu \|v\|^2 \leq v^T \nabla^2 f(x) v \leq L \|v\|^2$ for all v !
equivalent ($\lambda_1 = \operatorname{argmin} \frac{v^T \nabla^2 f(x) v}{\|v\|^2}$)

↳ roughly saying that the function will be above the blue hemisphere (μ) and below the red hemisphere (L) around x.

We can approximate a function f around the point x using the derivatives of the function at x.

let's say x, $f(x)$, $\frac{d f(x)}{d x}$, and $\frac{d^2 f(x)}{d x^2}$ are given, we then will have

$$f(y) \approx g(y) = \underbrace{f(x)}_{\text{Const}} + \underbrace{\frac{d f(x)}{d x}}_{\text{Const}} (y - x) + \underbrace{\frac{d^2 f(x)}{d x^2}}_{\text{Const}} (y-x)^2 \frac{1}{2}$$

$$f(y) \approx g(y) = \underbrace{f(x)}_{Const} + \underbrace{\frac{df}{dx}}_{Const}(y-x) + \underbrace{\frac{d^2 f}{dx^2}}_{Const}\frac{(y-x)^2}{2}$$

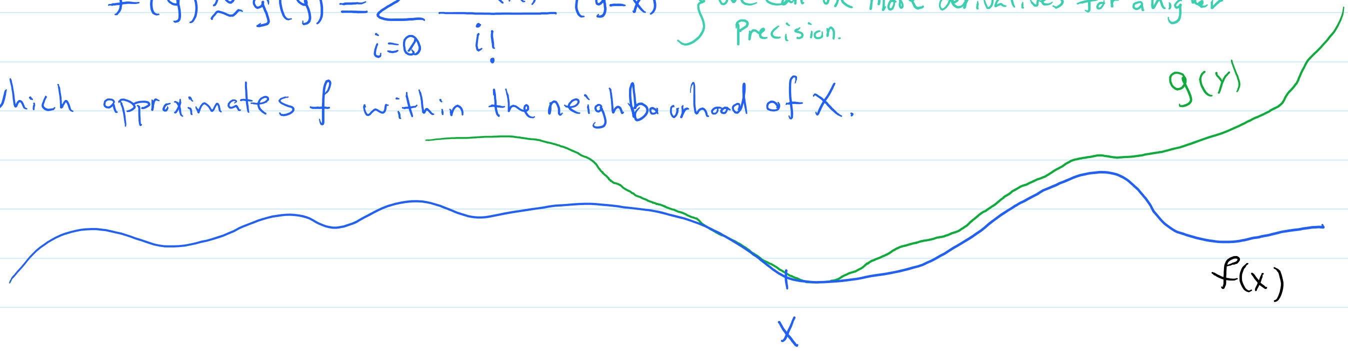
$\hookrightarrow$ g is a function of Y!

more generally a polynomial approximation $g(y)$ can be written as (Taylor expansion)

$$f(y) \approx g(y) = \sum_{i=0}^{\infty} \frac{f^{(i)}(x)}{i!}(y-x)^i$$

} We can use more derivatives for a higher precision.

Which approximates f within the neighborhood of X.



We use a generalization of this in the multivariate case

$$g(y) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(x)(y-x) + \overbrace{O(\|y-x\|^2)}^{\text{The discarded terms}}$$

Can be bounded by this.

or this particular variant that is applicable to our problem (convex optimization)

$$\exists z: \; f(y) = g(y) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}(y-x)^T \nabla^2 f(z)(y-x)$$

To prove the convergence rate of a gradient step $x^{t+1} = x^t - \frac{1}{L}\nabla f(x^t)$ we need to

① Guarantee progress ② Guarantee termination

$\hookrightarrow$ not enough by itself because I can make progress, but small enough that I only converge to a fixed point, e.g., $\sum_{i=0}^{\infty}\frac{1}{2^i} = 2$

$\hookrightarrow$ as a measure of $\boxed{f(x^t) - f(x^*)\ ?}$

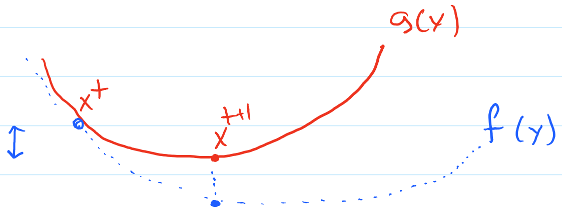
$\|x^t - x^*\|\ ?$

We assume $\underbrace{\mu I \preceq \nabla^2 f(x) \preceq LI}$ for all x.

① Progress. We start with the Taylor expansion

$$f(y) = f(x) + \nabla f(x)^T (y-x) + \frac{1}{2}\underbrace{\underbrace{(y-x)^T}_{v} \nabla^2 f(z) \underbrace{(y-x)}_{v}}_{< L\|v\|^2} \quad \text{for some } z.$$

$$\Rightarrow f(y) \le \underbrace{f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|y-x\|^2}_{g(y)}$$



g is a quadratic upper bound on f!

Since g is always above f, it is a safe bet if I minimize g instead of f!

Set $\nabla g(y) = 0$ to find the minimizer of g. note that $\nabla^2 g(y) \succ 0 \Rightarrow$ Convex

$$\Rightarrow \nabla g(y) = \nabla f(x) + L(y-x) = 0 \Rightarrow \boxed{y = x - \frac{1}{L}\nabla f(x)}$$

if I pick the next point x^{t+1}, I will be jumping to the minimizer of g, and we know the real function will be below g, so we're safe. So, how much progress have we made? plugging the value of y back, we'd get

$$f(y) \le f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|y-x\|^2$$

$$\Rightarrow f(x) - \frac{1}{L}\|\nabla f(x)\|^2 + \frac{1}{2L}\|\nabla f(x)\|^2 = f(x) - \frac{1}{2L}\|\nabla f(x)\|^2$$

So we have decreased $f(x^t)$ by jumping to $f(x^{t+1})$ by at least $\frac{1}{2L}\|\nabla f(x^t)\|^2$

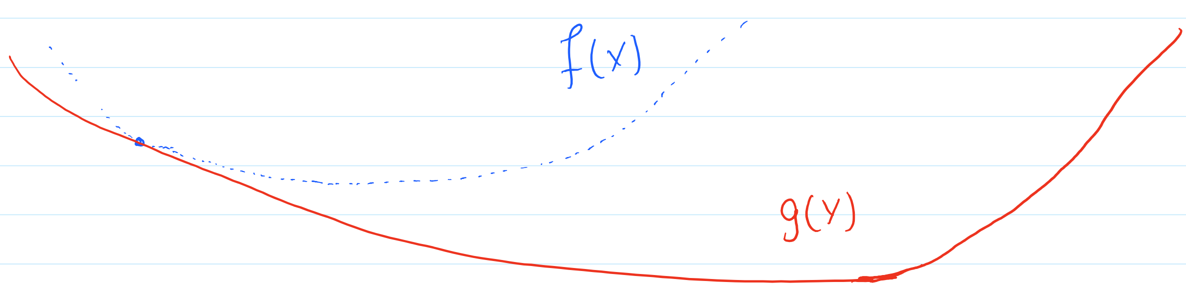
as long as $\|\nabla f(x^t)\|$ is not zero, progress is guaranteed!

---

now from the other side we have

$$f(y) \ge \underbrace{f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2}\|y-x\|^2}_{g(y)}$$

g is always below f



now we have a quadratic lower bound on f, which is useful to bound the maximum distance to the real minimizer because we can minimize $g(y)$ exactly.

if we minimize $g(y)$ we'd get $\underbrace{f(y)}_{\text{function}} \geq \underbrace{f(x) - \frac{1}{2\mu}\|\nabla f(x)\|^2}_{\text{a fixed value}}$ $*$

if $f(y)$ is always above $*$, then $f(y^*)$ is also above $*$

$$\Rightarrow \quad f(x^*) \geq f(x^t) - \frac{1}{2\mu}\|\nabla f(x^t)\|^2 \quad \text{I}$$

we previously had $f(x^{t+1}) \leq f(x^t) - \frac{1}{2L}\|\nabla f(x^t)\|^2$ II

and we wish to show $f(x^*) - f(x^t)$ goes to 0 as t goes to ∞.

$$\text{I} \Rightarrow \mu\left(f(x^*) - f(x^t)\right) \geq -\frac{1}{2}\|\nabla f(x^t)\|^2 + \text{II}$$

$$f(x^{t+1}) \leq f(x^t) + \frac{\mu}{L}\left(f(x^*) - f(x^t)\right) \quad \text{subtract } f(x^*)$$

$$f(x^{t+1}) - f(x^*) \leq \left(1 - \frac{\mu}{L}\right)\left(f(x^t) - f(x^*)\right) \quad \text{III}$$

$\Rightarrow$ This means we get closer to $f(x^*)$ by $\rho = (1 - \frac{\mu}{L})$ with each step.
by applying III to itself repeatedly, we'll have

$$f(x^t) - f(x^*) \leq \rho^t\left(f(x^0) - f(x^*)\right) \qquad \rho < 1$$

which means our distance to the optimal objective is at most ρ^t times our initial distance.

if we wish to get at least ε close to the optimal objective, we can say:

$$f(x^t) - f(x^*) \leq \underbrace{\rho^t\left(f(x^0) - f(x^*)\right)}\leq \varepsilon$$

$$\hookrightarrow \rho^t \leq \varepsilon \qquad \Rightarrow \left(\frac{1}{\rho}\right)^{-t} \leq \varepsilon$$

$$\Rightarrow -t \log\frac{1}{\rho} \leq \log\varepsilon \quad \Rightarrow t \geq \log\frac{1}{\varepsilon} \cdot C \quad \Rightarrow t = O\left(\log\frac{1}{\varepsilon}\right)$$