Lecture 14. Line search, Newton's method.

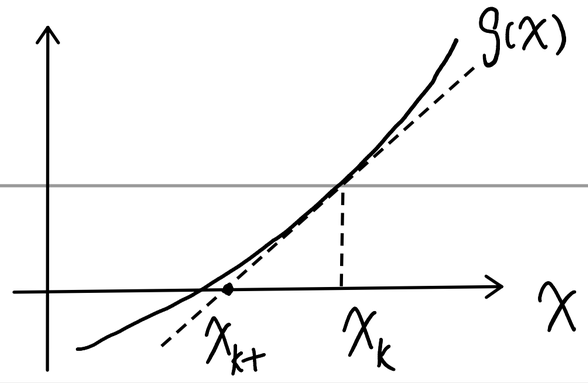
Exact line search: $x_{k+1} = x_k - t\nabla f(x_k)$. $t = \underset{s}{\operatorname{argmin}}\ f(x_k - s\nabla f(x_k))$.

Example. $f(x) = x^TQx + w^Tx$. $Q \succ 0$. let $d_k = \nabla f(x_k) = 2Qx_k + w$.

$$t = \underset{s}{\operatorname{argmin}}\ f(x_k - sd_k) = \underset{s}{\operatorname{argmin}}\ f(x_k) - 2sd_k^TQx_k + s^2d_k^TQd_k - sw^Td_k$$

$$= \underset{s}{\operatorname{argmin}}\ -sd_k^T(2Qx_k + w) + s^2d_k^TQd_k \qquad = \frac{d_k^Td_k}{2d_k^TQd_k}.$$

In general. find the root of gradients. binary search if $\mathbb{R} \to \mathbb{R}$.



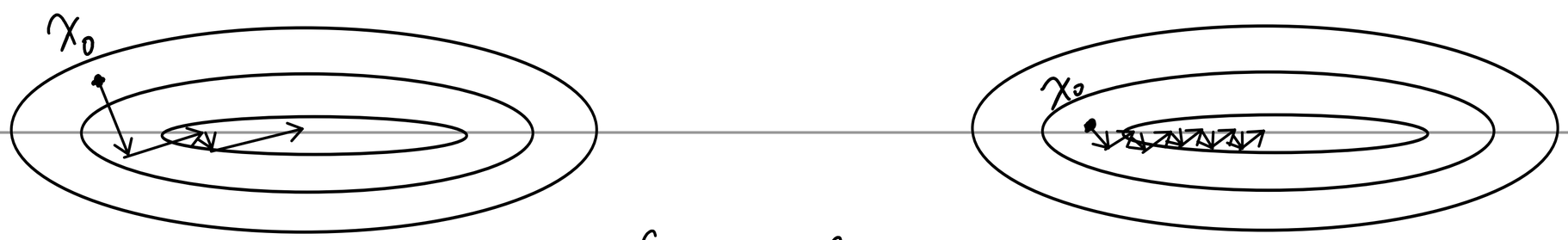
Newton's method: $x_{k+1} = x_k - \dfrac{g(x_k)}{g'(x_k)}$.

why? $g(x) \approx g(x_k) + g'(x_k)(x - x_k)$

Example. calculating $\dfrac{1}{\sqrt{x}}$ in Quake III Arena. 雷神之锤.

let $g(y) = \dfrac{1}{y^2} - x$. $g'(y) = -\dfrac{2}{y^3}$ return $y_0\left(\dfrac{3}{2} - \dfrac{x}{2}y_0^2\right)$.

Proposition. Successive gradient directions are always orthogonal.



$g(s) = f(x_k - s\nabla f(x_k))$

Proof. $0 = g'(t_k) = -\nabla f(x_k - t_k\nabla f(x_k))^T\nabla f(x_k) = -\nabla f(x_{k+1})^T\nabla f(x_k)$. □

Theorem. If f is m-strongly convex and L-smooth. $\{x_k\}$ given by

the gradient descent method with exact line search. then

$$f(x_k) - f^* \le \left(1 - \frac{m}{L}\right)^k (f(x_0) - f^*).$$

Proof. Let $g(s) = f(x_k - s\nabla f(x_k))$. $g(s)$ is unknown but can be bounded.

$$g(s) \le f(x_k) - s\|\nabla f(x_k)\|^2 + \frac{Ls^2}{2}\|\nabla f(x_k)\|^2 \triangleq h(s). \quad t_k = \arg\min_s g(s)$$

$$\Rightarrow f(x_{k+1}) = \min_s g(s) \le \min_s h(s) = h(1/L) = f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2$$

By m-strong convexity. $f(x^*) \ge f(x_k) + \nabla f(x_k)^T(x^* - x_k) + \frac{m}{2}\|x^* - x_k\|^2$

Let $\hat{f}(x) = f(x_k) + \nabla f(x_k)^T(x - x_k) + \frac{m}{2}\|x - x_k\|^2$ be quadratic.

$\hat{f}(x)$ has a minimum value $\hat{f}^*$. So $f(x^*) \ge \hat{f}(x^*) \ge \hat{f}^*$.

$$\nabla\hat{f}(x) = \nabla f(x_k) + m(x - x_k) \Rightarrow \hat{f}^* = \hat{f}\left(x_k - \frac{\nabla f(x_k)}{m}\right)$$

$$\Rightarrow f(x^*) \ge \hat{f}^* = f(x_k) - \frac{1}{m}\|\nabla f(x_k)\|^2 + \frac{1}{2m}\|\nabla f(x_k)\|^2.$$

$$\Rightarrow f(x_{k+1}) \le f(x_k) - \frac{1}{2L}\|\nabla f(x_k)\|^2 \le f(x_k) - \frac{m}{L}(f(x_k) - f^*). \quad \square.$$

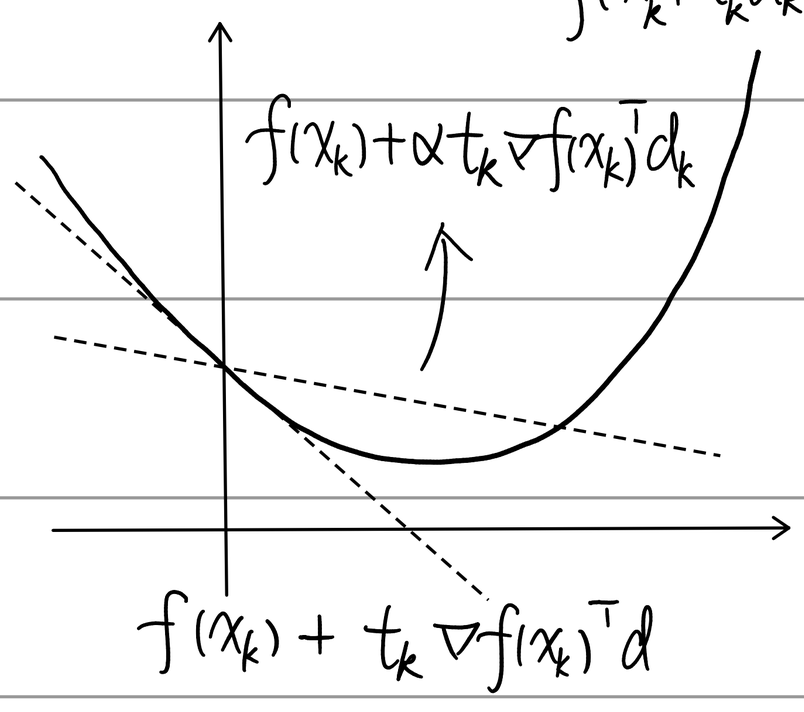
Remark. Exact line search is usually expensive. Is there a simpler way?

Backtracking line search : Armijo's rule

Given a descending direction d_k and $\alpha, \beta < 1$

while $f(x_k + t_k d_k) > f(x_k) + \alpha t_k \nabla f(x_k)^T d_k$

$$t_k = \beta \cdot t_k$$



In particular. $d_k = -\nabla f(x_k)$ in our setting. $\nabla f(x_k)^T d_k = -\|\nabla f(x_k)\|^2$.

Armijo used $\alpha = \beta = 1/2$. in textbook. $\alpha \in [0.01, 0.3]$. $\beta \in [0.1, 0.8]$.

Lower bound for t_k: initially set $t_k = 1$. assume f is L-smooth.

$$g(t) = f(x_k - t\nabla f(x_k)) \le f(x_k) - t\|\nabla f(x)\|^2 + \frac{L}{2}t^2\|\nabla f(x)\|^2.$$

$$\le f(x_k) - \frac{t}{2}\|\nabla f(x_k)\|^2 \quad (\text{if } t \le 1/L).$$

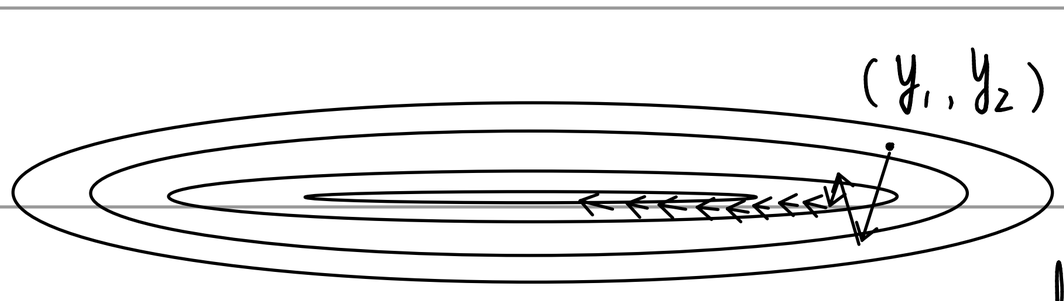
$$\le f(x_k) - \alpha t\|\nabla f(x_k)\|^2 \quad (\text{if } \alpha \le 1/2).$$

So for any $t \le 1/L$. Armijo's rule satisfies. (in general $t \le 2(1-\alpha)/L$ suffices). Thus backtracking terminates with $t_k = 1$, or $t_k > \beta/L$.

Theorem. If f is m-strongly convex and L-smooth. $\{x_k\}$ generated by the gradient descent with backtracking line search. where $0 < \alpha, \beta < 1$. then. $f(x_k) - f^* \le (1 - \min\{2m\alpha, 4m\alpha(1-\alpha)\beta/L\})^k (f(x_0) - f^*)$.

Recall if the condition number is large. gradien descent runs slowly.

Consider the function $f(x_1, x_2) = \frac{1}{100}x_1^2 + x_2^2$. at (y_1, y_2)



$$-\nabla f(y_1, y_2) = (-\frac{1}{50}y_1, -2y_2)^T.$$

locally decreases rapidly but not globally.

ideal descending direction: $(-y_1, -y_2) = -\begin{pmatrix} 50 & 0 \\ 0 & 1/2 \end{pmatrix}\nabla f(y_1, y_2)$.

In general. if $f(x) = x^TQx$. $\nabla f(x) = 2Qx$. $Q > 0$

at $x = (x_1, x_2)$ ideal direction $d = -(x_1, x_2) = -\frac{1}{2} Q^{-1} \nabla f(x)$

Recall Newton's method for finding roots. Consider Taylor series.

$$f(x) \approx f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2} (x - x_k) \nabla^2 f(x_k) (x - x_k) \triangleq g(x).$$

optimize quadratic $g(x)$: $\nabla g(x) = 0 \Rightarrow \nabla f(x_k) + \nabla^2 f(x_k)(x - x_k) = 0$.

Newton's method: $x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$. provided $\nabla^2 f(x_k) > 0$.

Note that if $\nabla^2 f(x_k) > 0$. so is $(\nabla^2 f(x_k))^{-1}$. thus if $\nabla f(x_k) \neq 0$.

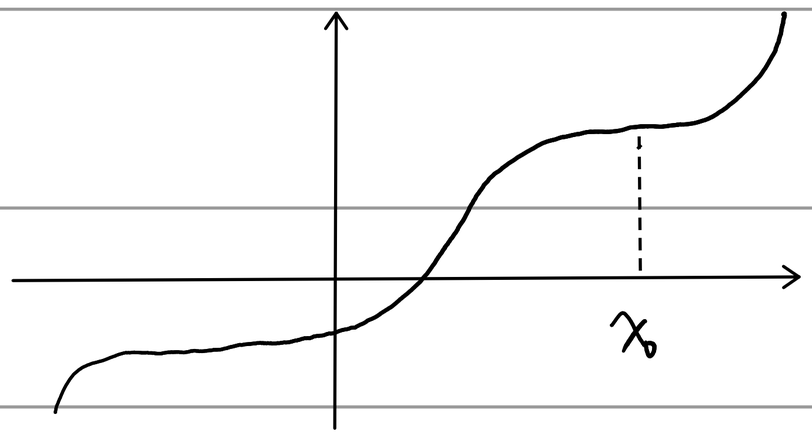
$$\nabla f(x_k)^T \left(-\nabla^2 f(x_k)^{-1} \nabla f(x_k) \right) = - \nabla f(x_k)^T \left(\nabla^2 f(x_k) \right)^{-1} \nabla f(x_k) < 0$$

$-(\nabla^2 f(x_k))^{-1} \nabla f(x_k)$ is a reasonable descending direction. so converge.

Question : step size ? convergence condition and rate ?

Intuitively step size = 1. since we optimize the second-order Taylor.

Recall $t = 1/L$ for GD since $f(x) \leq f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{L}{2} \| x - x_k \|^2$.



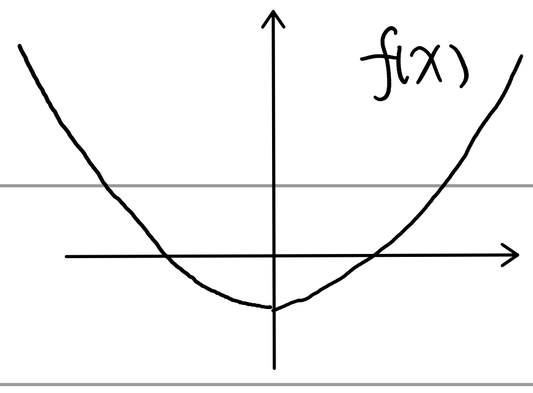
highly depends on the initial point

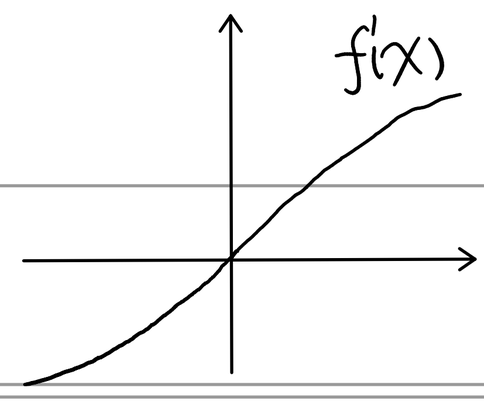
converge rapidly if starting from good point

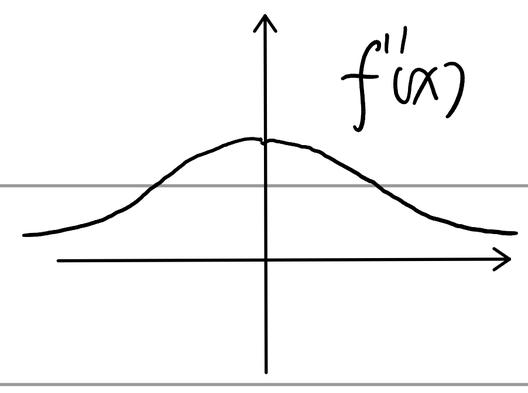
Example. $f(x) = x \sinh^{-1}(x) - \sqrt{1+x^2} = x \ln(x + \sqrt{x^2+1}) - \sqrt{x^2+1}$. $|x| \leq 10$.

$f'(x) = \sinh^{-1}(x)$

$f''(x) = 1/\sqrt{x^2+1}$.







$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)} \qquad |x_{k+1}| > |x_k| \text{ if } |x_k| \geq 4.$$

Taylor series approximate well in the neighborhood of x_k. but lose control.

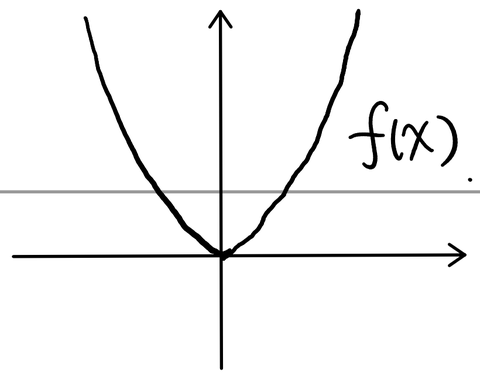
if $|x_k - x^*|$ is large. (L-smooth approximation is a global bound).

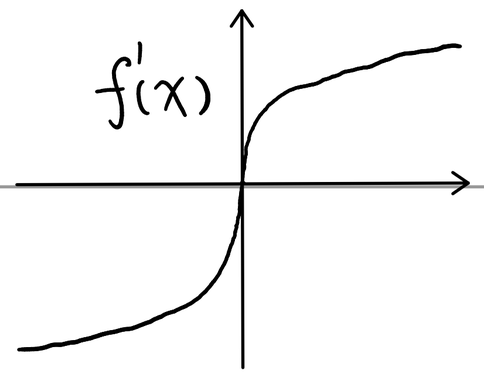
However, $|x_k - x^*|$ is not sufficient yet. Consider a strong version of $|x|$.

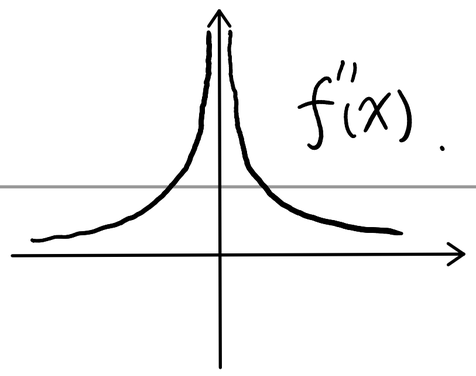
Example. $f(x) = x^{\frac{4}{3}}$. $f'(x) = \frac{4}{3} x^{\frac{1}{3}}$. $f''(x) = \frac{4}{9} x^{-\frac{2}{3}}$. $|x| \leq 1$.

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$$

$$= -2 x_k.$$







Roughly. we hope $(\nabla^2 f(x))^{-1} \nabla f(x)$ is 1-Lipschitz. if $f: \mathbb{R} \to \mathbb{R}$. hope

$\left(\frac{f'(x)}{f''(x)}\right)'$ is bounded $\iff |f''(x)/f''(x) - f'(x) f'''(x)/f''(x)^2|$ bounded.

If $f: \mathbb{R}^n \to \mathbb{R}$. we do not consider $f'''(x)$. so hope $\nabla^2 f(x)$ is Lipschitz.

Norm of matrices. e.g. Frobenius norm. $\|Q\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} Q_{ij}\right)^{1/2}$.

Operator norm: $Q \in \mathbb{R}^{m \times n}$ is a linear transform from $\mathbb{R}^n \to \mathbb{R}^m$ $x \mapsto Qx$.

Given $\|\cdot\|_a$ and $\|\cdot\|_b$ for $\mathbb{R}^n$ and $\mathbb{R}^m$. define operator norm of Q

$$\|Q\|_{a,b} = \max_{x \neq 0} \frac{\|Qx\|_b}{\|x\|_a} = \max_{\|x\|_a = 1} \|Qx\|_b = \max_{\|x\|_a \leq 1} \|Qx\|_b$$

Proposition. $\forall x \in \mathbb{R}^n$. $\|Qx\|_b \leq \|Q\|_{a,b} \|x\|_a$.

In particular. if $a = b = 2$. $\|\cdot\|_2 = \|\cdot\|_{a,b}$ is called the spectral norm.

Proposition. $\|Q\|_2 = (\lambda_{max}(Q^TQ))^{1/2}$ since $\|Qx\|^2 = x^TQ^TQx \leq \lambda_{max}\|x\|^2$

Definition: $\nabla^2 f(x)$ is M-Lipschitz if $\|\nabla^2 f(x) - \nabla^2 f(y)\|_2 \leq M\|x-y\|_2$.