Lecture 14. Exact and backtracking line search; Newton's method.

Proposition. successive gradient directions are always orthogonal. since

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

Theorem. If f is m-strongly convex and L-smooth. $\{x_n\}$ given by the exact line search. then $f(x_k) - f(x^*) \le (1-\frac{m}{L})^k (f(x_0) - f(x^*))$.

Proof. $g(s) = f(x_k - s\nabla f(x_k)) \le \underbrace{f(x_k) - s\|\nabla f(x_k)\|^2 + \frac{Ls^2}{2}\|\nabla f(x_k)\|^2}_{\triangleq h(s)}$.

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

By m-strongly convexity $f(x^*) \ge \underbrace{f(x_k) + \nabla f(x_k)^T(x^* - x_k) + \frac{m}{2}\|x^* - x_k\|^2}_{\triangleq \hat{f}(x^*)}$.

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

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

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

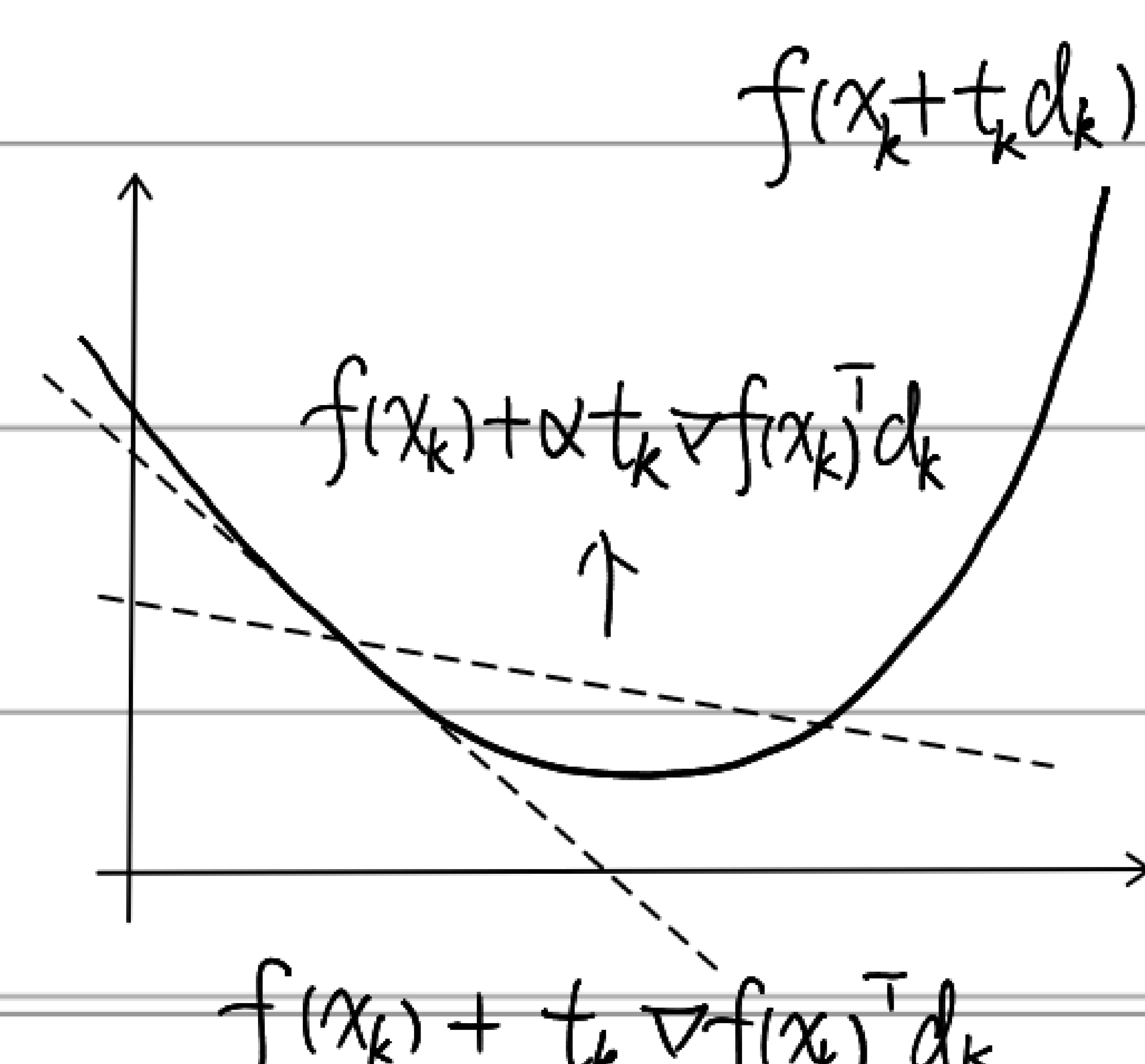
Remark: exact line search is usually expensive.

Backtracking line search: Armijo's rule.

given a descent direction d_k and $\alpha, \beta < 1$

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

$t_k \leftarrow \beta t_k$; $\quad x_{k+1} \leftarrow x_k + t_k d_k$



in particular. let $d_k = -\nabla f(x)$. for gradient descent method.

while $f(x_k - t_k \nabla f(x_k)) > f(x) - \alpha t_k \|\nabla f(x_k)\|^2$ $\quad t_k = \beta t_k$.

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

Convergence analysis for backtracking line search. assume $t_k = 1$ initially.

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

$$\le f(x_k) - \frac{t}{2}\|\nabla f(x_k)\|^2 \qquad (\text{assume } L\text{-smoothness and } \forall t \le 1/L).$$

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

in fact. for general α. we need $t \le 2(1-\alpha)/L$

so the backtracking line search terminates with $t = t_0 = 1$. or. $t \ge \beta/L$.

$$\Rightarrow f(x_{k+1}) \le f(x_k) - \alpha\|\nabla f(x_k)\|^2 \text{ or } f(x_{k+1}) \le f(x_k) - \frac{\alpha\beta}{L}\|\nabla f(x_k)\|^2.$$

$$\Rightarrow f(x_{k+1}) \le f(x_k) - \alpha \min\{1, \beta/L\}\|\nabla f(x_k)\|^2.$$

$$\Rightarrow f(x_{k+1}) - f(x^*) \le f(x_k) - f(x^*) - \alpha\min\{1, \beta/L\}\underbrace{\|\nabla f(x_k)\|^2}_{\ge 2m(f(x_k) - f(x^*))}$$

$$\le (1 - 2m\alpha\min\{1, \beta/L\})(f(x_k) - f(x^*)).$$

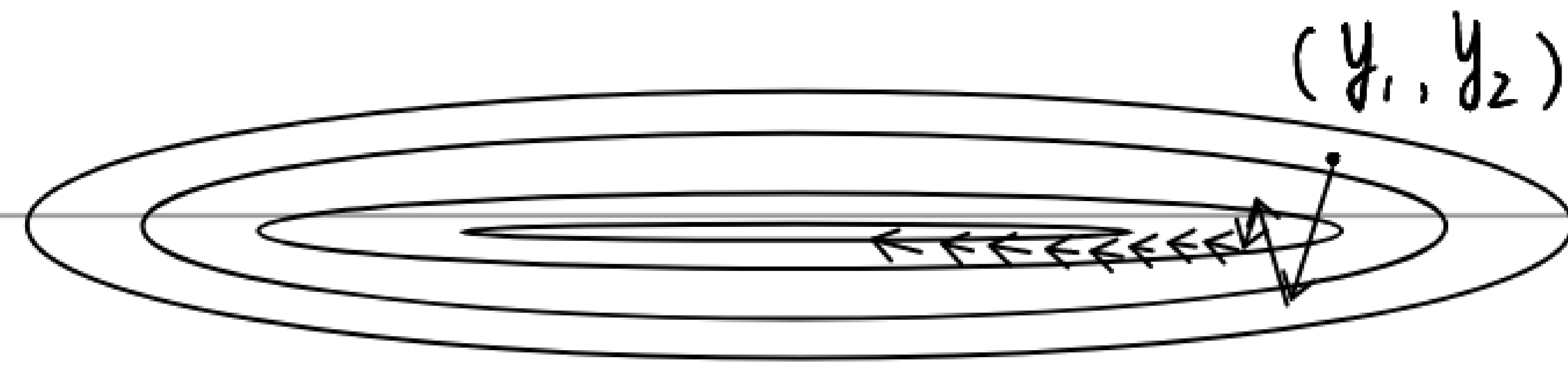
Remark. $2m\alpha\beta/L \le \beta m/L < 1$. and > 0 if $m > 0$. (worse than exact).

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. $f(x_k) - f(x^*) \le \left(1 - \min\{2m\alpha t_0, 4m\alpha(1-\alpha)\beta/L\}\right)^k (f(x_0) - f(x^*))$.

Better descent direction: Newton's method.

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



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

locally descend rapidly but not globally.

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

In general if $f(x) = x^T Q x$. $-\nabla f(x) = 2Qx$. hope $d_x = -\underbrace{\frac{1}{2} Q^{-1}}_{= (\nabla^2 f(x))^{-1}} \nabla f(x)$

Recall Newton's method for finding roots. $x \leftarrow x - \frac{g(x)}{g'(x)}$

By Taylor expansion. $f(x) \approx f(x_k) + \nabla f(x_k)^T (x - x_k) + \frac{1}{2}(x - x_k)^T \nabla^2 f(x_k)(x - x_k)$.

$\nabla f(x) = 0 \approx \nabla f(x_k) + \nabla^2 f(x_k)(x - x_k) \Rightarrow x \approx x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$.

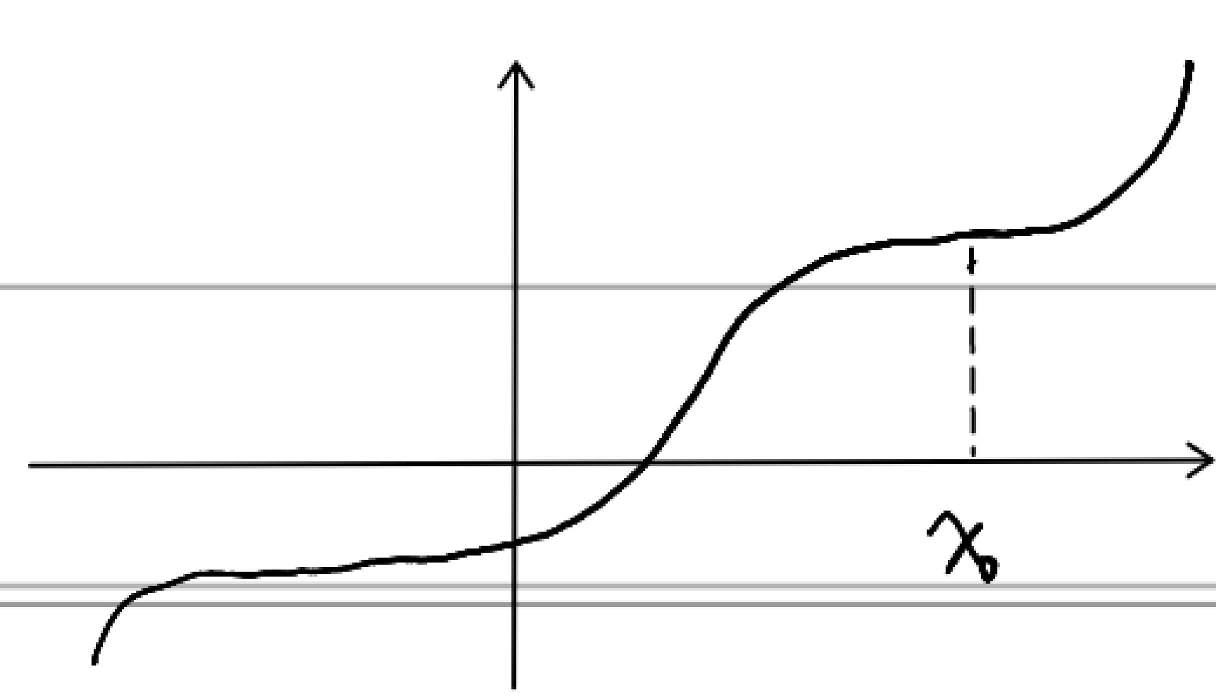
Newton's method: $x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$ provided $\nabla^2 f(x_k) \succ 0$.

Remark: if $f(x)$ is quadratic. Newton's method terminates in one step.

If $\nabla^2 f(x_k) \succ 0$. $-(\nabla^2 f(x_k))^{-1} \nabla f(x_k)$ is a descent direction. since

$(\nabla^2 f(x_k))^{-1} \succ 0$ thus. $-\nabla f(x_k)^T (\nabla^2 f(x_k))^{-1} \nabla f(x_k) < 0$ if $\nabla f(x_k) \neq 0$.

Question: convergence analysis of Newton's method.



highly depends on the initial point

converge rapidly if starting from good point

Consider function $f: \mathbb{R} \to \mathbb{R}$. $x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$.

$|x_{k+1} - x^*| = |x_k - x^* - (f''(x_k))^{-1} f'(x_k)|$ (note that $f'(x^*) = 0$).

$= |f''(x_k)|^{-1} \underbrace{|f''(x_k)(x_k - x^*) - (f'(x_k) - f'(x^*))|}$

$\approx f'''(\xi)(x_k - x^*)^2$ by Taylor expansion

$= |f''(x_k)|^{-1} \left| \int_{x^*}^{x_k} f''(x_k) - f''(y)\, dy \right|$

$\leq |f''(x_k)|^{-1} |x_k - x^*| \int_0^1 |f''(x_k) - f''(x^* + t(x_k - x^*))|\, dt.$

if m-strongly convex $\leq |x_k - x^*| / m \cdot \int_0^1 M(1-t)|x_k - x^*|\, dt$

$f''(x)$ is M-Lipschitz $= M/(2m)\, |x_k - x^*|^2$. (quadratic convergence).

If f is third-order continuous differentiable. and x_k close sufficiently to x^*. then $\exists M > 0$. $f''(x)$ near x^* is M-Lipschitz.

M/m cannot be too large. $f(x) = x^{\frac{4}{3}}$. $f'(x) = \frac{4}{3} x^{\frac{1}{3}}$. $f''(x) = \frac{4}{9} x^{-\frac{2}{3}}$.

$f'''(x) = -\frac{8}{27} x^{-\frac{5}{3}}$ $\quad x \in [-a, a]$ $\quad \underset{\to\infty}{M \gg \frac{8}{27} a^{-\frac{5}{3}}}$. $m = \frac{4}{9} a^{-\frac{2}{3}}$.

$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)} = -2x_k$. M/m too large if $a < 1$.

$|x_k - x^*|$ cannot be too large. $f'(x) = \operatorname{arcsinh}(x) = \ln(x + \sqrt{x^2+1})$. $|x| \leq 10$.

$f''(x) = 1/\sqrt{1+x^2}$ if $|x| \leq 10$. o.w $f''(x) = f''(10)$. $\quad f(x) = x \operatorname{arcsinh} x - \sqrt{1+x^2}$

$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)} < -x_k$ if $x_k > 5$ and $> -x_k$ if $x_k < -5$.

$\underset{f''(10)}{m = 101^{-1/2}}$ (strongly convex). $M = \max |f'''(x)| = \frac{2}{9}\sqrt{3}$.

Global convergence: exact / backtracking line search

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: for $Q \in \mathbb{R}^{m\times n}$ $\quad Q: \mathbb{R}^n \to \mathbb{R}^m \quad x \in \mathbb{R}^n \to Qx$

given $\|\cdot\|_a$ and $\|\cdot\|_b$ on $\mathbb{R}^n$ and $\mathbb{R}^m$. operator norm of Q is

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

Propositions. equivalence of definitions. furthermore. $\|Qx\|_b \leq \|Q\|_{a,b}\|x\|_a$

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

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

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

Theorem. If f is m-strongly convex. $\nabla^2 f$ is M-Lipschitz. $\{x_k\}$ produced by Newton's method. then $\|x_{k+1} - x^*\| \leq \frac{M}{2m}\|x_k - x^*\|^2$.

Remark: let $y_k = \frac{M}{2m}\|x_k - x^*\|$. then $y_{k+1} \leq y_k^2 \Rightarrow y_k \leq y_0^{2^k}$.

Proof. $\|x_{k+1} - x^*\|_2 \leq \|(\nabla^2 f(x_k))^{-1}\|_2 \, \|\nabla^2 f(x_k)(x_k - x^*) - (\nabla f(x_k) - \nabla f(x^*))\|_2$.

$$= \|(\nabla^2 f(x_k))^{-1}\|_2 \left\| \int_0^1 \left(\nabla^2 f(x_k) - \nabla^2 f(x^* + t(x_k - x^*))\right)(x_k - x^*)\, dt \right\|_2.$$

$$\leq \frac{1}{m}\|x_k - x^*\|_2 \cdot \int_0^1 \|\nabla^2 f(x_k) - \nabla^2 f(x^* + t(x_k - x^*))\|_2\, dt.$$

$$\leq \frac{M}{m}\|x_k - x^*\|_2 \cdot \int_0^1 (1-t)\|x_k - x^*\|_2\, dt = \frac{M}{2m}\|x_k - x^*\|_2^2. \quad \square$$