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

Proposition. successive gradient directions are always orthogonal. since.

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

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

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

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

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

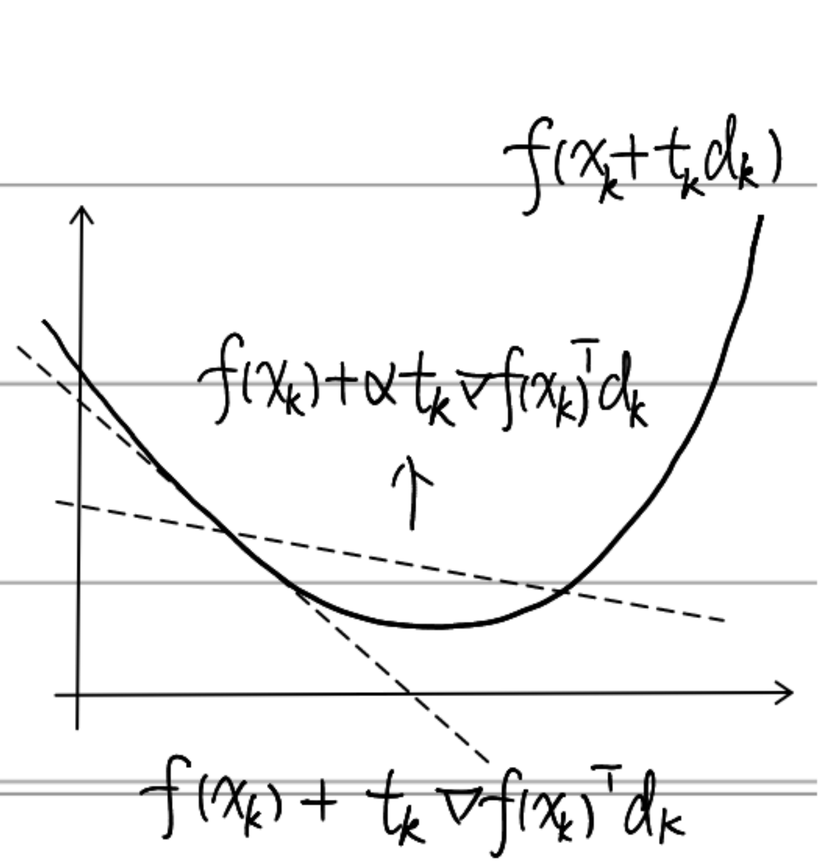
Remark: exact line search is usually expensive.

Backtracking line search: Armijo's rule.

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

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

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



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.

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

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

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

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 $t \ge \beta/L$.

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

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

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

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

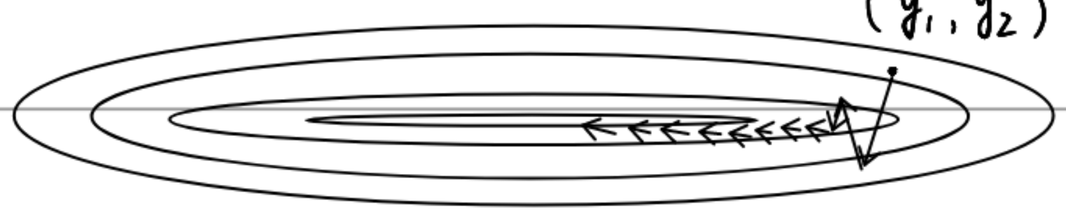
Remark. $2m\alpha\beta/L \le \beta m/L < 1$. and > 0 if $m > 0$. (worse than exact).

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

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 $d_k = -\frac{1}{2} Q^{-1} \nabla f(x)$

Recall Newton's method for finding roots. $x \leftarrow x - \frac{g(x)}{g'(x)}$ $= (\nabla^2 f(x))^{-1}$

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

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

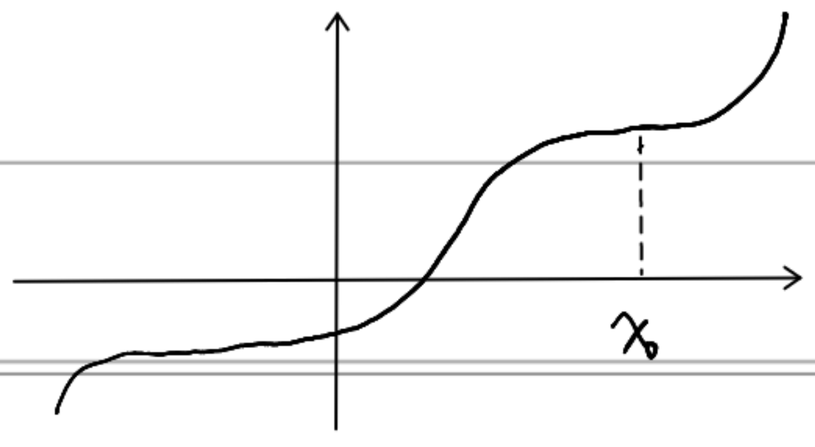
Newton's method: $x_{k+1} = x_k - (\nabla^2 f(x_k))^{-1} \nabla f(x_k)$ provided $\nabla^2 f(x_k) > 0$.

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

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

$(\nabla^2 f(x_k))^{-1} > 0$ 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