Lecture 13. Strongly convex. Condition number and exact line search.

Gradient descent with fixed step size t: $\sum_{k=1}^{T} f(x_k) - f(x^*) \le \frac{\|x_0 - x^*\|^2}{2t}$

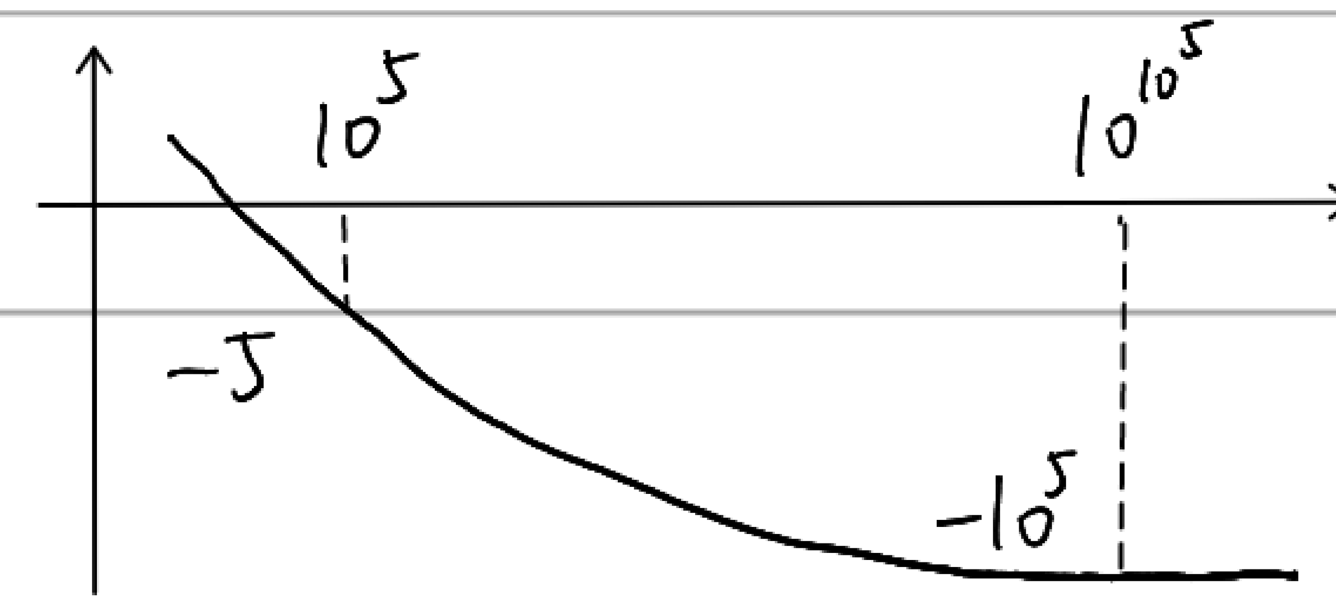
If $f(x_{k+1}) < f(x_k)$. $\Rightarrow f(x_T) - f(x^*) \le \frac{1}{2tT} \|x_0 - x^*\|^2$.

o.w. $f(\frac{1}{T}\sum x_k) \le \frac{1}{T}\sum f(x_k)$ by convexity. f is L-smooth.

$\Rightarrow f(\frac{1}{T}\sum x_k) - f(x^*) \le \frac{1}{2tT}\|x_0 - x^*\|^2$. $t < 1/L$.

Remark: rate of convergence is $O(1/T)$; $T = O(1/\varepsilon)$ to get ε-approximation.

Consider the following function:



$f(x) = -\log x$. $x < 10^{10^5}$. and -10^5 o.w.

$f'(x) = 1/x$. $x_{k+1} \leftarrow x_k + t/x_k$. stop at $|f'(x)| < 10^{-5}$. Bad!

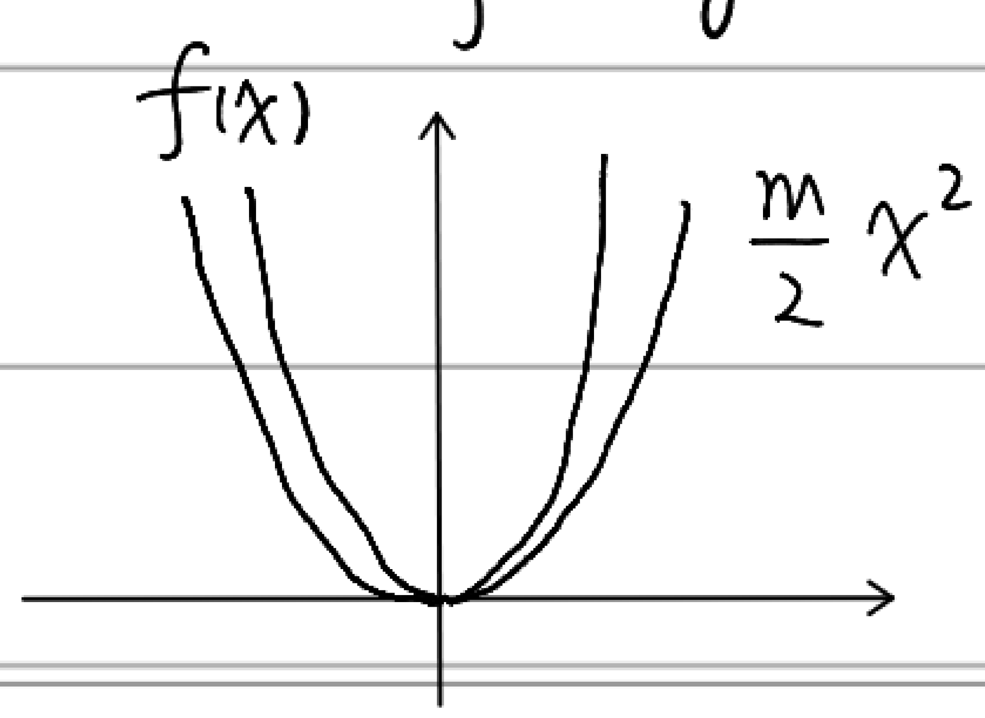
Another bad example: $f(x) = x^4$. if $|x| \le 1$. $4|x| - 3$. o.w.

$x_{k+1} \leftarrow x_k - 4x_k^3 \cdot t = x_k(1 - 4t x_k^2)$ $x_k \sim (8+k)^{-1/2}$. $f(x_k) \sim (8+k)^{-2}$

Good example: $f(x) = 6x^2$ $x_{k+1} = x_k(1-12t)$. $f(x_k) = 6x_0^2(1-12t)^{2k}$.

Strongly convex: f is strongly convex with $m > 0$. or m-strongly convex:

if $g(x) = f(x) - \frac{m}{2}\|x\|^2$ is convex.



should be strongly convex everywhere.

$f(x) - \frac{m}{2}\|x - y\|^2 = g(x) - \frac{m}{2}(\|y\|^2 - 2x^T y)$.

First-order condition: a differentiable f is m-strongly convex

iff $f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{m}{2} \|y-x\|^2$. $\forall x, y$.

Proof: $g(x) = f(x) - \frac{m}{2}\|x\|^2$ convex $\Leftrightarrow g(y) \geq g(x) + \nabla g(x)^T (y-x)$

$\Leftrightarrow f(y) - \frac{m}{2}\|y\|^2 \geq f(x) - \frac{m}{2}\|x\|^2 + (\nabla f(x) - mx)^T (y-x)$.

$\Leftrightarrow f(y) \geq f(x) + \nabla f(x)^T (y-x) + \frac{m}{2}(\|x\|^2 + \|y\|^2 - 2x^T y)$ $\square$

Remark: L-smoothness $\Rightarrow f(y) \leq f(x) + \nabla f(x)^T (y-x) + \frac{L}{2}\|y-x\|^2$

Second-order condition: a twice continuous differentiable f is

m-strongly convex iff $\nabla^2 f(x) \succeq mI$. i.e. $\lambda_{min}(\nabla^2 f(x)) \geq m$.

Example. x^4, $-\log x$ are not strongly convex. $f'' = 12x^2 - m$.

Example. $f(x) = w^T x$ is not strongly convex. $f'' = -m$.

$f(x) = x^T Q x$ is $2\lambda_{min}(Q)$-strongly convex if $Q \succ 0$.

Consider $f(x) = ax^2$. $a > 0$. $x_{k+1} = x_k(1-2t) = (1-2t)^{k+1} x_0$.

Theorem: If f is m-strongly convex and L-smooth. fix $t < 1/L$ and

assume $x^* = \arg\min f$. and $\{x_k\}$ given by the gradient descent method.

then. $\|x_k - x^*\|^2 \leq (1-mt)^k \|x_0 - x^*\|^2$

Remark: $\Rightarrow f(x_k) - f(x^*) \leq \frac{L}{2}(1-mt)^k \|x_0 - x^*\|^2$.

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

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

$\Rightarrow \|x_{k+1}-x^*\|^2 \leq (1-mt)\|x_k - x^*\|^2 + t^2\|\nabla f(x_k)\|^2 + 2t(f(x^*) - f(x_k))$.

(recall $f(x_{k+1}) \leq f(x_k) - \frac{t}{2}\|\nabla f(x_k)\|^2$ by L-smoothness)

$\Rightarrow \|x_{k+1}-x^*\|^2 \leq (1-mt)\|x_k - x^*\|^2 + 2t(f(x^*) - f(x_{k+1}))$. $\square$

Key difficulty: how to select step size t?

Convergence rate for quadratic function $f(x) = x^TQx$. $Q \succ 0$

$f(x)$ is $2\lambda_{max}$-smooth and $2\lambda_{min}$-strongly convex. select $t < 1/\lambda_{max}$.

let $Q = U^T \Lambda U$ where $\Lambda = \text{diag}\{\lambda_{min}, \cdots, \lambda_{max}\}$. $\nabla f(x) = 2Qx$.

$\Rightarrow x_{k+1} = (I - 2tQ)x = U^T\Lambda' U x_k$ where $\Lambda' = I - 2t\,\text{diag}\{\lambda_{min}, \cdots, \lambda_{max}\}$.

$\Rightarrow x_k = (U^T\Lambda'U)^k x_0 = U^T(\Lambda')^k U x_0$. let $y_k = Ux_k$. $y^* = Ux^* = 0$.

$y_k = (\Lambda')^k y_0 = (\text{diag}\{1-2t\lambda_{min}, \cdots, 1-2t\lambda_{max}\})^k y_0$

$= \text{diag}\{(1-2t\lambda_{min})^k, \cdots, (1-2t\lambda_{max})^k\} y_0$.

$\Rightarrow \|x_k\|^2 = \|y_k\|^2 = \sum_{i=1}^{n}(1-2t\lambda_i)^{2k} y_{0i}^2 \quad \lambda_{min} \leq \lambda_i \leq \lambda_{max}$

convergence rate $= \max\{(1-2t\lambda_i)^{2k}\} = \max\{(1-2t\lambda_{min})^{2k}, (1-2t\lambda_{max})^{2k}\}$.

select t to $\min \max\{|1-mt|, |1-Lt|\} \Rightarrow t = 2/(m+L)$.

$$\Rightarrow \max\ (1-2t\lambda_i)^{2k} \le \left(\frac{L-m}{L+m}\right)^{2k} \Rightarrow \|x_k\|^2 \le \left(\frac{L-m}{L+m}\right)^{2k} \|x_0\|^2.$$

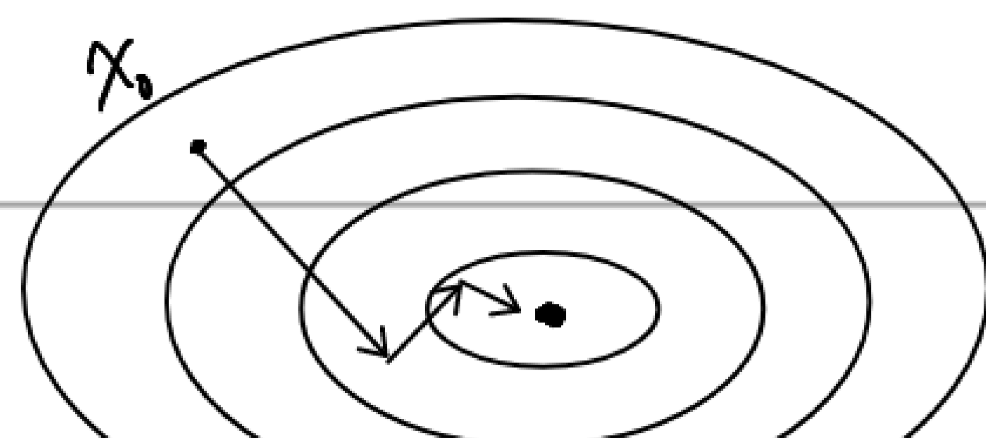
Condition number of Q: $K(Q) = \frac{\lambda_{max}(Q)}{\lambda_{min}(Q)} = \frac{L}{m} \geq 1$. for $Q > 0$.

Convergence rate of fixed step size gradient descent method:

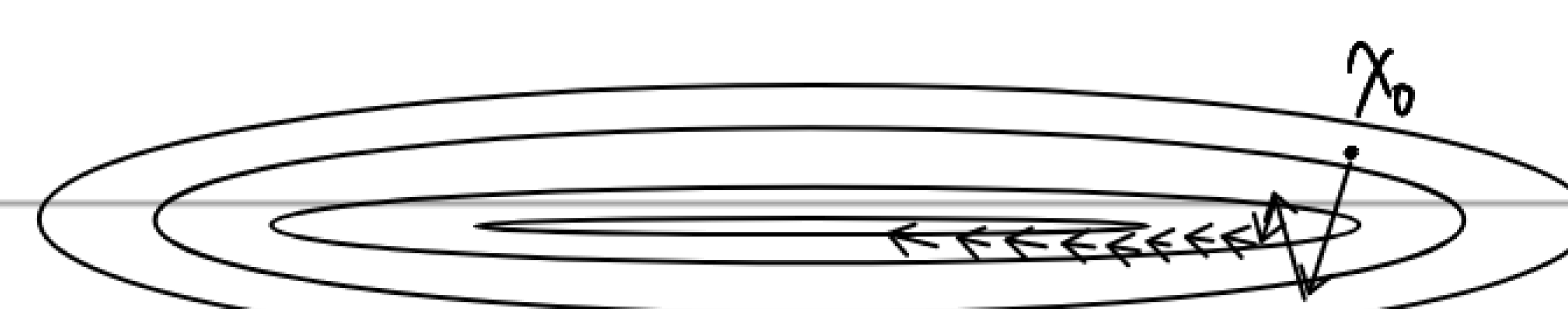
- for quadratic functions, rate depends on $\left(\frac{K-1}{K+1}\right)^2$.

- for nonquadratic functions. locally approximated by (Taylor's expansion)

  $\frac{1}{2} x^T \nabla^2 f(x^*) x$ + linear and constant terms. depends on $K(\nabla^2 f(x^*))$.



$Q = \text{diag}\{1/2, 1\}$
small $K = 2$.
well-conditioned



$Q = \text{diag}\{0.01, 1\}$
$x_0 = (a_1, a_2)$ ideal direction $-x_0$
actual direction $-(0.01 a_1, a_2)$
large $K = 100$.
ill-conditioned

Exact line search: $x_{k+1} \leftarrow x_k - t\nabla f(x_k)$, $t \leftarrow \arg\min_s f(x_k - s\nabla f(x_k))$.

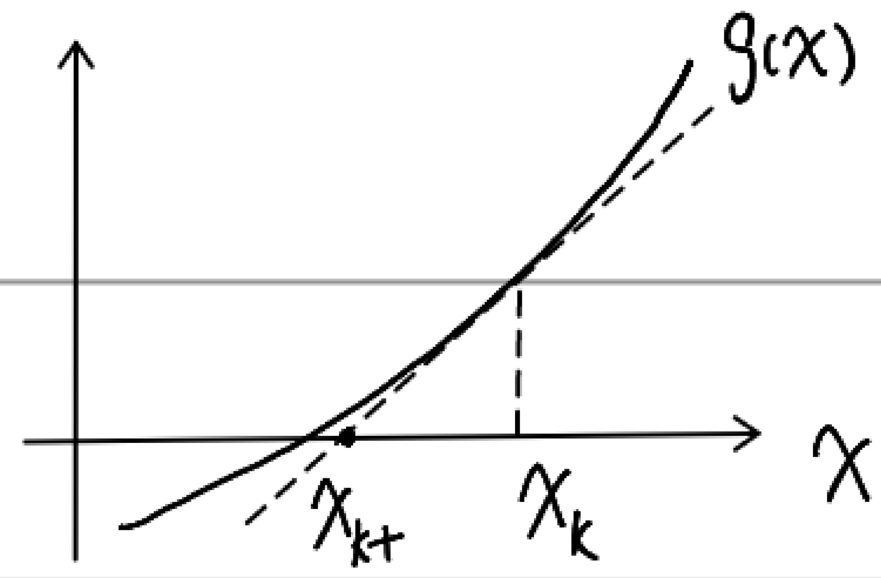
a convex function restricted on every line is also convex.

Example: $f(x) = x^T Q x + w^T x$. $Q > 0$. $d_k \triangleq \nabla f(x_k) = 2Qx_k + w$.

$$t = \arg\min_s g(s) = f(x_k - sd_k) = f(x_k) - 2sd_k^T Q x + s^2 d_k^T Q d_k - s w^T d_k.$$

$$= \arg\min_s f(x) - s d_k^T d_k + s^2 d_k^T Q d_k \quad = \frac{d_k^T d_k}{d_k^T Q d_k}.$$

In general. find the root of derivative: bisection. or Newton's method.



$g(x) \approx g(x_k) + g'(x_k)(x - x_k)$.

$x_{k+1} \leftarrow x_k - \frac{g(x_k)}{g'(x_k)}$

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

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

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

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

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

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

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