Lecture 13. Condition number; Line search

Theorem. If f is m-strongly convex and L-smooth. Fix $t \le 1/L$ and let $x^* = \arg\min f$. Suppose $\{x_k\}$ is given by the gradient descent method. Then $\|x_T - x^*\|^2 \le (1-mt)^T \|x_0 - x^*\|^2$.

Remark. m-strong convexity $\Rightarrow$ strict convexity. so x^* is unique.

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

Remark It further gives $f(x_T) - f^* \le \frac{L}{2}(1-mt)^T \|x_0 - x^*\|^2$ since

$$f(x_T) \le f(x^*) + \nabla f(x^*)^T(x_T - x^*) + \frac{L}{2}\|x_T - x^*\|^2 \text{ by } L\text{-smothness.}$$

Proof. Recall that $f(x_{k+1}) \le f(x_k) - \frac{t}{2}\|\nabla f(x_k)\|^2$. $(*)$ and

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

We use $\nabla f(x_k)^T(x_k - x^*) \ge f(x_k) - f(x^*)$ before by convexity.

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

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

$$(\text{plugging in } (*)) \quad \le (1-mt)\|x_k - x^*\|^2 - 2t(f(x_{k+1}) - f^*)$$

$$\le (1-mt)\|x_k - x^*\|^2 \qquad \square$$

Remark. Clearly $m \le L$. We choose t as large as possible $= 1/L$.

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

$f(x)$ is $2\lambda_{max}$ - smooth and $2\lambda_{min}$ - strongly convex.

Recall that $f(x_{k+1}) < f(x_k)$ if $t < 1/\lambda_{max}$. above result not apply.

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

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

$\Rightarrow x_k = (U \Lambda' U^T)^k x_0 = U \Lambda'^k U^T x_0$.

let $y = U^T x$. Clearly $x^* = 0$. So $y^* = U^T x^* = 0$.

$y_k = U^T x_k = \Lambda'^k U^T x_0 = \Lambda'^k y_0 = (I - 2t \,\text{diag}\{\lambda_{min} \dots \lambda_{max}\})^k y_0$

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

$\Rightarrow \|x_k - x^*\|^2 = \|x_k\|^2 = \|y_k\|^2 = \sum_{i=1}^{n} (1-2t\lambda_i)^{2k} (y_0)_i^2$.

$\leq \max\{(1-2t\lambda_i)^{2k}\} \|y_0\|^2 = \|x_0\|^2$.

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

select t to minimize $\max\{|1-mt|, |1-Lt|\} \Rightarrow t = \frac{2}{m+L}$.

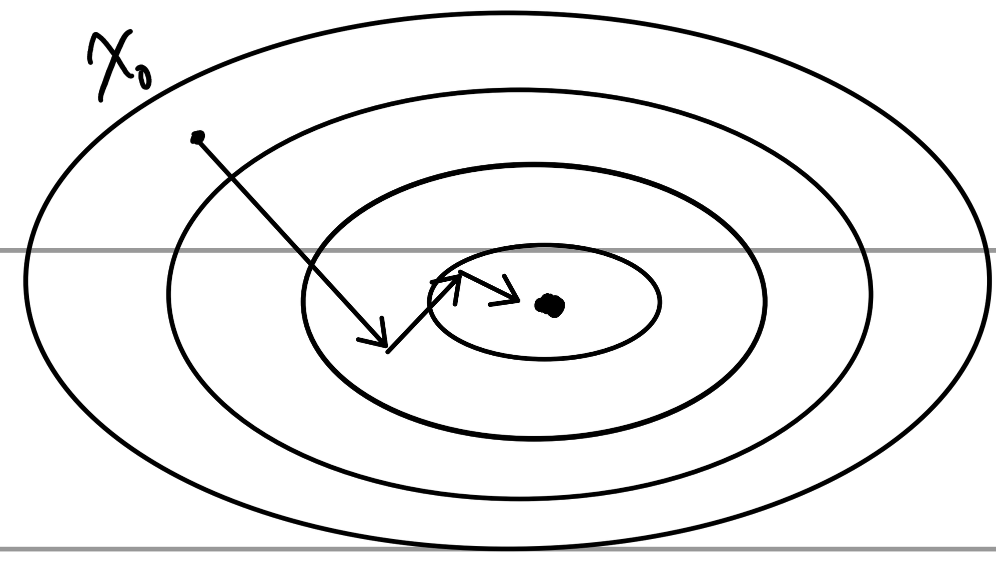
$\Rightarrow \max\{(1-2t\lambda_i)^{2k}\} \leq \left(\frac{L-m}{L+m}\right)^{2k}$. $\|x_k\|^2 \leq \left(\frac{L-m}{L+m}\right)^{2k} \|x_0\|^2$.

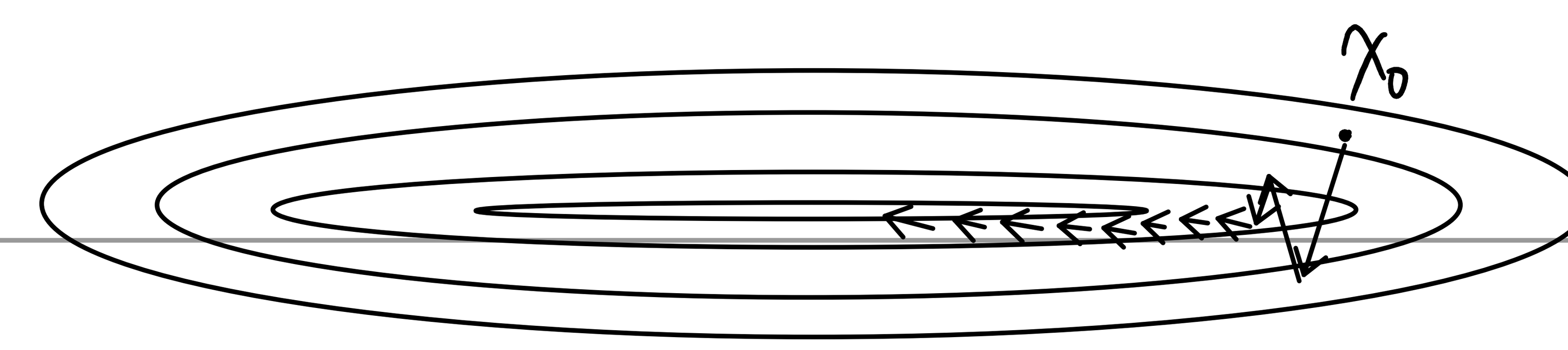
Definition (condition number). $\kappa(Q) = \frac{\lambda_{max}(Q)}{\lambda_{min}(Q)} = \frac{L}{m} \geq 1$ for $Q \succ 0$

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 series.

$\frac{1}{2}(x-x^*)^T \nabla^2 f(x^*)(x-x^*)$ + affine terms near x^*. $K(\nabla^2 f)$.





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

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

Floating step size: line search method.

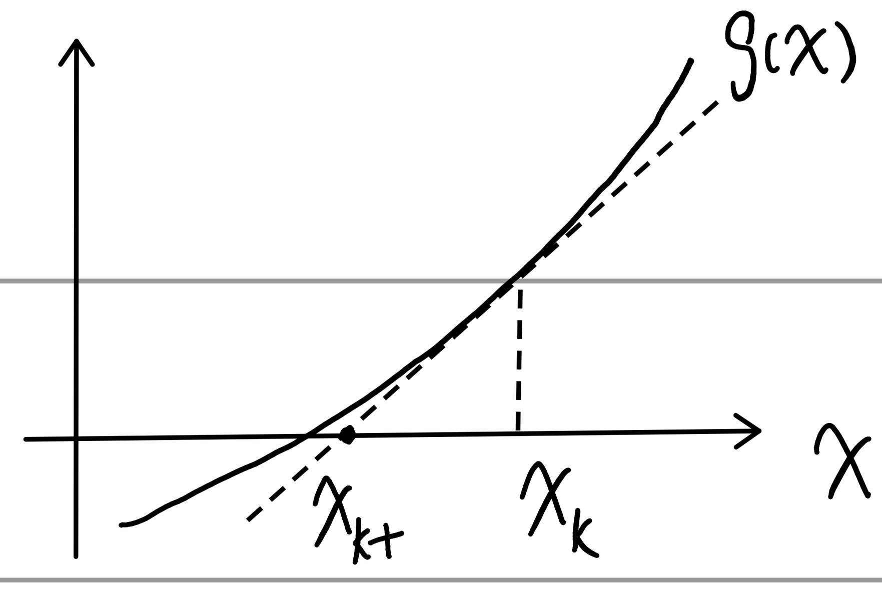
Exact line search: $x_{k+1} = x_k - t\nabla f(x_k)$. $t = \arg\min_s f(x_k - s\nabla f(x_k))$.

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

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

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

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



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

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

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