Lecture 12. Gradient descent method

Recall that x^* is optimal for $\min f(x) \iff \nabla f(x^*) = 0$.

Suppose $f(x) = ax^2 + bx + c$. $-f'(x)$ is the direction to optimal.

Intuitively. move from x to $x - t \cdot f'(x)$. when to stop?

Ideally. stop when $f'(x) = 0$. but inpractical.

$|f'(x)| < \delta$. $|f(x_{new}) - f(x_{old})| < \delta$. # of iterations $< T$. ...

Now consider $f: \mathbb{R}^n \to \mathbb{R}$. guess initial x_0. then how to move?

Naïve idea: $x_{k+1} \leftarrow x_k - t \cdot d_k$. s.t. $f(x_{k+1}) < f(x_k)$.

For convex differential f. $-d_k$ is a descent direction $\iff d_k^T \nabla f(x_k) > 0$.

Simply select $d_k = \nabla f(x_k)$. Advantage: max rate descending direction.

Assume $\|d_k\| = 1$. $\nabla_{d_k} f(x_k) = \lim_{t \downarrow 0} (f(x_k + t \cdot d_k) - f(x_k))/t = d_k^T \nabla f(x_k)$

By Cauchy-Schwarz. $d_k^T \cdot \nabla f(x_k) = \langle d_k, \nabla f(x_k) \rangle \leq \|d_k\| \cdot \|\nabla f(x_k)\|$

with equality iff $d_k = (\pm) \nabla f(x_k) / \|\nabla f(x_k)\|$. $\geq -\|d_k\| \cdot \|\nabla f(x_k)\|$.

Gradient descent: $x_{k+1} \leftarrow x_k - t \cdot \nabla f(x_k)$.

reasonable requirement: $f(x_{k+1}) < f(x_k)$. consider $f(x) = ax^2$. $a > 0$.

$\Rightarrow$

$f'(x) = 2ax \Rightarrow x_{k+1} = x_k - 2at\, x_k$. $(1-2at)^2 x_k^2 < x_k^2 \Rightarrow t < \frac{1}{a}$.

consider $f(x) = x^T Q x$ (symmetric $Q \in \mathbb{R}^{n\times n}$). $\nabla f(x) = 2Qx \quad x_{k+1} = (I - 2tQ)x_k$.

$f(x_{k+1}) = x_k^T Q x_k + 4t^2 (Qx_k)^T Q (Qx_k) - 4t (Qx_k)^T (Qx_k)$.

$f(x_{k+1}) < f(x_k) \Leftarrow t \cdot (Qx_k)^T Q (Qx_k) < (Qx_k)^T (Qx_k)$

Proposition. $\lambda_{min} \|x\|_2^2 \le x^T Q x \le \lambda_{max} \|x\|_2^2$. (so $t < 1/\lambda_{max}$ suffice)

Proof: $Q \in \mathbb{R}^{n\times n}$ symmetric $\Rightarrow$ orthogonally diagonalize $Q = U \Lambda U^T$

where $\Lambda = \mathrm{diag}\{\lambda_1, \dots \lambda_n\}$ and $U^T U = I$. let $x = Uy$.

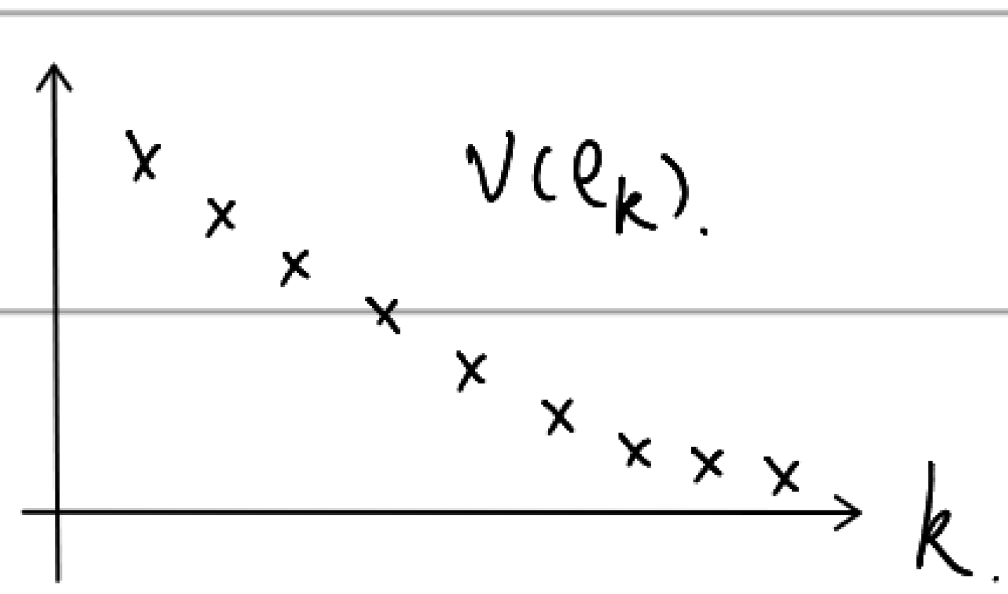
$x^T Q x = y^T U^T Q U y = y^T \Lambda y = \sum_{i=1}^{n} \lambda_i y_i^2 \le \max |\lambda_i| \cdot \|y\|_2^2$.

$\|y\|_2^2 = y^T y = y^T U^T U y = x^T x = \|x\|_2^2$. Similarly for λ_{min}. $\square$

Convergence: let $e_k = x_k - x^*$. and $V(e) = f(e + x^*) - f(x^*)$

$V(e) = 2e^T Q x^* + e^T Q e > 0$ (x^* optimal, $e \neq 0$). $V(0) = 0$.

$V(e_{k+1}) = f(x_{k+1}) - f(x^*) < f(x_k) - f(x^*) = V(e_k)$.



$V(e_k) \downarrow$ and lower-bounded $\Rightarrow \lim_{k\to\infty} V(e_k) = c$ exist

if $c = 0$ done. assume $c > 0$.

let $S = \{x : c \le V(x) \le V(e_0)\}$. S is compact.

note that $e_{k+1} = e_k + x^* - 2tQ(e_k + x^*) - x^* \triangleq \underbrace{g(e_k)}_{\text{continuous function.}}$

let $\delta = \min_{e \in S} |V(g(e)) - V(e)| > 0$ due to continuity and compactness

$0 \notin S \Rightarrow V(g(e)) < V(e).\ \forall e \in S \Rightarrow \delta > 0 \Rightarrow V(e_{k+1}) < V(e_k) - \delta.$ $\square$

Lyapunov's global stability theorem in discrete time:

$e_{k+1} = g(e_k)$ where $g: \mathbb{R}^n \to \mathbb{R}^n$ continuous. and $g(0) = 0$.

If there exists $V: \mathbb{R}^n \to \mathbb{R}$ continuous s.t. (Lyapunov function).

① $V(0) = 0$ $V(e) > 0$ $\forall e \neq 0$. (positivity).

② $V(e) \to \infty$ as $\|e\| \to \infty$ (radical unboundedness).

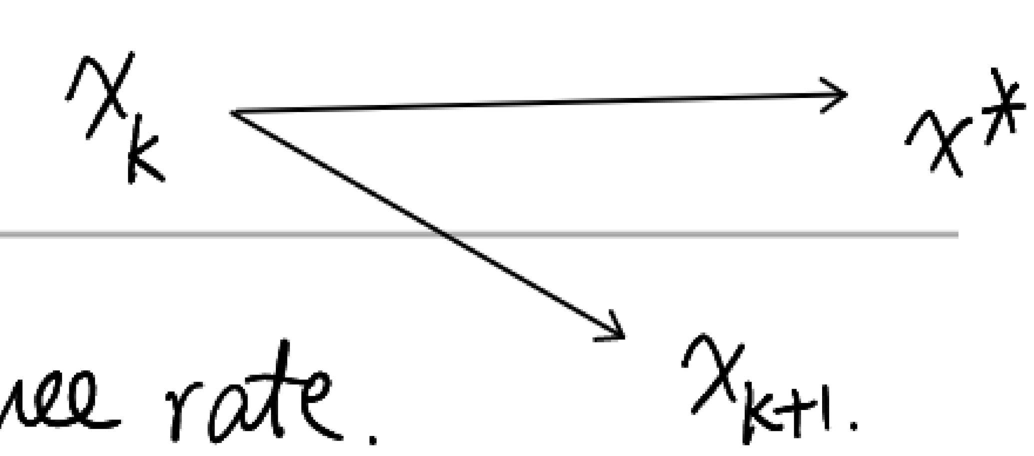
③ $V(g(e)) < V(e)$. $\forall e \neq 0$. (strict decrease).

Then $\forall e_0 \in \mathbb{R}^n$. we have $e_k \to 0$ as $k \to \infty$.

For gradient descent method. $e_k = x_k - x^*$. $g(e) = e - t \cdot \nabla f(e + x^*)$.

$V(e) = f(e + x^*) - f(x^*)$. $\exists$ unique optimal $x^* \Rightarrow$ convergence.

If we are in math dept. then we are done!



But in CS dept. we'd like to ask the convergence rate.

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

$\nabla f(x_k)^T(x_k - x^*) \geq f(x_k) - f(x^*)$. by convexity.

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

$\Rightarrow \|x_T - x^*\|^2 - \|x_0 - x^*\|^2 \leq t^2 \sum_{k=0}^{T-1} \|\nabla f(x_k)\|^2 - 2t \sum_{k=0}^{T-1} (f(x_k) - f(x^*))$.