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.

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

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

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

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

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

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

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

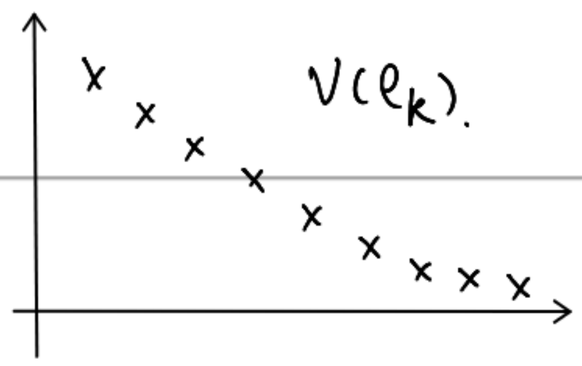
$$x^T Q x = y^T U^T Q U y = y^T \Lambda y = \sum_{i=1}^{n} \lambda_i y_i^2 \le \max|\lambda_i| \cdot \|y\|_2^2.$$

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

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. (continuous function)

note that $e_{k+1} = e_k + x^* - 2tQ(e_k + x^*) - x^* \triangleq g(e_k)$

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

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

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 $V(0) = 0 \quad V(e) > 0 \ \forall e \neq 0$. (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! $x_k \to x^*$ or $x_k \to x_{k+1}$.

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.

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

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

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

Consider $f(x) = |x|$ with $|f'(x)| = 1$. oscillation between $-\frac{t}{2}$ and $\frac{t}{2}$.

Idea: bound $\|\nabla f(x_k)\|^2$ by $f(x_k) - f(x_{k+1})$

Note that $f(x_{k+1}) - f(x_k) \ge \nabla f(x_k)^T \cdot (x_{k+1} - x_k) = -t\|\nabla f(x_k)\|^2$.

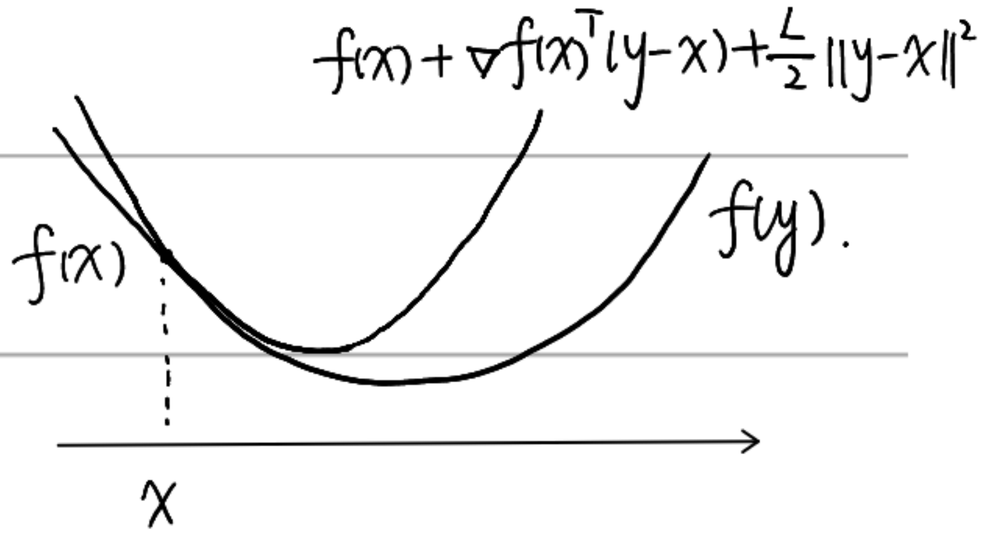
Hope LHS $\le -\frac{t}{2}\|\nabla f(x_k)\|^2 \Leftarrow f(y) \le f(x) + \nabla f(x)^T(y-x) + ?$

$? = \frac{t}{2}\|\nabla f(x)\|^2$ X

recall that $x_{k+1} - x_k = -t\nabla f(x)$

So $? = \frac{1}{2t}\|y-x\|^2$.

let $? = \frac{L}{2}\|y-x\|^2$. choose $t < \frac{1}{L}$



$f(y) \le f(x) + f'(x)(y-x) + \frac{L}{2}(y-x)^2 \Leftarrow f'(y) \le f'(x) + L(y-x)$.

Lipschitz Continuity: $f: \mathbb{R}^n \to \mathbb{R}^m$ is Lipschitz continuous with $L > 0$.

if $\|f(x) - f(y)\| \le L\|x-y\|$. $\forall x, y$. (or simply L-Lipschitz).

Example. $f(x) = w^T x$ is $\|w\|$-Lipschitz since. $|w^T(x-y)| \le \|w\| \cdot \|x-y\|$.

Example. $f(x) = Qx$ with $Q \succeq 0$ is $\lambda_{max}(Q)$-Lipschitz.

recall $\|f(x) - f(y)\| = \|(x-y)^T Q^2 (x-y)\| \le (\lambda_{max}(Q^2))^{1/2}\|x-y\|$.

remark: if $Q \not\succeq 0$. then is $\sigma_{max}(Q)$-Lipschitz (singular value).

L-smooth: $f: \mathbb{R}^n \to \mathbb{R}$ is L-smooth if ∇f is L-Lipschitz. i.e.,

$$\|\nabla f(x)-\nabla f(y)\| \le L\|x-y\| \quad \forall x, y.$$

Example. $f(x)=x^TQx$ with $Q\succeq 0$ is $2\lambda_{max}(Q)$-smooth.

Lemma. a twice continuous differentiable $f:\mathbb{R}^n\to\mathbb{R}$ is L-smooth

if $-LI \preceq \nabla^2 f(x) \preceq LI$. $\forall x$. or equivalently $|\lambda|\le L$. $\forall$ eigenvalues λ.

Proof. "$\Leftarrow$". assume $-LI \preceq \nabla^2 f(x) \preceq LI$. $\forall x$.

$\exists z = x+\xi(y-x)$.

By the mean value theorem. $\nabla f(x)-\nabla f(y) = \nabla^2 f(z)(x-y)$

$$\|\nabla f(x)-\nabla f(y)\| = \|\nabla^2 f(z)(x-y)\| = (x-y)^T(\nabla^2 f(z))^2(x-y) \le L\|x-y\|.$$

"$\Rightarrow$". assume f is L-smooth. fix $d\in\mathbb{R}^n$. let $g(\xi)=\nabla f(x+\xi d)$.

$$\Rightarrow \|g(\xi)-g(0)\| \le \xi L\|d\| \Rightarrow \|(g(\xi)-g(0))/\xi\| \le L\|d\|.$$

taking the limit and using the chain rule. $g'(0)=\nabla^2 f(x)\cdot d$

$$\Rightarrow \forall d\in\mathbb{R}^n \quad \|\nabla^2 f(x)\,d\| \le L\|d\| \Rightarrow |\lambda|\le L. \ \forall \text{ eigenvalue } \lambda. \quad \square$$

In particular. if f convex. f is L-smooth iff $\nabla^2 f(x)\preceq LI$.

Lemma. f is L-smooth $\Rightarrow f(y)\le f(x)+\nabla f(x)^T(y-x)+\frac{L}{2}\|x-y\|^2$.

Proof. Let $z(\xi)=x+\xi(y-x)$. $g(\xi)=f(z(\xi))$. so $g'(\xi)=\nabla f(z)^T(y-x)$

$$|g'(\xi)-g'(0)| = (\nabla f(z(\xi))-\nabla f(z(0)))^T(y-x) \le L\xi\|y-x\|^2.$$

$$f(y)-f(x)-\nabla f(x)^T(y-x) = g(1)-g(0)-g'(0) = \int_0^1 g'(\xi)-g'(0)\,d\xi \le \frac{L}{2}\|y-x\|^2.$$

$\square$.

Recall our goal is to bound $\sum_{k=0}^{T-1} \|\nabla f(x_k)\|^2$.

f is L-smooth $\Rightarrow f(x_{k+1}) \le f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2$.

$\Rightarrow f(x_{k+1}) \le f(x_k) - t\|\nabla f(x_k)\|^2 + \frac{t^2 L}{2}\|\nabla f(x_k)\|^2$. choose $0 < t < 1/L$.

$\Rightarrow f(x_{k+1}) \le f(x_k) - \frac{t}{2}\|\nabla f(x_k)\|^2 \Rightarrow \frac{t}{2}\sum_{k=0}^{T-1}\|\nabla f(x_k)\|^2 \le f(x_0) - f(x_T)$.

$$\sum_{k=0}^{T-1}(f(x_k) - f(x^*)) \le \frac{1}{2t}\left(\|x_0 - x^*\|^2 - \|x_T - x^*\|^2\right) + \frac{t}{2}\sum_{k=0}^{T-1}\|\nabla f(x_k)\|^2$$

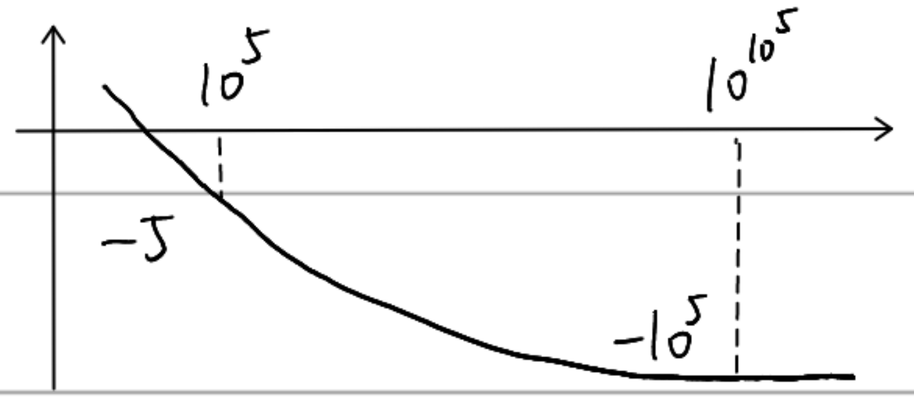
$$\Rightarrow \sum_{k=0}^{T-1}(f(x_k) - f(x^*)) \le \frac{1}{2t}\|x_0 - x^*\|^2 + f(x_0) - f(x_T)$$

$$\Rightarrow \sum_{k=1}^{T}(f(x_k) - f(x^*)) \le \frac{1}{2t}\|x_0 - x^*\|^2 \Rightarrow f(x_T) - f(x_0) \le \frac{\|x_0 - x^*\|^2}{2tT}$$

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 ow.



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