Lecture # 7 Gradient methods (unconstrained problems) E1 260

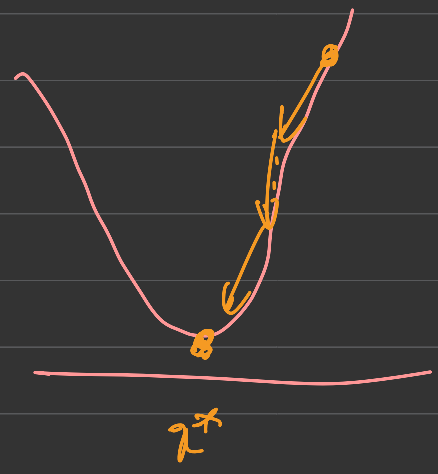
- Descent methods : Gradient, Steepest descent

- Gradient descent convergence :
  - Vanilla analysis (average iterates)
  - Convex & Lipschitz (average iterates $O(1/\sqrt{k})$
  - Convex & L-Smooth [last iterates $O(1/k)$]

Differntiable function minimization (unconstrained)

$$f : \mathbb{R}^n \longrightarrow \mathbb{R}$$

$$\boxed{\begin{array}{ll} \text{minimize} & f(\underline{x}) \\ \text{Subject to} & \underline{x} \in \mathbb{R}^n \end{array}}$$



Iterative "descent" algorithms :

Start with $\underline{x}_0$ and compute a sequence $\underline{x}_1, \underline{x}_2 \ldots$

such that

$$f(\underline{x}_{t+1}) < f(\underline{x}_t) \qquad t = 0, 1, \ldots$$

Descent direction:

$$f(\underline{x}_t + \underline{d}_t) \simeq f(\underline{x}_t) + \nabla f^T(\underline{x}_t)\,\underline{d}_t$$

directional derivative: Change in f for a small step $\underline{d}_t$

$$f'(\underline{x}_t ; \underline{d}_t) := \lim_{\tau \to 0} \frac{f(\underline{x}_t + \tau \underline{d}_t) - f(\underline{x}_t)}{\tau}$$

$$= \nabla f^T(\underline{x}_t)\,\underline{d}_t$$

- $\underline{d}_t$ is a descent direction at $\underline{x}_t$ if $\nabla f^T(\underline{x}_t)\,\underline{d}_t < 0$

- For convex f, (first-order condition)

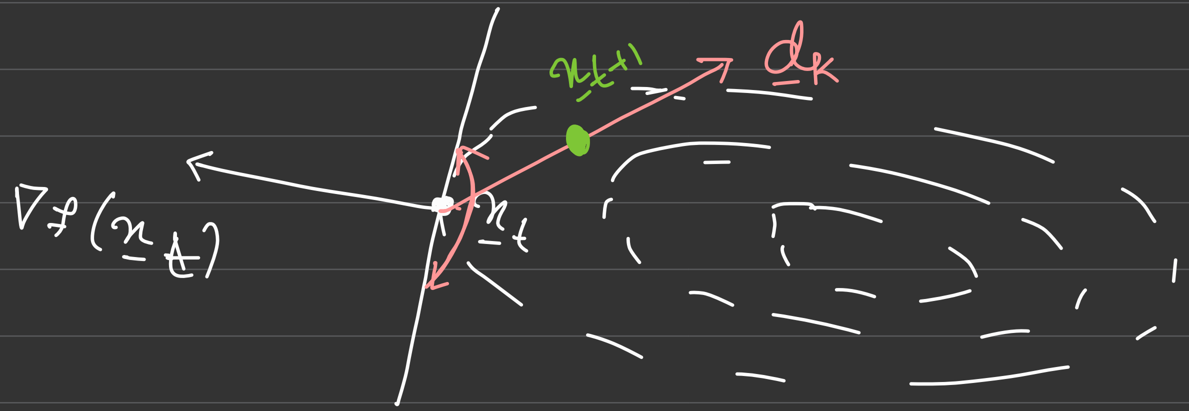
$$\nabla f^T(\underline{x}_t)\,(\underbrace{\underline{x}_{t+1} - \underline{x}_t}_{\underline{d}_t}) \geq 0 \implies f(\underline{x}_{t+1}) \geq f(\underline{x}_t)$$

So a descent direction must satisfy

$$\nabla f^T(\underline{x}_t)\,\underline{d}_t \leq 0$$

Iterative descent methods:

$$\underline{x}_{t+1} = \underline{x}_t + \eta_t \, \underline{d}_t \quad ; \quad \overbrace{\eta_t > 0}^{\text{Step size}}$$



Gradient descent: [Cauchy 1847]

$- \underline{\beta} \, \nabla f(x_t)$

$\underline{\beta > 0}$

$$\underline{x}_{t+1} = \underline{x}_t - \eta_t \nabla f(\underline{x}_t)$$

$$\underline{d}_t = -\nabla f(\underline{x}_t) \implies \nabla f^T(\underline{x}_t) \left(-\nabla f(\underline{x}_t) \right)$$

$$= -\|\nabla f(\underline{x}_t)\|_2^2 < 0$$

Steepest descent method:

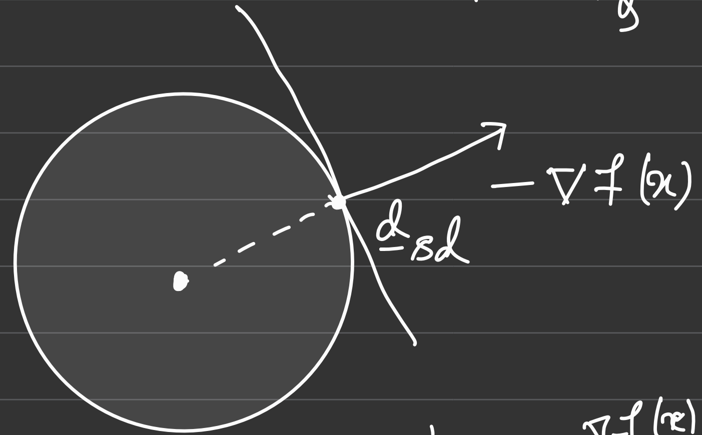
→ descent direction that is as negative as possible that yields the greatest rate of objective value improvement

$$\underline{d}_{sd} = \arg\min_{\underline{d}} \left\{ \nabla f^T(\underline{x})\, \underline{d} \;:\; \|\underline{d}\| \le 1 \right\}$$

$\|\underline{d}\|_g \le 1$

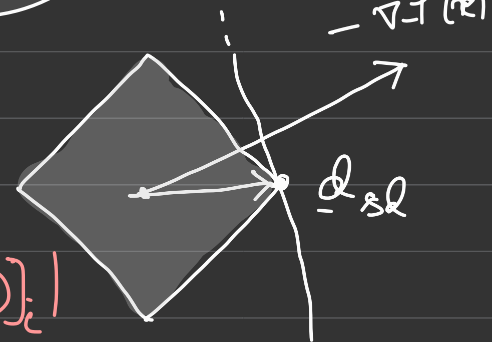
For Euclidean norm:

$$\underline{d}_{sd} = -\nabla f(\underline{x})$$



For l_1-norm:

$$\underline{d}_{sd} = \arg\min_{\underline{d}} \left\{ \nabla f^T(\underline{x})\, \underline{d} \;:\; \|\underline{d}\|_1 \le 1 \right\}$$



Let i be any index s.t. $\|\nabla f(\underline{x})\|_\infty = |[\nabla f(x)]_i|$

Then $\underline{d}_{sd} = -\operatorname{sign}\left(\dfrac{\partial f(\underline{x})}{\partial x_i}\right) \underline{e}_i$: coordinate descent.

Vanilla analysis: $\underline{x}_{t+1} = \underline{x}_t - \eta \nabla f(\underline{x}_t)$

$t = 0, \ldots, T-1$

Define: $\underline{g}_t = \nabla f(\underline{x}_t)$

so $\underline{g}_t = (x_{t+1} - x_t)/\eta$

Let us relate this vector to our current direction from an optimum $\underline{x}^*$: $x_t - \underline{x}^*$

$$\underline{g}_t^T(\underline{x}_t - \underline{x}^*) = \frac{1}{\eta}(\underline{x}_{t+1} - x_t)^T(\underline{x}_t - x^*)$$

Cosine theorem: $2\underline{v}^T\underline{w} = \|\underline{v}\|^2 + \|\underline{w}\|^2 - \|\underline{v} - \underline{w}\|_2^2$

$$\underline{g}_t^T(\underline{x}_t - \underline{x}^*) = \frac{1}{2\eta}\left[\|\underline{x}_{t+1} - \underline{x}_t\|_2^2 + \|\underline{x}_t - \underline{x}^*\|_2^2 - \|\underline{x}_{t+1} - x^*\|_2^2\right]$$

(*)

$$= \frac{1}{2\eta}\left[\eta^2\|\underline{g}_t\|_2^2 + \|\underline{x}_t - x^*\|_2^2 - \|\underline{x}_{t+1} - \underline{x}^*\|_2^2\right]$$

$$= \frac{\eta}{2}\|\underline{g}_t\|^2 + \frac{1}{2\eta}\left[\|\underline{x}_t - x^*\|_2^2 - \|\underline{x}_{t+1} - \underline{x}^*\|^2\right]$$

(*) Sum over the iteration t.

$$\boxed{a_n - a_{n+1}}$$

$$\sum_{t=0}^{T-1} \underline{g}_t^T (\underline{x}_t - \underline{x}^*) = \frac{\eta}{2} \sum_{t=0}^{T-1} \|\underline{g}_t\|^2$$

$$+ \frac{1}{2\eta} \left[\|\underline{x}_0 - \underline{x}^*\|_2^2 - \|\underline{x}_T - \underline{x}^*\|_2^2 \right]$$

$$\leq \frac{\eta}{2} \sum_{t=0}^{T-1} \|\underline{g}_t\|^2 + \frac{1}{2\eta} \|\underline{x}_0 - \underline{x}^*\|_2^2$$

Now, suppose f is convex : $f(\underline{y}) \geq f(x) + \nabla f^T(x)(y-x)$

$$\left.\begin{array}{l} y = x^* \\ x = x_t \end{array}\right\} \Rightarrow f(x_t) - f(x^*) \leq \nabla f^T(\underline{x}_t)(\underline{x}_t - x^*)$$

$$= \underline{g}_t^T (\underline{x}_t - \underline{x}^*)$$

- Upper bound on the average error: $f(\underline{x}_t) - f(\underline{x}^*)$

$$\sum_{t=0}^{T-1} f(x_t) - f(x^*) \leq \frac{\eta}{2} \sum_{t=0}^{T-1} \|\underline{g}_t\|^2 + \frac{1}{2\eta} \|\underline{x}_0 - \underline{x}^*\|_2^2$$

- Last iterate is not neccessarily the best one as "fixed" step size can make steps overshoot & increase function value

- For Lipschitz convex functions :

$f : X \longrightarrow \mathbb{R}^n$ is called Lipschitz continuous if there exists $B \geq 0$

$$\| f(\underline{x}) - f(\underline{y}) \| \leq B \, \| \underline{x} - \underline{y} \| \quad \forall \, \underline{x}, \underline{y} \in X$$

$$\iff \quad \| \nabla f(\underline{x}) \| \leq B \quad \forall \, \underline{x} \in X$$

$$f : \mathbb{R}^n \rightarrow \mathbb{R}$$

- $\| \underline{x}_0 - \underline{x}^* \| \leq R$

Then,

$$\sum_{t=0}^{T-1} \left(f(\underline{x}_t) - f(\underline{x}^*) \right) \leq \frac{\eta}{2} B^2 T + \frac{1}{2\eta} R^2 \quad (**)$$

So, choose η such that $q(\eta) = \frac{\eta}{2} B^2 T + \frac{1}{2\eta} R^2$ is minimized.

$$\frac{d}{d\eta} q(\eta) = 0 \implies \frac{1}{2} B^2 T - \frac{1}{2\eta^2} R^2 = 0$$

$$\implies \eta = \frac{R}{B\sqrt{T}} \quad \text{and} \quad q(\eta) = RB\sqrt{T}$$

$(**) \div T$

$$\frac{1}{T} \sum_{t=0}^{T-1} \left(f(\underline{x}_t) - f(\underline{x}^*) \right) \leq \frac{RB}{\sqrt{T}} \approx O\left(\frac{1}{\sqrt{T}}\right)$$

independent of 'n' but R & B does depend on n

So to obtain $\min_{t=0..T-1} f(\underline{x}_t) - f(\underline{x}^*) \leq \varepsilon$

we need $T \geq \frac{R^2 B^2}{\varepsilon^2}$. E.g. $\varepsilon = 10^{-8}$, R and B $\approx 10^d$??