From the previous lecture:

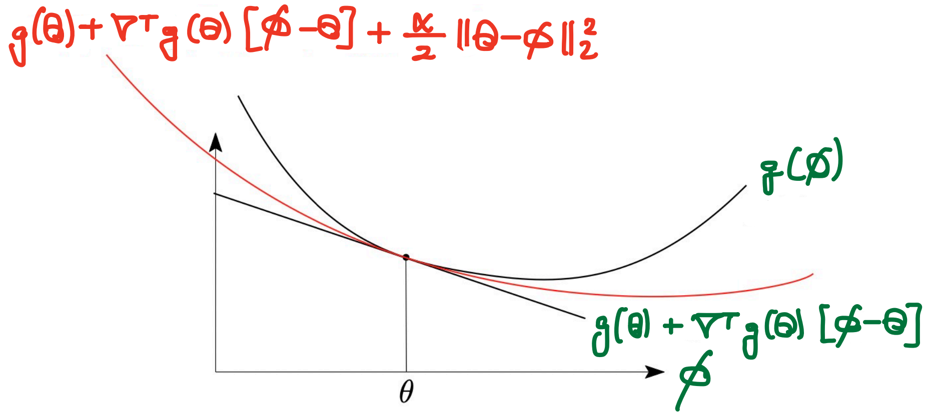
If $g:\mathbb{R}^d \to \mathbb{R}$ is convex and Lipschitz cont., then projected subgradient descent achieves $g\left(\frac{1}{T}\sum_{t=0}^{T-1}\theta_t\right) - g(\theta^*) \leq \frac{}{\sqrt{T}}$, which implies $O\left(\frac{1}{\varepsilon^2}\right)$ oracle complexity.

Today: improved convergence rates under strong convexity and smoothness.

DEF 1 (α-strong convexity)

$g:\mathbb{R}^d \to \mathbb{R}$ differentiable on $\mathrm{dom}(g)$ is α-strongly convex ($\alpha \geq 0$) if:

$$g(\theta) + \nabla^T g(\theta)[\theta'-\theta] \leq g(\theta') - \frac{\alpha}{2}\cdot\|\theta-\theta'\|_2^2, \quad \forall \theta,\theta' \in \mathrm{dom}(g).$$



PROP 1 (a) $g:\mathbb{R}^d \to \mathbb{R}$ diff. is α-strongly convex iff

$$\|\nabla g(\theta) - \nabla g(\theta')\|_2 \geq \alpha\cdot\|\theta-\theta'\|_2, \quad \forall \theta,\theta' \in \mathrm{dom}(g).$$

(b) $g:\mathbb{R}^d \to \mathbb{R}$ twice diff. is α-strongly convex iff

$$\nabla^2 g(\theta) \succeq \alpha I, \quad \forall \theta \in \mathrm{dom}(g). \quad (\text{i.e., } \nabla^2 g(\theta) - \alpha I \text{ is p.s.d.})$$

Proof: Exercise.

LEMMA 1 g is α-sc. iff $\theta \mapsto g(\theta) - \frac{\alpha}{2}\|\theta\|_2^2$ is convex.

LEMMA 2 $g: \mathbb{R}^d \to \mathbb{R}$ is strongly convex and continuous.

$\Rightarrow$ there exists a unique $\theta^* \in \underset{\theta \in \mathrm{dom}\, g}{\operatorname{argmin}}\; g(\theta)$.

The proofs of Lemma 1 and Lemma 2 can be found in the appendix.

The following ineq. is of fundamental importance.

LEMMA 3 (Lojasiewicz ineq.) $g: \mathbb{R}^d \to \mathbb{R}$ differentiable, α-sc.

Then,
$$\|\nabla g(\theta)\|_2^2 \geq 2\alpha \cdot \left[g(\theta) - \min_{\theta'} g(\theta') \right], \quad \forall \theta \in \mathrm{dom}(g).$$

Pf: Let $h(\phi) = g(\theta) + \nabla^T g(\theta)\,[\phi - \theta] + \frac{\alpha}{2}\|\theta - \phi\|_2^2$.

Then, h is minimized with $\phi^* = \theta - \frac{1}{\alpha}\nabla g(\theta)$ by the 1^{st}-order condition for optimality. Then,

$$g(\phi) \geq g(\theta) + \nabla^T g(\theta)\,[\phi - \theta] + \frac{\alpha}{2}\|\phi - \theta\|_2^2$$
$$\geq g(\theta) - \frac{1}{\alpha}\|\nabla g(\theta)\|_2^2 + \frac{1}{2\alpha}\|\nabla g(\theta)\|_2^2 = g(\theta) - \frac{1}{2\alpha}\|\nabla g(\theta)\|_2^2$$

Thus,
$$\inf_{\phi} g(\phi) \geq g(\theta) - \frac{1}{2\alpha}\|\nabla g(\theta)\|_2^2.$$

$$\Rightarrow \quad \|\nabla g(\theta)\|_2^2 \geq 2\alpha \cdot \left[g(\theta) - \inf_{\phi} g(\phi) \right], \ \forall \theta. \ \blacksquare$$

An immediate implication is that whenever $\nabla g(\theta) = 0$, $g(\theta) = \inf_{\phi} g(\phi)$.

PROPOSITION 2 (regularization)

Let $g : \mathbb{R}^d \to \mathbb{R}$ be a convex function, and $h : \mathbb{R}^d \to \mathbb{R}$ be an 1-sc function. Then, for any $\lambda \in \mathbb{R}_+$,

$$\theta \mapsto g(\theta) + \lambda h(\theta) \quad \text{is} \quad \lambda\text{-sc.}$$

Pf: $\varphi(\theta) = g(\theta) + \lambda h(\theta) \implies \nabla\varphi(\theta) = \nabla g(\theta) + \lambda \nabla h(\theta)$

Then,

$$\begin{aligned}\varphi(\theta) + \nabla^T\varphi(\theta)[\theta'-\theta] &= g(\theta) + \nabla^T g(\theta)[\theta'-\theta] + \lambda\left[h(\theta) + \nabla^T h(\theta)(\theta'-\theta)\right] \\ &\le g(\theta') + \lambda.\left[h(\theta') - \frac{1}{2}\|\theta-\theta'\|_2^2\right] \\ &= \varphi(\theta') - \frac{\lambda}{2}\|\theta-\theta'\|_2^2, \quad \forall \theta, \theta' \in \operatorname{dom} g \cap \operatorname{dom} h\end{aligned}$$

Note: A canonical example is Tikhonov regularization:

$$g(\theta) + \frac{\lambda}{2}\|\theta\|_2^2, \quad \lambda > 0 \quad \text{is} \quad \lambda\text{-sc.}$$

GRADIENT DESCENT FOR STRONGLY CONVEX FUNCTIONS

> **Algorithm: Gradient Descent**
>
> Inputs: $(\eta_t)_{t\ge 0}$ step-sizes
>
> $\theta_0 \in \operatorname{dom} g$
>
> for $t = 0, 1, \dots, T-1$
>
> $\theta_{t+1} = \theta_t - \eta_t \nabla g(\theta_t)$

What does GD do? Proximal form:

$$\theta_{t+1} \in \arg\min_{\theta} \Big\{ \underbrace{g(\theta_t) + \nabla^T g(\theta_t)[\theta-\theta_t]}_{\text{linear approximation around } g(\theta_t)} + \frac{1}{2\eta}\|\theta-\theta_t\|_2^2 \Big\}$$

$\|\theta-\theta_t\|_2^2$ penalizes moving from θ_t.

large step-size $\Rightarrow$ small penalty

$\Rightarrow$ move far from θ_t.

More generally,

$$\theta_{t+1} \in \arg\min_{\theta} \Big\{ g(\theta_t) + \nabla^T g(\theta_t)[\theta-\theta_t] + \frac{1}{\eta} \underbrace{D(\theta, \theta_t)}_{\text{Bregman divergence}} \Big\}$$

Mirror descent or non-Euclidean opt. (Nemirovski, Yudin, 1983)

THEOREM 1 (Convergence of GD for α-sc., L-Lipschitz functions)

$g : \mathbb{R}^d \to \mathbb{R}$ differentiable, α-sc. and L-Lipschitz.

Then, for any $T \geq 1$, with $\eta_t = \frac{1}{\alpha(t+1)}$,

(a) $$\|\theta_T - \theta^*\|_2^2 \leq \frac{L^2 \cdot (1+\log T)}{\beta^2 \cdot T},$$

(b) $$g\left(\frac{1}{T}\sum_{t=0}^{T-1}\theta_t\right) - g(\theta^*) \leq \frac{L^2}{2\beta T}[1+\log T],$$

where θ^* is the unique minimizer of g.

Pf: Lyapunov function : $\mathcal{L}(\theta) = \|\theta - \theta^*\|_2^2$, $\theta \in \mathbb{R}^d$.

Then, by α-sc. and L-Lipschitz continuity, $\forall t \geq 0$,

$$\mathcal{L}(\theta_{t+1}) = \mathcal{L}(\theta_t) - 2\eta_t \nabla^T g(\theta_t)[\theta_t - \theta^*] + \eta_t^2 \|\nabla g(\theta_t)\|_2^2$$

$$\leq \mathcal{L}(\theta_t) - 2\eta_t\left[g(\theta_t) - g(\theta^*) + \frac{\alpha}{2}\|\theta_t - \theta^*\|_2^2\right] + \eta_t^2 L^2$$

$$= \mathcal{L}(\theta_t)[1 - \eta_t\alpha] - 2\eta_t \Delta_t + \eta_t^2 L^2, \qquad \Delta_t := g(\theta_t) - g(\theta^*).$$

(a) Since $\Delta_t \geq 0$, $\forall t$, using $\eta_t = \frac{1}{\alpha(t+1)}$,

$$\mathcal{L}(\theta_T) \leq \frac{T-1}{T}\mathcal{L}(\theta_{T-1}) + \frac{L^2}{\alpha^2 T^2} \leq \frac{T-1}{T}\left[\frac{T-2}{T-1}\mathcal{L}(\theta_{T-2}) + \frac{L^2}{\alpha^2 (T-1)^2}\right] + \frac{L^2}{\alpha^2 T^2}$$

$$\leq \frac{T-2}{T-1}\mathcal{L}(\theta_{T-2}) + \frac{L^2}{\alpha^2 T}\left[\frac{1}{T} + \frac{1}{T-1}\right]$$

$$\vdots$$

$$\leq \frac{L^2}{\alpha^2 T}\left[1 + \frac{1}{2} + \cdots + \frac{1}{T}\right] \leq \frac{L^2}{\alpha^2 T}[\log T + 1].$$

(b) $$\mathcal{L}(\theta_T) \leq \frac{T-1}{T}\mathcal{L}(\theta_{T-1}) - \frac{2}{\alpha T}\Delta_{T-1} + \frac{L^2}{\alpha^2 T^2}$$

$$\leq -\frac{2}{\alpha T}\sum_{t=0}^{T-1}\Delta_t + \frac{L^2}{\alpha^2 T}(1+\log T) \quad \text{by induction.}$$

Then, $$\frac{1}{T}\sum_{t=0}^{T-1} g(\theta_t) - g(\theta^*) \leq \frac{L^2}{2\alpha T}[1+\log T] \quad \text{since } \mathcal{L} \geq 0.$$

Jensen's ineq. $\Rightarrow$ $$g\left(\frac{1}{T}\sum_{t=0}^{T-1}\theta_t\right) - g(\theta^*) \leq \frac{1}{T}\sum_{t=0} g(\theta_t) - g(\theta^*) \leq \frac{L^2(1+\log T)}{2\alpha T}$$

Remarks: (1) No need for projection when we have sc.

(2) Average-iterate rate: $g(\overline{\theta}_T) - g^* = \tilde{O}\left(\frac{1}{T}\right)$,

without sc. $g(\overline{\theta}_T) - g^* = \tilde{O}\left(\frac{1}{\sqrt{T}}\right)$.

(3) Last-iterate rate for parameter convergence:

$$\|\theta_T - \theta^*\|_2^2 = \tilde{O}\left(\frac{1}{T}\right).$$

Pf (LEMMA 1): Let $\varphi(\theta) = g(\theta) - \frac{\alpha}{2}\|\theta\|_2^2$, $\theta \in \operatorname{dom} g$. Then,

$$\nabla\varphi(\theta) = \nabla g(\theta) - \alpha\theta.$$

($\Rightarrow$)

$$\begin{aligned}
\varphi(\theta) + \nabla^T\varphi(\theta)[\theta'-\theta] &= g(\theta) - \frac{\alpha}{2}\|\theta\|_2^2 + [\nabla g(\theta) - \alpha\theta]^T(\theta'-\theta) \\
&= g(\theta) + \nabla^T g(\theta)[\theta'-\theta] + \frac{\alpha}{2}\|\theta\|_2^2 - \alpha\theta^T\theta' \\
&\le g(\theta') + \frac{\alpha}{2}\|\theta\|_2^2 - \alpha\theta^T\theta' - \frac{\alpha}{2}\|\theta-\theta'\|_2^2 \quad (g \text{ is } \alpha\text{-sc.}) \\
&= g(\theta') - \frac{\alpha}{2}\|\theta'\|_2^2 = \varphi(\theta').
\end{aligned}$$

($\Leftarrow$)

$$g(\theta) - \frac{\alpha}{2}\|\theta\|_2^2 + (\nabla g(\theta) - \alpha\theta)^T(\theta'-\theta) \le g(\theta') - \frac{\alpha}{2}\|\theta'\|_2^2$$

Then,

$$\begin{aligned}
g(\theta) + \nabla^T g(\theta)[\theta'-\theta] &\le g(\theta') - \frac{\alpha}{2}\|\theta'\|_2^2 + \alpha\theta^T(\theta'-\theta) + \frac{\alpha}{2}\|\theta\|_2^2 \\
&= g(\theta') - \frac{\alpha}{2}\|\theta'\|_2^2 + \alpha\theta^T\theta' - \frac{\alpha}{2}\|\theta\|_2^2 \\
&= g(\theta') - \frac{\alpha}{2}\|\theta'-\theta\|_2^2. \qquad \blacksquare
\end{aligned}$$