Convex Optimization

Why? Although almost all deep learning problems are highly nonconvex, most of the tools developed for convex optimization are directly or indirectly used in deep learning.

DEF (Convex set, convex function)

$\Theta \subset \mathbb{R}^d$ is a convex set if $\theta, \theta' \in \Theta \Rightarrow \gamma\theta + (1-\gamma)\theta' \in \Theta, \forall \gamma \in [0,1]$

$g : \mathbb{R}^d \to \mathbb{R}$ is a convex function if $dom(g)$ is a convex set, and

$$g(\gamma\theta + (1-\gamma)\theta') \le \gamma g(\theta) + (1-\gamma) g(\theta'), \quad \forall \theta, \theta' \in dom(g), \gamma \in [0,1].$$

As simple examples, $g(\theta) = a^T\theta + b$, $g(\theta) = \|\theta\|_p$ for $p \ge 1$,
$g(\theta) = \max\{0, 1 - a^T\theta\}$.

PROP (epigraph) Let $epi(g) = \{(\theta, t) \in \mathbb{R}^d \times \mathbb{R} : \theta \in dom(g), t \ge g(\theta)\}$.
Then, g is convex if and only if $epi(g)$ is convex.

Pf : Exercise.

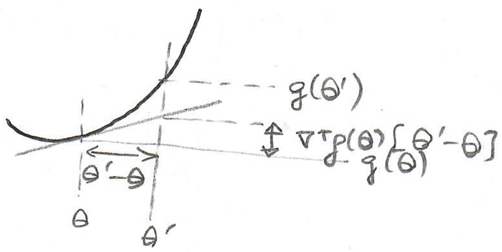
Recall that $g : \mathbb{R}^d \to \mathbb{R}$ is differentiable if $dom(g)$ is an open set, and

$$\nabla_\theta g(\theta) = \begin{bmatrix} \frac{\partial g}{\partial \theta_1} \\ \vdots \\ \frac{\partial g}{\partial \theta_d} \end{bmatrix}$$

exist for all $\theta \in dom(g)$.

PROP (1^{st}-order condition for convexity)
Suppose that $g : \mathbb{R}^d \to \mathbb{R}$ is differentiable with convex $dom(g)$.
Then,
if g is convex, $g(\theta) + \nabla^T g(\theta)[\theta' - \theta] \le g(\theta'), \forall \theta, \theta' \in dom\, g$.

Proof: For any $\theta, \theta' \in dom\, g$, $\gamma \in [0,1]$,



$$g(\gamma\theta' + (1-\gamma)\theta) \le \gamma g(\theta') + (1-\gamma) g(\theta)$$

Rearranging terms,

$$\frac{g(\theta + \gamma[\theta' - \theta]) - g(\theta)}{\gamma} + g(\theta) \le g(\theta').$$

g is diff. $\Rightarrow$ as $\gamma \to 0$, $\nabla^T g(\theta)[\theta' - \theta] + g(\theta) \le g(\theta')$

Note : The converse of the above is also true: if $\forall \theta, \theta' \in dom(g)$,
$g(\theta) + \nabla^T g(\theta)[\theta' - \theta] \le g(\theta')$, then g is convex.

PROP | (2nd-order condition for convexity)

Suppose that $g : \mathbb{R}^d \to \mathbb{R}$ is twice-differentiable, dom g is convex.
Then,

$$g \text{ is convex} \iff \nabla^2 g(\theta) \text{ is positive definite for all } \theta \in \text{dom } g.$$

PROP | (Some convexity-preserving operations)

(1) $h : \mathbb{R}^m \to \mathbb{R}$ is convex

$$\Rightarrow g(\theta) = h(A\theta + b) \text{ is convex for any } A \in \mathbb{R}^{m \times d},\ b \in \mathbb{R}^m.$$

(2) g_1, g_2 convex $\Rightarrow g_1 + g_2$ is convex.

$g_1, g_2, \dots, g_m$ convex $\Rightarrow$ $g(\theta) = \max_{i=1,2,\dots,m} g_i(\theta)$, $\theta \in \bigcap_{i=1}^m \text{dom}(g_i)$

is convex.

(3) $g : \mathbb{R}^d \to \mathbb{R}$, $h : \mathbb{R} \to \mathbb{R}$. Let $\varphi = h \circ g$.

φ is convex if
- g is convex, h is convex, h is nondecreasing
- g is concave, h is convex, h is nonincreasing.

Pf: (1)
$$\begin{aligned}
g(\gamma\theta + (1-\gamma)\theta') &= h(A(\gamma\theta + (1-\gamma)\theta') + b) \\
&= h(\gamma(A\theta + b) + (1-\gamma)(A\theta' + b)) \\
&\le \gamma h(A\theta + b) + (1-\gamma) h(A\theta' + b) \quad \text{by convexity of } h. \\
&= \gamma g(\theta) + (1-\gamma) g(\theta'). \Rightarrow g \text{ is convex.}
\end{aligned}$$

(2) $g_i(\gamma\theta + (1-\gamma)\theta') \le \gamma g_i(\theta) + (1-\gamma) g_i(\theta')$, $i = 1, 2$.

Then, $(g_1 + g_2)(\gamma\theta + (1-\gamma)\theta') \le \gamma(g_1+g_2)(\theta) + (1-\gamma)(g_1+g_2)(\theta')$.

$$\begin{aligned}
\max_i g_i(\gamma\theta + (1-\gamma)\theta') &\le \max_i \{\gamma g_i(\theta) + (1-\gamma) g_i(\theta')\} \\
&\le \max_i \gamma g_i(\theta) + \max_i (1-\gamma) g_i(\theta') \\
&\le \gamma \cdot \max_i g_i(\theta) + (1-\gamma)\max_i g_i(\theta')
\end{aligned}$$

(3) For $d = 1$. (Exercise : proof for $d > 1$)

$$\varphi'(\theta) = h'(g(\theta))\, g'(\theta)$$
$$\varphi''(\theta) = h''(g(\theta))[g'(\theta)]^2 + g''(\theta)\, h'(g(\theta)).$$

Use second-order condition for convexity.

ERM AS A CONVEX OPTIMIZATION PROBLEM

In some simple cases, ERM becomes a convex optimization problem.

EXAMPLE 1 $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$, $\Theta \subset \mathbb{R}^d$ compact and convex set.

$\mathcal{H}_\Theta = \{ x \mapsto \theta^T x \ : \ \theta \in \Theta, \ x \in \mathbb{R}^d \}$.

Then,

$$\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \theta^T x_i \right)^2$$ is a convex problem.

Pf: Exercise

Convexification of binary hypothesis testing:

Recall: $\hat{R}_S(f_\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{ f_\theta(x_i) \neq y_i \}$ for binary classification. Not convex, not continuous.

Idea: Use $h_\theta : \mathcal{X} \to \mathbb{R}$, $\theta \in \Theta$.

$$f_\theta(x) = \operatorname{sgn}(h_\theta(x)) \quad \text{for} \quad \operatorname{sgn}(z) = \begin{cases} -1, & z < 0 \\ 0, & z = 0 \\ 1, & z > 0. \end{cases}$$

Then,

$$\{ f_\theta(x_i) \neq y_i \} = \{ y_i h_\theta(x_i) \leq 0 \}$$

Thus, defining

$$l_{0-1}(z) = \mathbb{1}\{ z \leq 0 \},$$

$$\hat{R}_S(f_\theta) = \frac{1}{n} \sum_{i=1}^{n} l_{0-1}(y_i h_\theta(x_i)), \ \forall \theta \in \Theta.$$

Idea: Use convex surrogates for l_{0-1}.

Square loss: $l_{sq}(z) = (z-1)^2$,

Logistic loss: $l_{Log}(z) = \log(1+e^{-z})$,

Hinge loss: $l_{Hinge}(z) = \max\{0, 1-z\}$.

Then,

PROP For convex Θ, if $\theta \mapsto h_\theta$ is a concave function, then $\theta \mapsto \hat{R}_S(f_\theta)$ is a convex function with l_{Hinge} and l_{Log}.