Recall: $\Theta \subset \mathbb{R}^d$, $g : \Theta \to \mathbb{R}$. Suppose that g is differentiable over its domain. Then,

if g is convex, then

$$g(\theta) + \nabla^T g(\theta)[\theta' - \theta] \le g(\theta'), \quad \theta, \theta' \in \Theta$$

Non-Differentiable Convex Optimization

Note that a convex function is not always differentiable (everywhere) on its domain. E.g., $\underbrace{g(\theta) = |\theta|}$, $\underbrace{g(\theta) = \max\{0, \theta\}}_{\text{ReLU}}$.

Next, we generalize gradients.

DEF (Subgradients) $u \in \mathbb{R}^d$ is a subgradient of g at $\theta \in \Theta$ if

$$g(\theta') \ge g(\theta) + u^T[\theta' - \theta] \quad \text{for any } \theta' \in \Theta.$$

The set of subgradients of g at $\theta \in \Theta$ is called the subdifferential and is denoted as $\partial g(\theta)$.

PROP 1 (Existence of subgradients)

Let $\Theta \subset \mathbb{R}^d$ be convex, and $g : \Theta \to \mathbb{R}$.

(a) If $\partial g(\theta) \neq \emptyset$ for any $\theta \in \Theta$, then g is convex.

(b) If g is convex, then $\partial g(\theta) \neq \emptyset$ for any $\theta \in \Theta$.

Proof of PROP 1:

(a) Let $u \in \partial g(\gamma\theta + (1-\gamma)\theta')$. Then, by definition,

$$1-\gamma \,/ \quad g(\theta') \ge g(\gamma\theta + (1-\gamma)\theta') + \gamma \cdot u^T(\theta' - \theta)$$
$$\gamma \,/ \quad g(\theta) \ge g(\gamma\theta + (1-\gamma)\theta') - (1-\gamma)u^T(\theta' - \theta)$$

$$\gamma g(\theta) + (1-\gamma) g(\theta') \ge g(\theta\gamma + (1-\gamma)\theta'). \Rightarrow g \text{ is convex.}$$

(b) We will construct a subgradient of g at $\theta \in \operatorname{int} \operatorname{dom} g$. For this, we will use the supporting hyperplane theorem (see appendix).

Let $\theta \in \operatorname{int}(\Theta)$. Then, $(\theta, g(\theta)) \in \operatorname{epi}(g)$, indeed $(\theta, g(\theta)) \in \operatorname{bd}\operatorname{epi}(g)$. Since g is convex, $\operatorname{epi}(g)$ is convex. Thus, by the supporting hyperplane theorem, $\exists (w, v) \in \mathbb{R}^d \times \mathbb{R}$ s.t. $(w, v) \neq 0$, and

$$[w^T \;\; v] \begin{bmatrix} \theta \\ g(\theta) \end{bmatrix} \ge [w^T \;\; v] \begin{bmatrix} \theta' \\ t \end{bmatrix} \quad \text{for any } (\theta', t) \in \operatorname{epi}(g).$$

By tending $t \to \infty$, we see that $v \leq 0$ must hold.

Since $\theta \in \text{int}(\Theta)$, for sufficiently small $\varepsilon > 0$,

$$\tilde{\theta} = \theta + \varepsilon \omega \in \Theta.$$

$\Rightarrow v \neq 0$, since otherwise

$$\begin{bmatrix} \omega^T & 0 \end{bmatrix} \begin{bmatrix} \theta \\ g(\theta) \end{bmatrix} \geq \begin{bmatrix} \omega^T & 0 \end{bmatrix} \begin{bmatrix} \theta + \varepsilon\omega \\ t \end{bmatrix}$$

$$\Rightarrow \omega^T\theta \geq \omega^T\theta + \varepsilon\|\omega\|_2^2 \Rightarrow \omega = 0,$$ which would contradict with $(\omega, v) \neq 0$.

Let $t = g(\theta')$. Then, for some $v < 0$,

$$\begin{bmatrix} \omega^T & v \end{bmatrix} \begin{bmatrix} \theta \\ g(\theta) \end{bmatrix} \geq \begin{bmatrix} \omega^T & v \end{bmatrix} \begin{bmatrix} \theta' \\ g(\theta') \end{bmatrix}$$

$$\Rightarrow \omega^T(\theta - \theta') + v g(\theta) \geq v g(\theta'), \quad \theta' \in \Theta.$$

$$\Rightarrow g(\theta) + \frac{\omega^T}{v}(\theta - \theta') \leq g(\theta'), \quad \text{since } v < 0.$$

Thus, $\partial g(\theta) \neq \emptyset$ for $\theta \in \text{int}\,\Theta$. ∎

Remark: If g is differentiable, (b) of PROP 1 automatically holds since $\nabla g(\theta) \in \partial g(\theta)$ by the first-order condition for convexity.

Pf 1 (Existence of subgradients : detailed)

(A) $\partial f(\theta) \neq \emptyset, \forall \theta \Rightarrow f$ is convex.

(B) f is convex $\Rightarrow \partial f(\theta) \neq \emptyset, \forall \theta \in \Theta$.

Pf (B) Supporting hyperplane theorem:

Let $\Theta \subset \mathbb{R}^d$ be a convex set, and $\theta_0 \in \mathrm{bd}\,\Theta$.
Then, $\exists w \in \mathbb{R}^d$ s.t. $w \neq 0$ and

$$w^T \theta \leq w^T \theta_0, \forall \theta \in \Theta.$$

f is convex $\Rightarrow$ epi(f) is convex.

$\theta \in \mathrm{int}(\Theta)$. Then, $(\theta, f(\theta)) \in \mathrm{epi}(f)$, $(\theta, f(\theta)) \in \mathrm{bd\,epi}(f)$

By SuHT, $\exists (w, v) \in \mathbb{R}^d \times \mathbb{R}$ s.t. $(w, v) \neq 0$ and

$$[w^T \ v] \begin{bmatrix} \theta \\ f(\theta) \end{bmatrix} \geq [w^T \ v] \begin{bmatrix} \theta' \\ t \end{bmatrix}$$

for any $\begin{bmatrix} \theta' \\ t \end{bmatrix} \in \mathrm{epi}(f)$. Then, since $t \geq f(\theta')$, we

conclude that $v \leq 0$. Since $\theta \in \mathrm{int}(\Theta)$,

$\exists \varepsilon > 0$ s.t. $\theta + \varepsilon w \in \Theta$.

Thus, if $v = 0$,

$$[w^T \ 0] \begin{bmatrix} \theta \\ f(\theta) \end{bmatrix} \geq [w^T \ 0] \begin{bmatrix} \theta + \varepsilon w \\ t \end{bmatrix}$$

$$\Rightarrow w = 0 \quad \Rightarrow\Leftarrow$$

Hence, $v < 0$ must hold.

Then, let $t = f(\theta')$.

$$\Rightarrow w^T \theta + v f(\theta) \geq w^T \theta' + v f(\theta')$$

$$\Rightarrow f(\theta) + \frac{w^T}{v}(\theta - \theta') \leq f(\theta')$$

$$f(\theta) + u^T(\theta' - \theta) \leq f(\theta')$$

with $u = -\frac{w}{v}$.

DEF (Lipschitz continuity) Let $\Theta \subset \mathbb{R}^d$.

$g : \Theta \to \mathbb{R}$ is Lipschitz continuous if there exists a positive constant L s.t

$$|g(\theta) - g(\theta')| \leq L \cdot \|\theta - \theta'\|_2 , \quad \forall \theta, \theta' \in \Theta .$$

Exercise : If $g : \Theta \to \mathbb{R}^d$ is differentiable and $\|\nabla g(\theta)\|_2 \leq L$ for any $\theta \in \Theta$, then g is L-Lipschitz continuous.

THEOREM 1 (Lipschitz continuity of convex functions)

Let $g : \mathbb{R}^d \to \mathbb{R}$, and K be a compact set contained in $\operatorname{int} \operatorname{dom}(g)$. Then, g is Lipschitz continuous on K.

For the proof of Theorem 1, we will need the following lemmas.

Lemma 1 (Convex functions are locally bounded)

Let g be convex, and $\theta_0 \in \operatorname{int} \operatorname{dom}(g)$. Then, g is locally bounded : $\exists\, \varepsilon > 0$ and $M(\theta_0, \varepsilon) > 0$ s.t.

$$\max_{\theta \in B_2(\theta_0, \varepsilon)} g(\theta) \leq M(\theta_0, \varepsilon) .$$

Lemma 2 (Convex functions are locally Lipschitz)

Let g be convex, and $\theta_0 \in \operatorname{int} \operatorname{dom}(g)$. Then, g is locally Lipschitz : $\exists\, \varepsilon > 0$ and $L(\theta_0, \varepsilon) > 0$ s.t.

$$|g(\theta) - g(\theta_0)| \leq L(\theta_0, \varepsilon) \cdot \|\theta - \theta_0\|_2 , \quad \forall \theta \in B_2(\theta_0, \varepsilon).$$

Lemma 3 The following statements hold for a convex function :

(a) If $\theta_0 \in \operatorname{int} \operatorname{dom}(g)$, and g is locally Lipschitz, then

$$\|u\| \leq \overline{L}(\theta_0, \varepsilon) < \infty \text{ for any } u \in \partial g(\theta_0)$$

(b) If $\|u\|_2 \leq L(\theta_0, \varepsilon)$ for $u \in \partial g(\theta_0)$, then

$$g(\theta_0) - g(\theta) \leq L(\theta_0, \varepsilon) . \|\theta_0 - \theta\|_2 .$$

PROP 2 (Local minima are global minima under convexity)

Let g be convex.

(i) If θ^* is a local minimum of g, then θ^* is a global minimum of g.

(ii) The above happens if and only if $0 \in \partial g(\theta^*)$.

Proof: Clearly, $0 \in \partial g(\theta) \iff \theta$ is a global minimum of g.

Now, assume that θ is a local minimum of g:

$$\exists\, \varepsilon > 0 \quad \text{s.t.} \quad g(\theta') \geq g(\theta), \quad \forall \theta' \in B_2(\theta, \varepsilon).$$

Then, for small enough γ, for any $\tilde{\theta} \in \operatorname{dom} g$,

$$g(\theta) \leq g((1-\gamma)\theta + \gamma\tilde{\theta}) \leq (1-\gamma) g(\theta) + \gamma g(\tilde{\theta})$$

$$\Rightarrow \quad \gamma g(\theta) \leq \gamma g(\tilde{\theta}) \quad \Rightarrow \quad g(\theta) \leq g(\tilde{\theta}). \quad \blacksquare$$

PROP 3 (First-order optimality condition: constrained case)

Let g be convex, and $\Theta \subset \mathbb{R}^d$ a closed convex set on which g is differentiable. Then,

$$\theta^* \in \operatorname*{argmin}_{\theta \in \Theta} g(\theta),$$

if and only if

$$\nabla^T g(\theta^*) [\theta^* - \theta] \leq 0, \quad \forall \theta \in \Theta.$$

Proof: ($\Leftarrow$) $g(\theta^*) + \nabla^T g(\theta^*) [\theta - \theta^*] \leq g(\theta)$, $\forall \theta \in \Theta$ by convexity.

$$g(\theta) - g(\theta^*) \geq -\nabla^T g(\theta^*) [\theta^* - \theta] \geq 0, \forall \theta \in \Theta$$

$$\Rightarrow \quad \theta^* \in \operatorname*{argmin}_{\theta \in \Theta} g(\theta)$$

($\Rightarrow$) For $\theta \in \Theta$, let $h(t) := g(\theta^* + t \cdot [\theta - \theta^*])$, $t \in [0,1]$.

h is a convex function. Pf: Exercise.

Then,

$$\frac{dh(t)}{dt} = \nabla^T g(\theta^* + t[\theta - \theta^*]) \cdot [\theta - \theta^*].$$ Suppose to the contrary that $\left.\frac{dh(t)}{dt}\right|_{t=0} < 0$. Then, $\exists\, t_0 \in (0,1)$ small enough s.t. $h(t_0) < h(0)$. This is a contradiction since $\theta^* \in \operatorname{argmin} g(\theta)$.

NONDIFFERENTIABLE CONVEX OPTIMIZATION

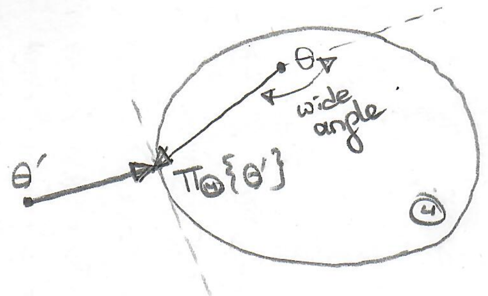
DEF | (Projection) Let $\Theta \subset \mathbb{R}^d$ be a compact and convex set. For any $\theta \in \mathbb{R}^d$, let

$$\Pi_{\Theta}\{\theta\} = \operatorname*{argmin}_{\theta' \in \Theta} \|\theta - \theta'\|_2 .$$

LEMMA | Let $\theta \in \Theta$, and $\theta' \in \mathbb{R}^d$. Then,

$$\left(\Pi_{\Theta}\{\theta'\} - \theta\right)^T \left(\Pi_{\Theta}\{\theta'\} - \theta'\right) \leq 0 .$$

Remark (Geometric intuition)



Proof: Let $h(\tilde{\theta}) = \frac{1}{2}\|\tilde{\theta} - \theta'\|_2^2$, $\tilde{\theta} \in \mathbb{R}^d$. Then, obviously, h is a convex function. By the first-order condition for optimality (PROP 3),

$$\nabla^T h\left(\Pi_{\Theta}\{\theta'\}\right)\left[\Pi_{\Theta}\{\theta'\} - \tilde{\theta}\right] \leq 0, \quad \forall \tilde{\theta} \in \Theta.$$

Thus,

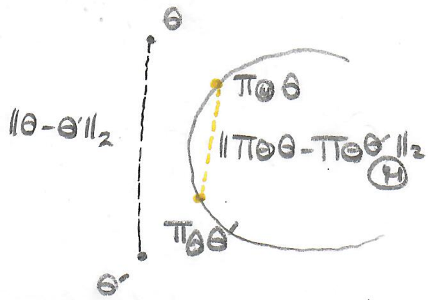
$$\left[\Pi_{\Theta}\{\theta'\} - \theta'\right] \cdot \left[\Pi_{\Theta}\{\theta'\} - \tilde{\theta}\right] \leq 0, \quad \forall \tilde{\theta} \in \Theta.$$

PROP 4 | (Projection is a nonexpansive operator)

Let $\Theta \subset \mathbb{R}^d$ be a compact and convex set. Then, for any $\theta, \theta' \in \mathbb{R}^d$,

$$\|\Pi_{\Theta}\{\theta\} - \Pi_{\Theta}\{\theta'\}\|_2 \leq \|\theta - \theta'\|_2 .$$

Proof: By LEMMA above,



Proof :

$$[\Pi_{\Theta}(\theta) - \theta]^T [\Pi_{\Theta}(\theta) - \underbrace{\Pi_{\Theta}(\theta')}_{\in \Theta}] \leq 0$$

$$[\Pi_{\Theta}(\theta') - \theta'] \; [\Pi_{\Theta}(\theta') - \Pi_{\Theta}(\theta)] \leq 0$$

Thus,

$$[\Pi_{\Theta}(\theta) - \Pi_{\Theta}(\theta')] \; [\Pi_{\Theta}(\theta) - \theta - (\Pi_{\Theta}(\theta') - \theta')]$$

$$= \|\Pi_{\Theta}(\theta) - \Pi_{\Theta}(\theta')\|_2^2 - (\Pi_{\Theta}(\theta) - \Pi_{\Theta}(\theta'))(\theta' - \theta) \leq 0$$

Hence, by using Cauchy - Schwarz,

$$\|\Pi_{\Theta}(\theta) - \Pi_{\Theta}(\theta')\|_2 \leq \|\theta - \theta'\|_2 .$$

ALGORITHM 1 $\Theta \subset \mathbb{R}^d$ compact and convex set

Inputs : $\theta_0 \in \Theta$, $\eta > 0$,

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

$$\tilde{\theta}_{t+1} = \theta_t - \eta u_t , \quad u_t \in \partial g(\theta_t),$$

$$\theta_{t+1} = \Pi_{\Theta}(\tilde{\theta}_{t+1}).$$

Return : $\bar{\theta}_T = \frac{1}{T} \sum_{t=0}^{T-1} \theta_t .$

THEOREM | (Lyapunov - deterministic)

Let $\Theta \subset \mathbb{R}^d$ be a compact and convex set.

For a given sequence $\theta_t \in \Theta$, $t \in \mathbb{Z}_+$, assume that there exist

$$\mathcal{L} : \Theta \to \mathbb{R}_+,$$
$$c : \Theta \to \mathbb{R}_+ \text{ (convex)},$$
$$\varepsilon \geq 0 \text{ and } V \in \mathbb{R} \text{ s.t.}$$

for any $\eta > 0$, $t \geq 0$,

$$\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_t) \leq -2\eta c(\theta_t) + \eta . \varepsilon + \eta^2 . V^2$$

Then, for any $T \geq 1$:

$$c\left(\frac{1}{T}\sum_{t=0}^{T-1} \theta_t\right) \leq \frac{\mathcal{L}(\theta_0)}{2\eta T} + \frac{\varepsilon}{2} + \frac{\eta V^2}{2}$$

COROLLARY | With $\eta = \sqrt{\frac{\mathcal{L}(\theta_0)}{T}} \cdot \frac{1}{V}$, we have:

$$c\left(\frac{1}{T}\sum_{t=0}^{T-1} \theta_t\right) \leq V\sqrt{\frac{\mathcal{L}(\theta_0)}{T}} + \frac{\varepsilon}{2}.$$

Proof By telescoping sum over $t = 0, 1, \dots, T-1$:

$$\mathcal{L}(\theta_T) - \mathcal{L}(\theta_0) \leq -2\eta \sum_{t=0}^{T-1} c(\theta_t) + \eta . T . \varepsilon + \eta^2 . T . V^2$$

Rearranging terms, since $\mathcal{L}(\theta_T) \geq 0$,

$$\frac{1}{T}\sum_{t=0}^{T-1} c(\theta_t) \leq \frac{\mathcal{L}(\theta_0)}{2\eta T} + \frac{\varepsilon}{2} + \eta\frac{V^2}{2}.$$

Since c is a convex function, by Jensen's inequality,

$$c\left(\frac{1}{T}\sum_{t=0}^{T-1} \theta_t\right) \leq \frac{1}{T}\sum_{t=0}^{T-1} c(\theta_t).$$

From the above inequalities, the result follows.

Performance analysis of Projected Subgradient Descent:

THEOREM (Near-optimality of Proj-GD) suppose $\sup_{\theta,\theta' \in \Theta} \|\theta-\theta'\|_2 \leq \rho$. Algorithm Proj-GD with $\eta = \dfrac{\rho}{L\sqrt{T}}$ satisfies

$$g\left(\frac{1}{T}\sum_{t=0}^{T-1}\theta_t\right) - \min_{\theta\in\Theta} g(\theta) \leq \frac{\rho L}{\sqrt{T}}$$

for any $T \geq 1$.

Pf: using $\mathcal{L}(\theta) = \|\theta-\theta^*\|_2^2$ where $\theta^* \in \arg\min_{\theta\in\Theta} g(\theta)$:

$$\begin{aligned}
\mathcal{L}(\theta_{t+1}) &= \|\theta_{t+1} - \theta^*\|_2^2 \\
&= \|\Pi_\Theta(\tilde\theta_{t+1}) - \Pi_\Theta(\theta^*)\|_2^2 \qquad \text{since } \theta^* \in \Theta \\
&\leq \|\tilde\theta_{t+1} - \theta^*\|_2^2 \qquad \text{by nonexpansiveness of } \Pi_\Theta \text{ (PROP 4)} \\
&= \|\theta_t - \eta u_t - \theta^*\|_2^2 \\
&= \|\theta_t - \theta^*\|_2^2 - 2\eta u_t^T(\theta_t-\theta^*) + \eta^2 \cdot \|u_t\|_2^2 \\
&\leq \mathcal{L}(\theta_t) - 2\eta u_t^T(\theta_t-\theta^*) + \eta^2 L^2
\end{aligned}$$

Then,

$$\mathcal{L}(\theta_{t+1}) - \mathcal{L}(\theta_t) \leq -2\eta \cdot u_t^T(\theta_t-\theta^*) + \eta^2 L^2.$$

By the defn of subgradient,

$$g(\theta_t) + u_t^T(\theta^* - \theta_t) \leq g(\theta^*).$$

Hence,

$$u_t^T(\theta_t-\theta^*) - g(\theta_t) \geq -g(\theta^*)$$

$$\Rightarrow u_t^T(\theta_t-\theta^*) \geq c(\theta_t)$$

where $c(\theta_t) = g(\theta_t) - g(\theta^*)$.

By Theorem (Lyapunov), since $\mathcal{L} \geq 0$ and $\theta \mapsto c(\theta)$ is convex, we have:

$$c(\bar\theta_T) \leq \frac{L\sqrt{\|\theta_0-\theta^*\|_2^2}}{\sqrt{T}}$$ since $\max_{\theta,\theta'\in\Theta} \|\theta-\theta'\|_2 \leq \rho$,

the result holds.