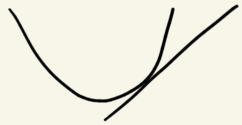


SGD

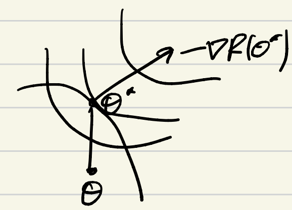
Def A function $R:\mathbb{R}^d \to \mathbb{R}$ is convex if $\forall \theta, \theta' \in \mathbb{R}^d$, $R(t\theta + (1-t)\theta') \le tR(\theta) + (1-t)R(\theta')$ $\forall t \in [0,1]$.

Lemma Let $R:\mathbb{R}^d \to \mathbb{R}$ be differentiable on the interior of its domain. R is convex iff
$\forall \theta, \theta' \in \mathbb{R}^d$ $R(\theta') \ge R(\theta) + \nabla R(\theta)^T(\theta' - \theta)$. ← 1st order approx is a global minorization

Pf) '⇒' From def of convexity, $R(\theta + t(\theta'-\theta)) \le R(\theta) + t(R(\theta') - R(\theta))$. $\equiv$ $R(\theta') - R(\theta) \ge \frac{1}{t}(R(\theta + t(\theta'-\theta)) - R(\theta))$. Send $t \downarrow 0$.
'⇐' Define $\theta_t = t\theta + (1-t)\theta'$. Combining $R(\theta) \ge R(\theta_t) + \nabla R(\theta_t)^T(\theta - \theta_t)$, $R(\theta') \ge R(\theta_t) + \nabla R(\theta_t)^T(\theta' - \theta_t)$,
$tR(\theta) + (1-t)R(\theta') \ge R(\theta_t) + \nabla R(\theta_t)^T(t\theta + (1-t)\theta' - \theta_t)$ $\forall t \in [0,1]$. ⊠

Rmk The latter def of convexity motivates generalization of gradients to nonsmooth, convex functions.

Optimality Consider $\min_{\theta \in \Theta} R(\theta)$, for $R:\mathbb{R}^d \to \mathbb{R}$ diff, convex.



Lemma $\theta^* = \operatorname{argmin}_{\theta \in \Theta} R(\theta)$ iff $\nabla R(\theta^*)^T(\theta - \theta^*) \ge 0$ $\forall \theta \in \Theta$

Pf) '⇐' From prev lemma, $R(\theta) - R(\theta^*) \ge \nabla R(\theta^*)^T(\theta - \theta^*) \ge 0$ $\forall \theta \in \Theta$.
'⇒' $\nabla R(\theta^*)^T(\theta - \theta^*) = \lim_{t \downarrow 0} \frac{1}{t}(R(\theta^* + t(\theta - \theta^*)) - R(\theta^*)) \ge 0$ $\forall \theta \in \Theta$. ⊠

Cor Let Θ be a closed convex set in $\mathbb{R}^d$. Define the projection operator $\Pi_\Theta(\theta) := \operatorname{argmin}_{\theta' \in \Theta} \|\theta - \theta'\|_2$.
Then, $\|\Pi_\Theta(\theta) - \theta'\|_2 \le \|\theta - \theta'\|_2$ $\forall \theta' \in \Theta$ $\forall \theta \in \mathbb{R}^d$.

Pf) From first order conditions for $\min_{\theta' \in \Theta} \|\theta - \theta'\|_2^2$,
$$0 \le (\Pi_\Theta(\theta) - \theta)^T(\theta' - \Pi_\Theta(\theta)) = (\Pi_\Theta(\theta) - \theta' + \theta' - \theta)^T(\theta' - \Pi_\Theta(\theta)) = -\|\theta' - \Pi_\Theta(\theta)\|_2^2 + (\theta' - \theta)^T(\theta' - \Pi_\Theta(\theta)).$$
From Cauchy-Schwarz, $\|\theta' - \Pi_\Theta(\theta)\|_2^2 \le \|\theta - \theta'\| \|\theta' - \Pi_\Theta(\theta)\|_2$. $\forall \theta' \in \Theta$ ⊠.

Stochastic gradients A stochastic gradient $G(\theta)$ is a RV s.t. $\mathbb{E}G(\theta) = \nabla R(\theta)$.
We study first-order optimization methods based on stoch. gradients.

(Canonical Problem)
$$\text{minimize}_{\theta \in \Theta} \{\mathbb{E}\,\ell(\theta; Z) =: R(\theta)\}.$$
If $\theta \mapsto \ell(\theta; z)$ is differentiable, then $\nabla_\theta \ell(\theta; Z)$ is a stochastic gradient if $\mathbb{E}$, ∇_θ can be interchanged.

- SGD Idea: Go in the direction of stoch. gradient, then project to Θ.
- Algo: Let $G_k(\theta)$ be a stoch. gradient of $R(\theta)$.
At each iteration k, $\theta_{k+1} = \Pi_\Theta(\theta_k - \alpha_k G_k(\theta_k))$ for some stepsize $\alpha_k > 0$.

We're implicitly assuming that projections are efficient to compute.

Rmk We can't even evaluate $\mathbb{E}\,\ell(\theta; Z)$. So SGD takes samples. In its simplest form, draw $Z_k \sim P$, then take $G(\theta_k) := \nabla_\theta \ell(\theta_k; Z_k)$. We could take multiple samples and average over them.

Rmk 2 We could consider ERM $\min_{\theta \in \Theta} \frac{1}{n}\sum \ell(\theta; z_i)$, and think of $\nabla_\theta \ell(\theta; z_i)$ as a stoch. gradient of the empirical loss. Our following convergence results still apply in this case.
The rationale for SGD w.r.t. empirical loss is purely computational:
instead of incurring $O(n)$ to evaluate each gradient, I want to compute an approximate gradient in $O(1)$.

Convergence Assume $\theta^* \in \arg\min_{\theta \in \Theta} R(\theta) > -\infty$ exists.

Theorem Let Θ be compact. Assume $\exists R>0$ s.t. $\sup_{\theta \in \Theta} \|\theta - \theta^*\|_2 \le D$, $\exists M>0$ s.t. $\mathbb{E}\|G(\theta)\|_2^2 \le M^2 \ \forall \theta \in \Theta$.
Let α_k be dec, pos. step sizes, and $\bar{\theta}_K = \frac{1}{K}\sum_1^K \theta_k$. Then,

$$\mathbb{E}[R(\bar{\theta}_K) - R(\theta^*)] \le \frac{D^2}{2K\alpha_K} + \frac{1}{2K}\sum_1^K \alpha_k M^2.$$

Pf) We expand on the error $\|\theta_{k+1} - \theta^*\|_2^2$.

$$\begin{aligned}\tfrac{1}{2}\|\theta_{k+1} - \theta^*\|_2^2 &= \tfrac{1}{2}\|\Pi_\Theta(\theta_k - \alpha_k G(\theta_k)) - \theta^*\|_2^2 \\ &\le \tfrac{1}{2}\|\theta_k - \alpha_k G(\theta_k) - \theta^*\|_2^2 \quad \text{by non-expansiveness of } \Pi_\Theta \\ &= \tfrac{1}{2}\|\theta_k - \theta^*\|_2^2 - \alpha_k \langle G(\theta_k), \theta_k - \theta^*\rangle + \tfrac{\alpha_k^2}{2}\|G(\theta_k)\|_2^2.\end{aligned}$$

Add & subtract $\alpha_k \langle \nabla R(\theta_k), \theta_k - \theta^*\rangle$ to get

$$\begin{aligned}&= \tfrac{1}{2}\|\theta_k - \theta^*\|_2^2 - \alpha_k\langle \nabla R(\theta_k), \theta_k - \theta^*\rangle + \tfrac{\alpha_k^2}{2}\|G(\theta_k)\|_2^2 - \alpha_k\langle G(\theta_k) - \nabla R(\theta_k), \theta_k - \theta^*\rangle \\ &\le \tfrac{1}{2}\|\theta_k - \theta^*\|_2^2 - \alpha_k(R(\theta_k) - R(\theta^*)) + \tfrac{\alpha_k^2}{2}\|G(\theta_k)\|_2^2 - \alpha_k\langle G(\theta_k) - \nabla R(\theta_k), \theta_k - \theta^*\rangle \quad \text{by convexity}\end{aligned}$$

Divide each side by α_k, and rearrange

$$R(\theta_k) - R(\theta^*) \le \frac{1}{2\alpha_k}\left(\|\theta_k - \theta^*\|_2^2 - \|\theta_{k+1} - \theta^*\|_2^2\right) + \frac{\alpha_k}{2}\|G(\theta_k)\|_2^2 - \langle G(\theta_k) - \nabla R(\theta_k), \theta_k - \theta^*\rangle \quad \cdots (*)$$

Now, note that
$$\begin{aligned}\sum_{k=1}^K \frac{1}{2\alpha_k}\left(\|\theta_k - \theta^*\|_2^2 - \|\theta_{k+1} - \theta^*\|_2^2\right) &= \frac{1}{2\alpha_1}\|\theta_1 - \theta^*\|_2^2 - \frac{1}{2\alpha_K}\|\theta_K - \theta^*\|_2^2 + \sum_{k=2}^K\left(\frac{1}{2\alpha_k} - \frac{1}{2\alpha_{k-1}}\right)\|\theta_k - \theta^*\|_2^2 \\ &\le \frac{D^2}{2\alpha_1} + \frac{D^2}{2}\sum_{k=2}^K\left(\frac{1}{\alpha_k} - \frac{1}{\alpha_{k-1}}\right) = \frac{D^2}{2\alpha_K}.\end{aligned}$$

So summing both sides of $(*)$,

$$\mathbb{E}\sum R(\theta_k) - R(\theta^*) \le \frac{D^2}{2\alpha_K} + \frac{1}{2}\sum \alpha_k M^2 - \sum_{k=1}^K \mathbb{E}\langle G(\theta_k) - \nabla R(\theta_k), \theta_k - \theta^*\rangle.$$

Taking expectations on both sides and noting

$$\begin{aligned}\mathbb{E}[\langle G(\theta_k) - \nabla R(\theta_k), \theta_k - \theta^*\rangle] &= \mathbb{E}[\mathbb{E}[\langle G(\theta_k) - \nabla R(\theta_k), \theta_k - \theta^*\rangle \mid \theta_k]] \\ &= \mathbb{E}[\langle \underbrace{\mathbb{E}[G(\theta_k)\mid\theta_k] - \nabla R(\theta_k)}_{0}, \theta_k - \theta^*\rangle] = 0,\end{aligned}$$

we get $\sum(R(\theta_k) - R(\theta^*)) \le \frac{D^2}{2\alpha_K} + \frac{1}{2}\sum \alpha_k M^2$. Noting $R(\bar{\theta}_K) \le \frac{1}{K}\sum R(\theta_k)$, we get the result. $\boxtimes$

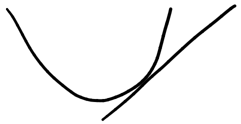
Cor For $\alpha_k = \frac{D}{M\sqrt{k}}$, $\mathbb{E}R(\bar{\theta}_K) - R(\theta^*) \le \frac{3DM}{2\sqrt{K}}$.

Pf) Noting $\sum_1^K \frac{1}{\sqrt{k}} \le \int_0^K \frac{1}{\sqrt{t}}\,dt = 2\sqrt{K}$, RHS $\le \frac{DM}{2\sqrt{K}} + \frac{DM}{\sqrt{K}}$. $\boxtimes$.

Rmk Think of K as # access to gradient oracle. If $G(\theta) = \nabla_\theta \ell(\theta; Z_i)$, then K = # samples.

Rmk Often, we iterate through data c times. This gives gains on empirical loss.
But population loss-wise, theory doesn't give gains as c grows.
In fact, we can't do better. We show this next class.

SGD



Def A function $R: \mathbb{R}^d \to \mathbb{R}$ is convex if $\forall \theta, \theta' \in \mathbb{R}^d$,

$$R(t\theta + (1-t)\theta') \leq tR(\theta) + (1-t)R(\theta') \quad \forall t \in [0,1].$$

Lemma Let $R: \mathbb{R}^d \to \mathbb{R}$ be differentiable on the interior of its domain.

R is convex iff $R(\theta') \geq R(\theta) + \nabla R(\theta)^T(\theta' - \theta)$, $\forall \theta, \theta' \in \mathbb{R}^d$

↑ 1st order approx is a global minorization

Optimality consider $\min_{\theta \in \Theta} R(\theta)$, for $R: \mathbb{R}^d \to \mathbb{R}$ diff; convex.

Lemma $\theta^* = \arg\min_{\theta \in \Theta} R(\theta)$ iff $\nabla R(\theta^*)^T(\theta - \theta^*) \geq 0 \quad \forall \theta \in \Theta$

Cor. Let Θ be a closed convex set in $\mathbb{R}^d$.

Define the projection operator $\Pi_{\Theta}(\theta) := \arg\min_{\theta' \in \Theta} \|\theta - \theta'\|_2$.

Then, $\|\Pi_{\Theta}(\theta) - \theta'\|_2 \leq \|\theta - \theta'\|_2 \quad \forall \theta' \in \Theta \quad \forall \theta \in \mathbb{R}^d$.

Stochastic gradients A stochastic gradient $G(\theta)$ is a RV s.t. $\mathbb{E}G(\theta) = \nabla R(\theta)$.

We study first-order optimization methods based on stoch. gradients.

(Canonical Problem) $\quad \text{minimize}_{\theta \in \Theta} \{ \mathbb{E}\, \ell(\theta; Z) =: R(\theta) \}$

- SGD Idea: Go in the direction of stoch. gradient, then project to Θ.
- Algo: let $G_k(\theta)$ be a stoch. gradient of $R(\theta)$.

At each iteration k,

$$\theta_{k+1} = \Pi_{\Theta}\left(\theta_k - \alpha_k G_k(\theta_k) \right) \quad \text{for some stepsize } \alpha_k > 0.$$

We're implicitly assuming that projections are efficient to compute.

Convergence Assume $\theta^* \in \arg\min_{\theta \in \Theta} R(\theta) > -\infty$ exists.

Theorem Let Θ be compact.

Assume $\exists D > 0$ s.t. $\sup_{\theta \in \Theta} \|\theta - \theta^*\|_2 \le D$, $\exists M > 0$ s.t. $\mathbb{E}\|G(\theta)\|_2^2 \le M^2$ $\forall \theta \in \Theta$.

Let α_k be dec, pos. step sizes, and $\bar{\theta}_K = \frac{1}{K}\sum_1^K \theta_k$

$$\mathbb{E}\left[R(\bar{\theta}_K) - R(\theta^*)\right] \le \frac{D^2}{2K\alpha_K} + \frac{1}{2K}\sum_1^K \alpha_k M^2.$$

Cor For $\alpha_k = \frac{D}{M\sqrt{k}}$, $\mathbb{E}R(\bar{\theta}_K) - R(\theta^*) \le \frac{3DM}{2\sqrt{K}}$.

Pf) Noting $\sum_1^K \frac{1}{\sqrt{k}} \le \int_0^K \frac{1}{\sqrt{t}}\,dt = 2\sqrt{K}$, RHS $\le \frac{DM}{2\sqrt{K}} + \frac{DM}{\sqrt{K}}$.