Remarks Contraction principle:

1) For $\phi:\mathbb{R}\to\mathbb{R}$, C_ϕ-Lip, $\mathbb{E}\left[\sup_{h\in\mathcal{H}} \frac{1}{n}\sum \sigma_i(\phi\circ h)(z_i) \,\middle|\, z_1^n\right] \le C_\phi \hat{R}_n\mathcal{H}$

2) For $\phi:\mathbb{R}\to\mathbb{R}$, C_ϕ-Lip & $\phi(0)=0$,

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}} \left|\frac{1}{n}\sum \sigma_i(\phi\circ h)(z_i)\right| \,\middle|\, z_1^n\right] \le 2C_\phi\, \mathbb{E}\left[\sup_{h\in\mathcal{H}} \left|\frac{1}{n}\sum \sigma_i h(z_i)\right| \,\middle|\, z_1^n\right]$$

Def A collection of zero mean RVs $\{V_h : h\in T\}$ is a sub-Gaussian process w.r.t. d if
$$\mathbb{E}e^{\lambda(V_h-V_{h'})} \le \exp\left(\frac{\lambda^2}{2} d(h,h')^2\right) \quad \forall h,h'\in T,\ \forall \lambda\in\mathbb{R}.$$
↳ tail of $V_h-V_{h'}$ is $d(h,h')^2$-sub G.

Key Lemma Let X_j be σ_i^2-sub-G RVs, $j=1,\dots,N$. Then, $\mathbb{E}\max_{1\le j\le N}|X_j| \le \max_{1\le j\le N}\sigma_j \cdot 2\sqrt{\log N}$, $N\ge 2$.

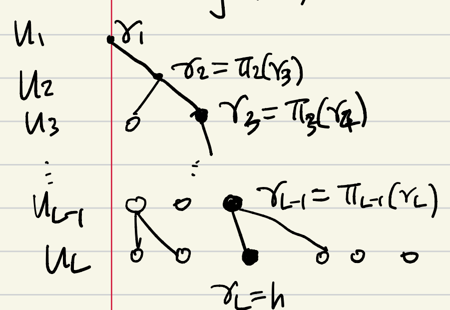
Theorem (Dudley's entropy integral) Let $\{V_h : h\in T\}$ be a sub-Gaussian process w.r.t. d on T.
For any $\delta\in[0,D]$,
$$\mathbb{E}\sup_{h\in T} V_h \le \mathbb{E}\left[\sup_{h,\tilde{h}\in T} V_h - V_{\tilde{h}}\right] \le 2\mathbb{E}\left[\sup_{\substack{r,r'\in T\\ d(r,r')\le\delta}} V_r - V_{r'}\right] + 32\int_{\delta/4}^{D}\sqrt{\log N(T,d,\varepsilon)}\, d\varepsilon$$

Pf) Let $N=N(T,d,\delta)$, and $\mathcal{U}:=\{h_j\}_{j=1}^N$ be a δ-cover of T. Fix an arbitrary $h\in T$.
There exists j s.t. $d(h,h_j)\le\delta$. Then,
$$V_h - V_{h_1} = V_h - V_{h_j} + V_{h_j} - V_{h_1} \le \sup_{\substack{r,r'\in T\\ d(r,r')\le\delta}} (V_r - V_{r'}) + \max_{1\le j\le N}|V_{h_j}-V_{h_1}|$$
Given another arbitrary $\tilde{h}\in T$, the same bound holds for $V_{h_1}-V_{\tilde{h}}$.
Adding the two, and taking supremum over $h,\tilde{h}\in T$
$$\sup_{h,\tilde{h}\in T} V_h - V_{\tilde{h}} \le 2\sup_{\substack{r,r'\in T\\ d(r,r')\le\delta}} (V_r - V_{r'}) + 2\max_{1\le j\le N}|V_{h_j}-V_{h_1}|$$
$$\le 2\sup_{\substack{r,r'\in T\\ d(r,r')\le\delta}} (V_r - V_{r'}) + 2\sup_{h,\tilde{h}\in\mathcal{U}}|V_h - V_{\tilde{h}}|.$$

Instead of bounding the last term via Lemma, we use a chaining argument.
For L s.t. $D2^{-L}\le\delta$, think of $\mathcal{U}_L=\mathcal{U}$ as a $(D2^{-L})$-cover of $\mathcal{U}$. Now, for each $m=1,\dots,L-1$,
define $\mathcal{U}_m :=$ minimal $(D\cdot 2^{-m})$-cover of $\mathcal{U}_{m+1}$ (we allow elements of T).

Best approx of $h\in\mathcal{U}$ in $\mathcal{U}_m$.

By def, $|\mathcal{U}_m| \le N(T,d,D\cdot 2^{-m})$.
For each $m=1,\dots,L-1$, define $\pi_m : \mathcal{U}_{m+1}\to\mathcal{U}_m$, $\pi_m(h) = \operatorname{argmin}_{h'\in\mathcal{U}_m} d(h,h')$
Using this, we construct a chain from any $h\in\mathcal{U}$. $\gamma_{m-1} = \pi_{m-1}(\gamma_m)$



$$V_h - V_{\gamma_1} = \sum_{m=2}^{L} V_{\gamma_m} - V_{\gamma_{m-1}} \quad \text{and}$$
$$\mathbb{E}|V_h - V_{\gamma_1}| \le \sum_{m=2}^{L}\sup_{r\in\mathcal{U}_m}|V_r - V_{\pi_{m-1}(r)}| \quad //$$

Similarly, for any other $\tilde{h}\in\mathcal{U}$, we have same bound with $\tilde{\gamma}_m$'s.

We arrive at $|V_h - V_{\tilde h}| = |V_h - V_{r_1} + V_{r_1} - V_{\tilde r_1} + V_{\tilde r_1} - V_{\tilde h}|$

$$\le |V_{r_1} - V_{\tilde r_1}| + \underbrace{|V_h - V_{r_1}| + |V_{\tilde r_1} - V_{\tilde h}|}_{\text{bound via chaining}}$$

$$\le \max_{r_1, \tilde r_1 \in U_1} |V_{r_1} - V_{\tilde r_1}| + 2\sum_{m=2}^{L} \max_{r \in U_m} |V_r - V_{\pi_{m-1}(r)}|$$

Now note that $\sup_{r_1, \tilde r_1 \in U_1} d(r_1, \tilde r_1) \le D$, and $V_{r_1} - V_{\tilde r_1}$ is $d(r_1, \tilde r_1)^2$-subGaussian.

and $\max_{r \in U_m} d(r, \pi_{m-1}(r)) \le D \cdot 2^{-(m-1)}$, and $|U_m| \le N(T, d, D \cdot 2^{-m})$.

From Lemma, $\mathbb{E} \max_{r_1, \tilde r_1 \in U_1} |V_{r_1} - V_{\tilde r_1}| \le 2D\sqrt{\log N(T, d, \frac{D}{2})}$, and

$$\mathbb{E} \max_{r \in U_m} |V_r - V_{\pi_{m-1}(r)}| \le 2D 2^{-(m-1)} \sqrt{\log N(T, d, D \cdot 2^{-m})}$$

Conclude that

$$\mathbb{E} \sup_{h, \tilde h \in U} |V_h - V_{\tilde h}| \le 4 \sum_{m=1}^{L} D \cdot 2^{-(m-1)} \sqrt{\log (T, d, D 2^{-m})}$$

Since $\delta \mapsto \log N(T, d, \delta)$ is dec, $D \cdot 2^{-m} \sqrt{\log N(T, d, D \cdot 2^{-m})} \le 2 \int_{D 2^{-(m+1)}}^{D 2^{-m}} \sqrt{\log N(T, d, \varepsilon)}\, d\varepsilon$

$$\Rightarrow 2\mathbb{E} \sup_{h, \tilde h \in T} |V_h - V_{\tilde h}| \le 32 \int_{\delta/4}^{D} \sqrt{\log N(T, d, \varepsilon)}\, d\varepsilon$$

Combining with $*$, we get the result. $\boxtimes$.

__Rmk__ Measurability, asymptotic versions $\sqrt{n}\left(\frac{1}{n}\sum \ell(\theta; z_i) - \mathbb{E}\ell(\theta; z)\right) \Rightarrow \mathbb{G}(h)$ unif in h

Asymptotics

We use the following notation.

Def RVs $X_n = O_p(1)$ if $\sup_{n\ge 1} \mathbb{P}(\|X_n\| \ge M) \to 0$ as $M \to \infty$

$X_n = o_p(1)$ if $\mathbb{P}(\|X_n\| \ge M) \to 0$ $\forall M > 0$.

We write $X_n = O_p(n)$ if $n^{-1} X_n = O_p(1)$.

ULLN We want to show that $\hat{\theta}_n \to \theta^*$, where $\theta^* = \operatorname{argmin}_{\theta \in \Theta} \mathbb{E}\ell(\theta; Z)$.

We use uniform law of large numbers $\sup_{\theta \in \Theta} \left| \frac{1}{n}\sum \ell(\theta; Z_i) - \mathbb{E}\ell(\theta; Z) \right| \xrightarrow{P} 0$ to prove this.

To simplify notation, let $R(\theta) := \mathbb{E}\ell(\theta; Z)$, $\hat{R}_n(\theta) := \frac{1}{n}\sum \ell(\theta; Z_i)$.

Prop If ULLN holds, and $\hat{\theta}_n$ is s.t. $\hat{R}_n(\hat{\theta}_n) \le \inf_{\theta \in \Theta} \hat{R}_n(\theta) + o_p(1)$, $R(\hat{\theta}_n) \xrightarrow{P} \inf_{\theta \in \Theta} R(\theta)$.

Pf) Let $\theta^* \in \operatorname{argmin}_{\theta \in \Theta} R(\theta)$. $R(\hat{\theta}_n) - R(\theta^*) = R(\hat{\theta}_n) - \hat{R}_n(\hat{\theta}_n) + \hat{R}_n(\hat{\theta}_n) - \hat{R}_n(\theta^*) + \hat{R}_n(\theta^*) - R(\theta^*)$

by hypothesis $\le \sup_{\theta \in \Theta} |\hat{R}_n(\theta) - R(\theta)| + o_p(1) + o_p(1)$ by ULLN $= o_p(1)$. ☒

Cor. Let $R(\cdot)$ be s.t. $\forall \varepsilon > 0$ $\exists \delta > 0$ s.t. $R(\theta) \ge R(\theta^*) + \delta$ whenever $\|\theta - \theta^*\| \ge \varepsilon$.

Under conditions of Prop, $\hat{\theta}_n \xrightarrow{P} \theta^*$.

Pf) $\mathbb{P}(\|\hat{\theta}_n - \theta^*\| \ge \varepsilon) \le \mathbb{P}(R(\hat{\theta}_n) - R(\theta^*) \ge \delta) \to 0$ by Prop. ☒.

Theorem Let H be an envelope function for $\mathcal{H}$: $\forall h \in \mathcal{H}$, $|h| \le H$. Let $\mathbb{E}|H(Z)| < \infty$, and define truncated version of $\mathcal{H}$: $\mathcal{H}_M := \{ \underbrace{h \mathbb{1}\{|h| \le M\}}_{=: h_M} : h \in \mathcal{H} \}$.

If $n^{-1} \log N(\mathcal{H}_M, \|\cdot\|_{L_1(\hat{P}_n)}, \varepsilon) \xrightarrow{P} 0$ for all fixed $\varepsilon > 0$, $M < \infty$,

then $\sup_{h \in \mathcal{H}} \left| \frac{1}{n}\sum h(Z_i) - \mathbb{E}h(Z) \right| \xrightarrow{P} 0$.

Pf) From symmetrization, $\mathbb{E} \sup_{h \in \mathcal{H}} \frac{1}{n}\sum h(Z_i) - \mathbb{E}h(Z) \le 2\mathbb{E}\mathcal{R}_n \mathcal{H}$

$\le 2\mathbb{E} \sup_{h \in \mathcal{H}} \frac{1}{n}\sum \sigma_i (h(Z_i) - h_M(Z_i)) + 2\mathbb{E}\mathcal{R}_n \mathcal{H}_M$

$\le 2\,\mathbb{E}\, H(Z)\mathbb{1}\{H(Z) > M\} + 2\mathbb{E}\mathcal{R}_n \mathcal{H}_M$

Take a ε-cover $\mathcal{H}_{M,\varepsilon}$ of $\mathcal{H}_M$ in $\|\cdot\|_{L^1(\hat{P}_n)}$. $\mathcal{R}_n \mathcal{H}_M \le \mathcal{R}_n \mathcal{H}_{M,\varepsilon} + \varepsilon$

Now, note that since $\sup_{h \in \mathcal{H}} \|h\|_{L_2(\hat{P}_n)} \le M$, Lemma gives

$\sqrt{n}\, \mathcal{R}_n \mathcal{H}_{M,\varepsilon} \le 2M \sqrt{\log N(\mathcal{H}_M, \|\cdot\|_{L_1(\hat{P}_n)}, \varepsilon)}$ $\Rightarrow$ $\mathcal{R}_n \mathcal{H}_{M,\varepsilon} \xrightarrow{P} 0$

Same bound holds for $\mathcal{R}_n(-\mathcal{H}_{M,\varepsilon})$.

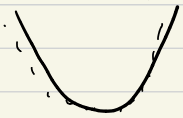
So $\mathbb{E} \sup_{h \in \mathcal{H}} \left| \frac{1}{n}\sum h(Z_i) - \mathbb{E}h(Z) \right| \le 4\,\mathbb{E}\, H(Z)\mathbb{1}\{H(Z) > M\} + 4\mathbb{E}\mathcal{R}_n \mathcal{H}_{M,\varepsilon} + \varepsilon$

$\Rightarrow \limsup_{n \to \infty} \mathbb{E} \sup_{h \in \mathcal{H}} \left| \frac{1}{n}\sum h(Z_i) - \mathbb{E}h(Z) \right| \le 4\,\mathbb{E}\, H(Z)\mathbb{1}\{H(Z) > M\} + \varepsilon$ $\forall \varepsilon > 0$, $M < \infty$.

Taking $M \to \infty$, $\varepsilon \downarrow 0$, we get the result. ☒.

Rates of convergence

We now characterize the rate of convergence for $\hat{\theta}_n \to \theta^*$.

Intuition  $R(\theta)$ If curvature of R is higher that perturbations $\hat{R}_n - R$, then we're good.

$$\hat{R}_n(\theta) - \hat{R}_n(\theta^*) = \underbrace{\hat{R}_n(\theta) - R(\theta) - (\hat{R}_n(\theta^*) - R(\theta^*))}_{=:\Delta_n(\theta) \text{ Fluctuation}} + \underbrace{R(\theta) - R(\theta^*)}_{\text{Growth}}$$

We call Δ_n the localized process.

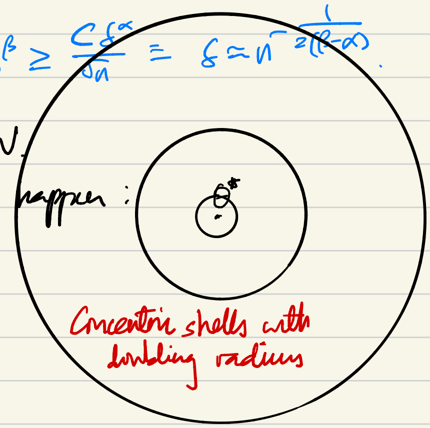
Def The modulus of continuity around θ^* is $W_n(\delta) := \sup_{\|\theta-\theta^*\|\le\delta} |\Delta_n(\theta)|$.

Theorem Let $\hat{\theta}_n \in \operatorname{argmin}_{\theta\in\Theta} \hat{R}_n(\theta)$, and assume $\hat{\theta}_n \xrightarrow{P} \theta^*$. We assume W_n is small compared to curvature.

Fluctuation $\mathbb{E}\, W_n(\delta) \le \frac{C}{\sqrt{n}} \delta^\alpha$ for some $M < \infty$, $\alpha > 0$.

Growth $\exists \beta \ge 1$, $\lambda > 0$, $\varepsilon > 0$ s.t. $R(\theta) \ge R(\theta^*) + \lambda \|\theta - \theta^*\|^\beta$ $\forall \theta$ s.t. $\|\theta - \theta^*\| \le \varepsilon$.

Then, $\|\hat{\theta}_n - \theta^*\| = O_p(n^{-\frac{1}{2(\beta-\alpha)}})$ if $\beta > \alpha$. Intuition: $\lambda\delta^\beta \ge \frac{C\delta^\alpha}{\sqrt{n}} \Rightarrow \delta \simeq n^{-\frac{1}{2(\beta-\alpha)}}$.



Pf) We use a peeling argument. Let $r_n := n^{\frac{1}{2(\beta-\alpha)}}$, and fix $M \in \mathbb{N}$.
If $r_n \|\hat{\theta}_n - \theta^*\| \ge 2^M$ then at least one of the following must happen:

- $\|\hat{\theta}_n - \theta^*\| > \varepsilon$ so growth condition doesn't apply
- $\exists_j$ s.t. $2^{j-1} < r_n \|\hat{\theta}_n - \theta^*\| \le 2^j$ cf. $2^j \le r_n \varepsilon$

  ↳ In this case, $\Delta_n(\hat{\theta}_n) = \underbrace{\hat{R}_n(\hat{\theta}_n) - \hat{R}_n(\theta^*)}_{\le 0} - \underbrace{(R(\hat{\theta}_n) - R(\theta^*))}_{\ge 0}$ satisfies

$$W_n(r_n^{-1}2^j) \ge |\Delta_n(\hat{\theta}_n)| \ge R(\hat{\theta}_n) - R(\theta^*) \ge \lambda \|\hat{\theta}_n - \theta^*\|^\beta \ge \lambda (r_n^{-1} \cdot 2^{j-1})^\beta.$$

So taking a union bound gives

$$\begin{aligned}
\mathbb{P}(r_n \|\hat{\theta}_n - \theta^*\| \ge 2^M) &\le \sum_{j\ge M,\, 2^j \le r_n\varepsilon} \mathbb{P}(r_n \|\hat{\theta}_n - \theta^*\| \in [2^{j-1}, 2^j]) + \mathbb{P}(\|\hat{\theta}_n - \theta^*\| \ge \varepsilon) \\
&\le \sum_{j\ge M,\, 2^j \le r_n\varepsilon} \mathbb{P}(W_n(r_n^{-1}2^j) \ge \lambda \cdot (r_n^{-1}2^{j-1})^\beta) + \mathbb{P}(\|\hat{\theta}_n - \theta^*\| \ge \varepsilon) \\
&\le \sum_{j\ge M,\, 2^j \le r_n\varepsilon} \frac{1}{\lambda (r_n^{-1}2^{j-1})^\beta} \mathbb{E}\, W_n(r_n^{-1}2^j) + \mathbb{P}(\|\hat{\theta}_n - \theta^*\| \ge \varepsilon) \\
&\le \frac{C}{\lambda\sqrt{n}} 2^\beta \cdot r_n^{\beta-\alpha} \sum_{j\ge M,\, 2^j \le r_n\varepsilon} \frac{1}{2^{j\cdot(\beta-\alpha)}} + \mathbb{P}(\|\hat{\theta}_n - \theta^*\| \ge \varepsilon) \\
&= \underbrace{\frac{C}{\lambda} \cdot 2^\beta \sum_{j\ge M} \frac{1}{2^{j(\beta-\alpha)}}}_{\to 0 \text{ as } M\to\infty} + \underbrace{\mathbb{P}(\|\hat{\theta}_n - \theta^*\| \ge \varepsilon)}_{\to 0 \text{ as } n\to\infty}
\end{aligned}$$

We've shown that $r_n \|\hat{\theta}_n - \theta^*\| = O_p(1)$. ☒.

Example Let $\theta \mapsto \ell(\theta; z)$ be C^3 $\forall z$, $\theta \mapsto \ell(\theta, z)$ is $L(z)$-Lipschitz $\forall z$, with $\mathbb{E} L(z)^2 < \infty$.
Assume $\nabla^2 R(\theta^*) \succ 0$. From Taylor expansion,

$$\begin{aligned}
R(\theta) &= R(\theta^*) + \nabla R(\theta^*)^T(\theta - \theta^*) + \tfrac{1}{2}(\theta-\theta^*)^T \nabla^2 R(\theta^*)(\theta - \theta^*) + O(\|\theta - \theta^*\|^3) \\
&\ge R(\theta^*) + \tfrac{1}{2}(\theta-\theta^*)^T \nabla^2 R(\theta^*)(\theta-\theta^*) + O(\|\theta-\theta^*\|^3) \\
&\ge R(\theta^*) + \tfrac{1}{4}\lambda_{\min}(\nabla^2 R(\theta^*)) \|\theta - \theta^*\|^2 \quad \text{for } \|\theta - \theta^*\| \text{ small enough.}
\end{aligned}$$

So $\theta \mapsto R(\theta)$ satisfies growth condition with $\beta = 2$ and $\lambda = \frac{1}{4}\lambda_{min}(\nabla^2 R(\theta^*))$.

To show Fluctuation, we use Dudley's entropy integral. From symmetrization,

$$\mathbb{E}[W_n(\delta) \mid Z_1^n] \le 2\,\mathbb{E}\Big[\sup_{\|\theta-\theta^*\|\le\delta} \Big| \frac{1}{n}\sum (\ell(\theta; z_i) - \ell(\theta^*; z_i))\sigma_i \Big| \,\Big|\, Z_1^n\Big]$$

Recall that ε-covering number of $\{z \mapsto \ell(\theta; z) - \ell(\theta^*; z) : \|\theta-\theta^*\| \le \delta\}$ is bdd by

$$N\Big(\{\theta : \|\theta-\theta^*\|\le\delta\}, \|\cdot\|, \frac{\varepsilon}{\|L\|_{L_2(\hat{P}_n)}}\Big) \le \Big(1 + \frac{\delta \cdot \|L\|_{L_2(\hat{P}_n)}}{\varepsilon}\Big)^d .$$

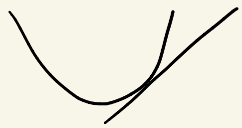
$$\mathbb{E}[W_n(\delta) \mid Z_1^n] \lesssim \frac{1}{\sqrt{n}} \int_0^{\delta\|L\|_{L_2(\hat{P}_n)}} \sqrt{d \log\Big(1 + \frac{\delta\|L\|_{L_2(\hat{P}_n)}}{\varepsilon}\Big)}\, d\varepsilon$$

$$\lesssim \sqrt{\frac{d}{n}} \cdot \delta \cdot \|L\|_{L_2(\hat{P}_n)}$$

Noting that $\mathbb{E}\|L\|_{L_2(\hat{P}_n)} \le \sqrt{\frac{1}{n}\sum \mathbb{E} L(Z_i)^2} = \sqrt{\mathbb{E} L(Z)^2}$, $\quad \mathbb{E} W_n(\delta) \lesssim \sqrt{\frac{d}{n}}\, \delta \cdot \sqrt{\mathbb{E} L(Z)^2}$.

$(\alpha = 1)$

We conclude that $\sqrt{n}\,\|\hat{\theta}_n - \theta^*\| = O_p(1)$.



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 $\forall \theta, \theta' \in \mathbb{R}^d$ $R(\theta') \ge R(\theta) + \nabla R(\theta)^\top(\theta' - \theta)$. ← 1st order approx is a global minorization

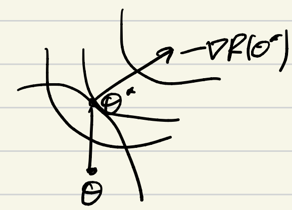
Pf) '⇒' From def of convexity, $R(\theta + t(\theta' - \theta)) \le R(\theta) + t(R(\theta') - R(\theta))$. ≡ $R(\theta') - R(\theta) \ge \frac{1}{t}(R(\theta + t(\theta' - \theta)) - R(\theta))$. Send $t \downarrow 0$.

'⇐' Define $\theta_t = t\theta + (1-t)\theta'$. Combining $R(\theta) \ge R(\theta_t) + \nabla R(\theta_t)^\top(\theta - \theta_t)$, $R(\theta') \ge R(\theta_t) + \nabla R(\theta_t)^\top(\theta' - \theta_t)$,

$tR(\theta) + (1-t)R(\theta') \ge R(\theta_t) + \nabla R(\theta_t)^\top(t\theta + (1-t)\theta' - \theta_t)$ $\forall t \in [0,1]$. ☒

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 $\theta^* = \arg\min_{\theta \in \Theta} R(\theta)$ iff $\nabla R(\theta^*)^\top(\theta - \theta^*) \ge 0$ $\forall \theta \in \Theta$

Pf) '⇐' From prev lemma, $R(\theta) - R(\theta^*) \ge \nabla R(\theta^*)^\top(\theta - \theta^*) \ge 0$ $\forall \theta \in \Theta$.

'⇒' $\nabla R(\theta^*)^\top(\theta - \theta^*) = \lim_{t \downarrow 0} \frac{1}{t}(R(\theta^* + t(\theta - \theta^*)) - R(\theta^*)) \ge 0$ $\forall \theta \in \Theta$. ☒

Cor Let Θ be a closed convex set in $\mathbb{R}^d$. Define the projection operator $\Pi_\Theta(\theta) := \arg\min_{\theta' \in \Theta} \|\theta - \theta'\|_2$.

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 $\min_{\theta' \in \Theta} \|\theta - \theta'\|_2^2$,

$$0 \le (\Pi_\Theta(\theta) - \theta)^\top(\theta' - \Pi_\Theta(\theta)) = (\Pi_\Theta(\theta) - \theta' + \theta' - \theta)^\top(\theta' - \Pi_\Theta(\theta)) = -\|\theta' - \Pi_\Theta(\theta)\|_2^2 + (\theta' - \theta)^\top(\theta' - \Pi_\Theta(\theta)).$$

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 $\mathbb{E}, \nabla_\theta$ 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 $\theta^* \in \operatorname{argmin}_{\theta \in \Theta} R(\theta) > -\infty$ 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.