Lecture 14: Smooth optimization

So far:

* Setting $\hat{R} = \mathcal{L} \circ H$ where
  $\mathcal{L}$ convex & Lipschitz
  H Lipschitz (windof) & "linearizable".

$$\langle \partial(\mathcal{L}\circ H)(w_s), z - w_s \rangle$$

$$(\partial H)\,\partial \mathcal{L}$$

$$\langle \partial \mathcal{L}(H(w_s)), \partial H(w_s)^T (z - w_s) \rangle$$

$$= \langle \partial \mathcal{L}(H(w_s)), H(z) - H(w_s) \rangle$$

$$+ \langle \partial \mathcal{L}(H(w_s)), \underbrace{H(w_s) - H(z) + \partial H(w_s)^T (z - w_s)}_{\text{Small near initialization}} \rangle$$

This time: $\mathcal{L} \circ H$ is smooth, meaning

$$\| \partial(\mathcal{L}\circ H)(a) - \partial(\mathcal{L}\circ H)(b) \| \le \beta \|a - b\|.$$

After this, one more near initialization setting: $\mathcal{L}$ strongly convex

Why smooth:
* last iterate "for free"
* guaranteed improvement in each iteration
* minimizes what we got from gradient flow
* gives local optimality guarantees with no further assumptions

$$\underbrace{\| \partial \hat{R}(a) - \partial \hat{R}(b) \| \le \beta \|a - b\|}_{\text{classical opt}} \Rightarrow \underbrace{\left| \hat{R}(a) - \hat{R}(b) - \langle \nabla \hat{R}(b), b - a \rangle \right| \le \frac{\beta}{2} \|a - b\|^2}_{\text{only use this version in proofs}}$$

if $\hat{R} = \mathcal{L} \circ H$, then

$$\| \partial(\mathcal{L}\circ H)(a) - \partial(\mathcal{L}\circ H)(b) \| = \| \partial H(a) \partial \mathcal{L}(H(a)) - \partial H(b)\, \partial \mathcal{L}(H(b)) \|$$

$$= \| \partial H(a) \partial \mathcal{L}(H(a)) - \partial H(a)\, \partial \mathcal{L}(H(b)) + \partial H(a)\, \partial \mathcal{L}(H(b)) - \partial H(b) \partial \mathcal{L}(H(b)) \|$$

$$\le \| \partial H(a) \left[\partial \mathcal{L}(H(a)) - \partial \mathcal{L}(H(b)) \right] \| + \| \left[\partial H(a) - \partial H(b) \right] \partial \mathcal{L}(H(b)) \|$$

$$\le \underbrace{\| \partial H(a) \|_{2\to 2}}_{\text{hard to satisfy (for now)}} \underbrace{\| \partial \mathcal{L}(H(a)) - \partial \mathcal{L}(H(b)) \|}_{\text{smooth}} + \underbrace{\| \partial \mathcal{L}(H(b)) \|_2}_{\text{smooth}} \cdot \underbrace{\| \partial H(a) - \partial H(b) \|_{2\to 2}}_{\text{smooth}}$$

$$\|Ax\| \le \|A\|_{2\to 2} \|x\|, \quad \text{where} \quad \|A\|_{2\to 2} = \max_{\|v\|_2 \le 1} \|Av\|_2$$

$$= \max_{v \neq 0} \frac{\|Av\|_2}{\|v\|_2}$$

$$\|A\|_{p\to q} = \max_{v \neq 0} \frac{\|Av\|_q}{\|v\|_p}$$

Consequences to GD: ① local optimality, ② last iterate guarantee when convex.

① Local optimality (without projection)

Recall GD: $w_{s+1} := w_s - \eta \nabla \hat{R}(w_s)$

By smoothness: $\hat{R}(w_{s+1}) = \hat{R}(w_s - \eta \nabla \hat{R}(w_s)) \le \hat{R}(w_s) + \langle \nabla \hat{R}(w_s), w_s - \eta \nabla \hat{R}(w_s) - w_s \rangle + \frac{\beta}{2} \| w_s - (w_s - \eta \nabla \hat{R}(w_s)) \|^2$

$= \hat{R}(w_s) - \eta \left(1 - \frac{\eta\beta}{2}\right) \| \nabla \hat{R}(w_s) \|^2.$

Note $\hat{R}(w_{s+1}) \le \hat{R}(w_s)$ when $\eta \le \frac{2}{\beta}$

Theorem Suppose $\hat{R}$ is β-smooth, not necessarily convex.

① If $\eta \in [0, \frac{2}{\beta}] \Rightarrow \hat{R}(w_s)$ is nonincreasing

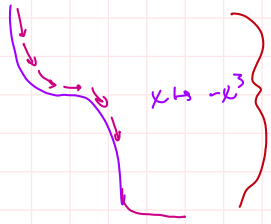
② If $\eta \in [0, \frac{2}{\beta}] \Rightarrow \min_{s \le t} \| \nabla \hat{R}(w_s) \|^2 \le \frac{1}{t} \sum_{s \le t} \| \nabla \hat{R}(w_s) \|^2$ (A)

$\overset{(A)}{\le} \frac{1}{t} \left(\frac{2}{\eta(2-\eta\beta)} \right) \left(\hat{R}(w_0) - \hat{R}(w_t) \right)$

$\overset{(B)}{\le} \frac{1}{t} \left(\frac{2}{\eta(2-\eta\beta)} \right) \left(\hat{R}(w_0) - \inf_{v \in \mathbb{R}^p} \hat{R}(v) \right).$

[Note: if $\eta = \frac{1}{\beta}$: $\min_{s \le t} \| \nabla \hat{R}(w_s) \|^2 \le \frac{2\beta}{t} \left(\hat{R}(w_0) - \hat{R}(w_t) \right) \le \frac{2\beta}{t} \left(\hat{R}(w_0) - \inf_{v \in \mathbb{R}^p} \hat{R}(v) \right)$ — e.g., ≥ 0]

Remark (interpretation). $\exists s \le t$ s.t. $\| \nabla \hat{R}(w_s) \|^2 = O(\frac{1}{t})$



might bounce out of local opt (gradient norm sequence need not be monotone)

$\hat{R}(w_{s+1})$

$\le \hat{R}(w_s) - \eta \left(1 - \frac{\eta\beta}{2}\right) \| \nabla \hat{R}(w_s) \|^2.$

2

Proof of (A). From before:

$\frac{\eta}{2}(2 - \eta\beta) \| \nabla \hat{R}(w_s) \|^2 \le \hat{R}(w_s) - \hat{R}(w_{s+1})$

→ apply $\frac{1}{t} \sum_{s \le t} (\cdot)$

$\frac{\eta}{2}(2 - \eta\beta) \frac{1}{t} \sum_{s \le t} \| \nabla \hat{R}(w_s) \|^2 \le \hat{R}(w_0) - \hat{R}(w_t)$

$\Rightarrow \frac{1}{t} \sum_{s \le t} \| \nabla \hat{R}(w_s) \|^2 \le \frac{2}{\eta(2-\eta\beta)} \left(\hat{R}(w_0) - \hat{R}(w_t) \right).$

Theorem. Suppose $\hat{R}$ is convex and β-smooth, & $\eta \le 1/\beta$.

$\forall z \in \mathbb{R}^d$,

$$\boxed{\frac{1}{2t\eta}\|w_t - z\|^2 + \hat{R}(w_t)} \le \frac{1}{2t\eta}\|w_t - z\|^2 + \frac{1}{t}\sum_{s=1}^{t}\hat{R}(w_s) \le \boxed{\frac{1}{2t\eta}\|w_0 - z\|^2 + \hat{R}(z)}.$$

Proof. Recall

$$\|w_t - z\|^2 + 2\eta\sum_{s\le t}\hat{R}(w_s) \le \|w_0 - z\|^2 + 2\eta t\hat{R}(z) + \sum_{s\le t}\eta^2\|\nabla\hat{R}(w_s)\|^2 \qquad \text{(MD lemma)}$$

$$\sum_{s\le t}\|\nabla\hat{R}(w_s)\|^2 \le \frac{2}{\eta(2-\eta\beta)}\left(\hat{R}(w_0) - \hat{R}(w_t)\right) \qquad \text{(smoothness guaranteed)}$$

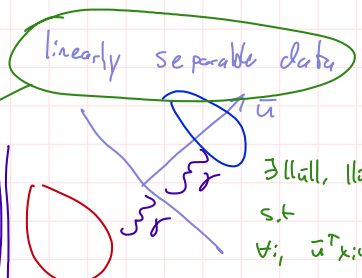
$$\le \frac{2}{\eta}\left(\hat{R}(w_0) - \hat{R}(w_t)\right)$$

combine: $\|w_t - z\|^2 + 2\eta\sum_{s\le t}\hat{R}(w_s) \le \|w_0 - z\|^2 + 2\eta t\hat{R}(z) + \eta^2\left[\frac{2}{\eta}\left(\hat{R}(w_0) - \hat{R}(w_t)\right)\right]$ (2η)

$$\Rightarrow \|w_0 - z\|^2 + 2\eta\sum_{s=1}^{t}\hat{R}(w_s) \le \|w_0 - z\|^2 + 2\eta t\hat{R}(z).$$

Then divide by $2\eta t$.

Example. logistic regression $\ell(z) = \ln(1+\exp(-z))$.



predictions $x \mapsto w^T x$.

Define $\gamma := \min_i \bar{u}^T x_i y_i > 0$, and $z := \bar{u}\frac{\ln t}{\gamma}$,

($\hat{R}$ has no minimum!)

whereby $\hat{R}(z) = \frac{1}{n}\sum_i \ln\left(1+\exp(-y_i x_i^T z)\right)$

$\le \frac{1}{n}\sum_i \exp(-y_i x_i^T z)$

$= \frac{1}{n}\sum_i \exp\left(\underbrace{-y_i x_i^T \bar{u}}_{\le -\gamma}\frac{\ln t}{\gamma}\right) \le \frac{1}{n}\sum_i \exp(-\ln t) = \frac{1}{t}.$

(Remark)
2012-2021: DL gets zero training error (aka interpolation)
2022~: large language model training; not zero training error.

By extremely cute theorem with step size $\eta = 1$, $w_0 = 0$,

$$\frac{1}{2t}\|w_t - z\|^2 + \hat{R}(w_t) \le \frac{1}{2t}\left\|\frac{\ln t}{\gamma}\right\|^2 + \hat{R}(z)$$

$$\le \frac{1}{t}\left(\frac{(\ln t)^2}{2\gamma^2} + 1\right).$$