Lecture 15: smooth & strongly convex optimization

Optimization near initialization so far

	rates	setting
Lipschitz & convex	$1/\sqrt{t}$	ReLU & smooth activation
	$1/t$	"realizable" & $\lvert\ell'\rvert \le \ell$ (e.g., logistic) (nonconvex!)
Smooth	min over $s \le t$ of $\lVert\nabla\hat{R}\rVert^2 \le O(\frac{1}{t})$	$R = 2 \cdot H$ smooth
Smooth & convex	$1/t$	$\hat{R} = 2 \cdot H$ smooth
smooth & strongly convex	$\exp(-\Theta(t))$	$\hat{R} = 2 \cdot H$ smooth

Recall from last time as a consequence of β-smoothness of $\hat{R}$

$$\|\partial(\ell\circ H)(w) - \partial(\ell\circ H)(v)\| \le \beta\|w-v\|$$

$\Rightarrow$ if $\eta \le \frac{1}{\beta}$, then $\hat{R}(w_{s+1}) \le \hat{R}(w_s - \eta\,\partial\hat{R}(w_s))$

$$\le \hat{R}(w_s) - \frac{\eta}{2}\|\partial\hat{R}(w_s)\|^2$$

$\Rightarrow$

- recurse $\Rightarrow$ $\min_{s\le t}\|\partial\hat{R}(w_s)\|^2 \le \frac{1}{t}\sum_{s\le t}\|\partial\hat{R}(w_s)\|^2 \le \frac{2}{t\eta}\left(\hat{R}(w_0) - \hat{R}(w_t)\right)$
  $\le \frac{2}{t\eta}\left(\hat{R}(w_0) - \inf\hat{R}(w_t)\right)$
- recurse & "main lemma" $\Rightarrow$ $\|w_t - z\|^2 + 2t\eta\,\hat{R}(w_t) \le \|w_0 - z\|^2 + 2t\eta\,\hat{R}(z)$.

Remark. today we'll get $\hat{R}(w_t) \le \hat{R}(w_0)\exp(-\sigma(t))$

$\Rightarrow$ $\|w_t - z\|^2 \le \|w_0 - z\|^2\exp(-\sigma(t))$ "contraction towards global optimum".

(common setting).

Process: $\hat{R}(w_{s+1}) \le \hat{R}(w_s) - \frac{\eta}{2}\|\partial\hat{R}(w_s)\|^2 = \hat{R}(w_s) - \frac{\eta}{2}\|\partial H\,\partial\ell(Hw_s)\|^2$

$$\le \hat{R}(w_s) - \frac{\eta}{2}\|\partial H\|_2^2\,\|\partial\ell(Hw_s)\|_2^2$$

$$\le \hat{R}(w_s) - \frac{\eta}{2}\sigma_n(\partial H)^2\,\|\partial\ell(Hw_s)\|_2^2$$

$\sigma_n(\partial H)^2 = \lambda_{\min}(\partial H^T\partial H)$, $\partial H^T\partial H \in \mathbb{R}^{n\times n}$ "kernel gram matrix"

$$(\partial H^T\partial H)_{ij} = \partial F(x_i;w_s)^T\,\partial F(x_j;w_s)$$

not w_0!

$\|\partial\ell(Hw_s)\|^2 \ge 2\lambda(\hat{R}(w_s) - \ell(\bar{y}))$ where $\bar{y} = \operatorname{argmin}_{\hat{y}}\ell(\hat{y})$

λ = strong convexity constant for ℓ

Looking ahead:

$$\hat{R}(w_{s+1}) - \ell(\bar{y}) \le \hat{R}(w_s) - \ell(\bar{y}) - \frac{\eta}{2}\sigma_n(\partial H)^2\left[2\lambda(\hat{R}(w_s) - \ell(\bar{y}))\right]$$

$$= \left[\hat{R}(w_s) - \ell(\bar{y})\right]\left(1 - \eta\lambda\sigma_n(\partial H)^2\right)$$

$$\le \left[\hat{R}(w_s) - \ell(\bar{y})\right]\exp\left(-\eta\lambda\,\sigma_{\le s}^2\right)$$

$\sigma_{\le s} := \min_{r\le s}\sigma_n(\partial H(w_r))$

$\Rightarrow$ $\hat{R}(w_t) - \ell(\bar{y}) \le \left[\hat{R}(w_0) - \ell(\bar{y})\right]\exp\left(-t\eta\lambda\,\sigma_{\le t}^2\right)$.

near initialization

"classical term"

Remark. ① contraction $w_t \to \bar{w}$ (inverse condition number)

① MD lemma / master lemma

② apply strong convexity of ℓ at end

② $t\eta\lambda \le \frac{t\lambda}{\beta}$ is classical; $\sigma_{\le t}^2$ is to handle NTK

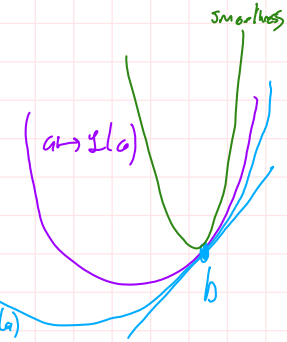
③

Strong convexity

Definition. $\mathcal{L}$ is λ-strongly-convex (λ-sc) if $\forall a, b$

$$\mathcal{L}(b) \geq \mathcal{L}(a) + \langle \partial \mathcal{L}(b), b-a \rangle + \frac{\lambda}{2}\|b-a\|^2$$

first-order convexity def.



Remark. If $\mathcal{L}$ is twice-diff,

$\mathcal{L}$ λ s.c. $\iff$ $\nabla^2 \mathcal{L} \geq \lambda I$.

i.e., $Q_b(a) \leq \mathcal{L}(a)$ $\forall a$,

& $Q_b(b) = \mathcal{L}(b)$

& $Q_b(b) + \langle \partial Q_b(b), b-a \rangle = \mathcal{L}(b) + \langle \partial \mathcal{L}(b), b-a \rangle$.

Example. ① If $\mathcal{L}$ is convex, then $\mathcal{L} + \frac{\lambda}{2}\|\cdot\|^2$ is λ-sc.

② $\hat{R}(w) = \frac{1}{2}\|Xw - y\|^2$ is $\sigma_n(X)^2$-sc.

Ⓐ note $\nabla^2 \hat{R} = X^T X$ Ⓑ

$$\frac{1}{2}\|Xa - y\|^2 = \frac{1}{2}\|X(a-b+b) - y\|^2 \quad \langle X^T X(b-a), a-b \rangle$$

$$= \frac{1}{2}\|Xb - y\|^2 + \langle Xb - y, X(a-b) \rangle + \frac{1}{2}\|X(a-b)\|^2$$

all equalities

$$\geq \hat{R}(b) + \langle \nabla \hat{R}(b), a-b \rangle + \frac{\sigma_{\min}(X)^2}{2}\|a-b\|^2.$$

Proposition If $\mathcal{L}$ is λ-sc, $\forall w$ $\|\partial \mathcal{L}(w)\|^2 \geq 2\lambda\left(\mathcal{L}(w) - \inf_{v \in \mathbb{R}^n} \mathcal{L}(v)\right)$.

Proof. Recall $Q_b(a) := \mathcal{L}(b) + \langle \partial \mathcal{L}(b), a-b \rangle + \frac{\lambda}{2}\|b-a\|^2 \leq \mathcal{L}(a)$.

Then $\inf_v \mathcal{L}(v) \geq \inf_v Q_w(v)$ optimized at $a = b - \frac{1}{\lambda}\partial \mathcal{L}(b)$

$$= \mathcal{L}(w) - \langle \partial \mathcal{L}(w), \tfrac{1}{\lambda}\partial \mathcal{L}(w) \rangle + \frac{\lambda}{2}\|\tfrac{1}{\lambda}\partial \mathcal{L}(w)\|^2$$

$$= \mathcal{L}(w) - \frac{1}{2\lambda}\|\partial \mathcal{L}(w)\|^2$$

$\partial \mathcal{L}(b) + \lambda(b-a) = 0 \Rightarrow a = b - \frac{1}{\lambda}\partial \mathcal{L}(b)$