Ch 13 part 2 Analysis and Stability

Thursday, March 19, 2020 4:10 PM

(following Shalev-Shwartz and Ben-David)

Analysis of II term

Recall II $= \mathbb{E}_S \hat{L}_S(A(S))$

$$A(S) = \operatorname*{argmin}_{w \in \mathcal{H}} \hat{L}_S(w) + \lambda \|w\|^2$$

The larger λ is,

the larger II becomes: this represents more bias or underfitting

So unlike I term,

now want λ small

To analyze, note that since $\lambda \|w\|^2 \geq 0$,

$$(\forall w) \quad \hat{L}_S(A(S)) \leq \hat{L}_S(A(S)) + \lambda \|A(S)\|^2$$
$$\leq \hat{L}_S(w) + \lambda \|w\|^2$$

hence

$$\text{II} := \mathbb{E}_S \hat{L}_S(A(S)) \leq \mathbb{E}_S \hat{L}_S(w) + \lambda \|w\|^2 \quad (\forall w \in \mathcal{H})$$
$$= L_D(w) + \lambda \|w\|^2$$

If w fixed, then draw $S \overset{iid}{\sim} D^m$,

$$\mathbb{E}_S \hat{L}_S(w) = L_D(w)$$

but $\mathbb{E}_S \hat{L}_S(A(S)) \neq L_D(A(S))$

$\mathbb{E}_S L_S(A(S)) \neq L_D(A(S))$

So we can choose the best (or at least good) w to minimize this bound (such a w is an "oracle")
(and can also choose λ as we wish...
though in practice choose λ via cross-validation or similar)

For example, and using a bound on I from last lecture,

Corollary 13.9 Let $(\mathcal{H}, Z, \ell)$ be a convex, ρ-Lipschitz, B-bounded (i.e. $\forall w \in \mathcal{H}, \|w\| \le B$), and $S \overset{iid}{\sim} \mathcal{D}^m$, then set $\lambda = \sqrt{\frac{2\rho^2}{B^2 m}}$ then if A is RLM w/ $R(w) = \lambda \|w\|^2$,

$$\mathbb{E}_S L_D(A(S)) \le \min_{w \in \mathcal{H}} L_D(w) + \underbrace{\rho B \sqrt{\frac{8}{m}}}_{\le \varepsilon \text{ if } m \ge \frac{8\rho^2 B^2}{\varepsilon^2}}$$

proof sketch:

$$\mathbb{E} L_D(A(S)) = \mathbb{E}\underbrace{\left(L_D(A(S)) - \hat{L}_S(A(S))\right)}_{\mathrm{I}} + \underbrace{\mathbb{E}\hat{L}_S(A(S))}_{\mathrm{II}}$$

$$\le \frac{2\rho^2}{\lambda m} + L_D(w) + \underbrace{\lambda\|w\|^2}_{\le \lambda B^2} \quad \forall w \in \mathcal{H}$$

$$= L_D(w) + \rho B \sqrt{8/m} \qquad \left(\text{since } \frac{2\rho^2 + \lambda^2 m B}{\lambda m} = \frac{4\rho^2}{\lambda m} = \sqrt{\frac{16 B^2 \rho^4 m}{2\rho^2 m}}\right)$$

and $\forall w \in \mathcal{H}$ so take min. $\square$

A similar bound, w/ different assumptions
(if X bounded, these are stronger assumptions, up to constants)

Corollary 13.11 Let $(\mathcal{H}, Z, \ell)$ be a convex, β-smooth, B-bounded, and $(\forall z \in Z)\ \ell(0,z) \le C \le 1$ then $(\forall \varepsilon > 0)(\forall \text{ distr } D)$, if $m \geq \frac{150\beta B^2}{\varepsilon^2}$ and $\lambda = \frac{\varepsilon}{3B^2}$, then

$$\mathbb{E}\, L_D(A(S)) \le \left(\min_{w \in \mathcal{H}} L_D(w)\right) + (1^1/_3) \cdot \varepsilon$$

proof:

$$\mathbb{E}\, L_D(A(S)) = \mathbb{E}\, \hat{L}_S(A(S)) + \mathrm{I}$$

bdd from prev. lecture

$$\le \frac{48\beta}{\lambda m} \mathbb{E}\, \hat{L}_S(A(S))$$

$$\left(\text{and } \mathbb{E}\, \hat{L}_S(A(S)) \le L_D(w) + \lambda \|w\|^2 \quad \forall w \in \mathcal{H}\right)$$

$$\le \left(L_D(w) + \lambda\|w\|^2\right) + \frac{48\beta}{\lambda m}\left(L_D(\bar{w}) + \lambda \|\bar{w}\|^2\right)$$

choose $w \in \operatorname{argmin} L_D(w)$ ... choose $\bar{w} = 0$

$$\le \min_{w \in \mathcal{H}} L_D(w) + \lambda B^2 + \frac{48\beta}{\lambda m}(C + 0)$$

$\lambda m \geq \frac{50\beta}{\varepsilon}$

$$\le \min_{w \in \mathcal{H}} L_D(w) + \underbrace{\frac{\varepsilon}{3} + \frac{48}{50}\varepsilon}_{\le 1^1/_3\, \varepsilon} \qquad \square$$

History of regularization + stability

Andrey Tikhonov 1906–1993 (sp. Tychonoff in topology)

ubiquitous in math

Stability for learning: Roger, Wagner '78 for k-NN

Bagging (bootstrap aggregating) by Breiman '96

is used to increase stability of unstable algo

Boosting create strong learner from weak ones

Bagging create stable learner from unstable ones

N...

Modern ex.

"Train faster, generalize better: Stability of stochastic gradient descent" (Hardt, Recht, Singer ICML '16)

Mohri et al.'s perspective

ch 14 "Algorithmic Stability"

Emphasize benefit of this analysis (vs uniform conv.) is tailored to algorithm

defines 'uniform stability' (can bound I)

and uses McDiarmid's ineq to get results (Thm 14.2)

Thm 14.2 loss $\le M$, algo A is β-uniformly stable then w.p. $\ge 1-\delta$,

$$L_D(A(S)) \le \hat{L}_S(A(S)) + \beta + (2m\beta + M)\sqrt{\frac{\log(1/\delta)}{2m}}$$

and Prop. 14.4 shows if $\|x\|^2 \le r^2 \; \forall x \in X$,

ℓ is convex and ρ-admissable (similar to Lipschitz)

then RLM w/ $R(w) = \lambda\|w\|^2$ is β-stable, $\beta \le \frac{\rho^2 r^2}{m\lambda}$

Trickier to see λ tradeoff

$\lambda \to \infty \Rightarrow \beta \to 0$ (good)

but $\hat{L}_S(A(S))$ not as close to $\min_{w \in \mathcal{H}} \hat{L}_S(w)$ (=ERM)