# Lecture 8 - scribbles - calibration

Friday, February 16, 2018
14:30

today: finish calibrated surrogate loss story

recall: $\ell(x, y; s) \triangleq \frac{1}{2k} \sum_{\tilde{y}} \left( s_x(\tilde{y}) - (-\ell(y, \tilde{y})) \right)^2$

$\ell_q(s) = \| s - (-L q) \|^2 + \text{cst.}$

$\ell_q(s) \triangleq \mathbb{E}_q \, \ell(x, y; s)$ $\quad s \in \mathbb{R}^K$

$\ell_q(s) - \min_{s \in \mathbb{R}^K} \ell_q(\tilde{s})$

$= \frac{1}{2k} \| s_x(\cdot) - (-\ell_{q_x}(\cdot)) \|_2^2$ $\quad \tilde{y}$

$\in \mathbb{R}^K$ $\quad \vec{\ell}_{q_x} = \sum_y q(y|x) \, \ell(y, \cdot)$

Let $\overleftrightarrow{L}$ be a $K \times K$ matrix where $L_{\tilde{y}, y} = \ell(y, \tilde{y})$

$\vec{\ell}_{q_x} = \overleftrightarrow{L} \vec{q}_x$

$s^* = -\overleftrightarrow{L} \vec{q}_x \in \text{span}(\overleftrightarrow{L})$ i.e. $\sum_y \alpha_y L(\cdot, y)$

column space

to get consistency, it is sufficient to only consider $s \in \text{span}(\overleftrightarrow{L})$

or that $s \in \text{span}(F) \supseteq \text{span}(\overleftrightarrow{L})$

restriction on scores $\quad F \in \mathbb{R}^{K \times r}$

$F$ can be chosen cleverly depending on $\overleftrightarrow{L}$

if $\text{span}(F) \supseteq \text{span}(L)$,

then have $\ell_q(\theta) - \min_{\theta \in \mathbb{R}^r} \ell_q(\theta) = \frac{1}{2k} \| \overleftrightarrow{F} \theta - (-\vec{\ell}_q) \|_2^2$

then have $f_q(\theta) - \min_{\theta \in \mathbb{R}^r} f_q(\theta) = \frac{1}{2k} \| F\theta - (-L_q) \|_2^2$

$s = F\theta$

thm. 7: $H_{\ell_{squared}, \ell, F}(\epsilon) \geq \frac{\epsilon^2}{2k \max_{i \neq j} \|P_F \Delta_{ij}\|_2^2} \geq \frac{\epsilon^2}{4k}$ ← lower bound → easiness result; this side is bad; $|\mathcal{Y}|$ e.g. $10^{23}$

if $\text{span}(F) \supseteq \text{span}(L)$

where $P_F$ is orthogonal projection on $\text{span}(F)$ $P_F = F(F^TF)^\dagger F^T$

$\Delta_{ij} \triangleq e_i - e_j \in \mathbb{R}^k$


---

in the paper, show that for 0-1 loss, $H(\epsilon) = \frac{\epsilon^2}{4k}$

thm. 8: if $\text{span}(F) = \mathbb{R}^k$ (ie. no constraints) ← "hardness result"

then $H(\epsilon) \leq \frac{\epsilon^2}{2k}$

ie. for any loss, we need exponential # of samples in the worst case

(caveat → all these are bounds and worst case)

* but for Hamming loss, if add constraint that $s(\tilde{y}) = \sum_{p \in \text{parts}} s_p(\tilde{y}_p)$

over T binary variables, then $H(\epsilon) = \frac{\epsilon^2}{8T}$ } not too big ⇒ we can learn?

---

missing link: going from calibration function to sample complexity

**setup:** let $S(x,\tilde{y})$ be of the form $F\theta(x)$

$\theta(x) \in \mathbb{R}^r$

$\theta_y(\cdot) \in \mathcal{H}$ (RKHS)

$\theta$ = optimization variable

$g(\theta) = \mathbb{E}_{(x,y)\sim p}\, g(x,y,\theta)$

run projected SGD on $g(\theta)$ i.e. $\theta^{(t+1)} \leftarrow \text{Proj}_{\theta^{(t)}}\left[\theta^{(t)} - \gamma \nabla_\theta g(x^{(t)}, y^{(t)}, \theta)\right]$

- Proj: on all ball of radius $D$
- $(x^{(t)}, y^{(t)}) \overset{\text{i.i.d.}}{\sim} p$
- $\nabla_\theta g(x^{(t)}, y^{(t)}, \theta) = F^T \nabla_S g(x^{(t)}, y^{(t)}, S)\, \Phi(x^{(t)})^T$
- $\Phi(x^{(t)})^T$: feature maps for $\mathcal{H}$, $k(x^{(t)}, \cdot)$

**convergence result:** (thm. 5)

if $\|\theta^*\|_{HS} \le D$

if $\mathbb{E}_{(x,y)\sim p} \|\nabla_\theta g(x,y,\theta)\|_{HS} \le M^2$

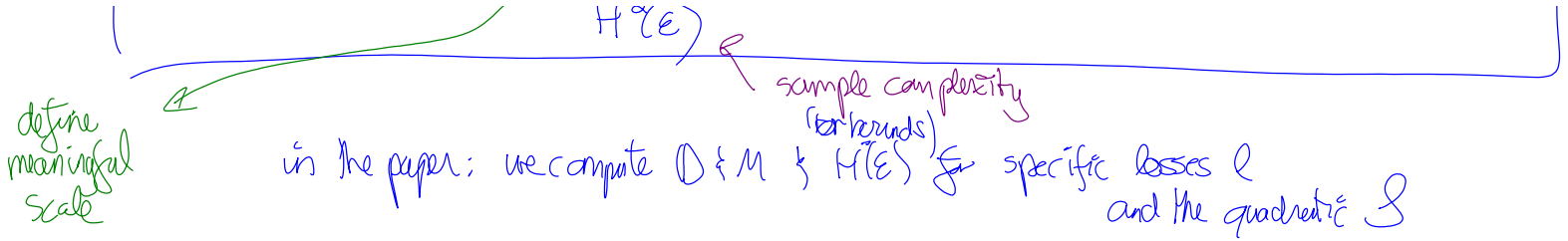
then that averaged SGD with step size $\gamma = \frac{2D}{M\sqrt{n}}$,

$$\mathbb{E}\left[g\left(\underbrace{\frac{1}{n}\sum_{t=1}^{n} \theta^{(t)}}_{\theta^{(n)}}\right)\right] - g(\theta^*) \le \frac{2DM}{\sqrt{n}} \quad \text{(convergence result)}$$

**thm. 6:** learning complexity

if $\theta^*$ which minimizes $L(\theta)$, has $\|\theta^*\|_{HS} \le D$

choosing $n > \frac{4D^2M^2}{\varepsilon^2}$ implies $\mathbb{E}[L(\theta^{(n)})] < L(\theta^*) + \varepsilon$




in the paper: we compute $D$ & $M$ & $H(\varepsilon)$ for specific losses $\ell$ and the quadratic $S$

to get sample complexity

Moral here:

- some losses are harder to learn than others (worse sample complexity)
  [0-1 is difficult in general]
- have linked computation to statistical performance in consistency framework
  → convex surrogate loss)
- could handle dependence on $x$ using RKHS

caveats: • distribution-free result (i.e. worst case over all distributions)

→ still need more theory? (e.g. role of $p(y|x)$?)

or other surrogates?