Comparison between two forms of Cauchy-Schwarz inequality

Miaoyan Wang, July 2, 2020

We have proposed two choices of function classes:

Case 1. bounded features + low rank coefficients

$$\mathcal{F}_1 = \mathcal{F}(r, M, G) = \{f\colon \boldsymbol{X} \mapsto \langle \boldsymbol{B}, \boldsymbol{X}\rangle \mid \boldsymbol{B}, \boldsymbol{X}, \in \mathbb{R}^{d_1\times d_2},\ \text{rank}(\boldsymbol{B}) \leq r, \|\boldsymbol{B}\|_{\text{sp}} \leq M, \|\boldsymbol{X}\|_F \leq G\}.$$

Case 2. unbounded, random features + low rank coefficients

$$\mathcal{F}_2 = \mathcal{F}(r, M) = \{f\colon \boldsymbol{X} \mapsto \langle \boldsymbol{B}, \boldsymbol{X}\rangle \mid \boldsymbol{B}, \boldsymbol{X} \in \mathbb{R}^{d_1\times d_2},\ \text{rank}(\boldsymbol{B}) \leq r, \|\boldsymbol{B}\|_F \leq M, \boldsymbol{X} \sim \mathcal{MN}(\mathbf{0}_{d_1\times d_2}, \boldsymbol{I}, \boldsymbol{I})\}.$$

Question: Can we provide a common approach to obtain sharp bounds for both cases?

Recall that the key step in the Rademacher bound is the Cauchy–Schwarz inequality,

$$\langle \boldsymbol{B}, \boldsymbol{S}_n\rangle \leq \|\boldsymbol{B}\|_p \|\boldsymbol{S}_n\|_q, \quad \text{for any } p, q \geq 0 \text{ satisfying } \frac{1}{p} + \frac{1}{q} = 1,$$

where $\boldsymbol{S}_n = \sum_{i=1}^n \sigma_i \boldsymbol{X}_i$ is a stochastically-weighted sum of feature matrices.

Approach 1 uses $p = q = 2$; i.e., F-norm for both $\boldsymbol{B}$ and $\boldsymbol{S}_n$.

Approach 2 uses $p = 0, q = \infty$; i.e., nuclear norm for $\boldsymbol{B}$ and spectral norm for $\boldsymbol{S}_n$.

Claim 1. *Approach 2 is always no worse than approach 1. In particular, both approaches give the same bounds in Case 1, and Approach 2 gives better bound in Case 2.*

In Case 2, substantially different results are obtained based on Approach 1 vs. 2.

- Applying Approach 1 to Case 2 gives polynomial growth in d:

$$\mathcal{R}_n(\mathcal{F}_2) \leq \frac{\|\boldsymbol{B}\|_F}{\sqrt{n}} \max_i \|\boldsymbol{X}_i\|_F \asymp \mathcal{O}\left(\sqrt{\frac{d_1 d_2}{n}}\right).$$

- Applying Approach 2 to Case 2 gives linear growth in d:

$$\mathcal{R}_n(\mathcal{F}_2) \leq \frac{1}{n}\|\boldsymbol{B}\|_* \mathbb{E}\|\boldsymbol{S}_n\|_{\text{sp}} \asymp \mathcal{O}\left(\sqrt{\frac{r(d_1+d_2)}{n}}\right), \text{ much sharper than Approach 1.}$$

It remains to show that, in Case 1, similar bounds are obtained based on Approaches 1 vs. 2.

- Applying Approach 1 to Case 1:

$$\mathcal{R}_n(\mathcal{F}_1) = \frac{\|\boldsymbol{B}\|_F}{\sqrt{n}} \max_i \|\boldsymbol{X}_i\|_F. \tag{1}$$

- Applying Approach 2 to Case 1:

$$\begin{aligned}\mathcal{R}'_n(\mathcal{F}_1) &\leq \frac{1}{n}\|\boldsymbol{B}\|_* \mathbb{E}\|\boldsymbol{S}_n\|_{\text{sp}} \\ &\leq 2\left(\sqrt{\frac{r}{n}}\log(d_1+d_2) + \sqrt{\log(d_1+d_2)}\right)\frac{\|\boldsymbol{B}\|_F}{\sqrt{n}} \max_i \|\boldsymbol{X}_i\|_{\text{sp}},\end{aligned} \tag{2}$$

  where the expectation is taken over i.i.d. Rademacher sequence $\sigma_i \sim_{\text{i.i.d}} \text{Bernoulli}(\frac{1}{2})$, and the second line comes from the matrix Bernstein inequality (c.f. Lemma 1).

Consider the high-dimensional regime as $n, d_1, d_2 \to \infty$ while holding r fixed. Note that the log term is smaller than any polynomial term, $\log(d_1+d_2) \leq o(d^\alpha)$ for any $\alpha > 0$. Henceforth, the bound (2) is no worse than (1),

$$\mathcal{R}'_n(\mathcal{F}_1) \ll o(d^{0.001})\frac{\|\boldsymbol{B}\|_F}{\sqrt{n}} \max_i \|\boldsymbol{X}_i\|_{\text{sp}} \leq \text{or} \ll \frac{\|\boldsymbol{B}\|_F}{\sqrt{n}} \max_i \|\boldsymbol{X}_i\|_F = \mathcal{R}_n(\mathcal{F}_1).$$

The gap in the last inequality can be substantial, e.g., by a factor of $\mathcal{O}(\sqrt{d})$ when $\boldsymbol{X}_i$ are approximately full rank. As a conclusion, we favor Approach 2 over Approach 1 in both cases.

Lemma 1 (Matrix Bernstein, Theorem 1.6.2 in Ref. [1])**.** *Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be independent, centered random matrices with common dimension $d_1 \times d_1$, and assume that each one is uniformly bounded,*

$$\mathbb{E}\mathbf{Y}_i = \mathbf{0} \quad \textit{and} \quad \|\mathbf{Y}_i\|_{sp} \leq L \quad \textit{for all } i \in [n].$$

Define the sum $\boldsymbol{S}_n = \sum_{i=1}^n \mathbf{Y}_i$, and let $v(\boldsymbol{S}_n)$ denote the matrix variance statistic of the sum:

$$v(\boldsymbol{S}_n) = \max\left\{\|\sum_{i=1}^n \mathbb{E}(\mathbf{Y}_i\mathbf{Y}_i^T)\|_{sp},\ \|\sum_{i=1}^n \mathbb{E}(\mathbf{Y}_i^T\mathbf{Y}_i)\|_{sp}\right\}.$$

Then

$$\mathbb{E}\|\boldsymbol{S}_n\|_{sp} \leq \sqrt{2v(\boldsymbol{S}_n)\log(d_1+d_2)} + \frac{1}{3}L\log(d_1+d_2). \tag{3}$$

Remark 1. In light of matrix Bernstein inequality, we probably do not need the Gaussian random feature assumption in the previous note.

Proof of bound (2)*.* We apply Bernstein inequality to $\boldsymbol{Y}_i = \sigma_i \boldsymbol{X}_i$, where $\boldsymbol{X}_i$ is a deterministic matrix and $\sigma_i \sim_{\text{i.i.d.}} \text{Ber}(1/2)$, for all $i \in [n]$. It ie easy to verify that $\boldsymbol{Y}_i$ are independent, centered random matrix with spectral norm bounded by $L = \max_i \|\boldsymbol{X}_i\|_{\mathrm{sp}}$. Furthermore, the matrix variance statistic

$$\begin{aligned} v(\boldsymbol{S}_n) &= \max\left\{ \|\sum_{i=1}^{n} \mathbb{E}\sigma_i^2(\boldsymbol{X}_i\boldsymbol{X}_i^T)\|_{\mathrm{sp}},\ \|\sum_{i=1}^{n} \mathbb{E}\sigma_i^2(\boldsymbol{X}_i\boldsymbol{X}_i^T)\|_{\mathrm{sp}} \right\} \\ &= \max\left\{ \|\sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{X}_i^T \mathbb{E}\sigma_i^2\|_{\mathrm{sp}},\ \|\sum_{i=1}^{n} \boldsymbol{X}_i\boldsymbol{X}_i^T \mathbb{E}\sigma_i^2\|_{\mathrm{sp}} \right\} \\ &= \max\left\{ \|\sum_{i=1} \boldsymbol{X}_i\boldsymbol{X}_i^T\|_{\mathrm{sp}},\ \|\sum_{i=1} \boldsymbol{X}_i^T\boldsymbol{X}_i\|_{\mathrm{sp}} \right\} \\ &\le n \max_i \|\boldsymbol{X}_i\|_{\mathrm{sp}}^2. \end{aligned} \tag{4}$$

Combining (4) into (3) gives

$$\mathbb{E}\|\boldsymbol{S}_n\|_{\mathrm{sp}} \le 2 \max_i \|\boldsymbol{X}_i\|_{\mathrm{sp}} \left(\sqrt{n \log(d_1 + d_2)} + \log(d_1 + d_2) \right). \tag{5}$$

The final conclusion (2) follows by plugging (5) and $\|\boldsymbol{B}\|_* \le \sqrt{r}\|\boldsymbol{B}\|_F$ into the first line of (2). $\square$

References

[1] Joel A Tropp. An introduction to matrix concentration inequalities. *arXiv preprint arXiv:1501.01571*, 2015.