Rademacher complexity and generalization error

Miaoyan Wang, July 1, 2020

1 Previous results

Define the linear function class

$$\mathcal{F} = \mathcal{F}(r, M) = \{f\colon \boldsymbol{X} \mapsto \langle \boldsymbol{B}, \boldsymbol{X}\rangle \;\big|\; \boldsymbol{B} \in \mathbb{R}^{d_1\times d_2},\ \mathrm{rank}(\boldsymbol{B}) \le r, \|\boldsymbol{B}\|_{\mathrm{sp}} \le M\},$$

where $\|\cdot\|_{\mathrm{sp}}$ denotes the matrix spectral norm.

Assumption 1 (Bounded feature)**.** *Let $\mathcal{X} \subset \mathbb{R}^{d_1\times d_2}$ be the feature space of interest. Assume $\|\boldsymbol{X}\|_F \le G$ for all $\boldsymbol{X} \in \mathcal{X}$, where $G > 0$ is a constant independent of the dimensions d_1, d_2.*

Lemma 1 (Rademacher complexity)**.** *Under Assumption 1, the Rademacher complexity of $\mathcal{F}$ with respect to a set of i.i.d. samples $\{\boldsymbol{X}_i \in \mathcal{X}\colon i = 1, \dots, n\}$ is*

$$\mathcal{R}_n(\mathcal{F}) \le 2MG\sqrt{\frac{r}{n}}.$$

Theorem 1.1. *Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be a L-Lipchitz loss function. Suppose ϕ entrywise dominates the $0/1$ loss, and the Assumption 1 holds. Then, with probability at least $1-\delta$, we have*

$$\mathbb{P}[Y \ne \mathit{sign}\ f(\boldsymbol{X})] \le \frac{1}{n}\sum_{i=1}^{n}\phi(y_i f(\boldsymbol{X}_i)) + 2LMG\sqrt{\frac{r}{n}} + \sqrt{\frac{\log(\frac{1}{\delta})}{2n}}, \quad \textit{for all } f \in \mathcal{F}.$$

2 Implications

Remark 1 (Connection to our estimator)**.** Theorem 1.1 immediately gives the statistical error bound for estimating f when restricted in the class $\mathcal{F}$. Specifically, note that our algorithm uses the 1-Lipchitz hinge loss, $\phi(t) = t_+$. Let $\hat{f} \in \mathcal{F}$ be the empirical risk minimizer (ERM) returned by our algorithm,

$$\hat{f} = \min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n}[y_i f(\boldsymbol{X}_i)]_+, \tag{1}$$

and let $f^* = \min_{f\in\mathcal{F}} \mathbb{E}_{(\boldsymbol{X},Y)}[Yf(\boldsymbol{X})]_+$ be the "best" population risk minimizer when restricted in the class $\mathcal{F}$. (In particular, $\mathrm{sign} f^*$ equals the Bayes classifier if $\mathcal{F}$ is rich enough.)

Corollary 1 (Excess risk)**.** *Under the assumption of Theorem 1.1, with very high probability,*

$$\underbrace{\mathbb{P}[Y \neq sign\ f^*(\boldsymbol{X})] - \mathbb{P}[Y \neq sign\ \hat{f}(\boldsymbol{X})]}_{\text{statistical error for estimating } f} \leq 4MG\sqrt{\frac{r}{n}}. \tag{2}$$

Remark 2. The sample requirement for consistent estimation is $n \gg \mathcal{O}(rM^2G^2)$.

Proof of Corollary 1. The bound (4) follows from the following observation,

$$\begin{aligned}
&\mathbb{P}[Y \neq \text{sign } f^*(\boldsymbol{X})] - \mathbb{P}[Y \neq \text{sign } \hat{f}(\boldsymbol{X})] \\
=&\underbrace{\left\{\mathbb{P}[Y \neq \text{sign } f^*(\boldsymbol{X})] - \frac{1}{n}\sum_{i=1}^n [y_i f^*(\boldsymbol{X}_i)]_+\right\}}_{\text{bounded by Theorem 1.1}} - \underbrace{\left\{\mathbb{P}[Y \neq \text{sign } \hat{f}(\boldsymbol{X})] - \frac{1}{n}\sum_{i=1}^n [y_i \hat{f}(\boldsymbol{X}_i)]_+\right\}}_{\text{bounded by Theorem 1.1}} \\
&+ \underbrace{\frac{1}{n}\sum_{i=1}^n [y_i f^*(\boldsymbol{X}_i)]_+ - \frac{1}{n}\sum_{i=1}^n [y_i \hat{f}(\boldsymbol{X}_i)]_+}_{\geq 0 \text{ by definition of } \hat{f}} \\
\leq& 4MG\sqrt{\frac{r}{n}},
\end{aligned}$$

□

Corollary 2. *Let sign* $f_{Bayes}\colon \mathcal{X} \to \{0,1\}$ *denote the Bayes classifier. Then*

$$\begin{aligned}
&\underbrace{\mathbb{P}[Y \neq sign\ f_{Bayes}(\boldsymbol{X})] - \mathbb{P}[Y \neq sign\ \hat{f}(\boldsymbol{X})]}_{total\ error} \\
\leq&\underbrace{\mathbb{P}[Y \neq sign\ f_{Bayes}(\boldsymbol{X})] - \mathbb{P}[Y \neq sign\ f^*(\boldsymbol{X})]}_{approximation\ error} + \underbrace{\mathbb{P}[Y \neq sign\ f^*(\boldsymbol{X})] - \mathbb{P}[Y \neq sign\ \hat{f}(\boldsymbol{X})]}_{statistical\ error}
\end{aligned}$$

3 Three caveats and remedies

1. The spectral-norm constraint $\|\boldsymbol{B}\|_{\text{sp}} \leq M$ was imposed to the analysis but not to the algorithm. Can we remove this constraint from $\mathcal{F}$?

   Q: Yes; use a different Cauchy-Schwart inequality in the Rademacher complexity bound.

2. In practice, our algorithm returns the penalized ERM $\hat{f}_{\text{pen}} = \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n [y_i f(\boldsymbol{X}_i)]_+ + \lambda\|F\|_F$, not (1). Can we modify the analysis to allow penalized ERM?

   Q: Yes. Note the equivalence between the following two optimizations:

$$\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n [y_i f(\boldsymbol{X}_i)]_+ + \lambda\|F\|_F. \quad \text{v.s} \quad \min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n [y_i f(\boldsymbol{X}_i)]_+, \text{ s.t. } \|F\|_F \leq C.$$

We define the penalized ERM $\hat{f}_{\text{pen}}$ by imposing the F-norm constraint to the class $\mathcal{F}$; i.e,

$$\hat{f}_{\text{pen}} = \underset{f \in \mathcal{F} \cap \{f\colon \|f\|_F \le C\}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} [y_i f(\boldsymbol{X}_i)]_+. \tag{3}$$

Theorem 4.1 gives the excess risk bound for (3).

3. The sample complexity for estimator (1) is $\mathcal{O}(rM^2G^2)$. We are interested in the high-dimensional regime as $d_1, d_2, n \to \infty$ while holding r fixed. Is it reasonable to assume a constant $G = \|\boldsymbol{X}\|_F > 0$ as $d_1, d_2 \to \infty$?

   Q: Depends. Consider the neuroimaging imaging application, where features $\boldsymbol{X} = [\![x_{pq}]\!] \in \mathbb{R}^{d_1 \times d_2}$ are brain images and the size $d_1 \times d_2$ represents the resolution. In the i.i.d. Gaussian random feature model $x_{qp} \sim_{\text{i.i.d.}} N(0,1)$, $G = \|\boldsymbol{X}\|_F = \mathcal{O}(\sqrt{d_1 d_2}) \to \infty$, which is bad. Our remedy for mitigating the growth rate is to use a different norm, $\|\boldsymbol{X}\|_{\text{sp}} = \mathcal{O}(\sqrt{d_1 + d_2}) \ll \mathcal{O}(\sqrt{d_1 d_2})$.

4 New results

Definition 1 (Gaussian feature with bounded variation)**.** The Gaussian matrix feature is defined as

$$\boldsymbol{X} \sim \mathcal{MN}(\boldsymbol{0}_{d_1 \times d_2}, \boldsymbol{U}, \boldsymbol{V}),$$

where $\boldsymbol{U} \in \mathbb{R}^{d_1 \times d_1}, \boldsymbol{V} \in \mathbb{R}^{d_2 \times d_2}$ denote the row- and column-wise covariance matrices, respectively. Equivalently, $\text{vec}(\boldsymbol{X}) \sim \mathcal{MN}(\boldsymbol{0}_{d_1 d_2}, \boldsymbol{U} \otimes \boldsymbol{V})$. We call $\boldsymbol{X}$ is Gaussian feature with bounded variation, if there exists a universal constant $c > 0$ such that $\|\boldsymbol{U}\|_{\text{sp}} \|\boldsymbol{V}\|_{\text{sp}} \le c$, as $d_1, d_2 \to \infty$.

Example 1. Let $\boldsymbol{X} = [\![x_{p,q}]\!] \in \mathbb{R}^{d_1 \times d_2}$ be a random matrix with i.i.d. Gaussian entries $x_{p,q} \sim_{\text{i.i.d}} N(0,1)$ for all $(p,q) \in [d_1] \times [d_2]$. Then, $\boldsymbol{X}$ is a Gaussian feature with bounded variation, because both $\boldsymbol{U}$ and $\boldsymbol{V}$ are identity matrices with spectral norm bounded by 1.

Proposition 1. *Let $\boldsymbol{X} \sim \mathcal{MN}(\boldsymbol{0}_{d_1 \times d_2}, \boldsymbol{U}, \boldsymbol{V})$ be Gaussian feature with bounded variation. Then, $\|\boldsymbol{X}\|_{sp} = \mathcal{O}(\sqrt{d_1 + d_2})$ and $\|\boldsymbol{X}\|_F = \mathcal{O}(\sqrt{d_1 d_2})$.*

Consider the modified linear function class

$$\mathcal{F} = \mathcal{F}(r, C) = \{f\colon \boldsymbol{X} \mapsto \langle \boldsymbol{B}, \boldsymbol{X} \rangle \;\big|\; \boldsymbol{B} \in \mathbb{R}^{d_1 \times d_2},\ \text{rank}(\boldsymbol{B}) \le r,\ \|\boldsymbol{B}\|_F \le C\}.$$

Lemma 2. *Let $\{\boldsymbol{X}_i\}_{i \in [n]}$ be a set of i.i.d. Gaussian features with bounded variation. The Rademacher*

Q: bounded feature + bounded coefficient —> whether low-rankness helps?

unbounded Gaussian feature + bounded coefficient —> yes, reduce product in d to linear in d

complexity of $\mathcal{F}$ *with respect to* $\{\boldsymbol{X}_i\}_{i\in[n]}$ *is*

$$\mathcal{R}_n(\mathcal{F}) \le 2C\sqrt{\frac{r(d_1+d_2)}{n}}.$$

Theorem 4.1 (Excess risk)**.** *Under Assumption 1, with very high probability*

$$\underbrace{\mathbb{P}[Y \neq sign\ f^*(\boldsymbol{X})] - \mathbb{P}[Y \neq sign\ \hat{f}_{pen}(\boldsymbol{X})]}_{\text{statistical error for estimating } f} \le \frac{4C\sqrt{r(d_1+d_2)}}{\sqrt{n}}. \tag{4}$$

Remark 3. The sample complexity for $\hat{p}_{\text{pen}}$ is $\mathcal{O}(r(d_1+d_2))$, which improves the sample complexity $\mathcal{O}(rd_1d_2)$ for $\hat{p}$ in (1).

<= |B||_F * ||sum of X||_F

Q: seems low-rankness does not play a role here ==> artifact of proof or it is a statistical fact?

Proof of Lemma 2.

F-norm <= nuclear norm <= \sqrt{r} * F-norm

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{(\sigma_i,\boldsymbol{X}_i)}\left\{\sup_{\text{rank}(\boldsymbol{B})\le r, \|\boldsymbol{B}\|_F\le C}\frac{1}{n}\sum_{i=1}^n \sigma_i\langle \boldsymbol{B}, \boldsymbol{X}_i\rangle\right\} &= \frac{2}{n}\mathbb{E}\left\{\sup_{\text{rank}(\boldsymbol{B})\le r, \|\boldsymbol{B}\|_F\le C}\langle \boldsymbol{B}, \sum_{i=1}^n \sigma_i\boldsymbol{X}_i\rangle\right\} \\
&\le \frac{1}{n}\mathbb{E}\left\{\sup_{\text{rank}(\boldsymbol{B})\le r, \|\boldsymbol{B}\|_F\le C}\|\boldsymbol{B}\|_*\,\|\sum_{i=1}^n \sigma_i\boldsymbol{X}_i\|_{\text{sp}}\right\} \\
&\le \frac{1}{n}\sqrt{r}C\mathbb{E}\left\{\|\sum_{i=1}^n \sigma_i\boldsymbol{X}_i\|_{\text{sp}}\right\},
\end{aligned}$$

where we have used the property $\|\boldsymbol{B}\|_* \le \sqrt{r}\|\boldsymbol{B}\|_F$ in the last line. Under the Gaussian feature assumption, $\boldsymbol{X}_i \overset{\mathcal{D}}{\sim} \sigma_i\boldsymbol{X}_i$ for all $i \in [n]$, and $\sqrt{n}\sum_{i=1}^n \sigma_i\boldsymbol{X}_i \overset{\mathcal{D}}{\sim} \mathcal{MN}(\boldsymbol{0}_{d_1\times d_2}, \boldsymbol{U}, \boldsymbol{V})$ (need to verify). The conclusion follows by noting $\mathbb{E}\{\|\sum_{i=1}^n \sigma_i\boldsymbol{X}_i\|_{\text{sp}}\} = \mathcal{O}(\sqrt{\frac{d_1+d_2}{n}})$. □