Theorems in high dimensional regime

Chanwoo Lee, July 5, 2020

1 Verification of Rademacher complexity Lemma 2.

In the proof of Lemma 2, the last inequality has $\frac{1}{n}\sqrt{r}C\mathbb{E}\left\{\|\sum_{i=1}^n \sigma_i \boldsymbol{X}_i\|_{\text{sp}}\right\}$. Notice $\boldsymbol{X}_i \overset{\mathcal{D}}{\sim} \sigma_i \boldsymbol{X}_i$. We have,

$$\text{Vec}(\sum_{i=1}^n \sigma_i \boldsymbol{X}_i) \overset{D}{\sim} \text{Vec}(\sum_{i=1}^n \boldsymbol{X}_i) \overset{D}{\sim} \mathcal{MN}(0_{d_1 d_2}, n\boldsymbol{U}\otimes\boldsymbol{V}) \overset{D}{\sim} \sqrt{n}\mathcal{MN}(0_{d_1 d_2}, \boldsymbol{U}\otimes\boldsymbol{V}).$$

Therefore, $\mathbb{E}\left\{\|\sum_{i=1}^n \sigma_i \boldsymbol{X}_i\|_{\text{sp}}\right\} = \mathcal{O}(\sqrt{n(d_1+d_2)})$, which proves the Rademacher complexity bound.

In more general case (This is from Cauchy-Schwartz note), where $E\boldsymbol{X}_i = 0$ and $\|\boldsymbol{X}_i\|_{\text{sp}} \le L$ for all $i \in [n]$ without Gaussian assumption, one can see that

$$\begin{aligned}\mathbb{E}\|\sum_{i=1}^n \sigma_i \boldsymbol{X}_i\|_{\text{sp}} &\le \sqrt{2v(\sum_{i=1}^n \sigma_i \boldsymbol{X}_i)\log(d_1+d_2)} + \frac{1}{3}L\log(d_1+d_2)\\ &\le L\left(\sqrt{2n\log(d_1+d_2)} + \frac{1}{3}\log(d_1+d_2)\right).\end{aligned}$$

from matrix Bernstein inequality. Based on this bound, we show the Rademacher complexity in more general case as,

$$\mathcal{R}_n(\mathcal{F}) \le \frac{1}{n}\|B\|_*\mathbb{E}\|S_n\|_{\text{sp}} \le \sqrt{r}\max_i \|\boldsymbol{X}_i\|_{\text{sp}}\|B\|_F\left(\sqrt{\frac{2\log(d_1+d_2)}{n}} + \frac{\log(d_1+d_2)}{3n}\right)$$

In this case, we expect $\max_i \|\boldsymbol{X}_i\|_{\text{sp}} \approx \mathcal{O}(d_1+d_2)$ but need to be verified.

2 Consistency of probability estimation with feature dimension term

Lemma 1. *Let $\mathcal{B}_r(k) = \{\boldsymbol{B}\in\mathbb{R}^{d_1\times d_2} : rank(B)\le r, \|\boldsymbol{B}\|\le k\}$. Then $N_2(\epsilon, \mathcal{B}_r(k)) \le \mathcal{O}\left(\left(\frac{k}{\epsilon}\right)^{r(d_1+d_2)}\right)$.*

Proof. Consider $\boldsymbol{B}\in\mathcal{B}_r(k)$ in the form of $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{V}^T$ where $\boldsymbol{U}\in\mathbb{R}^{d_1\times r}, \boldsymbol{V}\in\mathbb{R}^{d_2\times r}$ such that $\|\boldsymbol{U}\|\le\sqrt{k}$ and $\|\boldsymbol{V}\|\le\sqrt{k}$. We construct set of $\{U_i\}$ and $\{V_j\}$ such that for any $\boldsymbol{U},\boldsymbol{V}$, there exist i,j such that $\|\boldsymbol{U}-\boldsymbol{U}_i\|\le\epsilon/2\sqrt{k}$ and $\|\boldsymbol{V}-\boldsymbol{V}_j\|\le\epsilon/2\sqrt{k}$. Then, epsilon balls with centers in $\{\boldsymbol{U}\boldsymbol{V}^T : \boldsymbol{U}\in\{\boldsymbol{U}_i\}, \boldsymbol{V}\in\{\boldsymbol{V}_j\}\}$ can cover $\mathcal{B}_r(k)$ because for any $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{V}^T\in\mathcal{B}_r(k)$, we have $\boldsymbol{U}_i\boldsymbol{V}_j^T \in \{\boldsymbol{U}\boldsymbol{V}^T : \boldsymbol{U}\in\{\boldsymbol{U}_i\}, \boldsymbol{V}\in\{\boldsymbol{V}_j\}\}$ such that

$$\begin{aligned}\|\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{U}_i\boldsymbol{V}_j^T\| &\le \|\boldsymbol{U}\boldsymbol{V}^T - \boldsymbol{U}\boldsymbol{V}_j^T\| + \|\boldsymbol{U}\boldsymbol{V}_j^T - \boldsymbol{U}_i\boldsymbol{V}_j^T\|\\ &\le \|\boldsymbol{U}\|\|\boldsymbol{V}-\boldsymbol{V}_j\| + \|\boldsymbol{V}_j\|\|\boldsymbol{U}-\boldsymbol{U}_i\|\\ &\le \sqrt{k}\frac{\epsilon}{2\sqrt{k}} + \sqrt{k}\frac{\epsilon}{2\sqrt{k}} \le \epsilon.\end{aligned}$$

Therefore, the covering number of $N_2(\epsilon,\mathcal{B}_r(k)) \le \mathcal{O}\left(\left(\frac{k}{\epsilon}\right)^{r(d_1+d_2)}\right)$, where $\mathcal{O}\left(\left(\frac{k}{\epsilon}\right)^{r(d_1)}\right)$ comes from $\{\boldsymbol{U}_i\}$ and $\mathcal{O}\left(\left(\frac{k}{\epsilon}\right)^{r(d_2)}\right)$ from $\{\boldsymbol{V}_j\}$. □

Remark 1. This covering number bound is not the sharpest bound. There are several reasons for that. First, there are many representations of $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{V}^T$ i.e. the representation is not unique for given $\boldsymbol{B}$, which means there might be redundant centers in the set. In addition, when considered matrices are full rank ($r = \min(d_1, d_2)$), this bound is slightly greater than the covering number bound of coefficient $\mathcal{B}(k)$ only with norm constraint. However, the covering bound in Lemma 1 is small enough to show benefit of low rank structure.

Lemma 2. *Let* $\mathcal{F}_r(k) = \{f : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R} : f(\boldsymbol{X}) = \langle \boldsymbol{B}, \boldsymbol{X} \rangle$ *where* $\mathrm{rank}(B) \leq r, \|\boldsymbol{B}\| \leq k\}$. *Suppose that there exists* $G > 0$ *such that* $\sqrt{\mathbb{E}\|\boldsymbol{X}\|^2} \leq G$. *Then the covering number* $N_2(\epsilon, \mathcal{F}_r^V(k))$ *is bounded by*

$$\log N_2(\epsilon, \mathcal{F}_r^V(k)) \leq \mathcal{O}\left(r(d_1 + d_2) \log\left(\frac{Gk}{\epsilon}\right)\right).$$

Proof. Let $f_{\boldsymbol{B}}(\boldsymbol{X}) = \langle \boldsymbol{B}, \boldsymbol{X} \rangle$ and $K(\boldsymbol{X}, \boldsymbol{X}') = \langle \boldsymbol{X}, \boldsymbol{X}' \rangle$. Then for any $f_{\boldsymbol{B_1}}, f_{\boldsymbol{B_2}} \in \mathcal{F}_r(k)$,

$$\langle f_{\boldsymbol{B}_1}, f_{\boldsymbol{B}_2} \rangle = \langle K_{\boldsymbol{B}_1}, K_{\boldsymbol{b_2}} \rangle = K(\boldsymbol{B}_1, \boldsymbol{B}_2) = \langle \boldsymbol{B}_1, \boldsymbol{B}_2 \rangle.$$

Therefore, the metric space $(\mathcal{F}_r(k), \|\cdot\|_K)$ is isomorphic to $(\mathcal{B}_r(k), \|\cdot\|)$ where $\mathcal{B}_r(k) = \{\boldsymbol{B} \in \mathbb{R}^{d_1 \times d_2} : \mathrm{rank}(B) \leq r, \|\boldsymbol{B}\| \leq k\}$. From Lemma 1, we have the covering number $N_2(\epsilon, \mathcal{B}_r(k)) \leq \mathcal{O}\left(\left(\frac{k}{\epsilon}\right)^{r(d_1+d_2)}\right)$. Then Note that, for functions f_ℓ and f_u,

$$\begin{aligned}
\|V^T(f_\ell, \cdot) - V^T(f_u, \cdot)\|_2^2 &\leq \|f_\ell - f_u\|_2^2 = \mathbb{E}|\langle \boldsymbol{B}_\ell - \boldsymbol{B}_u, \boldsymbol{X} \rangle|^2 \\
&\leq \|\boldsymbol{B}_\ell - \boldsymbol{B}_u\|^2 \mathbb{E}\|\boldsymbol{X}\|^2 \\
&\leq \|\boldsymbol{B}_\ell - \boldsymbol{B}_u\|^2 G^2 = \|f_\ell - f_u\|_K G^2,
\end{aligned}$$

implying that $N_2(\epsilon, \mathcal{F}^V(k) \leq N_2(\epsilon, \mathcal{F}(k)) \leq N_{G\|\cdot\|_K}(\epsilon, \mathcal{F}(k)) \leq \mathcal{O}\left(r(d_1 + d_2) \log\left(\frac{Gk}{\epsilon}\right)\right).$ ☐

Lemma 3. *Let* $k > 0$ *be a given constant. If* $\frac{1}{Ke} > L > 0$, *we have*

$$\int_{\mathcal{O}(L)}^{\mathcal{O}(\sqrt{L})} \sqrt{\log\left(\frac{k}{\omega}\right)} d\omega \leq \mathcal{O}\left(\sqrt{L \log\left(\frac{k}{\sqrt{L}}\right)}\right).$$

Proof.

$$\begin{aligned}
\int_{\mathcal{O}(L)}^{\mathcal{O}(\sqrt{L})} \sqrt{\log\left(\frac{k}{\omega}\right)} - \frac{1}{2\sqrt{\log\left(\frac{k}{\omega}\right)}} d\omega &= k\left[\omega \sqrt{\log\left(\frac{1}{\omega}\right)}\right]_{\mathcal{O}(L/k)}^{\mathcal{O}(\sqrt{L}/k)} \\
&= \mathcal{O}\left(\sqrt{L \log\left(\frac{k}{\sqrt{L}}\right)}\right)
\end{aligned} \tag{1}$$

The first equality in (1) is from changing variable. Notice that

$$\int_{\mathcal{O}(L)}^{\mathcal{O}(\sqrt{L})} \sqrt{\log\left(\frac{k}{\omega}\right)} - \frac{1}{2\sqrt{\log\left(\frac{k}{\omega}\right)}} d\omega \geq \int_{\mathcal{O}(L)}^{\mathcal{O}(\sqrt{L})} \sqrt{\log\left(\frac{k}{\omega}\right)} - \mathcal{O}(1) d\omega, \tag{2}$$

from the condition on L. Combining Equation (1) and Equation (2) completes the proof. ☐

Lemma 4. $\sqrt{\frac{d}{L}\log\left(\frac{k}{\sqrt{L}}\right)} \leq \sqrt{n}$ *holds if* $L \leq \frac{log(n/d)+2log(k)}{n/d}$.

Proof. Suppose $L \leq \frac{log(n/d)+2log(k)}{n/d}$. By plugging in, we have

$$\begin{aligned}\sqrt{\frac{d}{L}\log\left(\frac{k}{\sqrt{L}}\right)} &\leq \sqrt{\frac{n}{log(n/d)+2log(k)}\left(\frac{\log(n/d)+2\log(k)-\log\log(nk^2/d)}{2}\right)} \\ &\leq \sqrt{n}.\end{aligned}$$

□

Let $\bar{f}_\pi$ be a Bayes rule. In addition, let $e_V(f,\bar{f}_\pi) = \mathbb{E}\{V(f,\boldsymbol{X},y) - V(\bar{f}_\pi,\boldsymbol{X},y)\}$ with $V(f,\boldsymbol{X},y) = S(y)L\{yf(\boldsymbol{X})\}$.

Based on function class $\mathcal{F}_r(M)$, we have the following theorem.

Theorem 2.1. *Assume that*

1. *For some positive sequence such that $s_n \to 0$ as $n \to \infty$, there exists $f^*_\pi \in \mathcal{F}_r(M)$ such that $e_V(f^*_\pi, \bar{f}_\pi) \leq s_n$.*
2. *For a given π, there exists $\eta > 0$ such that $|\mathbb{P}(y=1|\boldsymbol{X}) - \pi| \geq \eta$ almost surely.*
3. *Considered feature space is uniformly bounded such that there exists $0 < G < \infty$ satisfying $\sqrt{\mathbb{E}\|\boldsymbol{X}\|^2} \leq G$*

Then, for the estimator $\hat{p}$ obtained from our algorithm, there exists a constant c such that

$$\mathbb{P}\left\{\|\hat{p}-p\|_1 \geq \frac{1}{2m} + \frac{1}{2\eta}(m+1)\delta_n^2\right\} \leq 15\exp\{-cn(\lambda J^*_\pi)\},$$

provided that $\lambda^{-1} \geq \frac{GJ^*_\pi}{2\delta_n^2}$ *where* $J^*_\pi = \max(J(f^*_\pi), 1)$ *and* $\delta_n = \max\left(\mathcal{O}\left(\frac{\log(n/r(d_1+d_2))+2\log(GM)}{n/r(d_1+d_2)}\right), s_n\right)$.

Proof. We apply Theorem 3 in [1] to our case. We show that the Assumption 2 in [1] is satisfied. The first condition of the assumption is

$$\sup_{\{f\in\mathcal{F}: e_{V^T}(f,\bar{f}_\pi)\leq\delta\}} \|\text{sign}(f) - \text{sign}(\bar{f}_\pi)\| \leq a_1\delta^\alpha\}.$$

Notice that

$$\begin{aligned}e_{V^T}(f,\bar{f}_\pi) &= \mathbb{E}\left[S(y)L(yf(\boldsymbol{X}))\wedge T - S(y)L(y\bar{f}_\pi(\boldsymbol{X}))\right] \\ &\geq \mathbb{E}\left[S(y)(1-\text{sign}(yf(\boldsymbol{X})) - S(y)(1-\text{sign}(y\bar{f}_\pi(\boldsymbol{X}))\right] \\ &= \mathbb{E}\left[yS(y)\left(\text{sign}(\bar{f}_\pi) - \text{sign}(f)\right)\right] \\ &= \mathbb{E}\left[\mathbb{E}(yS(y)|\boldsymbol{X})\left(\text{sign}(\bar{f}_\pi) - \text{sign}(f)\right)\right] \\ &= \mathbb{E}\left[|\mathbb{P}(y=1|\boldsymbol{X}) - \pi||\text{sign}(\bar{f}_\pi) - \text{sign}(f)|\right] \\ &\geq \eta\mathbb{E}|\text{sign}(\bar{f}_\pi) - \text{sign}(f)| = \eta\|\text{sign}(\bar{f}_\pi) - \text{sign}(f)\|. \end{aligned} \tag{3}$$

Therefore, the first condition is satisfied with $a_1 = \frac{1}{\eta}$ and $\alpha = 1$. The second condition of the assumption is

$$\sup_{\{f\in\mathcal{F}:e_{V^T}(f,\bar{f}_\pi)\le\delta\}} \mathrm{var}\{V(f,\boldsymbol{X},y) - V(\bar{f}_\pi,\boldsymbol{X},y) \le a_2\delta^\beta\}.$$

Notice that

$$\begin{aligned}\mathrm{var}\{V^T(f,\boldsymbol{X},y) - V(\bar{f}_\pi,\boldsymbol{X},y)\} &\le \mathbb{E}|V(^T f,\boldsymbol{X},y) - V(\bar{f}_\pi,\boldsymbol{X},y)|^2 \\ &\le T\mathbb{E}|V^T(f,\boldsymbol{X},y) - V(\bar{f}_\pi,\boldsymbol{X},y)| \\ &= T\left(\lambda_1 + \lambda_2\right).\end{aligned}$$

where

$$\begin{aligned}\lambda_1 &= \mathbb{E}\left|S(y)(1-\mathrm{sign}(yf(\boldsymbol{X})) - V(\bar{f}_\pi,\boldsymbol{X},y)\right| = \mathbb{E}|S(y)||\mathrm{sign}(f) - \mathrm{sign}(\bar{f}_\pi)| \\ &\le \|\mathrm{sign}(f) - \mathrm{sign}(\bar{f}_\pi)|\|_1 \le \eta^{-1}e_{V^T}(f,\bar{f}_\pi) \quad \text{ from Equation (3)}.\end{aligned}$$

and

$$\begin{aligned}\lambda_2 &= \mathbb{E}\left[V^T(f,\boldsymbol{X},y) - S(y)(1-\mathrm{sign}(yf(\boldsymbol{X}))\right] \\ &\le e_{V^T}(f,\bar{f}_\pi) + \mathbb{E}\left\{V(\bar{f}_\pi,\boldsymbol{X},y) - S(y)(1-\mathrm{sign}(yf(\boldsymbol{X}))\right\} \\ &\le 2e_{V^T}(f,\bar{f}_\pi)\end{aligned}$$

Therefore, β in [1] can be replaced by 1.

Now we check Assumption 3 in [1]. From Lemma 2, we have

$$H_B(\epsilon,\mathcal{F}^V(k)) \le \mathcal{O}\left(r(d_1+d_2)\log\left(\frac{Gk}{\epsilon}\right)\right).$$

Therefore, we have the following equation from Lemma 3.

$$\phi(\epsilon,k) \approx \int_{\mathcal{O}(L)}^{\mathcal{O}(\sqrt{L})} \sqrt{r(d_1+d_2)\log\left(\frac{kG}{\omega}\right)}d\omega/L \lessapprox \mathcal{O}\left(\sqrt{r(d_1+d_2)}\left(\log\left(\frac{kG}{\sqrt{L}}\right)/L\right)^{1/2}\right),$$

where $L = \min\{\epsilon^2+\lambda(k/2-1)H^*_\pi, 1\}$. Solving Assumption 3 in [1] gives us $\epsilon_n^2 = \mathcal{O}\left(\frac{\log(n/r(d_1+d_2))+2\log(GM)}{n/r(d_1+d_2)}\right)$ by Lemma 4 when $\epsilon_n^2 \ge \lambda G J^*_\pi$. Plugging each variable into Theorem 3 proves the theorem. Notice that condition of λ is replaced because $\{\epsilon_n^2 \ge \lambda G J^*_\pi\} \subset \{\epsilon_n^2 \ge 2\lambda J^*_\pi\}$ when $rG \ge 2$. □

References

[1] Junhui Wang, Xiaotong Shen, and Yufeng Liu. Probability estimation for large-margin classifiers. *Biometrika*, 95(1):149–167, March 2008.