

# 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  $\frac{1}{n}\sqrt{r}C\mathbb{E}\{\|\sum_{i=1}^n \sigma_i \mathbf{X}_i\|_{\text{sp}}\}$ . Notice  $\mathbf{X}_i \stackrel{\mathcal{D}}{\sim} \sigma_i \mathbf{X}_i$ . We have,

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

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

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

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

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

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

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

d1+d2 -> sqrt(d1+d2) ?

## 2 Consistency of probability estimation with feature dimension term

Use \FnormSize{\cdot} in latex to denote F-norm

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

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

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

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

Good discussion.

In our case, sharp in leading term because  $r(d_1+d_2) \sim r(d_1+d_2) - r^2 = r(d_1+d_2) + o(d)$  as  $d \rightarrow \infty$

**Remark 1.** This covering number bound is not the sharpest bound. There are several reasons for that. First, there are many representations of  $\mathbf{B} = \mathbf{U}\mathbf{V}^T$  i.e. the representation is not unique for given  $\mathbf{B}$ , 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  $\mathcal{F}_r(k) = \{f : \mathbb{R}^{d_1 \times d_2} \rightarrow \mathbb{R} : f(\mathbf{X}) = \langle \mathbf{B}, \mathbf{X} \rangle \text{ where } \text{rank}(B) \leq r, \|\mathbf{B}\| \leq k\}$ . Suppose that there exists  $G > 0$  such that  $\sqrt{\mathbb{E}\|\mathbf{X}\|^2} \leq G$ . 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  $f_{\mathbf{B}}(\mathbf{X}) = \langle \mathbf{B}, \mathbf{X} \rangle$  and  $K(\mathbf{X}, \mathbf{X}') = \langle \mathbf{X}, \mathbf{X}' \rangle$ . Then for any  $f_{\mathbf{B}_1}, f_{\mathbf{B}_2} \in \mathcal{F}_r(k)$ ,

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

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

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

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

**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. □

Assume  $X \sim \text{Unif}[0,1]$  for simplicity —> Lebesgue measure = probability measure

1. statement of assumption 2 needs to be modified —> in order for the current proof to go through.

2. Is the current assumption 2 reasonable?

Partially yes. If the function  $p(x)$  is a step function with finite/zero-measured jumps, then the current assumption 2 holds.

3. Can we relax the assumption 2?

Partially yes. If the function  $p(x)$  is a linear function, then the current assumption 2 fails but the earlier smoothness assumption holds.

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

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

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

□

0-1 function?

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

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

**Interpretation of Assumption (2):** For a given  $\pi$ , your inequality is called the Tsypakov's noise condition.

==> The neighborhood of  $\pi$  falls outside of the image of  $p(x) := P(y=1|X)$  (which is a function from  $X$  to  $[0,1]$ )

**Theorem 2.1.** Assume that

==> Equivalently, the two regions  $\{p(x) > \pi\}$  and  $\{p(x) < \pi\}$  are well separated. (Draw a picture when  $\dim X = 1$ )

1. For some positive sequence such that  $s_n \rightarrow 0$  as  $n \rightarrow \infty$ , there exists  $f_\pi^* \in \mathcal{F}_r(M)$  such that  $e_V(f_\pi^*, \bar{f}_\pi) \leq s_n$ . hold for all  $\pi$
2. For a given  $\pi$ , there exists  $\eta > 0$  such that  $|\mathbb{P}(y = 1|\mathbf{X}) - \pi| \geq \eta$  almost surely. w.r.t. distribution over  $X$ .

3. Considered feature space is uniformly bounded such that there exists  $0 < G < \infty$  satisfying  $\sqrt{\mathbb{E}\|\mathbf{X}\|^2} \leq G$

Remark: Do we require Assumption (2) to hold uniformly over  $\pi \in [0,1]$ , or at least over all weights  $\pi_1$  up to  $\pi_m$ ? Then, for the estimator  $\hat{p}$  obtained from our algorithm, there exists a constant  $c$  such that

This seems to rather restrictive as  $m \rightarrow \infty$ ...considering the picture on separability between preimages.

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

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

log term is negligible

==>  $G$  is also allowed to increase with  $d$ , e.g. when  $X$  follows normal distribution

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

0-1 function by definition?

To-do:

1. Visualization

Notice that

2. What aspects of the program will affect

the smooth parameter/bound control?

Give a qualitative assessment:  $e_{V,T}(f, \bar{f}_\pi) = \mathbb{E}[S(y)L(yf(\mathbf{X})) \wedge T - S(y)L(y\bar{f}_\pi(\mathbf{X}))]$

e.g., richer class, simpler ground truth function,  $\geq \mathbb{E}[S(y)(1 - \text{sign}(yf(\mathbf{X}))) - S(y)(1 - \text{sign}(y\bar{f}_\pi(\mathbf{X})))]$

certain measure of functional complexity,

$$= \mathbb{E}[yS(y)(\text{sign}(\bar{f}_\pi) - \text{sign}(f))]$$

monotonic, # of sign-change?

$$= \mathbb{E}[\mathbb{E}(yS(y)|\mathbf{X})(\text{sign}(\bar{f}_\pi) - \text{sign}(f))]$$

$$\text{equal to } p_{\text{pi}}(x) = \mathbb{E}[|\mathbb{P}(y = 1|\mathbf{X}) - \pi||\text{sign}(\bar{f}_\pi) - \text{sign}(f)|]$$

$$> \eta |\mathbb{E}|\text{sign}(\bar{f}_\pi) - \text{sign}(f)| = \eta \|\text{sign}(\bar{f}_\pi) - \text{sign}(f)\|. \quad (3)$$

and  $f_{\text{bar}}$  have opposite signs.

$$\text{eta}(\pi, f) := \mathbb{E}[\text{p}_{\text{pi}}(x) * \mathbb{1}\{\mathbf{X}: \text{sign}(f(\mathbf{X})) \neq \text{sign}(\text{p}_{\text{pi}}(x))\}]$$

Q: How to visualize in 1-dim? (I have come up with a picture, but you should think on your own first). Try relating sign difference to

area in  $X$ ,  $e_V$  difference to area in  $(X, Y)$ . Under which scenario, is  $\text{eta}(\pi, f)$  uniformly bounded

Conjecture: suppose the ground truth probability  $P(y=1|X)$  is a non-zero linear function in  $X$ , and every function  $f(x)$  in  $F$  has finite change-of-sign points (e.g. in the linear function class). Under uniform assumption  $X \sim \text{Unif}[0,1]$ , the smooth parameter alpha = 1/2.

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

$$\sup_{\{f \in \mathcal{F} : e_{VT}(f, \bar{f}_\pi) \leq \delta\}} \text{var}\{V(f, \mathbf{X}, y) - V(\bar{f}_\pi, \mathbf{X}, y) \leq a_2 \delta^\beta\}.$$

Notice that

$$\begin{aligned} \text{var}\{V^T(f, \mathbf{X}, y) - V(\bar{f}_\pi, \mathbf{X}, y)\} &\leq \mathbb{E}|V^T(f, \mathbf{X}, y) - V(\bar{f}_\pi, \mathbf{X}, y)|^2 \\ &\leq T\mathbb{E}|V^T(f, \mathbf{X}, y) - V(\bar{f}_\pi, \mathbf{X}, y)| \\ &= T(\lambda_1 + \lambda_2). \end{aligned}$$

where

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

and

$$\begin{aligned} \lambda_2 &= \mathbb{E}[V^T(f, \mathbf{X}, y) - S(y)(1 - \text{sign}(yf(\mathbf{X})))] \\ &\leq e_{VT}(f, \bar{f}_\pi) + \mathbb{E}\{V(\bar{f}_\pi, \mathbf{X}, y) - S(y)(1 - \text{sign}(yf(\mathbf{X})))\} \\ &\leq 2e_{VT}(f, \bar{f}_\pi) \end{aligned}$$

Therefore,  $\beta$  in [1] can be replaced by 1.

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

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

Therefore, we have the following equation from Lemma 3.

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

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  $\epsilon_n^2 \geq \lambda G J_\pi^*$ . Plugging each variable into Theorem 3 proves the theorem. Notice that condition of  $\lambda$  is replaced because  $\{\epsilon_n^2 \geq \lambda G J_\pi^*\} \subset \{\epsilon_n^2 \geq 2\lambda J_\pi^*\}$  when  $rG \geq 2$ .  $\square$

## References

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

[2] Ref on Tsybakov's noise condition:

Statistical performance of Support Vector Machine, Blanchard, Bousquet, and Massart, AOS, 2008, 36, 2, 489-531  
Try searching the key word ‘‘Tsybakov’’ throughout the paper.