A Boost to the Upper Bound on the Variance in Liang and Rakhlin [2020]

Tengyuan Liang*[1]

[1]University of Chicago, Booth School of Business

January 6, 2021

Abstract

In this document, we provide a boost to the upper bound on the Variance, derived in Liang and Rakhlin [2020]. The boost to the upper bound provides easier interpretation, and further connects to the population eigenvalues of the covariance matrix.

Keywords— Minimum-norm interpolation, kernel rigeless regression.

In the Theorem 1 of Liang and Rakhlin [2020], the variance upper bound $\mathbf{V}$ can be boosted to

$$\mathbf{V} \leq \frac{C\sigma^2}{\gamma} \cdot \inf_{0\leq k\leq d} \left\{ \lambda_1(\Sigma)\frac{k}{n} + \lambda_k(\Sigma) \right\} , \tag{0.1}$$

where $\lambda_i(\Sigma), 1 \leq i \leq d$ are the population eigenvalues sorted in a non-increasing order. All the notations follow from the original paper.

To see this, let's only consider the case with $\alpha = 0$ and $\beta = 1$ (this can be done by centering and scaling the kernel). The full expression in $\mathbf{V}$ in Page 1339 of Liang and Rakhlin [2020] reads

$$\mathbf{V} \leq 8\sigma^2 \cdot \mathbf{E}_{\mathbf{x}\sim\mu} \left\| \left(\gamma I + \frac{XX^\star}{d}\right)^{-1} \frac{X\mathbf{x}}{d} \right\|^2 = 8\sigma^2 \cdot \mathrm{Tr}\left(\left(d\gamma I + XX^\star\right)^{-1} X\Sigma X^\star \left(d\gamma I + XX^\star\right)^{-1} \right) \tag{0.2}$$

Denote $\Sigma = \sum_{j=1}^{d} \lambda_j(\Sigma) \cdot u_j u_j^\star$ as the eigenvalue decomposition of the population covariance matrix. Take any $1 \leq k \leq d$. Denote $\Sigma_{>k} = \sum_{j>k} \lambda_j(\Sigma) \cdot u_j u_j^\star$, for this high frequency component we have

$$\begin{aligned} \mathrm{Tr}\left(\left(d\gamma I + XX^\star\right)^{-1} X\Sigma_{>k} X^\star \left(d\gamma I + XX^\star\right)^{-1} \right) &\leq \lambda_k(\Sigma) \cdot \mathrm{Tr}\left(\left(d\gamma I + XX^\star\right)^{-1} XX^\star \left(d\gamma I + XX^\star\right)^{-1} \right) && (0.3) \\ &\leq \lambda_k(\Sigma) \sum_{i=1}^{n} \frac{\lambda_i(XX^\top)}{(d\gamma + \lambda_i(XX^\top))^2} && (0.4) \\ &\leq \lambda_k(\Sigma) n \frac{1}{4d\gamma} \leq \frac{C}{4\gamma} \cdot \lambda_k(\Sigma) && (0.5) \end{aligned}$$

*tengyuan.liang@chicagobooth.edu. The author acknowledges the NSF Career Award and the George C. Tiao Fellowship for the support.

where the last line uses Remark 5.1 in Liang and Rakhlin [2020], $\frac{t}{(r+t)^2} \le \frac{1}{4r}$ for all $a, r > 0$. This proof is identical to that in Liang and Rakhlin [2020]. The last step also uses the fact $d \asymp n$.

Now for the low frequency component, $\Sigma_{\le k} = \sum_{j \le k} \lambda_j(\Sigma) \cdot u_j u_j^\star$. Denote $P_{u_j}^\perp := I - u_j u_j^\star \in \mathbb{R}^{d \times d}$ the projection matrix to the orthogonal complement of u_j, we have

$$\mathrm{Tr}\left(\left(d\gamma I + XX^\star\right)^{-1} X\Sigma_{\le k}X^\star\left(d\gamma I + XX^\star\right)^{-1}\right) \le \sum_{j \le k} \lambda_j(\Sigma) \cdot \|\left(d\gamma I + XX^\star\right)^{-1} Xu_j\|^2 \tag{0.6}$$

with the definition $v := Xu_j \in \mathbb{R}^n$, and $M := d\gamma I + XP_{u_j}^\perp X^\star$, we continue to bound

$$\begin{aligned}
\left\|\left(d\gamma I + XX^\star\right)^{-1} Xu_j\right\|^2 &= \left\|(M + vv^\star)^{-1} v\right\|^2 && (0.7)\\
&= \left\|\left(M^{-1} - \frac{M^{-1}vv^\star M^{-1}}{1 + v^\star M^{-1} v}\right)v\right\|^2 \quad \text{Woodbury formula} && (0.8)\\
&= \frac{v^\star M^{-2} v}{\left(1 + v^\star M^{-1} v\right)^2} \le \frac{1}{d\gamma} \frac{v^\star M^{-1} v}{\left(1 + v^\star M^{-1} v\right)^2} \quad \text{recall } \lambda_{\min}(M) > d\gamma && (0.9)\\
&\le \frac{1}{4\gamma}\frac{1}{d}\ , && (0.10)
\end{aligned}$$

where the last line again uses Remark 5.1 in Liang and Rakhlin [2020]. Therefore recalling $d \asymp n$

$$\begin{aligned}
\mathrm{Tr}\left(\left(d\gamma I + XX^\star\right)^{-1} X\Sigma_{\le k}X^\star\left(d\gamma I + XX^\star\right)^{-1}\right) &\le \sum_{j \le k} \lambda_j(\Sigma) \cdot \|\left(d\gamma I + XX^\star\right)^{-1} Xu_j\|^2 && (0.11)\\
&\le k\lambda_1(\Sigma) \cdot \frac{1}{4\gamma d} \le \frac{C}{4\gamma} \cdot \lambda_1(\Sigma)\frac{k}{n}\ . && (0.12)
\end{aligned}$$

The proof is now complete by combining Equations (0.5) and (0.12).

References

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel “Ridgeless” regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, June 2020. doi: 10.1214/19-AOS1849.