Supplementary Material for Partial Trace Regression and Low-Rank Kraus Decomposition

Hachem Kadri [1] **Stéphane Ayache** [1] **Riikka Huusari** [2] **Alain Rakotomamonjy** [3 4] **Liva Ralaivola** [4]

In this supplementary material, we prove Lemma 3 and Theorem 4 in Section 2.3 of the main paper. Let us first recall the definition of pseudo-dimension.

Definition 1 *(Shattering Mohri et al., 2018, Def. 10.1)*
Let G be a family of functions from X to $\mathbb{R}$. A set $\{x_1, ..., x_m\} \subset X$ is said to be shattered by G if there exist $t_1, ..., t_m \in \mathbb{R}$ such that,

$$f(x) = \left| \left\{ \begin{bmatrix} sign(g(x_1) - t_1) \\ \vdots \\ sign(g(x_m) - t_m) \end{bmatrix} : g \in G \right\} \right| = 2^m.$$

Definition 2 *(pseudo-dimension Mohri et al., 2018, Def. 10.2)*
Let G be a family of functions from X to $\mathbb{R}$. Then, the pseudo-dimension of G, denoted by $Pdim(G)$, is the size of the largest set shattered by G.

In the following we consider that the expected loss of any hypothesis $h \in \mathcal{F}$ is defined by $R(h) = \mathbb{E}_{(X,Y)}\big[\ell\big(Y, h(X)\big)\big]$ and its empirical loss by $\hat{R}(h) = \frac{1}{l}\sum_{i=1}^{l} \ell\big(Y, h(X)\big)$. To prove Lemma 3 and Theorem 4, we need the following two results.

Theorem 1 *(Srebro, 2004, Theorem 35)*
The number of sign configurations of m polynomials, each of degree at most d, over n variables is at most $\left(\frac{4edm}{n}\right)^n$ for all $m > n > 2$.

Theorem 2 *(Mohri et al., 2018, Theorem 10.6)*
Let H be a family of real-valued functions and let $G = \{x \mapsto L(h(x), f(x)) : h \in H\}$ be the family of loss functions associated to H. Assume that the pseudo-dimension of G is bounded by d and that the loss function L is bounded by M. Then, for any $\delta > 0$, with probability at least δ over the choice of a sample of size m, the following inequality holds for all $h \in H$:

$$R(h) \leq \hat{R}(h) + M\sqrt{\frac{2d \log\left(\frac{em}{d}\right)}{m}} + M\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2m}}.$$

1. Proof of Lemma 3

We now prove Lemma 3 in Section 2.3 of the main paper.

Lemma 3 *The pseudo-dimension of the real-valued function class $\tilde{\mathcal{F}}$ with domain $\mathbb{M}_p \times [q] \times [q]$ defined by*

$$\tilde{\mathcal{F}} = \{(X, s, t) \mapsto \big(\Phi(X)\big)_{st} : \Phi(X) = \sum_{j=1}^{r} A_j X A_j^\top\}$$

is upper bounded by $pqr \log(\frac{8epq}{r})$.

Proof: It is well known that the pseudo-dimension of a vector space of real-valued functions is equal to its dimension (Mohri et al., 2018, Theorem 10.5). Since $\tilde{\mathcal{F}}$ is a subspace of the p^2q^2-dimensional vector space

$$\big\{(X, s, t) \mapsto \big(\Phi(X)\big)_{st} \; : \; \Phi \in \mathcal{L}(\mathbb{M}_p; \mathbb{M}_q)\big\}$$

of real-valued functions with domain $\mathbb{M}_p \times [q] \times [q]$ the pseudo-dimension of $\tilde{\mathcal{F}}$ is bounded by p^2q^2.

Now, let $m \leq p^2q^2$ and let $\{(X_k, s_k, t_k)\}_{k=1}^m$ be a set of points that are pseudo-shattered by $\tilde{\mathcal{F}}$ with thresholds $t_1, \cdots, t_m \in \mathbb{R}$. Then for each binary labeling $(u_1, \cdots, u_m) \in \{-, +\}^m$, there exists $\tilde{\Phi} \in \tilde{\mathcal{F}}$ such that $sign(\tilde{\Phi}(X_k, s_k, t_k) - v_k) = u_k$. Any function $\tilde{\Phi} \in \tilde{\mathcal{F}}$ can be written as

$$\tilde{\Phi}(X, s, t) = \big(\sum_{j=1}^{r} A_j X A_j^\top\big)_{st}, \tag{1}$$

where $A_j \in \mathbb{M}_{q \times p}, \forall j \in [r]$. If we consider the pqr entries of A_j, $j = 1, \ldots, r$, as variables, the set $\{\tilde{\Phi}(X_k, s_k, t_k) - v_k\}_{k=1}^m$ can be seen (using Eq. 1) as a set of m polynomials

[1]Aix-Marseille University, CNRS, LIS, Marseille, France [2]Helsinki Institute for Information Technology HIIT, Department of Computer Science, Aalto University, Espoo, Finland [3]Université Rouen Normandie, LITIS, Rouen, France [4]Criteo AI Lab, Paris, France. Correspondence to: Hachem Kadri <hachem.kadri@univ-amu.fr>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

of degree 2 over these variables. Applying Theorem 1 above, we obtain that the number of sign configurations, which is equal to 2^m, is bounded by $\left(\frac{8em}{pqr}\right)^{pqr}$. The result follows since $m \leq p^2q^2$. ■

2. Proof of Theorem 4

In this section, we prove Theorem 4 in Section 2.3 of the main paper.

Theorem 4 *Let $\ell : \mathbb{M}_q \to \mathbb{R}$ be a loss function satisfying*

$$\ell(Y, Y') = \frac{1}{q^2} \sum_{s,t} \ell'(Y_{st}, Y'_{st})$$

for some loss function $\ell' : \mathbb{R} \to \mathbb{R}^+$ bounded by γ. Then for any $\delta > 0$, with probability at least $1 - \delta$ over the choice of a sample of size l, the following inequality holds for all $h \in \mathcal{F}$:

$$R(h) \leq \hat{R}(h) + \gamma \sqrt{\frac{pqr \log(\frac{8epq}{r}) \log(\frac{l}{pqr})}{l}} + \gamma \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2l}}.$$

Proof: For any $h : \mathbb{M}_p \to \mathbb{M}_q$ we define $\tilde{h} : \mathbb{M}_p \times [q] \times [q] \to \mathbb{R}$ by $\tilde{h}(X, s, t) = \big(h(X)\big)_{st}$. Let $\mathcal{D}$ denote the distribution of the input-output data. We have

$$\begin{aligned}
R(h) &= \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell(Y, h(X))] \\
&= \frac{1}{q^2} \sum_{s,t} \mathbb{E}_{(X,Y)\sim\mathcal{D}}[\ell'(Y_{st}, h(X)_{st}] \\
&= \mathbb{E}_{\substack{(X,Y)\sim\mathcal{D} \\ s,t\sim\mathcal{U}(q)}}[\ell'(Y_{st}, \tilde{h}(X, s, t))],
\end{aligned}$$

where $\mathcal{U}(q)$ denotes the discrete uniform distribution on $[q]$. It follows that $R(h) = R(\tilde{h})$. By the same way, we can show that $\hat{R}(h) = \hat{R}(\tilde{h})$. The generalization bound is then obtained using Theorem 2 above. ■

References

Mohri, M., Rostamizadeh, A., and Talwalkar, A. *Foundations of machine learning*. MIT press, 2018.

Srebro, N. *Learning with matrix factorizations*. PhD thesis, MIT, 2004.