Supplementary Material

On the Spectrum of Random Features Maps of High Dimensional Data

A. Proof of Theorem 1

To obtain the result presented in Theorem 1, we begin by recalling the expression of the average kernel matrix $\mathbf{\Phi}$, with

$$\mathbf{\Phi}_{i,j} = \mathbf{\Phi}(\mathbf{x}_i, \mathbf{x}_j) = \mathbb{E}_{\mathbf{w}} \mathbf{G}_{ij} = \mathbb{E}_{\mathbf{w}} \sigma(\mathbf{w}^\mathsf{T} \mathbf{x}_i) \sigma(\mathbf{w}^\mathsf{T} \mathbf{x}_j).$$

For $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$, one resorts to the integral calculus for standard Gaussian distribution in $\mathbb{R}^p$, which can be further reduced to a double integral as shown in (Williams, 1997; Louart et al., 2018) and results in the expressions in Table 1.

Then, from the discussion in Section 2, we have the following expansions for $\mathbf{x}_i \in \mathcal{C}_a$, $\mathbf{x}_j \in \mathcal{C}_b$, with $i \neq j$,

$$\mathbf{x}_i^\mathsf{T} \mathbf{x}_j = \underbrace{\boldsymbol{\omega}_i^\mathsf{T} \boldsymbol{\omega}_j}_{O(p^{-1/2})} + \underbrace{\boldsymbol{\mu}_a^\mathsf{T} \boldsymbol{\mu}_b / p + \boldsymbol{\mu}_a^\mathsf{T} \boldsymbol{\omega}_j / \sqrt{p} + \boldsymbol{\mu}_b^\mathsf{T} \boldsymbol{\omega}_i / \sqrt{p}}_{O(p^{-1})} \tag{3}$$

$$\|\mathbf{x}_i\|^2 = \underbrace{\tau}_{O(1)} + \underbrace{\operatorname{tr}(\mathbf{C}_a^\circ)/p + \|\boldsymbol{\omega}_i\|^2 - \operatorname{tr}(\mathbf{C}_a)/p}_{O(p^{-1/2})} + \underbrace{\|\boldsymbol{\mu}_a\|^2/p + 2\boldsymbol{\mu}_a^\mathsf{T} \boldsymbol{\omega}_i / \sqrt{p}}_{O(p^{-1})} \tag{4}$$

which further allows one to *linearize* the nonlinear function of $\mathbf{x}_i, \mathbf{x}_j$ in Table 1 via a Taylor expansion to obtain an *entry-wise* approximation of the key matrix $\mathbf{\Phi}$.

Nonetheless, this entry-wise approximation does not ensure a vanishing difference in terms of operator norm in the large p, T regime under consideration. Taking the popular Marčenko–Pastur law (Marčenko & Pastur, 1967) for example: consider a random matrix $\mathbf{W} \in \mathbb{R}^{n \times p}$ with i.i.d. standard Gaussian entries. Then, as $n, p \to \infty$ with $\frac{n}{p} \to c \in (0, \infty)$, entry-wisely we have that the entry (i, j) of the matrix $\frac{1}{p} \mathbf{W} \mathbf{W}^\mathsf{T}$ converge to 1 if $i = j$ and 0 otherwise, meaning that the sample covariance matrix $\frac{1}{p} \mathbf{W} \mathbf{W}^\mathsf{T}$ seemingly "converge to" an identity matrix (which is indeed the true population covariance). But it is well known (Marčenko & Pastur, 1967) that the eigenvalue distribution of $\frac{1}{p} \mathbf{W} \mathbf{W}^\mathsf{T}$ converges (almost surely so) to a continuous measure (the popular Marčenko–Pastur distribution) compactly supported on $[(1 - \sqrt{c})^2, (1 + \sqrt{c})^2]$, which is evidently different from the eigenvalue distribution $\delta_{x=1}$ of $\mathbf{I}_n$.

As a consequence, a more careful control of the entry-wise expansion of $\mathbf{\Phi}$ must be performed to ensure a vanishing expansion error in terms of operator norm. To this end, we follow the previous work of (El Karoui et al., 2010; Couillet et al., 2016) and consider the full matrix contribution.

In the following we proceed the aforementioned manipulations on the ReLU function as an example, derivations of other functions follow the same procedure and are thus omitted.

Proof of $\sigma(t) = \operatorname{ReLU}(t)$. We start with the computation of $\mathbf{\Phi}(\mathbf{a}, \mathbf{b})$. For $\sigma(t) = \operatorname{ReLU}(t)$, with the classical Gram-Schmidt process we obtain

$$\begin{aligned}
\mathbf{\Phi}(\mathbf{a}, \mathbf{b}) = \mathbb{E}_{\mathbf{w}} \sigma(\mathbf{w}^\mathsf{T} \mathbf{a}) \sigma(\mathbf{w}^\mathsf{T} \mathbf{b}) &= (2\pi)^{-\frac{p}{2}} \int_{\mathbb{R}^p} \sigma(\mathbf{w}^\mathsf{T} \mathbf{a}) \sigma(\mathbf{w}^\mathsf{T} \mathbf{b}) e^{-\frac{1}{2}\|\mathbf{w}\|^2} d\,\mathbf{w} \\
&= \frac{1}{2\pi} \int_{\mathbb{R}} \int_{\mathbb{R}} \sigma(\tilde{w}_1 \tilde{a}_1) \sigma(\tilde{w}_1 \tilde{b}_1 + \tilde{w}_2 \tilde{b}_2) e^{-\frac{1}{2}(\tilde{w}_1^2 + \tilde{w}_2^2)} d\,\tilde{w}_1 d\,\tilde{w}_2 \\
&= \frac{1}{2\pi} \int_{\mathbb{R}^2} \sigma(\tilde{\mathbf{w}}^\mathsf{T} \tilde{\mathbf{a}}) \sigma(\tilde{\mathbf{w}}^\mathsf{T} \tilde{\mathbf{b}}) e^{-\frac{1}{2}\|\tilde{\mathbf{w}}\|^2} d\,\tilde{\mathbf{w}} \\
&= \frac{1}{2\pi} \int_{\min(\tilde{\mathbf{w}}^\mathsf{T} \tilde{\mathbf{a}}, \tilde{\mathbf{w}}^\mathsf{T} \tilde{\mathbf{b}}) \geq 0} \tilde{\mathbf{w}}^\mathsf{T} \tilde{\mathbf{a}} \cdot \tilde{\mathbf{w}}^\mathsf{T} \tilde{\mathbf{b}} \cdot e^{-\frac{1}{2}\|\tilde{\mathbf{w}}\|^2} d\,\tilde{\mathbf{w}}
\end{aligned}$$

where $\tilde{a}_1 = \|\mathbf{a}\|$, $\tilde{b}_1 = \frac{\mathbf{a}^\mathsf{T} \mathbf{b}}{\|\mathbf{a}\|}$, $\tilde{b}_2 = \|\mathbf{b}\| \sqrt{1 - \frac{(\mathbf{a}^\mathsf{T} \mathbf{b})^2}{\|\mathbf{a}\|^2 \|\mathbf{b}\|^2}}$ and we denote $\tilde{\mathbf{w}} = [\tilde{w}_1, \tilde{w}_2]^\mathsf{T}$, $\tilde{\mathbf{a}} = [\tilde{a}_1, 0]^\mathsf{T}$ and $\tilde{\mathbf{b}} = [\tilde{b}_1, \tilde{b}_2]^\mathsf{T}$.

With a simple geometric representation we observe

$$\{\tilde{\mathbf{w}} \mid \min(\tilde{\mathbf{w}}^\mathsf{T}\tilde{\mathbf{a}}, \tilde{\mathbf{w}}^\mathsf{T}\tilde{\mathbf{b}}) \geq 0\} = \left\{ r\cos(\theta) + r\sin(\theta) \mid r \geq 0,\ \theta \in \left[\theta_0 - \frac{\pi}{2}, \frac{\pi}{2}\right]\right\}$$

with $\theta_0 \equiv \arccos\left(\frac{\tilde{b}_1}{\|\mathbf{b}\|}\right) = \frac{\pi}{2} - \arcsin\left(\frac{\tilde{b}_1}{\|\mathbf{b}\|}\right)$. Therefore with a polar coordinate change of variable we deduce, for $\sigma(t) = \mathrm{ReLU}(t)$ that

$$\begin{aligned}\mathbf{\Phi}(\mathbf{a}, \mathbf{b}) &= \tilde{a}_1 \frac{1}{2\pi} \int_{\theta_0 - \frac{\pi}{2}}^{\frac{\pi}{2}} \cos(\theta)\left(\tilde{b}_1 \cos(\theta) + \tilde{b}_2 \sin(\theta)\right) d\,\theta \int_{\mathbb{R}^+} r^3 e^{-\frac{1}{2}r^2} d\,r \\ &= \frac{1}{2\pi}\|\mathbf{a}\|\|\mathbf{b}\|\left(\sqrt{1 - \angle(\mathbf{a},\mathbf{b})^2} + \angle(\mathbf{a},\mathbf{b})\arccos\left(-\angle(\mathbf{a},\mathbf{b})\right)\right)\end{aligned}$$

with $\angle(\mathbf{a}, \mathbf{b}) \equiv \frac{\mathbf{a}^\mathsf{T}\mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$ as Table 1.

For a second step, with the expressions in (3) and (4) we perform a Taylor expansion to get

$$\angle(\mathbf{x}_i, \mathbf{x}_j) = \underbrace{\frac{1}{\tau}\boldsymbol{\omega}_i^\mathsf{T}\boldsymbol{\omega}_j}_{O(p^{-1/2})} + \underbrace{\frac{1}{\tau}\left(\frac{1}{p}\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\mu}_b + \frac{1}{\sqrt{p}}\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\omega}_j + \frac{1}{\sqrt{p}}\boldsymbol{\mu}_b^\mathsf{T}\boldsymbol{\omega}_i\right) - \frac{\boldsymbol{\omega}_i^\mathsf{T}\boldsymbol{\omega}_j}{2\tau^2}\left(\frac{1}{p}\operatorname{tr}\mathbf{C}_a^\circ + \boldsymbol{\phi}_i + \frac{1}{p}\operatorname{tr}\mathbf{C}_b^\circ + \boldsymbol{\phi}_j\right)}_{O(p^{-1})} + O(p^{-\frac{3}{2}})$$

where we recall $\boldsymbol{\phi}_i = \|\boldsymbol{\omega}_i\|^2 - \mathbb{E}\|\boldsymbol{\omega}_i\|^2 = \|\boldsymbol{\omega}_i\|^2 - \operatorname{tr}(\mathbf{C}_a)/p$ that is of order $O(p^{-1/2})$. Note that the third term $\boldsymbol{\omega}_i^\mathsf{T}\boldsymbol{\omega}_j\left(\operatorname{tr}\mathbf{C}_a^\circ/p + \boldsymbol{\phi}_i + \operatorname{tr}\mathbf{C}_b^\circ/p + \boldsymbol{\phi}_j\right)$, being of order $O(p^{-1})$, gives rise to a matrix of vanishing operator norm (Couillet et al., 2016), we thus conclude by stating that

$$\angle(\mathbf{x}_i, \mathbf{x}_j) = \underbrace{\frac{1}{\tau}\boldsymbol{\omega}_i^\mathsf{T}\boldsymbol{\omega}_j}_{O(p^{-1/2})} + \underbrace{\frac{1}{\tau}\left(\frac{1}{p}\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\mu}_b + \frac{1}{\sqrt{p}}\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\omega}_j + \frac{1}{\sqrt{p}}\boldsymbol{\mu}_b^\mathsf{T}\boldsymbol{\omega}_i\right)}_{O(p^{-1})} + O(p^{-\frac{3}{2}}).$$

Since $\angle(\mathbf{x}_i, \mathbf{x}_j) = O(p^{-\frac{1}{2}})$ we sequentially Taylor-expand $\sqrt{1 - \angle(\mathbf{x}_i, \mathbf{x}_j)^2}$ and $\arccos(-\angle(\mathbf{x}_i, \mathbf{x}_j))$ to the order of $O(p^{-\frac{3}{2}})$ as

$$\begin{aligned}\sqrt{1 - \angle(\mathbf{x}_i, \mathbf{x}_j)^2} &= 1 - \frac{1}{2}\angle(\mathbf{x}_i, \mathbf{x}_j)^2 + O(p^{-\frac{3}{2}}) \\ \arccos(-\angle(\mathbf{x}_i, \mathbf{x}_j)) &= \frac{\pi}{2} + \angle(\mathbf{x}_i, \mathbf{x}_j) + O(p^{-\frac{3}{2}}).\end{aligned}$$

As such, we conclude for $\sigma(t) = \mathrm{ReLU}(t)$ that

$$\begin{aligned}\mathbf{\Phi}(\mathbf{x}_i, \mathbf{x}_j) &= \frac{1}{2\pi}\|\mathbf{x}_i\|\|\mathbf{x}_j\|\left(\sqrt{1 - \angle(\mathbf{x}_i, \mathbf{x}_j)^2} + \angle(\mathbf{x}_i, \mathbf{x}_j)\arccos\left(-\angle(\mathbf{x}_i, \mathbf{x}_j)\right)\right) \\ &= \frac{1}{2\pi}\|\mathbf{x}_i\|\|\mathbf{x}_j\|\left(1 + \frac{\pi}{2}\angle(\mathbf{x}_i, \mathbf{x}_j) + \frac{1}{2}\angle(\mathbf{x}_i, \mathbf{x}_j)^2\right) + O(p^{-\frac{3}{2}}) \\ &= \frac{1}{4}\mathbf{x}_i^\mathsf{T}\mathbf{x}_j + \frac{1}{2\pi}\|\mathbf{x}_i\|\|\mathbf{x}_j\| + \frac{1}{4\pi}\mathbf{x}_i^\mathsf{T}\mathbf{x}_j\angle(\mathbf{x}_i, \mathbf{x}_j) + O(p^{-\frac{3}{2}}).\end{aligned}$$

Consequently we get the generic form for all functions $\sigma(\cdot)$ listed in Table 1 is given by

$$\begin{aligned}\mathbf{\Phi}_{i,j} \equiv \mathbf{\Phi}(\mathbf{x}_i, \mathbf{x}_j) = &\underbrace{c_0}_{O(1)} + \underbrace{c_1\left(\mathbf{t}_a/\sqrt{p} + \boldsymbol{\phi}_i + \mathbf{t}_b/\sqrt{p} + \boldsymbol{\phi}_j\right) + c_2\boldsymbol{\omega}_i^\mathsf{T}\boldsymbol{\omega}_j}_{O(p^{-1/2})} \\ &+ \underbrace{c_3\left(\mathbf{t}_a/\sqrt{p} + \boldsymbol{\phi}_i\right)\left(\mathbf{t}_b/\sqrt{p} + \boldsymbol{\phi}_j\right) + c_4\left((\mathbf{t}_a/\sqrt{p} + \boldsymbol{\phi}_i)^2 + (\mathbf{t}_b/\sqrt{p} + \boldsymbol{\phi}_j)^2\right) + c_5(\boldsymbol{\omega}_i^\mathsf{T}\boldsymbol{\omega}_j)^2}_{O(p^{-1})} \\ &+ \underbrace{c_7\left((\|\boldsymbol{\mu}_a\|^2 + \|\boldsymbol{\mu}_b\|^2)/p + 2(\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\omega}_i + \boldsymbol{\mu}_b^\mathsf{T}\boldsymbol{\omega}_j)/\sqrt{p}\right) + c_8\left(\boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\mu}_b/p + \boldsymbol{\mu}_a^\mathsf{T}\boldsymbol{\omega}_j/\sqrt{p} + \boldsymbol{\mu}_b^\mathsf{T}\boldsymbol{\omega}_i/\sqrt{p}\right)}_{O(p^{-1})} + O(p^{-\frac{3}{2}})\end{aligned}$$

where we recall the definition $\mathbf{t} \equiv \{\operatorname{tr} \mathbf{C}_a^\circ / \sqrt{p}\}_{a=1}^K$. In particular, we have for the ReLU nonlinearity $c_0 = \frac{\tau}{2\pi}$, $c_1 = c_7 = \frac{1}{4\pi}$, $c_2 = c_8 = \frac{1}{4}$, $c_3 = \frac{1}{8\pi\tau}$, $c_4 = -\frac{1}{16\pi\tau}$, $c_5 = \frac{1}{4\pi\tau}$ and $c_6 = 0$.

We then observe that, for all functions $\sigma(\cdot)$ listed in Table 1, we have $c_7 = c_1$ and $c_8 = c_2$. Besides, using the fact that

$$\left(\boldsymbol{\omega}_i^\top \boldsymbol{\omega}_j\right)^2 = \operatorname{tr}(\mathbf{C}_a \mathbf{C}_b)/p^2 + O(p^{-\frac{3}{2}})$$

and considering also the diagonal terms (with $i = j$) by adding the term $c_9 \mathbf{I}_T$, we finally get

$$\begin{aligned}
\boldsymbol{\Phi} &= c_0 \mathbf{1}_T \mathbf{1}_T^\top + c_1 \left(\boldsymbol{\phi} \mathbf{1}_T^\top + \mathbf{1}_T \boldsymbol{\phi}^\top + \left\{ \frac{\mathbf{t}_a \mathbf{1}_{T_a}}{\sqrt{p}} \right\}_{a=1}^K \mathbf{1}_T^\top + \mathbf{1}_T \left\{ \frac{\mathbf{t}_b \mathbf{1}_{T_b}^\top}{\sqrt{p}} \right\}_{b=1}^K \right) + c_2 \boldsymbol{\Omega}^\top \boldsymbol{\Omega} \\
&+ c_3 \left(\boldsymbol{\phi}\boldsymbol{\phi}^\top + \boldsymbol{\phi} \left\{ \frac{\mathbf{t}_b \mathbf{1}_{T_b}^\top}{\sqrt{p}} \right\}_{b=1}^K + \left\{ \frac{\mathbf{t}_a \mathbf{1}_{T_a}}{\sqrt{p}} \right\}_{a=1}^K \boldsymbol{\phi}^\top + \left\{ \mathbf{t}_a \mathbf{t}_b \frac{\mathbf{1}_{T_a} \mathbf{1}_{T_b}^\top}{p} \right\}_{a,b=1}^K \right) \\
&+ c_4 \left((\boldsymbol{\phi}^2) \mathbf{1}_T^\top + \mathbf{1}_T (\boldsymbol{\phi}^2)^\top + 2 \left(\mathcal{D} \left\{ \mathbf{t}_a \mathbf{1}_{T_a} \right\}_{a=1}^K \boldsymbol{\phi} \frac{\mathbf{1}_T^\top}{\sqrt{p}} \right) + 2 \left(\frac{\mathbf{1}_T}{\sqrt{p}} \boldsymbol{\phi}^\top \mathcal{D} \left\{ \mathbf{t}_b \mathbf{1}_{T_b} \right\}_{b=1}^K \right) \right. \\
&\left. + \left\{ \mathbf{t}_a^2 \frac{\mathbf{1}_{T_a}}{p} \right\}_{a=1}^K \mathbf{1}_T^\top + \mathbf{1}_T \left\{ \mathbf{t}_b^2 \frac{\mathbf{1}_{T_b}^\top}{p} \right\}_{b=1}^K \right) + c_5 \left\{ \operatorname{tr}(\mathbf{C}_a \mathbf{C}_b) \frac{\mathbf{1}_{T_a} \mathbf{1}_{T_b}^\top}{p^2} \right\}_{a,b=1}^K \\
&+ c_1 \left(\left\{ \|\boldsymbol{\mu}_a\|^2 \frac{\mathbf{1}_{T_a}}{p} \right\}_{a=1}^K \mathbf{1}_T^\top + \mathbf{1}_T \left\{ \|\boldsymbol{\mu}_b\|^2 \frac{\mathbf{1}_{T_b}^\top}{p} \right\}_{b=1}^K + \frac{2}{\sqrt{p}} \left\{ \boldsymbol{\Omega}_a^\top \boldsymbol{\mu}_a \mathbf{1}_{T_a}^\top \right\}_{a=1}^K + \frac{2}{\sqrt{p}} \left\{ \mathbf{1}_{T_b} \boldsymbol{\mu}_b^\top \boldsymbol{\Omega}_b \right\}_{b=1}^K \right) \\
&+ c_2 \left(\frac{1}{p} \left\{ \mathbf{1}_{T_a} \boldsymbol{\mu}_a^\top \boldsymbol{\mu}_b \mathbf{1}_{T_b}^\top \right\}_{a,b=1}^K + \frac{1}{\sqrt{p}} \left\{ \boldsymbol{\Omega}_a^\top \boldsymbol{\mu}_b \mathbf{1}_{T_b}^\top + \mathbf{1}_{T_a} \boldsymbol{\mu}_a^\top \boldsymbol{\Omega}_b \right\}_{a,b=1}^K \right) + c_9 \mathbf{I}_T + O_{\|\cdot\|}(p^{-\frac{1}{2}})
\end{aligned}$$

where we denote $\boldsymbol{\phi}^2 \equiv [\boldsymbol{\phi}_1^2, \ldots, \boldsymbol{\phi}_T^2]^\top$ and $O_{\|\cdot\|}(p^{-\frac{1}{2}})$ a matrix of operator norm of order $O(p^{-1/2})$ as $p \to \infty$.

Recalling that for $\mathbf{P} \equiv \mathbf{I}_T - \frac{1}{T} \mathbf{1}_T \mathbf{1}_T^\top$, we have $\mathbf{P}\mathbf{1}_T = \mathbf{1}_T \mathbf{P} = \mathbf{0}$ and therefore

$$\begin{aligned}
\boldsymbol{\Phi}_c \equiv \mathbf{P}\boldsymbol{\Phi}\mathbf{P} &= c_2 \mathbf{P} \boldsymbol{\Omega}^\top \boldsymbol{\Omega} \mathbf{P} + c_3 \mathbf{P} \left(\boldsymbol{\phi}\boldsymbol{\phi}^\top + \boldsymbol{\phi} \left\{ \frac{\mathbf{t}_b \mathbf{1}_{T_b}^\top}{\sqrt{p}} \right\}_{b=1}^K + \left\{ \frac{\mathbf{t}_a \mathbf{1}_{T_a}}{\sqrt{p}} \right\}_{a=1}^K \boldsymbol{\phi}^\top + \left\{ \mathbf{t}_a \mathbf{t}_b \frac{\mathbf{1}_{T_a} \mathbf{1}_{T_b}^\top}{p} \right\}_{a,b=1}^K \right) \mathbf{P} \\
&+ c_5 \mathbf{P} \left\{ \operatorname{tr}(\mathbf{C}_a \mathbf{C}_b) \frac{\mathbf{1}_{T_a} \mathbf{1}_{T_b}^\top}{p^2} \right\}_{a,b=1}^K \mathbf{P} + c_2 \mathbf{P} \left(\frac{1}{p} \left\{ \mathbf{1}_{T_a} \boldsymbol{\mu}_a^\top \boldsymbol{\mu}_b \mathbf{1}_{T_b}^\top \right\}_{a,b=1}^K + \frac{1}{\sqrt{p}} \left\{ \boldsymbol{\Omega}_a^\top \boldsymbol{\mu}_b \mathbf{1}_{T_b}^\top + \mathbf{1}_{T_a} \boldsymbol{\mu}_a^\top \boldsymbol{\Omega}_b \right\}_{a,b=1}^K \right) \mathbf{P} \\
&+ c_7 \mathbf{P} + O_{\|\cdot\|}(p^{-1/2}) \equiv \mathbf{P} \tilde{\boldsymbol{\Phi}} \mathbf{P} + O_{\|\cdot\|}(p^{-1/2}).
\end{aligned}$$

We further observe that, for all functions $\sigma(\cdot)$ listed in Table 1 we have $c_5 = 2c_3$ and let $d_0 = c_7,\ d_1 = c_2,\ d_2 = c_3 = c_5/2$ we obtain the expression of $\tilde{\boldsymbol{\Phi}}$ in Theorem 1, which concludes the proof. □