Supplementary Material

The Dynamics of Learning: A Random Matrix Approach

A. Proofs

A.1. Proofs of Theorem 1 and 2

Proof. We start with the proof of Theorem 1, since

$$\begin{aligned}\boldsymbol{\mu}^\mathsf{T}\mathbf{w}(t) &= \boldsymbol{\mu}^\mathsf{T} e^{-\frac{\alpha t}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}\mathbf{w}_0 + \boldsymbol{\mu}^\mathsf{T}\left(\mathbf{I}_p - e^{-\frac{\alpha t}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}\right)\mathbf{w}_{LS} \\ &= -\frac{1}{2\pi i}\oint_\gamma f_t(z)\boldsymbol{\mu}^\mathsf{T}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\mathbf{w}_0\,dz - \frac{1}{2\pi i}\oint_\gamma \frac{1-f_t(z)}{z}\boldsymbol{\mu}^\mathsf{T}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\frac{1}{n}\mathbf{X}\mathbf{y}\,dz\end{aligned}$$

with $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} = \frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} + \begin{bmatrix}\boldsymbol{\mu} & \frac{1}{n}\mathbf{Z}\mathbf{y}\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}^\mathsf{T}\\ \frac{1}{n}\mathbf{y}^\mathsf{T}\mathbf{Z}^\mathsf{T}\end{bmatrix}$ and therefore

$$\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p\right)^{-1} = \mathbf{Q}(z) - \mathbf{Q}(z)\begin{bmatrix}\boldsymbol{\mu} & \frac{1}{n}\mathbf{Z}\mathbf{y}\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}^\mathsf{T}\mathbf{Q}(z)\boldsymbol{\mu} & 1+\frac{1}{n}\boldsymbol{\mu}^\mathsf{T}\mathbf{Q}(z)\mathbf{Z}\mathbf{y}\\ 1+\frac{1}{n}\boldsymbol{\mu}^\mathsf{T}\mathbf{Q}(z)\mathbf{Z}\mathbf{y} & -1+\frac{1}{n}\mathbf{y}^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Q}(z)\frac{1}{n}\mathbf{Z}\mathbf{y}\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{\mu}^\mathsf{T}\\ \frac{1}{n}\mathbf{y}^\mathsf{T}\mathbf{Z}^\mathsf{T}\end{bmatrix}\mathbf{Q}(z).$$

We thus resort to the computation of the bilinear form $\mathbf{a}^\mathsf{T}\mathbf{Q}(z)\mathbf{b}$, for which we plug-in the deterministic equivalent of $\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_p$ to obtain the following estimations

$$\begin{aligned}\boldsymbol{\mu}^\mathsf{T}\mathbf{Q}(z)\boldsymbol{\mu} &= \|\boldsymbol{\mu}\|^2 m(z)\\ \frac{1}{n}\boldsymbol{\mu}^\mathsf{T}\mathbf{Q}(z)\mathbf{Z}\mathbf{y} &= o(1)\\ \frac{1}{n^2}\mathbf{y}^\mathsf{T}\mathbf{Z}^\mathsf{T}\mathbf{Q}(z)\mathbf{Z}\mathbf{y} &= \frac{1}{n^2}\mathbf{y}^\mathsf{T}\tilde{\mathbf{Q}}(z)\mathbf{Z}^\mathsf{T}\mathbf{Z}\mathbf{y} = \frac{1}{n}\mathbf{y}^\mathsf{T}\tilde{\mathbf{Q}}(z)\left(\frac{1}{n}\mathbf{Z}^\mathsf{T}\mathbf{Z} - z\mathbf{I}_n + z\mathbf{I}_n\right)\mathbf{y}\\ &= \frac{1}{n}\|\mathbf{y}\|^2 + z\frac{1}{n}\mathbf{y}^\mathsf{T}\tilde{\mathbf{Q}}(z)\mathbf{y} = 1 + z\frac{1}{n}\operatorname{tr}\tilde{\mathbf{Q}}(z) = 1 + z\tilde{m}(z)\end{aligned}$$

with the *co-resolvent* $\tilde{\mathbf{Q}}(z) = \left(\frac{1}{n}\mathbf{Z}^\mathsf{T}\mathbf{Z} - z\mathbf{I}_n\right)^{-1}$, $m(z)$ the *unique* solution of the Marčenko–Pastur equation (2) and $\tilde{m}(z) = \frac{1}{n}\operatorname{tr}\tilde{\mathbf{Q}}(z) + o(1)$ such that

$$cm(z) = \tilde{m}(z) + \frac{1}{z}(1-c)$$

which is a direct result of the fact that both $\mathbf{Z}^\mathsf{T}\mathbf{Z}$ and $\mathbf{Z}\mathbf{Z}^\mathsf{T}$ have the same eigenvalues except for the additional zeros eigenvalues for the larger matrix (which essentially depends on the sign of $1-c$).

We thus get, with the Schur complement lemma,

$$\begin{aligned}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p\right)^{-1} &= \mathbf{Q}(z) - \mathbf{Q}(z)\begin{bmatrix}\boldsymbol{\mu} & \frac{1}{n}\mathbf{Z}\mathbf{y}\end{bmatrix}\begin{bmatrix}\|\boldsymbol{\mu}\|^2 m(z) & 1\\ 1 & z\tilde{m}(z)\end{bmatrix}^{-1}\begin{bmatrix}\boldsymbol{\mu}^\mathsf{T}\\ \frac{1}{n}\mathbf{y}^\mathsf{T}\mathbf{Z}^\mathsf{T}\end{bmatrix}\mathbf{Q}(z) + o(1)\\ &= \mathbf{Q}(z) - \frac{\mathbf{Q}(z)}{z\|\boldsymbol{\mu}\|^2 m(z)\tilde{m}(z) - 1}\begin{bmatrix}\boldsymbol{\mu} & \frac{1}{n}\mathbf{Z}\mathbf{y}\end{bmatrix}\begin{bmatrix}z\tilde{m}(z) & -1\\ -1 & \|\boldsymbol{\mu}\|^2 m(z)\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}^\mathsf{T}\\ \frac{1}{n}\mathbf{y}^\mathsf{T}\mathbf{Z}^\mathsf{T}\end{bmatrix}\mathbf{Q}(z) + o(1)\end{aligned}$$

and the term $\boldsymbol{\mu}^\mathsf{T}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\frac{1}{n}\mathbf{X}\mathbf{y}$ is therefore given by

$$\begin{aligned}&\boldsymbol{\mu}^\mathsf{T}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\frac{1}{n}\mathbf{X}\mathbf{y} = \|\boldsymbol{\mu}\|^2 m(z) - \frac{\begin{bmatrix}\|\boldsymbol{\mu}\|^2 m(z) & 0\end{bmatrix}}{z\|\boldsymbol{\mu}\|^2 m(z)\tilde{m}(z) - 1}\begin{bmatrix}z\tilde{m}(z) & -1\\ -1 & \|\boldsymbol{\mu}\|^2 m(z)\end{bmatrix}\begin{bmatrix}\|\boldsymbol{\mu}\|^2 m(z)\\ 1+z\tilde{m}(z)\end{bmatrix} + o(1)\\ &= \frac{\|\boldsymbol{\mu}\|^2 m(z) z\tilde{m}(z)}{\|\boldsymbol{\mu}\|^2 m(z) z\tilde{m}(z) - 1} + o(1) = \frac{\|\boldsymbol{\mu}\|^2 (zm(z)+1)}{1 + \|\boldsymbol{\mu}\|^2(zm(z)+1)} + o(1) = \frac{\|\boldsymbol{\mu}\|^2 m(z)}{(\|\boldsymbol{\mu}\|^2 + c)m(z) + 1} + o(1)\end{aligned}$$

where we use the fact that $\tilde{m}(z) = cm(z) - \frac{1}{z}(1-c)$ and $(zm(z)+1)(cm(z)+1) = m$ from (2), while the term $\boldsymbol{\mu}^\mathsf{T}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\mathbf{w}_0 = O(n^{-\frac{1}{2}})$ due to the independence of $\mathbf{w}_0$ with respect to $\mathbf{Z}$ and can be check with a careful application of Lyapunov's central limit theorem (Billingsley, 2008).

Following the same arguments we have

$$\mathbf{w}(t)^\mathsf{T}\mathbf{w}(t) = -\frac{1}{2\pi i}\oint_\gamma f_t^2(z)\mathbf{w}_0\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\mathbf{w}_0\,dz - \frac{1}{\pi i}\oint_\gamma \frac{f_t(z)(1-f_t(z))}{z}\mathbf{w}_0\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\frac{1}{n}\mathbf{X}\mathbf{y}\,dz$$
$$-\frac{1}{2\pi i}\oint_\gamma \frac{(1-f_t(z))^2}{z^2}\frac{1}{n}\mathbf{y}^\mathsf{T}\mathbf{X}^\mathsf{T}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\frac{1}{n}\mathbf{X}\mathbf{y}\,dz$$

together with

$$\mathbf{w}_0\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\mathbf{w}_0 = \sigma^2 m(z) + o(1)$$
$$\mathbf{w}_0\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\frac{1}{n}\mathbf{X}\mathbf{y}^\mathsf{T} = o(1)$$
$$\frac{1}{n}\mathbf{y}^\mathsf{T}\mathbf{X}^\mathsf{T}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p\right)^{-1}\frac{1}{n}\mathbf{X}\mathbf{y} = 1 - \frac{1}{(\|\boldsymbol{\mu}\|^2 + c)m(z) + 1} + o(1).$$

It now remains to replace the different terms in $\boldsymbol{\mu}^\mathsf{T}\mathbf{w}(t)$ and $\mathbf{w}(t)^\mathsf{T}\mathbf{w}(t)$ by their asymptotic approximations. To this end, first note that all aforementioned approximations can be summarized as the fact that, for a generic $h(z)$, we have, as $n \to \infty$,

$$h(z) - \bar{h}(z) \to 0$$

almost surely for all z not an eigenvalue of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$. Therefore, there exists a probability one set Ω_z on which $h(z)$ is uniformly bounded for all large n, with a bound independent of z. Then by the Theorem of "no eigenvalues outside the support" (see for example (Bai & Silverstein, 1998)) we know that, with probability one, for all n, p large, no eigenvalue of $\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T}$ appears outside the interval $[\lambda_-, \lambda_+]$, where we recall $\lambda_- \equiv (1-\sqrt{c})^2$ and $\lambda_+ \equiv (1+\sqrt{c})^2$. As such, the set of intersection $\Omega = \cap_{z_i}\Omega_{z_i}$ for a finitely many z_i, is still a probability one set. Finally by Vitali convergence theorem, together with the analyticity of the function under consideration, we conclude the proof of Theorem 1. The proof of Theorem 2 follows exactly the same line of arguments and is thus omitted here. □

A.2. Detailed Derivation of (4)-(7)

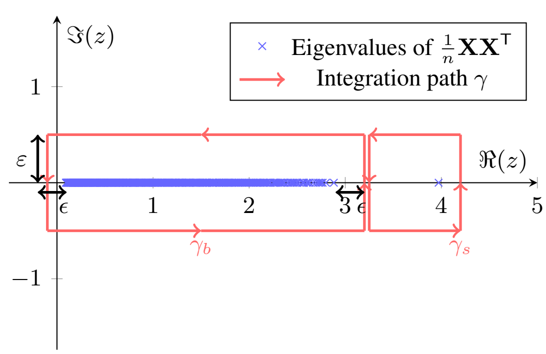

Figure 7. Eigenvalue distribution of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ for $\boldsymbol{\mu} = [1.5; \mathbf{0}_{p-1}]$, $p = 512$, $n = 1\,024$ and $c_1 = c_2 = 1/2$.

We first determine the location of the isolated eigenvalue λ (as shown in Figure 2). More concretely, we would like to find λ an eigenvalue of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ that lies outside the support of Marčenko–Pastur distribution (in fact, not an eigenvalue of $\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T}$).

Solving the following equation for $\lambda \in \mathbb{R}$,

$$
\begin{aligned}
&\det\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - \lambda \mathbf{I}_p\right) = 0 \\
&\Leftrightarrow \det\left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} - \lambda \mathbf{I}_p + \begin{bmatrix}\boldsymbol{\mu} & \frac{1}{n}\mathbf{Z}\mathbf{y}\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}^\mathsf{T} \\ \frac{1}{n}\mathbf{y}^\mathsf{T}\mathbf{Z}^\mathsf{T}\end{bmatrix}\right) = 0 \\
&\Leftrightarrow \det\left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} - \lambda \mathbf{I}_p\right)\det\left(\mathbf{I}_p + \mathbf{Q}(\lambda)\begin{bmatrix}\boldsymbol{\mu} & \frac{1}{n}\mathbf{Z}\mathbf{y}\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}^\mathsf{T} \\ \frac{1}{n}\mathbf{y}^\mathsf{T}\mathbf{Z}^\mathsf{T}\end{bmatrix}\right) = 0 \\
&\Leftrightarrow \det\left(\mathbf{I}_2 + \begin{bmatrix}1 & 1\\ 1 & 0\end{bmatrix}\begin{bmatrix}\boldsymbol{\mu}^\mathsf{T} \\ \frac{1}{n}\mathbf{y}^\mathsf{T}\mathbf{Z}^\mathsf{T}\end{bmatrix}\mathbf{Q}(\lambda)\begin{bmatrix}\boldsymbol{\mu} & \frac{1}{n}\mathbf{Z}\mathbf{y}\end{bmatrix}\right) = 0 \\
&\Leftrightarrow \det\begin{bmatrix}\|\boldsymbol{\mu}\|^2 m(\lambda) + 1 & 1 + z\tilde{m}(\lambda) \\ \|\boldsymbol{\mu}\|^2 m(\lambda) & 1\end{bmatrix} + o(1) = 0 \\
&\Leftrightarrow 1 + (\|\boldsymbol{\mu}\|^2 + c)m(\lambda) + o(1) = 0
\end{aligned}
$$

where we recall the definition $\mathbf{Q}(\lambda) \equiv \left(\frac{1}{n}\mathbf{Z}\mathbf{Z}^\mathsf{T} - \lambda\mathbf{I}_p\right)^{-1}$ and use the fact that $\det(\mathbf{A}\mathbf{B}) = \det(\mathbf{A})\det(\mathbf{B})$ as well as the Sylvester's determinant identity $\det(\mathbf{I}_p + \mathbf{A}\mathbf{B}) = \det(\mathbf{I}_n + \mathbf{B}\mathbf{A})$ for $\mathbf{A}, \mathbf{B}$ of appropriate dimension. Together with (2) we deduce the (empirical) isolated eigenvalue $\lambda = \lambda_s + o(1)$ with

$$
\lambda_s = c + 1 + \|\boldsymbol{\mu}\|^2 + \frac{c}{\|\boldsymbol{\mu}\|^2}
$$

which in fact gives the asymptotic location of the isolated eigenvalue as $n \to \infty$. In the following, we may thus use λ_s instead of λ throughout the computation. By splitting the path γ into $\gamma_b + \gamma_s$ that circles respectively around the main bulk between $[\lambda_-, \lambda_+]$ and the isolated eigenvalue λ_s, we easily deduce, with the residual theorem that $E = E_{\gamma_b} + E_{\gamma_s}$ with

$$
\begin{aligned}
E_{\gamma_s} &= -\frac{1}{2\pi i}\oint_{\gamma_s}\frac{1 - f_t(z)}{z}\frac{\|\boldsymbol{\mu}\|^2 m(z)}{1 + (\|\boldsymbol{\mu}\|^2 + c)m(z)}\,dz = -\mathrm{Res}\frac{1 - f_t(z)}{z}\frac{\|\boldsymbol{\mu}\|^2 m(z)}{1 + (\|\boldsymbol{\mu}\|^2 + c)m(z)} \\
&= -\lim_{z\to\lambda_s}(z - \lambda_s)\frac{1 - f_t(z)}{z}\frac{\|\boldsymbol{\mu}\|^2 m(z)}{1 + (\|\boldsymbol{\mu}\|^2 + c)m(z)} = -\frac{1 - f_t(\lambda_s)}{\lambda_s}\frac{\|\boldsymbol{\mu}\|^2 m(\lambda_s)}{(\|\boldsymbol{\mu}\|^2 + c)m'(\lambda_s)} \\
&= -\frac{\|\boldsymbol{\mu}\|^2}{\|\boldsymbol{\mu}\|^2 + c}\frac{1 - f_t(\lambda_s)}{\lambda_s}\frac{1 - c - \lambda_s - 2c\lambda_s m(\lambda_s)}{cm(\lambda_s) + 1} = \left(\|\boldsymbol{\mu}\|^2 - \frac{c}{\|\boldsymbol{\mu}\|^2}\right)\frac{1 - f_t(\lambda_s)}{\lambda_s}
\end{aligned}
\tag{11}
$$

with $m'(z)$ the derivative of $m(z)$ with respect to z and is obtained by taking the derivative of (2).

We now move on to handle the contour integration γ_b in the computation of E_{γ_b}. We follow the idea in (Bai & Silverstein, 2008) and choose the contour γ_b to be a rectangle with sides parallel to the axes, intersecting the real axis at 0 and λ_+ (in fact at $-\epsilon$ and $\lambda_+ + \epsilon$ so that the functions under consideration remain analytic) and the horizontal sides being a distance $\varepsilon \to 0$ away from the real axis. Since for nonzero $x \in \mathbb{R}$, the limit $\lim_{z\in\mathbb{Z}\to x} m(z) \equiv \check{m}(x)$ exists (Silverstein & Choi, 1995) and is given by

$$
\check{m}(x) = \frac{1 - c - x}{2cx} \pm \frac{i}{2cx}\sqrt{4cx - (1 - c - x)^2} = \frac{1 - c - x}{2cx} \pm \frac{i}{2cx}\sqrt{(x - \lambda_-)(\lambda_+ - x)}
$$

with the branch of $\pm$ is determined by the imaginary part of z such that $\Im(z)\cdot\Im m(z) > 0$ and we recall $\lambda_- \equiv (1 - \sqrt{c})^2$ and $\lambda_+ \equiv (1 + \sqrt{c})^2$. For simplicity we denote

$$
\Re\check{m} = \frac{1 - c - x}{2cx},\ \Im\check{m} = \frac{1}{2cx}\sqrt{(x - \lambda_-)(\lambda_+ - x)}
$$

and therefore

$$
\begin{aligned}
E_{\gamma_b} &= -\frac{1}{2\pi i}\oint_{\gamma_b}\frac{1 - f_t(z)}{z}\frac{\|\boldsymbol{\mu}\|^2 m(z)}{1 + (\|\boldsymbol{\mu}\|^2 + c)m(z)}\,dz \\
&= -\frac{\|\boldsymbol{\mu}\|^2}{\pi i}\int_{\lambda_-}^{\lambda_+}\frac{1 - f_t(x)}{x}\Im\left[\frac{\Re\check{m} - i\Im\check{m}}{1 + (\|\boldsymbol{\mu}\|^2 + c)(\Re\check{m} - i\Im\check{m})}\right]\,dx \\
&= -\frac{\|\boldsymbol{\mu}\|^2}{\pi i}\int_{\lambda_-}^{\lambda_+}\frac{1 - f_t(x)}{x}\Im\left[\frac{\Re\check{m} + \frac{\|\boldsymbol{\mu}\|^2 + c}{cx} - i\Im\check{m}}{1 + 2(\|\boldsymbol{\mu}\|^2 + c)\Re\check{m} + \frac{(\|\boldsymbol{\mu}\|^2 + c)^2}{cx}}\right]\,dx
\end{aligned}
$$

with $z = x \pm i\varepsilon$ and $\varepsilon \to 0$ (on different sides of the real axis) and the fact that $(\Re \check{m})^2 + (\Im \check{m})^2 = \frac{1}{cx}$. We take the imaginary part and result in

$$E_{\gamma_b} = \frac{\|\boldsymbol{\mu}\|^2}{\pi} \int_{\lambda_-}^{\lambda_+} \frac{1 - f_t(x)}{x} \frac{\Im \check{m}}{1 + 2(\|\boldsymbol{\mu}\|^2 + c)\Re \check{m} + \frac{(\|\boldsymbol{\mu}\|^2 + c)^2}{cx}} \, dx = \frac{1}{2\pi} \int_{\lambda_-}^{\lambda_+} \frac{1 - f_t(x)}{x} \frac{\sqrt{4cx - (1 - c - x)^2}}{\lambda_s - x} \, dx \tag{12}$$

where we recall the definition $\lambda_s \equiv c + 1 + \|\boldsymbol{\mu}\|^2 + \frac{c}{\|\boldsymbol{\mu}\|^2}$. Ultimately we assemble (11) and (12) to get the expression in (4). The derivations of (5)-(7) follow the same arguments and are thus omitted here.