High-dimensional Learning Dynamics of Deep Neural Nets in the Neural Tangent Regime

Yongqi Du YONGQI_DU@HUST.EDU.CN
Zenan Ling LINGZENAN@HUST.EDU.CN
Robert C. Qiu CAIMING@HUST.EDU.CN
Zhenyu Liao ZHENYU_LIAO@HUST.EDU.CN
Huazhong University of Science and Technology, Wuhan, China

Abstract

In this paper, built upon recent advances in the high-dimensional characterization of neural tangent kernels (NTKs) with random matrix techniques [5], we derive *precise* high-dimensional training dynamics for wide and deep neural networks (DNNs). By assuming a two-class Gaussian mixture model for the input data, *exact* expression of the training mean squared error (MSE) is derived, as a function of the dimension p, sample size n, and the statistics of input data, the depth L, as well as the nonlinear activation function in each layer of the network. The theoretical results provide novel insight into the inner mechanism of DNNs. Numerical experiments on not-so-wide networks are provided to validate the proposed asymptotic results.

1. Introduction

Large-scale machine learning models such as deep neural networks (DNNs) have made remarkable progress over the past decade, with a long list of successful applications ranging from computer vision [11], game [17], to natural language processing [20] and AI-generated content [7].

Despite the notable empirical success of DNNs, our theoretical understanding of them is progressing at a more modest pace. As a telling example, it still remains unclear why over-parameterized DNN models generally do *not* overfit, when trained on simple first-order methods such as the (stochastic) gradient descent [1, 13, 15]. The neural tangent kernel (NTK) [10], in this respect, provides a powerful tool in assessing the convergence and generalization properties of *deep and wide* neural networks, via a study of the associated neural tangent kernel functional space.

Yet, the practical computation and/or the theoretical assessment of NTKs remains *not easily accessible*, due to their mathematically involved and implicit form dependent on the input data and network architecture. This further poses technical challenges in applying NTK in the theoretical assessment of DNN models. These issues have been partially resolved in our recent paper [5], by providing, for Gaussian mixture data, precise high-dimensional characterization of the NTK matrices for a fully-connected DNN model. This further allows for:

(i) efficient numerical computation of the NTK matrix with complexity *independent* of the network width or depth, and only depends on input data as well as the activation in each layer through an iterative equation with a few parameters; and

(ii) theoretical assessment of, e.g., how different choice of activation functions affects the NTK, and the consequence of the convergence and generalization of the network.

In this paper, built upon recent progress in random matrix theory [3, 12], we derive *precise* training dynamics for deep and wide neural networks, when the input data are drawn from a high-dimensional two-class Gaussian mixture model. The obtained results can be used to "predict" the training dynamics of a given (wide and deep) network, without explicit computing the corresponding NTK matrix or its high-dimensional equivalent [5]. The proposed analysis provides novel insights into the interplay between the input data distribution and the network under study (for which a *phase transition* phenomenon might occur in the NTK eigenspectrum), and extends the analysis in [12] (performed on a toy linear regression model) to the more involved DNN model.

Notations. We denote scalars in lowercase letters, vectors in bold lowercase, and matrices in bold uppercase. We denote the transpose operator by $(\cdot)^\top$, and use $\|\cdot\|$ to denote the Euclidean norm for vectors and spectral/operator norm for matrices. For a random variable z, $\mathbb{E}[z]$ denotes the expectation of z. We use $\mathbf{1}_p$ and $\mathbf{I}_p$ for the vector of all ones of dimension p and the identity matrix of dimension $p \times p$, respectively. We use $O(\cdot)$, $o(\cdot)$ notations as in standard asymptotic statistics [19].

Organization of the paper. In the remainder of this paper, we present the GMM data and fully-connected DNN model under study, together with preliminary results in NTK in Section 2. Our main results on the precise training dynamics are given in Section 3. Numerical experiments are conducted in Section 4 to validate the proposed theoretical analysis.

2. System Model and Preliminaries

2.1. System model

We consider training data $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^p$ and their labels $y_1, \ldots, y_n \in \{\pm 1\}$ *independently* and *uniformly* drawn from the following binary Gaussian mixture model (class $\mathcal{C}_1$ versus $\mathcal{C}_2$):

$$\text{class } \mathcal{C}_a\colon \quad y_i = (-1)^a, \quad \sqrt{p}\mathbf{x}_i \sim \mathcal{N}(y_i \boldsymbol{\mu}, \mathbf{I}_p), \quad a \in \{1, 2\}, \tag{1}$$

for some deterministic mean vector $\boldsymbol{\mu} \in \mathbb{R}^p$. Collecting the n data vectors into a matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n] \in \mathbb{R}^{p \times n}$ and the associated labels into a column vector $\mathbf{y} = [y_1, y_2, \ldots, y_n]^\top \in \mathbb{R}^n$, this leads to the "signal-plus-noise" model for $(\mathbf{X}, \mathbf{y})$ as $\sqrt{p}\,\mathbf{X} = \boldsymbol{\mu}\mathbf{y}^\top + \mathbf{Z}$, with random matrix $\mathbf{Z} \in \mathbb{R}^{p \times n}$ having i.i.d. standard Gaussian entries.

We focus here on the training dynamics of a deep fully-connected neural network with a scalar output. Let $\mathbf{W}_1 \in \mathbb{R}^{d_1 \times p}, \cdots, \mathbf{W}_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ be the (intermediate) weight matrices and $\sigma_1, \cdots, \sigma_\ell$ be the *nonlinear* activation function of each layer, the network first maps the input data $\mathbf{X} \in \mathbb{R}^{p \times n}$ to their representations $\boldsymbol{\Sigma}_\ell(\mathbf{X}) \in \mathbb{R}^{d_\ell \times n}$ at layer ℓ, for $\ell = 1, \ldots, L$, given by (the columns of)

$$\boldsymbol{\Sigma}_\ell \equiv \boldsymbol{\Sigma}_\ell(\mathbf{X}) = \frac{1}{\sqrt{d_\ell}} \sigma_\ell \left(\frac{1}{\sqrt{d_{\ell-1}}} \mathbf{W}_\ell \sigma_{\ell-1} \left(\cdots \frac{1}{\sqrt{d_2}} \sigma_2 \left(\frac{1}{\sqrt{d_1}} \mathbf{W}_2 \sigma_1 \left(\mathbf{W}_1 \mathbf{X} \right) \right) \right) \right), \tag{2}$$

and then to the (scalar) output $f(\mathbf{X}) = \boldsymbol{\Sigma}_L(\mathbf{X})^\top \mathbf{w} \in \mathbb{R}^n$ via the readout weight vector $\mathbf{w} \in \mathbb{R}^{d_L}$.

Our objective is to evaluate the high-dimensional dynamics of the model in (2) in minimizing the following mean squared loss

$$L(\boldsymbol{\theta}) = \frac{1}{2} \| f(\mathbf{X}) - \mathbf{y} \|^2, \quad f(\mathbf{X}) = \boldsymbol{\Sigma}_L(\mathbf{X})^\top \mathbf{w}, \tag{3}$$

for $\boldsymbol{\theta} = [\mathrm{vec}(\mathbf{W}_1), \cdots, \mathrm{vec}(\mathbf{W}_L), \mathbf{w}]$, with full-batch gradient descent, in the *infinitely wide* NTK regime [10] as $d_1, \cdots, d_L \to \infty$, or more specifically, for $d_\ell \gg \max(n, p)$, $\ell = 1, \cdots, L$.

As in [3, 5, 10], we put ourselves under the following assumptions.

Assumption 1 (High-dimensional regime) *The data dimension p and sample size n are both large and comparable, that is, as $n \to \infty$, $n/p \to c \in (0, \infty)$ and $\|\boldsymbol{\mu}\| = O(1)$.*

Since the data are uniformly drawn from the binary GMM with prior probability $1/2$ each, the cardinalities n_1, n_2 for the two classes satisfy $n_a/n \to 1/2$ almost surely as $n \to \infty$, $a \in \{1, 2\}$.

Assumption 2 (Random weights initialization) *The weights $\mathbf{W}_1, \ldots, \mathbf{W}_L$, and $\mathbf{w}$ at $t = 0$ are independent and have i.i.d. zero-mean and unit-variance entries with finite eight-order moment.*

Assumption 3 (Activation functions) *The activations $\sigma_1, \ldots, \sigma_L$ are (at least) four-times differentiable for the standard normal measure, that is, $\max_{k \in \{0,1,2,3,4\}}\{|\mathbb{E}[\sigma_\ell^{(k)}(\xi)]|\} < C$ for some constant $C > 0, \xi \sim \mathcal{N}(0, 1)$, $\ell \in \{1, \ldots, L\}$.*

2.2. Learning dynamics, NTK, CK, and their high-dimensional equivalents

The dynamics of the network output $f_t(\mathbf{X})$ in (2), when trained using gradient descent with a sufficiently small learning rate, can be well described with the NTK, in the *infinitely wide* regime (also known as the NTK regime, with $d_\ell \gg \max(n, p)$), by the following differential equation [9, 10]:

$$\begin{aligned} \partial_t f_t(\mathbf{X}) &= \partial_{\boldsymbol{\theta}_t} f_t(\mathbf{X}) \cdot \partial_t \boldsymbol{\theta}_t = \partial_{\boldsymbol{\theta}_t} f_t(\mathbf{X}) \cdot (-\eta \cdot (\partial_{\boldsymbol{\theta}_t} f_t(\mathbf{X}))^\top \cdot \partial_{f_t} L) \\ &= -\eta \cdot \partial_{\boldsymbol{\theta}_t} f_t(\mathbf{X}) \cdot (\partial_{\boldsymbol{\theta}_t} f_t(\mathbf{X}))^\top \cdot (f_t(\mathbf{X}) - \mathbf{y}) = -\eta \cdot \mathbf{K}_{\mathrm{NTK},L} \cdot (f_t(\mathbf{X}) - \mathbf{y}). \end{aligned} \tag{4}$$

The solution to (4) is explicitly given by

$$f_t(\mathbf{X}) = e^{-\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \cdot f_0(\mathbf{X}) + (\mathbf{I}_n - e^{-\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}}) \cdot \mathbf{y}, \tag{5}$$

so that the (normalized) training MSE is given by

$$E_t \equiv \frac{1}{2n}\|f_t(\mathbf{X}) - \mathbf{y}\|^2 = \frac{1}{2n}(f_0(\mathbf{X}) - \mathbf{y})^\top \cdot e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \cdot (f_0(\mathbf{X}) - \mathbf{y}), \tag{6}$$

which will be our central object to study in this work.

As we shall see below in Theorem 3, the DNN training dynamic E_t in (6) can be expressed as an *explicit* function of the conjugate kernel $\mathbf{K}_{\mathrm{CK},\ell}$ and neural tangent kernel $\mathbf{K}_{\mathrm{NTK},\ell}$ matrices, defined respectively as [10]:

$$\mathbf{K}_{\mathrm{CK},\ell} = \mathbb{E}[\boldsymbol{\Sigma}_\ell^\top \boldsymbol{\Sigma}_\ell] \in \mathbb{R}^{n \times n}, \quad \mathbf{K}_{\mathrm{NTK},\ell} = \mathbf{K}_{\mathrm{CK},\ell} + \mathbf{K}_{\mathrm{NTK},\ell-1} \circ \mathbf{K}'_{\mathrm{CK},\ell}, \quad \ell \in \{1, \ldots, L\}, \tag{7}$$

with $\mathbf{K}_{\mathrm{NTK},0} = \mathbf{K}_{\mathrm{CK},0} = \mathbf{X}^\top \mathbf{X}$, '$\mathbf{A} \circ \mathbf{B}$' the Hadamard product between two matrices $\mathbf{A}, \mathbf{B}$, $\boldsymbol{\Sigma}_\ell \in \mathbb{R}^{d_\ell \times n}$ as defined in (2), and the matrix $\mathbf{K}'_{\mathrm{CK},\ell}$ is obtained by changing the activations σ_ℓ to σ'_ℓ in the definition of $\boldsymbol{\Sigma}_\ell$ in (2). (Note the expectation is taken with respect to the random weights $\mathbf{W}_1, \cdots, \mathbf{W}_\ell$).

In the following, we recall the precise high-dimensional equivalent results in [5], under the binary GMM setting of (1), for conjugate kernel (CK) and neural tangent kernel (NTK) matrix in Theorem 1-2, respectively.

Theorem 1 (High-dimensional equivalent for CK: two-class, [5, Theorem 1]) *Under Assumption 1–3, let* $\tau_0, \ldots, \tau_L \geq 0$ *be a sequence given by the following recursion:*

$$\tau_\ell = \sqrt{\mathbb{E}[\sigma_\ell^2(\tau_{\ell-1}\xi)]}, \quad \xi \sim \mathcal{N}(0,1), \quad \ell \in \{1, \ldots, L\}, \quad \tau_0 = 1, \tag{8}$$

with "centered" activation functions such that $\mathbb{E}[\sigma_\ell(\tau_{\ell-1}\xi)] = 0$, *we have, as* $n, p \to \infty$ *that* $\|\mathbf{K}_{\mathrm{CK},\ell} - \tilde{\mathbf{K}}_{\mathrm{CK},\ell}\| \to 0$ *almost surely, with*

$$\tilde{\mathbf{K}}_{\mathrm{CK},\ell} \equiv \alpha_{\ell,1}\mathbf{X}^\top\mathbf{X} + \alpha_{\ell,2}\boldsymbol{\psi}\boldsymbol{\psi}^\top/p + \alpha_{\ell,3}\mathbf{1}_n\mathbf{1}_n^\top/p + \alpha_{\ell,0}\mathbf{I}_n, \quad \alpha_{\ell,0} \equiv \tau_\ell^2 - \tau_0^2\alpha_{\ell,1} - \tau_0^4\alpha_{\ell,3}, \tag{9}$$

for data fluctuation $\boldsymbol{\psi} = \sqrt{p}\{\|\mathbf{x}_i - \mathbb{E}[\mathbf{x}_i]\|^2 - \mathbb{E}[\|\mathbf{x}_i - \mathbb{E}[\mathbf{x}_i]\|^2]\}_{i=1}^n \in \mathbb{R}^n$, *and* $\alpha_{\ell,0}, \alpha_{\ell,1}, \alpha_{\ell,2}, \alpha_{\ell,3} \geq 0$ *satisfying*

$$\alpha_{\ell,1} = \mathbb{E}[\sigma_\ell'(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,1}, \quad \alpha_{\ell,2} = \mathbb{E}[\sigma_\ell'(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,2} + \frac{1}{4}\mathbb{E}[\sigma_\ell''(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,4}^2, \tag{10}$$

$$\alpha_{\ell,3} = \mathbb{E}[\sigma_\ell'(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,3} + \frac{1}{2}\mathbb{E}[\sigma_\ell''(\tau_{\ell-1}\xi)]^2\alpha_{\ell-1,1}^2, \tag{11}$$

with $\alpha_{\ell,4} = \alpha_{\ell-1,4}\mathbb{E}\left[(\sigma_\ell'(\tau_{\ell-1}\xi))^2 + \sigma_\ell(\tau_{\ell-1}\xi)\sigma_\ell''(\tau_{\ell-1}\xi)\right]$ *for* $\xi \sim \mathcal{N}(0,1)$.

Theorem 2 (High-dimensional equivalent for NTK: two-class, [5, Theorem 2]) *Under the same settings and notations as in Theorem 1, we have* $\|\mathbf{K}_{\mathrm{NTK},\ell} - \tilde{\mathbf{K}}_{\mathrm{NTK},\ell}\| \to 0$ *almost surely, with*

$$\tilde{\mathbf{K}}_{\mathrm{NTK},\ell} \equiv \beta_{\ell,1}\mathbf{X}^\top\mathbf{X} + \beta_{\ell,2}\boldsymbol{\psi}\boldsymbol{\psi}^\top/p + \beta_{\ell,3}\mathbf{1}_n\mathbf{1}_n^\top/p + \beta_{\ell,0}\mathbf{I}_n, \quad \beta_{\ell,0} \equiv \kappa_\ell^2 - \tau_0^2\beta_{\ell,1} - \tau_0^4\beta_{\ell,3},$$

for $\boldsymbol{\psi} \in \mathbb{R}^n, \tau_\ell \geq 0$ *as in Theorem 1,* $\tau_\ell' \geq 0$ *as defined in (8) with activation* σ' *instead of* σ, $\kappa_\ell^2 = \tau_\ell^2 + \tau_\ell'^2$, *and non-negative scalars* $\beta_{\ell,0}, \beta_{\ell,1}, \beta_{\ell,2}, \beta_{\ell,3} \geq 0$ *satisfying*

$$\beta_{\ell,1} = \alpha_{\ell,1} + \mathbb{E}\left[\sigma_\ell'(\tau_{\ell-1}\xi)\right]^2\beta_{\ell-1,1}, \quad \beta_{\ell,2} = \alpha_{\ell,2} + \mathbb{E}\left[\sigma_\ell'(\tau_{\ell-1}\xi)\right]^2\beta_{\ell-1,2}, \tag{12}$$

$$\beta_{\ell,3} = \alpha_{\ell,3} + \mathbb{E}\left[\sigma_\ell'(\tau_{\ell-1}\xi)\right]^2\beta_{\ell-1,3} + \mathbb{E}\left[\sigma_\ell''(\tau_{\ell-1}\xi)\right]^2\alpha_{\ell-1,1}\beta_{\ell-1,1}. \tag{13}$$

Theorem 1 and 2 provide a precise characterization of the CK and NTK matrix, respectively, and pave the way for both efficient numerical computation *and* analytic assessment of these matrices and their functionals (e.g., the training and generalization dynamics of the network).

From a computational perspective, the evaluation of CKs and NTKs relies on either high-dimensional integration or averaging over a huge number of independent realizations of the network [16], and can be burdensome particularly when the data dimension/size and/or the network depth is large. In this vein, the high-dimensional equivalents in Theorem 1–2 provide, for high-dimensional GMM data, a computationally more accessible "proxy" to CK and NTK matrices, and their eigenspectral functionals such as the learning dynamics to be discussed in Section 3 below.

From a theoretical analysis perspective, the results in Theorem 1 and 2 provide a key enabler to assess, e.g., how the choice of activations impacts the NTK via the parameter β_ℓs. Notably, it is of interest to observe that even activations like $\sigma(t) = \cos(t)$ will lead to $\mathbb{E}[\sigma'(\tau\xi)] = 0$ so that $\beta_{\ell,1} = 0$ for all $\ell \geq 1$, and will thus predictably perform poorly for GMM data with different means as in (1); while odd activations such as $\sigma(t) = \sin(t)$ in general result in $\beta_{\ell,1} \neq 0$ and thereby avoid this issue. This observation is numerically supported by Figure 1 below, and we refer the readers to [5] for more detailed discussions and numerical evidence.

3. Main Results

With Theorem 1 and 2 at hand, we are ready to present the main results of this paper. Our first result is the following high-dimensional characterization of the training MSE dynamics of E_t in (6), the proof of which can be found in Appendix B.

Theorem 3 (High-dimensional equivalent for training dynamics) *Under Assumption 1–3 and as $n, p \to \infty$ with $d_\ell / \max(n,p) \to \infty$, we have, for training MSE E_t defined in (6) that*

$$E_t - \tilde{E}_t \to 0, \quad \tilde{E}_t \equiv \frac{1}{2n} \operatorname{tr} \left(e^{-2\eta t \tilde{\mathbf{K}}_{\mathrm{NTK},L}} \cdot \tilde{\mathbf{K}}_{\mathrm{CK},L} \right) + \frac{1}{2n} \mathbf{y}^\top \cdot e^{-2\eta t \tilde{\mathbf{K}}_{\mathrm{NTK},L}} \cdot \mathbf{y}$$

with $\tilde{\mathbf{K}}_{\mathrm{NTK},L}, \tilde{\mathbf{K}}_{\mathrm{CK},L}$ the high-dimensional equivalents in Theorem 1–2 with $\ell = L$, respectively.

As a consequence of the *explicit* expressions in Theorem 1 and 2, it follows from Theorem 3 that the training dynamics in (6) can be determined as an *explicit* function of the data Gram matrix $\mathbf{X}^\top \mathbf{X}$ (which is still random due to the randomness in $\mathbf{X}$, but is now *independent* of the weights $\mathbf{W}$) and a few *scalar* parameters (the αs and βs) pertinent to network activation functions.

We further show, in the high-dimensional regime for n, p large, that the random $\tilde{E}_t$ in Theorem 3 (and thus the training MSE E_t) can be well approximated by a *deterministic* quantity, the expression of which involves integration over the Marčenko-Pastur law [14] and is given as follows.

Theorem 4 (Precise high-dimensional training dynamics) *Under the settings and notations of Theorem 3, we have, for training MSE E_t defined in (6) that $E_t - \bar{E}_t \to 0$ almost surely, with* deterministic *$\bar{E}_t$ explicitly given by*

$$\bar{E}_t \equiv \frac{e^{-2\eta\beta_0 t}}{2} \int e^{-2\eta\beta_1 t x} \left[\left(\alpha_0 + \alpha_1 x + \frac{1}{1 + \|\boldsymbol{\mu}\|^2} \right) \mu(dx) + \frac{\nu(dx)}{1 + \|\boldsymbol{\mu}\|^{-2}} \right],$$

with the shortcuts $\alpha_k \equiv \alpha_{L,k}$, $\beta_k \equiv \beta_{L,k}$, $k \in \{0,1\}$ as defined in Theorem 1 and 2, two probability measures

$$\mu(dx) = \frac{\sqrt{(x - \lambda_-)^+ (\lambda_+ - x)^+}}{2\pi c x} dx + (1 - c^{-1})^+ \delta_0(x), \tag{14}$$

$$\nu(dx) = \frac{\sqrt{(x - \lambda_-)^+ (\lambda_+ - x)^+}}{2\pi c \|\boldsymbol{\mu}\|^2 (\lambda_{\boldsymbol{\mu}} - x)} dx + \frac{(\|\boldsymbol{\mu}\|^4 - c^{-1})^+}{\|\boldsymbol{\mu}\|^4} \delta_{\lambda_{\boldsymbol{\mu}}}(x), \tag{15}$$

for $(t)^+ \equiv \max(t, 0)$, $\lambda_\pm = (1 \pm \sqrt{c})^2$ the left and right edge of the popular Marčenko-Pastur law [14], as well as

$$\lambda_{\boldsymbol{\mu}} = \begin{cases} 1 + c + c\|\boldsymbol{\mu}\|^2 + \|\boldsymbol{\mu}\|^{-2} \geq \lambda_+ & \textit{if } \|\boldsymbol{\mu}\|^2 > 1/\sqrt{c}, \\ (1 + \sqrt{c})^2 = \lambda_+ & \textit{otherwise.} \end{cases} \tag{16}$$

Proof [Proof sketch of Theorem 4] By Theorem 3, our object of interest E_t can be well approximated by $\tilde{E}_t$ in the NTK regime, and it thus remains to evaluate the trace (i.e., $\operatorname{tr}(e^{-2\eta t \tilde{\mathbf{K}}_{\mathrm{NTK},L}} \tilde{\mathbf{K}}_{\mathrm{CK},L})$) and quadratic (i.e., $\mathbf{y}^\top e^{-2\eta t \tilde{\mathbf{K}}_{\mathrm{NTK},L}} \mathbf{y}$) functional forms of the two random kernel matrices $\tilde{\mathbf{K}}_{\mathrm{NTK}}$ and $\tilde{\mathbf{K}}_{\mathrm{CK}}$, that are (deterministic) low-rank perturbations from standard Wishart random matrices. As such, the proof of Theorem 4 can be achieved by (i) using Cauchy's integral formula to rewrite the aforementioned matrix functionals as a complex integration of the resolvent functionals of $\tilde{\mathbf{K}}_{\rm NTK}$ and $\tilde{\mathbf{K}}_{\rm CK}$; and (ii) approximating the random integrand by their *deterministic equivalents* [3, 6] (derived with the help of the Woodbury matrix identity in Lemma 2 and the Marčenko-Pastur deterministic equivalent in Lemma 5); then, (iii) the (now deterministic) complex integration can be explicitly computed by carefully decomposing the contour to encompass both the possible isolated eigenvalues (due to the low-rank perturbation terms of $\boldsymbol{\mu}, \boldsymbol{\psi}$ and $\mathbf{1}$ in Theorem 1 and 2) and the Marčenko-Pastur main bulk. The detailed derivations can be found in Appendix C. ■

The result in Theorem 4 provides novel insight into the impact of data (statistics) on the training dynamics of DNNs. In particular, depending on the signal-to-noise ratio (SNR) $\|\boldsymbol{\mu}\|^2$ of the GMM data in (1) and the dimension ratio $c = \lim n/p$, a "phase transition" can be observed for the largest eigenvalue of $\mathbf{X}^\top \mathbf{X}$, and thus in the NTK eigenspectrum: as long as $\|\boldsymbol{\mu}\|^2 > 1/\sqrt{c}$, a "spiked" eigenvalue (at $\lambda_{\boldsymbol{\mu}} \geq \lambda_+$) isolates from the main Marčenko-Pastur bulk, and gets larger as $\|\boldsymbol{\mu}\|^2$ increase. This has a direct consequence in the DNN learning dynamics and may dominate the dynamics in the initial stage of training, see Section 4 below for numerical evidence.

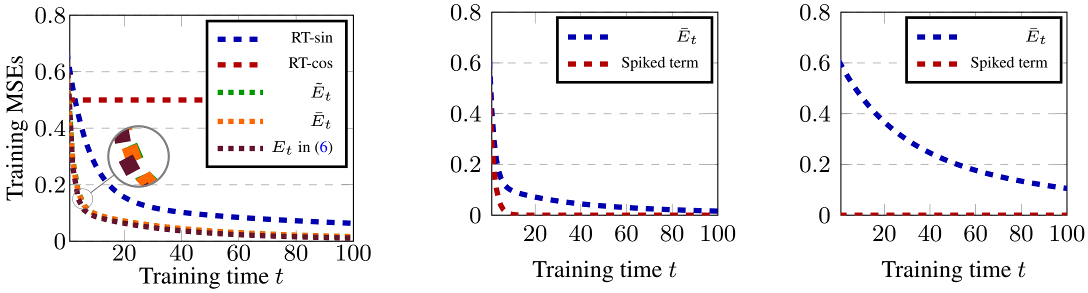
4. Numerical Experiments

In this section, we provide numerical results to validate the asymptotic analysis in Theorem 3 and 4. We train a fully-connected neural network model as defined in (2) with three hidden layers of width $d_\ell = 10\,000$ and sine or cosine activation $\sigma_\ell(t) = \sin(t), \cos(t)$, for all $\ell = 1, 2, 3$. We compare, in Figure 1(*a*), the temporal evolutions of the training MSE of (i) the actual training dynamics with gradient descent (of learning rate 0.01) and sine activation, denoted RT-sin in Figure 1(*a*); and (ii) its high-dimensional *random* equivalent $\tilde{E}_t$ given in Theorem 3; and (iii) its high-dimensional *deterministic* equivalent $\bar{E}_t$ given in Theorem 4; as well as (iv) the training dynamics E_t predicted by the NTK theory as in (6). We observe that the proposed Theorem 3 and 4 allow for a rather accurate assessment of the training dynamics within the reach of the NTK theory (i.e., $\tilde{E}_t$ and $\bar{E}_t$ to compare with E_t in (6) as proposed by the NTK theory), but nonetheless away from the actual gradient descent dynamics of finite width networks by a significant gap.

To illustrate the impact of different activations on DNN performance, Figure 1(*a*) also depicts the *actual* training dynamics of the same network, but with "centered" cosine activation. In this case, it is known (see Theorem 2 and the discussion thereafter) that the network is *not* trainable in the NTK regime and *cannot* yield satisfactory performance in classifying the Gaussian mixture in (1), as demonstrated by the plateau at large MSE (denoted RT-cos) in Figure 1(*a*).

Besides, as discussed after Theorem 4, the "signal strength" $\|\boldsymbol{\mu}\|$ of GMM data has a significant impact on the DNN training dynamics. Experiments are conducted, in Figure 1(*b*) and Figure 1(*c*), for GMM data with $\|\boldsymbol{\mu}\|^2 = 32$ and $\|\boldsymbol{\mu}\|^2 = 0.5$, respectively, with the same network and training procedure as above. We observe in Figure 1(*b*) that a larger $\|\boldsymbol{\mu}\|^2$ leads to a rapid initial drop in the training MSE, essentially due to the integration over ν, and more specifically, to the "spiked" eigenvalue $\lambda_{\boldsymbol{\mu}}$ in Theorem 4, referred to as the "Spiked term" in Figure 1(*b*) and Figure 1(*c*). It can indeed be shown that $\lambda_{\boldsymbol{\mu}}$ corresponds to the largest eigenvalue of the NTK matrix. These observations align with previous studies [4, 10] that investigate the impacts of the largest eigenvalues of the NTK. For $\|\boldsymbol{\mu}\|^2$ small, on the other hand, the training MSE decreases very slowly, and much more time is needed for the training, as illustrated in Figure 1(*c*).

To evaluate the computational benefit of the proposed exact training dynamics in Theorem 4, we compare in Table 1 the running time of (i) the actual gradient descent, (ii) the random equivalent



(a) sine versus cosine (b) $\|\boldsymbol{\mu}\|^2 = 32$ (c) $\|\boldsymbol{\mu}\|^2 = 0.5$

Figure 1: Training dynamics of a 3-hidden-layer fully-connected network with "centered" activations sine and cosine, layer widths $d_\ell = 10\,000$, for Gaussian mixture input data with $n = 5\,200, p = 5\,000$, with $\|\boldsymbol{\mu}\|^2 = 32$ and $1/2$. Gradient descent with step size 0.01.

training dynamic $\tilde{E}_t$ in Theorem 3, and (iii) the deterministic training dynamic $\bar{E}_t$ given in Theorem 4, for 500 times steps in the same setting as Figure 1(a). We see that the precise analysis in Theorem 4 greatly accelerates the training dynamics prediction, since the integration in Theorem 4 can be computed much more efficiently than matrix products of huge size.[1]

Table 1: Running time (mean ± standard deviation) of gradient descent (RT-sin, with GPU acceleration), random equivalent $\tilde{E}_t$, and deterministic equivalent $\bar{E}_t$ in the setting of Figure 1(a), for 500 time steps. Results are obtained by averaging over 10 independent runs.

	RT-sin	$\tilde{E}_t$	$\bar{E}_t$
Running time (s)	99.6 ± 1.4	65.1 ± 5.3	4.9 ± 0.4

Acknowledgments

Z. Liao would like to acknowledge the National Natural Science Foundation of China (via funds NSFC-62206101 and NSFC-12141107), the Fundamental Research Funds for the Central Universities of China (2021XXJS110), the Key Research and Development Program of Hubei (2021BAA037) and of Guangxi (GuiKe-AB21196034) for providing partial support.

References

[1] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, (48):30063–30070, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1907378117.

[2] Florent Benaych-Georges and Raj Rao Nadakuditi. The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices. *Advances in Mathematics*, 227(1):494–521, 2011. ISSN 0001-8708. doi: 10.1016/j.aim.2011.02.007.

[1] The running time for $\tilde{E}_t$ includes the computation time of $\tilde{\mathbf{K}}_{\rm CK}$ and $\tilde{\mathbf{K}}_{\rm NTK}$ as per Theorem 1 and 2.

[3] Romain Couillet and Zhenyu Liao. *Random Matrix Methods for Machine Learning*. Cambridge University Press, 2022. ISBN 9781009186742.

[4] Zhou Fan and Zhichao Wang. Spectra of the Conjugate Kernel and Neural Tangent Kernel for linear-width neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 7710–7721. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/572201a4497b0b9f02d4f279b09ec30d-Paper.pdf.

[5] Lingyu Gu, Yongqi Du, Yuan Zhang, Di Xie, Shiliang Pu, Robert Qiu, and Zhenyu Liao. "Lossless" Compression of Deep Neural Networks: A High-dimensional Neural Tangent Kernel Approach. In *Advances in Neural Information Processing Systems*, volume 35, pages 3774–3787. Curran Associates, Inc., 2022.

[6] Walid Hachem, Philippe Loubaton, and Jamal Najim. Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability*, 17(3):875–930, 2007. ISSN 1050-5164. doi: 10.1214/105051606000000925.

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

[8] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2 edition, 2012. ISBN 9780521548236. URL http://www.cambridge.org/9780521548236.

[9] Jiaoyang Huang and Horng-Tzer Yau. Dynamics of deep neural networks and neural tangent hierarchy. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4542–4551. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/huang20l.html.

[10] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems*, volume 31 of *NIPS'18*, pages 8571–8580. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017. ISSN 0001-0782. doi: 10.1145/3065386.

[12] Zhenyu Liao and Romain Couillet. The Dynamics of Learning: A Random Matrix Approach. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3072–3081, Stockholmsmässan, Stockholm Sweden, 2018. PMLR. URL http://proceedings.mlr.press/v80/liao18b.html.

[13] Zhenyu Liao, Romain Couillet, and Michael W Mahoney. A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent. In *Advances in Neural Information Processing Systems*, volume 33, pages 13939–13950. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/a03fa30821986dff10fc66647c84c9c3-Paper.pdf.

[14] Vladimir A Marcenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967. ISSN 0025-5734. doi: 10.1070/sm1967v001n04abeh001994.

[15] Song Mei and Andrea Montanari. The Generalization Error of Random Features Regression: Precise Asymptotics and the Double Descent Curve. *Communications on Pure and Applied Mathematics*, 2021. ISSN 0010-3640. doi: 10.1002/cpa.22008.

[16] Roman Novak, Lechao Xiao, Jiri Hron, Jaehoon Lee, Alexander A. Alemi, Jascha Sohl-Dickstein, and Samuel S. Schoenholz. Neural tangents: Fast and easy infinite neural networks in python. In *International Conference on Learning Representations*, 2020. URL https://github.com/google/neural-tangents.

[17] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016. ISSN 0028-0836. doi: 10.1038/nature16961.

[18] Jack W. Silverstein and Zhidong Bai. On the Empirical Distribution of Eigenvalues of a Class of Large Dimensional Random Matrices. *Journal of Multivariate Analysis*, 54(2):175–192, 1995. ISSN 0047-259X. doi: 10.1006/jmva.1995.1051.

[19] Aad W. Van der Vaart. *Asymptotic Statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, 2000. ISBN 9780521784504. doi: 10.1017/cbo9780511802256.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

In the appendix of this paper, we present in Appendix A the technical lemmas to be used throughout the proof; in Appendix B the proof of Theorem 3; and in Appendix C the proof of Theorem 4.

Appendix A. Useful lemmas

Lemma 1 (Quadratic-form-close-to-the-trace, trace lemma, [3, Lemma 2.11]) *Let* $\mathbf{A} \in \mathbb{R}^{n\times n}$ *be a deterministic matrix of bounded spectral norm and* $\mathbf{w} = [w_1, \ldots, w_n]^\top \in \mathbb{R}^n$ *be a random vector having independent entries with zero mean* $\mathbb{E}[w_i] = 0$*, unit variance* $\operatorname{Var}[w_i] = 1$*, and finite eighth-order moment* $\mathbb{E}[w_i^8] < \infty$*. Then,*

$$\frac{1}{n}\mathbf{w}^\top \mathbf{A}\mathbf{w} - \frac{1}{n}\operatorname{tr}\mathbf{A} \to 0, \tag{17}$$

almost surely as $n \to \infty$.

Lemma 2 (Woodbury matrix identity)

$$\left(\mathbf{A} + \mathbf{C}\mathbf{B}\mathbf{C}^\top\right)^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{C}\left(\mathbf{B}^{-1} + \mathbf{C}^\top\mathbf{A}^{-1}\mathbf{C}\right)^{-1}\mathbf{C}^\top\mathbf{A}^{-1}, \tag{18}$$

Lemma 3 ([18, Lemma 2.6]) *For* $\mathbf{A}, \mathbf{M} \in \mathbb{R}^{p\times p}$ *symmetric and nonnegative definite,* $\mathbf{u} \in \mathbb{R}^p$, $\tau \in \mathbb{R}$ *and* $z \in \mathbb{C}\setminus \operatorname{supp}\boldsymbol{\mu}\left(\mathbf{M} + \tau\mathbf{u}\mathbf{u}^\top\right)$,

$$\left|\operatorname{tr}\mathbf{A}\left(\mathbf{M} + \tau\mathbf{u}\mathbf{u}^\top - z\mathbf{I}_p\right)^{-1} - \operatorname{tr}\mathbf{A}\left(\mathbf{M} - z\mathbf{I}_p\right)^{-1}\right| \leq \frac{\|\mathbf{A}\|}{\operatorname{dist}(z, \operatorname{supp}\boldsymbol{\mu}\left(\mathbf{M} + \tau\mathbf{u}\mathbf{u}^\top\right))}.$$

with $\operatorname{dist}(z, \operatorname{supp}\boldsymbol{\mu}\left(\mathbf{M} + \tau\mathbf{u}\mathbf{u}^\top\right))$ the distance between z and the support of the eigenvalue distribution of $\mathbf{M} + \tau\mathbf{u}\mathbf{u}^\top$. This lemma establishes that low-rank perturbations have a negligible effect on the trace of the inverse of a matrix.

Lemma 4 (Cauchy's integral formula) *For* $\Gamma \subset \mathbb{C}$ *a positively (i.e., counterclockwise) oriented simple closed curve and a complex function* $f(z)$ *analytic in a region containing* Γ *and its inside, then if* $z_0 \in \mathbb{C}$ *is enclosed by* Γ,

$$f(z_0) = -\frac{1}{2\pi\imath}\oint_\Gamma \frac{f(z)}{z_0 - z}dz;$$

if not,

$$\frac{1}{2\pi\imath}\oint_\Gamma \frac{f(z)}{z_0 - z}dz = 0.$$

Lemma 5 ([3, Theorem 2.4, Marčenko-Pastur]) *Let* $\mathbf{X} \in \mathbb{R}^{p\times n}$ *with i.i.d. columns* $\mathbf{x}_i$ *such that* $\mathbf{x}_i$ *has independent entries with zero mean, unit variance, and some light tail condition t and denote* $\mathbf{Q}(z) = \left(\frac{1}{n}\mathbf{X}\mathbf{X}^\top - z\mathbf{I}_p\right)^{-1}$ *the resolvent of* $\frac{1}{n}\mathbf{X}\mathbf{X}^\top$*. Then, as* $n, p \to \infty$ *with* $p/n \to c \in (0, \infty)$,

$$\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z), \quad \bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_p,$$

$$zcm^2(z) - (1 - c - z)m(z) + 1 = 0$$

The function $m(z)$ is the unique solution as the Stieltjes transform of the probability measure $\boldsymbol{\mu}$ given explicitly by

$$\mu(dx) = \left(1 - c^{-1}\right)^{+} \delta_0(x) + \frac{1}{2\pi cx}\sqrt{(x - E_-)^{+}\,(E_+ - x)^{+}}dx$$

where $E_{\pm} = (1 \pm \sqrt{c})^2$ and $(x)^{+} = \max(0, x)$, and is known as the Marčenko-Pastur distribution. In particular, with probability one, the empirical spectral measure $\mu_{\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathrm{T}}}$ converges weakly to μ.

Lemma 6 (Weyl's inequality, [8, Theorem 4.3.1]) *Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{p\times p}$ be symmetric matrices and let the respective eigenvalues of $\mathbf{A}$, $\mathbf{B}$ and $\mathbf{A}+\mathbf{B}$ be arranged in nondecreasing order, i.e., $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_{p-1} \leq \lambda_p$. Then, for all $i \in \{1, \ldots, p\}$,*

$$\begin{aligned} &\lambda_i(\mathbf{A}+\mathbf{B}) \leq \lambda_{i+j}(\mathbf{A}) + \lambda_{p-j}(\mathbf{B}), \quad j = 0, 1, \ldots, p - i, \\ &\lambda_{i-j+1}(\mathbf{A}) + \lambda_j(\mathbf{B}) \leq \lambda_i(\mathbf{A}+\mathbf{B}), \quad j = 1, \ldots, i \end{aligned}$$

Lemma 7 (Eigenvalue phase transition, [2, Theorem 2.1]) *Let $\mathbf{X}_n$ be an $n \times n$ symmetric (or Hermitian) random matrix with ordered eigenvalues $\lambda_1(X_n) \geqslant \cdots \geqslant \lambda_n(X_n)$. Let $\mu_{\mathbf{X}_n}$ be the empirical eigenvalue distribution defined as*

$$\mu_{\mathbf{X}_n} = \frac{1}{n}\sum_{j=1}^{n} \delta_{\lambda_j(\mathbf{X}_n)}.$$

Assume the probability measure $\mu_{\mathbf{X}_n}$ converges almost surely weakly, as $n \to \infty$, to a nonrandom compactly supported probability measure $\mu_{\mathbf{X}}$. We suppose the smallest and largest eigenvalue of X_n converge almost surely to a and b.

Let P_n be an $n \times n$ symmetric (or Hermitian) random matrix having rank r with its r non-zero eigenvalues equal to $\theta_1, \ldots, \theta_r$, which are deterministic non-zero real numbers. And we define index $s \in \{0, \ldots, r\}$ such that $\theta_1 \geqslant \cdots \geqslant \theta_s > 0 > \theta_{s+1} \geqslant \cdots \geqslant \theta_r$.

Consider the rank r additive perturbation of the random matrix X_n given by

$$\widetilde{\mathbf{X}}_n = \mathbf{X}_n + \mathbf{P}_n.$$

For the extreme eigenvalues of $\widetilde{\mathbf{X}}_n$ each $1 \leqslant i \leqslant s$ as $n \to \infty$, we have

$$\lambda_i\left(\tilde{\mathbf{X}}_n\right) \xrightarrow{a.s.} \begin{cases} m_{\mu_{\mathbf{X}}}^{-1}(1/\theta_i) & \text{if } \theta_i > 1/m_{\mu_{\mathbf{X}}}\left(b^{+}\right), \\ b & \text{otherwise,} \end{cases}$$

while for each fixed $i > s$, $\lambda_i\left(\tilde{\mathbf{X}}_n\right) \xrightarrow{a.s.} b$.

Similarly, for the small eigenvalues, we have that for each $0 \leqslant j < r - s$,

$$\lambda_{n-j}\left(\widetilde{\mathbf{X}}_n\right) \xrightarrow{a.s.} \begin{cases} m_{\mu_{\mathbf{X}}}^{-1}(1/\theta_{r-j}) & \text{if } \theta_j < 1/m_{\mu_{\mathbf{X}}}\left(a^{-}\right), \\ a & \text{otherwise,} \end{cases}$$

while for each fixed $j \geqslant r - s$, $\lambda_{n-j}\left(\tilde{\mathbf{X}}_n\right) \xrightarrow{a.s.} a$.

Note that

$$m_{\mu_{\mathbf{X}}}(z) = \int \frac{1}{z-t}\,\mathrm{d}\mu_{\mathbf{X}}(t) \quad \text{for } z \notin \operatorname{supp}\mu_{\mathbf{X}},$$

is the Stieltjes transform of $\mu_{\mathbf{X}}$, $m_{\mu_X}^{-1}(\cdot)$ is its functional inverse.

Appendix B. Proof of Theorem 3

In this section, we consider the fully-connected DNN model as in (2) of depth L:

$$\boldsymbol{\Sigma}_\ell(\mathbf{X}) = \frac{1}{\sqrt{d_\ell}}\sigma_\ell\left(\frac{1}{\sqrt{d_{\ell-1}}}\mathbf{W}_\ell\sigma_{\ell-1}\left(\cdots\frac{1}{\sqrt{d_2}}\sigma_2\left(\frac{1}{\sqrt{d_1}}\mathbf{W}_2\sigma_1\left(\mathbf{W}_1\mathbf{x}\right)\right)\right)\right), \quad \ell = 1,\ldots,L, \tag{19}$$

and denote $\mathbf{W}(0), \mathbf{w}(0)$ the initial (random) states of the DNN weights $\mathbf{W}, \mathbf{w}$ at time $t = 0$, and collects all vectorized weights at initialization as

$$\mathbf{W}(0) = \left[\operatorname{vec}(\mathbf{W}_1(0)), \cdots, \operatorname{vec}(\mathbf{W}_L(0))\right],$$
$$\boldsymbol{\theta} = [\operatorname{vec}(\mathbf{W}_1(0)), \cdots, \operatorname{vec}(\mathbf{W}_L(0)), \mathbf{w}(0)].$$

For random x, we denote $\mathbb{E}_x[f(x)]$ the expectation of $f(x)$ with respect to x.

Organization of the proof of Theorem 3 Recall our object of interest here in Theorem 3 is the normalized training MSE E_t in the (infinitely wide) NTK regime given in (6) as

$$\begin{aligned} E_t &= \frac{1}{2n}\|f_t(\mathbf{X}) - \mathbf{y}\|_2^2 = \frac{1}{2n}(f_0(\mathbf{X}) - \mathbf{y})^\top \cdot e^{-2\eta t\cdot\mathbf{K}_{\mathrm{NTK},L}} \cdot (f_0(\mathbf{X}) - \mathbf{y}) \\ &= \underbrace{\frac{1}{2n}f_0(\mathbf{X})^\top e^{-2\eta t\cdot\mathbf{K}_{\mathrm{NTK},L}}f_0(\mathbf{X})}_{\equiv E_t^a} - \underbrace{\frac{1}{n}f_0(\mathbf{X})^\top e^{-2\eta t\mathbf{K}_{\mathrm{NTK},L}}\cdot\mathbf{y}}_{\equiv E_t^b} + \underbrace{\frac{1}{2n}\mathbf{y}^\top e^{-2\eta t\mathbf{K}_{\mathrm{NTK},L}}\mathbf{y}}_{\equiv E_t^c}, \end{aligned} \tag{20}$$

with $f_0(\mathbf{X})$ the output of DNN at time step $t = 0$ as

$$f_0(\mathbf{X}) = \boldsymbol{\Sigma}_L^\top(\mathbf{X})\mathbf{w}(0) \in \mathbb{R}^n, \tag{21}$$

with $\boldsymbol{\Sigma}_L(\mathbf{X})$ given in (19).

To prove Theorem 3, it suffices to show that under Assumption 1–3 and as $n, p \to \infty$ with $d_\ell/\max(n,p) \to \infty$, we have

1. $E_t^a - \tilde{E}_t^a \to 0$ with $\tilde{E}_t^a = \frac{1}{2n}\operatorname{tr}\left(e^{-2\eta t\cdot\tilde{\mathbf{K}}_{\mathrm{NTK},L}}\cdot\tilde{\mathbf{K}}_{\mathrm{CK},L}\right)$; and
2. $E_t^b \to 0$; and
3. $E_t^c - \tilde{E}_t^c \to 0$ with $\tilde{E}_t^c \equiv \frac{1}{2n}\mathbf{y}^\top \cdot e^{-2\eta t\tilde{\mathbf{K}}_{\mathrm{NTK},L}}\cdot\mathbf{y}$,

for the high-dimensional equivalent CK and NTK matrices $\tilde{\mathbf{K}}_{\mathrm{CK},L}$ and $\tilde{\mathbf{K}}_{\mathrm{NTK},L}$, given respectively in Theorem 1 and Theorem 2. This allows one to conclude that $E_t - \tilde{E}_t \to 0$ as $n, p \to \infty$ with $\tilde{E}_t = \tilde{E}_t^a + \tilde{E}_t^c$, and thus the conclusion of Theorem 3.

Detailed derivation of $E_t^a - \tilde{E}_t^a \to 0$ For the term $E_t^a = \frac{1}{2n}f_0(\mathbf{X})^\top \cdot e^{-2\eta t\cdot\mathbf{K}_{\mathrm{NTK},L}}\cdot f_0(\mathbf{X})$, by substituting $f_0(\mathbf{X})$ with the definition in (21), we get

$$E_t^a = \frac{1}{2n}f_0(\mathbf{X})^\top \cdot e^{-2\eta t\cdot\mathbf{K}_{\mathrm{NTK},L}}\cdot f_0(\mathbf{X}) = \frac{1}{2n}\mathbf{w}(0)^\top\boldsymbol{\Sigma}_L(\mathbf{X})e^{-2\eta t\cdot\mathbf{K}_{\mathrm{NTK},L}}\boldsymbol{\Sigma}_L(\mathbf{X})^\top\mathbf{w}(0),$$

for which we would like to apply Lemma 1 with $\mathbf{A} = \boldsymbol{\Sigma}_L(\mathbf{X})\cdot e^{-2\eta t\cdot\mathbf{K}_{\mathrm{NTK},L}}\cdot\boldsymbol{\Sigma}_L(\mathbf{X})^\top$ to conclude that $E_t^a \simeq \frac{1}{2n}\operatorname{tr}(\boldsymbol{\Sigma}_L(\mathbf{X})e^{-2\eta t\cdot\mathbf{K}_{\mathrm{NTK},L}}\boldsymbol{\Sigma}_L(\mathbf{X})^\top)$ under Assumption 2.

To that end, we need to first establish a spectral norm bound for the random matrix $\mathbf{\Sigma}_L(\mathbf{X}) \cdot e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \cdot \mathbf{\Sigma}_L(\mathbf{X})$ in the large d_ℓ, n, p regime. Note that with $d_\ell \gg \max(n,p)$ with $d_\ell, n, p \to \infty$, we have, by the law of large numbers that

$$\mathbf{\Sigma}_L(\mathbf{X})^\top \mathbf{\Sigma}_L(\mathbf{X}) - \mathbb{E}_{\mathbf{W}}\left[\mathbf{\Sigma}_L(\mathbf{X})^\top \mathbf{\Sigma}_L(\mathbf{X})\right] \to 0, \tag{22}$$

almost surely, so that

$$\mathbf{\Sigma}_L(\mathbf{X})^\top \mathbf{\Sigma}_L(\mathbf{X}) = \mathbb{E}_{\mathbf{W}}\left[\mathbf{\Sigma}_L(\mathbf{X})^\top \mathbf{\Sigma}_L(\mathbf{X})\right] + o_{\|\cdot\|_2}(1) = \mathbf{K}_{\mathrm{CK},L} + o_{\|\cdot\|_2}(1) = \tilde{\mathbf{K}}_{\mathrm{CK},L} + o_{\|\cdot\|_2}(1) \tag{23}$$

where we denote $o_{\|\cdot\|_2}(1)$ a matrix having (almost sure) vanishing spectral norm as $n, p \to \infty$ and used the definition of CK matrix $\mathbf{K}_{\mathrm{CK},L}$ in (7) in the second and Theorem 1 in the third approximation. We thus have

$$\begin{aligned}\|\mathbf{\Sigma}_L(\mathbf{X}) \cdot e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \cdot \mathbf{\Sigma}_L(\mathbf{X})^\top\|_2 &\leq \|\mathbf{\Sigma}_L(\mathbf{X})^\top \mathbf{\Sigma}_L(\mathbf{X})\|_2 \cdot \|e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}}\|_2 \\ &\leq \|\mathbf{\Sigma}_L(\mathbf{X})^\top \mathbf{\Sigma}_L(\mathbf{X})\|_2 = \|\tilde{\mathbf{K}}_{\mathrm{CK},L}\|_2 + o(1) = O(1)\end{aligned}$$

almost surely as $d_\ell, n, p \to \infty$, where we used the non-negative definiteness of $\mathbf{K}_{\mathrm{NTK},L}$ per its definition and the approximation in (23) in the second line. The conclusion that $\|\tilde{\mathbf{K}}_{\mathrm{CK},L}\|_2 = O(1)$ can be reached by using Theorem 1 and standard RMT arguments. This allows us to write with Lemma 1 that

$$\begin{aligned}E_t^a &= \frac{1}{2n}\mathbf{w}(0)^\top \mathbf{\Sigma}_L(\mathbf{X}) \cdot e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \cdot \mathbf{\Sigma}_L(\mathbf{X})^\top \mathbf{w}(0) \\ &= \frac{1}{2n}\operatorname{tr}\left(\mathbf{\Sigma}_L(\mathbf{X}) \cdot e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \cdot \mathbf{\Sigma}_L(\mathbf{X})^\top\right) + o(1) \\ &= \frac{1}{2n}\operatorname{tr}\left(e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \cdot \mathbf{K}_{\mathrm{CK},L}\right) + o(1) = \frac{1}{2n}\operatorname{tr}\left(e^{-2\eta t \cdot \tilde{\mathbf{K}}_{\mathrm{NTK},L}} \cdot \tilde{\mathbf{K}}_{\mathrm{CK},L}\right) + o(1) = E_t^a + o(1),\end{aligned}$$

where we used Theorem 1 and Theorem 2 in the last line. This concludes the proof of $E_t^a - \tilde{E}_t^a \to 0$.

Detailed derivation of $E_t^b \to 0$ For the term $E_t^b = \frac{1}{n} f_0(\mathbf{X})^\top \cdot e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \cdot \mathbf{y}$, we compute its mean and variance as follows. First, note that

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}[E_t^b] &= \frac{1}{n}\mathbb{E}_{\boldsymbol{\theta}}\left[f_0(\mathbf{X})^\top \cdot e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \cdot \mathbf{y}\right] \\ &= \frac{1}{n}\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbf{w}(0)^\top \mathbf{\Sigma}_L(\mathbf{X}) \cdot e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \cdot \mathbf{y}\right] &(24) \\ &= \frac{1}{n}\mathbb{E}_{\mathbf{w}(0)}\left[\mathbf{w}(0)^\top\right] \cdot \mathbb{E}_{\mathbf{W}(0)}\left[\mathbf{\Sigma}_L(\mathbf{X}) \cdot e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \cdot \mathbf{y}\right] = 0, &(25)\end{aligned}$$

where we used the independence between $\mathbf{w}(0)$ and $\mathbf{W}(0)$, together with the fact that $\mathbb{E}[\mathbf{w}(0)] = \mathbf{0}$. For the variance of E_t^b, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{\theta}}\left[(E_t^b)^2\right] &= \frac{1}{n^2}\mathbb{E}_{\boldsymbol{\theta}}\left[f_0(\mathbf{X})^\top \cdot e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \cdot \mathbf{y} \cdot \mathbf{y}^\top \cdot e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \cdot f_0(\mathbf{X})\right] \\ &= \frac{1}{n^2}\mathbb{E}_{\boldsymbol{\theta}}\left[\mathbf{w}(0)^\top \mathbf{\Sigma}_L(\mathbf{X}) \cdot e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \cdot \mathbf{y} \cdot \mathbf{y}^\top \cdot e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \cdot \mathbf{\Sigma}_L(\mathbf{X})^\top \mathbf{w}(0)\right]\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n^2} \operatorname{tr} \left(\mathbb{E}_{\mathbf{W}(0)} \left[\mathbf{\Sigma}_L(\mathbf{X}) \cdot e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \cdot \mathbf{y} \cdot \mathbf{y}^\top \cdot e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \cdot \mathbf{\Sigma}_L(\mathbf{X})^\top \right] \mathbf{I}_{d_\ell} \right) \\
&= \frac{1}{n^2} \operatorname{tr} \left(\mathbb{E}_{\mathbf{W}(0)} \left[e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \cdot \mathbf{y} \cdot \mathbf{y}^\top \cdot e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \cdot \mathbf{\Sigma}_L(\mathbf{X})^\top \mathbf{\Sigma}_L(\mathbf{X}) \right] \right) \\
&= \frac{1}{n^2} \operatorname{tr} \left(e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \cdot \mathbf{y} \cdot \mathbf{y}^\top \cdot e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \cdot \mathbf{K}_{\mathrm{CK},L} \right) \\
&\leq \frac{1}{n} \| e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \cdot \mathbf{K}_{\mathrm{CK},L} \cdot e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \|_2 = O(n^{-1}),
\end{aligned} \tag{26}
$$

where we used again the definition of CK matrix in the fourth, the fact that $\|\mathbf{y}\|_2^2 = n$ in the fifth, and $\|\mathbf{K}_{\mathrm{CK},L}\|_2 = \|\tilde{\mathbf{K}}_{\mathrm{CK},L}\|_2 + o(1) = O(1)$ in the last line. The results in (24) and (26) are sufficient to show that

$$
E_t^b = \frac{1}{n} f_0(\mathbf{X})^\top \cdot e^{-2\eta t \cdot \mathbf{K}_{\mathrm{NTK},L}} \cdot \mathbf{y} \to 0 \tag{27}
$$

in probability as $d_\ell, n, p \to \infty$. To establish an almost sure convergence result, one can similarly bound the fourth-order moment of E_t^b and apply Borel–Cantelli lemma.

Further, note that

$$
E_t^c = \frac{1}{2n} \mathbf{y}^\top e^{-2\eta t \mathbf{K}_{\mathrm{NTK},L}} \mathbf{y} = \frac{1}{2n} \mathbf{y}^\top e^{-2\eta t \tilde{\mathbf{K}}_{\mathrm{NTK},L}} \mathbf{y} + o(1) = \tilde{E}_t^c + o(1), \tag{28}
$$

with Theorem 2 and the fact that $\|\mathbf{y}\|_2^2 = n$. This concludes the proof of Theorem 3.

Appendix C. Precise high-dimensional training dynamics

The proof of Theorem 4 is shown in this section. We first introduce the process of the proof and elaborate on each step in Appendix C.1, which will be helpful for readers to follow the detailed proof shown in Appendix C.2. We follow notations in the main paragraphs and in Appendix B.

C.1. The calculation process with random matrix tools

Notations and Setup What we encounter in this work is to calculate a scalar mapping of a function $f(\tilde{\mathbf{M}})$ of $\tilde{\mathbf{M}} \in \mathbb{R}^{n\times n}$ (e.g., $\operatorname{tr}(f(\tilde{\mathbf{M}}))$ and $\mathbf{a}^\top f(\tilde{\mathbf{M}})\mathbf{b}$), with $\tilde{\mathbf{M}}$ a (deterministic) low-rank perturbation from a Wishart random matrix $\mathbf{M}$, in particular, both $\mathbf{K}_{\mathrm{CK}}$ and $\mathbf{K}_{\mathrm{NTK}}$ belong to this kind of matrix. And f is an analytic function at the area of interest.

C.1.1. COMPUTATIONAL PROCEDURE FOR THE CALCULATION OF A SCALAR FUNCTION $f(\tilde{\mathbf{M}})$

The calculation is performed by first converting the mapping of $f(\tilde{\mathbf{M}})$ to a complex integration of the resolvent $\mathbf{Q}_{\tilde{\mathbf{M}}(z)} = (\tilde{\mathbf{M}} - z\mathbf{I}_n)^{-1}$ per Cauchy's integral formula, then by replacing $\mathbf{Q}_{\tilde{\mathbf{M}}(z)}$ with its deterministic equivalent [2] matrix $\bar{\mathbf{Q}}_{\tilde{\mathbf{M}}(z)}$, we manage to convert the integral of a function of the *random matrix* $\mathbf{Q}_{\tilde{\mathbf{M}}(z)}$ into the integral of a function of the *deterministic matrix* $\bar{\mathbf{Q}}_{\tilde{\mathbf{M}}(z)}$, which also makes it feasible to perform the complex integral followed.

As an illustration, we consider the process for the calculation $\operatorname{tr}(f(\tilde{\mathbf{M}}))$, which is as follows:

$$\overbrace{\operatorname{tr}(f(\tilde{\mathbf{M}})) = -\frac{1}{2\pi i}\oint_\gamma f(z)\operatorname{tr}(\mathbf{Q}_{\tilde{\mathbf{M}}}(z))dz}^{\text{Cauchy's integral formula}} \underbrace{= -\frac{1}{2\pi i}\oint_\gamma f(z)\operatorname{tr}(\bar{\mathbf{Q}}_{\tilde{\mathbf{M}}}(z))dz}_{\text{Deterministic Equivalent}} \overbrace{= -\frac{1}{2\pi i}\oint_\gamma f(z)m(z)dz}^{\text{Complex integral}} \tag{29}$$

with $m(z) \equiv \operatorname{tr}(\bar{\mathbf{Q}}_{\tilde{\mathbf{M}}(z)})$ and γ the contour encompassing all eigenvalues of $\tilde{\mathbf{M}}$.

We will then elaborate on each step.

Cauchy's integral formula As a symmetric matrix, we can perform spectral decomposition on $\tilde{\mathbf{M}}$ and get $\tilde{\mathbf{M}} = \mathbf{U}\Lambda\mathbf{U}^\top$, with $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_n] \in \mathbb{R}^{n\times n}$ and $\Lambda = \operatorname{diag}\left\{\lambda_1(\tilde{\mathbf{M}}), \dots, \lambda_n(\tilde{\mathbf{M}})\right\}$. And thus

$$f(\tilde{\mathbf{M}}) = f(\mathbf{U}\Lambda\mathbf{U}^\top) = \mathbf{U}f(\Lambda)\mathbf{U}^\top = \mathbf{U}\operatorname{diag}\left\{f(\lambda_1(\tilde{\mathbf{M}})), \dots, f(\lambda_n(\tilde{\mathbf{M}}))\right\}\mathbf{U}^\top$$

then per Cauchy's integral formula as Lemma 4, we have

$$\begin{aligned} f(\tilde{\mathbf{M}}) &= -\frac{1}{2\pi\imath}\mathbf{U}\operatorname{diag}\left\{\oint_\Gamma \frac{f(z)}{\lambda_1(\tilde{\mathbf{M}}) - z}dz, \dots, \oint_\Gamma \frac{f(z)}{\lambda_n(\tilde{\mathbf{M}}) - z}dz\right\}\mathbf{U}^\top \\ &= -\frac{1}{2\pi\imath}\oint_\Gamma f(z)(\tilde{\mathbf{M}} - z\mathbf{I}_n)^{-1}dz = -\frac{1}{2\pi\imath}\oint_\Gamma f(z)\mathbf{Q}_{\tilde{\mathbf{M}}}(z)dz \end{aligned}$$

with Γ a contour encompassing all eigenvalues of $\tilde{\mathbf{M}}$ and thus we have

$$\operatorname{tr}(f(\tilde{\mathbf{M}})) = -\frac{1}{2\pi\imath}\oint_\Gamma f(z)\operatorname{tr}(\mathbf{Q}_{\tilde{\mathbf{M}}}(z))dz \tag{30}$$

[2] We follow notations of matrix equivalents in [3, Notation 1], that is, for $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n\times n}$, we denote $\mathbf{X} \leftrightarrow \mathbf{Y}$ if, for all unit norm $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, $\frac{1}{n}\operatorname{tr}\mathbf{A}(\mathbf{X} - \mathbf{Y}) \xrightarrow{a.s.} 0$, $\mathbf{a}^\top(\mathbf{X} - \mathbf{Y})\mathbf{b} \xrightarrow{a.s.} 0$ and $\|\mathbb{E}[\mathbf{X} - \mathbf{Y}]\| \to 0$

Deterministic Equivalent As noticed before, $\tilde{\mathbf{M}}$ is a (deterministic) low-rank perturbation from a Wishart random matrix $\mathbf{M}$, that is:

$$\tilde{\mathbf{M}} = \mathbf{M} + \mathbf{C}\mathbf{B}\mathbf{C}^\top \tag{31}$$

for $\mathbf{B} \in \mathbb{R}^{r\times r}$ a low-rank matrix. When it comes to finding the deterministic equivalent for $\mathbf{Q}_{\tilde{\mathbf{M}}}(z)$, one efficient way to leverage the Woodbury matrix identity and get

$$\mathbf{Q}_{\tilde{\mathbf{M}}}(z) = \left(\mathbf{M} + \mathbf{C}\mathbf{B}\mathbf{C}^\top - z\mathbf{I}_n\right)^{-1} = \mathbf{Q}_{\mathbf{M}}(z) - \mathbf{Q}_{\mathbf{M}}(z)\mathbf{C}\left(\mathbf{B}^{-1} + \mathbf{C}^\top\mathbf{Q}_{\mathbf{M}}(z)\mathbf{C}\right)^{-1}\mathbf{C}^\top\mathbf{Q}_{\mathbf{M}}(z),$$

to convert it to the deterministic equivalent $\bar{\mathbf{Q}}_{\mathbf{M}}(z)$ of the $\mathbf{Q}_{\mathbf{M}}(z)$ which has already existed, also known as part of the result the Marčenko-Pastur shown as in Lemma 5, that is:

$$\mathbf{Q}_{\mathbf{M}}(z) \leftrightarrow \bar{\mathbf{Q}}_{\mathbf{M}}(z), \quad \bar{\mathbf{Q}}_{\mathbf{M}}(z) = m(z)\mathbf{I}_p,$$

with $(z, m(z))$ the unique solution in $\mathcal{Z}\left(\mathbb{C}\backslash\left[(1-\sqrt{c})^2, (1+\sqrt{c})^2\right]\right)$ of

$$zcm^2(z) - (1-c-z)m(z) + 1 = 0$$

Complex Integral In the final step, a complex integral is performed with a contour encompassing all eigenvalues of $\tilde{\mathbf{M}}$, thus we focus on the spectral distribution of $\mathbf{M}$ first.

As a (deterministic) low-rank perturbation from a Wishart random matrix $\mathbf{M}$, the eigenvalues of $\tilde{\mathbf{M}}$ contain eigenvalues of $\mathbf{M}$, the "main bulk", and some "possible" isolated eigenvalues to the right of the "main bulk" (if the perturbations are nonnegative) per Weyl's Theorem shown in Lemma 6. The isolated eigenvalues are not always existing only when "Phase Transition" occurs, in other words, the disturbance will lead to isolated eigenvalues only when it is greater than the threshold of "Phase Transition", and the threshold is defined in Lemma 7, as for a Wishart random matrix $\mathbf{M}$, it is:

$$\lambda_i \xrightarrow{\text{a.s.}} \begin{cases} \frac{1}{cm(\lambda_i)+1} - \frac{1}{m(\lambda_i)} & \text{if } m(\lambda_i) > -\frac{1}{c+\sqrt{c}}, \\ (1+\sqrt{c})^2 & \text{otherwise,} \end{cases} \tag{32}$$

with $m(\lambda_i)$ obtained by letting $\det(\tilde{\mathbf{M}} - \lambda_i\mathbf{I}_n) = 0$, and c a constant associated with $\mathbf{M}$ as defined in the MP Lemma 5. We recommend readers to see [2] for more details about eigenvalues for low-rank disturbed random matrices.

After calculating eigenvalues of $\tilde{\mathbf{M}}$, the complex integral performed is then performed by dividing the eigenvalues into two groups and calculating them separately, including the "main bulk" of the eigenvalues calculated by contour integral but selecting a rectangular contour γ_a with upper and lower sides extremely close to the real axis and with the left and right sides holding a tiny distance ϵ from the edges of the main bulk; and some isolated eigenvalues of which integration can be tackled by the residue theorem, encompassed by a contour γ_b, as illustrated in the following figure.

With these preliminaries, we are prepared for the proof of Theorem 4.

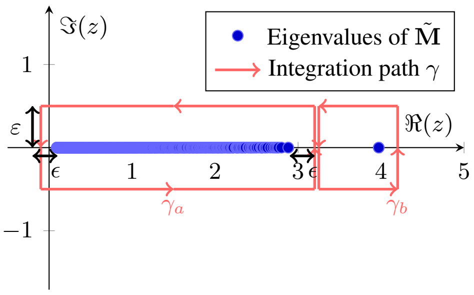


Figure 2: Eigenvalue distribution of $\tilde{\mathbf{M}}$ with the "main bulk" surrounded by γ_a and some isolated eigenvalues encompassed by γ_b.

C.2. Precise high-dimensional training dynamics

The proof is organized following the process introduced in Appendix C.1, and we will still introduce each step separately.

Cauchy's integral formula Recall results in Appendix B, we have that

$$\begin{aligned}
E_t &= \frac{1}{2n}\|f_t(\mathbf{X})-\mathbf{y}\|_2^2 = E_t^a + E_t^b + E_t^c = \tilde{E}_t^a + \tilde{E}_t^c + o(1) \\
&= \underbrace{\operatorname{tr}\left(\frac{1}{2n}e^{-2\eta t\cdot\tilde{\mathbf{K}}_{\mathrm{NTK},L}}\cdot\tilde{\mathbf{K}}_{\mathrm{CK}}(\mathbf{X},\mathbf{X})\right)}_{\tilde{E}_t^a} + \underbrace{\frac{1}{2n}\mathbf{y}^\top\cdot e^{-2\eta t\cdot\tilde{\mathbf{K}}_{\mathrm{NTK},L}}\cdot\mathbf{y}}_{\tilde{E}_t^c} + o(1).
\end{aligned}$$

And then substituting $\tilde{\mathbf{K}}_{\mathrm{NTK},L}, \tilde{\mathbf{K}}_{\mathrm{CK}}(\mathbf{X},\mathbf{X})$ with expressions in Theorem 1 and Theorem 2, we get:

$$\begin{aligned}
\tilde{E}_t^a &= \operatorname{tr}\left(\frac{1}{2n}e^{-2\eta t\cdot\tilde{\mathbf{K}}_{\mathrm{NTK},L}}\cdot\tilde{\mathbf{K}}_{\mathrm{CK}}(\mathbf{X},\mathbf{X})\right) = \operatorname{tr}\left(e^{-2\eta t\left(\beta_{L,1}\mathbf{X}^\top\mathbf{X}+\beta_{L,2}\frac{1}{p^2}\boldsymbol{\psi}\boldsymbol{\psi}^\top+\beta_{L,3}\frac{1}{p}\mathbf{1}_n\mathbf{1}_n^\top+\beta_{L,0}\mathbf{I}_n\right)}\cdot\right. \\
&\quad \left.\left[\frac{1}{2n}\left(\alpha_{L,1}\mathbf{X}^\top\mathbf{X}+\alpha_{L,2}\frac{1}{p}\boldsymbol{\psi}\boldsymbol{\psi}^\top+\alpha_{L,3}\frac{1}{p}\mathbf{1}_n\mathbf{1}_n^\top+\alpha_{L,0}\mathbf{I}_n\right)\right]\right) \\
&= \operatorname{tr}\left(e^{-2\eta t\beta_{L,0}\mathbf{I}_n}e^{-2\eta t\beta_{L,1}\left(\mathbf{A}^\top\mathbf{A}\right)}\cdot\left[\frac{1}{2n}\left(\frac{\alpha_{L,1}}{\beta_{L,1}}\left(\beta_{L,1}\mathbf{A}^\top\mathbf{A}-\beta_{L,2}\frac{1}{p}\boldsymbol{\psi}\boldsymbol{\psi}^\top-\beta_{L,3}\frac{1}{p}\mathbf{1}_n\mathbf{1}_n^\top\right)\right.\right.\right. \\
&\quad \left.\left.\left.+\alpha_{L,2}\frac{1}{p}\boldsymbol{\psi}\boldsymbol{\psi}^\top+\alpha_{L,3}\frac{1}{p}\mathbf{1}_n\mathbf{1}_n^\top+\alpha_{L,0}\mathbf{I}_n\right)\right]\right) \\
&= \operatorname{tr}\left(e^{-2\eta t\beta_{L,0}\mathbf{I}_n}e^{-2\eta t\beta_{L,1}\left(\mathbf{A}^\top\mathbf{A}\right)}\cdot\left(\frac{1}{2n}\alpha_{L,1}\mathbf{A}^\top\mathbf{A}+\frac{1}{2n}\alpha_{L,0}\mathbf{I}_n\right.\right. \\
&\quad \left.\left.+\frac{1}{2n}\left(\alpha_{L,2}-\frac{\alpha_{L,1}\beta_{L,2}}{\beta_{L,1}}\right)\frac{1}{p}\boldsymbol{\psi}\boldsymbol{\psi}^\top+\frac{1}{2n}\left(\alpha_{L,3}-\frac{\alpha_{L,1}\beta_{L,3}}{\beta_{L,1}}\right)\frac{1}{p}\mathbf{1}_n\mathbf{1}_n^\top\right)\right)+o(1) \\
&= e^{-2\eta\beta_{L,0}t}\left(\frac{1}{2n}\operatorname{tr}\left(\alpha_{L,1}e^{-2\eta t\beta_{L,1}\left(\mathbf{A}^\top\mathbf{A}\right)}\cdot\mathbf{A}^\top\mathbf{A}\right)+\frac{1}{2n}\operatorname{tr}\left(\alpha_{L,0}e^{-2\eta t\beta_{L,1}\left(\mathbf{A}^\top\mathbf{A}\right)}\cdot\mathbf{I}_n\right)\right)
\end{aligned}$$

$$+\, e^{-2\eta\beta_{L,0}t}\cdot (\underbrace{\frac{1}{2n}\cdot\frac{1}{p}\boldsymbol{\psi}^\top\left(\alpha_{L,2}-\frac{\alpha_{L,1}\beta_{L,2}}{\beta_{L,1}}\right)\boldsymbol{\psi}}_{o(n^{-1}) \text{ per Lemma 1}} + \underbrace{\frac{1}{2n}\cdot\frac{1}{p}\mathbf{1}_n^\top\left(\alpha_{L,3}-\frac{\alpha_{L,1}\beta_{L,3}}{\beta_{L,1}}\right)\mathbf{1}_n)}_{o(n^{-1}) \text{ per Lemma 1}}$$

$$= e^{-2\eta\beta_{L,0}t}\left[\frac{1}{2n}\operatorname{tr}\left(\alpha_{L,1}e^{-2\eta t\beta_{L,1}\left(\mathbf{A}^\top\mathbf{A}\right)}\cdot\mathbf{A}^\top\mathbf{A}\right)+\frac{1}{2n}\operatorname{tr}\left(\alpha_{L,0}e^{-2\eta t\beta_{L,1}\left(\mathbf{A}^\top\mathbf{A}\right)}\cdot\mathbf{I}_n\right)\right]+o(1)$$

where we denote $\beta_{L,1}\mathbf{A}^\top\mathbf{A}=\beta_{L,1}\mathbf{X}^\top\mathbf{X}+\beta_{L,2}\frac{1}{p}\boldsymbol{\psi}\boldsymbol{\psi}^\top+\beta_{L,3}\frac{1}{p}\mathbf{1}_n\mathbf{1}_n^\top$, and

$$\tilde{E}_t^c=\frac{1}{2n}\mathbf{y}^\top\cdot e^{-2\eta t\cdot\tilde{\mathbf{K}}_{\mathrm{NTK},L}}\cdot\mathbf{y}=\frac{1}{2n}\mathbf{y}^\top\cdot e^{-2\eta t\left(\beta_{L,1}\mathbf{X}^\top\mathbf{X}+\beta_{L,2}\frac{1}{p^2}\boldsymbol{\psi}\boldsymbol{\psi}^\top+\beta_{L,3}\frac{1}{p}\mathbf{1}_n\mathbf{1}_n^\top+\beta_{L,0}\mathbf{I}_n\right)}\cdot\mathbf{y}$$
$$=e^{-2\eta\beta_{L,0}t}\cdot\frac{1}{2n}\mathbf{y}^\top e^{-2\eta t\beta_{L,1}\left(\mathbf{A}^\top\mathbf{A}\right)}\cdot\mathbf{y}$$

And therefore we have

$$E_t=\tilde{E}_t^a+\tilde{E}_t^c+o(1)=\frac{1}{2n}e^{-2\eta\beta_{L,0}t}\operatorname{tr}\left(e^{-2\eta t\beta_{L,1}\left(\mathbf{A}^\top\mathbf{A}\right)}\cdot\left(\alpha_{L,1}\mathbf{A}^\top\mathbf{A}+\alpha_{L,0}\mathbf{I}_n\right)\right)$$
$$+\,e^{-2\eta\beta_{L,0}t}\cdot\frac{1}{2n}\mathbf{y}^\top e^{-2\eta t\beta_{L,1}\left(\mathbf{A}^\top\mathbf{A}\right)}\cdot\mathbf{y}+o(1)$$

And then with Cauchy's integral formula, we have:

$$E_t=-\frac{1}{2\pi\imath}\cdot\frac{1}{2n}e^{-2\eta\beta_{L,0}t}\oint_\gamma\left(\alpha_{L,1}z+\alpha_{L,0}\right)e^{-2\eta\beta_{L,1}tz}\cdot\operatorname{tr}\left(\left(\mathbf{A}^\top\mathbf{A}-z\mathbf{I}_n\right)^{-1}\right)dz$$
$$-\frac{1}{2\pi\imath}\cdot\frac{1}{2n}e^{-2\eta\beta_{L,0}t}\oint_\gamma e^{-2\eta\beta_{L,1}tz}\cdot\mathbf{y}^\top\cdot\left(\mathbf{A}^\top\mathbf{A}-z\mathbf{I}_n\right)^{-1}\cdot\mathbf{y}dz+o(1)$$
$$=\underbrace{-\frac{1}{2\pi\imath}\cdot e^{-2\eta\beta_{L,0}t}\oint_\gamma\left(\alpha_{L,1}z+\alpha_{L,0}\right)e^{-2\eta\beta_{L,1}tz}\cdot\boxed{\frac{1}{2n}\operatorname{tr}\left(\mathbf{Q}_{\mathbf{A}^\top\mathbf{A}}(z)\right)}dz}_{\tilde{E}_t^a}$$
$$\underbrace{-\frac{1}{2\pi\imath}\cdot e^{-2\eta\beta_{L,0}t}\oint_\gamma e^{-2\eta\beta_{L,1}tz}\cdot\boxed{\frac{1}{2n}\mathbf{y}^\top\cdot\mathbf{Q}_{\mathbf{A}^\top\mathbf{A}}(z)\cdot\mathbf{y}}dz}_{\tilde{E}_t^c}+o(1)\tag{33}$$

with and λ a contour containing all eigenvalues of $\mathbf{A}^\top\mathbf{A}$.

Remark 1 (On the limitations of β and its consequences on activation functions.) *In [5], different datasets may require different activation functions to achieve optimal classification performance. In particular, datasets with distinct means may require activations with corresponding non-zero $\beta_{\ell,1}$, while those with distinct variances may necessitate activations with non-zero $\beta_{\ell,2}$ or $\beta_{\ell,3}$. For a general analysis, we restrict $\beta_{\ell,1}$ to non-zero in Appendix C.*

And recall settings in (1), we are considering binary-class GMM data

$$\mathbf{X}=\frac{1}{\sqrt{p}}\cdot\left(\boldsymbol{\mu}\cdot\mathbf{y}^\top+\mathbf{Z}\right)\in\mathbb{R}^{p\times n},\tag{34}$$

Deterministic equivalence for $\mathbf{Q}_{\mathbf{A}^\top\mathbf{A}}(z)$ The subsequent procedure is to find the deterministic equivalent for $\mathbf{Q}_{\mathbf{A}^\top\mathbf{A}}(z)$. As declared in Appendix C.1, we leverage Woodbury Theorem to convert the identification of the deterministic equivalent of $\mathbf{Q}_{\mathbf{A}^\top\mathbf{A}}(z)$ to the deterministic equivalents of $\mathbf{Q}_{\frac{1}{p}\mathbf{Z}^\top\mathbf{Z}}(z)$, shown below.

$$
\begin{aligned}
&\mathbf{Q}_{\mathbf{A}^\top\mathbf{A}}(z) = \left(\mathbf{A}^\top\mathbf{A} - z\mathbf{I}_n\right)^{-1} \qquad (35)\\
&= \left(\frac{1}{\beta_{L,1}}\left(\beta_{L,1}\mathbf{X}^\top\mathbf{X} + \beta_{L,2}\frac{1}{p}\boldsymbol{\psi}\boldsymbol{\psi}^\top + \beta_{L,3}\frac{1}{p}\mathbf{1}_n\mathbf{1}_n^\top\right) - z\mathbf{I}_n\right)^{-1}\\
&= \left(\frac{1}{\beta_{L,1}}\left(\begin{bmatrix}\frac{1}{\sqrt{p}}\mathbf{y} & \frac{1}{\sqrt{p}}\mathbf{Z}^\top\boldsymbol{\mu}\end{bmatrix}\begin{bmatrix}\beta_{L,1}\|\boldsymbol{\mu}\|^2 & \beta_{L,1}\\ \beta_{L,1} & 0\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{p}}\mathbf{y}^\top\\ \frac{1}{\sqrt{p}}\boldsymbol{\mu}^\top\mathbf{Z}\end{bmatrix} + \beta_{L,2}\frac{1}{p}\boldsymbol{\psi}\boldsymbol{\psi}^\top + \beta_{L,3}\frac{1}{p}\mathbf{1}_n\mathbf{1}_n^\top\right)\right.\\
&\qquad\left. + \frac{1}{p}\mathbf{Z}^\top\mathbf{Z} - z\mathbf{I}_n\right)^{-1}\\
&= \left(\begin{bmatrix}\frac{1}{\sqrt{p}}\mathbf{y} & \frac{1}{\sqrt{p}}\mathbf{Z}^\top\boldsymbol{\mu} & \frac{1}{\sqrt{p}}\boldsymbol{\psi} & \frac{1}{\sqrt{p}}\mathbf{1}_n\end{bmatrix}\begin{bmatrix}\|\boldsymbol{\mu}\|^2 & 1 & 0 & 0\\ 1 & 0 & 0 & 0\\ 0 & 0 & \frac{\beta_{L,2}}{\beta_{L,1}} & 0\\ 0 & 0 & 0 & \frac{\beta_{L,3}}{\beta_{L,1}}\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{p}}\mathbf{y}^\top\\ \frac{1}{\sqrt{p}}\boldsymbol{\mu}^\top\mathbf{Z}\\ \frac{1}{\sqrt{p}}\boldsymbol{\psi}^\top\\ \frac{1}{\sqrt{p}}\mathbf{1}_n^\top\end{bmatrix} + \frac{1}{p}\mathbf{Z}^\top\mathbf{Z} - z\mathbf{I}_n\right)^{-1}\\
&= \mathbf{Q}(z) - \mathbf{Q}(z)\begin{bmatrix}\frac{1}{\sqrt{p}}\mathbf{y} & \frac{1}{\sqrt{p}}\mathbf{Z}^\top\boldsymbol{\mu} & \frac{1}{\sqrt{p}}\boldsymbol{\psi} & \frac{1}{\sqrt{p}}\mathbf{1}_n\end{bmatrix}\cdot\\
&\quad\begin{bmatrix}\frac{1}{p}\mathbf{y}^\top\mathbf{Q}(z)\mathbf{y} & \frac{1}{p}\mathbf{y}^\top\mathbf{Q}(z)\mathbf{Z}^\top\boldsymbol{\mu} + 1 & \frac{1}{p}\mathbf{y}^\top\mathbf{Q}(z)\boldsymbol{\psi} & \frac{1}{p}\mathbf{y}^\top\mathbf{Q}(z)\mathbf{1}_n\\ \frac{1}{p}\boldsymbol{\mu}^\top\mathbf{Z}\mathbf{Q}(z)\mathbf{y} + 1 & \frac{1}{p}\boldsymbol{\mu}^\top\mathbf{Z}\mathbf{Q}(z)\mathbf{Z}^\top\boldsymbol{\mu} - \|\boldsymbol{\mu}\|^2 & \frac{1}{p}\boldsymbol{\mu}^\top\mathbf{Z}\mathbf{Q}(z)\boldsymbol{\psi} & \frac{1}{p}\boldsymbol{\mu}^\top\mathbf{Z}\mathbf{Q}(z)\mathbf{1}_n\\ \frac{1}{p}\boldsymbol{\psi}^\top\mathbf{Q}(z)\mathbf{y} & \frac{1}{p}\boldsymbol{\psi}^\top\mathbf{Q}(z)\mathbf{Z}^\top\boldsymbol{\mu} & \frac{1}{p}\boldsymbol{\psi}^\top\mathbf{Q}(z)\boldsymbol{\psi} + \frac{\beta_{L,1}}{\beta_{L,2}} & \frac{1}{p}\boldsymbol{\psi}^\top\mathbf{Q}(z)\mathbf{1}_n\\ \frac{1}{p}\mathbf{1}_n^\top\mathbf{Q}(z)\mathbf{y} & \frac{1}{p}\mathbf{1}_n^\top\mathbf{Q}(z)\mathbf{Z}^\top\boldsymbol{\mu} & \frac{1}{p}\mathbf{1}_n^\top\mathbf{Q}(z)\boldsymbol{\psi} & \frac{1}{p}\mathbf{1}_n^\top\mathbf{Q}(z)\mathbf{1}_n + \frac{\beta_{L,1}}{\beta_{L,3}}\end{bmatrix}^{-1}\cdot\\
&\quad\begin{bmatrix}\frac{1}{\sqrt{p}}\mathbf{y}^\top\\ \frac{1}{\sqrt{p}}\boldsymbol{\mu}^\top\mathbf{Z}\\ \frac{1}{\sqrt{p}}\boldsymbol{\psi}^\top\\ \frac{1}{\sqrt{p}}\mathbf{1}_n^\top\end{bmatrix}\mathbf{Q}(z)
\end{aligned}
$$

for $\mathbf{Q}(z) = \left(\frac{1}{p}\mathbf{Z}^\top\mathbf{Z} - z\mathbf{I}_n\right)^{-1}$.

Then we resort to deterministic equivalent $\bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_n$ of $\mathbf{Q}(z)$ in Lemma 5, with $m(z)$ satisfying the following equation:

$$zcm^2(z) - (1 - c - z)m(z) + 1 = 0 \tag{36}$$

with $c = n/p$, and obtain:

$$
\begin{aligned}
&\tfrac{1}{p}\mathbf{y}^\top\mathbf{Q}(z)\mathbf{y} = cm(z), && \tfrac{1}{p}\mathbf{y}^\top\mathbf{Q}(z)\mathbf{Z}^\top\boldsymbol{\mu} = o(1), && \tfrac{1}{p}\mathbf{y}^\top\mathbf{Q}(z)\boldsymbol{\psi} = o(1),\\
&\tfrac{1}{p}\mathbf{y}^\top\mathbf{Q}(z)\mathbf{1}_n = \tfrac{1}{p}(n_a - n_b)m(z) = 0, && \tfrac{1}{p}\boldsymbol{\mu}^\top\mathbf{Z}\mathbf{Q}(z)\boldsymbol{\psi} = o(1), && \tfrac{1}{p}\boldsymbol{\mu}^\top\mathbf{Z}\mathbf{Q}(z)\mathbf{1}_n = o(1),\\
&\tfrac{1}{p}\boldsymbol{\psi}^\top\mathbf{Q}(z)\boldsymbol{\psi} = 2cm(z), && \tfrac{1}{p}\boldsymbol{\psi}^\top\mathbf{Q}(z)\mathbf{1}_n = o(1), && \tfrac{1}{p}\mathbf{1}_n^\top\mathbf{Q}(z)\mathbf{1}_n = cm(z),
\end{aligned}
$$

here we consider a "balanced sample" with $(n_a - n_b)$ in order $O(1)$ and thus $frac1p(n_a - n_b)m(z) = o(1)$, as well as

$$\begin{aligned}\frac{1}{p}\boldsymbol{\mu}^\top \mathbf{Z}\mathbf{Q}(z)\mathbf{Z}^\top \boldsymbol{\mu} = \frac{1}{p}\boldsymbol{\mu}^\top \tilde{\mathbf{Q}}(z)\mathbf{Z}\mathbf{Z}^\top \boldsymbol{\mu} &= \boldsymbol{\mu}^\top \tilde{\mathbf{Q}}(z)(\frac{1}{p}\mathbf{Z}\mathbf{Z}^\top - z\mathbf{I}_p + z\mathbf{I}_p)\boldsymbol{\mu} \\ &= \boldsymbol{\mu}^\top(\mathbf{I}_n + z\tilde{\mathbf{Q}}(z))\boldsymbol{\mu} = (\|\boldsymbol{\mu}\|^2\mathbf{I}_n + z\boldsymbol{\mu}^\top\tilde{\mathbf{Q}}(z)\boldsymbol{\mu}) \\ &= (\|\boldsymbol{\mu}\|^2 + z\tilde{m}(z)\|\boldsymbol{\mu}\|^2) \\ &= \|\boldsymbol{\mu}\|^2(1 + z\tilde{m}(z))\end{aligned}$$

for

$$\tilde{m}(z) = cm(z) + \frac{(c-1)}{z} \tag{37}$$

Then Equation (35) becomes:

$$\begin{aligned}&\mathbf{Q}_{\mathbf{A}^\top\mathbf{A}}(z) = \left(\mathbf{A}^\top\mathbf{A} - z\mathbf{I}_n\right)^{-1} \\ &= \mathbf{Q}(z) - \mathbf{Q}(z)\left[\tfrac{1}{\sqrt{p}}\mathbf{y} \quad \tfrac{1}{\sqrt{p}}\mathbf{Z}^\top\boldsymbol{\mu} \quad \tfrac{1}{\sqrt{p}}\boldsymbol{\psi} \quad \tfrac{1}{\sqrt{p}}\mathbf{1}_n\right]\cdot \\ &\begin{bmatrix} cm(z) & o(1)+1 & o(1) & o(1) \\ o(1)+1 & \|\boldsymbol{\mu}\|^2 z\tilde{m}(z) & o(1) & o(1) \\ o(1) & o(1) & 2cm(z) + \frac{\beta_{L,1}}{\beta_{L,2}} & o(1) \\ \frac{n_a - n_b}{p}m(z) & o(1) & o(1) & cm(z) + \frac{\beta_{L,1}}{\beta_{L,3}}\end{bmatrix}^{-1} \cdot \begin{bmatrix}\frac{1}{\sqrt{p}}\mathbf{y}^\top \\ \frac{1}{\sqrt{p}}\boldsymbol{\mu}^\top\mathbf{Z} \\ \frac{1}{\sqrt{p}}\boldsymbol{\psi}^\top \\ \frac{1}{\sqrt{p}}\mathbf{1}_n^\top\end{bmatrix}\mathbf{Q}(z) \\ &= \mathbf{Q}(z) - \mathbf{Q}(z)\left[\tfrac{1}{\sqrt{p}}\mathbf{y} \quad \tfrac{1}{\sqrt{p}}\mathbf{Z}^\top\boldsymbol{\mu} \quad \tfrac{1}{\sqrt{p}}\boldsymbol{\psi} \quad \tfrac{1}{\sqrt{p}}\mathbf{1}_n\right]\cdot \\ &\begin{bmatrix} cm(z) & 1 & 0 & 0 \\ 1 & \|\boldsymbol{\mu}\|^2 z\tilde{m}(z) & 0 & 0 \\ 0 & 0 & 2cm(z) + \frac{\beta_{L,1}}{\beta_{L,2}} & 0 \\ \frac{n_a - n_b}{p}m(z) & 0 & 0 & cm(z) + \frac{\beta_{L,1}}{\beta_{L,3}}\end{bmatrix}^{-1} \begin{bmatrix}\frac{1}{\sqrt{p}}\mathbf{y}^\top \\ \frac{1}{\sqrt{p}}\boldsymbol{\mu}^\top\mathbf{Z} \\ \frac{1}{\sqrt{p}}\boldsymbol{\psi}^\top \\ \frac{1}{\sqrt{p}}\mathbf{1}_n^\top\end{bmatrix}\mathbf{Q}(z) + o_{\|\cdot\|}(1) \\ &= \mathbf{Q}(z) - \mathbf{Q}(z)\left[\tfrac{1}{\sqrt{p}}\mathbf{y} \quad \tfrac{1}{\sqrt{p}}\mathbf{Z}^\top\boldsymbol{\mu} \quad \tfrac{1}{\sqrt{p}}\boldsymbol{\psi} \quad \tfrac{1}{\sqrt{p}}\mathbf{1}_n\right]\cdot\Lambda\cdot\begin{bmatrix}\frac{1}{\sqrt{p}}\mathbf{y}^\top \\ \frac{1}{\sqrt{p}}\boldsymbol{\mu}^\top\mathbf{Z} \\ \frac{1}{\sqrt{p}}\boldsymbol{\psi}^\top \\ \frac{1}{\sqrt{p}}\mathbf{1}_n^\top\end{bmatrix}\mathbf{Q}(z) + o_{\|\cdot\|}(1)\end{aligned} \tag{38}$$

in which we denote :

$$\Lambda = \begin{bmatrix}\frac{s_2}{-s_5^2 + s_1 s_2} & \frac{s_5}{s_5^2 - s_1 s_2} & 0 & 0 \\ \frac{s_5}{s_5^2 - s_1 s_2} & \frac{s_1}{-s_5^2 + s_1 s_2} & 0 & 0 \\ 0 & 0 & \frac{1}{s_3} & 0 \\ 0 & 0 & 0 & \frac{s_1 s_2 - s_5^2}{-s_4 s_5^2 + s_1 s_2 s_4}\end{bmatrix}, \tag{39}$$

for

$$s_1 = cm(z), \quad s_2 = \|\boldsymbol{\mu}\|^2 z\tilde{m}(z), \quad s_3 = 2cm(z) + \frac{\beta_{L,1}}{\beta_{L,2}},$$

$$s_4 = cm(z) + \frac{\beta_{L,1}}{\beta_{L,3}}, \quad s_5 = 1.$$

Thus for $\boxed{\frac{1}{2n}\mathbf{y}^\top \mathbf{Q}_{\mathbf{A}^\top\mathbf{A}}(z)\cdot\mathbf{y}}$, we have

$$\begin{aligned}
&\frac{1}{2n}\mathbf{y}^\top \mathbf{Q}_{\mathbf{A}^\top\mathbf{A}}(z)\cdot\mathbf{y} = \frac{1}{2n}\mathbf{y}^\top\left(\mathbf{A}^\top\mathbf{A} - z\mathbf{I}_n\right)^{-1}\cdot\mathbf{y}\\
&= \frac{1}{2n}\mathbf{y}^\top\mathbf{Q}(z)\mathbf{y} - \frac{1}{2c}\begin{bmatrix}\frac{1}{p}\mathbf{y}^\top\mathbf{Q}(z)\mathbf{y} & \frac{1}{p}\mathbf{y}^\top\mathbf{Q}(z)\mathbf{Z}^\top\boldsymbol{\mu} & \frac{1}{p}\mathbf{y}^\top\mathbf{Q}(z)\boldsymbol{\psi} & \frac{1}{p}\mathbf{y}^\top\mathbf{Q}(z)\mathbf{1}_n\end{bmatrix}\cdot\Lambda\cdot\\
&\begin{bmatrix}\frac{1}{p}\mathbf{y}^\top\mathbf{Q}(z)\mathbf{y}\\ \frac{1}{p}\boldsymbol{\mu}^\top\mathbf{Z}\mathbf{Q}(z)\mathbf{y}\\ \frac{1}{p}\boldsymbol{\psi}^\top\mathbf{Q}(z)\mathbf{y}\\ \frac{1}{p}\mathbf{1}_n^\top\mathbf{Q}(z)\mathbf{y}\end{bmatrix}\\
&= \frac{1}{2}m(z) - \frac{1}{2c}\begin{bmatrix}cm(z) & 0 & 0 & 0\end{bmatrix}\cdot\Lambda\cdot\begin{bmatrix}cm(z)\\0\\0\\0\end{bmatrix} + o(1)\\
&= \frac{1}{2}m(z) + \frac{cm^2(z)}{2}\cdot\frac{s_2}{-s_1s_2+s_5^2} = \frac{1}{2}m(z) + \frac{cm^2(z)}{2}\cdot\frac{s_2}{-s_1s_2+s_5^2}\\
&= \frac{1}{2}m(z) + \frac{cm^2(z)}{2}\cdot\frac{1}{-cm(z) - \frac{1}{\|\boldsymbol{\mu}\|^2}\cdot(cm(z)+1)}\\
&= \frac{1}{2}m(z) - \frac{cm^2(z)}{2}\cdot\frac{\|\boldsymbol{\mu}\|^2}{(1+\|\boldsymbol{\mu}\|^2)cm(z)-1}\\
&= \frac{1}{2}m(z)\cdot\left(1 - \frac{cm(z)\|\boldsymbol{\mu}\|^2}{(1+\|\boldsymbol{\mu}\|^2)cm(z)+1}\right)\\
&= \frac{1}{2}m(z)\cdot\frac{cm(z)+1}{(1+\|\boldsymbol{\mu}\|^2)cm(z)+1}
\end{aligned}$$

And for $\boxed{\frac{1}{2n}\operatorname{tr}\left(\mathbf{Q}_{\mathbf{A}^\top\mathbf{A}}(z)\right)}$, Lemma 3 shows that for $z\in\mathbb{C}\backslash\operatorname{support}\left(\boldsymbol{\mu}(\mathbf{A}\mathbf{A}^\top)\right)$:

$$\left|\frac{1}{2n}\operatorname{tr}\left(\left(\mathbf{A}^\top\mathbf{A} - z\mathbf{I}_n\right)^{-1}\right) - \frac{1}{2n}\operatorname{tr}\left(\mathbf{Q}(z)\right)\right| \le \underbrace{\frac{1}{2n}\frac{1}{\operatorname{dist}(z,\operatorname{support}(\boldsymbol{\mu}(\mathbf{A}\mathbf{A}^\top)))}}_{O(n^{-1})}.$$

For rectangular contour γ with its left and right sides slightly away from the edges of the "bulk", we can obtain:

$$\oint_\gamma \frac{1}{2n}\operatorname{tr}\left(\left(\mathbf{A}^\top\mathbf{A} - z\mathbf{I}_n\right)^{-1}\right) = \oint_\gamma \frac{1}{2n}\operatorname{tr}\left(\mathbf{Q}(z)\right) + o(1) = \oint_\gamma \frac{1}{2}m(z) + o(1)$$

Calculation of the eigenvalues and "Phase Transition" Then we narrow in on the complex integrations with the results above at hand, and we still focus our attention on each component separately below.

As declared in Appendix C.1, we need to obtain the isolated eigenvalues of $\mathbf{A}\mathbf{A}^\top$ first, shown below.

$$
\begin{aligned}
&\det\left(\mathbf{A}^\top\mathbf{A} - \lambda\mathbf{I}_n\right) = 0 \\
&\Leftrightarrow \det\left(\begin{bmatrix}\frac{1}{\sqrt{p}}\mathbf{y} & \frac{1}{\sqrt{p}}\mathbf{Z}^\top\boldsymbol{\mu} & \frac{1}{\sqrt{p}}\boldsymbol{\psi} & \frac{1}{\sqrt{p}}\mathbf{1}_n\end{bmatrix}\begin{bmatrix}\|\boldsymbol{\mu}\|^2 & 1 & 0 & 0\\ 1 & 0 & 0 & 0\\ 0 & 0 & \beta_{L,2}/\beta_{L,1} & 0\\ 0 & 0 & 0 & \beta_{L,3}/\beta_{L,1}\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{p}}\mathbf{y}^\top\\ \frac{1}{\sqrt{p}}\boldsymbol{\mu}^\top\mathbf{Z}\\ \frac{1}{\sqrt{p}}\boldsymbol{\psi}^\top\\ \frac{1}{\sqrt{p}}\mathbf{1}_n^\top\end{bmatrix}\right. \\
&\quad \left. + \frac{1}{p}\mathbf{Z}^\top\mathbf{Z} - \lambda\mathbf{I}_n\right) = 0 \\
&\Leftrightarrow \det\left(\mathbf{Q}(\lambda)\begin{bmatrix}\frac{1}{\sqrt{p}}\mathbf{y} & \frac{1}{\sqrt{p}}\mathbf{Z}^\top\boldsymbol{\mu} & \frac{1}{\sqrt{p}}\boldsymbol{\psi} & \frac{1}{\sqrt{p}}\mathbf{1}_n\end{bmatrix}\begin{bmatrix}\|\boldsymbol{\mu}\|^2 & 1 & 0 & 0\\ 1 & 0 & 0 & 0\\ 0 & 0 & \beta_{L,2}/\beta_{L,1} & 0\\ 0 & 0 & 0 & \beta_{L,3}/\beta_{L,1}\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{p}}\mathbf{y}^\top\\ \frac{1}{\sqrt{p}}\boldsymbol{\mu}^\top\mathbf{Z}\\ \frac{1}{\sqrt{p}}\boldsymbol{\psi}^\top\\ \frac{1}{\sqrt{p}}\mathbf{1}_n^\top\end{bmatrix}\right. \\
&\quad \left. + \mathbf{I}_n\right) \cdot \det\left(\frac{1}{p}\mathbf{Z}^\top\mathbf{Z} - \lambda\mathbf{I}_n\right) = 0 \\
&\Leftrightarrow \det\left(\begin{bmatrix}\|\boldsymbol{\mu}\|^2 & 1 & 0 & 0\\ 1 & 0 & 0 & 0\\ 0 & 0 & \beta_{L,2}/\beta_{L,1} & 0\\ 0 & 0 & 0 & \beta_{L,3}/\beta_{L,1}\end{bmatrix}\begin{bmatrix}\frac{1}{\sqrt{p}}\mathbf{y}^\top\\ \frac{1}{\sqrt{p}}\boldsymbol{\mu}^\top\mathbf{Z}\\ \frac{1}{\sqrt{p}}\boldsymbol{\psi}^\top\\ \frac{1}{\sqrt{p}}\mathbf{1}_n^\top\end{bmatrix}\mathbf{Q}(\lambda)\begin{bmatrix}\frac{1}{\sqrt{p}}\mathbf{y} & \frac{1}{\sqrt{p}}\mathbf{Z}^\top\boldsymbol{\mu} & \frac{1}{\sqrt{p}}\boldsymbol{\psi} & \frac{1}{\sqrt{p}}\mathbf{1}_n\end{bmatrix}\right. \\
&\quad \left. + \mathbf{I}_4\right) = 0 \\
&\Leftrightarrow \det\left(\mathbf{I}_4 + \begin{bmatrix}\|\boldsymbol{\mu}\|^2 & 1 & 0 & 0\\ 1 & 0 & 0 & 0\\ 0 & 0 & \beta_{L,2}/\beta_{L,1} & 0\\ 0 & 0 & 0 & \beta_{L,3}/\beta_{L,1}\end{bmatrix}\cdot\right. \\
&\quad \left.\begin{bmatrix}cm(\lambda) & 0 & 0 & 0\\ 0 & \|\boldsymbol{\mu}\|^2(1+\lambda\tilde{m}(\lambda)) & 0 & 0\\ 0 & 0 & 2cm(\lambda) & 0\\ 0 & 0 & 0 & cm(\lambda)\end{bmatrix}\right) = 0 \\
&\Leftrightarrow \det\begin{bmatrix}cm(\lambda)\|\boldsymbol{\mu}\|^2+1 & \|\boldsymbol{\mu}\|^2\lambda(1+\tilde{m}(\lambda)) & 0 & 0\\ cm(\lambda) & 1 & 0 & 0\\ 0 & 0 & 2cm(\lambda)\frac{\beta_{L,2}}{\beta_{L,1}}+1 & 0\\ 0 & 0 & 0 & cm(\lambda)\beta_{L,3}/\beta_{L,1}+1\end{bmatrix} = 0 \\
&\Leftrightarrow \left(1 + \frac{2c\beta_{L,2}m(\lambda)}{\beta_{L,1}}\right)\cdot\left(-\|\boldsymbol{\mu}\|^2(1+\lambda\tilde{m}(\lambda))\cdot\det\begin{bmatrix}cm(\lambda) & 0\\ 0 & cm(\lambda)\beta_{L,3}/\beta_{L,1}+1\end{bmatrix}\right.
\end{aligned}
$$

$$
\begin{aligned}
&+\det\begin{bmatrix} cm(\lambda)\|\boldsymbol{\mu}\|^2+1 & 0 \\ 0 & cm(\lambda)\beta_{L,3}/\beta_{L,1}+1\end{bmatrix}\Bigg)=0\\
\Leftrightarrow&\left(1+\frac{2c\beta_{L,2}m(\lambda)}{\beta_{L,1}}\right)\cdot\left(-\|\boldsymbol{\mu}\|^2(1+\lambda\tilde{m}(\lambda))\cdot\left(cm(\lambda)\cdot\left(cm(\lambda)\frac{\beta_{L,3}}{\beta_{L,1}}+1\right)\right)\right.\\
&\left.+\left(cm(\lambda)\|\boldsymbol{\mu}\|^2+1\right)\cdot\left(cm(\lambda)\frac{\beta_{L,3}}{\beta_{L,1}}+1\right)\right)=0\\
\Leftrightarrow&\left(1+\frac{2c\beta_{L,2}m(\lambda)}{\beta_{L,1}}\right)\cdot\left((cm(\lambda)\frac{\beta_{L,3}}{\beta_{L,1}}+1)-\lambda\tilde{m}(\lambda)\cdot(\|\boldsymbol{\mu}\|^2\cdot(cm(\lambda)\cdot\right.\\
&\left.\left(cm(\lambda)\frac{\beta_{L,3}}{\beta_{L,1}}+1\right)))\right)=0
\end{aligned}
$$

Refer to equation (37), and rearrange equation (36) as :

$$
(zm(z)+1)(cm(z)+1)=m(z). \tag{40}
$$

And therefore, we can get the following:

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

with $m(\lambda)\neq-\frac{1}{c}$. Thus the equation for the isolated eigenvalue mentioned earlier becomes:

$$
\begin{aligned}
&\left(1+\frac{2c\beta_{L,2}m(\lambda)}{\beta_{L,1}}\right)\cdot\left(\|\boldsymbol{\mu}\|^2\left(\frac{1}{cm(\lambda)+1}\right)\cdot\left(c^2m^2(\lambda)\beta_{L,3}/\beta_{L,1}+cm(\lambda)\right)\right.\\
&\quad\left.+cm(\lambda)\beta_{L,3}/\beta_{L,1}+1\right)=0.\\
&\left(1+\frac{2c\beta_{L,2}m(\lambda)}{\beta_{L,1}}\right)\cdot\left(\|\boldsymbol{\mu}\|^2\left(\frac{1}{cm(\lambda)+1}\right)\cdot(c^2m^2(\lambda)\frac{\beta_{L,3}}{\beta_{L,1}}+cm(\lambda))+cm(\lambda)\frac{\beta_{L,3}}{\beta_{L,1}}+1\right)=0.\\
\Leftrightarrow&\left(1+\frac{2c\beta_{L,2}m(\lambda)}{\beta_{L,1}}\right)\cdot\left(\|\boldsymbol{\mu}\|^2\left(\frac{cm(\lambda)}{cm(\lambda)+1}\right)+1\right)\cdot(cm(\lambda)\beta_{L,3}/\beta_{L,1}+1)=0\\
\Leftrightarrow&\left(1+\frac{2c\beta_{L,2}m(\lambda)}{\beta_{L,1}}\right)\cdot\left(\frac{\|\boldsymbol{\mu}\|^2cm(\lambda)+cm(\lambda)+1}{cm(\lambda)+1}\right)\cdot(cm(\lambda)\beta_{L,3}/\beta_{L,1}+1)=0
\end{aligned}
$$

Then we get the solution of the equation of eigenvalues as:

$$
m(\lambda_{\boldsymbol{\mu}})=-\frac{1}{c(\|\boldsymbol{\mu}\|^2+1)},\quad m(\lambda_{1_n})=-\frac{\beta_{L,1}}{c\beta_{L,3}},\quad m(\lambda_{\boldsymbol{\psi}})=-\frac{\beta_{L,1}}{2c\beta_{L,2}};
$$

then we shall consider the threshold of $m(\lambda)$ when the "phase transition" phenomenon [2] occurs, indicating that the corresponding λ is isolated from the "main-bulk", as declared in Appendix C.1 and Lemma 7, we thus get the corresponding eigenvalues as follows:

$$
\lambda_{\boldsymbol{\mu}}=\begin{cases}1+c+c\|\boldsymbol{\mu}\|^2+\frac{1}{\|\boldsymbol{\mu}\|^2} & \text{if } \|\boldsymbol{\mu}\|^2>\frac{1}{\sqrt{c}},\\ (1+\sqrt{c})^2 & \text{otherwise,}\end{cases},
$$

$$\lambda_{1_n} = \begin{cases} 1 + c + \frac{1}{\frac{\beta_{L,3}}{\beta_{L,1}} - 1} + c\left(\frac{\beta_{L,3}}{\beta_{L,1}} - 1\right) & \text{if } \frac{\beta_{L,3}}{\beta_{L,1}} > \frac{1}{\sqrt{c}} + 1, \\ (1+\sqrt{c})^2 & \text{otherwise,} \end{cases}, \tag{41}$$

$$\lambda_{\boldsymbol{\psi}} = \begin{cases} 1 + c + \frac{1}{\frac{2\beta_{L,2}}{\beta_{L,1}} - 1} + c\left(\frac{2\beta_{L,2}}{\beta_{L,1}} - 1\right) & \text{if } \frac{2\beta_{L,2}}{\beta_{L,1}} > \frac{1}{\sqrt{c}} + 1, \\ (1+\sqrt{c})^2 & \text{otherwise,} \end{cases}.$$

Integrals Then we are prepared for calculations of complex integral in (33), As declared in Appendix C.1, it is helpful to divide the eigenvalues into two groups and calculate them separately, the main bulk of the eigenvalues of matrix $\frac{1}{p}\mathbf{Z}^\top \mathbf{Z}$ (between $\lambda_- \equiv (1-\sqrt{c})^2$ and $\lambda_+ \equiv (1+\sqrt{c})^2$) of the Marčenko-Pastur distribution

$$\mu(dx) = \frac{\sqrt{(x-\lambda_-)^+ (\lambda_+ - x)^+}}{2\pi c x} dx + \left(1 - \frac{1}{c}\right)^+ \delta(x)),$$

and some isolated eigenvalues in (41). of which integration can be tackled by the residue theorem. And it boils down to the calculation of the following two items as mentioned above.

$$\begin{aligned} \frac{1}{2n}\|f_t(\mathbf{X}) - \mathbf{y}\|_2^2 &= -\frac{1}{2\pi\imath} \cdot e^{-2\eta\beta_{L,0}t} \oint_\gamma \left(\alpha_{L,1}z + \alpha_{L,0}\right) e^{-2\eta\beta_{L,1}tz} \cdot \boxed{\frac{1}{2n}\operatorname{tr}\left(\mathbf{Q}_{\mathbf{A}^\top\mathbf{A}}(z)\right)} dz \\ &- \frac{1}{2\pi\imath} \cdot e^{-2\eta\beta_{L,0}t} \oint_\gamma e^{-2\eta\beta_{L,1}tz} \cdot \boxed{\frac{1}{2n}\mathbf{y}^\top \cdot \mathbf{Q}_{\mathbf{A}^\top\mathbf{A}}(z) \cdot \mathbf{y}} dz + o(1) \\ &= \underbrace{-\frac{1}{2\pi i} \cdot e^{-2\eta\beta_{L,0}t} \oint_\gamma \left(\alpha_{L,1}z + \alpha_{L,0}\right) e^{-2\eta\beta_{L,1}tz} \cdot \frac{m(z)}{2} dz}_{\tilde{E}_t^a} \\ &+ \underbrace{-\frac{1}{2\pi i} \cdot e^{-2\eta\beta_{L,0}t} \oint_\gamma e^{-2\eta\beta_{L,1}tz} \cdot \left(\frac{m(z)}{2} \cdot \frac{cm(z)+1}{c(1+\|\boldsymbol{\mu}\|^2)m(z)+1}\right) dz}_{\tilde{E}_t^c} + o(1) \end{aligned} \tag{42}$$

with γ a contour encompassing all the eigenvalues of $\mathbf{A}^\top\mathbf{A}$, as mentioned above, we calculate the complex contour integrations by dividing the contour into two parts, one encompassing the "main-bulk" corresponding to the eigenvalues of $\frac{1}{p}\mathbf{Z}^\top\mathbf{Z}$, we call it γ_a and one surrounding the isolated eigenvalues which we have calculated in (41), we call it γ_b, and additionally, if $c > 1$, there will be isolated eigenvalue at 0 (see Lemma 5 for details), and we also consider this. We still compute $\tilde{E}_t^a$ and $\tilde{E}_t^c$ separately.

$$\begin{aligned} \tilde{E}_t^a &= -\frac{1}{2\pi i} \cdot e^{-2\eta\beta_{L,0}t} \oint_\gamma \left(\alpha_{L,1}z + \alpha_{L,0}\right) e^{-2\eta\beta_{L,1}tz} \cdot \frac{m(z)}{2} dz \\ &= \underbrace{-\frac{1}{2\pi i} \cdot e^{-2\eta\beta_{L,0}t} \oint_{\gamma_a} \left(\alpha_{L,1}z + \alpha_{L,0}\right) e^{-2\eta\beta_{L,1}tz} \cdot \frac{m(z)}{2} dz}_{\tilde{E}_t^{aa}} \\ &+ \underbrace{-\frac{1}{2\pi i} \cdot e^{-2\eta\beta_{L,0}t} \oint_{\gamma_b} \left(\alpha_{L,1}z + \alpha_{L,0}\right) e^{-2\eta\beta_{L,1}tz} \cdot \frac{m(z)}{2} dz,}_{\tilde{E}_t^{ab}} \end{aligned}$$

and

$$\begin{aligned}
\tilde{E}_t^{c} &= -\frac{1}{2\pi i}\cdot e^{-2\eta\beta_{L,0}t}\oint_{\gamma} e^{-2\eta\beta_{L,1}tz}\cdot\left(\frac{m(z)}{2}\cdot\frac{cm(z)+1}{c(1+\|\boldsymbol{\mu}\|^2)m(z)+1}\right)dz\\
&= \underbrace{-\frac{1}{2\pi i}\cdot e^{-2\eta\beta_{L,0}t}\oint_{\gamma_a} e^{-2\eta\beta_{L,1}tz}\cdot\left(\frac{m(z)}{2}\cdot\frac{cm(z)+1}{c(1+\|\boldsymbol{\mu}\|^2)m(z)+1}\right)dz}_{\tilde{E}_t^{ca}}\\
&\quad+\underbrace{-\frac{1}{2\pi i}\cdot e^{-2\eta\beta_{L,0}t}\oint_{\gamma_b} e^{-2\eta\beta_{L,1}tz}\cdot\left(\frac{m(z)}{2}\cdot\frac{cm(z)+1}{c(1+\|\boldsymbol{\mu}\|^2)m(z)+1}\right)dz}_{\tilde{E}_t^{cb}}\,.
\end{aligned}$$

Here we recall expressions of $m(z)$ as:

$$m(z)=\frac{1-c-z}{2cz}\pm\frac{\imath}{2cz}\sqrt{4cz-(1-c-z)^2}=\frac{1-c-z}{2cz}\pm\frac{\imath}{2cz}\sqrt{\left(z-\lambda_-\right)\left(\lambda_+-z\right)}$$

and thus for z extremely close to the real axis, we have

$$\Re(m(z))=\frac{1-c-z}{2cz},\qquad\qquad \Im(m(z))=\frac{1}{2cz}\sqrt{\left(z-\lambda_-\right)\left(\lambda_+-z\right)}$$

with the branch of ± is determined by the imaginary part of z such that $\Im(z)\cdot\Im m(z)>0$, illustrated in Remark 2. And thus we can get

$$\begin{aligned}
\tilde{E}_t^{aa} &= -\frac{1}{2\pi i}\cdot e^{-2\eta\beta_{L,0}t}\oint_{\gamma_a}\left(\alpha_{L,1}z+\alpha_{L,0}\right)e^{-2\eta\beta_{L,1}tz}\cdot\frac{m(z)}{2}dz\\
&= -\frac{1}{\pi}\cdot e^{-2\eta\beta_{L,0}t}\int_{\lambda_-}^{\lambda_+}\left(\alpha_{L,1}x+\alpha_{L,0}\right)e^{-2\eta\beta_{L,1}tx}\cdot\frac{-\Im m(x)}{2}dx\\
&\quad+\underbrace{-e^{-2\eta\beta_{L,0}t}\lim_{z\to 0}(z-0)\cdot e^{-2\eta\beta_{L,1}tz}\cdot\left(\alpha_{L,1}z+\alpha_{L,0}\right)\cdot\frac{m(z)}{2}}_{\text{per Residue Theorem}}\quad\text{if } c>1\\
&= \frac{1}{2}\cdot e^{-2\eta\beta_{L,0}t}\int_{\lambda_-}^{\lambda_+}e^{-2\eta\beta_{L,1}tx}\cdot\left(\alpha_{L,1}x+\alpha_{L,0}\right)\frac{\sqrt{\left(x-\lambda_-\right)\left(\lambda_+-x\right)}}{2\pi cx}dx\\
&\quad+e^{-2\eta\beta_{L,0}t}\alpha_{L,0}\cdot\frac{c-1}{2c}\quad\text{if } c>1
\end{aligned}$$

and per Residue Theorem, we have

$$\begin{aligned}
\tilde{E}_t^{ab} &= -\frac{1}{2\pi i}\cdot e^{-2\eta\beta_{L,0}t}\oint_{\gamma_b}\left(\alpha_{L,1}z+\alpha_{L,0}\right)e^{-2\eta\beta_{L,1}tz}\cdot\frac{m(z)}{2}dz\\
&= -\frac{1}{2\pi i}\cdot e^{-2\eta\beta_{L,0}t}\oint_{\gamma_{\lambda_{\boldsymbol{\mu}}}}\left(\alpha_{L,1}z+\alpha_{L,0}\right)e^{-2\eta\beta_{L,1}tz}\cdot\frac{m(z)}{2}dz\quad\text{if } \|\boldsymbol{\mu}\|^2>\frac{1}{\sqrt{c}},
\end{aligned}$$

$$
\begin{aligned}
&+ -\frac{1}{2\pi i}\cdot e^{-2\eta\beta_{L,0}t}\oint_{\gamma_{\lambda_{1_n}}}\left(\alpha_{L,1}z+\alpha_{L,0}\right)e^{-2\eta\beta_{L,1}tz}\cdot\frac{m(z)}{2}dz \quad \text{if } \frac{\beta_{L,3}}{\beta_{L,1}} > \frac{1}{\sqrt{c}}+1,\\
&+ -\frac{1}{2\pi i}\cdot e^{-2\eta\beta_{L,0}t}\oint_{\gamma_{\lambda_{\boldsymbol{\psi}}}}\left(\alpha_{L,1}z+\alpha_{L,0}\right)e^{-2\eta\beta_{L,1}tz}\cdot\frac{m(z)}{2}dz \quad \text{if } \frac{2\beta_{L,2}}{\beta_{L,1}} > \frac{1}{\sqrt{c}}+1,\\
= & -e^{-2\eta\beta_{L,0}t}\lim_{z\to\lambda_{\boldsymbol{\mu}}}\left(z-\lambda_{\boldsymbol{\mu}}\right)\left(\left(\alpha_{L,1}z+\alpha_{L,0}\right)e^{-2\eta\beta_{L,1}tz}\cdot\frac{m(z)}{2}\right) \quad \text{if } \|\boldsymbol{\mu}\|^2 > \frac{1}{\sqrt{c}},\\
&- e^{-2\eta\beta_{L,0}t}\lim_{z\to\lambda_{1_n}}\left(z-\lambda_{1_n}\right)\left(\alpha_{L,1}z+\alpha_{L,0}\right)e^{-2\eta\beta_{L,1}tz}\cdot\frac{m(z)}{2} \quad \text{if } \frac{\beta_{L,3}}{\beta_{L,1}} > \frac{1}{\sqrt{c}}+1,\\
&- e^{-2\eta\beta_{L,0}t}\lim_{z\to\lambda_{\boldsymbol{\psi}}}\left(z-\lambda_{\boldsymbol{\psi}}\right)\left(\alpha_{L,1}z+\alpha_{L,0}\right)e^{-2\eta\beta_{L,1}tz}\cdot\frac{m(z)}{2} \quad \text{if } \frac{2\beta_{L,2}}{\beta_{L,1}} > \frac{1}{\sqrt{c}}+1,\\
= & \,0,
\end{aligned}
$$

Similarly, when choose contour γ_a as declared in Appendix C.1, we have

$$
\begin{aligned}
\tilde{E}_t^{ca} = & -\frac{1}{2\pi i}\cdot e^{-2\eta\beta_{L,0}t}\oint_{\gamma_a} e^{-2\eta\beta_{L,1}tz}\cdot\left(\frac{m(z)}{2}\cdot\frac{cm(z)+1}{c(1+\|\boldsymbol{\mu}\|^2)m(z)+1}\right)dz\\
= & -\frac{1}{\pi}\cdot e^{-2\eta\beta_{L,0}t}\cdot\int_{\lambda_-}^{\lambda_+} e^{-2\eta\beta_{L,1}tz}\cdot\Im\left(\frac{m(x)}{2}\cdot\frac{cm(x)+1}{c(1+\|\boldsymbol{\mu}\|^2)m(x)+1}\right)dx\\
&+ \underbrace{-e^{-2\eta\beta_{L,0}t}\lim_{z\to 0}(z-0)\cdot e^{-2\eta\beta_{L,1}tz}\cdot\left(\frac{m(z)}{2}\cdot\frac{cm(z)+1}{c(1+\|\boldsymbol{\mu}\|^2)m(z)+1}\right)}_{\text{per Residue Theorem}} \quad \text{if } c>1\\
= & -\frac{1}{\pi}\cdot e^{-2\eta\beta_{L,0}t}\int_{\lambda_-}^{\lambda_+} e^{-2\eta\beta_{L,1}tx}\cdot\Im\left(\frac{m(x)}{2(1+\|\boldsymbol{\mu}\|^2)}+\frac{1}{2}\frac{(1-\frac{1}{1+\|\boldsymbol{\mu}\|^2})m(x)}{c(1+\|\boldsymbol{\mu}\|^2)m(x)+1}\right)dx\\
&- e^{-2\eta\beta_{L,0}t}\lim_{z\to 0}(z-0)\cdot e^{-2\eta\beta_{L,1}tz}\cdot\left(\frac{m(x)}{2\cdot(1+\|\boldsymbol{\mu}\|^2)}+\frac{1}{2}\frac{(1-\frac{1}{1+\|\boldsymbol{\mu}\|^2})m(x)}{c(1+\|\boldsymbol{\mu}\|^2)m(x)+1}\right) \quad \text{if } c>1\\
= & -\frac{1}{\pi}\cdot e^{-2\eta\beta_{L,0}t}\int_{\lambda_-}^{\lambda_+} e^{-2\eta\beta_{L,1}tx}\cdot\Bigg(\frac{-\Im m(x)}{2\cdot(1+\|\boldsymbol{\mu}\|^2)}\\
&+\frac{1}{2}\frac{\|\boldsymbol{\mu}\|^2}{1+\|\boldsymbol{\mu}\|^2}\Im\left(\frac{\Re m(x)-i\Im m(x)}{c(1+\|\boldsymbol{\mu}\|^2)\left(\Re m(x)-i\Im m(x)\right)+1}\right)\Bigg)dx\\
&- e^{-2\eta\beta_{L,0}t}\lim_{z\to 0}(z-0)\cdot e^{-2\eta\beta_{L,1}tz}\cdot\left(\frac{m(x)}{2\cdot(1+\|\boldsymbol{\mu}\|^2)}\right)+o(1) \quad \text{if } c>1\\
= & -\frac{1}{\pi}\cdot e^{-2\eta\beta_{L,0}t}\int_{\lambda_-}^{\lambda_+} e^{-2\eta\beta_{L,1}tx}\cdot\Bigg(\frac{-\Im m(x)}{2\cdot(1+\|\boldsymbol{\mu}\|^2)}\\
&+\frac{1}{2}\frac{\|\boldsymbol{\mu}\|^2}{1+\|\boldsymbol{\mu}\|^2}\frac{-\Im m(x)}{2c(1+\|\boldsymbol{\mu}\|^2)\Re m(x)+1+c^2(1+\|\boldsymbol{\mu}\|^2)^2\cdot\frac{1}{cx}}\Bigg)dx\\
&+ e^{-2\eta\beta_{L,0}t}\cdot\frac{c-1}{2c(1+\|\boldsymbol{\mu}\|^2)}+o(1) \quad \text{if } c>1\\
= & \frac{1}{2\pi(1+\|\boldsymbol{\mu}\|^2)}\cdot e^{-2\eta\beta_{L,0}t}\int_{\lambda_-}^{\lambda_+} e^{-2\eta\beta_{L,1}tx}\cdot\Im m(x)dx
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{\|\boldsymbol{\mu}\|^2 \cdot e^{-2\eta\beta_{L,0}t}}{2\pi(1+\|\boldsymbol{\mu}\|^2)} \int_{\lambda_-}^{\lambda_+} e^{-2\eta\beta_{L,1}tx} \cdot \frac{\Im m(x)}{2c(1+\|\boldsymbol{\mu}\|^2)\Re m(x) + 1 + c^2(1+\|\boldsymbol{\mu}\|^2)^2 \cdot \frac{1}{cx}} dx \\
&+ e^{-2\eta\beta_{L,0}t} \cdot \frac{c-1}{2c(1+\|\boldsymbol{\mu}\|^2)} + o(1) \quad \text{if } c > 1 \\
= &\frac{1}{2\pi(1+\|\boldsymbol{\mu}\|^2)} \cdot e^{-2\eta\beta_{L,0}t} \int_{\lambda_-}^{\lambda_+} e^{-2\eta\beta_{L,1}tx} \cdot \left(\frac{1}{2cx} \sqrt{(x-\lambda_-)(\lambda_+ - x)} \right) dz \\
&+ \frac{\|\boldsymbol{\mu}\|^2 \cdot e^{-2\eta\beta_{L,0}t}}{2\pi(1+\|\boldsymbol{\mu}\|^2)} \int_{\lambda_-}^{\lambda_+} e^{-2\eta\beta_{L,1}tx} \cdot \frac{\left(\frac{1}{2cx} \sqrt{(x-\lambda_-)(\lambda_+ - x)} \right)}{2c(1+\|\boldsymbol{\mu}\|^2)\left(\frac{1-c-x}{2cx}\right) + 1 + c^2(1+\|\boldsymbol{\mu}\|^2)^2 \cdot \frac{1}{cx}} dx \\
&+ e^{-2\eta\beta_{L,0}t} \cdot \frac{c-1}{2c(1+\|\boldsymbol{\mu}\|^2)} + o(1) \quad \text{if } c > 1 \\
= &\frac{1}{4\pi c(1+\|\boldsymbol{\mu}\|^2)} \cdot e^{-2\eta\beta_{L,0}t} \int_{\lambda_-}^{\lambda_+} e^{-2\eta\beta_{L,1}tx} \cdot \left(\frac{1}{x} \sqrt{(x-\lambda_-)(\lambda_+ - x)} \right) dx \\
&+ \frac{\|\boldsymbol{\mu}\|^2 \cdot e^{-2\eta\beta_{L,0}t}}{4\pi c(1+\|\boldsymbol{\mu}\|^2)} \int_{\lambda_-}^{\lambda_+} e^{-2\eta\beta_{L,1}tx} \cdot \frac{\sqrt{(x-\lambda_-)(\lambda_+ - x)}}{(1+\|\boldsymbol{\mu}\|^2)(1-c-x) + x + c(1+\|\boldsymbol{\mu}\|^2)^2} dx \\
&+ e^{-2\eta\beta_{L,0}t} \cdot \frac{c-1}{2c(1+\|\boldsymbol{\mu}\|^2)} + o(1) \quad \text{if } c > 1 \\
= &\frac{1}{2(1+\|\boldsymbol{\mu}\|^2)} \cdot e^{-2\eta\beta_{L,0}t} \int_{\lambda_-}^{\lambda_+} e^{-2\eta\beta_{L,1}tx} \cdot \frac{1}{2\pi cx} \sqrt{(x-\lambda_-)(\lambda_+ - x)} dx \\
&+ \frac{1}{2(1+\|\boldsymbol{\mu}\|^2)} \cdot e^{-2\eta\beta_{L,0}t} \int_{\lambda_-}^{\lambda_+} e^{-2\eta\beta_{L,1}tx} \cdot \frac{\sqrt{(x-\lambda_-)(\lambda_+ - x)}}{2\pi c(\lambda_{\boldsymbol{\mu}} - x)} dx \\
&+ e^{-2\eta\beta_{L,0}t} \cdot \frac{c-1}{2c(1+\|\boldsymbol{\mu}\|^2)} + o(1) \quad \text{if } c > 1
\end{aligned}
$$

and again, per Residue Theorem, we have

$$
\begin{aligned}
&\tilde{E}_t^{cb} = -\frac{1}{2\pi i} \cdot e^{-2\eta\beta_{L,0}t} \oint_{\gamma_b} e^{-2\eta\beta_{L,1}tz} \cdot \left(\frac{m(z)}{2} \cdot \frac{cm(z)+1}{c(1+\|\boldsymbol{\mu}\|^2)m(z)+1} \right) dz \\
&= -\frac{1}{2\pi i} \cdot e^{-2\eta\beta_{L,0}t} \oint_{\gamma_{\lambda_{\boldsymbol{\mu}}}} e^{-2\eta\beta_{L,1}tz} \cdot \left(\frac{m(z)}{2} \cdot \frac{cm(z)+1}{c(1+\|\boldsymbol{\mu}\|^2)m(z)+1} \right) dz \quad \text{if } \|\boldsymbol{\mu}\|^2 > \frac{1}{\sqrt{c}}, \\
&\quad - \frac{1}{2\pi i} \cdot e^{-2\eta\beta_{L,0}t} \oint_{\gamma_{\lambda_{1_n}}} e^{-2\eta\beta_{L,1}tz} \cdot \left(\frac{m(z)}{2} \cdot \frac{cm(z)+1}{c(1+\|\boldsymbol{\mu}\|^2)m(z)+1} \right) dz \quad \text{if } \frac{\beta_{L,3}}{\beta_{L,1}} > \frac{1}{\sqrt{c}} + 1, \\
&\quad - \frac{1}{2\pi i} \cdot e^{-2\eta\beta_{L,0}t} \oint_{\gamma_{\lambda_{\boldsymbol{\psi}}}} e^{-2\eta\beta_{L,1}tz} \cdot \left(\frac{m(z)}{2} \cdot \frac{cm(z)+1}{c(1+\|\boldsymbol{\mu}\|^2)m(z)+1} \right) dz \quad \text{if } \frac{2\beta_{L,2}}{\beta_{L,1}} > \frac{1}{\sqrt{c}} + 1, \\
&= -\frac{1}{2\pi i} \cdot e^{-2\eta\beta_{L,0}t} \oint_{\gamma_{\lambda_{\boldsymbol{\mu}}}} e^{-2\eta\beta_{L,1}tz} \cdot \left(\frac{m(z)}{2} \cdot \frac{cm(z)+1}{c(1+\|\boldsymbol{\mu}\|^2)m(z)+1} \right) dz \quad \text{if } \|\boldsymbol{\mu}\|^2 > \frac{1}{\sqrt{c}} \\
&\quad + 0 + 0 \\
&= -e^{-2\eta\beta_{L,0}t} \lim_{z \to \lambda_{\boldsymbol{\mu}}} (z - \lambda_{\boldsymbol{\mu}}) \cdot e^{-2\eta\beta_{L,1}tz} \cdot \left(\frac{m(z)}{2} \cdot \frac{cm(z)+1}{c(1+\|\boldsymbol{\mu}\|^2)m(z)+1} \right) \quad \text{if } \|\boldsymbol{\mu}\|^2 > \frac{1}{\sqrt{c}} \\
&= -e^{-2\eta\beta_{L,0}t} \cdot e^{-2\eta\beta_{L,1}\lambda_{\boldsymbol{\mu}}t} \cdot \left(\frac{m(\lambda_{\boldsymbol{\mu}})}{2} \cdot \frac{cm(\lambda_{\boldsymbol{\mu}})+1}{c(1+\|\boldsymbol{\mu}\|^2)m'(\lambda_{\boldsymbol{\mu}})} \right) \quad \text{if } \|\boldsymbol{\mu}\|^2 > \frac{1}{\sqrt{c}}
\end{aligned}
$$

$$
\begin{aligned}
&= -e^{-2\eta\beta_{L,0}t}\cdot e^{-2\eta\beta_{L,1}\lambda_{\boldsymbol{\mu}}t}\cdot \frac{-\frac{1}{c(\|\boldsymbol{\mu}\|^2+1)}\cdot\frac{\|\boldsymbol{\mu}\|^2}{(\|\boldsymbol{\mu}\|^2+1)}}{2c(1+\|\boldsymbol{\mu}\|^2)\left(\frac{\|\boldsymbol{\mu}\|^4}{(c\|\boldsymbol{\mu}\|^4-1)\cdot c(1+\|\boldsymbol{\mu}\|^2)^2}\right)} \quad \text{if } \|\boldsymbol{\mu}\|^2 > \frac{1}{\sqrt{c}} \\
&= -e^{-2\eta\beta_{L,0}t}\cdot e^{-2\eta\beta_{L,1}\lambda_{\boldsymbol{\mu}}t}\cdot \frac{-1}{2c(1+\|\boldsymbol{\mu}\|^2)\frac{\|\boldsymbol{\mu}\|^2}{(c\|\boldsymbol{\mu}\|^4-1)}} \quad \text{if } \|\boldsymbol{\mu}\|^2 > \frac{1}{\sqrt{c}} \\
&= -e^{-2\eta\beta_{L,0}t}\cdot e^{-2\eta\beta_{L,1}\lambda_{\boldsymbol{\mu}}t}\cdot \frac{-(c\|\boldsymbol{\mu}\|^4-1)}{2c(1+\|\boldsymbol{\mu}\|^2)\cdot\|\boldsymbol{\mu}\|^2} \quad \text{if } \|\boldsymbol{\mu}\|^2 > \frac{1}{\sqrt{c}}
\end{aligned}
$$

Note we calculate $m'(\lambda_{\boldsymbol{\mu}})$ by taking the derivative of both sides of the equation:

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

and we get:

$$
\begin{aligned}
1 &= \frac{m'(z)}{m^2(z)} - \frac{cm'(z)}{(cm(z)+1)^2}, \\
1 &= m'(z)\cdot\left(\frac{1}{m^2(z)} - \frac{c}{(cm(z)+1)^2}\right),
\end{aligned}
$$

then substitute $\lambda_{\boldsymbol{\mu}}$ in it, and additionally with

$$m(\lambda_{\boldsymbol{\mu}}) = -\frac{1}{c(\|\boldsymbol{\mu}\|^2+1)},$$

we get

$$m'(\lambda_{\boldsymbol{\mu}}) = \frac{\|\boldsymbol{\mu}\|^4}{(c\|\boldsymbol{\mu}\|^4-1)\cdot c(1+\|\boldsymbol{\mu}\|^2)^2}.$$

Remark 2 (The detail of determining the sign of the imaginary part of $m(z)$)
Take $z = (x+iy)$ with $y \uparrow 0$ for example, we have:

$$
\begin{aligned}
m(z) &= \frac{1-c-z}{2cz} \pm \frac{i}{2cz}\sqrt{(z-\lambda_-)(\lambda_+-z)} \\
&= \frac{1-c-(x+iy)}{2c(x+iy)} \pm \frac{i}{2c(x+iy)}\sqrt{((x+iy)-\lambda_-)(\lambda_+-(x+iy))} \\
&= \Re m(z) + \Im m(z)
\end{aligned}
$$

with

$$
\begin{aligned}
\Im m(z) &= \lim_{y\uparrow 0}\left(\frac{(c-1)y}{2c(x^2-y^2)} \pm \Im\left(\frac{i}{2c(x+iy)}\sqrt{(x-\lambda_-+iy)(\lambda_+-x-iy)}\right)\right) \\
&= \lim_{y\uparrow 0}\left(\frac{(c-1)y}{2c(x^2-y^2)} \pm \Im\left(\frac{ix+y}{2c(x^2-y^2)}\sqrt{(x-\lambda_-)(\lambda_+-x)+iy(\lambda_++\lambda_--2x)}\right)\right) \\
&= \Im\left(\pm\frac{i}{2cx}\sqrt{(x-\lambda_-)(\lambda_+-x)}\right) = \pm\frac{1}{2cx}\sqrt{(x-\lambda_-)(\lambda_+-x)}.
\end{aligned}
$$

then to satisfy $\Im(z)\cdot\Im m(z) > 0$, the minus is selected as the sign, and we finally get:

$$\Im m(z) = -\frac{1}{2cx}\sqrt{(x-\lambda_-)(\lambda_+-x)}, \quad \textit{for } z = (x+iy) \textit{ with } y\uparrow 0.$$

Combination and final result Combining results for E_t^{aa}, E_t^{ab}, E_t^{ca}, and E_t^{cb}, we finally get:

$$\begin{aligned}
E_t &= E_t^{aa} + E_t^{ab} + E_t^{ca} + E_t^{cb} \\
&= \frac{1}{2} \cdot e^{-2\eta\beta_{L,0}t} \int_{\lambda_-}^{\lambda_+} e^{-2\eta\beta_{L,1}tx} \cdot (\alpha_{L,1}x + \alpha_{L,0}) \frac{\sqrt{(x-\lambda_-)(\lambda_+ - x)}}{2\pi cx} dx \\
&+ \frac{1}{2(1+\|\boldsymbol{\mu}\|^2)} \cdot e^{-2\eta\beta_{L,0}t} \int_{\lambda_-}^{\lambda_+} e^{-2\eta\beta_{L,1}tx} \cdot \frac{1}{2\pi cx} \sqrt{(x-\lambda_-)(\lambda_+ - x)} dx \\
&+ \frac{1}{2(1+\|\boldsymbol{\mu}\|^2)} \cdot e^{-2\eta\beta_{L,0}t} \int_{\lambda_-}^{\lambda_+} e^{-2\eta\beta_{L,1}tx} \cdot \frac{\sqrt{(x-\lambda_-)(\lambda_+ - x)}}{2\pi c(\lambda_{\boldsymbol{\mu}} - x)} dx \\
&+ \frac{(\|\boldsymbol{\mu}\|^2 + 1/\sqrt{c})(\|\boldsymbol{\mu}\|^2 - 1/\sqrt{c})^+}{\|\boldsymbol{\mu}\|^4 + \|\boldsymbol{\mu}\|^2} + (\alpha_0 + (1+\|\boldsymbol{\mu}\|^2)^{-1})(1-c^{-1})^+ + o(1)
\end{aligned}$$

This eventually leads to $E_t - \bar{E}_t \to 0$ with

$$\begin{aligned}
\bar{E}_t = \frac{e^{-2\eta\beta_0 t}}{2} \int_{\lambda_-}^{\lambda_+} e^{-2\eta\beta_1 tx} \left[\alpha_1 x + \alpha_0 + \frac{1}{1+\|\boldsymbol{\mu}\|^2} + \frac{1}{1+\|\boldsymbol{\mu}\|^2} \cdot \frac{x}{\lambda_{\boldsymbol{\mu}} - x} \right] \frac{\sqrt{(x-\lambda_-)^+(\lambda_+ - x)^+}}{2\pi cx} dx \\
+ \frac{e^{-2\eta\beta_0 t}}{2} \left[\frac{(\|\boldsymbol{\mu}\|^2 + 1/\sqrt{c})(\|\boldsymbol{\mu}\|^2 - 1/\sqrt{c})^+}{\|\boldsymbol{\mu}\|^4 + \|\boldsymbol{\mu}\|^2} e^{-2\eta\beta_1\lambda_{\boldsymbol{\mu}} t} + (\alpha_0 + (1+\|\boldsymbol{\mu}\|^2)^{-1})(1-c^{-1})^+ \right],
\end{aligned}$$

where we recall $\lambda_{\boldsymbol{\mu}} \equiv 1 + c + c\|\boldsymbol{\mu}\|^2 + \|\boldsymbol{\mu}\|^{-2}$, which, by introducing the following two probability measures (similar to [12]) as defined in the statement of Theorem 4,

$$\mu(dx) = \frac{\sqrt{(x-\lambda_-)^+(\lambda_+ - x)^+}}{2\pi cx} dx + (1-c^{-1})^+ \delta_0(x), \tag{43}$$

$$\nu(dx) = \frac{\sqrt{(x-\lambda_-)^+(\lambda_+ - x)^+}}{2\pi c\|\boldsymbol{\mu}\|^2(\lambda_{\boldsymbol{\mu}} - x)} dx + \frac{(\|\boldsymbol{\mu}\|^4 - c^{-1})^+}{\|\boldsymbol{\mu}\|^4} \delta_{\lambda_{\boldsymbol{\mu}}}(x), \tag{44}$$

can be compactly written as

$$\boxed{\bar{E}_t = \frac{e^{-2\eta\beta_0 t}}{2} \int e^{-2\eta\beta_1 tx} \left[\left(\alpha_0 + \alpha_1 x + \frac{1}{1+\|\boldsymbol{\mu}\|^2} \right) \mu(dx) + \frac{\nu(dx)}{1+\|\boldsymbol{\mu}\|^{-2}} \right]}$$

with $(t)^+ \equiv \max(t, 0)$, the shortcuts $\alpha_k \equiv \alpha_{L,k}$, $\beta_k \equiv \beta_{L,k}$, $k \in \{0, 1\}$ as in Theorem 3. This concludes proof of Theorem 4.