Journal of Machine Learning Research 25 (2024) 1-49 Submitted 12/22; Revised 9/23; Published 1/24

Multiple Descent in the Multiple Random Feature Model

Xuran Meng XURANMENG@CONNECT.HKU.HK
Department of Statistics and Actuarial Science
The University of Hong Kong

Jianfeng Yao JEFFYAO@CUHK.EDU.CN
School of Data Science
The Chinese University of Hong Kong (Shenzhen)

Yuan Cao YUANCAO@HKU.HK
Department of Statistics and Actuarial Science
The University of Hong Kong

Editor: Daniel Hsu

Abstract

Recent works have demonstrated a *double descent* phenomenon in over-parameterized learning. Although this phenomenon has been investigated by recent works, it has not been fully understood in theory. In this paper, we investigate the multiple descent phenomenon in a class of multi-component prediction models. We first consider a "double random feature model" (DRFM) concatenating two types of random features, and study the excess risk achieved by the DRFM in ridge regression. We calculate the precise limit of the excess risk under the high dimensional framework where the training sample size, the dimension of data, and the dimension of random features tend to infinity proportionally. Based on the calculation, we further theoretically demonstrate that the risk curves of DRFMs can exhibit *triple descent*. We then provide a thorough experimental study to verify our theory. At last, we extend our study to the "multiple random feature model" (MRFM), and show that MRFMs ensembling K types of random features may exhibit $(K+1)$-*fold descent*. Our analysis points out that risk curves with a specific number of descent generally exist in learning multi-component prediction models.

Keywords: Over-parameterization, excess risk, multiple descent, double random feature model, multiple random feature model

1. Introduction

Modern machine learning models such as deep neural networks are usually highly over-parameterized so that they can be trained to exactly fit the training data. Such over-parameterized models have gained immense popularity and achieved state-of-the-art performance in various learning tasks. However, in classical statistical learning theory, over-parameterized models are believed to have high excess risks due to overfitting, and hence their success has not been fully explained in theory. This gap between theory and practice has motivated a number of recent works to study the success of over-parameterized models.

Recent works have pointed out a *double/multiple descent* phenomenon in over-parameterized learning: as the number of parameters in a model increases, the excess risk may increase and decrease multiple times (see Figure 1 for some examples). The double descent phenomenon was first demonstrated experimentally by Belkin et al. (2019) in random feature models, random forests and neural networks, and then studied theoretically by a series of works under different settings. Specifically, Belkin et al. (2020) theoretically demonstrated the double descent shape of the risk curve of the minimum norm predictor in learning linear models and Fourier series models. Wu and Xu (2020); Mel and Ganguli (2021); Hastie et al. (2022) studied the excess risk in linear regression under the setting where the dimension and sample size go to infinity preserving a fixed ratio, and showed that the risk decreases with respect to this ratio in the over-parameterized setting. Mei and Montanari (2022); Liao et al. (2020) further studied double descent in random feature models when the sample size, data dimension and the number of random features have fixed ratios and Adlam et al. (2022) extended the model by adding bias terms. Deng et al. (2022) studied double descent under logistic model. Emami et al. (2020) studied the asymptotic generalization error of generalized linear models. Liu et al. (2021) provided a precise characterization of generalization properties of high dimensional kernel ridge regression in both under-parameterized and over-parameterized regimes. Several recent works have also studied other learning settings under which the risk curves exhibit triple descent or multiple descent. Specifically, Mai et al. (2019) evaluated the asymptotic distribution of the logistic regression classifier in high dimension setting, and then provided the associated classification performance. Liang et al. (2020) gave an upper bound on the risk of the minimum-norm interpolants in a reproducing kernel Hilbert space and showed that it has a multiple descent shape with infinitely many peaks. Chen et al. (2021) showed that with different and well-designed data distributions in linear regression, the risk curve can have an arbitrary number of peaks at arbitrary locations as the data dimension increases. Mel and Ganguli (2021); Li and Wei (2021) showed that the risk curve of linear regression can exhibit multiple descent when learning anisotropic data. Adlam and Pennington (2020b) demonstrated triple descent for a specific random feature model associated with an over-parameterized two-layer neural network in the so-called "neural tangent kernel" (Jacot et al., 2018) regime. Misiakiewicz (2022); Xiao et al. (2022) showed that the risk curve of certain kernel predictors can exhibit multiple descent concerning the sample size and data dimension.

While recent works have provided valuable insights, the double, triple and multiple descent phenomena have not been fully understood in theory. Specifically, we note that various modern learning methods utilize multi-component predictors of a general form

$$f(\mathbf{x}) = f_1(\mathbf{x}) + f_2(\mathbf{x}) + \cdots + f_K(\mathbf{x}), \tag{1.1}$$

where $f_1(\mathbf{x}), \ldots, f_K(\mathbf{x})$ are individual prediction models. Such a multi-component formulation covers different learning methods. For example, ensemble methods (Hansen and Salamon, 1990; Dietterich, 2000; Krogh and Vedelsby, 1994) can naturally be formulated as (1.1); two-layer neural networks utilizing feature concatenation is also a summation of multiple components defined by different features; two-layer ResNet (He et al., 2016) models

can be formulated as (1.1) by treating the feedforward part and the skip-connection part of the model as two components; a class of semi-parametric methods consider a parametric component and a non-parametric component in the model (Zhao et al., 2016; Chernozhukov et al., 2018); Neural network models with the exact form of (1.1) can also be applied to solve partial differential equations (Liu, 2020).

In this work, we aim to study the double/multiple descent phenomenon in learning multi-component predictors. We theoretically demonstrate that

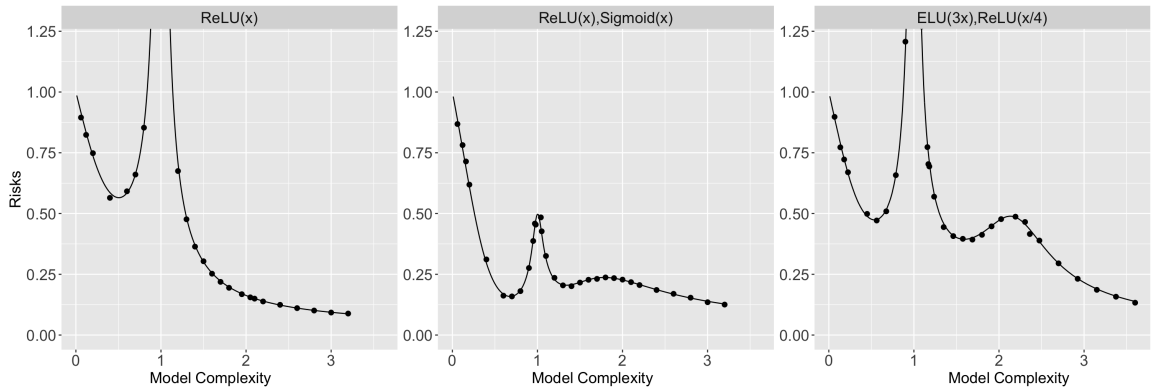
There exists a learning problem, such that for any $K \in \mathbb{N}_+$, there exists a K-component prediction model whose risk curve exhibits $(K+1)$-fold descent.

The learning problem mentioned in the claim above is the same learning problem where recent works have demonstrated double descent for random feature models Mei and Montanari (2022), and is also essentially the same learning problem (with slight modification) studied in Hastie et al. (2022) analyzing double descent in linear regression. Therefore, demonstrating this claim provides new insights into how complicated prediction model structures can affect the risk curve.

This paper aims to study the double/multiple descent phenomena in learning multi-component predictors of the form (1.1) through the lens of random feature models. Specifically, we introduce double and multiple random feature models (DRFMs and MRFMs), which ensemble two or more types of random features defined by different nonlinear activation functions. Under the setting where the training sample size, the dimension of data, and the dimension of random features tend to infinity proportionally, we establish an asymptotic limit of the excess risk achieved by DRFMs and MRFMs, and demonstrate that the risk curve of a DRFM can exhibit triple descent: an example for the DRFM with different activation functions is given in Figure 1. More generally, we also show that the risk curve of an MRFM with K types of random features can exhibit $(K+1)$-fold descent.

We summarize the contributions of this paper as follows.

1. Our first contribution is to demonstrate the existence of multiple descent in learning certain multi-component predictors. Specifically, we demonstrate that DRFMs may exhibit triple descent, and then extend the analysis to MRFMs and show that MRFMs consisting of K types of random features may have a risk curve with $(K+1)$-fold descent. To the best of our knowledge, such multiple descent risk curves with a specific number of peaks have not been well understood in random feature models or other multi-component learning models, and therefore we believe that DRFMs and MRFMs can serve as important examples in the literature of multiple descent.
2. We provide a natural and intuitive explanation of multiple descent in DRFMs and MRFMs. For example, for DRFMs, we point out that the existence of triple descent risk curves is predictable by considering the two extreme cases: (i) the DRFM uses two random features of the same type and scale, and (ii) one type of random feature in the DRFM has a very small scale and is thus negligible. This scale difference refers to a large gap in magnitude between the two random features, such as the activation pair $(\sigma_1(x), c_0\sigma_2(x))$ where the constant c_0 is small. We point out that these two cases both lead to double descent but with different peak locations. Therefore, for DRFMs where



(a) (b) (c)

Figure 1: Examples of double and triple descent. (a) gives the excess risk of a random feature model with ReLU activation function; (b) shows the excess risk of a double random feature model with ReLU and sigmoid activation functions; (c) shows the excess risk of a double random feature model with ELU and ReLU activation functions. The x-axis is the model complexity (number of parameters/sample size) and the y-axis is the excess risk. The curve gives our theoretical predictions, and the dots are our numerical results.

the scale difference between the two parts of random features is neither too big nor too small, we can expect triple descent to appear. Following this intuition, we successfully anticipate multiple descent in various simulations, and correctly predict the number of peaks in the risk curves and the locations of all peaks in the risk curves.

3. We also establish comprehensive theoretical results to back up our intuitive explanation. We calculate the precise limit of the excess risk achieved by DRFMs and MRFMs. This is an extension of the study of Mei and Montanari (2022) which analyzed the vanilla random feature model with a single activation function. We also establish a novel type of theoretical proof of multiple descent which is based on the comparison between excess risk values at different over-parameterization levels.

Our calculation of the theoretical limit of the excess risks of DRFMs and MRFMs follow the blueprint of Mei and Montanari (2022) and expand upon it by constructing new linear pencil matrices and giving new calculation for the related Stieltjes transforms. In essence, Mei and Montanari (2022) accomplished the following:

1. introduced a risk function decomposition and proved convergence in L_1 distance;
2. used a linear pencil matrix and its partial derivatives of logarithmic potential to express the decomposed terms;
3. provided an asymptotic approximation of the logarithmic potential and proved that partial derivatives are also approximated in L_1 distance;
4. calculated theoretical values from the asymptotic approximation.

Our theoretical analysis of the excess risk follow the decomposition method in 1, but due to the increased complexity of our model, we develop several new technical lemmas to overcome this higher complexity; such examples include Proposition A.2 in the Appendix and

Lemma I.3 in Online Supplementary. Moreover, the increased complexity in the main terms of the decomposition necessitates a more complex linear pencil matrix, as defined in Definition A.3 and C.3. Although the construction of the linear pencil matrices is inspired by Item 2 above, the higher complexity of our model results in a more intricate construction and a more complex calculation of the related Stieltjes transforms and their logarithmic potentials than in 3. Specifically, Proposition A.6 provides the calculation of the Stieltjes transforms in DRFM and serves as the inspiration for the calculation of the Stieltjes transforms in MRFM. In MRFM, we utilize mathematical induction to complete this calculation.

Besides the calculation of the theoretical limits of excess risks, this paper also presents a novel theory in the demonstration of multiple descent (given in Propositions 4.1 and 4.2). Instead of directly investigating the theoretical limits, our approach focuses on taking limits within specific parameter ranges to observe the resulting behavior. Specifically, we employ the following steps:

1. We give a fixed ratio between the number of training parameters and the sample size.
2. Within this ratio, we set the regularization parameter λ to approach zero, which allows us to approximate the implicit $\boldsymbol{\nu}$-system introduced later. We then replace the approximate solution with the theoretical limits.
3. To assess the impact of scale differences, we let one of the activation function scales tend towards zero, and examine the resulting theoretical limits.

By employing this method, we successfully utilize the $\varepsilon - \delta$ language to accurately depict the presence of two peaks and determine their precise locations.

The remaining of the paper is organized as follows. We first give some additional references and notations below. Section 2 introduces the problem settings. Section 3 establishes the theoretical limits of the excess risks of double random feature models. Section 4 gives theoretical analyses and simulations to demonstrate triple descent in some DRFMs. Section 5 extends the results to multiple random feature models and gives numerical simulations to demonstrate multiple descent. Finally, Section 6 concludes the paper and discusses some related questions for future investigation. Proofs of the main results and some additional experiments are presented in the appendix.

An online supplementary document (Meng et al.) is also available which gives some additional technical details of the proofs of the paper (with its sections numbered as I, II, III...).

1.1 Additional related works

Besides the works we previously discussed, a series of recent works have also studied the double and triple descent phenomena. Adlam and Pennington (2020a) developed a novel bias-variance decomposition, and utilized the decomposition to show double descent in random feature regression. d'Ascoli et al. (2020) developed a quantitative theory for the double descent phenomenon in the lazy learning regime of two-layer neural networks, and showed that overfitting is beneficial when the noise level in the data is low. Geiger et al. (2020) utilized the intuition of double descent to show that the smallest generalization error can sometimes be achieved by the ensemble of several neural networks of intermediate sizes.

Nakkiran et al. (2020); Patil et al. (2022) studied how an appropriately chosen parameters or suitable cross validation procedure can mitigate multiple descent in the prediction models. d'Ascoli et al. (2020) investigated the parameter-wise double descent and sample-wise triple descent phenomena in random feature regression. Deng et al. (2021) showed double descent phenomenon in logistic regression. Montanari and Zhong (2022) considered a two-layer neural network in the neural tangent regime, showed an interpolation phase transition, and gave a characterization of the generalization error which decreases with the number of training parameters.

Our paper is also closely related to the recent studies of the "benign overfitting" phenomenon. Tsigler and Bartlett (2023) showed that for certain regression problems, the risk achieved by the minimum norm linear interpolator can be asymptotically optimal. Bartlett et al. (2020) further extended the results in Tsigler and Bartlett (2023) to the setting of linear ridge regression. Chatterji and Long (2021) studied the risk of the maximum margin linear classifier in learning sub-Gaussian mixtures with additional label-flipping noises. Cao et al. (2021) established matching upper and lower bounds of the risk achieved by the maximum margin linear classifier. Frei et al. (2022) showed that fully-connected two-layer networks trained to achieve a zero training error can still achieve an asymptotically optimal test error. Cao et al. (2022) studied signal learning and noise memorization during the training of a two-layer convolutional neural network and revealed a phase transition between benign and harmful overfitting. Meng et al. (2023) further studied signal learning and noise memorization by two-layer convolutional neural networks when learning XOR data. Note that most studies along this line of research focus on the setting where the number of parameters N is much larger than the sample size n (e.g., $N = \Omega(n^2)$). In comparison, our work considers the setting where N and n go to infinity in comparable magnitudes, and studies how the excess risk changes with respect to their ratio.

1.2 Notations

We use lower case letters to denote scalars, and use bold face letters to denote vectors and matrices. For functions f, g and a probability measure ν, we denote $\langle f, g\rangle_\nu = \int f(\mathbf{x})g(\mathbf{x})\nu(d\mathbf{x})$. The ℓ_2-norm of a vector $\mathbf{v}$ is $\|\mathbf{v}\|_2$. For a matrix $\mathbf{A}$, we use $\|\mathbf{A}\|_\star$, $\|\mathbf{A}\|_{\max}$, $\|\mathbf{A}\|_{\mathrm{op}}$ and $\|\mathbf{A}\|_F$ to denote its nuclear norm, maximum norm, operator norm, and Frobinuous norm, respectively, and use $\mathrm{tr}(\mathbf{A})$ to denote its trace. A sub-matrix of $\mathbf{A}$ with row indices in I and column indices in J is denoted by $\mathbf{A}_{I,J}$, and $\mathrm{tr}_I(\mathbf{A}) = \mathrm{tr}(\mathbf{A}_{I,I})$ is the trace of the square sub-matrix with indices in I.

The sets of natural, real and complex numbers are denoted by $\mathbb{N}$, $\mathbb{R}$ and $\mathbb{C}$, respectively. For $z \in \mathbb{C}$, we use $\Re(z)$ and $\Im(z)$ to denote its real and imaginary part. $\mathbb{C}_+ = \{z \in \mathbb{C} : \Im(z) > 0\}$ denotes the upper half complex plane with positive imaginary part. Let $\mathrm{i} = \sqrt{-1}$ be the imaginary unit. The unit sphere of $\mathbb{R}^d$ is denoted by $\mathbb{S}^{d-1} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|_2 = 1\}$ and $c \cdot \mathbb{S}^{d-1}$ denotes the sphere with radius $c > 0$. The set of integers from n_1 to n_2 is denoted by $[n_1 : n_2] = \{n_1, \ldots, n_2\}$ and $[n] = [1 : n] = \{1, \ldots, n\}$. Moreover, $\mathbf{1}_q \in \mathbb{R}^q$ denotes q-dimensional all-one vectors.

We use the standard asymptotic notations $\Theta_d(\cdot)$, $O_d(\cdot)$, $o_d(\cdot)$ and $\Omega_d(\cdot)$, where the subscript d emphasizes the asymptotic variable. We write $X_1(d) = O_{\mathbb{P}}(X_2(d))$ if for any $\varepsilon > 0$, there exists $C > 0$ such that $\mathbb{P}(|X_1(d)/X_2(d)| > C) \le \varepsilon$ for all d. Similarly, we denote $X_1(d) = o_{\mathbb{P}}(X_2(d))$ if $\{X_1(d)/X_2(d)\}_d$ converges to 0 in probability.

2. The double random feature model

We consider regression problems where, for a data pair $(\mathbf{x}, y)$, the goal is to predict the scalar response y using the input vector $\mathbf{x} \in \mathbb{R}^d$. We analyze the prediction performance of a *double random feature model*, or DRFM, constructed as follows. The random features are based on two nonlinear activation functions σ_1, σ_2 and N random feature parameter vectors $\boldsymbol{\theta}_i \sim \mathrm{Unif}(\sqrt{d} \cdot \mathbb{S}^{d-1})$, $i \in [N]$. We let $a_i \in \mathbb{R}$, $i \in [N]$ be the linear combination coefficients of the random features, and denote $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N]^\top \in \mathbb{R}^{N \times d}$, $\mathbf{a} = [a_1, \dots, a_N]^\top \in \mathbb{R}^N$. Then a DRFM predictor has the form

$$\widehat{y} = f(\mathbf{x}; \mathbf{a}, \boldsymbol{\Theta}) = \sum_{i=1}^{N_1} a_i \sigma_1\big(\langle \boldsymbol{\theta}_i, \mathbf{x} \rangle / \sqrt{d}\big) + \sum_{i=N_1+1}^{N} a_i \sigma_2\big(\langle \boldsymbol{\theta}_i, \mathbf{x} \rangle / \sqrt{d}\big). \tag{2.1}$$

In (2.1), the first N_1 units use the activation function σ_1 and the first part of the random feature parameters $\boldsymbol{\Theta}_1 = [\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_{N_1}]^\top$, while the remaining $N_2 = N - N_1$ units use the second activation function σ_2 and the second part of the random feature parameters $\boldsymbol{\Theta}_2 = [\boldsymbol{\theta}_{N_1+1}, \dots, \boldsymbol{\theta}_N]^\top$. Note that the coefficients $a_1, \dots, a_N$ are the trainable parameters, while $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_N$ are randomly generated parameters to define the random features.

Note that in our definition of $f(\mathbf{x}; \mathbf{a}, \boldsymbol{\Theta})$, we have introduced the factor $1/\sqrt{d}$ inside the activation functions $\sigma_j(\cdot)$. This normalization facilitates our analysis using random matrix theory. Note also that the random feature parameters $\boldsymbol{\theta}_i$ are imposed to both have a fixed length $\sqrt{d}$, but the setting covers a more general situation where the parameters can have different lengths, say $c_1\sqrt{d}$ and $c_2\sqrt{d}$, respectively. Indeed, if $\|\boldsymbol{\theta}_i\|_2 = c_j\sqrt{d}$, we can introduce $\widetilde{\sigma}_j(z) = \sigma_j(c_j z)$ so that $\sigma_j\big(\langle \boldsymbol{\theta}_i, \mathbf{x} \rangle / \sqrt{d}\big) = \widetilde{\sigma}_j\big(\langle \boldsymbol{\tau}_i, \mathbf{x} \rangle / \sqrt{d}\big)$ where $\boldsymbol{\tau}_j = \boldsymbol{\theta}_j / c_j$ has length $\sqrt{d}$.

To go further, we specify the data we aim to learn with double random feature models. We assume the data are generated from a distribution defined as follows.

Definition 2.1 (Data generation model)**.** *The distribution of the data pair $(\mathbf{x}, y)$ is given as follows:*

1. *The input vector $\mathbf{x}$ follows the uniform distribution on the sphere $\sqrt{d} \cdot \mathbb{S}^{d-1}$ of raidus $\sqrt{d}$.*
2. *The output is $y = \langle \boldsymbol{\beta}_{1,d}, \mathbf{x} \rangle + F_0 + \varepsilon$, where $\boldsymbol{\beta}_{1,d} \in \mathbb{R}^d$, $F_0 \in \mathbb{R}$, and ε is a noise independent of $\mathbf{x}$. We assume that $\mathbb{E}(\varepsilon) = 0$, $\mathbb{E}(\varepsilon^2) = \tau^2$, and $\mathbb{E}(\varepsilon^4) < +\infty$.*

The parameters of the data generation model are $\boldsymbol{\beta}_d = [F_0, \boldsymbol{\beta}_{1,d}^\top]^\top$ and we hereafter denote by $\mathcal{D}(\boldsymbol{\beta}_d)$ the probability distribution of the pair $(\mathbf{x}, y)$.

This data generation model is standard in recent literature on double descent. Similar settings have been studied in a number of recent works (Hamsici and Martinez, 2007; Marinucci and Peccati, 2011; Di Marzio et al., 2014; Mei and Montanari, 2022).

Given a training data set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ consisting of n independent samples from the data generation model in Definition 2.1, we denote the data matrix by $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]^\top \in \mathbb{R}^{n\times d}$, the label vector by $\mathbf{y} = [y_1, ..., y_n]^\top$ and the noise vector by $\boldsymbol{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_n]^\top$. Then we fit a DRFM predictor $f(\cdot; \mathbf{a}, \boldsymbol{\Theta})$ based on the training data set S via the principle of ridge regression. Specifically, we learn the coefficient vector $\mathbf{a}$ by minimizing the ℓ_2-regularized square loss:

$$\widehat{\mathbf{a}} = \underset{\mathbf{a}}{\operatorname{argmin}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \Big(y_i - f(\mathbf{x}_i; \mathbf{a}, \boldsymbol{\Theta}) \Big)^2 + \frac{d}{n} \lambda \|\mathbf{a}\|_2^2 \right\}, \tag{2.2}$$

where $\lambda > 0$ is the regularization parameter. We here use the factor d/n in the regularization term to simplify our analysis. Removing the factor does not affect the results in this paper, because we consider the setting where d/n has a positive limit. This fact will be formally clarified in Section 3.

The excess risk of the predictor $f(\cdot; \widehat{\mathbf{a}}, \boldsymbol{\Theta})$ can be written as

$$R_d(\mathbf{X}, \boldsymbol{\Theta}, \lambda, \boldsymbol{\beta}_d, \boldsymbol{\varepsilon}) = \mathbb{E}_{\mathbf{x}\sim \mathrm{Unif}(\sqrt{d}\cdot\mathbb{S}^{d-1})} \big[F_0 + \mathbf{x}^\top \boldsymbol{\beta}_{1,d} - f(\mathbf{x}; \widehat{\mathbf{a}}, \boldsymbol{\Theta}) \big]^2. \tag{2.3}$$

This notation of the excess risk specifically highlights the dependency of the risk on $\mathbf{X}, \boldsymbol{\Theta}, \lambda, \boldsymbol{\beta}_d, \boldsymbol{\varepsilon}$. Note that we do not take average over the randomness of the training data $\mathbf{X}$, the noise vector $\boldsymbol{\varepsilon}$ or the random features $\boldsymbol{\Theta}$, but aim to show the convergence of the risk towards a fixed value as $d, N, n \to \infty$ in an appropriate manner.

3. Excess risks of double random feature models

In this section we present our main results on the excess risks of DRFMs. We first give a definition.

Definition 3.1. *The spherical moments of the activation functions σ_j ($j = 1, 2$) are*

$$\mu_{j,0} \triangleq \mathbb{E}\{\sigma_j(G)\}, \quad \mu_{j,1} \triangleq \mathbb{E}\{G\sigma_j(G)\}, \quad \mu_{j,2}^2 \triangleq \mathbb{E}\{\sigma_j(G)^2\} - \mu_{j,0}^2 - \mu_{j,1}^2,$$

where $G \sim \mathrm{N}(0, 1)$ is standard normal. We collect the six constants $\mu_{j,0}, \mu_{j,1}, \mu_{j,2}^2$, $j = 1, 2$ in a vector $\boldsymbol{\mu}$.

In Definition 3.1, the first index j points out the corresponding activation function, and the second index k links to the specific spherical moment. We now introduce the main assumptions in this paper.

Assumption 3.2. *The nonlinear activation functions $\sigma_j : \mathbb{R} \to \mathbb{R}$ ($j = 1, 2$) are weakly differentiable, with weak derivative σ_j'. Moreover, for some constants $0 < C_0, C_1 < +\infty$, $|\sigma_j(u)| \vee |\sigma_j'(u)| \le C_0 e^{C_1|u|}$, $u \in \mathbb{R}$.*

It is easy to see that commonly used activation functions such as ReLU, sigmoid, and hyperbolic tangent functions all satisfy Assumption 3.2. Therefore this is a mild assumption.

Assumption 3.3. *The data dimension d, random feature dimensions N_1, N_2, and sample size n are such that $d \to \infty$, $N_1 = N_1(d) \to \infty$, $N_2 = N_2(d) \to \infty$, $n = n(d) \to \infty$. Moreover, when $d \to \infty$, the following limits exist:*

$$\lim_{d\to+\infty} N_1/d = \psi_1 > 0, \quad \lim_{d\to+\infty} N_2/d = \psi_2 > 0, \quad \lim_{d\to+\infty} n/d = \psi_3 > 0.$$

Assumption 3.3 defines the asymptotic framework for our analysis where N_1, N_2, n, d go to infinity proportionally to each other. We let $\psi = \psi_1 + \psi_2$ and $\boldsymbol{\psi} = [\psi_1, \psi_2, \psi_3]$.

Assumption 3.4. *Let $F_{1,d} = \|\boldsymbol{\beta}_{1,d}\|_2$. Then $\lim_{d\to+\infty} F_{1,d} = F_1 > 0$. Moreover, if $F_0 \neq 0$, then $\mu_{1,0}^2 + \mu_{2,0}^2 > 0$.*

The condition $F_1 > 0$ fixes the asymptotic scale of $\boldsymbol{\beta}_{1,d}$. The second condition means that when $F_0 = \mathbb{E}(y) \neq 0$, we need either $\mu_{1,0}^2 > 0$ or $\mu_{2,0}^2 > 0$ so that the predictor $f(\mathbf{x}; \widehat{\mathbf{a}}, \boldsymbol{\Theta})$ can approximate the response y well when $d \to \infty$.

The statement of the main results needs some further preparation. For any $\xi \in \mathbb{C}_+$, we consider the following system of equations for the unknowns ν_1, ν_2, ν_3:

$$\begin{cases} \nu_1 \cdot \left(-\xi - \mu_{1,2}^2 \nu_3 - \dfrac{\mu_{1,1}^2 \nu_3}{1 - \mu_{2,1}^2 \nu_2 \nu_3 - \mu_{1,1}^2 \nu_1 \nu_3} \right) = \psi_1, \\ \nu_2 \cdot \left(-\xi - \mu_{2,2}^2 \nu_3 - \dfrac{\mu_{2,1}^2 \nu_3}{1 - \mu_{1,1}^2 \nu_1 \nu_3 - \mu_{2,1}^2 \nu_2 \nu_3} \right) = \psi_2, \\ \nu_3 \cdot \left(-\xi - \mu_{1,2}^2 \nu_1 - \mu_{2,2}^2 \nu_2 - \dfrac{\mu_{1,1}^2 \nu_1 + \mu_{2,1}^2 \nu_2}{1 - \mu_{1,1}^2 \nu_1 \nu_3 - \mu_{2,1}^2 \nu_2 \nu_3} \right) = \psi_3. \end{cases} \tag{3.1}$$

This system will be hereafter referred as the $\boldsymbol{\nu}$*-system*. For different values of $\xi \in \mathbb{C}_+$, the solutions of the above system can be viewed as functions of ξ. We let $\boldsymbol{\nu}(\xi) = [\nu_1, \nu_2, \nu_3]^\top(\xi) : \mathbb{C}_+ \to \mathbb{C}_+^3$ be the analytic function defined on $\mathbb{C}_+$ satisfying (i) for any $\xi \in \mathbb{C}_+$, $\boldsymbol{\nu}(\xi)$ is a solution to $\boldsymbol{\nu}$*-system* (3.1), (ii) there exists a sufficiently large constant ξ_0, such that $|\nu_j(\xi)| \leq 2\psi_j/\xi_0$, for all ξ with $\Im(\xi) \geq \xi_0$ and $j = 1, 2, 3$. It can be shown that such a function $\boldsymbol{\nu}$ exists and is unique, and therefore our definition of $\boldsymbol{\nu}$ is valid. The details are given in Proposition A.8. We hereafter denote $\boldsymbol{\nu} = \boldsymbol{\nu}(\xi, \boldsymbol{\mu})$ to emphasize the dependence in $\boldsymbol{\mu}$.

Definition 3.5 (Auxiliary matrices)**.** *Define $\xi^* = \sqrt{\lambda} \cdot \mathrm{i}$, and*

$$\nu_j^* \triangleq \nu_j(\xi^*; \boldsymbol{\mu}), \quad j = 1, 2, 3.$$

Here, ν_j is the solution of $\boldsymbol{\nu}$-system (3.1)*. Moreover, let $M_N \triangleq \nu_1^* \mu_{1,1}^2 + \nu_2^* \mu_{2,1}^2$, $M_D \triangleq \nu_3^* M_N - 1$, and define the matrices*

$$\mathbf{H} \triangleq \begin{bmatrix} -\frac{\nu_3^{*2}\mu_{1,1}^4}{M_D^2} + \frac{\psi_1}{\nu_1^{*2}} & -\frac{\nu_3^{*2}\mu_{1,1}^2\mu_{2,1}^2}{M_D^2} & -\frac{\mu_{1,1}^2}{M_D^2} - \mu_{1,2}^2 \\ * & -\frac{\nu_3^{*2}\mu_{2,1}^4}{M_D^2} + \frac{\psi_2}{\nu_2^{*2}} & -\frac{\mu_{2,1}^2}{M_D^2} - \mu_{2,2}^2 \\ * & * & -\frac{M_N^2}{M_D^2} + \frac{\psi_3}{\nu_3^{*2}} \end{bmatrix}, \quad \mathbf{V} \triangleq \begin{bmatrix} \mu_{1,2}^2 & 0 & \frac{\mu_{1,1}^2}{M_D^2} & \frac{\nu_3^{*2}\mu_{1,1}^2}{M_D^2} \\ \mu_{2,2}^2 & 0 & \frac{\mu_{2,1}^2}{M_D^2} & \frac{\nu_3^{*2}\mu_{2,1}^2}{M_D^2} \\ 0 & 1 & \frac{M_N^2}{M_D^2} & \frac{1}{M_D^2} \end{bmatrix},$$

*(**H** is symmetric). Finally, let* $\mathbf{L} \triangleq \mathbf{V}^\top \mathbf{H}^{-1} \mathbf{V}$.

See Proposition A.4 for the reason of selecting $\xi = \sqrt{\lambda} \cdot \mathrm{i}$. We are now in the position to state our main theorem which establishes the theoretical risk curve for the double random feature model.

Theorem 3.6. *Let the data matrix* $\mathbf{X}$*, noise vector* $\boldsymbol{\varepsilon}$*, and the DRFM model* $f(\cdot; \mathbf{a}, \boldsymbol{\Theta})$ *with random feature parameter matrix* $\boldsymbol{\Theta}$ *be defined as in Section 2. Moreover, let* M_D *and* $\mathbf{L}$ *be defined in Definition 3.5. Then under Assumptions 3.2, 3.3 and 3.4, for any regularization parameter* $\lambda > 0$*, the asymptotic excess risk* $R_d(\mathbf{X}, \boldsymbol{\Theta}, \lambda, \boldsymbol{\beta}_d, \boldsymbol{\varepsilon})$ *of the DRFM defined in* (2.3) *satisfies*

$$\mathbb{E}_{\mathbf{X}, \boldsymbol{\Theta}, \boldsymbol{\varepsilon}} \big| R_d(\mathbf{X}, \boldsymbol{\Theta}, \lambda, \boldsymbol{\beta}_d, \boldsymbol{\varepsilon}) - \mathcal{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, F_1, \tau) \big| = o_d(1),$$

where

$$\mathcal{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, F_1, \tau) = F_1^2 \bigg(\frac{1}{M_D^2} + \mathbf{L}_{3,4} + \mathbf{L}_{1,4} \bigg) + \tau^2 \big(\mathbf{L}_{2,3} + \mathbf{L}_{1,2} \big), \tag{3.2}$$

and $\mathbf{L}_{i,j}$ *are the elements in the matrix* $\mathbf{L}$ *which is defined in Definition 3.5.*

The proof of Theorem 3.6 is given in Appendix A. In Theorem 3.6, the regularization parameter λ is treated as a constant that does not depend on d, n, p. Note that the first three terms in (3.2) correspond to the estimation bias, and the last two terms are the variance terms. It can be checked that the values in $\boldsymbol{\nu}^* = [\nu_1^*, \nu_2^*, \nu_3^*]^\top$ are all purely imaginary numbers in $\mathbb{C}_+$. As the matrices $\mathbf{H}$ and $\mathbf{V}$ only depend on ν_j^{*2} (which are all negative), their elements are real-valued, so do the elements of the matrix $\mathbf{L}$. Moreover, given ν_j^*, $j = 1, 2, 3$, the terms $\mathbf{L}_{3,4}, \mathbf{L}_{1,4}, \mathbf{L}_{2,3}, \mathbf{L}_{1,2}$ in (3.2) all have closed form solutions. Due to the complexity of the solutions, we defer the calculation to Appendix A.

Remark 3.7. *By inspecting the expressions of the matrices* $\mathbf{H}$*,* $\mathbf{V}$ *and* $\mathbf{L}$*, we see that the dependence of the asymptotic excess risk* (3.2) *on the activation functions is expressed through their spherical moments* $\mu_{j,1}$ *and* $\mu_{j,2}$*,* $j = 1, 2$*. In particular, if we let* $\mu_{1,1} = \mu_{2,1}$ *and* $\mu_{1,2} = \mu_{2,2}$*, we are led to the case of a single activation function, and the asymptotic excess risk* (3.2) *coincides with the one found in Mei and Montanari (2022) for vanilla random feature models.*

Remark 3.8. *Theorem 3.6 shows that the excess risk converges to* $\mathcal{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, F_1, \tau)$ *in* L_1 *distance, which is a type of strong convergence. It directly implies convergence in probability: for any* $\rho, \delta > 0$*, there exists* $d_0 \in \mathbb{N}$ *such that for all* $d \geq d_0$*,* $\mathbb{P}\big(\big| R_d(\mathbf{X}, \boldsymbol{\Theta}, \lambda, \boldsymbol{\beta}_d, \boldsymbol{\varepsilon}) - \mathcal{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, F_1, \tau) \big| \leq \rho \big) \geq 1 - \delta$.

4. The phenomenon of triple descent in DRFMs

In this section we establish theoretical results showing the existence of DRFMs with triple descent risk curves and use simulations to verify our results.

Before we propose the detailed results, we first explain our intuition by considering the two extreme cases below:

- Case 1 (no scale difference): As discussed in Remark 3.7, if the two activation functions have identical spherical moments, that is, $\mu_{1,1} = \mu_{2,1}$ and $\mu_{1,2} = \mu_{2,2}$, the risk curve should be identical to that of a vanilla (single) random feature model. Hence, according to the study of vanilla random feature models in Mei and Montanari (2022), the risk curve commonly has a double descent shape, with the peak at the interpolation threshold $(N_1 + N_2)/n = 1$.
- Case 2 (large scale difference): If one of the two types of random features is too small in scale compared to the other, then we can expect that this small-scale part of random features is almost negligible. For example, under the extreme case that $N_1 = N_2$ and $\sigma_2(\cdot) \equiv 0$, the second type of random features can never contribute to the learned predictor, and this case also reduces to a vanilla random feature model. Therefore we can expect the risk curve to reach the peak at $N_1/n = 1$, that is, $(N_1 + N_2)/n = 2$.

We can see that the two extreme cases above both lead to double descent. However, in the first case, the peak is at $(N_1 + N_2)/n = 1$, while in the second case, the peak is at $(N_1 + N_2)/n = 2$. When the scales of the two parts of random features are neither too similar nor too different, we can expect the risk curve to exhibit certain characteristics from both extreme cases, possibly having two peaks at $(N_1 + N_2)/n = 1$ and $(N_1 + N_2)/n = 2$ respectively – this is exactly triple descent. This motivates us to conjecture that triple descent can occur when the two parts of random features have appropriate scale differences.

4.1 Triple descent: theoretical results

The asymptotic excess risk function $\mathcal{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, F_1, \tau)$ established in Theorem 3.6 can imply the existence of triple descent in double random feature models. Note that this risk function depends on several parameters including the smoothing parameter λ, the number of features in the model and some spherical moments of the involved activation functions. Here we focus on the "ridgeless regression" setting where $\lambda \to 0$, and we aim to construct specific configurations of $\boldsymbol{\mu}$ such that for any fixed values of F_1 and τ, the risk function exhibits (at least) triple descent as $(\psi_1 + \psi_2)/\psi_3$ increases.

For convenience, we use in this section the shorthand $\mathcal{R} := \mathcal{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, F_1, \tau)$. The following proposition demonstrates triple descent by considering the asymptotic regime where $\lambda \to 0$ and $\mu_{2,1}, \mu_{2,2} \to 0$: the former points to a limiting ridgeless regression model and the latter signifies the scale differences between two activation functions by shrinking the second activation to 0.

Proposition 4.1 ($\lambda \to 0$)**.** *Consider the same assumptions as in Theorem 3.6 and the asymptotic excess risk function $\mathcal{R} := \mathcal{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, F_1, \tau)$. For fixed $0 < \psi_1, \psi_2, \psi_3 < +\infty$ we have:*

1. *When $(\psi_1 + \psi_2)/\psi_3 = c_1 < 1$, $\lim_{\lambda \to 0} \mathcal{R} < +\infty$;*
2. *When $(\psi_1 + \psi_2)/\psi_3 = 1$, $\lim_{\lambda \to 0} \mathcal{R} = +\infty$;*
3. *When $1 < (\psi_1 + \psi_2)/\psi_3 = c_2 < 1 + \psi_2/\psi_1$, $\underline{\lim}_{\mu_{2,1}, \mu_{2,2} \to 0} \lim_{\lambda \to 0} \mathcal{R} < +\infty$;*
4. *When $(\psi_1 + \psi_2)/\psi_3 = 1 + \psi_2/\psi_1$, $\lim_{\mu_{2,1}, \mu_{2,2} \to 0} \lim_{\lambda \to 0} \mathcal{R} = +\infty$.*

The proof of Proposition 4.1 is given in Appendix B. This proposition theoretically demonstrate the existence of triple descent for certain DRFMs. Note that the risk function $\mathcal{R}$ depends on ψ_1, ψ_2, ψ_3. To simplify and standardize the setting, we specifically consider the case where ψ_3 and the ratio ψ_1/ψ_2 are both fixed. In this case, the change of model complexity has a single degree of freedom, which can be characterized by $c := (\psi_1+\psi_2)/\psi_3$. Now we can investigate the curve of the risk function with respect to c and see if the shape exhibits triple descent.

To see how Proposition 4.1 demonstrates triple descent, we pick two fixed "reference points" $0 < c_1 < 1$ and $1 < c_2 < 1+\psi_2/\psi_1$ (recall that we are considering the setting where ψ_2/ψ_1 is fixed.) By the third and fourth conclusions above, we can choose a large enough constant $M_1 > 0$ (Not related to ψ_1, ψ_2 and ψ_3), for which there exist $\mu_{2,1}$ and $\mu_{2,2}$ such that

$$\lim_{\lambda\to 0} \mathcal{R} > M_1 \text{ when } (\psi_1+\psi_2)/\psi_3 = 1+\psi_2/\psi_1, \text{ and } \lim_{\lambda\to 0} \mathcal{R} < M_1 \text{ when } (\psi_1+\psi_2)/\psi_3 = c_2.$$

For these chosen spectral moments $\mu_{2,1}$ and $\mu_{2,2}$ and by the first and second conclusions of the proposition, one can find a large constant $M_2 > M_1$ such that

$$\lim_{\lambda\to 0} \mathcal{R} > M_2 \text{ when } (\psi_1+\psi_2)/\psi_3 = 1, \quad \text{and} \quad \lim_{\lambda\to 0} \mathcal{R} < M_2 \text{ when } (\psi_1+\psi_2)/\psi_3 = c_1.$$

Recall that $\psi_1 \sim N_1/d$, $\psi_2 \sim N_2/d$ and $\psi_3 \sim n/d$ in the limits. It is customary to consider the asymptotic excess risk function $\mathcal{R}$ with respect to the "model complexity parameter" $c = \lim_{d\to+\infty}(N_1+N_2)/n = (\psi_1+\psi_2)/\psi_3$. Based on this analysis, we are able to find constants $0 < M_1 < M_2$ and $\mu_{2,1}$, $\mu_{2,2}$ that do not depend on ψ_1, ψ_2, ψ_3, so that the following four results hold:

1. $c = c_1$, $\lim_{\lambda\to 0} \mathcal{R} < M_2$;
2. $c = 1$, $\lim_{\lambda\to 0} \mathcal{R} > M_2$;
3. $c = c_2$, $\lim_{\lambda\to 0} \mathcal{R} < M_1$;
4. $c = 1+\psi_2/\psi_1$, $\lim_{\lambda\to 0} \mathcal{R} > M_1$.

These four cases above correspond to four situations with different model complexities: each case is for a specific value of $c = (\psi_1+\psi_2)/\psi_3 = \lim_{d\to+\infty}(N_1+N_2)/n$. The next proposition shows that the risk function has a finite limit when the model complexity parameter c tends to infinity, or in other words, in the infinitely over-parameterized regime.

Proposition 4.2 ($\psi_1, \psi_2 \to +\infty$)**.** *Consider the same assumptions as in Theorem 3.6 and the asymptotic excess risk function $\mathcal{R} := \mathcal{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, F_1, \tau)$ with non-degenerate activation functions. For fixed ψ_3 and $r_1, r_2 > 0$, we have*

$$\lim_{\substack{\psi_1,\psi_2\to+\infty \\ \psi_1/r_1=\psi_2/r_2}} \mathcal{R} = \frac{F_1^2\psi_3 + \tau^2\chi_0^2}{(\chi_0+1)^2\psi_3 - \chi_0^2},$$

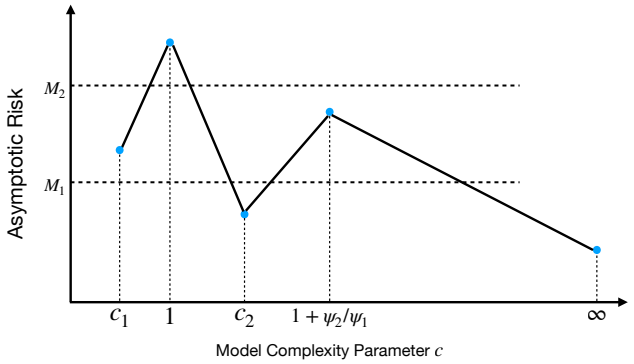


Figure 2: Existence of triple descent in a double random feature model: the four points c_1 to $1+\psi_2/\psi_1$ for the model complexity parameter $c=(\psi_1+\psi_2)/\psi_3$ are found in Proposition 4.1 and the last point depicts the limit found in Proposition 4.2 when $c\to\infty$.

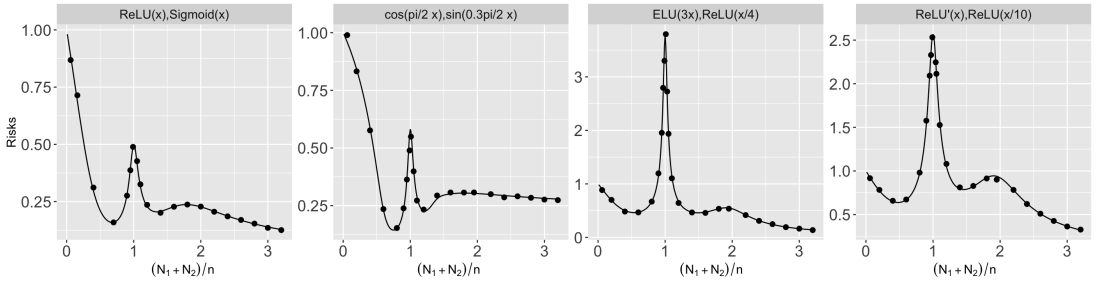
where

$$\chi_0 = \frac{(r_1\mu_{1,1}^2 + r_2\mu_{2,1}^2)\chi_1}{2\sum_{i,j=1}^{2} r_i r_j \mu_{i,1}^2 \mu_{j,2}^2},$$

$$\chi_1 = (\psi_3-1)\sum_{i=1}^{2} r_i\mu_{i,1}^2 - \sum_{i=1}^{2} r_i\mu_{i,2}^2 + \sqrt{\left((\psi_3-1)\sum_{i=1}^{2} r_i\mu_{i,1}^2 - \sum_{i=1}^{2} r_i\mu_{i,2}^2\right)^2 + 4\psi_3\sum_{i,j=1}^{2} r_i r_j \mu_{i,1}^2\mu_{j,2}^2}.$$

The proof of Proposition 4.2 can be found in Appendix B. By combining the derived limiting risk value from Proposition 4.2, where c tends to infinity, with the summary provided after Proposition 4.1, we can observe that the asymptotic excess risk function $\mathcal{R}$ exhibits (at least) triple descent with the chosen parameter values. This behavior occurs as the model complexity parameter c increases from 0 to $c_1, 1, c_2, 1+\psi_2/\psi_1$, and eventually tends to infinity. A visual representation of this phenomenon can be seen in Figure 2. Furthermore, when the model complexity is $c<1$, the asymptotic risk takes the form of a U shape, which aligns with classical theory.

Remark 4.3. *We can also consider the case where ψ_1, ψ_2 goes to zero, and this case corresponds to the setting where random feature model is almost reduced to a constant predictor. In this case, it is easy to show that $\lim_{\psi_1,\psi_2\to 0}\mathcal{R} = F_1^2$. In classical statistical theory, the first descent occurs here when the model complexity gradually increases: as the predictor becomes more complicated than a constant predictor, the asymptotic risk will first decrease below F_1^2.*



(a) (b) (c) (d)

Figure 3: Triple descent in double random feature models with different activation functions. The plots show both the asymptotic excess risks (curves) and empirical excess risks (dots). From (a) to (d), the activation functions are $\big(\mathrm{ReLU}(x), \mathrm{Sigmoid}(x)\big)$, $\big(\cos(\frac{\pi}{2}x), \sin(\frac{0.3\pi}{2}x)\big)$, $\big(\mathrm{ELU}(3x), \mathrm{ReLU}(x/4)\big)$ and $\big(\mathrm{ReLU}'(x), \mathrm{ReLU}(x/10)\big)$.

4.2 Triple descent: empirical evidence

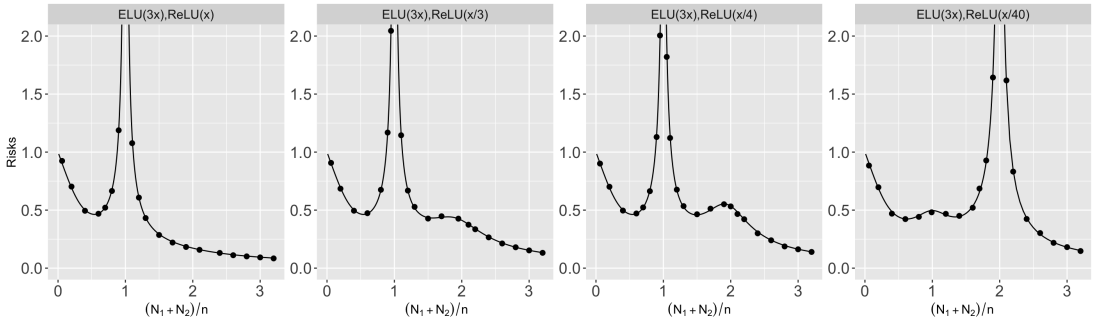
In this subsection, we empirically demonstrate the triple descent phenomenon in double random feature models. The simulation design is as follows.

- Training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are generated independently following Definition 2.1 with $\tau = 0.1$: each $\mathbf{x}_i$ is uniformly generated from the sphere $\sqrt{d} \cdot \mathbb{S}^{d-1}$, and the corresponding response is given as $y_i = \langle \boldsymbol{\beta}_1, \mathbf{x}_i \rangle + F_0 + \varepsilon_i$, where $\boldsymbol{\beta}_1$ is a randomly chosen unit vector;
- $F_0 = 0.2$, $\lambda = 10^{-5}$;
- Training sample size $n = 1000$, data dimension $d = 300$ and $N_1 = N_2$ varying from 0 to $1.6n$.

As we gradually increase the dimensions of random features $N_1 = N_2$ from 0 to $1.6n$, the model complexity parameter $c(d) = (N_1 + N_2)/n$ varies from 0 to 3.2. The empirical and finite-horizon values for the limiting excess risk $\mathcal{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, F_1, \tau)$ in Theorem 3.6 are obtained on a test data set of size 700 and averaged from 30 independent replications.

The results are given in Figure 3. In this figure (and all other figures of this section), the values of the asymptotic risk $\mathcal{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, F_1, \tau)$ are shown as continuous curves while empirical risk values are plotted using black dots. We consider activation functions $\mathrm{ReLU}(x) = x_+$, $\mathrm{ReLU}'(x) = \mathbb{1}\{x > 0\}$, $\mathrm{Sigmoid}(x) = 1/(1 + e^{-x})$, $\mathrm{ELU}(x) = x_+ - (1 - e^x)_-$, as well as trigonometric functions $\cos(x)$ and $\sin(x)$. We slightly scale the activation functions to show clearer shapes of triple descent: the four plots in Figure 3 represent DRFMs with activation pairs $\big(\mathrm{ReLU}(x), \mathrm{Sigmoid}(x)\big)$, $\big(\cos(\frac{\pi}{2}x), \sin(\frac{0.3\pi}{2}x)\big)$, $\big(\mathrm{ELU}(3x), \mathrm{ReLU}(x/4)\big)$ and $\big(\mathrm{ReLU}'(x), \mathrm{ReLU}(x/10)\big)$, respectively.

Clearly, the empirical risk values well match their theoretical counterparts in all the examined settings, which empirically validates the asymptotic risks established in Theorem 3.6. More importantly, these risk curves all exhibit triple descent as predicted by Propositions 4.1 and 4.2 (see also Figure 2), where the four critical constants have the



(a) (b) (c) (d)

Figure 4: Risk curves of DRFMs with scaled ReLU and ELU activation functions. The plots show both the asymptotic excess risks (curves) and empirical excess risks (dots). From (a) to (d), the activation functions are $(\mathrm{ELU}(3x), \mathrm{ReLU}(x))$, $(\mathrm{ELU}(3x), \mathrm{ReLU}(x/3))$, $(\mathrm{ELU}(3x), \mathrm{ReLU}(x/4))$ and $(\mathrm{ELU}(3x), \mathrm{ReLU}(x/40))$ respectively.

following values under the present experimental design:

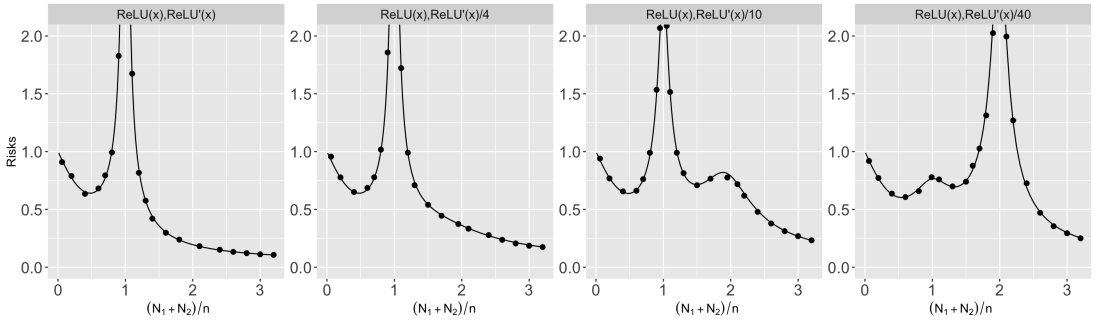
$$c_1 < 1, \quad c_2 = 1, \quad 1 < c_3 < 2, \quad c_4 = 2.$$

4.3 Impact of scale difference on triple descent

As demonstrated in Propositions 4.1 and 4.2, when the magnitude of a random feature is of a smaller order than the other feature, triple descent appears in a DRFM. In this section, we use our theoretical predictions as well as simulations to verify this result. The experiment setups are the same as the experiments in Section 4.2, except that here we use different pairs of activation functions. For two activation functions σ_1, σ_2, we gradually decrease the scale of σ_2 by using activation pairs $(\sigma_1(x), c_0\sigma_2(x))$ with a smaller and smaller factor c_0. Results for activation pairs $(\mathrm{ELU}, \mathrm{ReLU})$ and $(\mathrm{ReLU}, \mathrm{ReLU}')$ are reported in Figure 4 and Figure 5, respectively. Clearly, in both figures, the empirical errors (dots) well match their theoretical counterparts (curves). Moreover, In Figure 4 (a) and Figure 5 (a), we present the result when we appropriately balance the two activation functions such that the two parts of the random features have similar scales, and the resulting risk curves exhibit double descent with a peak at $(N_1 + N_2)/n = 1$. As the scale of the second random feature decreases, the risk curves transit from double descent curves to triple descent curves in Figure 4 (b), (c) and Figure 5 (b), (c). Finally, in Figure 4 (d) and Figure 5 (d) when the scale differences are large, the risk curves have a large peak near $c = 2$ but only a very small peak near $c = 1$. Clearly, these results perfectly match Proposition 4.1, and thus backs up the triple descent phenomena in DRFMs.

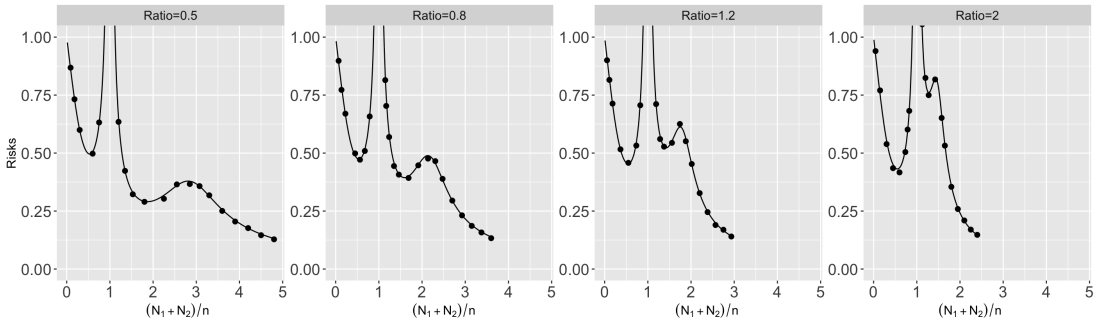
4.4 Impact of the ratio between random feature dimensions

Our previous experiments are all under the setting where $N_1 = N_2$, which corresponds to the case where the two parts of the random features have the same dimensions. In fact, we



(a) (b) (c) (d)

Figure 5: Risk curves of DRFMs with scaled ReLU and ReLU$'$ activation functions. The plots show both the asymptotic excess risks (curves) and empirical excess risks (dots). From (a) to (d), the activation functions are $(\mathrm{ReLU}(x), \mathrm{ReLU}'(x))$, $(\mathrm{ReLU}(x), \mathrm{ReLU}'(x)/4)$, $(\mathrm{ReLU}(x), \mathrm{ReLU}'(x)/10)$ and $(\mathrm{ReLU}(x), \mathrm{ReLU}'(x)/40)$ respectively.



(a) (b) (c) (d)

Figure 6: Risk curves of DRFMs with different ratios between random feature dimensions. The plots show both the asymptotic excess risks (curves) and empirical excess risks (dots). From (a) to (d), the ratios N_1/N_2 are 0.5, 0.8, 1.2 and 2, respectively. The activation functions are chosen as $\sigma_1(x) = \mathrm{ELU}(3x)$ and $\sigma_2(x) = \mathrm{ReLU}(x/4)$.

can study more general settings where N_1 and N_2 hold a ratio other than 1. Specifically, suppose that σ_1 has larger scale compared to σ_2. Then based on Proposition 4.1, it is clear that the first peak should be around $c = 1$, while the second peak should be around $c = 1 + \psi_2/\psi_1$.

We now consider the same experiment setup as in Section 4.2, except that here we focus on the activation pair $\big(\mathrm{ELU}(3x), \mathrm{ReLU}(x/4)\big)$, and no longer require $N_1 = N_2$. Instead, we consider the ratios $N_1/N_2 \in \{0.5, 0.8, 1.2, 2\}$ and plot the corresponding risk curves. Note that the coordinates in the first part of random features are about 10 times those in the second part (in magnitude), and the second peak in the risk curve is expected to be around the position $1 + (N_1/N_2)^{-1}$.

The simulation results are reported in Figure 6. It can be seen that the second peaks in Figure 6 (a), (b), (c), (d) are around $c = 1 + (N_1/N_2)^{-1} = 3, 9/4, 11/6, 3/2$, respectively.

This further verifies Proposition 4.1, and shows how one can design double random feature models with specific peak locations. We have also studied other key factors affecting the risk curve, such as the regularization parameter and the signal-to-noise ratio. Details of experimental results are reported in Appendix D.

4.5 Further discussion

Due to the complexity of the theoretical expressions involving almost 10 variables, it is difficult to provide a precise characterization such as under what conditions is the 2nd descent is lower than the 1st descent, or under what conditions is the peak of the second descent lower than the bottom of the 1st descent. While we have found empirically that the second peak tends to appear when the scale of σ_2 is small enough, it is hard to make a formal statement on the general conditions that guarantee this fact. We would also like to note that Appendix D provides some analysis on the effects of SNR and regularization on the multiple descent phenomenon. Specifically, we have observed that SNR affects the trend of the risks in the under-parameterized regime ($(N_1 + N_2)/n < 1$) and the highly over-parameterized regime ($(N_1 + N_2)/n > 2$, while λ affects the existence of the peak. Additionally, benign overfitting tends to occur when the SNR is high, while optimal regularization can help mitigate the multiple descent, as has been shown in previous literature (Nakkiran et al., 2020; Mei and Montanari, 2022).

5. The multiple random feature model

In the previous sections, we have studied double random feature models based on two activation functions. In this section, we extend our results to the case with K activation functions ($K \in \mathbb{N}_+$).

Suppose that for $j \in [K]$, there are N_j random feature units using activation function σ_j. Then we let $N = N_1 + \cdots + N_K$ be the total dimension of the random features. Moreover, we define the index set of the random feature units using the activation function σ_j as

$$\mathcal{N}_j = \left\{ i \in [N] : 1 + \sum_{r=1}^{j-1} N_r \le i \le \sum_{r=1}^{j} N_r \right\},\ j \in [K].$$

Let $\boldsymbol{\theta}_i \sim \mathrm{Unif}(\sqrt{d} \cdot \mathbb{S}^{d-1})$, $i \in [N]$ be the random feature parameter vectors and $a_i \in \mathbb{R}$, $i \in [N]$ be the linear combination coefficients of the random features. Then we denote $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_N]^\top \in \mathbb{R}^{N \times d}$, $\mathbf{a} = [a_1, \ldots, a_N]^\top \in \mathbb{R}^N$. A *multiple random feature model* (MRFM) predictor is defined as

$$f(\mathbf{x}; \mathbf{a}, \boldsymbol{\Theta}) = \sum_{j=1}^{K} \sum_{i \in \mathcal{N}_j} a_i \sigma_j\big(\langle \boldsymbol{\theta}_i, \mathbf{x} \rangle / \sqrt{d}\big).$$

We also denote by $\boldsymbol{\Theta}_j = [\boldsymbol{\theta}_{\mathcal{N}_j}]^\top \in \mathbb{R}^{N_j \times d}$ the collection of the random feature parameter vectors using the activation function σ_j. We learn the same data model in Definition 2.1

by fitting a training data set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with the function $f(\mathbf{x}; \mathbf{a}, \mathbf{\Theta})$ using ridge regression. Similar to Section 2, we learn the coefficient vector $\mathbf{a}$ by minimizing the ℓ_2-regularized square loss function:

$$\widehat{\mathbf{a}} = \operatorname*{argmin}_{\mathbf{a}} \left\{ \frac{1}{n} \sum_{j=1}^{n} \Big(y_j - f(\mathbf{x}_j; \mathbf{a}, \mathbf{\Theta}) \Big)^2 + \frac{d}{n} \lambda \|\mathbf{a}\|_2^2 \right\}.$$

The excess risk is denoted by $R_d(\mathbf{X}, \mathbf{\Theta}, \lambda, \boldsymbol{\beta}_d, \boldsymbol{\varepsilon})$ highlighting its dependence on $\mathbf{X}, \mathbf{\Theta}, \lambda, \boldsymbol{\beta}_d$ and $\boldsymbol{\varepsilon}$:

$$R_d(\mathbf{X}, \mathbf{\Theta}, \lambda, \boldsymbol{\beta}_d, \boldsymbol{\varepsilon}) = \mathbb{E}_{\mathbf{x} \sim \mathrm{Unif}(\sqrt{d} \cdot \mathbb{S}^{d-1})} \big[F_0 + \mathbf{x}^\top \boldsymbol{\beta}_{1,d} - f(\mathbf{x}; \widehat{\mathbf{a}}, \mathbf{\Theta}) \big]^2. \tag{5.1}$$

5.1 Excess risks of MRFMs

The definitions and assumptions below are similar to those previously used for DRFMs in Section 3.

Definition 5.1. *For $j = 1, 2, \ldots, K$ and $G \sim \mathrm{N}(0, 1)$, define*

$$\mu_{j,0} \triangleq \mathbb{E}\sigma_j(G), \qquad \mu_{j,1} \triangleq \mathbb{E}G\sigma_j(G), \qquad \mu_{j,2}^2 \triangleq \mathbb{E}\{\sigma_j^2(G)\} - \mu_{j,0}^2 - \mu_{j,1}^2.$$

These spherical moments are collected into a vector $\boldsymbol{\mu}$.

Assumption 5.2. *Let $\sigma_j : \mathbb{R} \to \mathbb{R}$ $(j = 1, 2, \ldots, K)$ be weakly differentiable, with weak derivative σ_j'. Assume $|\sigma_j(u)| \vee |\sigma_j'(u)| \le C_0 e^{C_1 |u|}$ for some constants $C_0, C_1 < +\infty$.*

Assumption 5.3. *We consider sequences of parameters $N_1, N_2, \ldots, N_K, n, d$ that go to infinity proportionally to each other. Without loss of generality, let the sequences be indexed by d, and assume for $j = 1, \ldots, K$, the following limits exist:*

$$\lim_{d \to +\infty} N_j/d = \psi_j \in (0, \infty), \qquad \lim_{d \to +\infty} n/d = \psi_{K+1} \in (0, \infty).$$

These limits are collected into the vector $\boldsymbol{\psi} = [\psi_1, \ldots, \psi_K, \psi_{K+1}]$.

Assumption 5.4. *Let $F_{1,d} = \|\boldsymbol{\beta}_{1,d}\|_2$. Then $\lim_{d \to +\infty} F_{1,d} = F_1 > 0$. Moreover, if $F_0 \neq 0$, then $\sum_{j=1}^{K} \mu_{j,0}^2 > 0$.*

All these assumptions are natural, and parallel Assumptions 3.2-3.4 in Section 3, respectively. The presentation of the results for the MRFM also relies on a system of self-consistent equations as follows. For $\xi \in \mathbb{C}_+$, consider the following system of equations with unknown

functions $(\nu_1, \ldots, \nu_{K+1})$: $\mathbb{C}_+ \to \mathbb{C}_+^{K+1}$ (as functions of the complex variable ξ):

$$\begin{cases} \nu_j \cdot \left(-\xi - \mu_{j,2}^2 \nu_{K+1} - \dfrac{\mu_{j,1}^2 \nu_{K+1}}{1 - \sum_{j=1}^K \mu_{j,1}^2 \nu_j \nu_{K+1}} \right) = \psi_j, \quad j = 1, \ldots, K \\ \nu_{K+1} \cdot \left(-\xi - \displaystyle\sum_{j=1}^K \mu_{j,2}^2 \nu_j - \dfrac{\sum_{j=1}^K \mu_{j,1}^2 \nu_j}{1 - \sum_{j=1}^K \mu_{j,1}^2 \nu_j \nu_{K+1}} \right) = \psi_{K+1}. \end{cases} \tag{5.2}$$

We let $\boldsymbol{\nu} = [\nu_1, \ldots, \nu_{K+1}]^\top : \mathbb{C}_+ \to \mathbb{C}_+^{K+1}$ be the analytic function defined on $\mathbb{C}_+$ satisfying (i) for any $\xi \in \mathbb{C}_+$, $\boldsymbol{\nu}(\xi)$ is a solution to $\boldsymbol{\nu}$-*system* (5.2), (ii) there exists a sufficiently large constant ξ_0, such that $|\nu_j(\xi)| \le 2\psi_j/\xi_0$ for all ξ with $\Im(\xi) \ge \xi_0$ and $j \in [K]$. It can be shown that such a function $\boldsymbol{\nu}$ exists and is unique, and therefore our definition of $\boldsymbol{\nu}$ is valid. The full justification is given in Proposition C.9 in the appendix. We also denote $\boldsymbol{\nu} = \boldsymbol{\nu}(\xi, \boldsymbol{\mu})$ to emphasize the dependence on $\boldsymbol{\mu}$.

Definition 5.5 (Auxiliary matrices)**.** *Define* $\xi^* = \sqrt{\lambda} \cdot \mathrm{i}$,

$$\boldsymbol{\nu}^* = [\nu_1^*, \ldots, \nu_{K+1}^*]^\top = \big[\nu_1, \ldots, \nu_{K+1}\big]^\top(\xi^*; \boldsymbol{\mu})$$

where ν_j *is the solution of* $\boldsymbol{\nu}$*-system* (5.2)*, and let*

$$M_N = \sum_{j=1}^K \mu_{j,1}^2 \nu_j^*, \quad M_D = \nu_{K+1}^* M_N - 1.$$

We then let $\mathbf{H} \in \mathbb{R}^{(K+1)\times(K+1)}$ *be a real symmetric matrix whose* (i,j)*-th entry* $(i \le j)$ *is*

$$\mathbf{H}_{i,j} = \begin{cases} -\dfrac{\nu_{K+1}^{*2} \mu_{i,1}^4}{M_D^2} + \dfrac{\psi_i}{\nu_i^{*2}}, & 1 \le i = j \le K, \\ -\dfrac{\nu_{K+1}^{*2} \mu_{i,1}^2 \mu_{j,1}^2}{M_D^2}, & 1 \le i < j \le K, \\ -\dfrac{\mu_{i,1}^2}{M_D^2} - \mu_{i,2}^2, & 1 \le i \le K,\ j = K+1, \\ -\dfrac{M_N^2}{M_D^2} + \dfrac{\psi_{K+1}}{\nu_{K+1}^{*2}}, & i = j = K+1. \end{cases}$$

Moreover, define $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3, \mathbf{v}_4] \in \mathbb{R}^{(K+1)\times 4}$*, where*

$$\mathbf{v}_1 = \Big[\mu_{1,2}^2, \mu_{2,2}^2, \ldots, \mu_{K,2}^2, 0\Big]^\top, \quad \mathbf{v}_2 = \Big[0, \ldots, 0, 1\Big]^\top,$$
$$\mathbf{v}_3 = \Big[\frac{\mu_{1,1}^2}{M_D^2}, \ldots, \frac{\mu_{K,1}^2}{M_D^2}, \frac{M_N^2}{M_D^2}\Big]^\top, \quad \mathbf{v}_4 = \Big[\nu_{K+1}^{*2} \frac{\mu_{1,1}^2}{M_D^2}, \ldots, \nu_{K+1}^{*2} \frac{\mu_{K,1}^2}{M_D^2}, \frac{1}{M_D^2}\Big]^\top.$$

Finally, let $\mathbf{L} = \mathbf{V}^\top \mathbf{H}^{-1} \mathbf{V} \in \mathbb{R}^{4\times 4}$.

It is clear that the above definitions are consistent with Definition 3.5 for the case of $K = 2$. Based on these definitions, the asymptotic limit of the excess risk can be expressed as function of the elements of the matrix $\mathbf{L}$. Our main result for MRFMs is given in the following theorem.

Theorem 5.6. *Let the data matrix $\mathbf{X}$ and the noise vector $\boldsymbol{\varepsilon}$ be generated as in Definition 2.1. Then under Assumptions 5.2, 5.3 and 5.4, for any regularization parameter $\lambda > 0$, the asymptotic excess risk $R_d(\mathbf{X}, \boldsymbol{\Theta}, \lambda, \boldsymbol{\beta}_d, \boldsymbol{\varepsilon})$ of the MRFM defined in* (5.1) *satisfies*

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon}}\big|R_d(\mathbf{X}, \boldsymbol{\Theta}, \lambda, \boldsymbol{\beta}_d, \boldsymbol{\varepsilon}) - \mathcal{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, F_1, \tau)\big| = o_d(1),$$

where, with M_D and the matrix $\mathbf{L}$ defined in Definition 5.5,

$$\mathcal{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, F_1, \tau) = F_1^2\bigg(\frac{1}{M_D^2} + \mathbf{L}_{3,4} + \mathbf{L}_{1,4}\bigg) + \tau^2\big(\mathbf{L}_{2,3} + \mathbf{L}_{1,2}\big). \tag{5.3}$$

Here, $\mathbf{L}_{i,j}$ are the elements in the matrix $\mathbf{L}$ which is defined in Definition 5.5.

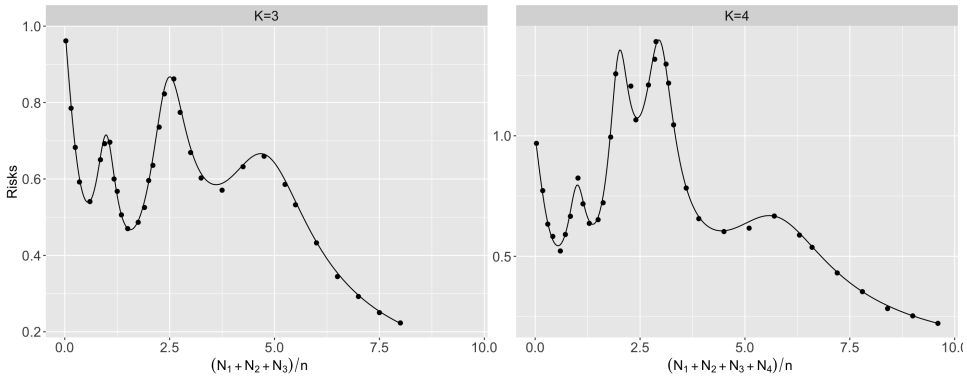
Theorem 5.6 is proved in Appendix C. The asymptotic excess risk for the MRFM given in Equation (5.3) is similar to (3.2) for the DRFM. It is clear that Theorem 5.6 covers Theorem 3.6 and the results in Mei and Montanari (2022) as special cases with $K = 2$ and $K = 1$, respectively.

5.2 Multiple descent in MRFMs

We now demonstrate the existence of multiple descent in MRFMs. The experimental setting is similar to the previous experiments reported in Section 4. We set $d = 300$, $n = 1000$, and $\lambda = 10^{-4}$. In simulation, the training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are generated independently according to Definition 2.1: each $\mathbf{x}_i$ is uniformly generated from the sphere $\sqrt{d} \cdot \mathbb{S}^{d-1}$, and the corresponding response is given as $y_i = \langle \boldsymbol{\beta}_1, \mathbf{x}_i \rangle + F_0 + \varepsilon_i$, where $\boldsymbol{\beta}_1$ is a randomly chosen unit vector, $F_0 = 0.2$ and $\tau = 0.1$. We estimate the excess risks of the MRFMs with a test data set of size 700, and take average over 30 independent runs. We consider two MRFMs with $K = 3$ and $K = 4$, respectively. For the case $K = 3$, we consider three activation functions $\sigma_1(x) = \mathrm{ReLU}(9x)$, $\sigma_2(x) = \mathrm{ReLU}(x)$ and $\sigma_3(x) = \mathrm{ReLU}(0.1x)$, and set the ratios between dimensions of random features as $N_1 = N_2 = N_3/3$. For the case $K = 4$, we use four activation functions $\sigma_1(x) = \mathrm{ReLU}(80x)$, $\sigma_2(x) = \mathrm{ReLU}(9x)$, $\sigma_3(x) = \mathrm{ReLU}(x)$ and $\sigma_4(x) = \mathrm{ReLU}(0.1x)$, and keep the ratios $N_1 = N_2 = N_3 = N_4/3$.

The results are given in Figure 7. We can see that the simulation results (dots) well match the theoretically derived risks (curves), which validates our results in Theorem 5.6. Moreover, Figure 7 (a) (where we use three different activation functions) shows quadruple descent, while Figure 7 (b) (where we use four different activation functions) shows quintuple descent. With these observations, we believe an MRFM using K activation functions may exhibit $(K + 1)$-fold descent.

Following a similar analysis as in Section 4, we can also study the locations of each peak in the risk curves as follows. First consider the experiment with $K = 3$. Clearly,



(a) (b)

Figure 7: Multiple descent in multiple random feature models. (a) gives the risk curve for the MRFM with three activation functions, which exhibits quadruple descent; (b) shows the risk curve for the MRFM with four activation functions, which exhibits quintuple descent.

the first peak always locates around $(N_1 + N_2 + N_3)/n = 1$. Regarding the second peak, note that the scales of the activation functions are set in descending order. Under this case, the first two types of random features will mainly contribute to the predictor and the third type of random features is negligible, therefore we have $(N_1 + N_2)/n = 1$ around the second peak. Since $N_1 = N_2 = N_3/3$, we have $N_1 = N_2 = n/2$ and $N_3 = 3n/2$. Hence we conclude that the second peak should be around $(N_1 + N_2 + N_3)/n = 2.5$. Similarly, regarding the third peak, we have $N_1/n = 1$, which indicates that the peak locates around $(N_1+N_2+N_3)/n = 5$. These predicted locations clearly match the results shown in Figure 7 (a). For the case $K = 4$, with a similar argument, we can expect that the four peaks are located around $1, 2, 3, 6$, respectively. This also matches the result in Figure 7 (b).

6. Conclusion

This paper considers the learning of double random feature models and multiple random feature models. We give the explicit formulas for the asymptotic excess risks achieved by DRFMs and MRFMs. These theoretical results are further well confirmed by empirical simulations in various settings. We provide an explanation of the triple descent and multiple descent phenomena based on the scale difference between activation functions, and discuss how the ratio between random feature dimensions control the location of the second peaks in the risk curves. By showing that MRFMs with K types of random features may exhibit $(K + 1)$-fold descent, we demonstrate that risk curves with a specific number of descent generally exist in random feature based regression.

An immediate future work direction is to study ridge-less regression where $\lambda = 0$. Moreover, our result can help future studies on the advantages and disadvantages of overfitting by quantitatively comparing the risks achieved by over-parameterized/under-parameterized models with different regularization levels. Extending our findings to deep learning is another important future direction.

Acknowledgment

We would like to thank the reviewers and the Editor for their helpful comments which improved the manuscript significantly. Jianfeng Yao's research is partially supported by the NSFC RFIS Grant (Senior Scholar) 12350710179.

Appendix A. Proof of Theorem 3.6

The proof is presented in the following four steps.

1. We first develop a decomposition of the risk and find an asymptotic approximation whose main terms are expressed as traces of several random matrices, see Proposition A.2;
2. We then create a new random matrix called the linear pencil matrix, which includes all the fundamental random matrices involved in the asymptotic approximation found in the first step, so that the needed traces are all functions of the limiting spectrum of the linear pencil matrix, see Proposition A.4;
3. Next, we find the key limiting spectral functions of the linear pencil matrix including its Stieltjes transform and logarithmic potential, and show that the needed traces converge to some specific partial derivatives of the limiting logarithmic potential, see Propositions A.6 and A.7.
4. The last step collects the results of the previous three steps and establishes the limit of the excess risk (with respect to the L_1 distance).

The four steps are given in the following subsections, respectively. A few technical lemmas and propositions used in these steps are stated without proofs; these proofs are deferred to the online supplementary material (Meng et al.). Before proceeding further, we remind the reader the following notations: $\mathbf{X} = [\mathbf{x}_1, ..., \mathbf{x}_n]^\top \in \mathbb{R}^{n\times d}$ with $(\mathbf{x}_i)_{i\in[n]} \sim \text{Unif}(\sqrt{d}\cdot\mathbb{S}^{d-1})$, $\mathbf{y} = [y_1, ..., y_n]^\top$, $\mathbf{\Theta} = [\mathbf{\Theta}_1^\top, \mathbf{\Theta}_2^\top]^\top = [\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_N]^\top \in \mathbb{R}^{N\times d}$ with $(\boldsymbol{\theta}_i)_{i\in[N]} \sim \text{Unif}(\sqrt{d}\cdot\mathbb{S}^{d-1})$. Some new notations are given in the following definition.

Definition A.1. *Define*

$$\begin{aligned}
&\mathbf{Z}_j = \sigma_1\left(\mathbf{X}\mathbf{\Theta}_j^\top/\sqrt{d}\right)/\sqrt{d} \in \mathbb{R}^{n\times N_j},\ j=1,2, \quad \mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2) \in \mathbb{R}^{n\times N},\\
&\mathbf{\Upsilon} = (\mathbf{Z}^\top\mathbf{Z} + \lambda\mathbf{I}_N)^{-1}; \quad \boldsymbol{\sigma}(\mathbf{x}) = \big(\sigma_1(\mathbf{x}^\top\mathbf{\Theta}_1^\top/\sqrt{d}), \sigma_2(\mathbf{x}^\top\mathbf{\Theta}_2^\top/\sqrt{d})\big)^\top \in \mathbb{R}^N;\\
&\mathbf{M}_1 = \text{diag}\big(\mu_{1,1}\mathbf{I}_{N_1}, \mu_{2,1}\mathbf{I}_{N_2}\big), \quad \mathbf{M}_2 = \text{diag}\big(\mu_{1,2}\mathbf{I}_{N_1}, \mu_{2,2}\mathbf{I}_{N_2}\big).
\end{aligned}$$

Furthermore, for any matrix $\mathbf{W} \in \mathbb{R}^{N\times N}$, we define a bracket $[\mathbf{W}]_{\mathbf{Z}} \triangleq \mathbf{Z}\mathbf{\Upsilon}\mathbf{W}\mathbf{\Upsilon}\mathbf{Z}^\top$.

A.1 Step 1: bias-variance decomposition of the excess risk

By the definition of $\widehat{\mathbf{a}}$ in (2.2), we have

$$\widehat{\mathbf{a}} = \underset{\mathbf{a}}{\text{argmin}} \left\{ \frac{1}{n}\sum_{j=1}^n \Big(y_j - f(\mathbf{x}_j; \mathbf{a}, \mathbf{\Theta})\Big)^2 + \frac{d}{n}\lambda\|\mathbf{a}\|_2^2 \right\} = \frac{1}{\sqrt{d}}\mathbf{\Upsilon}\mathbf{Z}^\top\mathbf{y}. \tag{A.1}$$

The excess risk is then of the form

$$R_d(\mathbf{X}, \boldsymbol{\Theta}, \lambda, \boldsymbol{\beta}_d, \boldsymbol{\varepsilon}) = \mathbb{E}_{\mathbf{x}}\big[\mathbf{x}^\top \boldsymbol{\beta}_{1,d} + F_0 - \widehat{\mathbf{a}}^\top \boldsymbol{\sigma}(\mathbf{x})\big]^2.$$

The goal of Theorem 3.6 is to calculate this risk. One of the major challenges in this calculation is the nonlinearities of the activation functions. To overcome this challenge, we introduce a decomposition of the risk in the proposition below. We remind readers that $F_{1,d} = \|\boldsymbol{\beta}_{1,d}\|_2$.

Proposition A.2. *For any $\lambda > 0$, let*

$$\overline{R}_d(\mathbf{X}, \boldsymbol{\Theta}, \lambda, F_{1,d}, \tau) = F_{1,d}^2 - \frac{2F_{1,d}^2}{d}\mathrm{tr}\bigg(\mathbf{M}_1 \frac{\boldsymbol{\Theta}\mathbf{X}^\top}{d}\mathbf{Z}\boldsymbol{\Upsilon}\bigg) + \frac{F_{1,d}^2}{d}\mathrm{tr}\bigg(\big[\widetilde{\mathbf{U}}\big]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\bigg) + \frac{\tau^2}{d}\mathrm{tr}(\big[\widetilde{\mathbf{U}}\big]_{\mathbf{Z}}),$$

where $\widetilde{\mathbf{U}} = \mathbf{M}_1\boldsymbol{\Theta}\boldsymbol{\Theta}^\top\mathbf{M}_1/d + \mathbf{M}_2\mathbf{M}_2$. Then under the same conditions as Theorem 3.6,

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon}}\Big|R_d(\mathbf{X}, \boldsymbol{\Theta}, \lambda, \boldsymbol{\beta}_d, \boldsymbol{\varepsilon}) - \overline{R}_d(\mathbf{X}, \boldsymbol{\Theta}, \lambda, F_{1,d}, \tau)\Big| = o_d(1).$$

The proof of Proposition A.2 is given in Section I in the online supplementary material (Meng et al.). It presents the bias-variance decomposition as the sum of four terms: the first three terms with $F_{1,d}^2$ together give the bias in the asymptotic excess risk, while the last term with τ^2 is the variance.

A.2 Step 2: approximation of the risk decomposition via a linear pencil matrix

The approximating function $\overline{R}_d(\mathbf{X}, \boldsymbol{\Theta}, \lambda, F_{1,d}, \tau)$ found in Proposition A.2 depends on three traces of certain random matrices. In this step, we calculate these traces via a special random matrix, namely the linear pencil matrix defined as follows.

Definition A.3. *(1) Let*

$$\mathcal{Q} := \{\mathbf{q} = [q_1, q_2, q_3, q_4, q_5] \in \mathbb{R}_+^5 : \ q_4, q_5 \le (1+q_1)/2, \ \|\mathbf{q}\|_2 \le 1\}.$$

Depending on $\mathbf{q} \in \mathcal{Q}$ and $\boldsymbol{\mu}$, the linear pencil matrix $\mathbf{A}(\mathbf{q}, \boldsymbol{\mu})$ is

$$\mathbf{A}(\mathbf{q}, \boldsymbol{\mu}) = \begin{bmatrix} q_2\mu_{1,2}^2\mathbf{I}_{N_1} + q_4\mu_{1,1}^2\frac{\boldsymbol{\Theta}_1\boldsymbol{\Theta}_1^\top}{d} & q_4\mu_{1,1}\mu_{2,1}\frac{\boldsymbol{\Theta}_1\boldsymbol{\Theta}_2^\top}{d} & \mathbf{Z}_1^\top + q_1\widetilde{\mathbf{Z}}_1^\top \\ q_4\mu_{1,1}\mu_{2,1}\frac{\boldsymbol{\Theta}_2\boldsymbol{\Theta}_1^\top}{d} & q_2\mu_{2,2}^2\mathbf{I}_{N_2} + q_4\mu_{2,1}^2\frac{\boldsymbol{\Theta}_2\boldsymbol{\Theta}_2^\top}{d} & \mathbf{Z}_2^\top + q_1\widetilde{\mathbf{Z}}_2^\top \\ \mathbf{Z}_1 + q_1\widetilde{\mathbf{Z}}_1 & \mathbf{Z}_2 + q_1\widetilde{\mathbf{Z}}_2 & q_3\mathbf{I}_n + q_5\frac{\mathbf{X}\mathbf{X}^\top}{d} \end{bmatrix} \in \mathbb{R}^{P\times P},$$

where $P = N + n$, and $\widetilde{\mathbf{Z}}_j = \frac{\mu_{j,1}}{d}\mathbf{X}\boldsymbol{\Theta}_j^\top$ for $j = 1, 2$.

(2) The Stieltjes transform of the empirical eigenvalue distribution of $\mathbf{A} = \mathbf{A}(\mathbf{q}, \boldsymbol{\mu})$ (up to the factor P/d) is

$$M_d(\xi; \mathbf{q}, \boldsymbol{\mu}) = \frac{1}{d}\mathrm{tr}\big[(\mathbf{A} - \xi\mathbf{I}_P)^{-1}\big], \quad \xi \in \mathbb{C}_+,$$

and its logarithmic potential is

$$G_d(\xi;\mathbf{q},\boldsymbol{\mu})=\frac{1}{d}\log\det(\mathbf{A}-\xi\mathbf{I}_P)=\frac{1}{d}\sum_{i=1}^{P}\log(\lambda_i(\mathbf{A})-\xi),\quad \xi\in\mathbb{C}_+.$$

Here $\lambda_1(\mathbf{A})\geq\cdots\geq\lambda_P(\mathbf{A})$ *are the eigenvalues of* $\mathbf{A}$*, and* $\log(z):=\log(|z|)+\mathrm{i}\arg(z)$*, for* $z\in\mathbb{C}$*,* $-\pi<\arg(z)\leq\pi$ *is the principal value of a complex logarithmic function.*

We assume that $\mathbf{q}\in\mathcal{Q}$ throughout the paper. The three traces in the definition of $\overline{R}_d(\mathbf{X},\boldsymbol{\Theta},\lambda,F_{1,d},\tau)$ in Proposition A.2 are now expressed as partial derivatives of the logarithmic potential G_d as shown in the proposition below.

Proposition A.4. *Let* $\widetilde{\mathbf{U}}$ *be defined in Proposition A.2. Then we have*

$$\begin{aligned}\frac{1}{d}\mathrm{tr}\left(\mathbf{M}_1\frac{\boldsymbol{\Theta}\mathbf{X}^\top}{d}\mathbf{Z}\boldsymbol{\Upsilon}\right)&=\frac{1}{2}\partial_{q_1}G_d(\xi^*;\mathbf{q},\boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}},\\ \frac{1}{d}\mathrm{tr}\left(\left[\widetilde{\mathbf{U}}\right]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\right)&=-\partial^2_{q_4,q_5}G_d(\xi^*;\mathbf{q},\boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}}-\partial^2_{q_5,q_2}G_d(\xi^*;\mathbf{q},\boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}},\\ \frac{1}{d}\mathrm{tr}(\left[\widetilde{\mathbf{U}}\right]_{\mathbf{Z}})&=-\partial^2_{q_4,q_3}G_d(\xi^*;\mathbf{q},\boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}}-\partial^2_{q_2,q_3}G_d(\xi^*;\mathbf{q},\boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}}.\end{aligned}$$

We remind readers that $\xi^*=\sqrt{\lambda}\cdot\mathrm{i}$. The proof of Proposition A.4 is given in Section II in the online supplementary material (Meng et al.).

A.3 Step 3: key limiting spectral functions of the linear pencil matrix

Proposition A.4 shows that the excess risk can be calculated based on $G_d(\xi^*;\mathbf{q},\boldsymbol{\mu})$. Moreover, by Definition A.3, we have $\frac{d}{d\xi}G_d(\xi;\mathbf{q},\boldsymbol{\mu})=-M_d(\xi;\mathbf{q},\boldsymbol{\mu})$, which shows that $G_d(\xi;\mathbf{q},\boldsymbol{\mu})$ is related to $M_d(\xi;\mathbf{q},\boldsymbol{\mu})$. Therefore, we study the Stieltjes transform $M_d(\xi;\mathbf{q},\boldsymbol{\mu})$ and calculate its limit as $d,n,N\to\infty$. To do so, we define the following system of equations.

Definition A.5. *For* $\xi\in\mathbb{C}_+$*, define a function* $\mathsf{F}(\cdot;\xi,\mathbf{q},\boldsymbol{\mu})$ *from* $\mathbb{C}^3$ *to* $\mathbb{C}^3$ *by*

$$\mathbf{m}=[m_1,m_2,m_3]\longmapsto\mathsf{F}(\mathbf{m};\xi,\mathbf{q},\boldsymbol{\mu})=\begin{bmatrix}\psi_1\Big\{-\xi+q_2\mu_{1,2}^2-\mu_{1,2}^2m_3+\frac{H_1}{H_D}\Big\}^{-1}\\ \psi_2\Big\{-\xi+q_2\mu_{2,2}^2-\mu_{2,2}^2m_3+\frac{H_2}{H_D}\Big\}^{-1}\\ \psi_3\Big\{-\xi+q_3-\mu_{1,2}^2m_1-\mu_{2,2}^2m_2+\frac{H_3}{H_D}\Big\}^{-1}\end{bmatrix},$$

where

$$\begin{aligned}H_1=&\mu_{1,1}^2q_4(1+m_3q_5)-\mu_{1,1}^2(1+q_1)^2m_3,\\ H_2=&\mu_{2,1}^2q_4(1+m_3q_5)-\mu_{2,1}^2(1+q_1)^2m_3,\\ H_3=&q_5(1+\mu_{1,1}^2m_1q_4+\mu_{2,1}^2m_2q_4)-\mu_{2,1}^2(1+q_1)^2m_2-\mu_{1,1}^2(1+q_1)^2m_1,\\ H_D=&(1+\mu_{1,1}^2m_1q_4+\mu_{2,1}^2m_2q_4)(1+m_3q_5)-\mu_{2,1}^2(1+q_1)^2m_2m_3-\mu_{1,1}^2(1+q_1)^2m_1m_3.\end{aligned}$$

We write the three coordinates of F *as* $\mathsf{F}(\mathbf{m};\xi,\mathbf{q},\boldsymbol{\mu})=[\mathsf{F}_1,\mathsf{F}_2,\mathsf{F}_3]^\top(\mathbf{m};\xi,\mathbf{q},\boldsymbol{\mu})$.

We give in Section III in the online supplementary material (Meng et al.) some properties of the function $\mathbf{F}$. In particular, we show that there exists a constant $\xi_0 > 0$, such that for all ξ with $\Im(\xi) > \xi_0$ and $\mathbf{q} \in \mathcal{Q}$, $\mathbf{F}(\cdot; \xi, \mathbf{q}, \boldsymbol{\mu})$ has a unique fixed point $\mathbf{m}(\xi; \mathbf{q}, \boldsymbol{\mu}) = [m_1, m_2, m_3]^\top(\xi; \mathbf{q}, \boldsymbol{\mu})$ satisfying $|m_j(\xi)| \leq 2\psi_j/\xi_0$ for $j = 1, 2, 3$. Note that this fixed point result only defines $\mathbf{m}(\xi; \mathbf{q}, \boldsymbol{\mu})$ on $\{\xi : \Im(\xi) > \xi_0\}$. To extend its definition to $\mathbb{C}_+$, we aim to show that $\mathbf{m}$ is an analytic function on $\{\xi : \Im(\xi) > \xi_0\}$, and its analytic continuation to $\mathbb{C}_+$ is still a fixed point of $\mathbf{F}(\cdot; \xi, \mathbf{q}, \boldsymbol{\mu})$, i.e.,

$$\mathbf{m}(\xi; \mathbf{q}, \boldsymbol{\mu}) \equiv \mathbf{F}[\mathbf{m}(\xi; \mathbf{q}, \boldsymbol{\mu}); \xi, \mathbf{q}, \boldsymbol{\mu}] \tag{A.2}$$

for all $\xi \in \mathbb{C}_+$. More importantly, by using random matrix theory, we also aim to show that the limiting spectral distribution (LSD) of the matrix $\mathbf{A}$ exists and its Stieltjes transform is

$$m(\xi; \mathbf{q}, \boldsymbol{\mu}) = \sum_{i=1}^{3} m_i(\xi; \mathbf{q}, \boldsymbol{\mu}).$$

These results are formally given in the following proposition.

Proposition A.6. *Under Assumptions 3.2 and 3.3, $\mathbf{m}(\xi; \mathbf{q}, \boldsymbol{\mu})$ is analytic on $\{\xi : \Im(\xi) > \xi_0\}$, and has a unique analytic continuation to $\mathbb{C}_+$. Moreover, this analytic continuation (still denoted as $\mathbf{m}(\xi; \mathbf{q}, \boldsymbol{\mu})$) satisfies the following properties:*

1. *$\mathbf{m}(\xi; \mathbf{q}, \boldsymbol{\mu}) \in \mathbb{C}_+^3$ for all $\xi \in \mathbb{C}_+$.*
2. *$\mathbf{m}(\xi, \mathbf{q}, \boldsymbol{\mu}) \equiv \mathbf{F}[\mathbf{m}(\xi, \mathbf{q}, \boldsymbol{\mu}); \xi, \mathbf{q}, \boldsymbol{\mu}]$ for all $\xi \in \mathbb{C}_+$.*
3. *Let $M_d(\xi; \mathbf{q}, \boldsymbol{\mu})$ be defined in Definition A.3. Then for any compact set $\Omega \subset \mathbb{C}_+$,*

$$\lim_{d \to +\infty} \mathbb{E}\left[\sup_{\xi \in \Omega} \left| M_d(\xi; \mathbf{q}, \boldsymbol{\mu}) - m(\xi; \mathbf{q}, \boldsymbol{\mu}) \right| \right] = 0.$$

The proof of Proposition A.6 is given in Section IV in the online supplementary material (Meng et al.). It shows that $M_d(\xi; \mathbf{q}, \boldsymbol{\mu})$ has a deterministic limit equal to $m(\xi; \mathbf{q}, \boldsymbol{\mu})$. This result, together with the connection between $M_d(\xi; \mathbf{q}, \boldsymbol{\mu})$ and the logarithmic potential $G_d(\xi; \mathbf{q}, \boldsymbol{\mu})$ in Definition A.3, further indicates that G_d may also have a deterministic limit, and its deterministic limit can possibly be expressed as a function of $\mathbf{m}(\xi; \mathbf{q}, \boldsymbol{\mu})$. In fact, this limit is found to be

$$g(\xi; \mathbf{q}, \boldsymbol{\mu}) \triangleq L(\xi, m_1(\xi; \mathbf{q}, \boldsymbol{\mu}), m_2(\xi; \mathbf{q}, \boldsymbol{\mu}), m_3(\xi; \mathbf{q}, \boldsymbol{\mu}); \mathbf{q}, \boldsymbol{\mu}), \tag{A.3}$$

where

$$\begin{aligned} &L(\xi, z_1, z_2, z_3; \mathbf{q}, \boldsymbol{\mu}) \triangleq \\ &\log\left[(1 + \mu_{1,1}^2 z_1 q_4 + \mu_{2,1}^2 z_2 q_4)(1 + z_3 q_5) - \mu_{1,1}^2 (1 + q_1)^2 z_1 z_3 - \mu_{2,1}^2 (1 + q_1)^2 z_2 z_3\right] \\ &- \mu_{1,2}^2 z_1 z_3 - \mu_{2,2}^2 z_2 z_3 + q_2 \mu_{1,2}^2 z_1 + q_2 \mu_{2,2}^2 z_2 + q_3 z_3 - \xi(z_1 + z_2 + z_3) \\ &- \psi_1 \log(z_1/\psi_1) - \psi_2 \log(z_2/\psi_2) - \psi_3 \log(z_3/\psi_3) - \psi_1 - \psi_2 - \psi_3. \end{aligned} \tag{A.4}$$

The following proposition formally shows that $g(\xi;\mathbf{q},\boldsymbol{\mu})$ and its partial derivatives are the deterministic limit of the G_d and the partial derivatives of G_d, respectively.

Proposition A.7. *Let $G_d(\xi;\mathbf{q},\boldsymbol{\mu})$ be defined in Definition A.3 and $g(\xi;\mathbf{q},\boldsymbol{\mu})$ defined in* (A.3)*. Then for any fixed $\xi\in\mathbb{C}_+$, $\mathbf{q}\in\mathcal{Q}$ and $u\in\mathbb{R}_+$,*

$$\begin{aligned}
&\lim_{d\to+\infty}\mathbb{E}[|G_d(\xi;\mathbf{q},\boldsymbol{\mu})-g(\xi;\mathbf{q},\boldsymbol{\mu})|]=0,\\
&\lim_{d\to+\infty}\mathbb{E}[\|\nabla_{\mathbf{q}}G_d(\mathrm{i}u;\mathbf{q},\boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}}-\nabla_{\mathbf{q}}g(\mathrm{i}u;\mathbf{q},\boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}}\|_2]=0,\\
&\lim_{d\to+\infty}\mathbb{E}[\|\nabla^2_{\mathbf{q}}G_d(\mathrm{i}u;\mathbf{q},\boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}}-\nabla^2_{\mathbf{q}}g(\mathrm{i}u;\mathbf{q},\boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}}\|_{\mathrm{op}}]=0.
\end{aligned}$$

Proposition A.7 is proved in Section V in the online supplementary material (Meng et al.).

A.4 Step 4: completion of the proof

According to Propositions A.2, A.4, and A.7, the key terms in the excess risk can be calculated as the partial derivatives of the function $g(\xi;\mathbf{q},\boldsymbol{\mu})$ at $\mathbf{q}=\mathbf{0}$. However, $g(\xi;\mathbf{q},\boldsymbol{\mu})$ is based on $\mathbf{m}(\xi;\mathbf{q},\boldsymbol{\mu})$, and the calculation of the partial derivatives of $g(\xi;\mathbf{q},\boldsymbol{\mu})$ is non-trivial: $\mathbf{m}(\xi;\mathbf{q},\boldsymbol{\mu})$ is originally defined on $\{\xi:\Im(\xi)>\xi_0\}$ as the fixed point of $\mathbf{F}$, and its definition is then extended to $\mathbb{C}_+$ in Proposition A.6. To finalize the proof, we first present the following proposition relating $\mathbf{m}(\xi;\mathbf{q},\boldsymbol{\mu})$ to the function $\nu(\xi;\boldsymbol{\mu})$ defined in Section 3.

Proposition A.8. *There exists a unique analytic function $\boldsymbol{\nu}=[\nu_1,\nu_2,\nu_3]^\top:\mathbb{C}_+\to\mathbb{C}_+^3$ such that:*

1. *For any $\xi\in\mathbb{C}_+$, $\boldsymbol{\nu}(\xi;\boldsymbol{\mu})$ is a solution to $\boldsymbol{\nu}$-system* (3.1)*.*
2. *There exists $\xi_0>0$, such that $|\nu_j(\xi;\boldsymbol{\mu})|\le 2\psi_j/\xi_0$, for all ξ with $\Im(\xi)\ge\xi_0$ and $j=1,2,3$. Moreover, it holds that $\boldsymbol{\nu}(\xi;\boldsymbol{\mu})=\mathbf{m}(\xi;\mathbf{0},\boldsymbol{\mu})$ for all $\xi\in\mathbb{C}_+$.*
3. *$\boldsymbol{\nu}^*=\boldsymbol{\nu}(\sqrt{\lambda}\cdot\mathrm{i};\boldsymbol{\mu})$ in Definition 3.5 satisfies $\nu_j^*=b_j\cdot\mathrm{i}$ with $b_j>0$ for all $j=1,2,3$.*

The proof of Proposition A.8 is given in Section VI in the online supplementary material (Meng et al.). The proposition thus justifies the definition of $\boldsymbol{\nu}(\xi;\boldsymbol{\mu})$ in Section 3 by demonstrating its existence and uniqueness. Moreover, it also relates $\boldsymbol{\nu}(\xi;\boldsymbol{\mu})$ to the function $\mathbf{m}(\xi;\mathbf{q},\boldsymbol{\mu})$ introduced in step 3 of the proof. With this result, we can finalize the proof of Theorem 3.6 as follows.

Proof [Proof of Theorem 3.6] Let

$$\begin{aligned}
\mathcal{R}(\lambda,\boldsymbol{\psi},\boldsymbol{\mu},F_1,\tau)=F_1^2\cdot\big[1-\partial_{q_1}g(\xi^*;\mathbf{q},\boldsymbol{\mu})-\partial^2_{q_4,q_5}g(\xi^*;\mathbf{q},\boldsymbol{\mu})-\partial^2_{q_2,q_5}g(\xi^*;\mathbf{q},\boldsymbol{\mu})\big]\big|_{\mathbf{q}=\mathbf{0}}\\
-\tau^2\cdot\big[\partial^2_{q_3,q_4}g(\xi^*;\mathbf{q},\boldsymbol{\mu})+\partial^2_{q_2,q_3}g(\xi^*;\mathbf{q},\boldsymbol{\mu})\big]\big|_{\mathbf{q}=\mathbf{0}},
\end{aligned}\tag{A.5}$$

where g is defined in (A.3), and $\xi^*=\sqrt{\lambda}\cdot\mathrm{i}$ is given in Definition A.1. Then by Propositions A.2, A.4 and A.7, we have

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon}}\Big|R_d(\mathbf{X},\boldsymbol{\Theta},\lambda,\boldsymbol{\beta}_d,\boldsymbol{\varepsilon})-\mathcal{R}(\lambda,\boldsymbol{\psi},\boldsymbol{\mu},F_1,\tau)\Big|=o_d(1).$$

Therefore to complete the proof, it suffices to calculate the partial derivative terms of $g(\xi^*; \mathbf{q}, \boldsymbol{\mu})$ at $\mathbf{q} = \mathbf{0}$. For this calculation, we first note that by the definition of $L(\xi, \mathbf{z}; \mathbf{q}, \boldsymbol{\mu})$ in (A.4) and the definition of $\mathbf{m}$ in (A.2) as the fixed point of $\mathsf{F}(\cdot; \xi, \mathbf{q}, \boldsymbol{\mu})$, we have that

$$\nabla_{\mathbf{z}} L(\xi, \mathbf{z}; \mathbf{q}, \boldsymbol{\mu})|_{\mathbf{z}=\mathbf{m}} \equiv \mathbf{0}. \tag{A.6}$$

Readers can refer to Lemma V.3 and its proof in the online supplementary material (Meng et al.) for the detailed derivation of (A.6). Let $\mathbf{m}^*(\mathbf{q}, \boldsymbol{\mu}) = [m_1(\xi^*; \mathbf{q}, \boldsymbol{\mu}), m_2(\xi^*; \mathbf{q}, \boldsymbol{\mu}), m_3(\xi^*; \mathbf{q}, \boldsymbol{\mu})]^\top$. Then by Proposition A.8, we have $\boldsymbol{\nu}^* = \mathbf{m}^*(\mathbf{0}, \boldsymbol{\mu})$. Therefore,

$$\begin{aligned}
\partial_{q_1} g(\xi^*; \mathbf{q}, \boldsymbol{\mu})\big|_{\mathbf{q}=\mathbf{0}} &= \partial_{q_1}\big[L(\xi^*, \mathbf{m}^*(\mathbf{q}, \boldsymbol{\mu}); \mathbf{q}, \boldsymbol{\mu})\big]\big|_{\mathbf{q}=\mathbf{0}} \\
&= \big[\big\langle \nabla_{\mathbf{z}} L(\xi^*, \mathbf{z}; \mathbf{q}, \boldsymbol{\mu})|_{\mathbf{z}=\mathbf{m}^*}, \partial_{q_1}\mathbf{m}^*\big\rangle + \partial_{q_1} L(\xi^*, \mathbf{z}; \mathbf{q}, \boldsymbol{\mu})|_{\mathbf{z}=\mathbf{m}^*}\big]\big|_{\mathbf{q}=\mathbf{0}} \\
&= 0 + \partial_{q_1} L(\xi^*, \mathbf{z}; \mathbf{q}, \boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}, \mathbf{z}=\boldsymbol{\nu}^*} = \frac{2\nu_3^* M_N}{M_D},
\end{aligned} \tag{A.7}$$

where the first equality is by the definition of g, the second equality follows by the chain rule, the third equality follows by (A.6), and the last equality is by direct calculation and the definition that $M_N = \nu_1^* \mu_{1,1}^2 + \nu_2^* \mu_{2,1}^2$, $M_D = \nu_3^* M_N - 1$.

For the second order derivatives, let q_i q_j be the i^{th} and j^{th} element in $\mathbf{q}$ for $i, j = 2, 3, 4, 5$. Then by (A.6), with similar calculation as (A.7), we have

$$\frac{\partial^2 g(\xi^*; \mathbf{q}, \boldsymbol{\mu})}{\partial q_i \partial q_j} = \frac{\partial^2 L(\xi^*, \mathbf{z}; \mathbf{q}, \boldsymbol{\mu})}{\partial q_i \partial q_j}\bigg|_{\mathbf{z}=\mathbf{m}^*} + \bigg\langle \nabla_{\mathbf{z}} \bigg[\frac{\partial L(\xi^*, \mathbf{z}; \mathbf{q}, \boldsymbol{\mu})}{\partial q_i}\bigg]\bigg|_{\mathbf{z}=\mathbf{m}^*}, \frac{\partial \mathbf{m}^*}{\partial q_i}\bigg\rangle. \tag{A.8}$$

Moreover, by (A.6) and the formula for implicit differentiation, we have

$$\frac{\partial \mathbf{m}^*}{\partial q_i} = -\big[\big(\nabla_{\mathbf{z}}^2 L(\xi^*, \mathbf{z}; \mathbf{q}, \boldsymbol{\mu})\big)\big|_{\mathbf{z}=\mathbf{m}^*}\big]^{-1} \frac{\partial\big[\nabla_{\mathbf{z}} L(\xi^*, \mathbf{z}; \mathbf{q}, \boldsymbol{\mu})\big]}{\partial q_i}\bigg|_{\mathbf{z}=\mathbf{m}^*}. \tag{A.9}$$

In addition, we let $\mathbf{u} = [q_2, q_3, q_4, q_5, z_1, z_2, z_3]^\top$, and define the symmetric matrix

$$\begin{aligned}
\mathbf{W} &= \mathbf{W}(\boldsymbol{\nu}^*, \boldsymbol{\mu}) = \nabla_{\mathbf{u}}^2 L(\xi, \mathbf{z}; \mathbf{q}, \boldsymbol{\mu})|_{\mathbf{z}=\boldsymbol{\nu}^*, \mathbf{q}=0} \\
&= \begin{bmatrix}
0 & 0 & 0 & 0 & \mu_{1,2}^2 & \mu_{2,2}^2 & 0 \\
* & 0 & 0 & 0 & 0 & 0 & 1 \\
* & * & -\frac{M_N^2}{M_D^2} & -\frac{\nu_3^2 M_N^2}{M_D^2} & \frac{\mu_{1,1}^2}{M_D^2} & \frac{\mu_{2,1}^2}{M_D^2} & \frac{M_N^2}{M_D^2} \\
* & * & * & -\frac{\nu_3^2}{M_D^2} & \frac{\nu_3^2 \mu_{1,1}^2}{M_D^2} & \frac{\nu_3^2 \mu_{2,1}^2}{M_D^2} & \frac{1}{M_D^2} \\
* & * & * & * & -\frac{\nu_3^2 \mu_{1,1}^4}{M_D^2} + \frac{\psi_1}{\nu_1^2} & -\frac{\nu_3^2 \mu_{1,1}^2 \mu_{2,1}^2}{M_D^2} & -\frac{\mu_{1,1}^2}{M_D^2} - \mu_{1,2}^2 \\
* & * & * & * & * & -\frac{\nu_3^2 \mu_{2,1}^4}{M_D^2} + \frac{\psi_2}{\nu_2^2} & -\frac{\mu_{2,1}^2}{M_D^2} - \mu_{2,2}^2 \\
* & * & * & * & * & * & -\frac{M_N^2}{M_D^2} + \frac{\psi_3}{\nu_3^2}
\end{bmatrix}
\end{aligned} \tag{A.10}$$

Then by (A.8), (A.9) and (A.10), we have

$$\frac{\partial^2 g(\xi^*;\mathbf{q},\boldsymbol{\mu})}{\partial q_2 \partial q_5}\Big|_{\mathbf{q}=0} = \mathbf{W}_{1,4} - \mathbf{W}_{1,[5:7]}\Big(\mathbf{W}_{[5:7],[5:7]}\Big)^{-1}\mathbf{W}_{[5:7],4}, \tag{A.11}$$

$$\frac{\partial^2 g(\xi^*;\mathbf{q},\boldsymbol{\mu})}{\partial q_3 \partial q_4}\Big|_{\mathbf{q}=0} = \mathbf{W}_{2,3} - \mathbf{W}_{2,[5:7]}\Big(\mathbf{W}_{[5:7],[5:7]}\Big)^{-1}\mathbf{W}_{[5:7],3}, \tag{A.12}$$

$$\frac{\partial^2 g(\xi^*;\mathbf{q},\boldsymbol{\mu})}{\partial q_2 \partial q_3}\Big|_{\mathbf{q}=0} = \mathbf{W}_{1,2} - \mathbf{W}_{1,[5:7]}\Big(\mathbf{W}_{[5:7],[5:7]}\Big)^{-1}\mathbf{W}_{[5:7],2}, \tag{A.13}$$

$$\frac{\partial^2 g(\xi^*;\mathbf{q},\boldsymbol{\mu})}{\partial q_4 \partial q_5}\Big|_{\mathbf{q}=0} = \mathbf{W}_{3,4} - \mathbf{W}_{3,[5:7]}\Big(\mathbf{W}_{[5:7],[5:7]}\Big)^{-1}\mathbf{W}_{[5:7],4}. \tag{A.14}$$

Now the terms on the right hand side above can be directly calculated: (recalling $\mathbf{V}, \mathbf{H}$ given in Definition 3.5, and ν_j^* is the solution of ***ν-system*** (3.1) given $\xi = \sqrt{\lambda}\cdot \mathrm{i}$) we have

$$\mathbf{W}_{1,4} = \mathbf{W}_{2,3} = \mathbf{W}_{1,2} = 0, \quad \mathbf{W}_{3,4} = -\frac{\nu_3^{*2} M_N^2}{M_D^2},$$
$$\mathbf{W}_{[5:7],[1:4]} = \mathbf{W}_{[1:4],[5:7]}^\top = \mathbf{V}, \quad \text{and} \quad \mathbf{W}_{[5:7],[5:7]} = \mathbf{H}.$$

Plugging (A.7) and (A.11)-(A.14) into (A.5) completes the proof of Theorem 3.6. ■

Finally, recall $\mathbf{L} = \mathbf{V}^\top \mathbf{H}^{-1}\mathbf{V}$, we give the closed form expression for the terms $\mathbf{L}_{1,4}, \mathbf{L}_{2,3}, \mathbf{L}_{1,2}, \mathbf{L}_{3,4}$ in Theorem 3.6. Let $[\nu_1^*, \nu_2^*, \nu_3^*]$, M_N and M_D be defined in Definition 3.5, and

$$\begin{aligned} S =& \nu_3^{*4}\Big(\nu_2^{*2} M_N^2 \mu_{2,1}^4 \psi_1 + \nu_1^{*2} M_N^2 \mu_{1,1}^4 \psi_2 + \nu_1^{*2}\nu_2^{*2} M_D^2 \big(\mu_{1,2}^2\mu_{2,1}^2 - \mu_{1,1}^2\mu_{2,2}^2\big)^2\Big) \\ &- \nu_3^{*2}\nu_2^{*2}\psi_1\big(2M_D^2\mu_{2,1}^2\mu_{2,2}^2 + M_D^4\mu_{2,2}^4 + \mu_{2,1}^4(1 + M_D^2\psi_3)\big) \\ &- \nu_3^{*2}\nu_1^{*2}\psi_2\big(2M_D^2\mu_{1,1}^2\mu_{1,2}^2 + M_D^4\mu_{1,2}^4 + \mu_{1,1}^4(1 + M_D^2\psi_3)\big) \\ &- \nu_3^{*2}\psi_1\psi_2 M_D^2 M_N^2 + M_D^4\psi_1\psi_2\psi_3. \end{aligned} \tag{A.15}$$

Then by direct calculation, the terms $\mathbf{L}_{1,4}, \mathbf{L}_{2,3}, \mathbf{L}_{1,2}, \mathbf{L}_{3,4}$ satisfy the following equations:

$$\begin{aligned}
\frac{S\cdot\mathbf{L}_{1,4}}{\nu_3^{*2}} =& -\nu_3^{*2}M_N^2\Big(\nu_2^{*2}\mu_{2,1}^2\mu_{2,2}^2\psi_1+\nu_1^{*2}\mu_{1,1}^2\mu_{1,2}^2\psi_2\Big)\\
&+\nu_1^{*2}\mu_{1,2}^2\psi_2\Big(M_D^2\mu_{1,2}^2+\mu_{1,1}^2\big(1+M_D^2\psi_3\big)\Big)\\
&+\nu_2^{*2}\mu_{2,2}^2\psi_1\Big(M_D^2\mu_{2,2}^2+\mu_{2,1}^2\big(1+M_D^2\psi_3\big)\Big),\\
\frac{S\cdot\mathbf{L}_{2,3}}{\nu_3^{*2}} =&\nu_2^{*2}\mu_{2,1}^2\big(\mu_{2,1}^2+M_D^2\mu_{2,2}^2\big)\psi_1+\nu_1^{*2}\mu_{1,1}^2\big(\mu_{1,1}^2+M_D^2\mu_{1,2}^2\big)\psi_2\\
&-\nu_3^{*2}M_N^2\big(\nu_2^{*2}\mu_{2,1}^4\psi_1+\nu_1^{*2}\mu_{1,1}^4\psi_2\big)+M_D^2M_N^2\psi_1\psi_2,\\
\frac{S\cdot\mathbf{L}_{1,2}}{\nu_3^{*2}} =&M_D^2\Big(\nu_2^{*2}\mu_{2,2}^2\big(\mu_{2,1}^2+M_D^2\mu_{2,2}^2\big)\psi_1+\nu_1^{*2}\mu_{1,2}^2\big(\mu_{1,1}^2+M_D^2\mu_{1,2}^2\big)\psi_2\\
&-\nu_1^{*2}\nu_2^{*2}\nu_3^{*2}\big(\mu_{1,2}^2\mu_{2,1}^2-\mu_{1,1}^2\mu_{2,2}^2\big)^2\Big),\\
\frac{M_D^2S\cdot\mathbf{L}_{3,4}}{\nu_3^{*2}} =&\nu_3^{*2}\Big(\nu_2^{*2}M_N^2\mu_{2,1}^2\big(M_D^2\mu_{2,2}^2-\mu_{2,1}^2\big)\psi_1+\nu_1^{*2}M_N^2\mu_{1,1}^2\big(M_D^2\mu_{1,2}^2-\mu_{1,1}^2\big)\psi_2\Big)\\
&+\psi_1\psi_2M_D^2M_N^2-\nu_1^{*2}\nu_2^{*2}\nu_3^{*2}M_D^2\big(\mu_{1,2}^2\mu_{2,1}^2-\mu_{1,1}^2\mu_{2,2}^2\big)^2\\
&+\nu_2^{*2}\mu_{2,1}^2\psi_1\big(M_D^2\mu_{2,2}^2+\mu_{2,1}^2+M_D^2\mu_{2,1}^2\psi_3\big)\\
&+\nu_1^{*2}\mu_{1,1}^2\psi_2\big(M_D^2\mu_{1,2}^2+\mu_{1,1}^2+M_D^2\mu_{1,1}^2\psi_3\big).
\end{aligned}\tag{A.16}$$

Clearly, the equations above give explicit calculations of $\mathbf{L}_{1,4}, \mathbf{L}_{2,3}, \mathbf{L}_{1,2}, \mathbf{L}_{3,4}$ given the solution $[\nu_1^*, \nu_2^*, \nu_3^*]$ of the self consistent system $\boldsymbol{\nu}$-*system* (3.1). Readers may keep in mind that ν_j^{*2} is negative since ν_j^* is purely imaginary.

A.5 Discussion on the proof of Theorem 3.6

In this section, we briefly discuss the proof of Theorem 3.6 and highlight the novel challenges we encountered in our own proof and setting compared to Mei and Montanari (2022). We compare the differences in the various steps of the proof to better understand the unique aspects of our extension.

1. The first step in our proof is to directly calculate the excess risk according to its definition, and identify key terms which require further analysis. To do so, we perform a bias-variance decomposition of the risk and find an asymptotic approximation whose main terms are expressed as traces of several random matrices, as detailed in Proposition A.2. Compared to Mei and Montanari (2022), our analysis on the DRFM addresses the impact of different activation functions. As shown in Lemma I.3, the terms become more complex for DRFMs, and it requires additional treatment and careful justification to prove the specific negligible terms. Furthermore, the decomposition in Proposition A.2 includes additional diagonal matrices $\mathbf{M}_1$ and $\mathbf{M}_2$, whereas in Mei and Montanari (2022), it is only a scalar. To address this difference, we have extended several technical lemmas to accommodate the inclusion of $\mathbf{M}_1$ and $\mathbf{M}_2$. For further details, please refer to Section I in the online supplementary material (Meng et al.).

2. The second step involves the construction of a new random matrix, known as the linear pencil matrix. While a similar technique is also used in Mei and Montanari (2022), the linear pencil matrix is more intricate in our setting, see Definition A.3, due to the greater complexity of the terms involving $\mathbf{M}_1$ and $\mathbf{M}_2$. Specifically, the linear pencil matrix in our case is a 3 by 3 block matrix with a more complicated structure than the matrix proposed in Mei and Montanari (2022).
3. The third step is a standard procedure that involves identifying the critical limiting spectral functions of the linear pencil matrix, including its Stieltjes transform and logarithmic potential. However, in our case, these calculations differ from the reference and require additional investigation due to the increased complexity of the linear pencil matrix and the more intricate formula of the related implicit equations. Furthermore, these new calculations of the Stieltjes transform and logarithmic potential provide inspiration for the study of the multiple random feature models and we find a mathematical induction method to complete the study.

Appendix B. Proof of Propositions 4.1 and 4.2

In this section we present the detailed proofs of Propositions 4.1 and 4.2. We denote by $\boldsymbol{\nu}^* = \boldsymbol{\nu}(\sqrt{\lambda}\cdot \mathrm{i}; \boldsymbol{\mu}) = \mathbf{m}(\sqrt{\lambda}\cdot \mathrm{i}; \mathbf{0}, \boldsymbol{\mu})$, Proposition A.8 shows that the three numbers ν_j^*, $j = 1, 2, 3$, are all purely imaginary with positive imaginary parts, that is, $\nu_j^* = \mathrm{i}\nu_j$ where $\nu_j > 0$. Moreover by $\boldsymbol{\nu}$-*system* (3.1), we also have the following self-consistent equations:

$$\begin{cases} \sqrt{\lambda}\nu_1 + \mu_{1,2}^2\nu_1\nu_3 + \dfrac{\mu_{1,1}^2\nu_1\nu_3}{1+\mu_{1,1}^2\nu_1\nu_3+\mu_{2,1}^2\nu_2\nu_3} = \psi_1, \\ \sqrt{\lambda}\nu_2 + \mu_{2,2}^2\nu_2\nu_3 + \dfrac{\mu_{2,1}^2\nu_2\nu_3}{1+\mu_{1,1}^2\nu_1\nu_3+\mu_{2,1}^2\nu_2\nu_3} = \psi_2, \\ \sqrt{\lambda}\nu_3 + \mu_{1,2}^2\nu_1\nu_3 + \mu_{2,2}^2\nu_2\nu_3 + \dfrac{\mu_{1,1}^2\nu_1\nu_3+\mu_{2,1}^2\nu_2\nu_3}{1+\mu_{1,1}^2\nu_1\nu_3+\mu_{2,1}^2\nu_2\nu_3} = \psi_3. \end{cases} \tag{B.1}$$

The system (B.1) can be further rewritten as

$$\begin{cases} \lambda\nu_1\nu_3 = \left(\psi_1 - \mu_{1,2}^2\nu_1\nu_3 - \dfrac{\mu_{1,1}^2\nu_1\nu_3}{1+\mu_{1,1}^2\nu_1\nu_3+\mu_{2,1}^2\nu_2\nu_3}\right) \\ \qquad\qquad \cdot \left(\psi_3 - \mu_{1,2}^2\nu_1\nu_3 - \mu_{2,2}^2\nu_2\nu_3 - \dfrac{\mu_{1,1}^2\nu_1\nu_3+\mu_{2,1}^2\nu_2\nu_3}{1+\mu_{1,1}^2\nu_1\nu_3+\mu_{2,1}^2\nu_2\nu_3}\right), \\ \lambda\nu_2\nu_3 = \left(\psi_2 - \mu_{2,2}^2\nu_2\nu_3 - \dfrac{\mu_{2,1}^2\nu_2\nu_3}{1+\mu_{1,1}^2\nu_1\nu_3+\mu_{2,1}^2\nu_2\nu_3}\right) \\ \qquad\qquad \cdot \left(\psi_3 - \mu_{1,2}^2\nu_1\nu_3 - \mu_{2,2}^2\nu_2\nu_3 - \dfrac{\mu_{1,1}^2\nu_1\nu_3+\mu_{2,1}^2\nu_2\nu_3}{1+\mu_{1,1}^2\nu_1\nu_3+\mu_{2,1}^2\nu_2\nu_3}\right), \\ \sqrt{\lambda}(\nu_1+\nu_2-\nu_3) = \psi_1+\psi_2-\psi_3. \end{cases} \tag{B.2}$$

Our proofs of Propositions 4.1 and 4.2 mainly study the asymptotic properties of $\nu_1\nu_3$ and $\nu_2\nu_3$ based on (B.2). Specifically, we define

$$\chi_1(\boldsymbol{\mu}) = \lim_{\lambda\to 0} \nu_1\nu_3, \quad \chi_2(\boldsymbol{\mu}) = \lim_{\lambda\to 0} \nu_2\nu_3.$$

Note that the existence of these limits with values in $[0,+\infty)\cup\{+\infty\}$ is guaranteed by the property of Stieltjes transform, and the limit value $\chi_1(\boldsymbol{\mu})$, $\chi_2(\boldsymbol{\mu})$ are related with the moment vector $\boldsymbol{\mu}$. In the following proof, we drop the argument $\boldsymbol{\mu}$ in χ_1, χ_2 for simplicity.

B.1 Proof of Proposition 4.1

We first prove the second and fourth conclusions of Proposition 4.1 where the excess risk tends to infinity, and then we prove its first and third conclusions. Readers may keep in mind that when we let $\lambda\to 0$, the moment vector $\boldsymbol{\mu}$ is fixed.

Proof [**Second conclusion**] If $\psi_3 = \psi_1 + \psi_2$, then by (B.2) we have $\nu_1 + \nu_2 = \nu_3$. We first use a proof by contradiction to show that $\chi_1 = \lim_{\lambda\to 0}\nu_1\nu_3 > 0$. It is obvious by definition that $\chi_1 \geq 0$. If $\chi_1 = 0$, then from the first equation in (B.2) we have

$$\begin{aligned}
0 = \lim_{\lambda\to 0}\lambda\nu_1\nu_3 &= \lim_{\lambda\to 0}\left(\psi_1 - \mu_{1,2}^2\nu_1\nu_3 - \frac{\mu_{1,1}^2\nu_1\nu_3}{1+\mu_{1,1}^2\nu_1\nu_3+\mu_{2,1}^2\nu_2\nu_3}\right)\\
&\qquad\cdot\left(\psi_3 - \mu_{1,2}^2\nu_1\nu_3 - \mu_{2,2}^2\nu_2\nu_3 - \frac{\mu_{1,1}^2\nu_1\nu_3+\mu_{2,1}^2\nu_2\nu_3}{1+\mu_{1,1}^2\nu_1\nu_3+\mu_{2,1}^2\nu_2\nu_3}\right),\\
&= \psi_1\cdot\lim_{\lambda\to 0}(\psi_1+\sqrt{\lambda}\nu_2) \geq \psi_1^2.
\end{aligned}$$

This is impossible and hence we have $\chi_1 > 0$. Moreover, if $\chi_1 = +\infty$, we have

$$\begin{aligned}
0 = \lim_{\lambda\to 0}\lambda = \lim_{\lambda\to 0}&\left(\psi_1/(\nu_1\nu_3) - \mu_{1,2}^2 - \frac{\mu_{1,1}^2}{1+\mu_{1,1}^2\nu_1\nu_3+\mu_{2,1}^2\nu_2\nu_3}\right)\\
&\cdot\left(\psi_3 - \mu_{1,2}^2\nu_1\nu_3 - \mu_{2,2}^2\nu_2\nu_3 - \frac{\mu_{1,1}^2\nu_1\nu_3+\mu_{2,1}^2\nu_2\nu_3}{1+\mu_{1,1}^2\nu_1\nu_3+\mu_{2,1}^2\nu_2\nu_3}\right) \gg 0,
\end{aligned}$$

which is also a contradiction. Therefore $0 < \chi_1 < \infty$. Similarly we conclude that $0 < \chi_2 < \infty$.

Furthermore, the relation $\nu_1 + \nu_2 = \nu_3$ implies that $\nu_1, \nu_2 < \nu_3$. Then we have $\lim_{\lambda\to 0}\nu_1, \nu_2 < +\infty$ and $\sqrt{\lambda}\nu_1, \sqrt{\lambda}\nu_2 \to 0$ when $\lambda\to 0$. Therefore (B.1) gives us the following equations when $\lambda\to 0$:

$$\begin{cases}
\mu_{1,2}^2\chi_1 + \dfrac{\mu_{1,1}^2\chi_1}{1+\mu_{1,1}^2\chi_1+\mu_{2,1}^2\chi_2} = \psi_1,\\
\mu_{2,2}^2\chi_2 + \dfrac{\mu_{2,1}^2\chi_2}{1+\mu_{1,1}^2\chi_1+\mu_{2,1}^2\chi_2} = \psi_2.
\end{cases} \tag{B.3}$$

By (B.3), we can express ψ_1, ψ_2 and $\psi_3 = \psi_1 + \psi_2$ by χ_1 and χ_2. Moreover, note that when $\lambda \to 0$,

$$\nu_1^*\nu_3^* = -\chi_1,\ \nu_2^*\nu_3^* = -\chi_2,\ M_N\nu_3^* = -\mu_{1,1}^2\chi_1 - \mu_{2,1}^2\chi_2,\ M_D = -\mu_{1,1}^2\chi_1 - \mu_{2,1}^2\chi_2 - 1. \tag{B.4}$$

Therefore S and $\mathbf{L}_{i,j}$ in (A.15) and (A.16) can also be expressed by χ_1 and χ_2 when $\lambda \to 0$. With direct algebraic calculations, we obtain

$$\lim_{\lambda\to 0} S = 0, \quad \lim_{\lambda\to 0} S\cdot(\mathbf{L}_{3,4} + \mathbf{L}_{1,4}) \neq 0, \quad \lim_{\lambda\to 0} S\cdot(\mathbf{L}_{2,3} + \mathbf{L}_{1,2}) \neq 0.$$

This implies that $\mathbf{L}_{3,4} + \mathbf{L}_{1,4} \to \infty$ and $\mathbf{L}_{2,3} + \mathbf{L}_{1,2} \to \infty$ when $\lambda \to 0$. Since $\mathcal{R} \geq 0$, we have

$$\lim_{\lambda\to 0} \mathcal{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, F_1, \tau) = \lim_{\lambda\to 0} F_1^2\left(\frac{1}{M_D^2} + \mathbf{L}_{3,4} + \mathbf{L}_{1,4}\right) + \tau^2(\mathbf{L}_{2,3} + \mathbf{L}_{1,2}) = +\infty.$$

This gives the second conclusion in Proposition 4.1. ■

Proof [**Fourth conclusion**] If $(\psi_1 + \psi_2)/\psi_3 = 1 + \psi_2/\psi_1$, then $\psi_1 = \psi_3$, and (B.2) gives $\sqrt{\lambda}(\nu_1 + \nu_2 - \nu_3) = \psi_2$. By substitution of $\sqrt{\lambda}(\nu_1 + \nu_2 - \nu_3) = \psi_2$ into the second equation in (B.1) we obtain

$$\sqrt{\lambda}\nu_3 + \mu_{2,2}^2\nu_2\nu_3 + \frac{\mu_{2,1}^2\nu_2\nu_3}{1 + \mu_{1,1}^2\nu_1\nu_3 + \mu_{2,1}^2\nu_2\nu_3} = \sqrt{\lambda}\nu_1.$$

Thus $\nu_3 < \nu_1$. Moreover, if $\chi_1 = \lim_{\lambda\to 0}\nu_1\nu_3 = +\infty$, then the first equation in (B.2) indicates that

$$\begin{aligned} 0 = \lim_{\lambda\to 0}\lambda = \lim_{\lambda\to 0}\Bigg(&\psi_1/(\nu_1\nu_3) - \mu_{1,2}^2 - \frac{\mu_{1,1}^2}{1 + \mu_{1,1}^2\nu_1\nu_3 + \mu_{2,1}^2\nu_2\nu_3}\Bigg) \\ &\cdot\left(\psi_3 - \mu_{1,2}^2\nu_1\nu_3 - \mu_{2,2}^2\nu_2\nu_3 - \frac{\mu_{1,1}^2\nu_1\nu_3 + \mu_{2,1}^2\nu_2\nu_3}{1 + \mu_{1,1}^2\nu_1\nu_3 + \mu_{2,1}^2\nu_2\nu_3}\right) \gg 0, \end{aligned}$$

which is impossible. Therefore $\chi_1 < +\infty$. Similarly, the second equation in (B.2) gives $\chi_2 = \lim_{\lambda\to 0}\nu_2\nu_3 < +\infty$. Here, $\chi_1, \chi_2 < +\infty$ is obtained under a given moment vector $\boldsymbol{\mu}$. Combined $\chi_1 < +\infty$ with $\nu_3 < \nu_1$, we get $\sqrt{\lambda}\nu_3 \to 0$ as $\lambda \to 0$. Therefore the third equation in (B.1) gives us

$$\psi_3 = \mu_{1,2}^2\chi_1 + \mu_{2,2}^2\chi_2 + \frac{\mu_{1,1}^2\chi_1 + \mu_{2,1}^2\chi_2}{1 + \mu_{1,1}^2\chi_1 + \mu_{2,1}^2\chi_2}. \tag{B.5}$$

We remind the readers that we aim at proving

$$\lim_{\mu_{2,1},\mu_{2,2}\to 0}\lim_{\lambda\to 0}\mathcal{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, F_1, \tau) = +\infty. \tag{B.6}$$

To show this, we rely on the following claim (recall that χ_2 depends on $\mu_{2,1}, \mu_{2,2}$):

$$\lim_{\mu_{2,1},\mu_{2,2}\to 0} \mu_{2,2}^2\chi_2 + \mu_{2,1}^2\chi_2 = 0. \tag{B.7}$$

In the following, we first explain how (B.7) can be used to show (B.6), then give the proof of (B.7).

Proof of (B.6) *based on* (B.7). By (B.5) and (B.7), we have

$$\psi_3 = \lim_{\mu_{2,1},\mu_{2,2}\to 0} \mu_{1,2}^2\chi_1 + \frac{\mu_{1,1}^2\chi_1}{1+\mu_{1,1}^2\chi_1}. \tag{B.8}$$

Recall that in Theorem 3.6, $\mathcal{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, F_1, \tau)$ is defined based on the quantities S and $\mathbf{L}_{i,j}$, $i, j = 1, \ldots, 4$. The analytical expressions of these quantities are given in (A.15) and (A.16) respectively. We replace the terms $\boldsymbol{\psi}$ and $\boldsymbol{\nu}^*$ in (A.15) and (A.16) with terms consisting of χ_1 and χ_2 by using equations (B.4), (B.7), (B.8), and then get that

$$\lim_{\mu_{2,1},\mu_{2,2}\to 0} \lim_{\lambda\to 0} S = 0, \quad \lim_{\mu_{2,1},\mu_{2,2}\to 0} \lim_{\lambda\to 0} S\cdot(\mathbf{L}_{3,4}+\mathbf{L}_{1,4}) \neq 0, \quad \lim_{\mu_{2,1},\mu_{2,2}\to 0} \lim_{\lambda\to 0} S\cdot(\mathbf{L}_{2,3}+\mathbf{L}_{1,2}) \neq 0.$$

Therefore the limits $\mathbf{L}_{3,4}+\mathbf{L}_{1,4} = \infty$ and $\mathbf{L}_{2,3}+\mathbf{L}_{1,2} = \infty$ when $\lambda \to 0$ and $\mu_{2,1}, \mu_{2,2} \to 0$. Since $\mathcal{R} > 0$, we have

$$\begin{aligned}
&\lim_{\mu_{2,1},\mu_{2,2}\to 0} \lim_{\lambda\to 0} \mathcal{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, F_1, \tau) \\
&\quad = \lim_{\mu_{2,1},\mu_{2,2}\to 0} \lim_{\lambda\to 0} F_1^2\left(\frac{1}{M_D^2} + \mathbf{L}_{3,4} + \mathbf{L}_{1,4}\right) + \tau^2(\mathbf{L}_{2,3}+\mathbf{L}_{1,2}) = +\infty.
\end{aligned}$$

Proof of (B.7). We first show that $\lim_{\lambda\to 0} \nu_3 = 0$. From the analysis above, we have $\lim_{\lambda\to 0} \nu_3 < +\infty$ due to $\nu_3 < \nu_1$ and $\chi_1 = \lim_{\lambda\to 0} \nu_1\nu_3 < +\infty$. If $\lim_{\lambda\to 0} \nu_3 > 0$, then combined with $\lim_{\lambda\to 0} \nu_1\nu_3 < +\infty$ we have $\sqrt{\lambda}\nu_1$, $\sqrt{\lambda}\nu_2 \to 0$, the first and second equations in (B.1) give us

$$\begin{cases}
\mu_{1,2}^2\chi_1 + \dfrac{\mu_{1,1}^2\chi_1}{1+\mu_{1,1}^2\chi_1+\mu_{2,1}^2\chi_2} = \psi_1, \\
\mu_{2,2}^2\chi_2 + \dfrac{\mu_{2,1}^2\chi_2}{1+\mu_{1,1}^2\chi_1+\mu_{2,1}^2\chi_2} = \psi_2.
\end{cases}$$

Combining the equations above with (B.5), we have $\psi_1 + \psi_2 = \psi_3$, which is a contradiction to the condition $\psi_1 = \psi_3$. Therefore $\lim_{\lambda\to 0} \nu_3 = 0$.

Combining the limit $\lim_{\lambda\to 0} \nu_3 = 0$ with (B.2) yields that $\lim_{\lambda\to 0} \sqrt{\lambda}(\nu_1+\nu_2) = \psi_2$. (B.1) further indicates the existence of $\lim_{\lambda\to 0} \sqrt{\lambda}\nu_1$ and $\lim_{\lambda\to 0} \sqrt{\lambda}\nu_2$ respectively due to the existence of χ_1 and χ_2 (The existence can also be guaranteed by the property of Stieltjes transform). Next we show that $\lim_{\lambda\to 0} \sqrt{\lambda}\nu_1, \lim_{\lambda\to 0} \sqrt{\lambda}\nu_2 > 0$. We use a proof by contradiction:

- If $\lim_{\lambda\to 0}\sqrt{\lambda}\nu_1 = 0$, then it holds that $\lim_{\lambda\to 0}\sqrt{\lambda}\nu_2 = \psi_2$, then we conclude that $\nu_2 \gg \nu_1$, and $\lim_{\lambda\to 0}\nu_1\nu_3 > 0$, $\lim_{\lambda\to 0}\nu_2\nu_3 = 0$ from (B.1). This is a contradiction because $\lim_{\lambda\to 0}\nu_1\nu_3 > 0$ and $\lim_{\lambda\to 0}\nu_2\nu_3 = 0$ indicate $\nu_1 \gg \nu_2$.
- If $\lim_{\lambda\to 0}\sqrt{\lambda}\nu_2 = 0$, we have $\lim_{\lambda\to 0}\sqrt{\lambda}\nu_1 = \psi_2$, then the second equation in (B.1) indicates that $\lim_{\lambda\to 0}\nu_2\nu_3 > 0$. Moreover, $\lim_{\lambda\to 0}\sqrt{\lambda}\nu_2 = 0$ and $\lim_{\lambda\to 0}\sqrt{\lambda}\nu_1 = \psi_2$ indicate that $\nu_1 \gg \nu_2$, therefore $\lim_{\lambda\to 0}\nu_1\nu_3 = +\infty$, which contradicts to the conclusion $\chi_1 = \lim_{\lambda\to 0}\nu_1\nu_3 < +\infty$ above.

From the analysis above we prove that ν_1 and ν_2 have the same order when $\lambda \to 0$. If $\chi_1 = \lim_{\lambda\to 0}\nu_1\nu_3 = 0$, then $\chi_2 = \lim_{\lambda\to 0}\nu_2\nu_3 = 0$. The first and second equations in (B.1) give us $\lim_{\lambda\to 0}\sqrt{\lambda}(\nu_1+\nu_2) \to \psi_1+\psi_2$ which contradicts the third equation in (B.2) which indicates that $\lim_{\lambda\to 0}\sqrt{\lambda}(\nu_1+\nu_2) \to \psi_2$. Therefore we have $\chi_1, \chi_2 > 0$. Here, we utilize the fact $\lim_{\lambda\to 0}\nu_3 = 0$. Finally we have

$$\nu_1 = \Theta\left(\frac{1}{\sqrt{\lambda}}\right),\ \nu_2 = \Theta\left(\frac{1}{\sqrt{\lambda}}\right),\ \nu_3 = \Theta(\sqrt{\lambda}).$$

Then we can assume that

$$\lim_{\lambda\to 0}\sqrt{\lambda}\nu_1 = \psi_1 - n_1,\ \ \lim_{\lambda\to 0}\sqrt{\lambda}\nu_2 = \psi_2 - n_2,\ \ \lim_{\lambda\to 0}\nu_3/\sqrt{\lambda} = k,$$

where $0 \le n_1 < \psi_1$, $0 \le n_2 < \min(\psi_1, \psi_2)$, $k > 0$ and n_1, n_2, k satisfy

$$\begin{cases} \mu_{1,2}^2(\psi_1 - n_1)k + \dfrac{\mu_{1,1}^2(\psi_1-n_1)k}{1+\mu_{1,1}^2(\psi_1-n_1)k+\mu_{2,1}^2(\psi_2-n_2)k} = n_1, \\ \mu_{2,2}^2(\psi_2 - n_2)k + \dfrac{\mu_{2,1}^2(\psi_2-n_2)k}{1+\mu_{1,1}^2(\psi_1-n_1)k+\mu_{2,1}^2(\psi_2-n_2)k} = n_2, \\ n_1 + n_2 = \psi_3 = \psi_1. \end{cases} \tag{B.9}$$

It is easy to see that $\chi_1 = (\psi_1 - n_1)\cdot k$ and $\chi_2 = (\psi_2 - n_2)\cdot k$. We can also show that $\lim_{\mu_{2,1},\mu_{2,2}\to 0} n_2 = 0$. Indeed if $\limsup_{\mu_{2,1},\mu_{2,2}\to 0} n_2 > 0$, the second equation in (B.9) gives $k \to +\infty$. However, $n_2 \cdot k \to +\infty$ leads to a contradiction to the first equation in (B.9). Next, using the second equation in (B.9), we have $\lim_{\mu_{2,1},\mu_{2,2}\to 0}\mu_{2,2}^2\chi_2 = 0$.

As for $\lim_{\mu_{2,1},\mu_{2,2}\to 0}\mu_{2,1}^2\chi_2$, note that χ_1 is bounded by the inequality $\chi_1 < \psi_3/\mu_{1,2}^2$ due to the first equation in (B.9), thus we have

$$0 = \lim_{\mu_{2,1},\mu_{2,2}\to 0}\frac{\mu_{2,1}^2(\psi_2-n_2)k}{1+\mu_{1,1}^2(\psi_1-n_1)k+\mu_{2,1}^2(\psi_2-n_2)k} = \lim_{\mu_{2,1},\mu_{2,2}\to 0}\frac{\mu_{2,1}^2\chi_2}{1+\mu_{1,1}^2\chi_1+\mu_{2,1}^2\chi_2},$$

which indicates that $\lim_{\mu_{2,1},\mu_{2,2}\to 0}\mu_{2,1}^2\chi_2 = 0$. Hence the claim (B.7) holds and the proof is complete. ■

Proof [**Third conclusion**] Let $r = 1-(c_2-1)\psi_3/\psi_2$, then $\psi_3 = \psi_1 + r\psi_2$ with $0 < r < 1$. An analysis similar to the previous case of $\psi_3 = \psi_1$ leads to $\nu_3 < \nu_1 + \nu_2$, and $\nu_1 = \Theta(\frac{1}{\sqrt{\lambda}}), \nu_2 = \Theta(\frac{1}{\sqrt{\lambda}})$, and $\nu_3 = \Theta(\sqrt{\lambda})$. We still assume that

$$\lim_{\lambda\to 0}\sqrt{\lambda}\nu_1 = \psi_1 - n_1,\ \ \lim_{\lambda\to 0}\sqrt{\lambda}\nu_2 = \psi_2 - n_2,\ \ \lim_{\lambda\to 0}\nu_3/\sqrt{\lambda} = k,$$

where $0 \le n_1 < \psi_1$, $r\psi_2 \le n_2 < \psi_2$, $k > 0$ and n_1, n_2, k satisfy

$$\begin{cases} \mu_{1,2}^2(\psi_1 - n_1)k + \dfrac{\mu_{1,1}^2(\psi_1-n_1)k}{1+\mu_{1,1}^2(\psi_1-n_1)k + \mu_{2,1}^2(\psi_2-n_2)k} = n_1, \\ \mu_{2,2}^2(\psi_2 - n_2)k + \dfrac{\mu_{2,1}^2(\psi_2-n_2)k}{1+\mu_{1,1}^2(\psi_1-n_1)k + \mu_{2,1}^2(\psi_2-n_2)k} = n_2, \\ n_1 + n_2 = \psi_3 = \psi_1 + r\psi_2. \end{cases} \tag{B.10}$$

It is easy to see that $\chi_1 = (\psi_1 - n_1)\cdot k$ and $\chi_2 = (\psi_2 - n_2)\cdot k$. Let $\mu_{2,1}, \mu_{2,2} \to 0$ and note that $n_2 \ge r\psi_2$. We must have $k \to +\infty$ by the second equation in (B.10). Therefore the first equation in (B.10) indicates that $n_1 = \psi_1$ and $n_2 = r\psi_2$ as $\mu_{2,1}, \mu_{2,2} \to 0$. Now it is easy to prove the third conclusion in Proposition 4.1 if we further assume that $\mu_{2,1}/\mu_{2,2} \to 0$ due to

$$\varliminf_{\mu_{2,1},\mu_{2,2}\to 0}\lim_{\lambda\to 0}\mathcal{R}(\lambda,\boldsymbol{\psi},\boldsymbol{\mu},F_1,\tau) \le \lim_{\substack{\mu_{2,1},\mu_{2,2}\to 0\\ \mu_{2,1}/\mu_{2,2}\to 0}}\mathcal{R}(\lambda,\boldsymbol{\psi},\boldsymbol{\mu},F_1,\tau).$$

Define $\bar{\chi}_1 = \lim_{\substack{\mu_{2,1},\mu_{2,2}\to 0\\ \mu_{2,1}/\mu_{2,2}\to 0}}\chi_1$. Then we have

$$\lim_{\substack{\mu_{2,1},\mu_{2,2}\to 0\\ \mu_{2,1}/\mu_{2,2}\to 0}}\mu_{2,2}^2\chi_2 = r\psi_2,\ \ \mu_{1,2}^2\bar{\chi}_1 + \frac{\mu_{1,1}^2\bar{\chi}_1}{1+\mu_{1,1}^2\bar{\chi}_1} = \psi_1.$$

Combining the expression of S and $\mathbf{L}_{i,j}$ in (A.15) and (A.16) gives us

$$\lim_{\substack{\mu_{2,1},\mu_{2,2}\to 0\\ \mu_{2,1}/\mu_{2,2}\to 0}}\lim_{\lambda\to 0} S = (1-r)r\bar{\chi}_1(\psi_2 + \mu_{1,1}^2\psi_2\bar{\chi}_1)^2(\mu_{1,2}^2 + \mu_{1,1}^4\mu_{1,2}^2\bar{\chi}_1^2 + \mu_{1,1}^2(1+2\mu_{1,2}^2\bar{\chi}_1)) > 0,$$

$$\lim_{\substack{\mu_{2,1},\mu_{2,2}\to 0\\ \mu_{2,1}/\mu_{2,2}\to 0}}\lim_{\lambda\to 0}|S\cdot(\mathbf{L}_{3,4}M_D^2 + \mathbf{L}_{1,4})| < +\infty,\quad \lim_{\substack{\mu_{2,1},\mu_{2,2}\to 0\\ \mu_{2,1}/\mu_{2,2}\to 0}}\lim_{\lambda\to 0}|S\cdot(\mathbf{L}_{2,3} + \mathbf{L}_{1,2})| < +\infty.$$

Therefore $\varliminf_{\mu_{2,1},\mu_{2,2}\to 0}\lim_{\lambda\to 0}\mathcal{R}(\lambda,\boldsymbol{\psi},\boldsymbol{\mu},F_1,\tau) \le \lim_{\substack{\mu_{2,1},\mu_{2,2}\to 0\\ \mu_{2,1}/\mu_{2,2}\to 0}}\mathcal{R}(\lambda,\boldsymbol{\psi},\boldsymbol{\mu},F_1,\tau) < +\infty$. This completes the proof of the third conclusion in Proposition 4.1. ■

Proof [**First conclusion**] Let $r = 1 + (1 - c_1)\psi_3/\psi_2$, then we have $\psi_3 = \psi_1 + r\psi_2$ with $r > 1$. Similarly to the previous arguments, we obtain $\nu_1\nu_3 = \boldsymbol{\Theta}_\lambda(1)$ and $\nu_2\nu_3 = \boldsymbol{\Theta}_\lambda(1)$. Also note that $\nu_1 + \nu_2 < \nu_3$, therefore it holds that $\sqrt{\lambda}\nu_1 \to 0$ and $\sqrt{\lambda}\nu_1 \to 0$. Recall that we defined $\chi_1 = \lim_{\lambda\to 0}\nu_1\nu_3$ and $\chi_2 = \lim_{\lambda\to 0}\nu_2\nu_3$, and the system (B.3) still holds in the current case. Substituting (B.3) into (A.15) and (A.16), and after some simple calculation, we obtain

$$\lim_{\lambda\to 0} S > (r-1)\mu_{1,1}^4\mu_{2,1}^2(1+\mu_{2,2}^2\chi_2)\chi_1^2\chi_2^2 > 0,\ \lim_{\lambda\to 0} S\cdot(\mathbf{L}_{3,4}M_D^2 + \mathbf{L}_{1,4}) < +\infty, \lim_{\lambda\to 0} S\cdot(\mathbf{L}_{2,3} + \mathbf{L}_{1,2}) < +\infty.$$

Therefore when $\psi_3 = \psi_1 + r\psi_2$, $r > 1$, $\lim_{\lambda\to 0}\mathcal{R}(\lambda, \boldsymbol{\psi}, \boldsymbol{\mu}, F_1, \tau) < +\infty$. This completes the proof of the first conclusion in Proposition 4.1. ■

B.2 Proof of Proposition 4.2

For this proposition, we let $\psi_0 = \psi_1/r_1 = \psi_2/r_2 \to +\infty$. By the system (B.1) we have

$$\sqrt{\lambda}\nu_3 = \psi_3 - \mu_{1,2}^2\nu_1\nu_3 - \mu_{2,2}^2\nu_2\nu_3 - \frac{\mu_{1,1}^2\nu_1\nu_3 + \mu_{2,1}^2\nu_2\nu_3}{1 + \mu_{1,1}^2\nu_1\nu_3 + \mu_{2,1}^2\nu_2\nu_3}. \tag{B.11}$$

Therefore $\nu_3 < \psi_3/\sqrt{\lambda}$ with fixed ψ_3. Then from the first and second equations in (B.1) we easily get that $\lim_{\psi_0\to+\infty}\nu_1$, $\lim_{\psi_0\to+\infty}\nu_2 = +\infty$. If $\overline{\lim}_{\psi_0\to+\infty}\nu_3 > 0$, further from (B.11) we will get

$$\overline{\lim_{\psi_0\to+\infty}}\sqrt{\lambda}\left(\nu_3 + \mu_{1,2}^2\nu_1\nu_3 + \mu_{2,2}^2\nu_2\nu_3 + \frac{\mu_{1,1}^2\nu_1\nu_3 + \mu_{2,1}^2\nu_2\nu_3}{1 + \mu_{1,1}^2\nu_1\nu_3 + \mu_{2,1}^2\nu_2\nu_3}\right) = \psi_3.$$

This is a contradiction because the left hand side of the equation above tends to infinity while the right hand side is fixed. Therefore we have $\lim_{\psi_0\to+\infty}\nu_3 = 0$. Combining this result with

$$\sqrt{\lambda}\nu_1 + \mu_{1,2}^2\nu_1\nu_3 + \frac{\mu_{1,1}^2\nu_1\nu_3}{1 + \mu_{1,1}^2\nu_1\nu_3 + \mu_{2,1}^2\nu_2\nu_3} = \psi_1,$$
$$\sqrt{\lambda}\nu_2 + \mu_{2,2}^2\nu_2\nu_3 + \frac{\mu_{2,1}^2\nu_2\nu_3}{1 + \mu_{1,1}^2\nu_1\nu_3 + \mu_{2,1}^2\nu_2\nu_3} = \psi_2,$$

we conclude that

$$\lim_{\psi_0\to+\infty}\nu_1/\psi_0 = r_1/\sqrt{\lambda},\quad \lim_{\psi_0\to+\infty}\nu_2/\psi_0 = r_2/\sqrt{\lambda}.$$

We further define

$$\overline{\lim_{\psi_0\to+\infty}}\ \nu_3\psi_0=\overline{\chi},\quad \underline{\lim_{\psi_0\to+\infty}}\ \nu_3\psi_0=\underline{\chi}.$$

Then we have

$$\overline{\lim_{\psi_0\to+\infty}}\ \nu_1\nu_3=r_1\overline{\chi},\quad \overline{\lim_{\psi_0\to+\infty}}\ \nu_2\nu_3=r_2\overline{\chi},\quad \underline{\lim_{\psi_0\to+\infty}}\ \nu_1\nu_3=r_1\underline{\chi},\quad \underline{\lim_{\psi_0\to+\infty}}\ \nu_2\nu_3=r_2\underline{\chi}.$$

Taking the superior and inferior limit when $\psi_0\to+\infty$ in the third equation of (B.1), we have

$$\begin{cases}\psi_3=\mu_{1,2}^2r_1\overline{\chi}+\mu_{2,2}^2r_2\overline{\chi}+\dfrac{\mu_{1,1}^2r_1\overline{\chi}+\mu_{2,1}^2r_2\overline{\chi}}{1+\mu_{1,1}^2r_1\overline{\chi}+\mu_{2,1}^2r_2\overline{\chi}},\\ \psi_3=\mu_{1,2}^2r_1\underline{\chi}+\mu_{2,2}^2r_2\underline{\chi}+\dfrac{\mu_{1,1}^2r_1\underline{\chi}+\mu_{2,1}^2r_2\underline{\chi}}{1+\mu_{1,1}^2r_1\underline{\chi}+\mu_{2,1}^2r_2\underline{\chi}}.\end{cases}$$

Therefore $\overline{\chi}$ and $\underline{\chi}$ are both the solution of the equation

$$\psi_3(1+\mu_{1,1}^2r_1x+\mu_{2,1}^2r_2x)=(\mu_{1,2}^2r_1x+\mu_{2,2}^2r_2x)(1+\mu_{1,1}^2r_1x+\mu_{2,1}^2r_2x)+\mu_{1,1}^2r_1x+\mu_{2,1}^2r_2x. \tag{B.12}$$

Note that $\overline{\chi}$ and $\underline{\chi}$ are both positive, and the equation above only has one positive root. Therefore we conclude that $\overline{\chi}=\underline{\chi}$, and we can write $\chi:=\overline{\chi}=\underline{\chi}$. By calculating the positive root of (B.12), we easily see that $(r_1\mu_{1,1}^2+r_2\mu_{2,1}^2)\chi=\chi_0$ where χ_0 is defined in Proposition 4.2. Plugging the limits $\nu_1\nu_3\to r_1\chi$ and $\nu_2\nu_3\to r_2\chi$ into M_D and M_N in (A.15) and (A.16), we obtain $M_D\to-\chi_0-1$, $\nu_3^*M_N\to-\chi_0$ when $\psi_0\to+\infty$. Direct algebraic calculation then gives

$$\mathbf{L}_{2,3}\to\frac{\chi_0^2}{(\chi_0+1)^2\psi_3-\chi_0^2},\qquad \mathbf{L}_{3,4}\to\frac{\chi_0^2}{(\chi_0+1)^4\psi_3-\chi_0^2(\chi_0+1)^2},\qquad \mathbf{L}_{1,2},\mathbf{L}_{1,4}\to0$$

when $\psi_0\to+\infty$. Then we have

$$\begin{aligned}\lim_{\psi_0\to\infty}\mathcal{R}(\lambda,\boldsymbol{\psi},\boldsymbol{\mu},F_1,\tau)&=\lim_{\psi_0\to\infty}F_1^2\bigg(\frac{1}{M_D^2}+\mathbf{L}_{3,4}+\mathbf{L}_{1,4}\bigg)+\tau^2(\mathbf{L}_{2,3}+\mathbf{L}_{1,2})\\&=F_1^2\bigg(\frac{1}{(\chi_0+1)^2}+\frac{\chi_0^2}{(\chi_0+1)^4\psi_3-\chi_0^2(\chi_0+1)^2}\bigg)+\tau^2\bigg(\frac{\chi_0^2}{(\chi_0+1)^2\psi_3-\chi_0^2}\bigg)\\&=\frac{F_1^2\psi_3+\tau^2\chi_0^2}{(\chi_0+1)^2\psi_3-\chi_0^2}.\end{aligned}$$

This proves Proposition 4.2.

Appendix C. Proof of Theorem 5.6

Here, we provide the proof of Theorem 5.6 for the MRFM. The proof for MRFM bears significant resemblance to the previous proof of Theorem 3.6. In this section, we offer a brief overview of the proof for MRFM, highlighting the key distinctions between these two theorems. Here we will focus on several key steps in the proof that are significantly different from the proof of DRFMs.

C.1 Step 1: bias-variance decomposition of the excess risk

We first give some notations as follows.

Definition C.1. *Define*

$$\mathbf{Z}_j = \sigma_j\left(\mathbf{X}\boldsymbol{\Theta}_j^\top/\sqrt{d}\right)/\sqrt{d} \in \mathbb{R}^{n\times N_j},\ \mathbf{Z} = \big[\mathbf{Z}_1,\ldots,\mathbf{Z}_K\big],$$
$$\boldsymbol{\sigma}(\mathbf{x}) = \big[\sigma_1(\mathbf{x}^\top\boldsymbol{\Theta}_1^\top/\sqrt{d}),\ldots,\sigma_K(\mathbf{x}^\top\boldsymbol{\Theta}_K^\top/\sqrt{d})\big]^\top \in \mathbb{R}^N,\ \boldsymbol{\Upsilon} = (\mathbf{Z}^\top\mathbf{Z}+\lambda\mathbf{I}_N)^{-1},$$
$$\mathbf{V}_0(F_0) = \mathbb{E}_{\mathbf{x}}\big[\boldsymbol{\sigma}(\mathbf{x})F_0\big] \in \mathbb{R}^{N\times 1},\ \mathbf{V}(\boldsymbol{\beta}_{1,d}) = \mathbb{E}_{\mathbf{x}}\big[\boldsymbol{\sigma}(\mathbf{x})\mathbf{x}^\top\boldsymbol{\beta}_{1,d}\big] \in \mathbb{R}^{N\times 1}, \mathbf{U} = \mathbb{E}_{\mathbf{x}}\big[\boldsymbol{\sigma}(\mathbf{x})\boldsymbol{\sigma}(\mathbf{x})^\top\big] \in \mathbb{R}^{N\times N}.$$

Clearly, these notations are consistent with Definition A.1 and Proposition A.2. Based on these notations with direct calculation, we can express the excess risk $R_d(\mathbf{X},\boldsymbol{\Theta},\lambda,\boldsymbol{\beta}_d,\boldsymbol{\varepsilon})$ as

$$R_d(\mathbf{X},\boldsymbol{\Theta},\lambda,\boldsymbol{\beta}_d,\boldsymbol{\varepsilon}) = F_0^2 + F_{1,d}^2 - 2\mathbf{y}^\top\mathbf{Z}\boldsymbol{\Upsilon}[\mathbf{V}(\boldsymbol{\beta}_{1,d})+\mathbf{V}_0(F_0)]/\sqrt{d} + \mathbf{y}^\top\big[\mathbf{U}\big]_{\mathbf{Z}}\mathbf{y}/d. \quad \text{(C.1)}$$

To continue the calculation, we consider the Gegenbauer decompositions of the activation functions. Suppose that the Gegenbauer decompositions of $\sigma_j(\cdot)$, $j=1,\ldots,K$, are

$$\sigma_j(x) = \sum_{k=0}^{+\infty}\lambda_{d,k}\big(\sigma_j\big)B(d,k)\cdot Q_k^{(d)}(\sqrt{d}\cdot x), \quad j=1,\ldots,K,$$

where $\lambda_{d,k}(\sigma_j)$ are the decomposition coefficients, $Q_k^{(d)}$, $k\in\mathbb{N}$ are the Gegenbauer polynomials, and $B(d,0)=1$, $B(d,k)=k^{-1}(2k+d-2)\binom{k+d-3}{k-1}$ for $k\geq 1$. Let

$$\boldsymbol{\Lambda}_{d,k} = \mathrm{diag}\big(\lambda_{d,k}(\sigma_1)\mathbf{I}_{N_1},\ldots,\lambda_{d,k}(\sigma_K)\mathbf{I}_{N_K}\big), \quad k\in\mathbb{N}=\{0,1,...\}, \quad \text{(C.2)}$$
$$\mathbf{M}_1 = \mathrm{diag}\big(\mu_{1,1}\mathbf{I}_{N_1},\ldots,\mu_{K,1}\mathbf{I}_{N_K}\big), \quad \mathbf{M}_2 = \mathrm{diag}\big(\mu_{1,2}\mathbf{I}_{N_1},\ldots,\mu_{K,2}\mathbf{I}_{N_K}\big). \quad \text{(C.3)}$$

Now we present Proposition C.2 below, which is the counterpart of Proposition A.2.

Proposition C.2. *For any given λ, let*

$$\overline{R}_d(\mathbf{X},\boldsymbol{\Theta},\lambda,F_{1,d},\tau) = F_{1,d}^2 - \frac{2F_{1,d}^2}{d}\mathrm{tr}\mathbf{M}_1\frac{\boldsymbol{\Theta}\mathbf{X}^\top}{d}\mathbf{Z}\boldsymbol{\Upsilon} + \frac{F_{1,d}^2}{d}\mathrm{tr}\Big(\big[\widetilde{\mathbf{U}}\big]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}\Big) + \frac{\tau^2}{d}\mathrm{tr}(\big[\widetilde{\mathbf{U}}\big]_{\mathbf{Z}}),$$

where $\widetilde{\mathbf{U}} = \mathbf{M}_1\boldsymbol{\Theta}\boldsymbol{\Theta}^\top\mathbf{M}_1/d + \mathbf{M}_2\mathbf{M}_2$. Then under the same conditions as Theorem 5.6,

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon}}\Big|R_d(\mathbf{X},\boldsymbol{\Theta},\lambda,\boldsymbol{\beta}_d,\boldsymbol{\varepsilon}) - \overline{R}_d(\mathbf{X},\boldsymbol{\Theta},\lambda,F_{1,d},\tau)\Big| = o_d(1).$$

The proof for Proposition C.2 is exactly the same as the proof for Proposition A.2, except the definitions of $\mathbf{\Lambda}_{d,k}$, $\mathbf{M}_1$ and $\mathbf{M}_2$ are changed. We therefore omit the proof details.

C.2 Step 2: approximation of the risk decomposition via a linear pencil matrix

The approximating function $\overline{R}_d(\mathbf{X}, \mathbf{\Theta}, \lambda, F_{1,d}, \tau)$ established in Proposition C.2 again depends on traces of several random matrices. These traces are next evaluated using a new linear pencil matrix, which is a bit more involved compared with the linear pencil matrix for DRFMs.

Definition C.3. *(1) Let* $\mathcal{Q} := \{\mathbf{q} = [q_1, q_2, q_3, q_4, q_5] \in \mathbb{R}_+^5 : q_4, q_5 \le (1+q_1)/2, \|\mathbf{q}\|_2 \le 1\}$. *Depending on* $\mathbf{q} \in \mathcal{Q}$ *and* $\boldsymbol{\mu}$*, the linear pencil matrix* $\mathbf{A}(\mathbf{q}, \boldsymbol{\mu}) \in \mathbb{R}^{P\times P}$ *(*$P = N + n$*) is*

$$\begin{aligned}\mathbf{A}(\mathbf{q}, \boldsymbol{\mu}) &= \begin{bmatrix} q_2\mathbf{M}_2\mathbf{M}_2 + q_4\mathbf{M}_1\frac{\mathbf{\Theta}\mathbf{\Theta}^\top}{d}\mathbf{M}_1 & \mathbf{Z}^\top + q_1\widetilde{\mathbf{Z}}^\top \\ \mathbf{Z} + q_1\widetilde{\mathbf{Z}} & q_3\mathbf{I}_n + q_5\frac{\mathbf{X}\mathbf{X}^\top}{d}\end{bmatrix} \\ &= \begin{bmatrix} q_2\mu_{1,2}^2\mathbf{I}_{N_1} + q_4\mu_{1,1}^2\frac{\mathbf{\Theta}_1\mathbf{\Theta}_1^\top}{d} & \cdots & q_4\mu_{1,1}\mu_{K,1}\frac{\mathbf{\Theta}_1\mathbf{\Theta}_K^\top}{d} & \mathbf{Z}_1^\top + q_1\widetilde{\mathbf{Z}}_1^\top \\ \vdots & \ddots & \vdots & \vdots \\ q_4\mu_{K,1}\mu_{1,1}\frac{\mathbf{\Theta}_K\mathbf{\Theta}_1^\top}{d} & \cdots & q_2\mu_{K,2}^2\mathbf{I}_{N_K} + q_4\mu_{K,1}^2\frac{\mathbf{\Theta}_K\mathbf{\Theta}_K^\top}{d} & \mathbf{Z}_K^\top + q_1\widetilde{\mathbf{Z}}_K^\top \\ \mathbf{Z}_1 + q_1\widetilde{\mathbf{Z}}_1 & \cdots & \mathbf{Z}_K + q_1\widetilde{\mathbf{Z}}_K & q_3\mathbf{I}_n + q_5\frac{\mathbf{X}\mathbf{X}^\top}{d}\end{bmatrix},\end{aligned}$$

where $\widetilde{\mathbf{Z}}_j = \frac{\mu_{j,1}}{d}\mathbf{X}\mathbf{\Theta}_j^\top$, $j = 1, \ldots, K+1$.

(2) The Stieltjes transform of the empirical eigenvalue distribution of $\mathbf{A}$ *(up to a* P/d *factor) is*

$$M_d(\xi; \mathbf{q}, \boldsymbol{\mu}) = \frac{1}{d}\mathrm{tr}\big[(\mathbf{A} - \xi\mathbf{I}_P)^{-1}\big], \quad \xi \in \mathbb{C}_+,$$

and its logarithmic potential is

$$G_d(\xi; \mathbf{q}, \boldsymbol{\mu}) = \frac{1}{d}\log\det(\mathbf{A} - \xi\mathbf{I}_P) = \frac{1}{d}\sum_{i=1}^{P}\log(\lambda_i(\mathbf{A}) - \xi), \quad \xi \in \mathbb{C}_+.$$

Here $\{\lambda_i(\mathbf{A})\}_{i\in[P]}$ *are the eigenvalues of* $\mathbf{A}$ *in decreasing order, and* $\log(z) := \log(|z|) + \mathrm{i}\arg(z)$*, for* $z \in \mathbb{C}$*,* $-\pi < \arg(z) \le \pi$ *is the principal value of a complex logarithmic function.*

The three traces appearing in the definition of $\overline{R}_d(\mathbf{X}, \mathbf{\Theta}, \lambda, F_{1,d}, \tau)$ in Proposition C.2 are now expressed as partial derivatives of the logarithmic potential G_d as shown in the proposition below.

Proposition C.4. *Let ξ^* be defined in Definition C.1 and $\widetilde{\mathbf{U}}$ be defined in Proposition C.2. Then*

$$\begin{aligned}
\frac{1}{d}\mathrm{tr}\mathbf{M}_1\frac{\mathbf{\Theta}\mathbf{X}^\top}{d}\mathbf{Z}\mathbf{\Upsilon} &= \frac{1}{2}\partial_{q_1}G_d(\xi^*;\mathbf{q},\boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}},\\
\frac{1}{d}\mathrm{tr}(\big[\widetilde{\mathbf{U}}\big]_{\mathbf{Z}}\frac{\mathbf{X}\mathbf{X}^\top}{d}) &= -\partial^2_{q_4,q_5}G_d(\xi^*;\mathbf{q},\boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}} - \partial^2_{q_2,q_5}G_d(\xi^*;\mathbf{q},\boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}},\\
\frac{1}{d}\mathrm{tr}(\big[\widetilde{\mathbf{U}}\big]_{\mathbf{Z}}) &= -\partial^2_{q_3,q_4}G_d(\xi^*;\mathbf{q},\boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}} - \partial^2_{q_2,q_3}G_d(\xi^*;\mathbf{q},\boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}}.
\end{aligned}$$

The proof for Proposition C.4 is the same as the proof for Proposition A.4. We omit the details for simplicity.

C.3 Step 3: key limiting spectral functions of the linear pencil matrix

Proposition C.4 shows that the excess risk depends on the limiting spectral properties of the linear pencil matrix $\mathbf{A}$. Therefore we study the Stieltjes transform $M_d(\xi;\mathbf{q},\boldsymbol{\mu})$ of the empirical eigenvalue distribution of $\mathbf{A}$ and calculate its limit as $d,n,N\to\infty$. We first give the following definition.

Definition C.5. *Define* $\mathsf{F}(\cdot;\xi,\mathbf{q},\boldsymbol{\mu}) = [\mathsf{F}_1(\cdot;\xi,\mathbf{q},\boldsymbol{\mu}),\cdots,\mathsf{F}_{K+1}(\cdot;\xi,\mathbf{q},\boldsymbol{\mu})]^\top : \mathbb{C}^{K+1}\to\mathbb{C}^{K+1}$ *as*

$$\begin{aligned}
\mathsf{F}_j(\mathbf{m};\xi,\mathbf{q},\boldsymbol{\mu}) &= \psi_j\bigg\{-\xi+q_2\mu_{j,2}^2-\mu_{j,2}^2m_{K+1}+\frac{H_j}{H_D}\bigg\}^{-1},\quad j=1,\ldots,K,\\
\mathsf{F}_{K+1}(\mathbf{m};\xi,\mathbf{q},\boldsymbol{\mu}) &= \psi_{K+1}\bigg\{-\xi+q_3-\sum_{j=1}^K\mu_{j,2}^2m_j+\frac{H_{K+1}}{H_D}\bigg\}^{-1},
\end{aligned}$$

where $\xi\in\mathbb{C}_+$, $\mathbf{m}=[m_1,\ldots,m_{K+1}]\in\mathbb{C}^{K+1}$, *and*

$$\begin{aligned}
H_j &= \mu_{j,1}^2q_4(1+m_{K+1}q_5)-\mu_{j,1}^2(1+q_1)^2m_{K+1},\quad j=1,\ldots,K,\\
H_{K+1} &= q_5\Big(1+\sum_{j=1}^K\mu_{j,1}^2m_jq_4\Big)-(1+q_1)^2\sum_{j=1}^K\mu_{j,1}^2m_j,\\
H_D &= \Big(1+\sum_{j=1}^K\mu_{j,1}^2m_jq_4\Big)(1+m_{K+1}q_5)-(1+q_1)^2\sum_{j=1}^K\mu_{j,1}^2m_jm_{K+1}.
\end{aligned}$$

Note that the function $\mathsf{F}(\mathbf{m};\xi,\mathbf{q},\boldsymbol{\mu})$ in Definition C.5 above is not related to d. Lemma C.6 below ensures the existence and uniqueness of the fixed point of $\mathsf{F}(\mathbf{m};\xi,\mathbf{q},\boldsymbol{\mu})$ for $\xi\in\{\xi\in\mathbb{C}:\Im(\xi)>\xi_0\}$ with some sufficiently large constant ξ_0.

Lemma C.6. *For $\mathsf{F}(\mathbf{m};\xi,\mathbf{q},\boldsymbol{\mu})$ in Definition C.5, there exists $\xi_0>0$ such that, for any $\xi\in\mathbb{C}_+$ with $\Im(\xi)>\xi_0$, the equation $\mathbf{m}=\mathsf{F}(\mathbf{m};\xi,\mathbf{q},\boldsymbol{\mu})$ admits a unique solution in $\mathbb{D}(2\psi_1/\xi_0)\times\ldots\times\mathbb{D}(2\psi_{K+1}/\xi_0)$.*

The proof of Lemma C.6 is given in Section VII.1 in the online supplementary material (Meng et al.). Define the fixed point of $\mathbf{F}(\mathbf{m};\xi,\mathbf{q},\boldsymbol{\mu})$ as the function of ξ on $\{\xi:\Im(\xi)>\xi_0\}$:

$$\mathbf{m}(\xi;\mathbf{q},\boldsymbol{\mu})=\begin{bmatrix} m_1(\xi;\mathbf{q},\boldsymbol{\mu}) \\ \vdots \\ m_{K+1}(\xi;\mathbf{q},\boldsymbol{\mu}) \end{bmatrix} \tag{C.4}$$

The following proposition shows that $\mathbf{m}$ is an analytic function on $\{\xi:\Im(\xi)>\xi_0\}$, and its analytic continuation to $\mathbb{C}_+$ is still a fixed point of $\mathbf{F}(\cdot;\xi,\mathbf{q},\boldsymbol{\mu})$.

Proposition C.7. *Under Assumptions 5.2 and 5.3, $\mathbf{m}(\xi;\mathbf{q},\boldsymbol{\mu})$ is analytic on $\{\xi:\Im(\xi)>\xi_0\}$, and has a unique analytic continuation to $\mathbb{C}_+$. Moreover, this analytic continuation (still denoted as $\mathbf{m}(\xi;\mathbf{q},\boldsymbol{\mu})$) satisfies the following properties:*

1. *$\mathbf{m}(\xi;\mathbf{q},\boldsymbol{\mu})\in\mathbb{C}_+^{K+1}$ for all $\xi\in\mathbb{C}_+$.*
2. *$\mathbf{m}(\xi,\mathbf{q},\boldsymbol{\mu})\equiv\mathbf{F}[\mathbf{m}(\xi,\mathbf{q},\boldsymbol{\mu});\xi,\mathbf{q},\boldsymbol{\mu}]$ for all $\xi\in\mathbb{C}_+$.*
3. *Let $M_d(\xi;\mathbf{q},\boldsymbol{\mu})$ be defined in Definition C.3. Then for any compact set $\Omega\subset\mathbb{C}_+$,*

$$\lim_{d\to+\infty}\mathbb{E}\bigg[\sup_{\xi\in\Omega}\big|M_d(\xi;\mathbf{q},\boldsymbol{\mu})-\sum_{j=1}^{K+1}m_j(\xi;\mathbf{q},\boldsymbol{\mu})\big|\bigg]=0.$$

The proof of Proposition C.7 is given in Section VII.2 in the online supplementary material (Meng et al.). The study of the limiting spectral distribution also leads to a deterministic limit for the logarithmic potential G_d. This limit logarithmic potential is found to be

$$g(\xi;\mathbf{q},\boldsymbol{\mu})\triangleq L(\xi,m_1(\xi;\mathbf{q},\boldsymbol{\mu}),\ldots,m_{K+1}(\xi;\mathbf{q},\boldsymbol{\mu});\mathbf{q},\boldsymbol{\mu}), \tag{C.5}$$

where the function L is

$$\begin{aligned} &L(\xi,z_1,\ldots,z_{K+1};\mathbf{q},\boldsymbol{\mu})\triangleq \\ &\quad\log\bigg[\Big(1+q_4\sum_{j=1}^{K}\mu_{j,1}^2z_j\Big)(1+z_{K+1}q_5)-\sum_{j=1}^{K}\mu_{j,1}^2(1+q_1)^2z_jz_{K+1}\bigg]-\sum_{j=1}^{K}\mu_{j,2}^2z_jz_{K+1} \\ &\quad+q_2\sum_{j=1}^{K}\mu_{j,2}^2z_j+q_3z_{K+1}-\sum_{j=1}^{K+1}\psi_j\log(z_j/\psi_j)-\xi\Big(\sum_{j=1}^{K+1}z_j\Big)-\sum_{j=1}^{K+1}\psi_j. \end{aligned} \tag{C.6}$$

This convergence, together with those of the partial derivatives of our interest, are formally established in the following proposition.

Proposition C.8. *Let $G_d(\xi;\mathbf{q},\boldsymbol{\mu})$ be defined in Definition C.3, and $g(\xi;\mathbf{q},\boldsymbol{\mu})$ be defined in equation (C.5). For any fixed $\mathbf{q}\in\mathcal{Q}$, $\xi\in\mathbb{C}_+$ and $u\in\mathbb{R}_+$,*

$$\begin{aligned} &\lim_{d\to+\infty}\mathbb{E}[|G_d(\xi;\mathbf{q},\boldsymbol{\mu})-g(\xi;\mathbf{q},\boldsymbol{\mu})|]=0, \\ &\lim_{d\to+\infty}\mathbb{E}[\|\nabla_{\mathbf{q}}G_d(\mathrm{i}u;\mathbf{q},\boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}}-\nabla_{\mathbf{q}}g(\mathrm{i}u;\mathbf{q},\boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}}\|_2]=0, \end{aligned}$$

$$\lim_{d\to+\infty} \mathbb{E}[\|\nabla_{\mathbf{q}}^2 G_d(iu;\mathbf{q},\boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}} - \nabla_{\mathbf{q}}^2 g(iu;\mathbf{q},\boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}}\|_{\mathrm{op}}] = 0.$$

The proof of Proposition C.8 utilizes the key observation that $\nabla_{\mathbf{z}} L(\xi,\mathbf{z};\mathbf{q},\boldsymbol{\mu})|_{\mathbf{z}=\mathbf{m}} \equiv \mathbf{0}$. We omit the details here since it is similar to the proof of Proposition A.7.

C.4 Step 4: complete the proof

Similar to the previous proof of Theorem 3.6, we give the following proposition to ensure the existence and uniqueness of $\boldsymbol{\nu}$ defined in Section 5.

Proposition C.9. *There exists a unique analytic function $\boldsymbol{\nu} = [\nu_1,\ldots,\nu_{K+1}]^\top : \mathbb{C}_+ \to \mathbb{C}_+^{K+1}$ such that:*

1. *For any $\xi \in \mathbb{C}_+$, $\boldsymbol{\nu}(\xi;\boldsymbol{\mu})$ is a solution to $\boldsymbol{\nu}$-system (5.2).*
2. *There exists $\xi_0 > 0$, such that $|\nu_j(\xi;\boldsymbol{\mu})| \le 2\psi_j/\xi_0$, for all ξ with $\Im(\xi) \ge \xi_0$ and $j = 1,\ldots,K+1$.*

Moreover, it holds that $\boldsymbol{\nu}(\xi;\boldsymbol{\mu}) = \mathbf{m}(\xi;\mathbf{0},\boldsymbol{\mu})$ for all $\xi \in \mathbb{C}_+$.

Proof [Proof of Proposition C.9] By Proposition C.7, the existence is directly verified as $\mathbf{m}(\xi;\mathbf{0},\boldsymbol{\mu})$ is a solution. For the uniqueness of $\boldsymbol{\nu}$, note that $\boldsymbol{\nu}(\xi;\boldsymbol{\mu})$ and $\mathbf{m}(\xi;\mathbf{0},\boldsymbol{\mu})$ are analytic. By Lemma C.6, they are identical on $\{\xi : \Im(\xi) > \xi_0\}$ with some sufficiently large ξ_0. The uniqueness of $\boldsymbol{\nu}$ thus results from the uniqueness of the analytic continuation. ■

Proposition C.9 justifies the definition of $\boldsymbol{\nu}(\xi;\boldsymbol{\mu})$ in Section 5 by demonstrating its existence and uniqueness. Moreover, it also relates $\boldsymbol{\nu}(\xi;\boldsymbol{\mu})$ to the function $\mathbf{m}(\xi;\mathbf{q},\boldsymbol{\mu})$ introduced in step 3 of the proof. With this result, we can finalize the proof of Theorem 5.6 as follows.

Proof [Proof of Theorem 5.6] Let

$$\begin{aligned}\mathcal{R}(\lambda,\boldsymbol{\psi},\boldsymbol{\mu},F_1,\tau) = F_1^2 \cdot \big[1 - \partial_{q_1} g(\xi^*;\mathbf{q},\boldsymbol{\mu}) - \partial_{q_4,q_5}^2 g(\xi^*;\mathbf{q},\boldsymbol{\mu}) - \partial_{q_2,q_5}^2 g(\xi^*;\mathbf{q},\boldsymbol{\mu})\big]\big|_{\mathbf{q}=\mathbf{0}} \\ - \tau^2 \cdot \big[\partial_{q_3,q_4}^2 g(\xi^*;\mathbf{q},\boldsymbol{\mu}) + \partial_{q_2,q_3}^2 g(\xi^*;\mathbf{q},\boldsymbol{\mu})\big]\big|_{\mathbf{q}=\mathbf{0}},\end{aligned} \tag{C.7}$$

where g is defined in (C.5). Then by Propositions C.2, C.4 and C.8, we have

$$\mathbb{E}_{\mathbf{X},\boldsymbol{\Theta},\boldsymbol{\varepsilon}}\Big|R_d(\mathbf{X},\boldsymbol{\Theta},\lambda,\boldsymbol{\beta}_d,\boldsymbol{\varepsilon}) - \mathcal{R}(\lambda,\boldsymbol{\psi},\boldsymbol{\mu},F_1,\tau)\Big| = o_d(1).$$

Recall equations (C.5) and (C.6), for any $\xi \in \mathbb{C}_+$ we have

$$\nabla_{\mathbf{z}} L(\xi,\mathbf{z};\mathbf{q},\boldsymbol{\mu})|_{\mathbf{z}=\mathbf{m}} = \mathbf{0}.$$

Here $\mathbf{z} = [z_1,\ldots,z_{K+1}]^\top$. Then from the formula for implicit differentiation, we have

$$\partial_{q_1} g(\xi^*;\mathbf{q},\boldsymbol{\mu})|_{\mathbf{q}=\mathbf{0}} = \partial_{q_1} L(\xi^*,\mathbf{z};\mathbf{q},\boldsymbol{\mu})|_{\mathbf{z}=\boldsymbol{\nu}^*,\mathbf{q}=\mathbf{0}} = \frac{2\nu_{K+1}^* M_N}{M_D}. \tag{C.8}$$

We remind readers that $M_N = \sum_{j=1}^{K} \nu_j^* \mu_{j,1}^2$, $M_D = \nu_{K+1}^* M_N - 1$ and $\boldsymbol{\nu}^* = \mathbf{m}(\xi^*;\mathbf{0},\boldsymbol{\mu})$. Denote $\mathbf{u} = (q_2,q_3,q_4,q_5,\mathbf{z})$, and construct the matrix $\mathbf{W}(\boldsymbol{\nu}^*,\boldsymbol{\mu}) = \nabla_{\mathbf{u}}^2 L(\xi^*,\mathbf{z};\mathbf{q},\boldsymbol{\mu})|_{\mathbf{z}=\boldsymbol{\nu}^*,\mathbf{q}=\mathbf{0}}$. Note

that (A.8) and (A.9) in our proof of DRFMs still hold for the case of MRFMs. Therefore we have (to simplify the writing, we drop the arguments in the matrix $\mathbf{W}$):

$$\frac{\partial^2 g(\xi^*;\mathbf{q},\boldsymbol{\mu})}{\partial q_2 \partial q_5}\Big|_{\mathbf{q}=0} = \mathbf{W}_{1,4} - \mathbf{W}_{1,[5:(K+5)]}\Big(\mathbf{W}_{[5:(K+5)],[5:(K+5)]}\Big)^{-1}\mathbf{W}_{[5:(K+5)],4}, \tag{C.9}$$

$$\frac{\partial^2 g(\xi^*;\mathbf{q},\boldsymbol{\mu})}{\partial q_3 \partial q_4}\Big|_{\mathbf{q}=0} = \mathbf{W}_{2,3} - \mathbf{W}_{2,[5:(K+5)]}\Big(\mathbf{W}_{[5:(K+5)],[5:(K+5)]}\Big)^{-1}\mathbf{W}_{[5:(K+5)],3}, \tag{C.10}$$

$$\frac{\partial^2 g(\xi^*;\mathbf{q},\boldsymbol{\mu})}{\partial q_2 \partial q_3}\Big|_{\mathbf{q}=0} = \mathbf{W}_{1,2} - \mathbf{W}_{1,[5:(K+5)]}\Big(\mathbf{W}_{[5:(K+5)],[5:(K+5)]}\Big)^{-1}\mathbf{W}_{[5:(K+5)],2}, \tag{C.11}$$

$$\frac{\partial^2 g(\xi^*;\mathbf{q},\boldsymbol{\mu})}{\partial q_4 \partial q_5}\Big|_{\mathbf{q}=0} = \mathbf{W}_{3,4} - \mathbf{W}_{3,[5:(K+5)]}\Big(\mathbf{W}_{[5:(K+5)],[5:(K+5)]}\Big)^{-1}\mathbf{W}_{[5:(K+5)],4}. \tag{C.12}$$

Similar to the case of DRFMs, we have

$$\mathbf{W}_{1,4} = \mathbf{W}_{2,3} = \mathbf{W}_{1,2} = 0, \quad \mathbf{W}_{3,4} = -\frac{\nu_{K+1}^{*2} M_N^2}{M_D^2},$$
$$\mathbf{V} = \mathbf{W}_{[5:(K+5)],[1:4]} = \mathbf{W}_{[1:4],[5:(K+5)]}^\top, \quad \text{and} \quad \mathbf{H} = \Big(\mathbf{W}_{[5:(K+5)],[5:(K+5)]}\Big).$$

Plugging (C.8) and (C.9)-(C.12) into (C.7) proves Theorem 5.6. ■

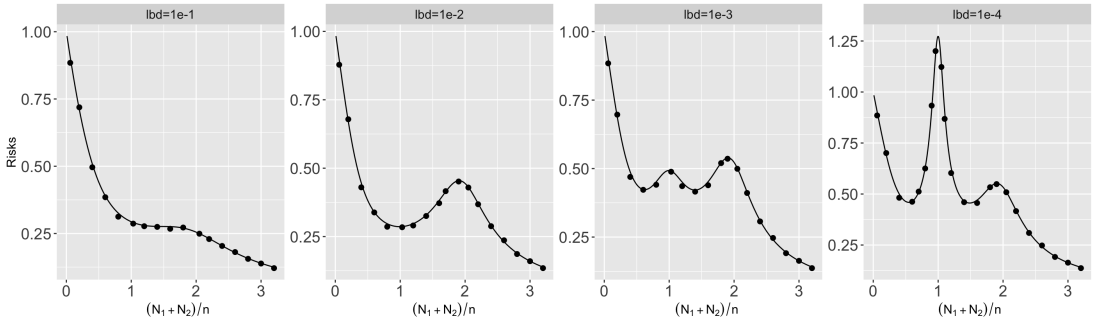
Appendix D. Other key factors affecting the risk curve

Here we investigate several other factors that affect the shape of the risk curve. By studying how these factors affect the risk, we aim to provide a clearer understanding of Proposition 4.1, Proposition 4.2 and the triple descent phenomena. Our analysis also shows how we can design DRFMs to achieve a specific risk curve shape. Unlike Chen et al. (2021) which requires designing a specific data distribution, our study shows that various risk curves can be achieved by different random feature models on a fixed data distribution.

The regularization parameter λ. We investigate how the regularization parameter λ affect the shape of the risk curve. We again use the same experiment setup as in Section 4.2, expect that we focus on activation functions ELU$(3x)$ and ReLU$(x/4)$, and calculate the risk curves w.r.t. different regularization parameters $\lambda = 10^{-1}, 10^{-2}, 10^{-3}$ and 10^{-4}.

The results are given in Figure 8. Note that Proposition 4.1 holds under the condition λ tends to 0. When the regularization parameter λ is large, the risk decreases with the model complexity parameter $c \sim (N_1+N_2)/n$. As λ decreases, the peak at $c = 2$ first appears, and then the peak at $c = 1$ also appears when $\lambda = 10^{-3}$. Finally when $\lambda = 10^{-4}$, the risk around $c = 1$ becomes very high. From these experiments, we can conclude that (i) Double/triple descent happens particularly when there is no regularization or when the regularization is very weak. (ii) the risk value of the first peak around $c = 1$ is more sensitive to λ then that of the second peak.

Signal-to-noise ratio. We also study how the signal-to-noise ratio (SNR) in the data, which we define as $\|\boldsymbol{\beta}_1\|_2/\tau$, affects the shape of the risk curve. We again use the same experimental setup as in Section 4.2, except that (i) we focus on activation functions (ELU$(3x)$

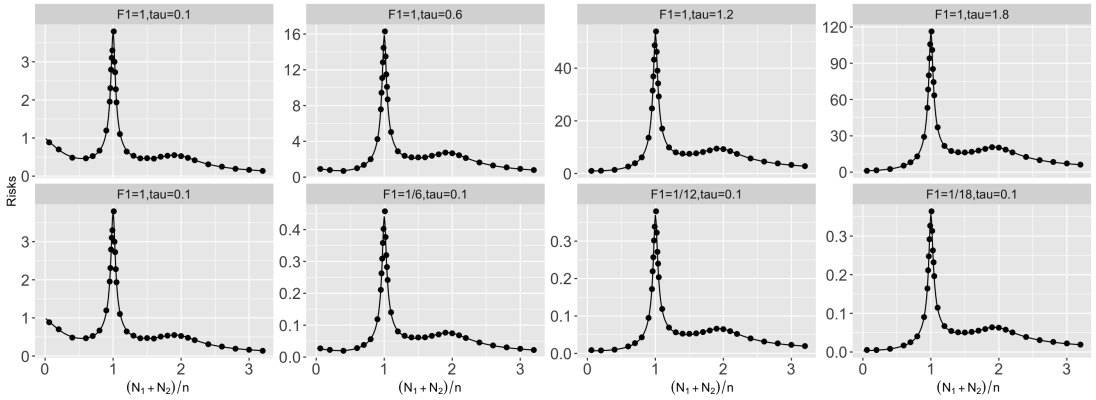


(a) (b) (c) (d)

Figure 8: Risk curves of DRFMs trained with different regularization parameters. The plots show both the asymptotic excess risks (curves) and empirical excess risks (dots). From (a) to (d), we set $\lambda = 10^{-1}, 10^{-2}, 10^{-3}$ and 10^{-4}, respectively. The activation functions are chosen as $\sigma_1(x) = \mathrm{ELU}(3x)$ and $\sigma_2(x) = \mathrm{ReLU}(x/4)$ in all these experiments.

and $\mathrm{ReLU}(x/4)$), and (ii) we perform experiments with different values of $\|\boldsymbol{\beta}_1\|_2 = F_1$ and τ.

The results are given in Figure 9. We first see that the risk curves in each column have the same shapes. This matches our theoretical result that the risk has the form $R = \tau^2(a \cdot \mathrm{SNR} + b)$ for some positive functions a, b depending on the other parameters. Moreover, the SNR has a particularly high impact on the trend of the risks in the under-parameterized regime ($(N_1 + N_2)/n < 1$) and the highly over-parameterized regime ($(N_1 + N_2)/n > 2$, shown in Proposition 4.2). Specifically, in column (a) when the SNR is large, we can see that the lowest risk is achieved in the highly over-parameterized regime; on the other hand, in columns (c) and (d) when the SNR is relatively small, the lowest risk is achieved in the under-parameterized regime.



(a) (b) (c) (d)

Figure 9: Risk curves of DRFMs under different SNRs. The plots show both the asymptotic excess risks (curves) and empirical excess risks (dots). In the top row, we set $\|\boldsymbol{\beta}_1\|_2 = 1$ and $\tau = 0.1, 0.6, 1.2$ and 1.8 (from (a) to (d)). In the bottom row, we set $\tau = 0.1$ and $\|\boldsymbol{\beta}_1\|_2 = 1, 1/6, 1/12$ and $1/18$ (from (a) to (d)). The parameter values are chosen such that the two figures in each column have the same SNR.

References

Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. *Advances In Neural Information Processing Systems*, 33: 11022–11032, 2020a.

Ben Adlam and Jeffrey Pennington. The neural tangent kernel in high dimensions: triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, volume 119, pages 74–84, 2020b.

Ben Adlam, Jake A Levinson, and Jeffrey Pennington. A random matrix perspective on mixtures of nonlinearities in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 3434–3457. PMLR, 2022.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

Yuan Cao, Quanquan Gu, and Mikhail Belkin. Risk bounds for over-parameterized maximum margin classification on sub-gaussian mixtures. *Advances in Neural Information Processing Systems*, 34:8407–8418, 2021.

Yuan Cao, Zixiang Chen, Misha Belkin, and Quanquan Gu. Benign overfitting in two-layer convolutional neural networks. *Advances in neural information processing systems*, 35: 25237–25250, 2022.

Niladri S Chatterji and Philip M Long. Finite-sample analysis of interpolating linear classifiers in the overparameterized regime. *The Journal of Machine Learning Research*, 22 (1):5721–5750, 2021.

Lin Chen, Yifei Min, Mikhail Belkin, and Amin Karbasi. Multiple descent: design your own generalization curve. *Advances in Neural Information Processing Systems*, 34, 2021.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters: Double/debiased machine learning. *The Econometrics Journal*, 21(1), 2018.

Stéphane d'Ascoli, Levent Sagun, and Giulio Biroli. Triple descent and the two kinds of overfitting: where & why do they appear? *Advances in Neural Information Processing Systems*, 33:3058–3069, 2020.

Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 11:435–495, 2021.

Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA*, 11(2):435–495, 2022.

Marco Di Marzio, Agnese Panzera, and Charles C Taylor. Nonparametric regression for spherical data. *Journal of the American Statistical Association*, 109(506):748–763, 2014.

Thomas G Dietterich. Ensemble methods in machine learning. In *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings 1*, pages 1–15. Springer, 2000.

Stéphane d'Ascoli, Maria Refinetti, Giulio Biroli, and Florent Krzakala. Double trouble in double descent: Bias and variance (s) in the lazy regime. In *International Conference on Machine Learning*, pages 2280–2290. PMLR, 2020.

Melikasadat Emami, Mojtaba Sahraee-Ardakan, Parthe Pandit, Sundeep Rangan, and Alyson Fletcher. Generalization error of generalized linear models in high dimensions. In *International Conference on Machine Learning*, pages 2892–2901. PMLR, 2020.

Spencer Frei, Niladri S Chatterji, and Peter Bartlett. Benign overfitting without linearity: Neural network classifiers trained by gradient descent for noisy linear data. In *Conference on Learning Theory*, pages 2668–2703. PMLR, 2022.

Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stéphane d'Ascoli, Giulio Biroli, Clément Hongler, and Matthieu Wyart. Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401, 2020.

Onur C Hamsici and Aleix M Martinez. Spherical-homoscedastic distributions: the equivalency of spherical and normal distributions in classification. *Journal of Machine Learning Research*, 8(7), 2007.

Lars Kai Hansen and Peter Salamon. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence*, 12(10):993–1001, 1990.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2): 949–986, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. *Advances in neural information processing systems*, 7, 1994.

Yue Li and Yuting Wei. Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent, 2021.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *Conference on Learning Theory*, pages 2683–2711. PMLR, 2020.

Zhenyu Liao, Romain Couillet, and Michael Mahoney. A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent. In *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020.

Fanghui Liu, Zhenyu Liao, and Johan Suykens. Kernel regression in high dimensions: Refined analysis beyond double descent. In *International Conference on Artificial Intelligence and Statistics*, pages 649–657. PMLR, 2021.

Ziqi Liu. Multi-scale deep neural network (MscaleDNN) for solving poisson-boltzmann equation in complex domains. *Communications in Computational Physics*, 28(5):1970–2001, 2020.

Xiaoyi Mai, Zhenyu Liao, and Romain Couillet. A large scale analysis of logistic regression: Asymptotic performance and new insights. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3357–3361. IEEE, 2019.

Domenico Marinucci and Giovanni Peccati. *Random Fields on the Sphere: Representation, Limit Theorems and Cosmological Applications*. Cambridge University Press, 2011.

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*, 75(4):667–766, 2022.

Gabriel Mel and Surya Ganguli. A theory of high dimensional regression with arbitrary correlations between input features and target functions: sample complexity, multiple descent curves and a hierarchy of phase transitions. In *International Conference on Machine Learning*, pages 7578–7587. PMLR, 2021.

Xuran Meng, Jianfeng Yao, and Yuan Cao. Online supplementary material to "multiple descent in the multiple random feature model". URL `https://github.com/XuranMeng/Multipledescent/blob/main/onlinesupplementary.pdf`.

Xuran Meng, Difan Zou, and Yuan Cao. Benign overfitting in two-layer relu convolutional neural networks for xor data. *arXiv preprint arXiv:2310.01975*, 2023.

Theodor Misiakiewicz. Spectrum of inner-product kernel matrices in the polynomial regime and multiple descent phenomenon in kernel ridge regression. *arXiv preprint arXiv:2204.10425*, 2022.

Andrea Montanari and Yiqiao Zhong. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *The Annals of Statistics*, 50(5): 2816–2847, 2022.

Preetum Nakkiran, Prayaag Venkat, Sham M Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. In *International Conference on Learning Representations*, 2020.

Pratik Patil, Arun Kumar Kuchibhotla, Yuting Wei, and Alessandro Rinaldo. Mitigating multiple descents: A model-agnostic framework for risk monotonization, 2022.

Alexander Tsigler and Peter L Bartlett. Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123):1–76, 2023.

Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33, 2020.

Lechao Xiao, Hong Hu, Theodor Misiakiewicz, Yue M Lu, and Jeffrey Pennington. Precise learning curves and higher-order scaling limits for dot product kernel regression. In *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

Tianqi Zhao, Guang Cheng, and Han Liu. A partially linear framework for massive heterogeneous data. *Annals of statistics*, 44(4):1400, 2016.