A Random Matrix Theory Perspective on the Spectrum of Learned Features and Asymptotic Generalization Capabilities

Yatin Dandi
SPOC Laboratory, EPFL,
IdePHICS Laboratory, EPFL,
Lausanne, Switzerland

Luca Pesce
IdePHICS Laboratory, EPFL,
Lausanne, Switzerland

Hugo Cui
CMSA, Harvard University,
SPOC Laboratory, EPFL,
Lausanne, Switzerland

Florent Krzakala
IdePHICS Laboratory, EPFL,
Lausanne, Switzerland

Yue M.Lu
John A. Paulson School
of Engineering and Applied Sciences,
Harvard University

Bruno Loureiro
Département d'Informatique,
École Normale Supérieure (ENS),
PSL & CNRS, F-75230 Paris cedex 05,
France

Abstract

A key property of neural networks is their capacity to adapt to data during training. Yet, our current mathematical understanding of feature learning and its relationship to generalization remains limited. In this work, we provide a random matrix analysis of how fully-connected two-layer neural networks adapt to the target function after a single, but aggressive, gradient descent step. We rigorously establish the equivalence between the updated features and an isotropic spiked random feature model, in the limit of large batch size. For the latter model, we derive a *deterministic equivalent* description of the feature empirical covariance matrix in terms of certain low-dimensional operators. This allows us to sharply characterize the impact of training on the asymptotic feature spectrum, and in particular, provides a theoretical grounding for how the tails of the feature spectrum modify with training. The deterministic equivalent further yields the exact asymptotic generalization error, shedding light on the mechanisms behind its improvement in the presence of feature learning. Our result goes beyond standard random matrix ensembles, and therefore we believe it is of independent technical interest. Different from previous work, our result holds in the challenging maximal learning rate regime, is fully rigorous and allows for finitely supported second layer initialization, which turns out to be crucial for studying the functional expressivity of the learned features. This provides a sharp description of the impact of feature learning on the generalization of two-layer neural networks, beyond the random features and lazy training regimes.

1 INTRODUCTION

An essential property of neural networks is their capacity to extract relevant low-dimensional features from high-dimensional data. This *feature learning* is usually signaled by an array of telltale phenomena — such as the improvement of the test error over non-adaptive methods [Bach, 2021], or the lengthening of the tails in the spectra of the network weights and activations [Martin and Mahoney, 2021, Martin et al., 2021, Wang et al., 2024a]. Yet, a precise theoretical characterization of the learned features, and how they translate into the aforementioned generalization and spectral properties, is still largely lacking, and arguably constitutes one of the key open questions in machine learning theory.

In this work, we provide a rigorous answer to these questions for two layer neural networks trained with a single but large gradient step. More precisely,

- we provide an exact characterization of statistics associated to the learned features, and in particular the achieved test error;

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

- we quantitatively characterize how feature learning results in modified tails in the spectrum of the feature covariance matrix, as illustrated in Fig. 1.

Our results provide a sharp mathematical description of feature learning in this context and allow us to explore the interplay between representational learning and generalization. Before exposing our main technical results, we first offer an overview of known results on feature learning (or the lack thereof) in two-layer neural networks, so as to put our work in context.

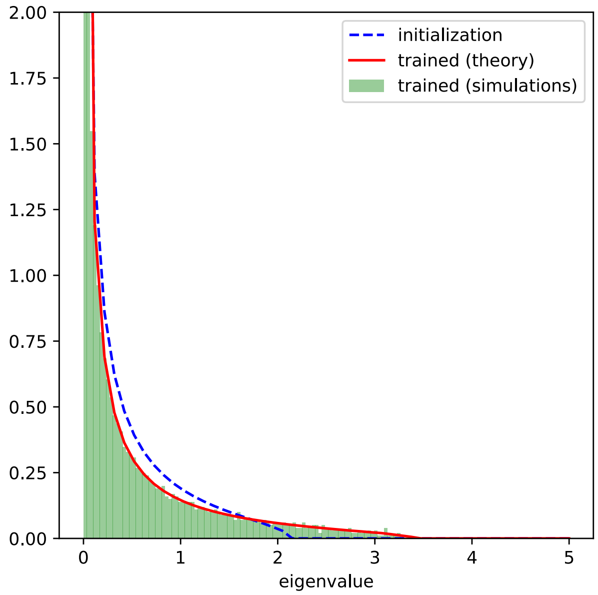

Figure 1: **Bulk spectrum of the empirical features covariance** at initialization (dashed blue) and after training (green); the red line corresponds to the theoretical characterization derived in this manuscript.

Models — The present manuscript addresses the simplest class of neural network architectures (used in Fig. 1), namely fully-connected, shallow two-layer neural networks:

$$f(x; W, a) = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} a_j \sigma(w_j^\top x), \tag{1}$$

where $W = \{w_j, j \in [p]\} \in \mathbb{R}^{p\times d}$ and $a = \{a_j, j \in [p]\} \in \mathbb{R}^p$ denote the first and second layer weights, respectively, and σ is an activation function. Motivated by the lazy regime of large-width networks [Jacot et al., 2018, Chizat et al., 2019], the generalization properties of two-layer neural networks have been thoroughly investigated in the simple case where only the second-layer weights a are trained (typically by ridge regression), while the first-layer weights $W = W^0$ are fixed at (typically random) initialization. This model, which is equivalent to the *Random Features* (RF) approximation of kernel methods introduced by Rahimi and Recht [2007], is particularly amenable to mathematical treatment. The reason is that, besides being a convex problem, in the asymptotic regime where the number of samples n, width p and dimension d scale as $n, p = \Theta(d)$ with $d \to \infty$, the random feature map $\varphi(x) = \sigma(w^\top x)$ statistically behaves as a *linear* function with additive noise — a result often referred to as the Gaussian Equivalence Principle (GEP) [Goldt et al., 2022, Hu and Lu, 2022, Mei and Montanari, 2022]. While this surprising property makes the problem tractable with random matrix theory arguments, it implies that in this regime random features can learn, at best, a linear approximation of the underlying target function. This sets a benchmark for the fundamental limitation of not adapting the first-layer weights to the data.

Gradient descent — Going (literally) one step beyond random features, Ba et al. [2022] showed that training the first layer weights with a single Gradient Descent (GD) step on a batch of data $\{(x_\mu, y_\mu) : \mu \in [n_0]\}$ from the same target distribution:

$$w_j^1 = w_j^0 - \eta \nabla_{w_j} \frac{1}{2n_0} \sum_{\mu \in [n_0]} \left(y_\mu - f(x_\mu; a^0, w^0)\right)^2 \tag{2}$$

followed by ridge regression on the second layer weights a with n fresh samples can drastically change the story above, depending on the scaling of the learning rate η where $n \propto d, n_0 \propto d$. More precisely, they showed that for $\eta = \Theta_d(\sqrt{d})$, Gaussian Equivalence asymptotically holds as $d \to \infty$, implying that an exact asymptotic treatment based on the GEP still holds. However, in the regime of an aggressive learning rate $\eta = \Theta_d(d)$ (the so-called Maximal Update parametrization [Yang et al., 2022]) they showed that the first-layer weights adapt to the data distribution, translating into an improvement over the RF lower-bound.

This finding has sparked considerable interest in exactly characterizing the asymptotic generalization error achieved under this single, large step setting. Dandi et al. [2023] proved a novel *Conditional Gaussian Equivalence Principle* (cGEP), and derived more general lower-bounds for the performance of two-layer neural networks after the gradient step. However, these bounds do not provide a fine-grained description of what is learned in the feature learning regime. Reaching such a sharp description instead requires addressing the challenging Random Matrix Theory (RMT) problem of characterizing the non-linear transformation of a highly structured random matrix $\sigma\left(W - \eta G\right)$ with G denoting the gradient matrix, in the regime $\Theta_d(\|\eta G\|_F) = \Theta_d(||W||_F)$. Moniri et al. [2023] provided the first result in this direction for the intermediate learning rate regime $\eta = \Theta_d(d^{1/2+\zeta})$, with $\zeta \in (0, 1/2)$, proving a polynomial GEP together with an exact RMT analysis of the asymptotic generalization error. Their findings, however, do not hold in the more challenging Maximal Update regime $\eta = \Theta(d)$. Leveraging the results on the low-rank approximation of the Gradient matrix [Ba et al., 2022] and the cGEP from [Dandi et al., 2023], Cui et al. [2024] studied the latter regime approximating the two-layer network by a Spiked Random Features Model (SRFM) with the non-rigorous replica method [Mezard et al., 1986]. Though such heuristic arguments are inspirational, they often lack interpretability when compared to other approaches.

Summary of main results — In this work, we provide a rigorous RMT treatment of feature learning after a single GD step in the challenging Maximal Update step size regime. Beyond proving the conjectured results from Cui et al. [2024], our analysis extends these findings in several key directions, enabling a quantitative exploration and deeper understanding of fundamental aspects of feature learning, particularly with respect to generalization and spectral properties. Specifically:

- We prove a *deterministic equivalent* (Equation 16 in Theorem 4.5) description for the empirical feature matrix after the gradient step. This characterization is non-asymptotic in the problem dimensions, and in particular sharply holds in the Maximal-update scaling in the proportional regime where $n, p, \eta = \Theta(d)$ and $n_0 = \Omega(d^{1+\epsilon})$ as $d \to \infty$, for some $\epsilon > 0$. This result characterizes the spectral properties as well as the generalization error upon updating the second layer.

- Our proof proceeds through multiple stages of deterministic equivalences and the asymptotic description of the high-dimensional features through low-dimensional non-linear functions.

- We derive an exact asymptotic formula for the generalization error of ridge regression on features updated via a gradient step in the proportional high-dimensional regime, where $n, p, \eta = \Theta(d)$ with $d \to \infty$. Our result offers a rigorous proof of the conjectures in Cui et al. [2024], while extending them in several directions, as it applies to finitely supported second-layer initialization and structured first-layer initialization.

- Using the deterministic equivalent, we demonstrate how the "spikes" in the weights resulting from feature learning alter the entire shape of the feature covariance spectrum and its tail behavior. This observation aligns with the empirical findings of Wang et al. [2024a] and provides a rigorous foundation for them in the context of our setting.

- Finally, we precisely characterize the effect of the variability in the second-layer initialization on the functional expressivity of the network after one GD step.

These findings provide a detailed understanding of the consequences of feature learning in our setting, helping to establish several widely accepted intuitions on a more quantitative and rigorous basis.

Further related works

Fixed feature methods – A plethora of works characterized the generalization capabilities of two-layer networks in the high-dimensional regime when the first hidden layer is not trained, with the most prominent example of such fixed feature method being kernel machines [Bordelon et al., 2020, Canatar et al., 2021, Cui et al., 2021, 2023, Dietrich et al., 1999, Donhauser et al., 2021, Ghorbani et al., 2019, 2020, Opper and Urbanczik, 2001, Xiao et al., 2022]. This class of algorithms is amenable to theoretical analysis and comes with sharp generalization guarantees. However, these methods adapt to relevant low-dimensional structures at much higher sample complexities than fully trained two-layer networks. In simple terms, the sample complexity of kernel methods is not driven by the presence (or lack thereof) of a low-dimensional target subspace. Identical considerations hold for Random Feature Models, where the number of samples n in the generalization guarantees is replaced with $\min(n, p)$, with p being the number of random features [Gerace et al., 2020, Mei and Montanari, 2022, Mei et al., 2022, Hu et al., 2024, Aguirre-López et al., 2024].

Feature learning – The discussion above portrays the limitations of fixed feature methods. Inspired by this, a large body of work has studied the sharp separation between the generalization capabilities of such methods versus fully trained two-layer networks that learn features through gradient-based training. Many of these works fall under the umbrella of the so-called mean-field regime [Chizat and Bach, 2018, Mei et al., 2018, Rotskoff and Vanden-Eijnden, 2022, Sirignano and Spiliopoulos, 2020]. The authors mapped the optimization of two-layer shallow networks onto a convex problem in the space of measures on the weights and paved the way for understanding how features are learned in the high-dimensional regime. The arguably most popular data model in the theoretical community for addressing this question is the multi-index data model with isotropic Gaussian data. This setting has attracted considerable attention in the theoretical community with numerous works that have analyzed the feature learning capabilities of shallow networks trained with gradient-based schemes [Ben Arous et al., 2021, Abbe et al., 2023, Ba et al., 2024, Bardone and Goldt, 2024, Berthier et al., 2023, Bietti et al., 2023, Damian et al., 2024, Dandi et al., 2023, Paquette et al., 2021, Veiga et al., 2022, Zweig and Bruna, 2023, Dandi et al., 2024].

Deterministic equivalents – Deterministic equivalents of large empirical covariance matrices have been extensively studied, beginning with the seminal work of Marchenko and Pastur [1967]. This was extended by Burda et al. [2004], Knowles and Yin [2017] to separable data covariances, and further by Bai and Zhou [2008], Louart and Couillet [2018], Chouard [2022] for non-separable covariances. These methodologies have enabled precise asymptotic characterizations of the learning dynamics in single-layer neural networks [Louart et al., 2018] and deep random feature models [Schröder et al., 2023, 2024, Bosch et al., 2023, Chouard, 2023]. The propagation of spiked eigenstructure in the input data through a deep random feature model was recently studied by Wang et al. [2024b].

2 NOTATIONS and SETTING

Consider a supervised learning problem with training data $\mathcal{D} = \{(x_\mu, y_\mu) \in \mathbb{R}^{d+1}, \mu \in [N]\}$. As motivated in the introduction, our goal is to study the problem of feature learning with two-layer neural networks defined in (1). A widespread intuition in the machine learning literature for why learning is possible despite the curse of dimensionality is that real data distributions typically exhibit low-dimensional latent structures [Bellman et al., 1957]. To reflect and model this intuition, we assume our training data have been independently drawn from an isotropic *Gaussian single-index model*:

$$y_\mu = f_\star(x_\mu) = g(x_\mu^\top w^\star), \qquad x_\mu \sim \mathcal{N}(0, I_d), \tag{3}$$

where the unit norm vector $w^\star$ denotes the target weights and $g : \mathbb{R} \to \mathbb{R}$ is the link function. We assume the mapping g to be determistic for simplicity, and more general stochastic mappings can be incorporated into our analysis straightforwardly. Note that in 3, the high-dimensional covariates $X = (x_\mu)_{\mu\in[n]} \in \mathbb{R}^{n\times d}$ are isotropic in $\mathbb{R}^d$, and therefore the structure in the data distribution is in the conditional distribution of the labels $y|x$, which depends on the covariates only through their projection onto a 1-dimensional subspace of $\mathbb{R}^d$. Therefore, learning features in this model translate to learning the target weight $w^\star$.

Given a batch of N samples $\mathcal{D} = \{(x_\mu, y_\mu), \mu \in [N]\}$ independently drawn from model (3), we are interested in studying how our two-layer neural network (1) learns the target feature $w^\star$ through the *Empirical Risk Minimization* (ERM) on the training data. We follow the same procedure as in [Ba et al., 2022, Dandi et al., 2023, Moniri et al., 2023, Cui et al., 2024] and consider the following two-step training procedure:

1. Let W^0 and a^0 denote the first and second layer weights at initialization. Consider a partition of the training data $\mathcal{D} = \mathcal{D}_0 \cup \mathcal{D}_1$ in two disjoint sets of size n_0 and $n := N - n_0$, respectively. First, we apply a single gradient step on the square loss for the first-layer weights, keeping the 2nd layer a^0 fixed:

$$\begin{aligned} w_j^1 &= w_j^0 - \eta g_j^0 \\ g_j^0 &= \frac{1}{n_0\sqrt{p}} \sum_{\mu\in[n_0]} \Big(f(x_\mu; W^0, a^0) - y_\mu\Big)\, a_j^0 x_\mu \sigma'({w_j^0}^\top x_\mu) \end{aligned} \tag{4}$$

   In this first *representation learning* step, the hidden layer weights adapt to the low-dimensional relevant features from the data.

2. Given the updated weights W^1, we update the second-layer weights via ridge regression on the remaining data $\mathcal{D}_1$:

$$\begin{aligned} \hat{a}_\lambda &= \underset{a\in\mathbb{R}^p}{\mathrm{argmin}} \sum_{\mu\in[n]} \Big(y_\mu - f(x_\mu; a, W^1)\Big)^2 + \lambda||a||_2^2 \\ &= \Big({}^{\Phi^\top\Phi}/_p + \lambda I_n\Big)^{-1} {}^{\Phi^\top y}/_{\sqrt{p}} \end{aligned} \tag{5}$$

   where we defined the feature matrix $\Phi \in \mathbb{R}^{n\times p}$ with elements $\phi_{\mu j} = \sigma(x_\mu^\top w_j^1)$ and the label vector $y = (y_\mu)_{\mu\in[n]}$.

In the following, we will assume the following initial conditions for the training protocol above:

Assumption 2.1 (Initialization)**.** We assume the first layer weights are initialized uniformly at random from the hypersphere $w_j^0 \sim \mathrm{Unif}(\mathbb{S}^{d-1}(1))$ and the second-layer weights read $a_j^0 = {}^{\tilde{a}_j^0}/_{\sqrt{p}}$, where the $\{\tilde{a}_j^0, j \in [p]\}$ are $O_d(1)$ scalars initialized i.i.d. by sampling from a dimension-independent vocabulary of size k, with probabilities $\pi = (\pi_q)_{q\in[k]}$.

Our goal in the following is two-fold. First, to characterize the properties of the empirical feature covariance matrix $\Phi^\top\Phi$. Second, to characterize the generalization error associated with the minimizer of equation (5), which is defined as:

$$\varepsilon_{\mathrm{gen}} = \mathbb{E}_{\mathcal{D}_1}\mathbb{E}_{y_{\mathrm{new}},\boldsymbol{x}_{\mathrm{new}}}\left[\Big(y_{\mathrm{new}} - f(\boldsymbol{x}_{\mathrm{new}}; W^1, \hat{a}_\lambda)\Big)^2\right] \tag{6}$$

where the expectation is over the joint distribution defined by the model in 3. In particular, we will focus on the proportional high-dimensional regime with Maximal Update scaling, which we formalize in the following assumption.

Assumption 2.2 (High-dimensional regime)**.** We assume that $n_0 = \Omega(d^{1+\epsilon})$ for some $\epsilon > 0$. We work under the proportional regime with Maximal Update scaling, defined as the limit where $n, p, \eta, d \to \infty$ at fixed ratios:

$$\alpha := \frac{n}{d}, \qquad \beta := \frac{p}{d}, \qquad \tilde{\eta} := \frac{\eta}{d} \tag{7}$$

Assumption 2.3 (Activation function)**.** σ is odd, uniformly Lipschitz such that σ'', σ''' exist almost surely and are bounded in absolute value by some constant C almost surely with respect to the Lebesgue measure. Furthermore, g is uniformly bounded and Lipschitz with $\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[g(z)\right]=0$ and $\mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[g'(z)\right]\neq 0$.

For non-odd σ, the gradient update may contain certain non-vanishing residual terms due to a non-zero mean and second-order Hermite coefficients (see Prop. F.2 for details). Therefore, we restricted ourselves to odd σ so that Lemma 3.1 applies and the bulk in the weight matrix is asymptotically isotropic. However, under the exact spike model $W(1)=W^{(0)}+u(w^\star)^\top$ (i.e assuming the validity of Lemma 3.1), our exact analysis continues to hold for non-odd activations. We illustrate this in Figures 1, 2 and 3 where our predictions hold for $\sigma=$ ReLu.

3 ASYMPTOTICS OF THE FIRST GRADIENT STEP

We introduce in this section our first result that establishes a rigorous framework for studying the exact asymptotics after one GD step. This question has been the subject of intense theoretical scrutiny in recent years. First, Ba et al. [2022] proved that in the high-dimensional regime, specified in Assumption 2.2, the hidden layer weights after one GD step are approximately low rank:

$$W^1=W^0+uv^\top+\Delta \tag{8}$$

where the spiked structure is identified by: i) $u=\eta c_1 c_1^\star a^0/\sqrt{p}$ is proportional to the second layer at initialization a^0, the learning rate η, and the first Hermite coefficients of the network activation σ and the target activation g, i.e, $c_1=\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[\sigma(\xi)\xi\right]$, $c_1^\star=\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}\left[g(\xi)\xi\right]$; ii) v lies along the first Hermite coefficient of the target $f_\star(\cdot)$, i.e., $v\propto\sum_{\mu\in[n]}y_\mu x_\mu$. The spike component v is correlated with the target vector $w^\star$ in eq. (3). It is precisely the presence of such correlated components that enables the trained model to surpass the Random Features performance [Damian et al., 2022, Abbe et al., 2022, 2023, Dandi et al., 2023]. The "noise" term Δ, arises from higher-order components of the activations (see details in F.2)
The decomposition in eq. (8) underlies the analysis of Ba et al. [2022], Moniri et al. [2023], Cui et al. [2024] in the propotional regime $p\propto d, p\propto n$. Heuristically, one could hope to analyze the trained weights by mapping the problem to a Spiked Random Feature Model (SRFM), where the weights F in a feature map $\sigma(Fx)$ — can be decomposed as the sum of a random bulk (F_0) and a spike:

$$F=F_0+uv^\top \tag{9}$$

Along these lines, Cui et al. [2024] approximate the noise term Δ in eq. (8) as an isotropic Gaussian matrix to reach an asymptotic description of the equivalent SRFM model using non-rigorous tools from Statistical Physics. It however remained unclear whether the uniform/Gaussian isotropic description of $W^0+\Delta$ in eq. (8) is an accurate approximation of the bulk in the actual GD step. We offer a rigorous answer to the validity of the approximation.

Anisotropic bulk covariance – We provide the asymptotic description of the covariance for the bulk weights in eq. (8) after one GD step. For $n_0/d=\Theta(1)$, we unveil the presence of *anisotropic* components that contrast with the uniform approximation considered in Cui et al. [2024], which corresponds to taking a diagonal approximation of the covariance. We refer to Appendix F for formal results and additional investigations.

Diagonal approximation regime – On the other hand, we provably show that the diagonal approximation considered in Cui et al. [2024] for the covariance of the bulk weights is valid as soon as the number of samples n_0 used in the GD step is sufficiently large ($n_0=\Theta(d^{1+\epsilon})$ for any $\epsilon>0$). In this regime, the spike v in Equation 8 can be further replaced by the signal w^*:

Lemma 3.1. *Let $W^{(1)}\in\mathbb{R}^{p\times d}$ denote the weight matrix after the first gradient step. Then, under Assumptions 2.1, 2.2, and 2.3:*

$$\left\|W^{(1)}-\left(W^{(0)}+u(w^\star)^\top\right)\right\|_2\xrightarrow[d\to\infty]{a.s}0. \tag{10}$$

where $w^\star$ is the target vector in eq. (3), and $u=\eta c_1c_1^\star a^0/\sqrt{p}$ as defined in eq. (8), with η being the learning rate.

From eq. (8), we see that the finite support assumption for the second layer (2.1) translates to finite support of the entries u_i and we denote with $A_u=\{\zeta_1^u,\dots,\zeta_k^u\}$ its vocabulary with the corresponding probabilities $\pi=(\pi_q)_{q\in[k]}$. The above Lemma rigorously characterizes the regimes in which the isotropic SRFM approximation of Cui et al. [2024] is justified and correctly describes the network after one GD step.

4 MAIN RESULTS

We are now in the position to state our main technical results. Namely, a rigorous characterization of the empirical feature matrix after one gradient step through a *deterministic equivalent* description. This picture enables the characterization of the features spectrum and the resulting asymptotic generalization error.

Extended features and resolvent – As is well established in random matrix theory, the resolvent $G(z)=(\frac{1}{p}\Phi^\top\Phi-zI)^{-1}$, where $\Phi=\sigma(X(W^1)^\top)$, allows the extraction of a large class of summary statistics related to the spectrum of Φ [Bai and Zhou, 2008, Anderson et al., 2010]. To additionally characterize the generalization error, and capture the mean dependence of $\{\phi_\mu,\mu\in[n]\}$ on the spike components $\kappa_\mu \overset{\text{def}}{=} x_\mu^\top w^\star$, we construct an augmented version of $G(z)$. We find that the relevant statistics in our setup are captured by the resolvent of certain *extended* features that we introduce below:

Definition 4.1 (Extended resolvent)**.** Let (X, y) denote a batch of data drawn from the Gaussian single-index model in Eq. (3), and consider the feature matrix $\Phi = \sigma(X(W^1)^\top)$ after the first gradient step (Eq. 4), with $\kappa_\mu = x_\mu^\top w^\star$. Let s_q denote the subset of coordinates such that $u_j = \zeta_q^u$ and the "mean" $\bar{\phi}_\mu^q = \frac{1}{|s_q|}\sum_{j \in s_q} \Phi_{\mu,j}$. We define the *extended features* $\phi_\mu^e \in \mathbb{R}^{(p+k+1)}$ and the *extended resolvent* $G_e(z) \in \mathbb{R}^{(p+k+1)\times(p+k+1)}$ for $z \in \mathbb{C}/\mathbb{R}^+$ as:

$$\phi_\mu^e = \begin{pmatrix} y_\mu \\ \bar{\phi}_\mu \\ \tilde{\phi}_\mu \end{pmatrix}, \; G_e(z) \overset{\text{def}}{=} \left(\frac{(\Phi^e)^\top \Phi^e}{p} - zI \right)^{-1} \tag{11}$$

where $\tilde{\phi}_{\mu,j} = \phi_{\mu,j} - \bar{\phi}_j$, $\Phi^e = \{\phi_\mu^e \in \mathbb{R}^{p+k+1}, \mu \in [n]\}$

A few comments about the definition of the extended features ϕ_μ^e and the resolvent $G_e(z)$ in Definition 4.1 are in order. First, due to the extensive spike and the finite support over u by Assumption 2.1, each subset s_q possesses a non-zero mean $\bar{\phi}_\mu^q = \frac{1}{|s_q|}\sum_{j \in s_q} \Phi_{\mu j}$, asymptotically converging to $c_0(\kappa_\mu, \zeta_q^u)$ defined below.

Definition 4.2 (Shifted Hermite coefficient)**.** We define the shifted Hermite coefficient $c_\ell(\kappa, \zeta)$ of the activation $\sigma(\cdot)$

$$c_\ell(\kappa, \zeta) = \mathbb{E}_{z \sim \mathcal{N}(0,1)}[\sigma(z + \kappa\zeta) h_\ell(z)] \tag{12}$$

where $(h_\ell)_{\ell > 0}$ denote the Hermite polynomials.

Therefore, unlike typical random-matrix ensembles, the means $\bar{\phi}_\mu^q$ have $\mathcal{O}(1)$ fluctuations due to a large dependence on κ. Note that the means $\bar{\phi}_\mu^q$ can be equivalently described as the projections of ϕ_μ along the directions $(e^1, \cdots, e^k)$ defined as:

$$e_j^q = \frac{1}{\sqrt{p}} \begin{cases} 1 & \text{if } u_j = \zeta_q^u \\ 0 & \text{otherwise} \end{cases}, \qquad j \in [p], \quad q \in [k]. \tag{13}$$

By decomposing ϕ_μ as $\sum_{q=1}^k \bar{\phi}_\mu^q e^q + \tilde{\phi}_\mu$, we realize that the first term varies only along a k-dimensional subspace with variations governed by κ_μ, while the second term contributes to the "bulk" of the feature covariance. The surrogate form ϕ_μ^e splits the features into $\bar{\phi}_\mu, \tilde{\phi}_\mu$ precisely to account for these different scales of fluctuations.

Deterministic Equivalent – The extended resolvent $G_e(z)$ is a high-dimensional random matrix, inheriting the randomness from the training data (X, y) and the initialization weights W^0. To reach the deterministic equivalent $\mathcal{G}_e$ for the above extended resolvent $G_e(z)$, our proof proceeds by subsequently addressing and removing the randomness over the data X, and the weights W^0, eventually obtaining an equivalent dependent only on the coefficients u_i and the projections of W^0 on $w^\star$ denoted as $\theta := W^0 w^\star \in \mathbb{R}^p$. This "special" dependence on u_i and θ_i is expected, since by Lemma 3.1, these determine the component along the spike in the updated matrix W^1, while the remaining directions in the weights maintain isotropic dependence and are averaged out. The resulting description is characterized through low-dimensional kernels and functions. Concretely, we show that for a large class of functions $\mathcal{F}$ associated to the feature matrix covariance $\Phi^\top\Phi$ and labels $y \in \mathbb{R}^n$, with entries $\{y_\mu\}_{\mu=1}^n$, $\mathcal{F}(\Phi^\top\Phi, y) \xrightarrow{a.s} \mathcal{F}^\star(\theta, u)$ In contrast to the high dimensional matrices (X, W^1), $\mathcal{F}^\star(\theta, u)$ depends only on sequences of scalars (θ, u), turning $\mathcal{F}^\star(\theta, u)$ into finite-dimensional expectations.

As stated in Assumption 2.1, we consider a finitely supported second layer, leading to the entries of u being supported on finitely-many values $A_u = \{\zeta_1^u, \cdots, \zeta_k^u\}$ with probabilities $\pi = (\pi_q)_{q \in [k]}$. From Lemma 3.1 and equation 8, we see that neurons with identical values of u_i contain identical contributions from the spike. This leads to the deterministic equivalent $\mathcal{G}_e$ of the extended resolvent inheriting a block structure, with blocks corresponding to different values of u_i. Let $p_1, \cdots, p_k$ denote the number of neurons with u_i taking values $\zeta_1^u, \cdots, \zeta_k^u$ respectively. Then, by the strong law of large numbers $\frac{p_q}{p} \xrightarrow{a.s} \pi_q$ as $p \to \infty$. Without loss of generality, we assume that the neurons are arranged such that:

$$[u_1, \cdots, u_p] = [\zeta_1^u 1_{1\times p_1}, \zeta_2^u 1_{1\times p_2}, \dots \zeta_k^u 1_{1\times p_k}] \tag{14}$$

To compactly express this block structure, we introduce a notation for block-structured matrices and vectors:

Definition 4.3. Let $p_1, \cdots p_k$ be the sequence defined above with $\frac{p_q}{p} \xrightarrow{a.s} \pi_q$. Let $C \in \mathbb{R}^{k\times k}$ be a fixed matrix. We define the extended matrix C_e as:

$$C_e = \begin{pmatrix} C_{11}1_{p_1\times p_1}, & \cdots & \cdots & C_{1k}1_{p_1\times p_k} \\ C_{21}1_{p_2\times p_1}, & C_{22}1_{p_2\times p_2}, & \cdots & C_{2k}1_{p_2\times p_k} \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix},$$

We are now ready to state the definition of the extended deterministic equivalent:

Definition 4.4 (Deterministic equivalent)**.** Let $\mathbb{C}^+, \mathbb{C}^-$ denote the set of complex numbers with positive and negative imaginary parts respectively. Suppose that $z \in \mathbb{C}/\mathbb{R}^+$. Let $V^\star \in \mathbb{C}^{k\times k}, \nu^\star \in \mathbb{C}^k, b^\star \in \mathbb{C}^k$ be uniquely defined through the following conditions:

(i) $V^\star, \nu^\star, b^\star$ satisfy the following set of self-consistent equations:

$$V^\star_{qq'}(z) = \mathbb{E}_\kappa\left[\alpha\frac{c_1(\kappa,\zeta^u_q)c_1(\kappa,\zeta^u_{q'})}{1+\chi\left(z;\kappa\right)}\right]$$

$$\nu^\star_q(z) = \mathbb{E}_\kappa\left[\sum_{\ell\geq 2}\frac{\alpha c^2_\ell(\kappa,\zeta^u_q)}{1+\chi(z;\kappa)}\right]$$

$$b^\star_q(z) = \pi_q\beta\left(L_{q,q}(z) + (\mathrm{diag}(\nu^\star(z)) - zI_k))_{q,q}\right)^{-1},$$

where $\kappa \sim \mathcal{N}(0,1)$, $(c_\ell(\kappa,\zeta))_{\ell>0}$ are defined in 4.2 and $(\chi(z;\kappa), L(z))$ read as follows:

$$\beta\chi(z;\kappa) = \sum_{q,q'\in[k]}\psi_{qq'}c_1(\kappa,\zeta^u_q)c_1(\kappa,\zeta^u_{q'})+$$
$$+\sum_{q\in[k]}b^\star_q\sum_{\ell\geq 2}c^2_\ell(\kappa,\zeta^u_q),$$
$$L(z) = (V^\star(z))^{1/2}Q(z)(V^\star(z))^{1/2},$$

where $\psi(z), Q(z) \in \mathbb{R}^{k\times k}$ are defined as:

$$\psi(z) = b^\star(z) - L(z)\odot(b^\star(z)(b^\star(z))^\top),$$
$$Q(z) = \left(I_k + (V^\star(z))^{1/2}\,\mathrm{diag}(b^\star(z))(V^\star(z))^{1/2}\right)^{-1}$$

(ii) $V^\star, \nu^\star, b^\star$ are analytic mappings satisfying $V^\star_{i,j} : \mathbb{C}^+ \to \mathbb{C}^-$ for $i,j \in [k], \nu^\star_i : \mathbb{C}^+ \to \mathbb{C}^-$ for $i \in [k]$, $b^\star : \mathbb{C}^+_i \to \mathbb{C}^+$ for $i \in [k]$. For $z \in \mathbb{C}^+$ with imaginary part $\zeta > 0$, $\left|b^\star_i(z)\right| \leq \frac{\pi_i}{\beta\zeta}$.

We define the *deterministic equivalent extended resolvent* $\mathcal{G}_e(z) \in \mathbb{R}^{(p+1)\times(p+1)}$ as:

$$\mathcal{G}_e(z) = \begin{bmatrix} A^*_{11} - zI_k & (A^*_{21})^\top \odot \theta^\top \\ \theta\odot A^*_{21} & A^*_{22} + \alpha S^*_e \odot \theta\theta^\top\end{bmatrix}^{-1}, \tag{15}$$

where $\theta = Ww^\star \in \mathbb{R}^p$, and $S^\star_e \in \mathbb{R}^{p\times p}$, $A^*_{11} \in \mathbb{R}^{(k+1)\times(k+1)}$, $A^*_{21} \in \mathbb{R}^{p\times(k+1)}$ are defined as:

$$S^* = \mathbb{E}_\kappa\left[(\kappa^2-1)\frac{c_1(\kappa,\zeta^u_i)c_1(\kappa,\zeta^u_j)}{1+\chi(z;\kappa)}\right]$$
$$A^*_{11} = \mathbb{E}_\kappa\left[\frac{\alpha}{1+\chi(z;\kappa)}\iota\;\iota^\top\right],$$
$$A^*_{21}[j,:] = \alpha\,\mathbb{E}_\kappa\left[\frac{c_1(\kappa,u_j)}{1+\chi(z;\kappa)}\kappa\iota^\top\right], \forall j\in[p]$$
$$A^*_{22} = \left(\mathrm{diag}\left(\frac{\pi}{\beta b^\star}\right) - \mathrm{diag}\left(\frac{\pi}{\beta b^\star}+z\pi\right)\odot\frac{11^\top}{p}\right)_e,$$

where $\kappa \sim \mathcal{N}(0,1), \iota = (g(\kappa), c_0(\kappa,\zeta^u_1), \cdots c_0(\kappa,\zeta^u_k))^\top$ and the subscript e in $S^\star_e$, $A^\star_{22}$ refers to the block matrix notation in Definition 4.3.

We are now in a position to state our main result, which states that $\mathcal{G}_e(z)$ approximates $G_e(z)$ for “typical" linear functionals.

Theorem 4.5 (Deterministic equivalent)**.** *Consider the extended resolvent $G_e(z)$ (Definition 4.1) associated to a batch of training data (X,y) Eq. (3) after one gradient step. Let $\mathcal{G}_e(z)$ denote the deterministic equivalent (4.4). Then, under Assumptions 2.1, 2.2, and 2.3, with neurons arranged as Equation 14, for any $z \in \mathbb{C}/\mathbb{R}^+$ and sequence of deterministic matrices $A \in \mathbb{C}^{(p+1)\times(p+1)}$ with $\|A\|_{\mathrm{tr}} = \mathrm{tr}\Big((AA^*)^{1/2}\Big)$ uniformly bounded in d:*

$$\mathrm{Tr}(AG_e(z)) \xrightarrow[d\to\infty]{a.s} \mathrm{Tr}(A\mathcal{G}_e(z)). \tag{16}$$

The class of linear functionals A characterized above includes weighted traces as well as low-rank projections [Rubio and Mestre, 2011]. A direct consequence of Theorem 4.5 is that it yields the Stieltjes transform of the bulk covariance:

Corollary 4.6 (Stieltjes transform)**.** *Let μ_d denote the empirical spectral measure of the bulk covariance ${}^{\tilde\Phi^\top\tilde\Phi}/_p$. Let $m_d(z)$ denote the Stieltjes transform $m_d(z) = \int\frac{1}{\lambda-z}d\mu_d(\lambda)$. Let $b^\star_q(z)$ be as defined in Definition 4.4, then:*

$$m_d(z) \xrightarrow[d\to\infty]{a.s} \beta\sum_{q=1}^k b^\star_q(z). \tag{17}$$

Equipped with the above result and the Stieljtes-inversion formula we can directly characterize the support of the empirical spectral measure of the covariance matrix (see Fig. 1 for an example where to bulk appears to be bounded). Furthermore, as discussed previously, the deterministic equivalent $\mathcal{G}_e(z)$ contains all the necessary summary statistics for a full asymptotic characterization of the generalization error. This is the objective of the following theorem.

Theorem 4.7 (Generalization Error)**.** *Under the proportional asymptotics 2.2, the generalization error Eq. (6) is given by the following low-dimensional, deterministic formula:*

$$\lim_{n,d,p\to\infty}\mathbb{E}[\varepsilon_{\mathrm{gen}}] = \mathbb{E}_\kappa\left[\Lambda_\kappa(\{\tau_{0,q},\tau_{1,q}\}_{q\in[k]},\tau_2,\tau_3)\right] \tag{18}$$

where $\{\tau_{0,q},\tau_{1,q},,q\in[k]\},\tau_2,\tau_3$ are certain scalar deterministic functions of $V^\star_{qq'}(-\lambda), \nu^\star(-\lambda)$ reported in App. E along with the precise expression of the function $\Lambda_\kappa(\cdot)$.

Remark: We note that the effect of the step-size η and the scale of initialization of a^0 on the generalization error and spectrum is captured by the above equivalence through the dependence of $\mathcal{G}_e$ on $\{\zeta^u_q\}_{q\in[k]}$.

Intepretation and proof sketch – Unlike the high-dimensional interactions in the true feature covariance $\Phi^\top\Phi$, the interactions between neurons i,j in $\mathcal{G}_e(z)$ depend only on the scalars (u_i,θ). $\mathcal{G}_e(z)$ therefore reflects the structure of a non-linear low-dimensional Kernel on (u_i,θ). Note that the dimensions of the order-parameters $V^\star \in \mathbb{C}^{k\times k}, \nu^\star \in \mathbb{C}^k, b^\star \in \mathbb{C}^k$ grows with the support size k. We refer to Lemma D.14 for the analysis of the conditions under which we guarantee existence of solutions. In the continuous support limit $k \to \infty$, the self-consistent equations are rather cumbersome to solve and we expect $\{V^\star, \nu^\star, b^\star\}$ to converge to certain limiting Kernels and functions respectively, satisfying *functional* fixed point equations. The precise characterization of this regime constitutes an interesting avenue for future research where we expect an analytical ansatz to pave the way for an efficient solution.

As mentioned earlier, our proof proceeds through two stages of deterministic equivalent, successively eliminating the randomness over X and W^0 respectively. A crucial aspect of our analysis is to decouple the randomness of X, W^0 along the spike w^* and the orthogonal subspace – this is due to the fact that the dependence of the features ϕ_μ on $\kappa_\mu = x_\mu^\top w^\star$ is of a larger order than on the components of x_μ in the orthogonal space. Conditioned on κ_μ, we show that the covariance of ϕ_μ can be well-approximated through an equivalent linear model. However, due to the variability in κ, the description of the resolvent does not reduce to a standard random matrix theory ensemble. Lastly, we obtain the generalization error through the introduction of certain perturbation terms into the deterministic equivalent, with the resolvent acting as a “generating function" for additional relevant statistics. The detailed proofs are provided in the Appendices.

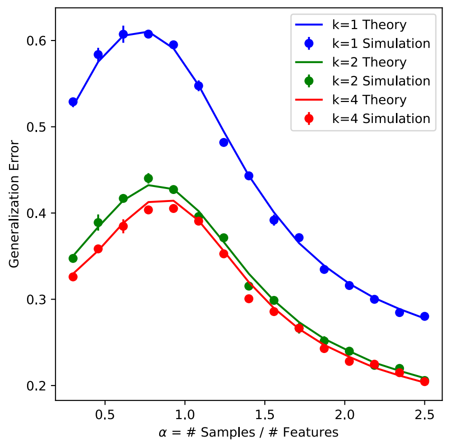

Figure 2: **Increase fitting accuracy through second layer variability:** Illustration of the benefits of larger support for the 2^{nd} layer values $\sigma = \text{ReLu}, \sigma_\star = \tanh$. Theoretical (continuous lines) and numerical (dots) predictions for the generalization error as a function of the number of samples per dimension α for different values of the second layer vocabulary size $k \in (1, 2, 4)$. The numerical simulations consider $\lambda = 0.01, \gamma = 0.5, \beta = 1.5, p = 2048$. Note the significant drop in the generalization error for $k > 1$. The choice of the probabilities $\pi = \{\pi_q\}_{q\in[k]}$ and the vocabulary $\zeta = \{\zeta_q\}_{q\in[k]}$ are: a) $\mathbf{k} = \mathbf{1} : \pi = \{1\}, \zeta = \{1\}$; b) $\mathbf{k} = \mathbf{2} : \pi = \{0.9, 0.1\}, \zeta = \{1, -1\}$; c) $\mathbf{k} = \mathbf{4} : \pi = \{0.7, 0.1, 0.1, 0.1\}, \zeta = \{1, -0.5, 1.5 - 2\}$

5 CONSEQUENCES OF THE MAIN RESULTS

Tight characterization of the feature covariance spectrum – The singular values of the features after one, but non-maximal gradient step – $\eta \asymp p^{\zeta + 1/2}$ with $\zeta < 1/2$ – have been characterized in Moniri et al. [2023], showing how a series of ℓ spikes appear after the step, in addition to the RF bulk corresponding to the features at initialization, as characterized in e.g. [Pennington and Worah, 2017, Benigni and Péché, 2021, 2022, Fan and Wang, 2020, Louart et al., 2018]. The number of spikes is given by the integer ℓ such that $\ell{-}1/2\ell < \zeta < \ell/2\ell{+}2$. Note that in the maximal step size limit $\zeta = 1/2$, the corresponding ℓ diverges, and it is largely unclear how the bulk and spikes recombine into the limiting features covariance spectrum.

Our results further provide a tight asymptotic characterization of the bulk spectrum (represented in red in Fig. 1 for $k = 1, \alpha = 0.8, \sigma = \text{ReLu}, g = \sin, \tilde{\eta} = 3.3$). Note how this bulk is modified from the unspiked RF spectrum (the dashed blue line in Fig. 1), displaying in particular a wider support and longer tails in this particular instance. The spectrum also exhibits outlying eigenvalues of order $\Theta(d)$, arising from the means $\bar{\phi}_\mu$, not represented in Fig. 1 for readability. This theoretical result ties in with numerous previous empirical observations [Martin and Mahoney, 2021, Martin et al., 2021, Wang et al., 2024a] that heavier tails can emerge after feature learning, a behaviour which further tends to correlate with better generalization abilities. Interestingly, this phenomenon persists even when the network is trained with *multiple* large stochastic GD steps, or with adaptive optimizers such as Adam [Kingma, 2014], as empirically observed by Wang et al. [2024a].

On a qualitative level, the departure of the bulk from its untrained shape can be intuitively seen as a result of the recombination between the untrained bulk and some of the spikes predicted by Moniri et al. [2023], as they proliferate when $\zeta \to 1/2$. We stress that in addition to these spikes, the feature covariance spectrum includes k additional spikes related to the non-zero mean property of the features Φ. Such “spurious spikes” are not illustrated in Fig. 1 and are present due to the finite support assumption for the second layer (Ass. 2.1). It is an interesting avenue of future research to study the recombination of such spurious spikes with the bulk in the limit $k \to \infty$.

Precise characterization of the learned features – While two-layer neural networks are known to be universal approximators [Cybenko, 1989, Hornik et al., 1989], a precise characterization of the approximation space spanned by a given trained neural network feature maps remains to a large extent elusive. As we discuss in this paragraph, for features resulting from a maximal update on the first layer weights, the *diversity* of the second-layer initialization a^0, namely the number of different values its components take, plays a crucial role in allowing for expressive feature maps. Indeed, there is a net increase of the expressivity of the neural network for larger vocabulary sizes k. More precisely, the network is able to express non-linear functions in κ – the projection on the spike. This is reflected in Def. 4.4 containing non-linear dependence on κ along the functional basis $\{c_0(\kappa, \zeta_q), \kappa c_1(\kappa, \zeta_q)\}_{q\in[k]}$. This functional basis is larger, and thus the neural network more expressive, for larger vocabulary sizes k, i.e. when the second layer is initialized with more variability. Furthermore, we note that the non-linear dependence on κ of the basis is modulated by $\{\zeta_q\}_{q\in[k]}$ that relates to the learning rate strength η as per Lemma 3.1.

To illustrate this point concretely, consider for definiteness the case of an error function activation $\sigma(\cdot) = \mathrm{erf}(\cdot)$. Then, $\{c_0(\cdot, \zeta_q)\}_{q\in[k]} = \{\mathrm{erf}(\zeta_q/\sqrt{3}\cdot)\}_{q\in[k]}$. While a uniform second layer initialization $a^0 \propto 1_p$ ($k = 1$) only allows the network to express monotonic sigmoid functions, allowing the second initialization to take $k = 2$ values already allows to express non-monotonic function with a derivative which can change sign twice. Pushing the second layer diversity to vocabularies of size $k \geq 3$ further enriches the pool of expressible functions with further non-monotonic functions. Depending on the functional form of the target activation $\sigma_\star$, the variability of the second layer can thus prove particularly instrumental in reaching a good approximation and learning. In Fig. 2, we illustrate the role of the vocabulary size k, for the simplest possible setting (single-index target, $\sigma = \mathrm{relu}, \sigma_\star = \tanh$), by plotting the test error as a function of the sample complexity α for varying vocabulary sizes $k \in \{1, 2, 4\}$. We observe a net decrease in the test error performance with increasing second layer variability k at initialization. Furthermore, we exemplify in Fig. 3 the comparison between two-layer networks and kernel methods. In the proportional sample regime probed in our analysis (Ass. 2.2) fixed-feature methods attains asymptotically the errror achieved best linear predictor (see e.g. Mei and Montanari [2022]). In contrast, learning features with two-layer nets enables to fit non-linear function of the projection on the spike κ (Def. 4.4) and therefore surpassing such limitations.

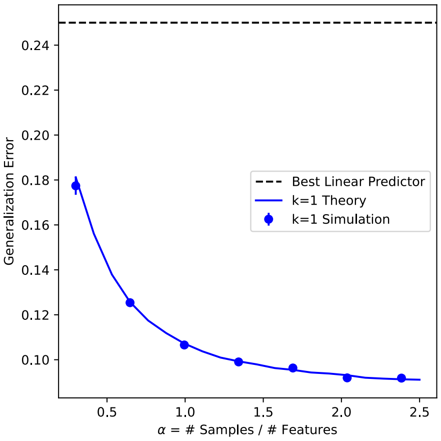

Figure 3: **Surpassing Kernel Methods:** Illustration of the superiority of two-layer networks over kernel methods. Learning features enables two-layer networks to surpass the lowest achievable test error by fixed-feature method, corresponding to the best linear predictor (dashed horizontal black line). The numerical simulations consider $\lambda = 0.1, \gamma = 0.5, \beta = 4, p = 1024, k = 1$ and $\sigma = \sigma_\star = \mathrm{ReLu}$.

6 CONCLUSIONS AND LIMITATIONS

We present a rigorous random matrix theory analysis of feature learning in a two-layer neural network, when the first layer is trained with a single, but aggressive, gradient step, in the limit where the number of samples n, the input dimension d, the hidden layer width p and the learning rate η, jointly tend to infinity at proportional rates. We rigorously justify how the trained neural network can be approximated by a spiked random features model in the limit of large batch sizes. We derive a deterministic equivalent for the empirical covariance matrix of the resulting features. We further provide a tight asymptotic characterization of the test error, when the second layer is subsequently trained with ridge regression. Our results provide a rigorous proof to the heuristic work of Cui et al. [2024], while extending it in multiple aspects. In particular, we allow for non-uniform initialization for the second-layer weights. We discuss how the second-layer variability enhances the expressivity of the trained network, and its ability to fit a single-index target. Among the limitations are the finitely supported second layer initialization, the use of Gaussian data, and asymptotic nature of the results. We note, however, that numerical experiments shows that predictions are accurate even at fairly moderate sizes. Secondly, a number of universality results shows that such ensembles extend over larger datasets [Dudeja et al., 2023, Gerace et al., 2024, Loureiro et al., 2021, Pesce et al., 2023, Wang et al., 2022].

ACKNOWLEDGEMENTS

We would like to thank Denny Wu, Edgar Dobriban, Lenka Zdeborová and Ludovic Stephan for stimulating discussions. We also thank the Institut d'Études Scientifiques de Cargèse for the hospitality during the "Statistical physics & machine learning back together again" workshop, where this work started. BL acknowledges funding from the French government, managed by the National Research Agency (ANR), under the France 2030 program with the reference "ANR-23-IACL-0008" and the Choose France - CNRS AI Rising Talents program. YD and HC acknowledge support from the Swiss National Science Foundation grant SNFS SMArtNet (grant number 212049). YD, LP, and FZ acknowledge support from the SNFS grant OperaGOST (grant number 200390). The work of YML was supported in part by the Harvard FAS Dean's Fund for Promising Scholarship and by a Harvard College Professorship. HC acknowledges support from the Center of Mathematical Sciences and Applications (CMSA) of Harvard University.

References

Francis Bach. The quest for adaptivity, 2021. URL https://francisbach.com/quest-for-adaptivity/.

Charles H Martin and Michael W Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021.

Charles H Martin, Tongsu Peng, and Michael W Mahoney. Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. *Nature Communications*, 12(1):4122, 2021.

Zhichao Wang, Andrew Engel, Anand D Sarwate, Ioana Dumitriu, and Tony Chiang. Spectral evolution and invariance in linear-width neural networks. *Advances in Neural Information Processing Systems*, 36, 2024a.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. *Advances in neural information processing systems*, 20, 2007.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, 145: 426–471, 2022.

Hong Hu and Yue M Lu. Universality laws for high-dimensional learning with random features. *IEEE Transactions on Information Theory*, 2022.

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.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, volume 35, pages 37932–37946, 2022.

Greg Yang, Edward J Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. *arXiv preprint arXiv:2203.03466*, 2022.

Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. How two-layer neural networks learn, one (giant) step at a time. *arXiv preprint arXiv:2305.18270*, 2023.

Behrad Moniri, Donghwan Lee, Hamed Hassani, and Edgar Dobriban. A theory of non-linear feature learning with one gradient step in two-layer neural networks. *arXiv preprint arXiv:2310.07891*, 2023.

Hugo Cui, Luca Pesce, Yatin Dandi, Florent Krzakala, Yue M Lu, Lenka Zdeborová, and Bruno Loureiro. Asymptotics of feature learning in two-layer networks after one gradient-step. *arXiv preprint arXiv:2402.04980*, 2024.

M Mezard, G Parisi, and M Virasoro. *Spin Glass Theory and Beyond*. WORLD SCIENTIFIC, 1986. doi: 10.1142/0271.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 1024–1034, 2020.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature Communications*, 12(1):2914, 2021.

Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. *Advances in Neural Information Processing Systems*, 34: 10131–10143, 2021.

Hugo Cui, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Error scaling laws for kernel classification under source and capacity conditions. *Machine Learning: Science and Technology*, 4(3):035033, 2023.

Rainer Dietrich, Manfred Opper, and Haim Sompolinsky. Statistical mechanics of support vector networks. *Phys. Rev. Lett.*, 82:2975–2978, Apr 1999. doi: 10.1103/PhysRevLett.82.2975.

Konstantin Donhauser, Mingqi Wu, and Fanny Yang. How rotational invariance of common kernels prevents generalization in high dimensions. In *International Conference on Machine Learning*, pages 2804–2814. PMLR, 2021.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Limitations of lazy training of two-layers neural network. *Advances in Neural Information Processing Systems*, 32, 2019.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? *Advances in Neural Information Processing Systems*, 33:14820–14830, 2020.

M. Opper and R. Urbanczik. Universal learning curves of support vector machines. *Phys. Rev. Lett.*, 86:4410–4413, 2001.

Lechao Xiao, Hong Hu, Theodor Misiakiewicz, Yue Lu, and Jeffrey Pennington. Precise learning curves and higher-order scalings for dot-product kernel regression. *Advances in Neural Information Processing Systems*, 35: 4558–4570, 2022.

Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Generalisation error in learning with random features and the hidden manifold model. In *International Conference on Machine Learning*, pages 3452–3462. PMLR, 2020.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022.

Hong Hu, Yue M Lu, and Theodor Misiakiewicz. Asymptotics of random feature regression beyond the linear scaling regime. *arXiv preprint arXiv:2403.08160*, 2024.

Fabián Aguirre-López, Silvio Franz, and Mauro Pastore. Random features and polynomial rules. *arXiv preprint arXiv:2402.10164*, 2024.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

Grant Rotskoff and Eric Vanden-Eijnden. Trainability and accuracy of artificial neural networks: An interacting particle system approach. *Communications on Pure and Applied Mathematics*, 75(9):1889–1935, 2022.

Justin Sirignano and Konstantinos Spiliopoulos. Mean field analysis of neural networks: A central limit theorem. *Stochastic Processes and their Applications*, 130(3):1820–1852, 2020.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *The Journal of Machine Learning Research*, 22(1):4788–4838, 2021.

Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2552–2623. PMLR, 2023.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, and Denny Wu. Learning in the presence of low-dimensional structure: a spiked random matrix perspective. *Advances in Neural Information Processing Systems*, 36, 2024.

Lorenzo Bardone and Sebastian Goldt. Sliding down the stairs: how correlated latent variables accelerate learning with neural networks. *arXiv preprint arXiv:2404.08602*, 2024.

Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks. *arXiv preprint arXiv:2303.00055*, 2023.

Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning gaussian multi-index models with gradient flow. *arXiv preprint arXiv:2310.19793*, 2023.

Alex Damian, Eshaan Nichani, Rong Ge, and Jason D Lee. Smoothing the landscape boosts the signal for sgd: Optimal sample complexity for learning single index models. *Advances in Neural Information Processing Systems*, 36, 2024.

Courtney Paquette, Kiwon Lee, Fabian Pedregosa, and Elliot Paquette. Sgd in the large: Average-case analysis, asymptotics, and stepsize criticality. In *Annual Conference Computational Learning Theory*, 2021.

Rodrigo Veiga, Ludovic Stephan, Bruno Loureiro, Florent Krzakala, and Lenka Zdeborová. Phase diagram of stochastic gradient descent in high-dimensional two-layer neural networks. *Advances in Neural Information Processing Systems*, 35:23244–23255, 2022.

Aaron Zweig and Joan Bruna. Symmetric single index learning. *arXiv preprint arXiv:2310.02117*, 2023.

Yatin Dandi, Emanuele Troiani, Luca Arnaboldi, Luca Pesce, Lenka Zdeborova, and Florent Krzakala. The benefits of reusing batches for gradient descent in two-layer networks: Breaking the curse of information and leap exponents. In *Forty-first International Conference on Machine Learning*, 2024.

Vladimir Alexandrovich Marchenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Matematicheskii Sbornik*, 114(4):507–536, 1967.

Zdzislaw Burda, A Görlich, Andrzej Jarosz, and Jerzy Jurkiewicz. Signal and noise in correlation matrix. *Physica A: Statistical Mechanics and its Applications*, 343:295–310, 2004.

Antti Knowles and Jun Yin. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169:257–352, 2017.

Zhidong Bai and Wang Zhou. Large sample covariance matrices without independence structures in columns. *Statistica Sinica*, pages 425–442, 2008.

Cosme Louart and Romain Couillet. Concentration of measure and large random matrices with an application to sample covariance matrices. *arXiv preprint arXiv:1805.08295*, 2018.

Clément Chouard. Quantitative deterministic equivalent of sample covariance matrices with a general dependence structure. *arXiv preprint arXiv:2211.13044*, 2022.

Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *The Annals of Applied Probability*, 28(2):1190–1248, 2018.

Dominik Schröder, Hugo Cui, Daniil Dmitriev, and Bruno Loureiro. Deterministic equivalent and error universality of deep random features learning. In *International Conference on Machine Learning*, pages 30285–30320. PMLR, 2023.

Dominik Schröder, Daniil Dmitriev, Hugo Cui, and Bruno Loureiro. Asymptotics of learning with deep structured (random) features. *arXiv preprint arXiv:2402.13999*, 2024.

David Bosch, Ashkan Panahi, and Babak Hassibi. Precise asymptotic analysis of deep random feature models. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4132–4179. PMLR, 2023.

Clément Chouard. Deterministic equivalent of the conjugate kernel matrix associated to artificial neural networks, 2023.

Zhichao Wang, Denny Wu, and Zhou Fan. Nonlinear spiked covariance matrices and signal propagation in deep neural networks. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 4891–4957. PMLR, 2024b.

R. Bellman, Rand Corporation, and Karreman Mathematics Research Collection. *Dynamic Programming*. Rand Corporation research study. Princeton University Press, 1957.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 5413–5452, 2022.

Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. The merged-staircase property: a necessary and nearly sufficient condition for sgd learning of sparse functions on two-layer neural networks. In *Conference on Learning Theory*, pages 4782–4887. PMLR, 2022.

Greg W Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices.* Number 118. Cambridge university press, 2010.

Francisco Rubio and Xavier Mestre. Spectral convergence for a general class of random matrices. *Statistics & probability letters*, 81(5):592–602, 2011.

Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. *Advances in neural information processing systems*, 30, 2017.

Lucas Benigni and Sandrine Péché. Eigenvalue distribution of some nonlinear models of random matrices. *Electronic Journal of Probability*, 26:1–37, 2021.

Lucas Benigni and Sandrine Péché. Largest eigenvalues of the conjugate kernel of single-layered neural networks. *arXiv preprint arXiv:2201.04753*, 2022.

Zhou Fan and Zhichao Wang. Spectra of the conjugate kernel and neural tangent kernel for linear-width neural networks. *Advances in neural information processing systems*, 33:7710–7721, 2020.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

Rishabh Dudeja, Yue M. Lu, and Subhabrata Sen. Universality of approximate message passing with semirandom matrices. *The Annals of Probability*, 51(5):1616–1683, 2023.

Federica Gerace, Florent Krzakala, Bruno Loureiro, Ludovic Stephan, and Lenka Zdeborová. Gaussian universality of perceptrons with random labels. *Physical Review E*, 109(3):034305, 2024.

Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151, 2021.

Luca Pesce, Florent Krzakala, Bruno Loureiro, and Ludovic Stephan. Are gaussian data all you need? the extents and limits of universality in high-dimensional generalized linear estimation. In *International Conference on Machine Learning*, pages 27680–27708. PMLR, 2023.

Tianhao Wang, Xinyi Zhong, and Zhou Fan. Universality of approximate message passing algorithms and tensor networks. *arXiv preprint arXiv:2206.13037*, 2022.

Yue M Lu and Horng-Tzer Yau. An equivalence principle for the spectrum of random inner-product kernel matrices with polynomial scalings. *arXiv preprint arXiv:2205.06308*, 2022.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12:389–434, 2012.

Andrea Montanari and Basil N. Saeed. Universality of empirical risk minimization. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4310–4312. PMLR, 02–05 Jul 2022.

Xiuyuan Cheng and Amit Singer. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 2(04):1350010, 2013.

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

Romain Couillet and Zhenyu Liao. *Random matrix methods for machine learning*. Cambridge University Press, 2022.

CHECKLIST

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model.
   [Yes]. The descrption of the setting and relevant assumptions is performed in Section 2.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm.
   [Yes]. We precisely clarify the algorithmic routine in Section 2.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.
   [No]. The interest of this work is primarily theoretical.
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results.
   [Yes]. The main results are accompanied always with the full set of assumptions.
   (b) Complete proofs of all theoretical results.
   [Yes]. The detailed proof of the different main results is carried over in the appendices.
   (c) Clear explanations of any assumptions.
   [Yes]. We provide explanations for the introduction of the different assumptions.
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL).
   [Yes]. Although the code has not been provided, we give the details of the simulations in the captions and relative appendices.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen).
   [Yes]. We provide the details used in the simulations, e.g. activation functions and hyperparameters, in the main text.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times).
   [Yes]. We clarify the specific statistics used to illustrate the results.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider).
   [No]. Our main results involve solutions of self-consistent equations that can be solved using CPUs on a personal computer.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]

   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

SUPPLEMENTARY MATERIAL

A Structure of the Appendix

The Appendix is organized as follows:

- In section B, we list some notations, definitions and preliminary results utilized throughout the proof.
- In section C, we prove the the isotropic-spike approximation of the gradient and show that in the limit $\alpha \to \infty$, it suffices to establish the deterministic equivalent and generalization error under the isotropic-spike approximation.
- In section D, we prove the main Theorem 4.5 characterizing the asymptotic deterministic equivalent of the sample covariance of the extended features.
- In section E, we show how the generalization error can be expressed through certain functionals of the deterministic equivalent and finally obtain Theorem 4.5.
- Finally in Section F, we provide proofs of certain auxiliary results used in the analysis along with additional theoretical investigations

B Preliminaries

B.1 Stochastic Domination

Throughout the analysis, we use the following notation for controlling high-probability bounds over stochastic error terms:

Definition B.1. [Stochastic dominance [Lu and Yau, 2022]] We say that a sequence of real or complex random variables X_d is stochastically dominated by another sequence Y_d if for all $\epsilon > 0$ and k, the following holds for large enough d:

$$\Pr\left[|X|_d > d^{\epsilon} |Y|_d\right] \leq d^{-k}. \tag{19}$$

We denote the above relation through the following notation:

$$X = \mathcal{O}_{\prec}(Y). \tag{20}$$

We further denote:

$$|X - Y| = \mathcal{O}_{\prec}(Z), \tag{21}$$

as:

$$X - Y = \mathcal{O}_{\prec}(Z). \tag{22}$$

Similarly for vectors $X, Y \in \mathbb{R}^{k_d}$ for some sequence of dimensions k_d, we use the shorthand:

$$X - Y = \mathcal{O}_{\prec}(Z), \tag{23}$$

to denote:

$$\|X - Y\| = \mathcal{O}_{\prec}(Z), \tag{24}$$

It is easy to check that stochastic dominance is closed under unions of polynomially many events in d. We will often exploit this while taking unions over $p = \mathcal{O}(d)$ neurons and $n = \mathcal{O}(d)$ samples. Furthermore, $\prec$ absorbs polylogarithmic factors i.e:

$$X = \mathcal{O}_{\prec}(Y) \implies X = \mathcal{O}_{\prec}((\operatorname{polylog} d)Y) \tag{25}$$

Furthermore, it subsumes exponential tail bounds of the form:

$$\Pr[X_d > tY_d] \leq e^{-t^{\alpha}}, \tag{26}$$

for some $\alpha > 0$, as well as polynomial tails of arbitrarily large degree:

$$\Pr[X_d > tY_d] \leq \frac{C_k}{t^k}, \tag{27}$$

for some sequence of constants C_k dependent on k.

Definition B.2 (Hermite Expansion). Let $f : \mathbb{R} \to \mathbb{R}$ be square-integrable function w.r.t the Gaussian measure. Then, f admits a series expansion in the orthonormal basis of Hermite polynomials given by:

$$f(x) = \sum_{i=0}^{\infty} a_i h_i(x), \tag{28}$$

where the convergence holds in L_2 w.r.t the Gaussian measure.

Lemma B.3 (Resolvent Identity). *Let, $A, B \in \mathbb{R}^{p \times p}$ be two invertible matrices, then:*

$$A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1} \tag{29}$$

Lemma B.4. *Let $z \in \mathbb{C}/\mathbb{R}^{+}$ be arbitary and let $A \in \mathbb{R}^{N \times N}$ denote a p.s.d matrix. Let $\zeta = \max(\left|Im(z)\right|, -Re(z)$. Then:*

$$\left\|(A - zI_N)\right\| \leq \frac{1}{\zeta} \tag{30}$$

Lemma B.5 (Burkholder's inequality (Lemma 2.12 in Bai and Zhou [2008])). *Let $X_i, i = 1, \cdots n$ be a complex-valued Martingale difference sequence w.r.t a filtration $\mathcal{F}_i$, then for any $p \geq 1$, there exists a constant K_p such that:*

$$\mathbb{E}\left[\left|\sum_{i=1}^{n} X_i\right|^p\right] \leq K_p \mathbb{E}\left[(\sum_{i=1}^{n} X_i^2)^{p/2}\right] \tag{31}$$

Definition B.6 (Lipschitz concentration [Louart and Couillet, 2018]). A random variable $X \in \mathbb{R}^d$ is said to be α-Lipschitz concentrated if for any 1-Lipschitz function f:

$$\mathbb{P}[f(X) \geq t] \leq \alpha(t) \tag{32}$$

Definition B.7 (Trace norm). For any $A \in \mathbb{C}^{p \times p}$, the Trace norm is defined as:

$$\|A\|_{tr} = \operatorname{Tr}\left(\sqrt{A^{\star} A}\right), \tag{33}$$

where $A^{\star}$ denotes Hermitian conjugate of A.

Lemma B.8 (Trace-norm inequality). *For any $A, B \in \mathbb{C}^{p \times p}$:*

$$\|AB\|_{tr} \leq \|A\|_{tr} \|B\|, \tag{34}$$

where $\|B\|$ denotes the operator norm.

Definition B.9 (Schur complement). Let p, q two non-negative integers such that $p + q > 0$, consider the matrix $A \in \mathbb{R}^{(p+q) \times (p+q)}$:

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

If A_{22} is invertible, the Schur complement of the block A_{22} of the matrix A is the matrix $A/A_{22} \in \mathbb{R}^{p \times p}$ defined as:

$$A/A_{22} = A_{11} - A_{12} A_{22}^{-1} A_{21} \tag{35}$$

Similarly, the Schur complement of the block A_{11} of the matrix A is the matrix $A/A_{11} \in \mathbb{R}^{q \times q}$ defined as:

$$A/A_{11} = A_{22} - A_{21} A_{11}^{-1} A_{12} \tag{36}$$

Lemma B.10. *[Theorem 5.2.2 in Vershynin [2018]] Let $z \sim \mathcal{N}(0, I_N)$ denote a standard Gaussian random vector in R^N and let $f : \mathbb{R}^N \to \mathbb{R}$ be a Lipschitz-function with Lipschiz-constant bounded by L. Then:*

$$\Pr\left[\left|f(z) - \mathbb{E}\left[f(z)\right]\right| \geq t\right] \leq 2e^{-ct^2/L} \tag{37}$$

Lemma B.11. *[Theorem 5.1.3 in [Vershynin, 2018]] Let $z \sim \mathcal{U}(\mathbb{S}^{N-1}(\sqrt{d}))$ denote a random vector uniformly sampled from the N dimensional sphere of radius $\sqrt{N}$. Then, for any Lipschitz-function $f : \mathbb{R}^N \to \mathbb{R}$ with Lipschiz-constant bounded by L:*

$$\Pr\left[\left|f(z) - \mathbb{E}\left[f(z)\right]\right| \geq t\right] \leq 2e^{-ct^2/L} \tag{38}$$

Lemma B.12. *[Theorem 5.3 in Vershynin [2010] and Theorem 4.3.5 in [Vershynin, 2018]] Under assumptions 2.1, the operator norms $\|W_0\|, \|X\|$ satisfy, for some constants C_1, C_2*

$$\|W_0\| \leq \frac{1}{\sqrt{d}} C_1(\sqrt{d} + \sqrt{p}) + \mathcal{O}_{\prec}(\frac{1}{\sqrt{d}})$$
$$\|X\| \leq C_2(\sqrt{n} + \sqrt{d}) + \mathcal{O}_{\prec}(1)$$

Lemma B.13. *[O'Donnell, Proposition 11.33, p. 338] The Hermite polynomials $(h_\alpha)_{\alpha\in\mathbb{N}}$ form a complete orthonormal basis. Further, for any $\rho \in [-1, 1]$ and two standard normal Gaussian random variables z, z' that are ρ-correlated, we have*

$$\mathbb{E}_{z,z'}\left[h_\alpha(z)h_\beta(z')\right] = \begin{cases} \rho^\alpha & \textit{if } \alpha = \beta, \\ 0 & \textit{if } \alpha \neq \beta. \end{cases} \tag{39}$$

Lemma B.14. *Let $f_1(z)$ and $f_2(z)$ be two twice-differentiable functions with the first and second derivatives bounded almost surely. Suppose that $f_1(z), f_2(z)$ admit the following Hermite expansions:*

$$f_1(z) = \sum_{k\geq 0} c_k h_k(z) \qquad f_2(z) = \sum_{k\geq 0} \tilde{c}_k h_k(z), \tag{40}$$

where $\{h_k(z)\}_k$ denote the set of normalized Hermite polynomials. Given three unit norm vectors w_1, w_2, w' such that:

$$\left|w_i^\top w'\right| = \mathcal{O}_{\prec}(\frac{1}{\sqrt{d}}) \tag{41}$$

for $i = 1, 2$. Then for $g \sim \mathcal{N}(0, I)$, the following holds:

$$\mathbb{E}\left[f_1(w_1^\top g)f_2(w_2^\top g)g^\top w'\right] = c_0\tilde{c}_1(w_2^\top w') + \tilde{c}_0 c_1(w_1^\top w') + \mathcal{O}_{\prec}(\frac{1}{d}), \tag{42}$$

$$\mathbb{E}\left[f_1(w_1^\top g)f_2(w_2^\top g)(g^\top w')^2\right] = c_0\tilde{c}_0 + \mathcal{O}(d^{-1/2}), \tag{43}$$

C Spiked isotropic approximation

In this section, we justify the assumption of an isotropic bulk in the regime $\alpha_0 \to \infty$, We first establish Lemma 3.1 and subsequently control the resulting approximation error for deterministic equivalence and generalization errors.

C.1 Proof of Lemma 3.1

Let $\hat{y}_1, \cdots, \hat{y}_n$ denote the outputs of the network at initialization, i.e:

$$\hat{y}_\mu = f(x_\mu; W^{(0)}, a) \tag{44}$$

We start by expressing the gradient as:

$$G = \frac{1}{n\sqrt{p}} \operatorname{diag}\{a_1, \ldots a_p\} \sigma'(W^0 X) \operatorname{diag}\{y_1 - \hat{y}_1, \ldots, y_n - \hat{y}_n\} X \in \mathbb{R}^{p\times d}, \tag{45}$$

where $X \in n \times d$ denotes the matrix with rows the data vectors in the first batch, $\{y_i\}_{i\in[n]}$ are the corresponding labels, and $\sigma'(\cdot)$ is the derivative of the student activation function.

Next, we decompose σ' as:

$$\sigma'(x) = c_1 + \sigma'_{>1}(x), \tag{46}$$

where

$$\sigma'_{>1}(x) \overset{\text{def}}{=} \sum_{k\geq 2} \sqrt{k!}\,\mu_k\, h_{k-1}(x), \tag{47}$$

and c_k denote the Hermite coefficients of σ. Using the decomposition in (281), we can rewrite the weight matrix W^1 as

$$W^1 = W^0 + uv^\top + \underbrace{\frac{\eta}{n}\operatorname{diag}\{\sqrt{p}a_1, \dots \sqrt{p}a_p\}\sigma'_{>1}(W^0X)\operatorname{diag}\{y_1, \dots, y_n\}X^T}_{\Delta_1}$$
$$- \underbrace{\frac{\eta}{n}\operatorname{diag}\{\sqrt{p}a_1, \dots \sqrt{p}a_p\}\sigma'_{>1}(W^0Z)\operatorname{diag}\{\hat{y}_1, \dots, \hat{y}_n\}Z^T}_{\Delta_2}$$

where $u = \mu_1\eta\sqrt{p}a$ and $v = X^\top y/n$.

Our goal is therefore to bound the contributions from Δ_1, Δ_2. We start by expressing Δ_1 as a sum of rank one terms:

$$\Delta_1 = \frac{1}{n}\sum_{\mu=1}^{n_0} b_\mu x_\mu^\top, \tag{48}$$

where $b_\mu^\top = y_\mu\sigma'(W^0X)[\sqrt{p}a_1, \dots \sqrt{p}a_p]$.

Let $\{c^\star_k\}_{k\in\mathbb{N}}$ denote the Hermite coefficients of g. An application of Stein's Lemma and Hermite expansion yields (See Lemma 7 in Damian et al. [2022] or Lemma 4 in Dandi et al. [2023]) yields the following expansion for the i_{th} row of Δ_1:

$$\mathbb{E}_{x_\mu}\left[b^i_\mu x_\mu\right] = \sum_{k=1}^{\infty} c_{k+1}\mu^\star_{k+1}\langle (w^\star)^{\otimes k+1}, w_i^{\otimes k}\rangle + \sum_{k=1}^{\infty} c_{k+2}c^\star_k\langle (w^\star)^{\otimes k}, w_i^{\otimes k}w_i\rangle \tag{49}$$

By assumption $c_2 = 0$, therefore the term $c_2, c^\star_2 w^\star\langle w^\star, w_i\rangle$ vanishes. Since $\langle w^\star, w_i\rangle = \mathcal{O}_\prec(\frac{1}{\sqrt{d}})$ for all $i \in [p]$, the remaining terms are bounded in term by $\mathcal{O}_\prec(\frac{1}{d^{1.5}})$. We obtain:

$$\left\|\mathbb{E}_{x_\mu}\left[b^i_\mu x_\mu\right]\right\|_F = \mathcal{O}_\prec(\frac{1}{d}), \tag{50}$$

which results in the following bound on the operator norm:

$$\left\|\mathbb{E}_x\left[\Delta_1\right]\right\| = \mathcal{O}_\prec(\frac{1}{\sqrt{d}}). \tag{51}$$

Next, note that the uniform boundedness of σ', σ^* imply that $\left\|b_\mu\right\|^2 < Cp$ for some constant $C > 0$. Consider the symmetrized matrix:

$$B_\mu = \begin{bmatrix} 0 & b_\mu x_\mu^\top \\ x_\mu b_\mu^\top & 0 \end{bmatrix} \in \mathbb{R}^{(p+d)\times(p+d)} \tag{52}$$

Then, for all $k \geq 1$:

$$B_\mu^{2k+1} = \begin{bmatrix} 0 & \left\|x_\mu\right\|^{2k}\left\|b_\mu\right\|^{2k} b_\mu x_\mu^\top \\ \left\|x_\mu\right\|^{2k}\left\|b_\mu\right\|^{2k} x_\mu b_\mu^\top & 0 \end{bmatrix} \tag{53}$$

$$B_\mu^{2k} = \begin{bmatrix} \left\|x_\mu\right\|^{2k}\left\|b_\mu\right\|^{2k-2} b_\mu b_\mu^\top & 0 \\ 0 & \left\|x_\mu\right\|^{2k-2}\left\|b_\mu\right\|^{2k} x_\mu x_\mu^\top \end{bmatrix} \tag{54}$$

Now, let $z \in \mathbb{R}^{p+d}$ with $\|z\| = 1$ and let $z_p = z[: p]$, $z_d = z[p :]$ denote the first p and the remaining components of z respectively. From Equation (53), we have:

$$\mathbb{E}\left[z^\top B_\mu^{2k+1} z\right] = \mathbb{E}\left[\left\|x_\mu\right\|^{2k}\left\|b_\mu\right\|^{2k}\langle b_\mu, z_p\rangle\langle x_\mu, z_d\rangle\right] + \mathbb{E}\left[\left\|x_\mu\right\|^{2k}\left\|b_\mu\right\|^{2k}\langle b_\mu, z_p\rangle\langle x_\mu, z_d\rangle\right]. \tag{55}$$

Applying Cauchy–Schwarz to each term in the RHS yields:

$$\mathbb{E}\left[z^\top B_\mu^{2k+1} z\right] \leq \mathbb{E}\left[\left\|x_\mu\right\|^{4k}\left\|b_\mu\right\|^{4k}\right]^{1/2} \mathbb{E}\left[(\langle b_\mu, z_p\rangle)^2(\langle x_\mu, z_d\rangle)^2\right]^{1/2} + \mathbb{E}\left[\left\|x_\mu\right\|^{4k}\left\|b_\mu\right\|^{4k}\right]^{1/2} \mathbb{E}\left[(\langle b_\mu, z_p\rangle)^2(\langle x_\mu, z_d\rangle)^2\right]^{1/2} \tag{56}$$

Now, the boundedness of $\sigma', \sigma^\star$ imply that $\left\|b_\mu\right\|^{4k} \leq C_1^{4k} p^{2k}$ for some constant C_1. While $\left\|x_\mu\right\|^2$ is a sub-exponential random variable with parameter d and therefore [Vershynin, 2018]:

$$\mathbb{E}\left[\left\|x_\mu\right\|^{4k}\right] \leq C_2^{2k} d^{2k} 2k^{2k}, \tag{57}$$

for some constant $C_2 > 0$. Therefore:

$$\mathbb{E}\left[\left\|x_\mu\right\|^{4k}\left\|b_\mu\right\|^{4k}\right] \leq C_2^{2k} d^{2k} 2k^{2k} \times C_1^{4k} p^{2k} \tag{58}$$

Finally, by assumption 2.3, σ', g are uniformly-lipschitz. Furthermore, with high probabilty over W, $\|W\| \leq C_3$ for some constant C_4. We therefore obtain that $x \to \langle b_\mu, z_p\rangle$ is uniformly lipschitz in x with high probabilty over W.

Furthermore, applying Lemma B.13 to σ', g and using $c_2 = 0$, yields:

$$\mathbb{E}\left[\langle b_\mu, z_p\rangle\right] = \mathcal{O}(\frac{1}{\sqrt{d}}). \tag{59}$$

Therefore, $\mathbb{E}\left[(\langle b_\mu, z_p\rangle)^2(\langle x_\mu, z_d\rangle)^2\right]^{1/2}$ is further bounded by some constant $C_4 > 0$.

Since $p/d = \beta$ is a constant, substituting in Equation (56) we obtain:

$$\mathbb{E}\left[z^\top B_\mu^{2k+1} z\right] \leq (C_4 d)^{2k-1} d. \tag{60}$$

Similarly,

$$\mathbb{E}\left[z^\top B_\mu^{2k} z\right] \leq (C_5 d)^{2k-2} d, \tag{61}$$

for some constant $C_5 > 0$.

Therefore:

$$\mathbb{E}\left[B_\mu^k\right] \prec (C_4 d)^{k-2} d I_{p+d}, \tag{62}$$

for some constant C_4.

Subsequently, we apply the matrix-Bernstein inequality for self-adjoint matrices with subexponential tails (Theorem 6.2 in Tropp [2012]) to obtain that:

$$\Pr\left[\left\|\frac{1}{n}\sum_{\mu=1}^{n} B_\mu - \left\|\mathbb{E}\left[B_\mu\right]\right\|\right\| \geq t\right] \leq d e^{-t(\log d)^2/(c_1+c_2 t)}, \tag{63}$$

for some constants $c_1, c_2 > 0$. Borel-Cantelli Lemma then implies:

$$\|\Delta_1\| \xrightarrow[a.s]{d\to\infty} 0. \tag{64}$$

Now, to bound Δ_2, we use:

$$\|\Delta_2\| \leq \frac{\eta}{n}\left\|\operatorname{diag}\{\sqrt{p}a_1, \ldots \sqrt{p}a_p\}\right\|\left\|\sigma'_{>1}(W^0 Z)\right\|\left\|\operatorname{diag}\{\hat{y}_1, \ldots, \hat{y}_n\}\right\|\|X\|. \tag{65}$$

We show that:

$$\left\|\operatorname{diag}\{\hat{y}_1, \ldots, \hat{y}_n\}\right\| = \mathcal{O}_\prec(\frac{1}{\sqrt{d}}) \tag{66}$$

Recall that:

$$\hat{y}_\mu = \frac{1}{\sqrt{p}} \sum_{j=1}^{p} a_j \sigma(w_j^\top x_\mu) \tag{67}$$

By Theorem 3.1.1 in Vershynin [2018], $\left| \|x_\mu\| - \sqrt{d} \right|$ for $\mu \in [n]$ are independent sub-Gaussian random variables. Therefore:

$$\sup_{\mu \in [n]} \|x_\mu\| = \sqrt{d} + \mathcal{O}_\prec(\sqrt{\log n}). \tag{68}$$

Therefore, we may condition on the high-probability event:

$$\mathcal{E}_n = \{ \sup_{\mu \in [n]} \|x_\mu\| \leq C\sqrt{d} \}, \tag{69}$$

for some $C > 1$. Conditioned on the event $\mathcal{E}_n$, Lemma B.11 and assumption 2.3 imply that for each $\mu \in [n]$, $\sqrt{p} a_j \{\sigma(w_j^\top x_\mu), j \in [p]\}$ are independent sub-Gaussian random variables with mean 0. (Recall that by assumption 2.3, σ is odd). Therefore $\forall \mu \in [n]$:

$$\hat{y}_\mu = \mathcal{O}_\prec(\frac{1}{\sqrt{d}}), \tag{70}$$

where we absorbed polylogarithmic factors through the stochastic domination notation (Definition B.1).

Lastly, it remains to show that the spike v converges to $w^\star$. Recall that $v := \frac{1}{c_0^\star} \frac{X^\top y}{n}$. By assumptions 2.3, $y_\mu x_\mu^i$ are independent sub-Gaussian random variables for $i \in [d]$. Therefore:

$$\|v - w^\star\| = \mathcal{O}_\prec(\sqrt{d/n_0}). \tag{71}$$

Since, $n_0 = \mathcal{O}(d^{1+\epsilon})$, we obtain:

$$v \xrightarrow[a.s]{} w^\star \tag{72}$$

It remains to show that the equivalence in the sense of Lemma 3.1 extends to generalization error and deterministic equivalence:

Proposition C.1. *Let G^e, $(G_\star)^e$ denote the extended resolvents with features $W^{(1)}$ and $\tilde{W} = W^0 + \eta u (w^\star)^\top$ respectively. Analogously, let $\hat{a}, \hat{a}_\star$ denote the ridge-regression predictors corresponding to $W^{(1)}$ and $\tilde{W}$ respectively. Then,for sequence of deterministic matrices $A \in \mathbb{C}^{(p+1)\times(p+1)}$ with $\|A\|_{\mathrm{tr}} = \mathrm{Tr}\Big((AA^*)^{1/2}\Big)$ uniformly bounded in d:*

$$\left| \mathrm{Tr}\big(AG_e(z)\big) - \mathrm{Tr}(AG_\star)^e \right| \xrightarrow[d\to\infty]{a.s} 0. \tag{73}$$

$$\left| e_{gen}(\hat{a}, W^{(1)}) - e_{gen}(\hat{a}_\star, W^{(1)}) \right| \xrightarrow[d\to\infty]{a.s} 0. \tag{74}$$

Proof. Consider the matrix:

$$\Xi = \sigma(XW^{(1\top)}) - \sigma(X\tilde{W}^\top) \in \mathbb{R}^{n\times p}. \tag{75}$$

A straightforward consequence of the proof of Lemma 3.1 and Assumption 2.3, Lemma B.10 is that each row of Ξ is a sub-Gaussian random vector with sub-Gaussian norms $\mathcal{O}(\frac{\mathrm{polylog}\, d}{d}^\epsilon)$. Therefore, through tail bounds on the operator norms of matrices with sub-Gaussian rows [Vershynin, 2010], we obtain:

$$\frac{1}{\sqrt{p}} \|\Xi\| \xrightarrow[a.s]{} 0 \tag{76}$$

The claim in Eq (73) then follows since by Lemma B.8:

$$\begin{aligned} \left| \mathrm{Tr}\big(AG_e(z)\big) - \mathrm{Tr}(AG_\star)^e \right| &\leq \|A\|_{\mathrm{Tr}} \big\|G_e(z) - G_\star\big\| \\ &\leq \|A\|_{\mathrm{Tr}} \frac{C}{\zeta^2} \left\| \sigma(XW^{(1)^\top}) - \sigma(X\tilde{W}^\top) \right\|, \end{aligned}$$

for some constant C. In the last line, we used Lemma B.3 and ζ is defined as in Lemma B.4. Similarly, Equation (76) implies the almost sure convergence for the generalization error. □

D Deterministic Equivalent

The proof of Theorem 4.5 proceeds in three parts:

- Approximation of the covariance $\mathbb{E}\left[\phi\phi^\top\right]$ through the covariance $R^\star_\kappa$ of a conditional-Gaussian equivalent model at fixed values of $\boldsymbol{W}$. $R^\star_\kappa$ possesses a block-structure due to the finite-support assumption over u_i
- First-stage deterministic equivalent: here we average over the randomness in the inputs $\boldsymbol{X}$ and construct a deterministic equivalent dependent on $\boldsymbol{W}$ and expressed as a functional of $R^\star_\kappa$.
- Second-stage deterministic equivalent: Here we further average over the randomness in $\boldsymbol{W}$ to establish deterministic equivalents for $R^\star_\kappa$ and associated transforms.

We describe each of these stages below:

D.1 Gaussian-equivalent covariance approximation

We start by showing that conditioned on κ, the covariance of the extended features ϕ^e_μ can be well-approximated through that of an equivalent linear model. This approximation is similar to the *conditional Gaussian equaivalence* in Dandi et al. [2023], Cui et al. [2024]. However, we emphasize that we do not directly utilize any universality result on generalization errors established in recent works [Lu and Yau, 2022, Montanari and Saeed, 2022, Dandi et al., 2023]. Instead, we will directly utilize the approximation of the covariance in our analysis of the spectrum and generalization errors in the subsequent sections.

Recall the definition of the surrogate feature vectors in equation (11):

$$\phi^e_\mu = \begin{pmatrix} y_\mu \\ \bar{\phi}_\mu \\ \tilde{\phi}_\mu \end{pmatrix} \tag{77}$$

We observe that the sample covariance matrix of z_μ posseses the following block structure:

$$(\Phi^e_\mu)^\top(\Phi^e_\mu) = \begin{bmatrix} \boldsymbol{y}\boldsymbol{y}^\top & \boldsymbol{y}\bar{\Phi}^\top & \boldsymbol{y}\tilde{\Phi}^\top \\ \bar{\Phi}\boldsymbol{y}^\top & \bar{\Phi}\bar{\Phi}^\top & \tilde{\Phi}\bar{\Phi}^\top \\ \tilde{\Phi}\boldsymbol{y}^\top & \bar{\Phi}\tilde{\Phi}^\top & \tilde{\Phi}\tilde{\Phi}^\top \end{bmatrix} \tag{78}$$

Define:

$$\Sigma_{11} := \begin{bmatrix} \boldsymbol{y}\boldsymbol{y}^\top & \boldsymbol{y}\bar{\Phi}^\top \\ \bar{\Phi}\boldsymbol{y}^\top & \bar{\Phi}\bar{\Phi}^\top \end{bmatrix} \quad \Sigma_{21} := \begin{bmatrix} \tilde{\Phi}y^\top \\ \tilde{\Phi}\bar{\Phi}^\top \end{bmatrix} \quad \Sigma_{22} := \begin{bmatrix} \tilde{\Phi}\tilde{\Phi}^\top \end{bmatrix} \tag{79}$$

Proposition D.1. *For a vector $u \in \mathbb{R}^m$, let $c_j(u,\kappa) \in \mathbb{R}^m$ denote the vector with entries $c_j(u_i,\kappa)$. Let $\mathbb{E}_\kappa\left[\cdot\right]$ denote the expectation w.r.t x conditioned on the sigma-algebra generated by $\kappa := x^\top w^*$. Define:*

$$R^\star_{11}(\kappa) = \begin{bmatrix} g^2(\kappa) & g(\kappa)c_0(\kappa,u_\pi)^\top \\ g(\kappa)c_0(\kappa,u_\pi) & c_0(\kappa,u_\pi)c_0(\kappa,u_\pi)^\top \end{bmatrix}, \tag{80}$$

where u_π denotes the vector $(u_1,\cdots,u_k)$

$$R^\star_{21}(\kappa) = c_1(\kappa,u) \odot \boldsymbol{\theta}\kappa\left[\sigma_\star(\kappa) \quad c_0(\kappa,u_1)\cdots \quad c_0(\kappa,u_k)\right], \tag{81}$$

$$R^\star_{22}(\kappa) = (c_1(\kappa,u)c_1(\kappa,u)^\top) \odot WW^\top) + \mathrm{diag}(\left(\sum_{k\geq 2} c_k^2(\kappa,u)\right) + (\kappa^2-1)c_1(\kappa,u)c_1(\kappa,u)^\top \odot \boldsymbol{\theta}\boldsymbol{\theta}^\top$$
$$+\frac{1}{d}c_2(\kappa,u)c_2(\kappa,u)^\top$$

Then, there exists an $\ell \in \mathbb{N}$ such that, the following holds almost surely over κ:

$$\sup_{\kappa\in\mathbb{R}/\mathcal{E}_0} \left\|\mathbb{E}_\kappa\left[\Sigma_{11}\right] - R^\star_{11}(\kappa)\right\| = \mathcal{O}_\prec(\frac{\kappa^\ell}{\sqrt{d}}) \tag{82}$$

$$\sup_{\kappa\in\mathbb{R}/\mathcal{E}_0} \left\|\mathbb{E}_\kappa\left[\Sigma_{21}\right] - \mathbb{R}^\star_{21}(\kappa)\right\| = \mathcal{O}_\prec(\frac{\kappa^\ell}{\sqrt{d}}) \tag{83}$$

$$\sup_{\kappa\in\mathbb{R}/\mathcal{E}_0} \left\|\mathbb{E}_\kappa\left[\Sigma_{22}\right] - R^\star_{22}(\kappa)\right\| = \mathcal{O}_\prec(\frac{\kappa^\ell}{\sqrt{d}}) \tag{84}$$

Proof. Let w_i denote the ith row of W and $x \sim \mathcal{N}(0, I_d)$. By the rotational invariance of the Gaussian measure, we can express x as:

$$x = \kappa w^* + (I_d - w^*(w^*)^\top)\xi, \tag{85}$$

where $\xi \sim \mathcal{N}(0, I_d)$ is independent of x

We obtain:

$$x^\top w^1 = \kappa u_i + w_i^\top \xi + w_i^\top w^*(\kappa - \xi^\top w^*) \tag{86}$$

Therefore,

$$\phi_i(x) = \sigma(\kappa u_i + w_i^\top \xi + w_i^\top w^*(\kappa - \xi^\top w^*)). \tag{87}$$

Since

$$\left|w_i^\top w^*(\kappa - \xi^\top w^*)\right| = \mathcal{O}_\prec(\frac{\kappa}{\sqrt{d}}), \tag{88}$$

by assumption 2.3 an application of the Taylor's theorem yields:

$$\begin{aligned}\phi_i = \sigma(\kappa u_i + w_i^\top \xi) &+ \sigma'(\kappa u_i + w_i^\top \xi)\left[w_i^\top w^*(\kappa - \xi^\top w^*)\right] \\ &+ \frac{1}{2}\sigma''(\kappa u_i + w_i^\top \xi)\left[w_i^\top w^*(\kappa - \xi^\top w^*)\right]^2 + \mathcal{O}_\prec(\kappa^3 d^{-3/2}).\end{aligned}. \tag{89}$$

To simplify the expectations of the second, third terms, we leverage the Hermite expansion and Lemmas B.13, B.14.

We begin by expanding σ along the Hermite-basis for a fixed value of κu_i:

$$\sigma(\kappa u_i + z) = c_0(\kappa u_i) + c_1(\kappa u_i)h_1(z) + c_2(\kappa u_i)h_2(z) + \ldots + \tag{90}$$

Using the identity $h_k'(z) = \sqrt{k}h_{k-1}(z)$, we also have

$$\sigma'(u_i\kappa + z) = c_1(\kappa, u_i) + \sqrt{2}c_2(\kappa, u_i)h_1(z) + \sqrt{3}c_3(\kappa, u_i)h_2(z) + \ldots \tag{91}$$

and

$$\sigma''(u_i\kappa + z) = \sqrt{2}c_2(\kappa, u_i) + \sqrt{6}c_3(\kappa, u_i)h_1(z) + \ldots \tag{92}$$

Consider the second term in Equation (89):

$$\mathbb{E}_\kappa\left[\sigma'(\kappa u_i + w_i^\top \xi)\left[w_i^\top w^*(\kappa - \xi^\top w^*)\right]\right] = \kappa(w_i^\top w^*)\mathbb{E}\left[\sigma'(\kappa u_i + w_i^\top \xi)\right] - (w_i^\top w^*)\mathbb{E}\left[\sigma'(\kappa u_i + w_i^\top \xi)\xi^\top w^*\right] \tag{93}$$

From (91), the the first term in the RHS equals $\kappa(w_i^\top w^*)c_1(\kappa, u_i)$ while from Lemma B.13, the second term equals

$$\mathbb{E}_\kappa\left[\sigma'(\kappa u_i + w_i^\top \xi)\xi^\top w^*\right] = \sqrt{2}c_2(\kappa, u_i)w_i^\top w^*. \tag{94}$$

Combining, we obtain:

$$\mathbb{E}_\kappa\left[\sigma'(\kappa u_i + w_i^\top \xi)\left[w_i^\top w^*(\kappa - \xi^\top w^*)\right]\right] = \kappa(w_i^\top w^*)c_1(\kappa, u_i) - \sqrt{2}c_2(\kappa, u_i)(w_i^\top w^*)^2. \tag{95}$$

Next, consider the third term in Equation (89):

$$\mathbb{E}_\kappa\left[\frac{1}{2}\sigma''(\kappa u_i + w_i^\top \xi)\left[w_i^\top w^*(\kappa - \xi^\top w^*)\right]^2\right] = (w_i^\top w^*)^2\frac{1}{2}\mathbb{E}_\kappa\left[\sigma''(\kappa u_i + w_i^\top \xi)\left[\kappa^2 - 2\xi^\top w^* + (\xi^\top w^*)^2\right]\right] \tag{96}$$

Again applying Equation (92) and Lemma B.13 to each term in the RHS yields, simplifies as follows:

$$\begin{aligned}\mathbb{E}_\kappa\left[\frac{1}{2}\kappa^2(w_i^\top w^*)^2\sigma''(\kappa u_i + w_i^\top \xi)\right] &= \frac{1}{\sqrt{2}}\kappa^2 c_2(\kappa, u_i)\\ \mathbb{E}_\kappa\left[(w_i^\top w^*)^2\sigma''(\kappa u_i + w_i^\top \xi)\xi^\top w^*\right] &= \sqrt{6}c_3(\kappa, u_i)(w_i^\top w^*)^3 = \mathcal{O}_\prec(\frac{1}{d^{3/2}})\\ \mathbb{E}_\kappa\left[\frac{1}{2}(w_i^\top w^*)^2\sigma''(\kappa u_i + w_i^\top \xi)(\xi^\top w^*)^2\right] &= \frac{1}{\sqrt{2}}c_2(\kappa, u_i)(w_i^\top w^*)^2 + \mathcal{O}_\prec(\frac{1}{d^2}),\end{aligned}$$

where in the second equation, we used that $c_3(\kappa, u_i)$ is uniformly bounded in κ by Assumption 2.3 and in the last equation, we used $(\xi^\top w^*)^2 = \sqrt{2}h_2(\xi^\top w^*) + 1$.

Combining, we obtain:

$$\mathbb{E}_\kappa\left[\phi_i\right] = c_0(\kappa, \boldsymbol{u}) + c_1(\kappa)kw_i^\top w^\star + \frac{c_2(\kappa, \boldsymbol{u})}{\sqrt{2}}\left[(w_i^\top w^*)^2(\kappa^2 - 1)\right] + \mathcal{O}_\prec(d^{-3/2}). \tag{97}$$

gives us:

$$\mathbb{E}_\kappa\left[\phi\right] = c_0(\kappa, u)1_p + c_1(\kappa, u)\odot W\left(\kappa w^\star\right) + \mathcal{O}_\prec(d^{-1}) \tag{98}$$

Now, since $\langle w_1, w^\star\rangle, \cdots, \langle w_p, w^\star\rangle \in \mathbb{R}^d$ are i.i.d. sub-Gaussian random variables, we obtain that:

$$\mathbb{E}_\kappa\left[\bar{\phi}\right]_q = c_0(\kappa, \zeta_q^u). \tag{99}$$

Furthermore, by Assumption 2.3 and since $\left\lVert W\right\rVert_2 < C$ a.s for large enough C, we obtain that $\bar{\phi}_q$ is an $\mathcal{O}(\frac{1}{\sqrt{d}})$-Lipschitz function of x, we obtain:

$$\bar{\phi} - \mathbb{E}_\kappa\left[\bar{\phi}\right] = \mathcal{O}_\prec(\frac{1}{\sqrt{d}}) \tag{100}$$

Therefore, in what follows, we will utilize:

$$\tilde{\phi}_i = \phi_i - c_0(\kappa, u_i) + \mathcal{O}_\prec(\frac{1}{\sqrt{d}}). \tag{101}$$

Combining the above approximation with Equation (97) and using the uniform boundedness of g directly yields the estimates for Σ_{11}, Σ_{21} in Equations (82) and (83).

To study the block Σ_{22}, we separate the case of diagonal and off-diagonal entries. For the former, we apply Lemma B.13 to Equation (89) to obtain:

$$\mathbb{E}_\kappa\left[\tilde{\phi}_i^2\right] = \sum_{k\geq 1} c_k^2(\kappa, u_i) + \mathcal{O}_\prec(\kappa/\sqrt{d}). \tag{102}$$

For the off-diagonal terms with $i \neq j$, we start by noting that

$$\mathbb{E}_\kappa\left[\tilde{\phi}\tilde{\phi}^\top\right] = (I_p - P)\mathbb{E}_\kappa\left[\phi\phi^\top\right](I_p - P)^\top, \tag{103}$$

where P denotes the projection on the directions $e^1, \cdots e^k$ defined in Equation (13). The Taylor expansion in (89) and the approximation (101) then gives us:

$$\mathbb{E}_\kappa\left[\tilde{\phi}_i\tilde{\phi}_j\right] = \underbrace{\mathbb{E}_\kappa\left[(\sigma(\kappa u_i + w_i^\top \xi) - \bar{\phi}_i)(\sigma(\kappa u_j + w_j^\top \xi) - \bar{\phi}_j))\right]}_{(A)} \tag{104}$$

$$+ \underbrace{\mathbb{E}_\kappa\left[(\sigma(\kappa u_i + w_i^\top \xi) - c_0(\kappa, u_i))\sigma'(\kappa u_j + w_j^\top \xi)\left[w_j^\top w^\star(\kappa - \xi^\top w^\star)\right]\right]_{(B)}} + (B)_{i\leftrightarrow j} \tag{105}$$

$$+ \underbrace{\frac{1}{2}\mathbb{E}_\kappa\left[(\sigma(\kappa u_i + w_i^\top \xi) - c_0(\kappa, u_i))\sigma''(\kappa u_j + w_j^\top \xi)\left[w_j^\top w^\star(\kappa - \xi^\top w^\star)\right]^2\right]}_{(C)} + (C)_{i\leftrightarrow j} \tag{106}$$

$$+ \underbrace{\mathbb{E}_\kappa\left[\sigma'(\kappa u_i + w_i^\top \xi)\left[w_i^\top w^\star(\kappa - \xi^\top w^\star)\right]\sigma'(\kappa u_j + w_j^\top \xi)\left[w_j^\top w^\star(\kappa - \xi^\top w^\star)\right]\right]}_{(D)} \tag{107}$$

$$+ \mathcal{O}(\kappa^3 d^{-3/2}), \tag{108}$$

where $(B)_{i\leftrightarrow j}, (C)_{i\leftrightarrow j}$ denote the corresponding terms with the roles of i, j interchanged. Next, we consider each term on the right-hand side. Applying Lemma B.13, we obtain that the term A results in:

$$(A) = (I_p - P)\left(c_1(\kappa, u)c_1(\kappa, u)^\top \odot W^\top W + c_2(\kappa, u)c_2(\kappa, u)^\top \odot (W^\top W)^2\right)(I_p - P)^\top + \mathcal{O}_\prec(d^{-3/2}). \tag{109}$$

For terms in (B), we apply Lemma B.14 with $w_1, w_2 = w_i, w_j$ and $w' = w^\star$ to obtain:

$$\begin{aligned}(B) + (B)_{i\leftrightarrow j} = &\left(\sqrt{2}c_1(\kappa, u_i)c_2(\kappa, u_j)w_i^\top w_j\right)(w_i + w_j)^\top(\kappa w^\star) \\ &- 2c_1(\kappa, u_i)c_1(\kappa, u_j)\left[(w_i^\top w^\star)(w_j^\top w^\star)\right] + \mathcal{O}_\prec(d^{-3/2}).\end{aligned} \tag{110}$$

Similarly, for terms in C, Lemma B.14 directly implies that:

$$(C) + (C)_{i\leftrightarrow j} = \mathcal{O}_\prec(d^{-3/2}). \tag{111}$$

Finally,

$$(D) = c_1(\kappa, u_i)c_1(\kappa, u_j)(\kappa^2 + 1)(w_i^\top w^\star)(w_j^\top w^\star) + \mathcal{O}_\prec(d^{-3/2}) \tag{112}$$

We are now ready to establish an approximation of the correlation matrix $\mathbb{E}\left[\tilde{\phi}\tilde{\phi}^\top\right]$, up to an error term whose operator norm is of size $\mathcal{O}_\prec(d^{-1/2})$. Starting with (A), since $w_i^\top e^q = \mathcal{O}_\prec(\frac{1}{\sqrt{d}})$ for $i \in [p], q \in [k]$, we have:

$$\left\|(I_p - P)\left(c_1(\kappa, u)c_1(\kappa, u)^\top \odot W^\top W\right)(I_p - P)^\top - c_1(\kappa, u)c_1(\kappa, u)^\top \odot W^\top W\right\| = \mathcal{O}_\prec(\frac{1}{\sqrt{d}}). \tag{113}$$

Therefore, the extra term compared to $R_{22}^\star(\kappa)$ is the one containing $(w_i^\top w_j)^2$. Write

$$(w_i^\top w_j)^2 = \frac{1}{d} + \sqrt{\frac{2p}{d^2}}(H_2)_{ij}, \tag{114}$$

$H_2 \in \mathbb{R}^{p\times p}$ is a symmetric matrix such that

$$(H_2)_{ij} = \frac{(\sqrt{d}w_i^\top w_j)^2 - 1}{\sqrt{2p}} 1_{i\neq j}. \tag{115}$$

This is in fact a Kernel gram matrix where the "kernel" function is the second Hermit polynomial $h_2(z)$. As studied in e.g. [Zhou & Montanari, 2013], [Lu & Yau, 2023], the spectrum of this kernel matrix is asymptotically equal to the semicircle law. Moreover, the operator norm of this matrix is bounded by $\mathcal{O}_\prec(1)$. It follows that the term $\sqrt{\frac{2p}{d^2}}(H_2)_{ij}$ in (114) can be ignored due to its vanishing operator norm. Thus, what is left is the contribution from the term $\frac{1}{d}$, which vanishes since $(I_p - P)c_2(\kappa, u)c_2(\kappa, u)^\top(I_p - P)^\top = 0$.

For each k, let $h_{k,w^\star}$ be the vector in $\mathbb{R}^p$, defined as

$$(h_{k,w^\star})_i \overset{\text{def}}{=} h_k(\sqrt{d}w_i^\top w^\star) \tag{116}$$

with h_k being the kth Hermite polynomial. By substituting $(w_i^\top w^\star)^2 = \frac{1}{d}(h_2(\sqrt{d}dw_i^\top w^\star) - 1)$

Next, we examine $(B) + (B)_{i\leftrightarrow j}$ and capture its main terms. First, note that $(w_i^\top w_j)w_i^\top(\kappa w^\star)$ can be ignored, since it corresponds to a matrix

$$W \operatorname{diag}\left\{\kappa w_i^\top w^\star\right\} W^\top, \tag{117}$$

whose operator norm is bounded by $\mathcal{O}_\prec(d^{-1/2})$ by Lemma B.12.

Let $\theta_i = W^\star w_i$ for $i \in [p]$ The remaining components in $(B) + (B)_{i\leftrightarrow j}$ can be expressed as:

$$-2c_1(\kappa, u)c_1(\kappa, u)\theta_i\theta_j$$

It is also easy to verify that the last component (D) can be expressed as follows:

$$(D) = c_1(\kappa, u)c_1(\kappa, u)(\kappa^2 + 1)\theta_i\theta_j, \tag{118}$$

The terms B and D combine to give the component:

$$(\kappa^2 - 1)c_1(\kappa, u)c_1(\kappa, u)^\top \odot \boldsymbol{\theta}\boldsymbol{\theta}^\top \tag{119}$$

Combining with the contribution from A and the diagonal terms in Equation (102), we obtain Equation (84). □

Before moving further, we note down a bound on $R^\star_\kappa$ that will prove useful late:

Proposition D.2. *Let $R^\star_\kappa$ be as defined in Proposition D.1. Then, $\exists \ell \in \mathbb{N}$ such that:*

$$\|R^\star_\kappa\| = \mathcal{O}_\prec(\kappa^\ell) \tag{120}$$

.

Proof. The above is a direct consequence of Proposition D.1 and Assumption 2.3, with the later implying that $\left|c_o(\kappa, u)\right| \leq K|u\kappa|$ for some constant K while $\left|c_j(\kappa, u)\right|$ are uniformly bounded for $j = 1, 2, 3$. □

D.2 First-Stage Equivalent

We now proceed with establishing the first stage of deterministic equivalence, aiming to remove the randomness w.r.t X. Concretely, our goal in this section is to construct a sequence of matrices $\mathcal{G}^e_W$ dependent on W but not X such that $\mathcal{G}^e_W$ is deterministically equivalent to $\mathcal{G}^e$ for all suitably-bounded linear functions independent of X (but possibly depending on W). Ideally $\mathcal{G}^e_W$ will possess a simpler structure than $\mathcal{G}^e$, allowing us to further simplify it in the next stage.

For subsequent usage, we shall establish a more quantitative version of deterministic equivalent than Theorem 4.5, with an explicit error bound of $\mathcal{O}_\prec(\frac{1}{\sqrt{d}})$.

Proposition D.3. *Let X_d, W_d be a sequence of random matrices generated as in Assumption 2.1. Let $z \in \mathbb{C}/\mathbb{R}^+$ be arbitrary. Let $\mathcal{G}_W(z)$ denote the the following sequence of random matrices dependent only on W:*

$$\mathcal{G}_W(z) := \left(\frac{\alpha}{p}\Sigma^\star + \zeta L - zI\right)^{-1}, \tag{121}$$

where:

$$\chi_d(\kappa) \overset{\textit{def}}{=} \frac{\mathbb{E}_{W,X}\left[\operatorname{Tr}(G^\mu_e R^\star_\kappa)\right]}{p}, \tag{122}$$

and the expectation is w.r.t both W, X conditioned on the projection κ_μ along the spike $\boldsymbol{v}$ as fixed. Since G^μ_e is independent of κ, $\chi_d(\kappa)$ is a fixed function of κ alone. The subscript d is to remind us that $\chi_d(\kappa)$ defined as in Equation (122) is an unknown dimension-dependent function.

We further define:

$$\Sigma^\star = \mathbb{E}_{\kappa\sim\mathcal{N}(0,1)}\left[\frac{1}{1+\chi(\kappa)}R^\star(\kappa_\mu)\right], \tag{123}$$

with $R^\star(\kappa)$ denoting the matrix:

$$R^\star(\kappa) = \begin{pmatrix} R^\star(\kappa)_{11} & R^\star(\kappa)_{12} \\ R^\star(\kappa)_{21} & R^\star(\kappa)_{22} \end{pmatrix} \tag{124}$$

and $R^\star(\kappa)_{11}, R^\star(\kappa)_{12}, R^\star(\kappa)_{21}, R^\star(\kappa)_{22}$ are as defined in Proposition D.1.

Then for any sequence of matrices A, possibly dependent on W, but not X:

$$\left|\mathrm{Tr}\{A\mathcal{G}_W(z)\} - \mathrm{Tr}\{AG_e(z)\}\right| = \mathcal{O}_\prec(\frac{\|A\|_{\mathrm{Tr}}}{\sqrt{d}}) \tag{125}$$

Proof. We start by defining:

$$\chi_{X,W}(\mu) := \frac{1}{p}(\phi^e_\mu)^\top G^\mu_e \phi^e_\mu, \tag{126}$$

Note that unlike $\chi(\kappa)$ which is solely a function of κ, $\chi_{X,W}(\mu)$ is a random variable depending on X, W, z_μ. We further define:

$$R_\mu = \mathbb{E}\left[\phi^e_\mu(\phi^e_\mu)^\top | \kappa = \kappa_\mu\right] \tag{127}$$

We proceed through a leave-one-out argument, analogous to the one for Wishart-type matrices [Bai and Zhou, 2008]. By definition, G_e satisfies:

$$G_e(\frac{1}{p}\sum_{\mu=1}^n \phi^e_\mu(\phi^e_\mu)^\top - zI) = I. \tag{128}$$

Let $G^\mu_e := (\sum_{\nu\neq\mu}\phi^e_\mu(\phi^e_\mu)^\top + \lambda I)^{-1}$ for $\mu\in[n]$ denote the resolvent with the sample μ removed.

Using the Sherman-Morrison formula, we can express G_e for any $\mu\in[n]$ as:

$$G_e = G^\mu_e - \frac{1}{1+\frac{1}{p}(\phi^e_\mu)^\top G^\mu_e \phi^e_\mu}\phi^e_\mu(\phi^e_\mu)^\top. \tag{129}$$

Next, for each $\mu\in[n]$, we replace G_e in the term $G_e(\phi^e_\mu(\phi^e_\mu)^\top)$ by the above decomposition to obtain:

$$\sum_{\mu=1}^n \frac{1}{p}(G^\mu_e - \frac{1}{1+\frac{1}{p}(\phi^e_\mu)^\top G^\mu_e \phi^e_\mu)}\phi^e_\mu(\phi^e_\mu)^\top)\phi^e_\mu(\phi^e_\mu)^\top - zIG_e = I. \tag{130}$$

The first set of terms simplify as:

$$\begin{aligned} G_e(\phi^e_\mu(\phi^e_\mu)^\top) &= (G^\mu_e - \frac{1}{1+\frac{1}{p}(\phi^e_\mu)^\top G^\mu_e \phi^e_\mu}G^\mu_e\phi^e_\mu(\phi^e_\mu)^\top G^\mu_e)\phi^e_\mu(\phi^e_\mu)^\top \\ &= \frac{1}{1+\chi_{X,W}(\mu)}G^\mu_e\phi^e_\mu(\phi^e_\mu)^\top. \end{aligned}$$

Substituting in (130) yields:

$$\frac{1}{p}\sum_{\mu=1}^n \frac{1}{1+\chi_{X,W}(\mu)}G^\mu_e\phi^e_\mu(\phi^e_\mu)^\top - zIG_e = I, \tag{131}$$

By concentration of $\chi_{X,W}(\mu)$ and averaging over $\phi^e_\mu(\phi^e_\mu)^\top$, we expect the first term to be well-approximated by $\sum_{\mu\in[n]}(\frac{1}{1+\chi_d(\kappa)}G^\mu_e\phi^e_\mu(\phi^e_\mu)^\top)$. We therefore, isolate the errors in $\chi_{X,W}(\mu)$ to obtain:

$$\frac{1}{p}\sum_{\mu\in[n]}(\frac{1}{1+\chi_d(\mu)}G^\mu_e\phi^e_\mu(\phi^e_\mu)^\top + \mathcal{E}_{1,\mu}G^\mu_e\phi^e_\mu(\phi^e_\mu)^\top + \mathcal{E}_{2,\mu}G^\mu_e\phi^e_\mu(\phi^e_\mu)^\top + \mathcal{E}_{3,\mu}G^\mu_e R^\star(\kappa_\mu)) - zG_e = I. \tag{132}$$

where $\mathcal{E}_{1,\mu}, \mathcal{E}_{2,\mu}, \mathcal{E}_{3,\mu}$ account for the error in $\chi_{X,W}(\mu)$ due to randomness over ϕ_μ, approximation of covariance R_μ by $R^\star_\mu$ and the concentration of $\operatorname{Tr}\left(G^\mu_e R^\star_\mu\right)$ to $\chi(\kappa)$ respectively.

$$\mathcal{E}_{1,\mu} = \frac{(\phi^e_\mu)^\top G^\mu_e \phi^e_\mu) - \frac{\operatorname{Tr}\left(G^\mu_e R_\mu\right)}{p}}{(1+\frac{1}{p}(\phi^e_\mu)^\top G^\mu_e \phi^e_\mu)(1+\frac{\operatorname{Tr}\left(G^\mu_e R_\mu\right)}{p})}$$
$$\mathcal{E}_{2,\mu} = \frac{\frac{\operatorname{Tr}\left(G^\mu_e R_\mu\right)}{p} - \frac{\operatorname{Tr}\left(G^\mu_e R^\star_\mu\right)}{p}}{(1+\frac{\operatorname{Tr}\left(G^\mu_e R^\star_\mu\right)}{p})(1+\frac{\operatorname{Tr}\left(G^\mu_e R^\star_\mu\right)}{p})}$$
$$\mathcal{E}_{3,\mu} = \frac{\frac{\operatorname{Tr}\left(G^\mu_e R^\star_\mu\right)}{p} - \chi(\kappa)}{(1+\chi(\kappa))(1+\frac{\operatorname{Tr}\left(G^\mu_e R^\star_\mu\right)}{p})}$$

Next, due to the average over n samples and the small error in replacing G^μ_e by G_e, we further expect $\sum_{\mu\in[n]}(\frac{1}{1+\chi(\kappa)} G^\mu_e \phi^e_\mu (\phi^e_\mu)^\top)$ to be well-approximated (in the sense of deterministic equivalence) by $G_e(\sum_{\mu\in[n]} \frac{1}{1+\chi(\kappa)} R^\star_\mu)$.

Therefore, we explicitly introduce the above term into Equation (132) at the cost of additional errors:

$$\frac{1}{p} G_e(\sum_{\mu\in[n]} \frac{1}{1+\chi(\kappa)} R^\star_\mu) - zG_e = I + \Delta, \tag{133}$$

where:

$$\Delta = \sum_{\mu\in[n]} \Delta_\mu, \tag{134}$$

where:

$$\Delta_\mu = \frac{1}{p}\left(\frac{1}{1+\chi_W(\mu)} G_e R^\star_\mu - G^\mu_e \frac{1}{1+\chi_W(\mu)} \phi^e_\mu (\phi^e_\mu)^\top - G^\mu_e \mathcal{E}_{1,\mu} \phi^e_\mu (\phi^e_\mu)^\top - G^\mu_e \mathcal{E}_{2,\mu} \phi^e_\mu (\phi^e_\mu)^\top - G^\mu_e \mathcal{E}_{3,\mu} \phi^e_\mu (\phi^e_\mu)^\top\right) \tag{135}$$

With the above error terms defined, we're now ready to analyze the linear functional $\operatorname{Tr}\left(AG_e(z)\right)$ in Equation (125). We perform the analysis in two-stages:

- Show that $\operatorname{Tr}\left\{AG_e(z)\right\} \to \operatorname{Tr}\left\{A\mathcal{G}_w(z)\right\}$.
- Show that $\operatorname{Tr}\left\{AG_e(z)\right\}$ concentrates around its expectation.

We start with the first stage. We have:

$$\mathbb{E}\left[\operatorname{Tr}(AG_e)\right] = \mathbb{E}\left[\frac{1}{p}\operatorname{Tr}\left\{(A(I+\Delta)(\alpha R^\star - zI)^{-1})\right\}\right], \tag{136}$$

where $R^\star := \sum_{\mu\in[n]} \frac{1}{1+\chi_W(\mu)} R_\mu$.

Our next goal is therefore to bound the contribution from Δ.

The independence of G^μ_e and ϕ^e_μ, implies that:

$$\mathbb{E}\left[G^\mu_e \frac{1}{1+\chi_W(\mu)} \phi^e_\mu (\phi^e_\mu)^\top\right] = \mathbb{E}\left[G^\mu_e \frac{1}{1+\chi_W(\mu)} R_\mu\right] \tag{137}$$

Thus, we have:

$$\begin{aligned}\mathbb{E}\left[\frac{1}{p}\operatorname{Tr}\left(A\Delta_\mu\right)\right] &= \underbrace{\mathbb{E}\left[\operatorname{Tr}\left(A(\mathcal{E}_{1,\mu}+\mathcal{E}_{2,\mu}+\mathcal{E}_{3,\mu})G^\mu_e \phi^e_\mu (\phi^e_\mu)^\top\right)\right]}_{T_1} \\ &+ \underbrace{\frac{1}{1+\chi_W(\mu)}\frac{1}{p}\mathbb{E}\left[\operatorname{Tr}\left(AG_e R^\star_\mu\right)\right] - \frac{1}{p}\frac{1}{1+\chi_W(\mu)}\mathbb{E}\left[\operatorname{Tr}\left(AG^\mu_e R_\mu\right)\right]}_{T_2}\end{aligned} \tag{138}$$

We start with the term T_1. We proceeed by bounding the contributions from $\mathcal{E}_{1,\mu}, \mathcal{E}_{2,\mu}, \mathcal{E}_{3,\mu}$.

First, note that cyclicity of the trace implies that for $i = 1, 2, 3$:

$$\operatorname{Tr}\left(\frac{1}{p}A(\mathcal{E}_{i,\mu})G_e^\mu \phi_\mu^e (\phi_\mu^e)^\top\right) = (\mathcal{E}_{i,\mu})\frac{1}{p}\operatorname{Tr}\Big((\phi_\mu^e)^\top A G_e^\mu \phi_\mu^e\Big) \tag{139}$$

Our strategy is to bound the tail behavior of $\left|\mathcal{E}_{i,\mu}\right|, \operatorname{Tr}\left(\frac{1}{p}(\phi_\mu^e)^\top A G_e^\mu \phi_\mu^e\right)$ and translate them into expectation bounds on the product.

Let $\epsilon_{1,\mu} \coloneqq \frac{1}{p}(\phi_\mu^e)^\top G_e^\mu (\phi_\mu^e) - \frac{\operatorname{Tr}\left(G_e^\mu R_\mu\right)}{p}, \epsilon_{2,\mu} \coloneqq \frac{\operatorname{Tr}\left(G_e^\mu R_\mu\right)}{p} - \frac{\operatorname{Tr}\left(G_e^\mu R_\mu^\star\right)}{p}, \epsilon_{3,\mu} \coloneqq \frac{\operatorname{Tr}\left(G_e^\mu R_\mu^\star\right)}{p} - \chi_d(\kappa)$. Then:

$$\begin{aligned}
\mathcal{E}_{1,\mu} &= \frac{\epsilon_{1,\mu}}{(1 + \frac{\operatorname{Tr}\left(G_e^\mu R_\mu\right)}{p} + \epsilon_{1,\mu})(1 + \frac{\operatorname{Tr}\left(G_e^\mu R_\mu\right)}{p})} \\
&= \frac{\epsilon_{1,\mu}}{(1 + \frac{\operatorname{Tr}\left(G_e^\mu R_\mu\right)}{p})^2} + \frac{\epsilon_{1,\mu}^2}{(1 + \frac{\operatorname{Tr}\left(G_e^\mu R_\mu\right)}{p} + \epsilon_{1,\mu})(1 + \frac{\operatorname{Tr}\left(G_e^\mu R_\mu\right)}{p})},
\end{aligned}$$

and analogous relations hold for $\mathcal{E}_{2,\mu}, \mathcal{E}_{3,\mu}$.

We start with bounding the term $\operatorname{Tr}\left(\frac{1}{p}(\phi_\mu^e)^\top A G_e^\mu \phi_\mu^e\right)$. Since $\operatorname{Tr}\left(\frac{1}{p}(\phi_\mu^e)^\top A G_e^\mu \phi_\mu^e\right)$ is a quadratic form in the features ϕ_μ^e, one could hope to exploit a Hanson-Wright type inequality. This requires ϕ_μ^e to be “well-concentrated" in some sense. Fortunately, assumption 2.3 directly yields the following regularity property:

Proposition D.4. *Let $\xi_\mu \sim \mathcal{N}(0, I_d)$ be defined through the decomposition:*

$$x_\mu = \kappa w^\star + (I - w^\star(w^\star)^\top)\xi_\mu \tag{140}$$

Then, w.h.p as $d \to \infty$, The map $\xi_\mu \to \phi_\mu^e$ is Lipschitz-continuous

The Lipschitz-continuity of ϕ_μ allows us to apply a generalization of the Hanson-wright inequality for Lipschitz-maps of independent Gaussian i.i.d vectors [Louart and Couillet, 2018], leading to the following result:

Lemma D.5. *For any sequence of matrices $M \in \mathbb{C}^{p\times p}$, independent of ϕ_μ^e:*

$$\left|(\phi_\mu^e)^\top M \phi_\mu^e - \operatorname{Tr}\big(M R_\mu\big)\right| = \mathcal{O}_\prec(\|M\|_F), \tag{141}$$

Applying the above Lemma with $M = A G_e^\mu$ and noting that:

$$\begin{aligned}
\|A G_e^\mu\|_F &\leq \|A G_e^\mu\|_{\operatorname{Tr}} \\
&\leq \|A\|_{\operatorname{Tr}} \|G_e^\mu\| \\
&\leq \frac{1}{\zeta}\|A\|_{\operatorname{Tr}}
\end{aligned}$$

Therefore, Lemma D.5 yields:

$$\operatorname{Tr}\left\{(\frac{1}{p}(\phi_\mu^e)^\top A G_e^\mu \phi_\mu^e)\right\} = \mathcal{O}_\prec(\frac{\|A\|_{\operatorname{Tr}}\kappa^\ell}{p}). \tag{142}$$

We next claim that to obtain the desired bound on $\operatorname{Tr}(A\Delta)$, it suffices to show that:

$$\epsilon_{i,\mu} = \mathcal{O}_\prec(\frac{\kappa^\ell}{\sqrt{d}}), \tag{143}$$

for $i = 1, 2, 3$. To see this, note that $R_\mu^\star$ is positive semi-definite and G_e^μ has a positive Hermitian component. Therefore, the absolute values of the term $\frac{1}{(1+\frac{\operatorname{Tr}\left(G_e^\mu R_\mu\right)}{p}+\epsilon_{1,\mu})(1+\frac{\operatorname{Tr}\left(G_e^\mu R_\mu\right)}{p})}$ in $\mathcal{E}_{1,\mu}$ is uniformly bounded by 1, and similarly for $\mathcal{E}_{2,\mu}, \mathcal{E}_{3,\mu}$. Therefore, for $i = 1, 2, 3$:

$$\mathcal{E}_{i,\mu} = \mathcal{O}_\prec(\epsilon_{i,\mu} + \epsilon_{i,\mu}^2). \tag{144}$$

Applying Cauchy-Shwartz inequality to the sequence of variables $\mathcal{E}_{i,\mu}, (z_\mu^\top A z_\mu)$ yields:

$$\mathbb{E}\left[\left|\mathcal{E}_{i,\mu}, (z_\mu^\top A z_\mu)\right|\right] = \mathcal{O}(\kappa^\ell \frac{1}{p}\frac{1}{\sqrt{d}}), \tag{145}$$

for some $\ell \in \mathbb{N}$ and $i = 1, 2, 3$. Subsequently, combining with $\sup_\mu(\kappa_\mu) = \mathcal{O}(\sqrt{\log n})$ and summing over $\mu = 1 \cdots n$, we obtain:

$$\mathbb{E}\left[T_1\right] = \mathcal{O}(\frac{\text{polylog}\, d}{\sqrt{d}}) \tag{146}$$

In light of the above discussion, we now move to bounding $\epsilon_{i,\mu}$ for $i = 1, 2, 3$:

Proposition D.6. *Let $\epsilon_{1,\mu} := \frac{1}{p}(\phi_\mu^e)^\top G_e^\mu(\phi_\mu^e) - \frac{\text{Tr}\left(G_e^\mu R_\mu\right)}{p}$, $\epsilon_{2,\mu} := \frac{\text{Tr}\left(G_e^\mu R_\mu\right)}{p} - \frac{\text{Tr}\left(G_e^\mu R_\mu^\star\right)}{p}$, $\epsilon_{3,\mu} := \frac{\text{Tr}\left(G_e^\mu R_\mu^\star\right)}{p} - \chi_d(\kappa)$. Then for $i = 1, 2, 3$, $\exists \ell \in \mathbb{N}$ such that:*

$$\epsilon_{i,\mu} = \mathcal{O}_\prec(\frac{\kappa^\ell}{\sqrt{d}}) \tag{147}$$

Proof. We start with $\epsilon_{1,\mu}$. Since $\frac{1}{p}(\phi_\mu^e)^\top G_e^\mu(\phi_\mu^e) - \frac{\text{Tr}\left(G_e^\mu R_\mu\right)}{p}$ corresponds to the deviation of the quadratic form $(\phi_\mu^e)^\top G_e^\mu(\phi_\mu^e)$ from it's mean, we may again apply the generalized Hanson-Right inequality (Lemma D.5). Note that $\|G_e^\mu\|_2 \leq \frac{1}{\zeta}$ by Lemma B.4 implying $\|G_e^\mu\|_F \leq \frac{\sqrt{p}}{\zeta}$. Sine G_e^μ is independent of ϕ_μ^e, Lemma D.5 yields:

$$\epsilon_{1,\mu} = \mathcal{O}_\prec(\frac{\kappa^\ell}{\sqrt{d}}) \tag{148}$$

$\epsilon_{2,\mu}$ is directly bounded through Proposition D.1

Bounding the third term containing $\epsilon_{3,\mu}$ is more challenging and will take up the bulk of the remaining discussion.

To bound $\epsilon_{3,\mu}$, we require establishing the concentration of $\text{Tr}\left(G_e^\mu R_\mu^\star\right)$ around $\chi_d(\kappa)$, both w.r.t X, W (recall the definition of $\chi(\kappa)$ in Equation (122)).

A standard way of achieving this is through the use of Martingale-based arguments [Bai and Zhou, 2008, Cheng and Singer, 2013, Hastie et al., 2022]. We proceed through a similar technique, however, our analysis is complicated by the presence of structured covariance $R_\mu^\star$ and the joint-randomness over X, W.

Lemma D.7. *Almost surely over $\kappa_\mu \sim \mathcal{N}(0, 1)$:*

$$\left|\text{Tr}\left(G_e^\mu R_\mu^\star\right) - \chi_d(\kappa_\mu)\right| = \mathcal{O}_\prec(\frac{\kappa_\mu^\ell}{\sqrt{d}}), \tag{149}$$

Proof. As mentioned above, the proof follows through a martingale argument. Concretely, proceed by succesively applying Burkholder's inequality w.r.t the doob martingales on filtrations generated by X, W respectively. We first start by conditioning on the following high-probability event over W:

$$\mathcal{E}_w = \{W : \|W\|_2 = \mathcal{O}(1), \langle w_i, w_j\rangle = \mathcal{O}_\prec(\frac{1}{\sqrt{d}})\}. \tag{150}$$

It is easy to check that the moments of $\chi_d(\kappa)$ are bounded, and therefore, the error in the expectation upon restriction to $\mathcal{E}_w$ can be bounded by $\mathcal{O}_\prec(\frac{1}{d}^k)$ for arbitrarly large k.

Let $\mathbb{E}_\mu$ denote the conditional expectation w.r.t the sigma-algebra generated by $x_{\mu_{\mu'<\mu}}$. We apply the following martingale decomposition:

$$\begin{aligned}&\text{Tr}(G_e R_\kappa^\star) - \mathbb{E}\left[\text{Tr}(G_e R_\kappa^\star)\right]\\&= \sum_{\mu=1}^n \mathbb{E}_\mu \text{Tr}(G_e R_\kappa^\star) - \mathbb{E}_{\mu-1}\text{Tr}(G_e R_\kappa^\star).\end{aligned}$$

Let $e_\mu = \mathbb{E}_\mu \mathrm{Tr}(G_e R^\star_\kappa) - \mathbb{E}_{\mu-1}\mathrm{Tr}(G_e R^\star_\kappa)$. We have:

$$\mathbb{E}_\mu(\mathrm{Tr}(G_e R^\star_\kappa) - \mathbb{E}_{\mu-1}\mathrm{Tr}(G_e R^\star_\kappa)) = \mathbb{E}_\mu(\mathrm{Tr}(G_e R^\star_\kappa) - \mathrm{Tr}(G^\mu_e R^\star_\kappa)) - \mathbb{E}_{\mu-1}(\mathrm{Tr}(G_e R^\star_\kappa) - \mathrm{Tr}(G^\mu_e R^\star_\kappa)), \tag{151}$$

where we used that $\mathbb{E}_{\mu-1}\mathrm{Tr}(G^\mu_e R^\star_\kappa)) = \mathbb{E}_\mu \mathrm{Tr}(G^\mu_e R^\star_\kappa))$ since G^μ_e does not depend on x_μ

Applying the Sherman-Morrison formula yields:

$$\mathbb{E}_\mu \mathrm{Tr}(G_e R^\star_\kappa) - \mathrm{Tr}(G^\mu_e R^\star_\kappa) = -\frac{1}{p}\mathbb{E}_\mu \frac{1}{1+(\phi^e_\mu)^\top G^\mu_e \phi^e_\mu}(\phi^e_\mu)^\top R^\star_\kappa \phi^e_\mu. \tag{152}$$

Now, $\left|\frac{1}{1+(\phi^e_\mu)^\top G^\mu_e \phi^e_\mu}\right| \leq 1$ while $(\phi^e_\mu)^\top R^\star_\kappa \phi^e_\mu$ has moments bounded polynomially in κ by Lemma D.5.

Therefore, Lemma B.5 combined with Markov's inequality for large enough p, implies:

$$\left|\frac{1}{p}\,\mathrm{Tr}\Big(G^\mu_e R^\star_\mu\Big) - \chi_W(\kappa)\right| = \mathcal{O}_\prec(\frac{1}{\sqrt{d}}) \tag{153}$$

Where:

$$\chi_W(\kappa) \overset{\text{def}}{=} \mathbb{E}_X\left[\mathrm{Tr}(G_e R^\star_\kappa)\right], \tag{154}$$

where the expectation is only w.r.t the matrix X.

It remains to show by establishing concentration w.r.t W that:

$$\big|\chi_W(\kappa) - \chi(\kappa)\big| = \mathcal{O}_\prec(\frac{1}{\sqrt{d}}), \tag{155}$$

We again condition on the following high-probability event:

$$\mathcal{E}_x = \{X : \|W\|_2 = \mathcal{O}(\sqrt{d}), \langle w_i, w_j\rangle = \mathcal{O}_\prec(1)\}. \tag{156}$$

We again apply a martingale argument, except that unlike the concentration w.r.t X, both $G_e, R^\star$ now depend on W.

Let $\mathbb{E}_i$ denote the conditional expectation w.r.t the sigma-algebra generated by $w_{j_{j<i}}$. Let $G_e(-i)$ denote the extended resolvent obtained after the removal of the $(k+1+i)_{\text{th}}$ row and column in $(\Phi^e)^\top\Phi^e$, except for the diagonal $(k+1+i),(k+1+i)$ entry. Note that this corresponds to a finite-rank perturbation:

$$\frac{1}{p}(\Phi^e)^\top\Phi^e - e_{(k+1+i)}b_i^\top - b_i e^\top_{(k+1+i)}, \tag{157}$$

where $b_i^\top$ denotes the (normalized) $(k+1+i)_{\text{th}}$ row of $(\Phi^e)^\top\Phi^e$ given by:

$$b_i^\top = \frac{1}{p}(\Phi^e_{i:})^\top(\Phi^e) \tag{158}$$

Analogously, let $R^\star_\kappa(-i)$ be obtained by the removal of the $(k+1+i)_{\text{th}}$ row and column in $R^\star_\kappa$ (except for the diagonal $(k+1+i),(k+1+i)$ entry).

Using the Woodbury-matrix identity, we have:

$$G_e = G_e(-i) - G_e(-i)\begin{bmatrix} e_{(k+1+i)} & b_i\end{bmatrix}\Big(\begin{bmatrix}0 & 1\\ 1 & 0\end{bmatrix} + \Psi\Big)^{-1}\begin{bmatrix} e_{(k+1+i)} & b_i\end{bmatrix}^\top G_e(-i), \tag{159}$$

where $\Psi = \begin{bmatrix} e_{(k+1+i)} & b_i\end{bmatrix}^\top G_e(-i)\begin{bmatrix} e_{(k+1+i)} & b_i\end{bmatrix} \in \mathbb{R}^{2\times 2}$

Let $e_i = \mathbb{E}_i\mathbb{E}_x\left[\mathrm{Tr}(G_e R^\star_\kappa)\right] - \mathbb{E}_{i-1}\mathbb{E}_x\left[\mathrm{Tr}(G_e R^\star_\kappa)\right]$. We express $\mathrm{Tr}(G_e R^\star_\kappa)$ as:

$$\mathrm{Tr}(G_e R^\star_\kappa) = \mathrm{Tr}\big(G_e(-i)R^\star_\kappa(-i)\big) + \Delta,$$

where Δ arises from the low-rank projections.

To Apply Lemma B.5, it remains to bound the second moments of the residual terms Δ. It is easy to check that conditioned on the high-probability events $\mathcal{E}_w, \mathcal{E}_x$, for any uniformly Lipschitz $f : \mathbb{R}^p \to \mathbb{R}$, $f(\frac{1}{\sqrt{p}}b_i)$ is a uniformly Lipschitz map of w_i. Therefore, Lemma B.11 implies that $\sqrt{p}b_i$ is a Lipschitz-concentrated vector in the sense of Definition 9 in Couillet and Liao [2022]. Therefore the generalized Hanson-Wright inequality implies that:

$$\left| b_i^\top G_e(-i) b_i - \mathbb{E}\left[b_i^\top G_e(-i) b_i\right] \right| = \mathcal{O}_\prec(\frac{1}{\sqrt{d}}). \tag{160}$$

Similarly, we obtain that the remaining entries of Ψ concentrate around their expectations. As a result, δ decomposes into quadratic forms in $e_{(k+1+i)}, b_i$ which can be bounded as $\mathcal{O}_\prec(\frac{1}{p})$. Applying Lemma B.5 then completes the proof. □

The proof of Lemma D.7 completes the proof of Proposition D.6 and thus bounds T_1 in Equation (138) □

We now return to bounding the term T_2 in Equation (138). Let ζ be as defined in Lemma B.4 First, Proposition D.1 and Lemmas B.4, B.8 imply that:

$$\left| \mathrm{Tr}\big(AG_e^\mu R_\mu\big) - \mathrm{Tr}\Big(AG_e^\mu R_\mu^\star\Big) \right| \leq \mathcal{O}_\prec(\frac{1}{\sqrt{d}}). \tag{161}$$

Therefore, we may replace R_μ by $R_\mu^\star$ at the cost of an error $\mathcal{O}_\prec(\frac{1}{\sqrt{d}})$.

Next, we apply the Sherman-Morrison formula to $G_e - G_e^\mu$ to obtain

$$\mathbb{E}\left[\mathrm{Tr}\Big(AG_e R_\mu^\star\Big)\right] - \mathbb{E}\left[\mathrm{Tr}\Big(AG_e^\mu R_\mu^\star\Big)\right] = \frac{1}{p}\mathbb{E}\left[\frac{1}{1+(\phi_\mu^e)^\top G_e^\mu \phi_\mu^e}\mathrm{Tr}\Big(A\phi_\mu^e(\phi_\mu^e)^\top R_\mu^\star\Big)\right] \tag{162}$$

By the cyclicity of trace, the second term can be expressed as the quadratic form $(\phi_\mu^e)^\top R_\mu^\star A \phi_\mu^e$

We start by noting that, Lemmas B.4 B.8 imply:

$$\mathbb{E}\left[\mathrm{Tr}\big(AG_e R_\mu\big)\right] \leq \left\|AG_e R_\mu\right\|_{\mathrm{Tr}} \leq \|A\|_{\mathrm{Tr}}\left\|G_e R_\mu\right\| \leq \frac{C}{\zeta}, \tag{163}$$

for some constant C.

Therefore, using Lemmas B.3 and B.8, we obtain::

$$\left| \frac{1}{1+\chi_W(\mu)}\mathbb{E}\left[\mathrm{Tr}\big(AG_e R_\mu\big)\right] - \frac{1}{1+\chi_W(\mu)}\mathbb{E}\left[\mathrm{Tr}\big(AG_e^\mu R_\mu\big)\right] \right| = \mathcal{O}_\prec(\frac{\|A\|_{\mathrm{Tr}}}{p}). \tag{164}$$

We therefore obtain:

$$\mathbb{E}\left[\mathrm{Tr}\big(A\Delta_\mu\big)\right] = \mathcal{O}(\frac{1}{\sqrt{d}}) \tag{165}$$

The convexity of || then implies that

$$\mathbb{E}\left[\mathrm{Tr}(AG_e)\right] = \mathbb{E}\left[\mathrm{Tr}\Big\{\big(A(\alpha R^\star + \lambda I)^{-1}\big)\Big\}\right] + \mathcal{O}(\frac{\mathrm{polylog}\, d}{\sqrt{d}}). \tag{166}$$

Next, we show that the averaging over κ allows us to replace $R^\star$ with $\Sigma^\star$:

Lemma D.8. *There exists $\ell \in \mathbb{N}$ such that almost surely over κ:*

$$\chi_\mu(\kappa) = \mathcal{O}(\kappa^\ell) \tag{167}$$

Proof. Recall that G_μ^e doesn't depend on κ, while $R_\mu^\star$ depends on κ only through scalars, $c_j(\kappa, u_i)$ for $j = 1, 2, 3$. The claim then directly follows using assumption 2.3. □

Since κ_μ for $\mu \in [n]$ are independent Gaussians, the above Lemma implies that the covariance $\sum_{\mu\in[n]} \frac{1}{1+\chi_W(\mu)} R^\star(\kappa_\mu)$ can further by replaced by $\Sigma^\star$ in proposition D.3 with additional error $\mathcal{O}(\frac{1}{\sqrt{d}})$.

It remains to show that $\mathrm{Tr}(G_e A)$ concentrates around its expectation w.r.t X. This follows from a martingale argument identical to the first part of the proof of Lemma D.7, with the role of $R^\star$ being replaced by A. Then, we have the following martingale decomposition:

$$\mathrm{Tr}(G_e A) - \mathbb{E}\left[\mathrm{Tr}(G_e A)\right] = \sum_{\mu=1}^{n} \mathbb{E}_\mu \mathrm{Tr}(G_e A) - \mathbb{E}_{\mu-1}\mathrm{Tr}(G_e A).$$

Let $e_\mu = \mathbb{E}_\mu \mathrm{Tr}(G_e A) - \mathbb{E}_{\mu-1}\mathrm{Tr}(G_e A)$. We have:

$$\mathbb{E}_\mu(\mathrm{Tr}(G_e A) - \mathbb{E}_{\mu-1}\mathrm{Tr}(G_e A)) = \mathbb{E}_\mu(\mathrm{Tr}(G_e A) - \mathrm{Tr}(G_e^\mu A)) - \mathbb{E}_{\mu-1}(\mathrm{Tr}(G_e A) - \mathrm{Tr}(G_e^\mu A)) \tag{168}$$

where G_e^μ denotes the resolvent with the μ_{th} example removed. Next, exactly as in Lemma D.7, we apply the generalized Hanson-Wright inequality to show that:

$$(\mathrm{Tr}(G_e A) - \mathrm{Tr}(G_e^\mu A)) = \mathcal{O}_\prec(\frac{1}{p}). \tag{169}$$

Lemma B.5 for $p = 4$ and Markov's inequality then imply that, with high-probability:

$$\mathrm{Tr}(G_e A) = \mathbb{E}\left[\mathrm{Tr}(G_e A)\right] + \mathcal{O}_\prec(\frac{\|A\|_{\mathrm{Tr}}}{\sqrt{d}}), \tag{170}$$

completing the proof of Proposition D.3

□

D.3 Second stage of Deterministic Equivalent (Averaging over W)

Substituting $R^\star_\mu$ from Proposition D.1, we obtain that Σ^* in proposition D.3 posseses the following structure:

$$\Sigma^* = \begin{pmatrix} \boldsymbol{A}^*_{11} & (\boldsymbol{A}^*_{21})^\top \odot \boldsymbol{\theta}^\top \\ \boldsymbol{\theta} \odot \boldsymbol{A}^*_{21} & {\scriptstyle (\boldsymbol{V}_e \odot \boldsymbol{W}\boldsymbol{W}^\top + \mathrm{diag}(\nu_e))) + \alpha S_e \odot \theta\theta^\top,} \end{pmatrix} \tag{171}$$

where:

$$\begin{aligned} \boldsymbol{A}^*_{11} &= \mathbb{E}_\kappa\left[\frac{\alpha}{1+\chi(z;\kappa)} R^\star_{11}(\kappa)\right] \\ &= \mathbb{E}_\kappa\left[\frac{\alpha}{1+\chi(z;\kappa)} \begin{bmatrix} \sigma^2_\star(\kappa) & \sigma_\star(\kappa)c_0(\kappa,u_\pi)^\top \\ \sigma_\star(\kappa)c_0(\kappa,u_\pi) & c_0(\kappa,u_\pi)c_0(\kappa,u_\pi)^\top \end{bmatrix}\right] \end{aligned}$$

$$\boldsymbol{A}^*_{11} = \mathbb{E}_\kappa\left[\frac{\alpha}{1+\chi(z;\kappa)} R^\star_{21}(\kappa)\right]$$

and $\boldsymbol{V}, \nu_e, S$ are defined as:

$$\begin{aligned} V^{(d)}_{i,j}(\zeta) &= \mathbb{E}_\kappa\left[\alpha \frac{(c_1(\kappa,u_i)c_1(\kappa,u_j)}{1+\chi(s,m)}\right] \\ \nu^{(d)}_i(\rho) &= \mathbb{E}_\kappa\left[(\sum_{k\geq 2}\frac{\alpha c_k^2(\kappa,u_i)}{1+\chi(\kappa,W)}\right] \\ S &= \mathbb{E}\left[(\kappa^2-1)\mathbb{E}_\kappa\left[\frac{(c_1(\kappa,u_i)c_1(\kappa,u_j)}{1+\chi(\kappa)}\right]\right], \end{aligned}$$

with $\kappa \sim \mathcal{N}(0,1)$ throughout.

Note that $\boldsymbol{A}^*_{11}, \boldsymbol{A}^*_{21}$ are deterministic, while the terms dependent on $\boldsymbol{\theta}$, including δ contribute finite-rank spikes. Therefore, the bulk statistics of $\Sigma^\star$ arise out of the term $\boldsymbol{V}_e \odot \boldsymbol{W}\boldsymbol{W}$.

To average-out the randomness over $\boldsymbol{W}$ in , it will be convenient to first extract a deterministic equivalent for the following matrix:

$$M^* = (\boldsymbol{V}_e^{(d)} \odot \boldsymbol{W}\boldsymbol{W}^\top + \mathrm{diag}(\nu_e^{(d)}) - z\boldsymbol{I}_p)^{-1} \tag{172}$$

where $V_e, \mathrm{diag}(\nu^{(d)})_e$ denote the extended block-structured matrices as per definition 4.3.

Additionally, to express $\chi_W(\kappa)$ self-consistently in terms of V, ν, we will require the following additional functional:

$$1/d * \mathrm{Tr}\big(e_i \odot WM^*W \odot e_j\big) \tag{173}$$

Lemma D.9 (Deterministic Equivalent for Block-Structured Wishart)**.** *Let $C(z) \in \mathbb{C}^{k\times k}$, $D(z) \in \mathbb{C}^k$ be analytic mappings such that $C(z) : \mathbb{C}^+ \to \mathbb{C}^-$ and $D(z) : \mathbb{C}^+ \to \mathbb{C}^-$ entry-wise with $\left|C_{i,j}\right|, |D_i|$ uniformly bounded by some constant independent of ζ. Furthermore, suppose that $D(z)$ is diagonal. Let $C_e(z), D_e \in \mathbb{C}^{p\times p}$ denote the extended matrices as defined in Definition 4.3. Let $R_{C,D}$ denote the block structured resolvent defined as:*

$$R_{C,D} \overset{def}{=} ((C_e) \odot \tilde{W}^0(\tilde{W}^0)^\top + \mathrm{diag}(D_e) - zI_p)^{-1}. \tag{174}$$

Define $\mathcal{M}^*(C_e, D_e)$ as the diagonal matrix:

$$\mathcal{M}^*(C_e, D_e) := \mathrm{diag}(\frac{b^\star}{\pi\beta}), \tag{175}$$

where above, $b^\star, \pi \in \mathbb{R}^k$ are divided element-wise and $b^\star$ satisfies the following self-consistent equation:

$$b^\star(C,D)_q = \pi_{u_i}\beta((C^{-1} + \mathrm{diag}(b^\star))^{-1} + (\mathrm{diag}(D) - zI_p))^{-1}_{q,q}, \tag{176}$$

for $q \in [p]$. Then for any sequence of matrices $A \in \mathbb{C}^{p\times p}$:

$$\left|\mathrm{Tr}\{A\mathcal{M}^\star\} - \mathrm{Tr}\big\{AR_{C,D}\big\}\right| = \mathcal{O}_\prec(\frac{\|A\|_{\mathrm{Tr}}}{\sqrt{d}}) \tag{177}$$

Furthermore, the expression:

$$K^*_{i,j}(C,D) = 1/d * \mathrm{Tr}\Big(e^i \odot WM^*W \odot e^j\Big) \tag{178}$$

satisfies the following deterministic equivalence:

$$K^* = \psi(C,D) + \mathcal{O}_\prec(\frac{1}{\sqrt{d}})$$

where $\psi(C,D)$ is defined as:

$$\psi(C,D) = b^\star(C,D) - L(C,D) \odot (b^\star(C,D)(b^\star(C,D))^\top), \tag{179}$$

with:

$$L(C,D) = \Big((C)^{-1} + \mathrm{diag}(b^\star(C,D))\Big)^{-1}. \tag{180}$$

Proof. The proof relies on the observation that due to the block structure in V, removing a coordinate in $\boldsymbol{W}\boldsymbol{W}^\top$ results in a finite rank perturbation to $M^\star$. We then proceed through a leave-one out argument similar to section D. The proof is deferred to section F. □

In light of Lemma D.9, we obtain the following candidate for the deterministic equivalent to $G_w(z)$:

$$(\frac{\alpha}{p}\tilde{\Sigma}^\star - zI), \tag{181}$$

where $\tilde{\Sigma^\star}$ is obtained by replacing $(\boldsymbol{V}_e^{(d)} \odot \boldsymbol{W}\boldsymbol{W}^\top + \mathrm{diag}(\nu_e^{(d)}) + \lambda \boldsymbol{I}_p) = (M^*)^{-1}$ by $\mathcal{M}^\star(V, \mathrm{diag}(\nu))^{-1}$.

However, while obtaining $\mathcal{M}^\star(V, \mathrm{diag}(\nu))^{-1}$, we averaged over the dependence on $\theta = Ww^\star$ in addition to the remaining components of W. This is undesirable since the dependence on θ captures the correlations amongst blocks Σ_{22}, Σ_{21} in the extended resolvent which will be relevant for the characterization of the generalization error. We obtain back this dependence through in the following result:

Lemma D.10. *Let $M_d^\star = (\boldsymbol{V}_e^{(d)} \odot \boldsymbol{W}\boldsymbol{W}^\top + \mathrm{diag}(\nu_e^{(d)}) + \lambda \boldsymbol{I}_p)^{-1}$. Define:*

$$M_d^\theta := (\mathrm{diag}(\frac{b^\star}{\pi\beta})_e - (\boldsymbol{V}_d^{-1} + \mathrm{diag}(b^\star))^{-1} \odot (\frac{b^\star}{\pi\beta})_e (\frac{b^\star}{\pi\beta})_e^\top \odot \theta\theta^\top), \tag{182}$$

where $\psi(V_d, \nu_d)$ is as defined in Lemma D.9. For any sequence of matrices $A \in \mathbb{C}^{p \times p}$, possibly dependent on $\theta = Ww^\star$:

$$\mathrm{Tr}\{(M_d^\star A)\} = \mathrm{Tr}\Big\{(M_d^\theta A)\Big\} + \mathcal{O}_\prec(\frac{\|A\|_{\mathrm{Tr}}}{\sqrt{d}}). \tag{183}$$

Proof. To obtain back the dependence on θ, we write:

$$M_d^* = (\boldsymbol{V}_e^{(d)} \odot (\boldsymbol{W}_\perp (\boldsymbol{W}_\perp)^\top) + \boldsymbol{V}_e^{(d)} \odot \theta\theta^\top) + \mathrm{diag}(\nu_e^{(d)}) + \lambda \boldsymbol{I}_p)^{-1}, \tag{184}$$

where $\boldsymbol{W}_\perp = \boldsymbol{W} - \theta(w^\star)^\top$ denotes the components of the weights upon the removal of the components θ along $w^\star$.

Next, note that $\boldsymbol{V}_e^{(d)} \odot \theta\theta^\top$ is a finite-rank perturbation along directions $e^1, \cdots, e^k$ defined in Equation (13). Concretely, let $E \in \mathbb{R}^{p \times k}$ denote the matrix with columns $e^1, \cdots e^k$ and define $E_\theta = \theta \odot E$. Then, $\boldsymbol{V}_e^{(d)} \odot \theta\theta^\top$ can be expressed as:

$$\boldsymbol{V}_e^{(d)} \odot \theta\theta^\top = E_\theta \boldsymbol{V}_d E_\theta^\top. \tag{185}$$

Therefore, we apply the Woodbury matrix identity to obtain:

$$M^* = \tilde{M} - \tilde{M} E_\theta (\boldsymbol{V}_d^{-1} + E_\theta^\top \tilde{M} E_\theta)^{-1} E_\theta \tilde{M} \tag{186}$$

, where $\tilde{M} = (\boldsymbol{V}_e^{(d)} \odot (\boldsymbol{W}_\perp (\boldsymbol{W}_\perp)^\top) + \mathrm{diag}(\nu_e^{(d)}) - z\boldsymbol{I}_p)^{-1}$ denotes the inverse upon the removal of components alog θ. By the rotational invariance of w_i, $\tilde{M}$ asymptotically shares the deterministic equivalent for $M^\star$ described in Lemma D.9. Since $\tilde{M}$ is independent of θ, and θ_i are asymptotically distributed as $\mathcal{N}(0, \frac{1}{d})$, $E_\theta^\top \tilde{M} E_\theta$ in turn simplifies to $\mathrm{diag}(b^\star(V_d, \nu_d))$. We then replace the occurances of $\tilde{M}$ with $\mathcal{M}^\star(V_d, \nu_d) = \frac{b^\star}{\pi\beta}$ to obtain:

$$\mathrm{diag}(\frac{b^\star}{\pi\beta}) - \mathrm{diag}(\frac{b^\star}{\pi\beta})_e E_\theta (\boldsymbol{V}_d^{-1} + \mathrm{diag}(b^\star))^{-1} E_\theta \, \mathrm{diag}(\frac{b^\star}{\pi\beta})_e. \tag{187}$$

Subsequently, analogous to Equation (185), we have the following equivalent representation of the middle-block:

$$E_\theta (\boldsymbol{V}_d^{-1} + \mathrm{diag}(b^\star))^{-1} E_\theta = (\boldsymbol{V}_d^{-1} + \mathrm{diag}(b^\star))_e^{-1} \odot \theta\theta^\top. \tag{188}$$

We thus obtain:

$$M^\star \simeq \mathrm{diag}(\frac{b^\star}{\pi\beta}) - (\boldsymbol{V}_d^{-1} + \mathrm{diag}(b^\star))^{-1} \odot (\frac{b^\star}{\pi\beta})_e (\frac{b^\star}{\pi\beta})_e^\top \odot \theta\theta^\top \tag{189}$$

where $\simeq$ denotes deterministic equivalence in the sense of Equation (183), which follows from proposition D.9. □

We can finally claim the second level of deterministic equivalence, by replacing $\mathcal{G}_W(z)$ with a sequence of matrices depending only on θ, u:

Proposition D.11. *Consider the sequence of (random) resolvents defined by:*

$$\mathcal{G}_W(z) = \left(\frac{\alpha}{p}\Sigma^\star - zI\right)^{-1}, \tag{190}$$

, and:

$$\tilde{\mathcal{G}}_e(z) = \begin{pmatrix} \boldsymbol{A}_{11}^* - zI_{k+1} & (\boldsymbol{A}_{21}^*)^\top \odot \boldsymbol{\theta}^\top \\ \boldsymbol{\theta} \odot \boldsymbol{A}_{21}^* & (\frac{b^\star}{\pi\beta} - \frac{b^\star}{\pi\beta}(\boldsymbol{V}_d^{-1} + \mathrm{diag}(\frac{b^\star}{\pi\beta}))_e^{-1} \odot \theta\theta^\top \frac{b^\star}{\pi\beta}) \end{pmatrix} \tag{191}$$

Then, for any sequence of deterministic matrices A:

$$\left|\mathrm{Tr}\big(\mathcal{G}_W(z)A\big) - \mathrm{Tr}\Big(\tilde{\mathcal{G}}_e(z)A\Big)\right| = \mathcal{O}_\prec(\frac{\|A\|_{tr}}{\sqrt{d}}) \tag{192}$$

Proof. Applying Lemma B.3 twice, we obtain:

$$\begin{aligned}\mathcal{G}_W(z) - \tilde{\mathcal{G}}_e(z) &= \frac{\alpha}{p}\mathcal{G}_W(z)((M_d^*)^{-1} - M^{\theta_d^{-1}})\tilde{\mathcal{G}}_e(z)\\ &\frac{\alpha}{p}\mathcal{G}_W(z)(M_d^*)^{-1}(M^* - M^{\theta_d^{-1}})M^{\theta_d^{-1}}\tilde{\mathcal{G}}_e(z)\end{aligned}$$

By Lemma B.4, $\left\|\mathcal{G}_W(z)\right\|\left\|\tilde{\mathcal{G}}_e(z)\right\|$ are bounded by $\frac{1}{\zeta}$ while the norms of $\left\|(M_d^*)^{-1}\right\|$, $\left\|(M^\theta)_d^{-1}\right\|$ are bounded by constants due to Lemma B.12. Therefore:

$$\left|\operatorname{Tr}\big(\mathcal{G}_W(z)A\big) - \operatorname{Tr}\Big(\tilde{\mathcal{G}}_e(z)A\Big)\right| \leq \frac{C}{\zeta^2}\left|\operatorname{Tr}\big(\mathcal{G}_W(z)M^\star\big) - \operatorname{Tr}\Big(\tilde{\mathcal{G}}_e(z)M_d^\theta\Big)\right| \tag{193}$$

Applying Lemma D.10 then completes the proof. □

D.4 Self-consistent Equation for $\chi(\kappa)$

$\tilde{\mathcal{G}}(z)$ is almost identical to the desired equivalent matrix $\mathcal{G}_e(z)$ in Theorem 4.5, except for $\chi_d(\kappa)$ still being dimension-dependent unknowns. We resolve this by utilizing D.3 to obtain a self-consistent equation for $\chi(\kappa)$

Lemma D.12. *Let $\psi, b^\star$ be as defined in Lemma D.9. Then, almost surely over $\kappa \sim \mathcal{N}(0,1)$, the following holds:*

$$\chi_d(\kappa, z) = \sum_{q,q' \in [k]} \psi_{qq'}(V_d, \nu_d) c_1(\kappa, \zeta_q^u) c_1(\kappa, \zeta_{q'}^u) + \sum_{q \in [k]} b_q^\star((V_d, \nu_d)) \sum_{\ell \geq 2} c_\ell^2(\kappa, \zeta_q^u) + \mathcal{O}_\prec(\frac{\kappa^\ell}{\sqrt{d}}).$$

Proof. We first recall that $\|R_\kappa^\star\|_2 = \mathcal{O}_\prec(\kappa^\ell)$ for some $\ell \in \mathbb{N}$. Now, note that $\chi_d(\kappa) = \mathbb{E}\left[\operatorname{Tr}\Big(\tilde{\mathcal{G}}_e(z)R_\kappa^\star\Big)\right]$ involves dependency on W in $R_\kappa^\star$. Therefore, we cannot directly apply Proposition D.3 and must resort to Proposition D.3. we have:

$$\chi_d(\kappa) = \mathbb{E}\left[\frac{1}{p}\operatorname{Tr}\big(\mathcal{G}_W(z)R_\kappa^\star\big)\right] + \mathcal{O}_\prec(\frac{\kappa^\ell}{\sqrt{d}}) \tag{194}$$

We start by expressing $\mathcal{G}_W(z)$ through a Schur-complement decomposition:

$$\mathcal{G}_W(z) = \begin{bmatrix} C_W & -C_W Q_W^\top \\ -Q_W C_W & \tilde{G}_W + Q_Q C_W Q_W^\top, \end{bmatrix}, \tag{195}$$

, where:

$$\tilde{G}_W = ((\boldsymbol{V}_e \odot \boldsymbol{W}\boldsymbol{W}^\top + \operatorname{diag}\nu_d + \lambda \boldsymbol{I}_p)) + \alpha S \odot \theta\theta^\top - zI)^{-1}, \tag{196}$$

and C_W, Q_W in Equation (195) contribute only finite-rank components of bounded operator norm. Since $\|R_\kappa^\star\|_2 = \mathcal{O}_\prec(\kappa^\ell)$, such components contribute $\mathcal{O}(\frac{1}{p})$ to $\mathbb{E}\left[\frac{1}{p}\operatorname{Tr}\big\{\mathcal{G}_W(z)R_{22}^\star(\kappa)\big\}\right]$ and can therefore be ignored. We're left with:

$$\frac{1}{p}\mathbb{E}\left[\operatorname{Tr}\Big\{\tilde{G}_W R_{22}^\star(\kappa)\Big\}\right] \tag{197}$$

, which evaluates to:

$$\underbrace{\mathbb{E}\left[\operatorname{Tr}\Big\{(c_1(\kappa,u)c_1(\kappa,u)^\top) \odot WW^\top)(\boldsymbol{V}_e \odot \boldsymbol{W}\boldsymbol{W}^\top + \operatorname{diag}(\nu_d) + \lambda \boldsymbol{I}_p)^{-1}\Big\}\right]}_{T_1}$$

$$+ \underbrace{\mathbb{E}\left[\operatorname{Tr}\left\{\operatorname{diag}(\Big(\sum_{k \geq 2} c_k^2(\kappa,u)\Big)(\boldsymbol{V}_e \odot \boldsymbol{W}\boldsymbol{W}^\top + \operatorname{diag}(\nu_d) + \lambda \boldsymbol{I}_p)^{-1}\right\}\right]}_{T_2} + \mathcal{O}(\frac{1}{p}),$$

where we again supressed contributions from finite-rank terms. Since $\mathrm{diag}(\Big(\sum_{k\geq 2} c_k^2(\kappa, u)\Big)$ is independent of W, by Lemma D.9, T_2 simplifies to:

$$\sum_{q\in[k]} b_q^\star((V_d, \nu_d)) \sum_{\ell\geq 2} c_\ell^2(\kappa, \zeta_q^u) \tag{198}$$

While T_1 can be decomposed into terms of the form $\mathrm{Tr}\big(e^i \odot WM^*W \odot e^j\big)$, which converge to $\psi(V,\nu)$ by Lemma D.9.

□

D.5 Self-consistent equations for $V^\star, D^\star$

Using Lemma D.9, we obtain a deterministic equivalent $\mathcal{G}_e'$ that doesn't depend on the realizations of X, W. However, the quantities V_d, D_d still depend on dimension d due to the dimension dependent definition of $\chi_d(\kappa)$. While D.12 defines a self-consistent equation for $\chi(\kappa)$, it remains an infinite-dimensional (functional) order parameter. Instead, we show that we can directly construct dimension-independent self-consistent equations on V_d, ν_d:

Proposition D.13. *For $C, D \in \mathbb{R}^{k,k}$, let $\psi_{i,j}$ be as defined in Theorem 4.5. Define:*

$$\chi_{C,D}(\kappa) = \frac{1}{\beta} \sum_{i=1,j=1}^{k} \psi_{i,j}(C,D) c_1(u_i,\kappa) c_1(u_j,\kappa) + \frac{1}{\beta}\sum_{i=1}^{k} (b_{ii}(\sum_{k\geq 2} c_k^2(\kappa, u_i)) \tag{199}$$

Then, V, D defined as:

$$V = \mathbb{E}\left[\frac{\alpha}{1+\chi_d(\kappa)} c_1(u_i,\kappa) c_1(u_J,\kappa)\right],$$

$$D = \mathbb{E}\left[\frac{\alpha}{1+\chi_d(\kappa)} \sum_{k\geq 2} c_k^2(u_i,\kappa)\right]$$

satisfy:

$$V_{i,j} = \mathbb{E}\left[\frac{\alpha}{1+\chi_{V,D}(\kappa)} c_1(u_i,\kappa) c_1(u_J,\kappa)\right] + \mathcal{O}_\prec(\frac{1}{\sqrt{d}}) \tag{200}$$

$$D_{i,i} = \mathbb{E}\left[\frac{\alpha}{1+\chi_{V,D}(\kappa)} \sum_{k\geq 2} c_k^2(u_i,\kappa)\right] + \mathcal{O}_\prec(\frac{1}{\sqrt{d}}). \tag{201}$$

Proof. Note that $\frac{\alpha}{1+\chi_{V,D}(\kappa)}$ is uniformly Lipschitz in κ. Consider independent sample $\kappa_1, \cdots, \kappa_n'$ for some $n' \propto d$. Then $\sup|\kappa_i| = \sqrt{\log d}$ and:

$$\mathbb{E}\left[\frac{\alpha}{1+\chi_{V,D}(\kappa)} c_1(u_i,\kappa) c_1(u_J,\kappa)\right] = \frac{1}{n'} \sum_{i=1}^{n'} \frac{\alpha}{1+\chi_{V,D}(\kappa_i)} c_1(u_i,\kappa) c_1(u_J,\kappa) + \mathcal{O}_\prec(\frac{1}{\sqrt{d}}). \tag{202}$$

Proposition D.13 then follows from Proposition D.11 and the self-consistent equation for $\chi(\kappa)$ in Lemma D.12. □

Next, we show that the above fixed point equations are contractive, allowing us to translate approximate satisfiability to distance of V_d, D_d from the unique fixed points satisfied by $V^\star, D^\star$ defined in Theorem 4.5:

Lemma D.14. *Let $z \in \mathbb{C}^+$. Define $\mathcal{S}(z)$ as the set of $C, D \in \mathbb{C}^{-k\times k} \times \mathbb{C}^{-k}$ satisfying:*

- $b^\star(C,D) \in \mathbb{C}^+$.
- $\left|b^\star(C,D)\right| \leq \frac{\pi_i}{\beta\zeta}$

Then, there exists $C > 0$ such that for $\zeta = Im(z) > C$, the fixed point iteration defined in Proposition D.13 i.e:

$$F : \mathbb{C}^{-k\times k} \times \mathbb{C}^{-k} \to \mathbb{C}^{-k\times k} \times \mathbb{C}^{-k}$$
$$F(C, D) = \mathbb{E}\left[\frac{\alpha}{1 + \chi_{V,D}(\kappa)} c_1(\boldsymbol{u}, \kappa) c_1(\boldsymbol{u}, \kappa)^\top\right], \mathbb{E}\left[\frac{\alpha}{1 + \chi_{C,D}(\kappa)} \sum_{k\geq 2} c_k^2(\boldsymbol{u}, \kappa)\right],$$

is contractive in $\mathcal{S}(z)$, i.e for any C, D:

$$\big\|F(C_1, D_1) - F(C_2, D_2)\big\| \leq \big\|(C_1, D_1) - (C_2, D_2)\big\| \tag{203}$$

Proof. From the definition of $\chi(\kappa)$, we directly obtain that for any C, D such that $|b| \leq \frac{\beta\pi}{\zeta}$

$$\big|\chi(\kappa, b_1) - \chi(\kappa, b_2)\big| \leq C_1\|b_1 - b_2\| + C_2|b|\|C_1, D_1 - C_2, D_2\| \tag{204}$$

The restriction $\big|b_i^\star(z)\big| \leq \frac{\pi_i}{\beta\zeta}$ therefore implies that $\big|\chi_{C,D}(\kappa)\big| \leq \frac{K}{\zeta}$ for some $K > 0$. Now, since:

$$\left|\frac{1}{1 + \chi_{V,D}(\kappa)}\right| \leq 1 - \big|\chi_{V,D}(\kappa)\big|, \tag{205}$$

for small enough $\big|\chi_{V,D}(\kappa)\big|$, we obtain the following entry-wise upper -bound for any-feasible solution of C, D:

$$|C| \leq K', |D| \leq K', \tag{206}$$

for some $K' > 0$. This ensures that the boundedness condition on C, D in Lemma D.9 applies.

Recall the definition of $b^\star(C, D)$:

$$b^\star(C, D) = \pi_{u_i}\beta((C^{-1} + \mathrm{diag}(b))^{-1} + (\mathrm{diag}(D) - zI_p))^{-1}_{p,p} \tag{207}$$

Subsequently, we may apply Lemma B.3 to obtain: (See also Lemma F.1 in the proof of Lemma D.9)

$$\big\|b(C_1, D_1) - b(C_2, D_2)\big\| \leq \frac{K''}{\zeta^2}\|C_1, D_1 - C_2, D_2\|, \tag{208}$$

, for some constant $K'' > 0$ Combining the above with 209 yields:

$$\big|\chi(\kappa, C_1, D_1) - \chi(\kappa, C_2, D_2)\big| \leq \frac{K'''}{\zeta}\|C_1, D_1 - C_2, D_2\| \tag{209}$$

, for some constant $K''' > 0$.

□

D.6 Proof of Theorem 4.5

Lemma D.14 and the Banach-fixed point theorem imply that $F(C, D)$ admit unique fixed points $V^\star(z), \nu^\star(z)$ for $\zeta > C$ and that:

$$V_d(z) \xrightarrow[a.s]{} V^\star(z), \quad \nu_d(z) \xrightarrow[a.s]{\nu}{}^\star(z) \tag{210}$$

Next, we note that from Proposition D.11, for each $q \in [k]$, $b^\star(z)_q$ can be expressed as:

$$b^\star(z)_q = \sum_{i=1}^{p} \frac{(v_i^\top e^q)^2}{z - \lambda_i} + \mathcal{O}_\prec(\frac{1}{\sqrt{d}}). \tag{211}$$

Theorem 4.5 then follows by noting that by the standard properties of Stieltjes transforms Bai and Zhou [2008], $b^\star(z), V^\star(z), \nu^\star(z)$ admit unique analytic-continuations to $\mathbb{C}/\mathbb{R}^+$.

E Generalization Error

Having obtained the full-deterministic equivalent in Theorem 4.5, we now show how it can be exploited to yield the asymptotic generalization error after a gradient step.

We will proceed as follows:

- Use the covariance approximation in Proposition D.1 to identify certain “statistics" involving the resolvent that characterize the asymptotic generalization error.
- Relate these statistics to projections of the extendent resolvent onto certain deterministic matrices. Introduce perturbation terms to extract more complicated functionals.
- Replace these statistics with the corresponding quantities obtained through the deterministic equivalent $\mathcal{G}_e$, and where necessary average over θ to obtain the asymptotic generalization as a function of the parameters $V^\star, \nu^\star, b^\star$ in Theorem 4.5.

E.1 Order Parameters for Generalization Error

We start by using the covariance approximation D.1 to simplify the expression for the generalization error with an arbitrary fixed choice of $a \in \mathbb{R}^p$

Lemma E.1. *Let $a \in \mathbb{R}^p$ be a fixed vector such that $\|a\|/\sqrt{p} = \mathcal{O}(1)$ and let $e^1, \cdots, e^k$ denote the “spike" directions defined in Equation (13). Define for $i \in [k]$:*

$$\tau^d_{0,i} \overset{def}{=} \frac{a^\top e_i}{\sqrt{p}}, \qquad \tau^d_{1,i} \overset{def}{=} \frac{a^\top e_i \odot w^\star}{\sqrt{p}}, \qquad \tau^d_2 \overset{def}{=} \frac{a^\top C \odot WW^\top a}{p} \quad \textit{and} \quad \tau^d_4 \overset{def}{=} \frac{a^\top D a}{p}, \tag{212}$$

where $C = \mathbb{E}\left[c_1(\kappa, u) c_1(\kappa, u)^\top\right]$ and $D_{j,j,} = \mathbb{E}_\kappa\left[\sum_{k \geq 2} c_k^2(\kappa, u_j)\right]$. Then, the generalization error can be expressed as:

$$\begin{aligned}\mathbb{E}\left[e_g\right] = \mathbb{E}&\left[\left[\sigma_\star(s) - \sum_{j=1}^k c_0(\kappa, u_j)\tau_{0,j} - \sum_{j=1}^k c_1(\kappa, u_j)\kappa\tau_{1,j}\right]^2\right] + \tau_3 \\ &- \mathbb{E}\left[(\sum_{j=1}^k c_1(\kappa, u_j)\tau_{1,j})^2 + (\sum_{j=1}^k c_1(\kappa, u_j)\tau_{2,j})^2\right] + \tau_3 + \mathcal{O}_\prec(d^{-1/2}).\end{aligned} \tag{213}$$

Proof. The prediction at a point x under the simplified updated weights is given by:

$$f(x, \tilde{W}, a) = \sum_{j=1}^k \tau_{0,j}\bar{\phi}^j_x + a^\top \tilde{\phi}_x \tag{214}$$

The generalization error is then given by:

$$\mathbb{E}\left[e_g(a)\right] = \mathbb{E}\left[[\sigma_\star(s) - f(x, \tilde{W}, a)]^2\right]$$

$\mathbb{E}\left[e_g\right]$ can be equivalent expressed through the following quadratic form applied to the extended features ϕ^e_μ:

$$\begin{aligned}\mathbb{E}\left[e_g(a)\right] &= \mathbb{E}\left[(\phi^e_\mu)^\top u_a u_a^\top \phi^e_\mu\right] \\ &= \mathbb{E}\left[\mathrm{Tr}\Big(u_a u_a^\top \phi^e_\mu (\phi^e_\mu)^\top\Big)\right]\end{aligned}$$

where $u_a = [1, -\tau_0, \tilde{a}] \in \mathbb{R}^{p+k+1}$. By Proposition D.1, we have:

$$\mathbb{E}\left[\mathrm{Tr}\Big(u_a u_a^\top \phi^e_\mu (\phi^e_\mu)^\top\Big)\right] = \mathbb{E}\left[\mathrm{Tr}\Big(u_a u_a^\top R^\star_\kappa\Big)\right] + \mathcal{O}(\frac{\operatorname{polylog} d}{\sqrt{d}}). \tag{215}$$

Expanding each of the terms in $R^\star_\kappa$ then yields Equation (213). □

E.2 Extended Resolvent to Generalization Error

From Lemma E.1, we note that obtaining the limiting generalization error requires characterizing the functionals $\tau_{0,i}^d, \tau_{1,i}^d, \cdots$ when $\hat{a}$ is set as the ridge-regression estimator:

$$\begin{aligned}\hat{\boldsymbol{a}}_\lambda &= \underset{\boldsymbol{a}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i\in[n]} \Big(y_i - f(\boldsymbol{x}_i; \boldsymbol{a}, \boldsymbol{W}^1)\Big)^2 + \lambda||\boldsymbol{a}||_2^2 \\ &= \Big(\Phi^\top\Phi + \lambda\boldsymbol{I}_n\Big)^{-1}\Phi^\top\boldsymbol{y}. \end{aligned} \tag{216}$$

To extract these relevant functionals, we introduce certain perturbation terms in the extended resolvent.

$$\boldsymbol{G}_e(\rho_1,\rho_2) \overset{\text{def}}{=} \Big((\Phi^e)^\top(\Phi^e) + \lambda\boldsymbol{I} + \rho_1\boldsymbol{L} + \rho_2\boldsymbol{D'}\Big)^{-1} \in \mathbb{R}^{(p+1)\times(p+1)}, \tag{217}$$

where the matrices $\boldsymbol{D'}, \boldsymbol{L} \in \mathbb{R}^{(p+1)\times(p+1)}$ are given by:

$$\boldsymbol{L} = \begin{bmatrix} \boldsymbol{0}_{k+1,k+1} & \boldsymbol{0}_{k+1,p} \\ \boldsymbol{0}_{p,k+1} & \boldsymbol{C}\odot\boldsymbol{W}\boldsymbol{W}^\top \end{bmatrix}, \qquad \boldsymbol{D'} = \begin{bmatrix} \boldsymbol{0}_{k+1,k+1} & \boldsymbol{0}_{k+1,p} \\ \boldsymbol{0}_{p,k+1} & \boldsymbol{D} \end{bmatrix}, \tag{218}$$

with $C = \mathbb{E}\left[c_1(\kappa,u)c_1(\kappa,u)^\top\right]$ and $D_{j,j,} = \mathbb{E}_\kappa\left[\sum_{k\geq 2} c_k^2(\kappa,u_j)\right]$. We will demonstrate that the introduction of the perturbation terms $\boldsymbol{L}, \boldsymbol{D'}$ allows extraction of all the relevant functionals in Lemma E.1.

Proposition E.2. *Let $\hat{a}_\lambda$ denote the ridge-regression minimizer with regularization $\lambda > 0$ i.e:*

$$\begin{aligned}\hat{\boldsymbol{a}}_\lambda &= \underset{\boldsymbol{a}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i\in[n]} \Big(y_i - f(\boldsymbol{x}_i; \boldsymbol{a}, \boldsymbol{W}^1)\Big)^2 + \lambda||\boldsymbol{a}||_2^2 \\ &= \Big(\Phi^\top\Phi + \lambda\boldsymbol{I}_n\Big)^{-1}\Phi^\top\boldsymbol{y}. \end{aligned} \tag{219}$$

Let $\tilde{G}(\rho_1,\rho_2)$ denote the resolvent of the "bulk" component i.e:

$$\tilde{G}(\rho_1,\rho_2) = (\tilde{\phi}\tilde{\phi}^\top + \lambda I_n)^{-1} \tag{220}$$

Consider the following Schur-complement decomposition of G_e:

$$G_e(\rho_1,\rho_2) = \begin{bmatrix} C & -CQ^\top \\ -QC & P \end{bmatrix}, \tag{221}$$

where

$$P = \tilde{G}(\rho_1,\rho_2) + QCQ^\top, \tag{222}$$

$$Q = \tilde{G}(\rho_1,\rho_2)\Sigma_{21}/p, \tag{223}$$

where $\Sigma_{12} = \tilde{\Phi}\begin{bmatrix} y & \bar{\Phi}\end{bmatrix}$ with $y \overset{def}{=} [y^1, y^2, \ldots, y^n]$, $\bar{\Phi} \overset{def}{=} [\bar{\phi}^1, \bar{\phi}^2, \ldots, [\bar{\phi}^n]$, and

$$C^{-1} = \begin{bmatrix} \|y\|^2/p + \lambda & y^\top\bar{\Phi}/p \\ y^\top\bar{\Phi}/p & \bar{\Phi}^\top\bar{\Phi}/p + \lambda \end{bmatrix} - \frac{1}{p}\begin{bmatrix} y^\top \end{bmatrix}\tilde{\Phi}^\top G(\rho_1,\rho_2)\tilde{\Phi}\begin{bmatrix} y & \bar{\Phi}\end{bmatrix}. \tag{224}$$

Let $\hat{\tau}_0, \hat{\tau}_{1i}, \hat{\tau}_{2i}, \hat{\tau}_3$ be as defined in Lemma E.1, then:

$$\hat{\tau}_0 = (C^{-1})_{01}((C^{-1})_{11} + \lambda(1 - \frac{1}{p}\operatorname{diag}(\pi_u)))^{-1} \tag{225}$$

$$\hat{\tau}_{1,j} = (\theta\odot e^j)^\top Q\begin{bmatrix} 1 \\ -\hat{\tau}_0 \end{bmatrix} + \mathcal{O}(\frac{\hat{\tau}_0}{\sqrt{p}}) \tag{226}$$

$$\hat{\tau}_2 = \begin{bmatrix}1 & -\tau_0^*\end{bmatrix} \frac{\partial}{\partial \rho_1} \left[(C)^{-1}\right]_{\rho=0} \begin{bmatrix}1 \\ -\hat{\tau}_0\end{bmatrix} + \mathcal{O}(\frac{\hat{\tau}_0^2}{p}) \tag{227}$$

and

$$\hat{\tau}_3 = \begin{bmatrix}1 & -\tau_0^*\end{bmatrix} \frac{\partial}{\partial \rho_2} \left[(C)^{-1}\right]_{\rho=0} \begin{bmatrix}1 \\ -\hat{\tau}_0\end{bmatrix} + \mathcal{O}(\frac{\hat{\tau}_0^2}{p}). \tag{228}$$

, where $\mathcal{O}(\hat{\tau}_0)$ denotes error bounded as $C\|\hat{\tau}_0\|$ for some constant $C > 0$.

Proof. We decompose $\hat{a}_\lambda$ into projections along $\boldsymbol{e}^1, \cdots, \boldsymbol{e}^k$ and the orthogonal complement, denoted as $\tilde{a} = (I - \Pi)a$, where Π denotes the projection operator along $\boldsymbol{e}^1, \cdots, \boldsymbol{e}^k$.

The ridge-regression objective can be re-expressed as:

$$\mathcal{R}(\tau_0, \tilde{a}) = \min_{\tau_0 \in \mathbb{R}^k, \tilde{a}} \sum_{\mu=1}^{n} \left(y_\mu - \bar{\phi}_\mu^\top \tau_0 - \tilde{a}^\top \bar{\phi}_\mu\right)^2 + \lambda \left\| \frac{1}{\sqrt{\pi}} \cdot \tau_0 \right\|^2 + \lambda \|\tilde{a}\|^2, \tag{229}$$

where $\cdot$ denotes element-wise multiplication by $\pi \in \mathbb{R}^k$, accounting for the variability in the number of entries with value ζ_q^u for $q \in [k]$.

The optimality condition can similarly be expressed in terms of $\tau_0, \tilde{a}$. Differentiating the objective in Eq. (229) w.r.t τ_0 yields:

$$\hat{\tau}_0 = (\sum_\mu \bar{\phi}_\mu y_\mu - \sum_{\mu,\nu} \bar{\phi}_\mu \tilde{\phi}_\mu^\top \tilde{G} \tilde{\phi}_\mu y_\nu)(\sum_\mu \bar{\phi}_\mu \bar{\phi}_\mu^\top - \sum_{\mu,\nu} \bar{\phi}_\mu \tilde{\phi}_\mu \tilde{G} \tilde{\phi}_\mu \bar{\phi}_\mu + \lambda \operatorname{diag}(\pi))^{-1}.$$

Similarly, differentiating w.r.t $\tilde{a}$, we obtain:

$$\sum_\mu \bar{\phi}_\mu \bar{\phi}_\mu^\top \tau_0 + \lambda \operatorname{diag}(1/\pi) \tau_0 = \textstyle\sum_\mu \bar{\phi}_\mu y_\mu - \sum_\mu \bar{\phi}_\mu \tilde{\phi}_\mu \tilde{a} \tag{230}$$

Simplifying, we obtain that orthogonal component of the minimizer $\tilde{a}$ is given by:

$$\tilde{a} = \tilde{G} \sum_\mu \tilde{\phi}_\mu (y_\mu - \bar{\phi}_\mu^\top \tau_0), \tag{231}$$

where $\tilde{G}$ is defined as before:

$$\tilde{G} \overset{\text{def}}{=} (\sum_\mu \tilde{\phi}_\mu (\tilde{\phi}_\mu)^\top + \lambda I)^{-1}. \tag{232}$$

Therefore, we obtain:

$$\tilde{a} = Q \begin{bmatrix}1 \\ -\hat{\tau}_0\end{bmatrix}, \tag{233}$$

Isolating the contributions from the projections along $e^1, \cdots, e^k$ in τ_1, i, we obtain:

$$\frac{a^\top e_i \odot w^\star}{\sqrt{p}} = \frac{\tilde{a}^\top e_i \odot w^\star}{\sqrt{p}} + \mathcal{O}(\frac{\hat{\tau}_0}{\sqrt{p}})$$

Comparing the above with Equation (224), we obtain Equation (225).

The expressions for $\hat{\tau}_{2i}, \hat{\tau}_3$ then follow directly from the expressions for $\frac{\partial}{\partial \rho_1} \left[(C)^{-1}\right]_{\rho=0}, \frac{\partial}{\partial \rho_2} \left[(C)^{-1}\right]_{\rho=0}$ □

E.3 Deterministic Equivalent for the perturbed Resolvent

Note that the relations established in Proposition E.2 still involve the full high-dimensional matrices X, W. Our goal now will be to use the deterministic equivalence we established earlier to obtain deterministic limits for each of the quantities.

We begin by incorporating the perturbation terms into our deterministic equivalence. Fortunately, this simply corresponds to rescaling certain terms by constants, as we explain below:

Theorem E.3. *Let $V^\star \in \mathbb{C}^{k\times k}, \nu^\star \in \mathbb{C}^k, b^\star \in \mathbb{C}^k$ be uniquely defined through the following conditions:*

- *$V^\star, \nu^\star, b^\star$ satisfy the following set of self-consistent equations:*

$$V^\star_{qq'}(z) = \mathbb{E}_\kappa\left[\alpha\frac{c_1(\kappa,\zeta^u_q)c_1(\kappa,\zeta^u_{q'}) + \rho_1 + \rho_1\chi_{V,d,b}}{1+\chi_{V,d,b}(\kappa)}\right]$$
$$\nu^\star_q(z) = \mathbb{E}_\kappa\left[\sum_{\ell\geq 2}\frac{\alpha c^2_\ell(\kappa,\zeta^u_q)}{1+\chi_{V,d,b}(\kappa)} + \rho_2\sum_{\ell\geq 2}c^2_\ell(\kappa,\zeta^u_q)\right]..,$$
$$b^\star_q(z) = \pi_q\beta((V^\star)^{-1} + \mathrm{diag}(b^\star) + (\mathrm{diag}(\nu^\star) - zI_p))^{-1}_{q,q}$$

where $\kappa \sim \mathcal{N}(0,1)$ $(c_\ell(\kappa,\zeta))_{\ell>0}$ are defined in 4.2 and where $\kappa \sim \mathcal{N}(0,1)$, $(c_\ell(\kappa,\zeta))_{\ell>0}$ are defined in 4.2 and $(\chi(z;\kappa), L(z))$ read as follows:

$$\beta\chi(z;\kappa) = \sum_{q,q'\in[k]}\psi_{qq'}c_1(\kappa,\zeta^u_q)c_1(\kappa,\zeta^u_{q'}) + \sum_{q\in[k]}b^\star_q\sum_{\ell\geq 2}c^2_\ell(\kappa,\zeta^u_q),$$
$$L(z) = \Big(V^\star(z)^{-1} + \mathrm{diag}(b^\star(z))\Big)^{-1},$$

where $\psi(z) \in \mathbb{R}^{k\times k}$ is defined as:

$$\psi(z) = b^\star(z) - L(z)\odot(b^\star(z)(b^\star(z))^\top), \tag{234}$$

- *$V^\star, \nu^\star, b^\star$ are analytic mappings satisfying $V^\star_{i,j}: \mathbb{C}^+ \to \mathbb{C}^-$ for $i,j \in [k]$, $\nu^\star_i : \mathbb{C}^+ \to \mathbb{C}^-$ for $i \in [k]$, $b^\star : \mathbb{C}^+_i \to \mathbb{C}^+$ for $i \in [k]$.*
- *$\left|b^\star_i(z)\right| \leq \frac{\pi_i}{\beta\zeta}$.*

Define $\mathcal{G}_e(z,\rho_1,\rho_2) \in \mathbb{R}^{(p+1)\times(p+1)}$ as:

$$\mathcal{G}_e(z) = \begin{bmatrix} A^*_{11} - zI_k & (A^*_{21})^\top\odot\theta^\top \\ \theta\odot A^*_{21} & A^*_{22} + \alpha S^*_e\odot\theta\theta^\top\end{bmatrix}^{-1}, \tag{235}$$

where A^*_{11}, A^*_{12}, S_e are identical to Theorem 4.5.

Then, for any $z \in \mathbb{C}/\mathbb{R}^+$ and sequence of deterministic matrices $A \in \mathbb{C}^{(p+1)\times(p+1)}$ with $\|A\|_{\mathrm{tr}} = \mathrm{Tr}\Big((AA^*)^{1/2}\Big)$ uniformly bounded in d:

$$\mathrm{Tr}(AG_e(z,\rho_1,\rho_2)) \xrightarrow[d\to\infty]{a.s} \mathrm{Tr}(A\mathcal{G}_e(z,\rho_1,\rho_2))). \tag{236}$$

Proof. We observe that the introduction of perturbation terms ρ_1, ρ_2 simply amounts to shifting $V^\star, \nu^\star$ by constants. To see this, first note that the perturbation matrices $C\odot WW^\top, D$ do not depend on X. Therefore the proof of Proposition D.3 applies to $G_e(z,\rho_1,\rho_2)$ with the only modification being the inclusion of L, D' along with $-zI$ as constant terms. Subsequently, in the second stage, the structure of L, D' allows them to be directly absorbed into the component $M^\star$ with changes upto constants in V_d, ν_d. □

As a direct corollary, we obtain the deterministic equivalents of the Schur-complement Representations in the following sense:

Corollary E.4. *Let $\mathcal{G}_e(\rho_1,\rho_2)$ denote the deterministic equivalent for the perturbed resolvent as per Theorem E.3 with z set as $-\lambda$. Define the following block-matrix representations for $G_e, \mathcal{G}_e$:*

$$G_e(\rho_1,\rho_2) = \begin{bmatrix} C & -CQ^\top \\ -QC & P\end{bmatrix}. \tag{237}$$

$$\mathcal{G}_e(\rho_1, \rho_2) = \begin{bmatrix} \mathcal{C} & -\mathcal{C}\mathcal{Q}^\top \\ -\mathcal{Q}\mathcal{C} & \mathcal{P} \end{bmatrix}. \tag{238}$$

Then:

$$C = \mathcal{C} + \mathcal{O}_\prec(\frac{1}{\sqrt{d}}), \tag{239}$$

Furthermore, for any fixed $\boldsymbol{r} \in \mathbb{R}^{p-k}$ with $\|\boldsymbol{r}\| = \mathcal{O}(1)$:

$$Q\boldsymbol{r} = \mathcal{Q}\boldsymbol{r} + \mathcal{O}_\prec(\frac{1}{\sqrt{d}}), \tag{240}$$

Proof. The above equivalences follow from Theorem 4.5 by setting A as matrices acting on the individual blocks of $G_e(\zeta, \rho)$. Concretely, to obtain $C_{i,j}$ we set A as the matrix with $A_{i,j} = 1$ and 0 otherwise. □

We derive the resulting expressions for $\mathcal{C}, \mathcal{Q}$ below for subsequent usage:

Lemma E.5. *Consider the Schur-decomposition of the equivalent resolvent $\mathcal{G}_e$ (with $z = -\lambda$) defined by Equation (238) and let ψ, S be as defined in Theorem E.3. The matrices $\mathcal{C}, \mathcal{Q}$ satisfy:*

$$\begin{aligned} \mathcal{C}^{-1} &= A^*_{11} + \lambda I_{k+1} - (A^*_{21})^\top ((\psi)^{-1} + \alpha S)^{-1} A^*_{21} + \mathcal{O}_\prec(\frac{1}{\sqrt{p}}) \\ \mathcal{Q} &= ((\mathrm{diag}(\frac{b^\star}{\pi\beta})_e - \frac{1}{\beta^2}(\mathrm{diag}(b^\star) - \psi(V_d, \nu_d))_e \odot \frac{\theta}{\pi}\frac{\theta}{\pi}^\top)^{-1} + \alpha S_e \odot \theta\theta^\top)^{-1} \theta \odot A^\star_{21} + \mathcal{O}_\prec(\frac{1}{\sqrt{p}}), \end{aligned} \tag{241}$$

where in the expression for $\mathcal{C}$, we've further averaged over θ.

Proof. Throughout, we shall use the following simplification:

$$(e^i \odot \theta)(e^j \odot \theta) = \pi \delta_{i=j} + \mathcal{O}_\prec(\frac{1}{\sqrt{d}}) \tag{242}$$

, which follows from the definition of e^i, π and θ_i being independent sub-Gaussian random variables.

By Theorem E.3 and the definition of $\mathcal{C}$, we have:

$$\begin{aligned} \mathcal{C} &= A^*_{11} + \lambda I_{k+1} \\ &\quad \underbrace{- (A^\star_{21})^\top \odot \theta^\top ((\mathrm{diag}(\frac{b^\star}{\pi\beta})_e - \frac{1}{\beta^2}(\mathrm{diag}(b^\star) - \psi(V_d, \nu_d))_e \odot \frac{\theta}{\pi}\frac{\theta}{\pi}^\top)^{-1} + \alpha S_e \odot \theta\theta^\top)^{-1} \theta \odot A^\star_{21}}_{T} \end{aligned}$$

,

To simplify the above expression, we apply the matrix-Woodbury identity to the term T. We first recognize from the definition of ψ in Equation (179) that the following relation holds:

$$(\boldsymbol{V}_d^{-1} + \mathrm{diag}(b^\star))^{-1} \odot (\frac{b^\star}{\pi\beta})_e (\frac{b^\star}{\pi\beta})_e^\top \odot \theta\theta^\top = \frac{1}{\beta^2}(\mathrm{diag}(b^\star) - \psi(V_d, \nu_d))_e \odot \frac{\theta}{\pi}\frac{\theta}{\pi}^\top \tag{243}$$

Further, noting that $\alpha S_e \odot \theta\theta^\top$ corresponds to a finite-rank perturbation $E_\theta S_e E_\theta^\top$ along with additional simplifications by Equation (242), we obtain:

$$\begin{aligned} T &= (A^\star_{21})^\top \odot \theta^\top ((\mathrm{diag}(\frac{b^\star}{\pi\beta})_e - \frac{1}{\beta^2}(\mathrm{diag}(b^\star) - \psi(V_d, \nu_d))_e \odot \frac{\theta}{\pi}\frac{\theta}{\pi}^\top) \theta \odot A^\star_{21} \\ &\quad - (A^\star_{21})^\top \psi ((\alpha S)^{-1} + \psi)^{-1} \psi A^\star_{21} + \mathcal{O}_\prec(\frac{1}{\sqrt{d}}) \\ &= (A^\star_{21})^\top \psi A^\star_{21} - (A^\star_{21})^\top \psi ((\alpha S)^{-1} + \psi)^{-1} \psi A^\star_{21} + \mathcal{O}_\prec(\frac{1}{\sqrt{d}}) \\ &= (A^\star_{21})^\top ((\alpha S) + \psi^{-1}) A^\star_{21} + \mathcal{O}_\prec(\frac{1}{\sqrt{d}}) \end{aligned}$$

Next, for $\mathcal{Q}$, the expression follows by direction substitution. Recall from Proposition E.2, that $Q = \tilde{G}(\rho_1, \rho_2)\Sigma_{21}/p$. From the structure of $\mathcal{G}(\rho_1, \rho_2)$ in Theorem E.3, we obtain that the corresponding matrix $\mathcal{Q}$, in the equivalent $\mathcal{G}(\rho_1, \rho_2)$ is similarly given by:

$$((\mathrm{diag}(\frac{b^\star}{\pi\beta})_e - \frac{1}{\beta^2}(\mathrm{diag}(b^\star) - \psi(V_d, \nu_d))_e \odot \frac{\theta}{\pi}\frac{\theta}{\pi}^\top)^{-1} + \alpha S_e \odot \theta\theta^\top)^{-1}\theta \odot A^\star_{21}. \tag{244}$$

□

E.4 Extracting order-parameters through the deterministic equivalent

Armed with the deterministic equivalent for the perturbed resolvent (Theorem E.3), it remains to justify and extract the quantities $\tau_0, \tau_1, \tau_2, \tau_3$.

Lemma E.6. *Let $C^{-1}, \mathcal{C}^{-1}$ denote the corresponding blocks of G_e and $\mathcal{G}_e$ respectively. Then, w.h.p as $d \to \infty$*

$$\frac{\partial}{\partial \rho_1}\left(C^{-1}\right)_{\rho=0} = \frac{\partial}{\partial \rho_1}\left(\mathcal{C}^{-1}\right)_{\rho=0} + \mathcal{O}_\prec(\frac{1}{d^{1/4}}) \tag{245}$$

$$\frac{\partial}{\partial \rho_2}\left(C^{-1}\right)_{\rho=0} = \frac{\partial}{\partial \rho_2}\left(\mathcal{C}^{-1}\right)_{\rho=0} + \mathcal{O}(\frac{1}{d^{1/4}}) \tag{246}$$

Proof. We have:

$$\frac{\partial^2}{\partial \zeta^2}\left(C^{-1}\right)_\zeta = \frac{1}{p}\begin{bmatrix} y^\top \\ \bar{\Phi}^\top \end{bmatrix} \tilde{\Phi}^\top \tilde{G}(WW^\top)^2 G\tilde{\Phi}\begin{bmatrix} y & \bar{\Phi} \end{bmatrix} + \frac{1}{p}\begin{bmatrix} y^\top \\ \bar{\Phi}^\top \end{bmatrix} \tilde{\Phi}^\top \tilde{G}(WW^\top)G(WW^\top)\tilde{\Phi}\begin{bmatrix} y & \bar{\phi} \end{bmatrix}. \tag{247}$$

Recall that $\|G\|_\mu < \frac{1}{\gamma}$ and $\left\|WW^\top\right\| = \mathcal{O}_\prec(1)$ by Lemma B.4. By the submultiplicity of operator norms, we have that $norm\frac{\partial^2}{\partial \zeta^2}\left(C^{-1}\right)_\zeta$ is uniformly bounded in ζ, d with high-probability as $d \to \infty$. Therefore, applying the mean-value-theorem entry-wise to C^{-1} and $C_\star^{-1}$ yields that for any $\zeta > 0$:

$$\frac{\partial}{\partial \zeta}\left(C^{-1}\right)_{\zeta=0} = \frac{(C^{-1}(\zeta) - C^{-1}(0)}{\zeta} + K\zeta, \tag{248}$$

and

$$\frac{\partial}{\partial \zeta}\left(\mathcal{C}^{-1}\right)_{\zeta=0} = \frac{(\mathcal{C}^{-1}(\zeta) - \mathcal{C}^{-1}(0)}{\zeta} + K\zeta, \tag{249}$$

for some constant $K > 0$.

Combining the above with the bounds from Corollary E.4 yields:

$$\frac{\partial}{\partial \rho_1}\left(C^{-1}\right)_{\rho=0} = \frac{\partial}{\partial \rho_1}\left(\mathcal{C}^{-1}\right)_{\rho=0} + K\zeta + \frac{1}{\rho}\mathcal{O}_\prec(\frac{1}{\sqrt{d}}) \tag{250}$$

Therefore, setting $\zeta = \frac{1}{d^{1/4}}$ yields:

$$\frac{\partial}{\partial \rho_2}\left(C^{-1}\right)_{\rho=0} = \frac{\partial}{\partial \rho_2}\left(\mathcal{C}^{-1}\right)_{\rho=0} + \mathcal{O}_\prec(\frac{1}{d^{1/4}}) \tag{251}$$

□

E.5 Proof of Theorem 4.7

The proof of Theorem 4.7 follows directly from Lemma E.1 combined with the next result, which defines the parameters $\hat{\tau}_0, \hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_3$ using the fixed point parameters $V^\star, \nu^\star$:

Proposition E.7. *Let $\hat{\boldsymbol{a}}_\lambda = \underset{\boldsymbol{a}\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i\in[n]} \Big(y_i - f(\boldsymbol{x}_i; \boldsymbol{a}, \boldsymbol{W}^1)\Big)^2 + \lambda||\boldsymbol{a}||_2^2$ denote the ridge-regression estimator after a gradient update to the first layer. Let $\hat{\tau}_0, \hat{\tau}_1, \hat{\tau}_2, \hat{\tau}_3$ be defined as in E.1 with $a = \hat{a}_\lambda$ and let $\mathcal{C}^\star(V^\star, \nu^\star), \psi(V^\star, \nu^\star)$ be functions of $V^\star, \nu^\star$ defined in Theorems 4.5, Lemma D.9 with z set as $-\lambda$ and let $S \in \mathbb{R}^{k\times k}$ be as defined in Theorem 4.5. Then, with $V^\star, \nu^\star$ being the unique solutions to the fixed-point equations defined in Theorem 4.7.*

$$\hat{\tau}_0 = (\mathcal{C}^{-1})_{01}((\mathcal{C}^{-1})_{11} + \lambda(1 - \frac{1}{p}\operatorname{diag}(\pi_u)))^{-1} + \mathcal{O}_\prec(\frac{1}{\sqrt{d}}) \tag{252}$$

$$(\hat{\tau}_1) = ((\psi)^{-1} + \alpha S)^{-1}\widetilde{A}^*_{21}\begin{bmatrix}1\\ -\hat{\tau}_{0,1}\\ \vdots\\ -\hat{\tau}_{0,k}\end{bmatrix} + \mathcal{O}_\prec(\frac{1}{\sqrt{d}}) \tag{253}$$

$$\hat{\tau}_2 = \begin{bmatrix}1 & -\hat{\tau}_0\end{bmatrix}\frac{\partial}{\partial\rho_1}\Big[(\mathcal{C})^{-1}\Big]_{\rho=0}\begin{bmatrix}1\\ -\hat{\tau}_0\end{bmatrix} + \mathcal{O}_\prec(\frac{1}{\sqrt{d}}) \tag{254}$$

and

$$\hat{\tau}_2 = \begin{bmatrix}1 & -\hat{\tau}_0\end{bmatrix}\frac{\partial}{\partial\rho_2}\Big[(\mathcal{C})^{-1}\Big]_{\rho=0}\begin{bmatrix}1\\ -\hat{\tau}_0\end{bmatrix} + \mathcal{O}_\prec(\frac{1}{\sqrt{d}}) \tag{255}$$

where:

$$\mathcal{C}^{-1} = A^*_{11} + \lambda I_{k+1} - (A^*_{21})^\top((\psi)^{-1} + \alpha S)^{-1}A^*_{21}, \tag{256}$$

and $\widetilde{A}^*_{21} \in \mathbb{R}^{k\times k+1}$ is defined analogous to A^*_{21} in Theorem 4.5 but with $u_1, \cdots, u_p$ replaced by $\zeta^u_1, \cdots, \zeta^u_k$ i.e:

$$A^*_{21}[j,:] = \alpha\,\mathbb{E}_\kappa\left[\frac{c_1(\kappa, \zeta^u_j)}{1 + \chi(z;\kappa)}\kappa\iota^\top\right],\ \forall j \in [k] \tag{257}$$

Proof. For τ^*_0, the result follows directly from Proposition E.2, Corollary E.4 and Lemma E.5. This further implies that $\hat{\tau}_0$ has entries $\mathcal{O}(1)$.

Therefore, we may again apply Proposition E.2 to obtain:

$$\hat{\tau}_2 = \begin{bmatrix}1 & -\tau^*_0\end{bmatrix}\frac{\partial}{\partial\rho_1}\Big[(C)^{-1}\Big]_{\rho=0}\begin{bmatrix}1\\ -\hat{\tau}_0\end{bmatrix} + \mathcal{O}(\frac{1}{p}) \tag{258}$$

and

$$\hat{\tau}_3 = \begin{bmatrix}1 & -\tau^*_0\end{bmatrix}\frac{\partial}{\partial\rho_2}\Big[(C)^{-1}\Big]_{\rho=0}\begin{bmatrix}1\\ -\hat{\tau}_0\end{bmatrix} + \mathcal{O}(\frac{1}{p}). \tag{259}$$

The expressions for $\hat{\tau}_2, \hat{\tau}_3$ in Equations (254),(258) are then direct consequences of Corollary E.4, Lemma E.5 and Lemma E.6 (to justify differentiating through the deterministic equivalent). It remains to obtain the resulting expression for $\hat{\tau}_1$. Applying Proposition E.2, and using $\hat{\tau}_0 = \mathcal{O}(1)$, we obtain:

$$\hat{\tau}_{1,j} = (\theta \odot e^j)^\top Q\begin{bmatrix}1\\ -\hat{\tau}_0\end{bmatrix} + \mathcal{O}(\frac{1}{\sqrt{p}}). \tag{260}$$

Next, applying Lemma E.5 and Corollary E.4 with $r = (\theta \odot e^j)$ yields:

$$\hat{\tau}_{1,j} = (\theta \odot e^j)^\top((\operatorname{diag}(\frac{b^\star}{\pi\beta})_e - \frac{1}{\beta^2}(\operatorname{diag}(b^\star) - \psi(V_d, \nu_d))_e \odot \frac{\theta}{\pi}\frac{\theta}{\pi}^\top)^{-1} + \alpha S_e \odot \theta\theta^\top)^{-1}\theta \odot A^\star_{21} \tag{261}$$

Subsequently, analogous to the derivation of $\mathcal{C}$ in Lemma E.5, we simplify the above expression using the Woodbury identity to obtain Equation (253). □

F Proofs of auxilary results

F.1 Proof of Lemma D.9

We proceed using the leave-one out argument for the Wishart spectrum. Let W_{-i} denote the weight matrix with the i_{th} column removed and define:

$$E_{w_i} \coloneqq w_i \odot [e_1, \cdots, e_k] \in \mathbb{R}^{p \times k}. \tag{262}$$

Then:

$$C_e \odot WW^\top = C_e \odot W_{-i}W_{-i}^\top + C_e \odot w_i w_i^\top. \tag{263}$$

Similar to the term $V_e \odot \theta\theta^\top$ term in Section D.3, we observe that $C_e \odot w_i w_i^\top$ is a rank k matrix generated through the vectors $e_1, \cdots, e_k$:

$$C_e \odot w_i w_i^\top = E_{w_i} C E_{w_i}^\top, \tag{264}$$

The Woodbury matrix identity yields:

$$R = R_{-i} - R_{-i}E_{w_i}((C)^{-1} + E_{w_i}^\top R_{-i}E_{w_i})^{-1}E_{w_i}^\top R_{-i}, \tag{265}$$

where:

$$R_{-i} = (C_e \odot W_{-i}W_{-i}^\top + (\operatorname{diag}(D)_e - zI_p))^{-1} \tag{266}$$

We substitute the above into the relation:

$$R(C_e \odot WW^\top + (\operatorname{diag}(D)_e - zI_p)) = I \tag{267}$$

to obtain:

$$\sum_{i=1}^{d} R_{-i}E_{w_i}CE_{w_i}^\top - \sum_{i=1}^{d} R_{-i}E_{w_i}((C)^{-1} + B_m)^{-1}B_m CE_{w_i}^\top + R(\operatorname{diag}(D)_e - zI_p) = I \tag{268}$$

Therefore, we obtain:

$$dR\mathbb{E}\left[E_w CE_w^\top\right] - \sum_{i=1}^{d} R_{-i}E_{w_i}(C^{-1} + B_m)^{-1}B_m CE_{w_i}^\top + R(\operatorname{diag}(D)_e - zI_p) = I + \Delta \tag{269}$$

where Δ includes error terms and B_m is an $k \times k$ matrix given by:

$$B_m = \mathbb{E}\left[E_w^\top R_{-i}E_w\right]. \tag{270}$$

Bounding the error terms exactly as in Proposition D.3, we obtain that contributions from δ are of order $\mathcal{O}_\prec(\frac{\|A\|_{\mathrm{Tr}}}{\sqrt{d}})$

Now, since $(C^{-1} + B_m)^{-1}B_m = I - (C^{-1} + B_m)^{-1}(C)^{-1}$,we obtain:

$$dRE_{w_i}((C)^{-1} + B_m)^{-1}E_{w_i}^\top + R(\operatorname{diag} D_e - zI_p) = I + \mathcal{O}_\prec(\frac{\|A\|_{\mathrm{Tr}}}{\sqrt{d}}).$$

which leads to the following deterministic equivalent for $R(C, D)$:

$$\mathcal{R} = (\mathcal{M} + (\operatorname{diag} D_e - zI_p))^{-1}, \tag{271}$$

where:

$$\mathcal{M} = d\mathbb{E}\left[E_w(C^{-1} + B_m)^{-1}E_w^\top\right], \tag{272}$$

To obtain the deterministic equivalent, it remains to next derive a self-consistent equation for the matrix B_m itself. Substituting Equation (272) into Equation (270) we obtain that B_m is diagonal with entries satisfying

$$b_q \coloneqq B_{q,q} = \pi_q\beta(((C)^{-1} + \operatorname{diag}(b))^{-1} + (\operatorname{diag}(D) - zI_p))_{q,q}^{-1}, \tag{273}$$

for $q \in [p]$.

Lastly, it remains to establish the uniqueness of the fixed points of b_q:

Lemma F.1. *Let $F_b : \mathbb{C}^+ \to \mathbb{C}^+$ denote the mapping:*

$$F(b)_q = \pi_q \beta(((C)^{-1} + \operatorname{diag}(b))^{-1} + (\operatorname{diag}(D) - zI_p))^{-1}_{q,q}, \tag{274}$$

for $q \in [k]$. Then for large enough ζ and C, D with entries in $\mathbb{C}^-$ satisfying $|C|_{ij} \leq K$ for some constant K independent of ζ :

$$\left\|F(b') - F(b)\right\|_2 < \frac{K'}{\zeta^2}\|b - b\|_2 \tag{275}$$

, for b, b' satisfying $|b|_q \leq \frac{\pi}{\beta\zeta}$ and some constant $K > 0$.

Proof. Lemma B.3 and Lemma B.12 along with the bounds on b, C, D imply:

$$\left\|F(b') - F(b)\right\|_2 \leq \frac{K_1}{\zeta^2}\left\|((C)^{-1} + \operatorname{diag}(b'))^{-1} - ((C)^{-1} + \operatorname{diag}(b))^{-1}\right\|, \tag{276}$$

for some constant $K_1 > 0$. Again applying Lemma B.3 and using the bounds on entries of C yields:

$$\left\|F(b') - F(b)\right\|_2 \leq \frac{K_2}{\zeta^2}\left\|b' - b\right\|_2 \tag{277}$$

□

The control of all the error terms and concentration bounds follows ezactly as in Proposition D.3

F.2 Equivalence between gradient update and the spiked random feature model

In this section, we discuss the regime $n_0 = \alpha_0 d$ for some constant α_0. In this regime, we show that the “bulk" in the gradient update possesses a non-isotropic component.

The starting point is the following relationship between the initial weight matrix W^0 and the new version W^1 after one gradient step:

$$W^1 = W^0 + (\eta p)G, \tag{278}$$

where G is the gradient matrix constructed as

$$G = \frac{1}{n\sqrt{p}} \operatorname{diag}\{a_1, \dots a_p\} \sigma'(W^0 X) \operatorname{diag}\{y_1, \dots, y_n\} X^T, \tag{279}$$

where X is an $d \times n_0$ matrix whose columns are the data vectors in the first batch, $\{y_i\}_{i\in[n]}$ are the corresponding labels, and $\sigma'(\cdot)$ is the derivative of the student activation function.

Consider the Hermite expansion of σ:

$$\sigma(x) = c_0 + c_1 h_1(x) + c_2 h_2(x) + \dots. \tag{280}$$

We can then write

$$\sigma'(x) = c_1 + \sigma'_{>1}(x), \tag{281}$$

where

$$\sigma'_{>1}(x) \overset{\text{def}}{=} \sum_{k\geq 2} \sqrt{k!} c_k \, h_{k-1}(x). \tag{282}$$

Using the decomposition in (281), we can rewrite the weight matrix W^1 as

$$W^1 = \underbrace{W^0 + \frac{\eta}{n_0} \operatorname{diag}\{\sqrt{p}a_1, \dots \sqrt{p}a_p\} \sigma'_{>1}(W^0 X) \operatorname{diag}\{y_1, \dots, y_n\} X^T}_{\widetilde{W}} + uv^\top, \tag{283}$$

where $u = c_1 c_1^\star \eta \sqrt{p} a$ and $v = \frac{1}{c_1^\star} Xy/n_0$, where $c_1^\star$ denotes the first Hermite coefficient of the target activation g. Next, we examine $\widetilde{W}$, which is the “bulk” of the weight matrix W^1 after one gradient step.

Observe that, conditioned on X and y, the rows of $\widetilde{W}$ are independent centered random vectors. For $i \in [n]$, the ith row of $\widetilde{W}$, denoted by $b_i \in \mathbb{R}^d$, is

$$b_i = w_i + \frac{\eta(\sqrt{p}a_i)}{n_0} X \operatorname{diag}\{y_1, \ldots, y_n\} \sigma'_{>1}(X^\top w_i). \tag{284}$$

Proposition F.2. *Let $\mathbb{E}_{w_i}[\cdot]$ denote the conditional expectation with respect to w_i, with X, $\{y_j\}_{j\in[n_0]}$ and a_i kept fixed. Then*

$$\begin{aligned} d \cdot \mathbb{E}_{w_i}\left[b_i b_i^\top\right] &= I + \frac{2\sqrt{2}c_2\eta(\sqrt{p}a_i)}{n_0} X \operatorname{diag}\{y_1, \ldots, y_n\} X^\top \\ &+ 2\eta^2 c_2^2 (\sqrt{p}a_i)^2 \left(\frac{1}{n} X \operatorname{diag}\{y_1, \ldots, y_{n_0}\} X^\top\right)^2 + \frac{6c_3^2\eta^2(\sqrt{p}a_i)^2}{n_0^2} X y y^\top X^\top \\ &+ \left(\sum_{k\geq 3} k! c_k^2\right) \eta^2 (\sqrt{p}a_i)^2 (d/n_0) \left(\frac{1}{n\,0} X \operatorname{diag}\left\{y_1^2, \ldots, y_{n_0}^2\right\} X^\top\right) + \Delta, \end{aligned} \tag{285}$$

where $\Delta \in \mathbb{R}^{d\times d}$ is small error term such that

$$\|\Delta\|_{\mathsf{op}} \prec d^{-1/2}. \tag{286}$$

Proof. From (284),

$$\begin{aligned} \mathbb{E}_{w_i}\left[b_i b_i^\top\right] &= \mathbb{E}_{w_i}\left[w_i w_i^\top\right] + \underbrace{\frac{\eta(\sqrt{p}a_i)}{n_0} X \operatorname{diag}\{y_1, \ldots, y_{n_0}\} \mathbb{E}_{w_i}\left[\sigma'_{>1}(X^\top w_i) w_i^\top\right]}_{(*)} + (*)^\top \\ &+ \frac{\eta^2(\sqrt{p}a_i)^2}{n_0^2} X \operatorname{diag}\{y_1, \ldots, y_{n_0}\} \mathbb{E}_{w_i}\left[\sigma'_{>1}(X^\top w_i)(\sigma'_{>1}(X^\top w_i))^\top\right] \operatorname{diag}\{y_1, \ldots, y_{n_0}\} X^\top. \end{aligned} \tag{287}$$

By symmetry, it is straightforward to check that $\mathbb{E}_{w_i}\left[w_i w_i^\top\right] = I/d$. To compute the last three terms on the right-hand side of the above equation, we first make the simplifying assumptions that $w_i \sim \mathcal{N}(0, I/d)$ and that each column of Z has a fixed norm equal to $\sqrt{d}$. Note that these assumptions are asymptotically accurate as $d \to \infty$. For finite d, we can absorb the errors introduced by these assumptions into the error matrix Δ. Under the above assumptions, it is easy to check that:

$$\mathbb{E}_{w_i}\left[\sigma'_{>1}(X^\top w_i) w_i^\top\right] = \frac{\sqrt{2}c_2}{d} X^\top \tag{288}$$

and

$$\mathbb{E}_{w_i}\left[\sigma'_{>1}(X^\top w_i)(\sigma'_{>1}(X^\top w_i))^\top\right] = \frac{2c_2^2}{d} X^\top X + \frac{6c_3^2}{d} 1 1^\top + \left(\sum_{k\geq 3} k! c_k^3\right) I_n + \widetilde{\Delta}, \tag{289}$$

where $\left\|\widetilde{\Delta}\right\|_{\mathsf{op}} \prec d^{-1/2}$. Substituting these estimates into (287) gives us (285), with

$$\Delta = \eta^2 (\sqrt{p}a_i)^2 (d/n_0) \left(\frac{1}{n\,0} X \operatorname{diag}\{y_1, \ldots, y_n\} \widetilde{\Delta} \operatorname{diag}\{y_1, \ldots, y_n\} X^\top\right). \tag{290}$$

Let us further evaluate the trace of the expression (285) in the setting of Cui et al. [2024], for odd activations (implying in particular $\mu_2 = 0$), and uniform initializations $\forall i, \ \sqrt{p}a_i = 1$. Since

$$\begin{aligned} \frac{1}{d} \operatorname{Tr}\left[\frac{1}{n_0} X \operatorname{diag}\left\{y_1^2, \ldots, y_n^2\right\} X^\top\right] &= \frac{1}{d} \sum_{i=1}^d \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}[g(w_\star^\top x)^2 x_i^2] \\ &= \frac{1}{d} \sum_{i=1}^d \left(2w_i^\star \mathbb{E}_{x\sim\mathcal{N}(0,I_d)}[g(w_\star^\top x) g'(w_\star^\top x) x_i] + \mathbb{E}_{z\sim\mathcal{N}(0,I_d)}[g(w_\star^\top z)^2]\right) \\ &\asymp \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}[g(\xi)^2] \end{aligned} \tag{291}$$

Thus,

$$\frac{1}{d}\operatorname{Tr}\left[\mathbb{E}_{w_i}\left[b_i b_i^\top\right]\right] = 1 + \left(\sum_{k\geq 3} k! c_k^2\right)\eta^2 \frac{1}{\alpha_0}\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}[g(\xi)^2] = 1 + \mathbb{E}_{\xi\sim\mathcal{N}(0,1)}[\sigma'_{>1}(\xi)^2]\eta^2\frac{1}{\alpha_0}\mathbb{E}_{\xi\sim\mathcal{N}(0,1)}, \tag{292}$$

which corresponds to the term c in Cui et al. [2024] up to conventions. □

F.3 Proof of Lemma B.14

Proof. We start by rewriting $g = z + w'(s - z^\top w')$, where $s \sim \mathcal{N}(0,1)$ and $z \sim \mathcal{N}(0,I)$ are independent. It follows that

$$f_1(w_1^\top g) f_2(w_2^\top g) g^\top w' = s f_1\big(w_1^\top z + (w_1^\top w')(s - z^\top w')\big) f_2\big(w_2^\top z + (w_2^\top w')(s - z^\top w')\big) \tag{293}$$

$$= s f_1(w_1^\top z) f_2(w_2^\top z) + s f_1(w_1^\top z) f_2'(w_2^\top x)(w_2^\top w')(s - z^\top w') \tag{294}$$

$$+ s f_2(w_2^\top z) f_1'(w_1^\top x)(w_1^\top w')(s - z^\top w') + \mathcal{O}_\prec(d^{-1}). \tag{295}$$

Taking expectation, and applying Lemma B.13, we get (42). The estimate in (43) can be treated similarly. Note that

$$f_1(w_1^\top g) f_2(w_2^\top g)(\theta^\top g)^2 = s^2 f_1(w_1^\top x) f_2(w_2^\top z) + \mathcal{O}_\prec(d^{-1/2}), \tag{296}$$

which immediately leads to (43). □