Asymptotics of Feature Learning in Two-layer Networks after One Gradient-step

Hugo Cui [1] **Luca Pesce** [2] **Yatin Dandi** [1 2] **Florent Krzakala** [2] **Yue M. Lu** [3] **Lenka Zdeborová** [1] **Bruno Loureiro** [4]

Abstract

In this manuscript, we investigate the problem of how two-layer neural networks learn features from data, and improve over the kernel regime, after being trained with a single gradient descent step. Leveraging the insight from (Ba et al., 2022), we model the trained network by a spiked Random Features (sRF) model. Further building on recent progress on Gaussian universality (Dandi et al., 2023), we provide an exact asymptotic description of the generalization error of the sRF in the high-dimensional limit where the number of samples, the width, and the input dimension grow at a proportional rate. The resulting characterization for sRFs also captures closely the learning curves of the original network model. This enables us to understand how adapting to the data is crucial for the network to efficiently learn non-linear functions in the direction of the gradient – where at initialization it can only express linear functions in this regime.

1. Introduction

A common deep learning intuition behind the unreasonable effectiveness of neural networks is their capacity to effectively adapt to the training data which makes them superior to kernel methods. While kernel methods and their finite width approximations are known to be data-hungry (see e.g. (Adlam et al., 2023)), neural networks have proven themselves to be flexible and efficient in practice. Their adaptivity and capacity to learn features from data are behind the success in efficiently solving problems from image classification to text generation. A large part of our current theoretical understanding of neural networks stems from the investigation of their lazy regime where features are *not* learned during training. This includes a set of works that fall under the umbrella of Gaussian processes (Neal, 1996; Lee et al., 2018), the Neural Tangent Kernel (NTK) (Jacot et al., 2018) and the Lazy regime (Chizat et al., 2019). A crucial question in the theoretical machine learning community is thus to characterize the advantages of two-layer neural networks beyond these convex optimization approaches.

Different theoretical works have offered sharp separation results between kernel and feature learning regimes (see e.g. (Ghorbani et al., 2019; Refinetti et al., 2021b; Damian et al., 2022; Abbe et al., 2023; Shi et al., 2022)). In particular, (Ba et al., 2022; Damian et al., 2022), and later (Dandi et al., 2023), discussed the advantage of neural networks when training with only *one single step* of large-batch gradient descent with a large learning rate. Specifically, (Ba et al., 2022) highlighted that the weight matrix after the first training step can be decomposed in a bulk plus a rank-one spike, effectively mapping the learned features to a *spiked Random Features model* (sRF), defined in eq. 6. This observation has fueled many further studies on the effect of spiked structure, see e.g. (Damian et al., 2022; Dandi et al., 2023; Ba et al., 2024; Mousavi-Hosseini et al., 2023; Moniri et al., 2023).

In this paper, we follow this line of work and provide an exact high-dimensional description of the test error achieved by a sRF, which we use to model a two-layer network after a *single, large, gradient step*. Our work provides a sharp asymptotic treatment of a setting where feature learning is modeled in a non-perturbative, high-dimensional regime, with a model able to express non-linear functions beyond polynomials. This analysis quantitatively illustrates the benefits of feature learning over the lazy regime.

Our **main contributions** in this paper are the following:

- **Exact asymptotics for sRF** – We provide a sharp asymptotic characterization of the test error, alongside a set of summary statistics, for a sRF. We discuss analytically, and provide numerical support, why sRFs constitute good approximations for two-layer neural networks with first layer weights trained with a large learning rate gradient step. The derivation leverages the (non-rigorous) replica method from statistical physics (Parisi, 1979; 1983), and provides a set of

---

*Equal contribution [1] Statistical Physics Of Computation lab., École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland [2]Information Learning & Physics lab., École Polytechnique Fédérale de Lausanne (EPFL), 1015 Lausanne, Switzerland [3]Harvard University, School of Engineering and Applied Sciences [4]Département d'Informatique, École Normale Supérieure (ENS) - PSL & CNRS, F-75230 Paris cedex 05, France. Correspondence to: Hugo Cui <hugo.cui@epfl.ch>.

scalar self-consistent equations for the generalization error.

• **Conditional Gaussian Equivalence** – Building upon (Dandi et al., 2023), we show (and provide strong numerical evidence to support) that the learning properties of the sRF model are asymptotically equivalent to a simple *conditional* Gaussian model in the high-dimensional proportional regime. The conditional Gaussian distribution is characterized by the projections of the input data on the spike in the weight matrix. This mapping constitutes the extension of related theoretical results that unveiled a similar Gaussian equivalence property for the training and generalization error for non-spiked vanilla RFs (Goldt et al., 2022; 2020; Gerace et al., 2020; Hu & Lu, 2022; Dhifallah & Lu, 2020; Mei & Montanari, 2022; Cui et al., 2023; Schröder et al., 2023; Bosch et al., 2023; Dandi et al., 2024).

• **Feature learning** – We provide an extensive discussion on how feature learning leads to a drastic improvement in the generalization performance over random features in a data-limited regime, demonstrating a clear and quantitative separation with respect to kernel method and random feature models. In particular, we derive both upper and lower bounds on the generalization error and discuss under which conditions they are tight.

The code used in the present work is available here.

2. Setting, Motivation and Related Work

Setting — We study fully-connected two-layer networks

$$f_{W,\boldsymbol{a}}(\boldsymbol{x}) = \frac{1}{\sqrt{p}} \sum_{i=1}^{p} a_i \sigma(\boldsymbol{w}_i^\top \boldsymbol{x}), \tag{1}$$

and their capacity to learn a single-index target function of isotropic Gaussian covariates:

$$f_\star(\boldsymbol{x}) = \sigma_\star(\boldsymbol{\theta}^\top \boldsymbol{x}/\sqrt{d}), \quad \boldsymbol{x} \sim \mathcal{N}(0, I_d) \tag{2}$$

from finite batch $\mathcal{D} = \{(\boldsymbol{x}^\mu, y^\mu)_{\mu=1}^n\}$ of n independently drawn training samples. While we do not consider bias terms for conciseness, we discuss in Appendix C how such terms can be treated. We further assume the activation σ to be odd. We consider a layer-wise training procedure where the first layer weights $W \in \mathbb{R}^{p \times d}$ are trained for a single gradient step:

$$\boldsymbol{w}_i^{(1)} = \boldsymbol{w}_i^{(0)} - \eta \boldsymbol{g}_i \tag{3}$$

$$\boldsymbol{g}_i = \frac{1}{\sqrt{p} n_0} \sum_{\mu=1}^{n_0} (y^\mu - f_{W^{(0)}, \boldsymbol{a}^{(0)}}(\boldsymbol{x}^\mu)) a_i^{(0)} \sigma'(\boldsymbol{w}_i^{(0)\top} \boldsymbol{x}^\mu) \boldsymbol{x}^\mu$$

on a subset $\mathcal{D}_0 \subset \mathcal{D}$ of size n_0, where $(W^{(0)}, \boldsymbol{a}^{(0)})$ denote the initial weights and $\eta > 0$ the learning rate. For simplicity, we assume $\boldsymbol{a}^{(0)} = \mathbb{1}_p/\sqrt{p}$ (uniform initialization) and $\boldsymbol{w}_i^{(0)}$ with unit norm $\|\boldsymbol{w}_i^{(0)}\| = 1$ and weak correlation $\boldsymbol{w}_i^{(0)} \cdot$

$\boldsymbol{w}_j^{(0)} = O_d(\text{polylog}(d)/\sqrt{d})$ for $i \neq j$ (e.g. uniformly drawn from the unit sphere $\mathbb{S}^{d-1}$). Finally, we assume that the second layer initialization is not informed, in the sense that its overlap with the target weights $\boldsymbol{\theta}$ is asymptotically small, $\boldsymbol{a}^{(0)} \cdot \boldsymbol{\theta}/\sqrt{p} = O_d(1/\sqrt{d})$. Given the updated weights $W^{(1)}$, we train the read-out layer on the remaining data $\mathcal{D}_1 = \mathcal{D} \setminus \mathcal{D}_0$:

$$\hat{\boldsymbol{a}}_\lambda = \underset{\boldsymbol{a} \in \mathbb{R}^p}{\arg\min} \frac{1}{2} \sum_{\mu=1}^{n_1} \left(y^\mu - f_{W^{(1)}, \boldsymbol{a}}(\boldsymbol{x}^\mu) \right)^2 + \frac{\lambda}{2} \|\boldsymbol{a}\|^2. \tag{4}$$

with $\lambda \in \mathbb{R}_+$ being a regularization parameter. Note that the layer-wise training procedure considered here is commonly studied in the theoretical machine learning literature (Ba et al., 2022; Damian et al., 2022; Abbe et al., 2023; Berthier et al., 2023; Dandi et al., 2023; Moniri et al., 2023) due to its mathematical tractability.

Our main goal in the following is to describe the generalization error:

$$\epsilon_g = \mathbb{E}_{\mathcal{D}, \boldsymbol{x}} \left(f_\star(\boldsymbol{x}) - f_{W^{(1)}, \hat{\boldsymbol{a}}_\lambda}(\boldsymbol{x}) \right)^2. \tag{5}$$

in the high-dimensional proportional limit where $n_0, n_1, d, p, \eta \to \infty$ at fixed ratios $\alpha_0 = n_0/d$, $\alpha = n_1/d, \beta = p/d, \tilde{\eta} = \eta/d$.

Motivation — Driven by the lazy-training regime of learning of large-width networks (Chizat et al., 2019), a large body of literature has been dedicated to the particular case where the first layer weights are fixed at initialization $W^{(0)}$ ($\eta = 0$), also known as the Random Features (RF) model (Rahimi & Recht, 2007). In particular, (Ghorbani et al., 2019; 2020; Mei et al., 2022) have shown that with $n_1 = \Theta_d(d)$ samples $f_{W^{(0)}, \hat{\boldsymbol{a}}_\lambda}$ can only approximate, at best, a linear function of $\boldsymbol{x}^\mu$, with the non-linear part playing a role akin to additive label noise. This is a strong limitation: RF network requires *polynomials* number of data and neurons to fit a simple polynomial (Mei et al., 2022; Xiao et al., 2022). It is one of our motivations here to discuss how, with a single gradient step, these limitations are lifted.

Behind the scenes in this effective linearity of random features is a Gaussian Equivalent Principle (GEP) (Goldt et al., 2022; Mei & Montanari, 2022; Hu & Lu, 2022; Montanari & Saeed, 2022; Dandi et al., 2024), which states that in this regime the random feature map $\varphi = \sigma(W^{(0)} \boldsymbol{x})$ is statistically equivalent to a rescaled stochastic linear map $\varphi^g(\boldsymbol{x}) \asymp \mu_0 \mathbb{1} + \mu_1 W^{(0)} \boldsymbol{x} + \mu_\star \boldsymbol{z}$, with $\boldsymbol{z} \sim \mathcal{N}(0, I_p)$. This surprising universality result allows to go beyond lower bounds for the generalization performance, making the problem amenable to a tight high-dimensional characterization of all relevant statistics in these models (Mei & Montanari, 2022; Gerace et al., 2020; Dhifallah & Lu, 2020).

Fig. 1 illustrates this fundamental limitation of RF models contrasted with the function $f(x_\theta) =$
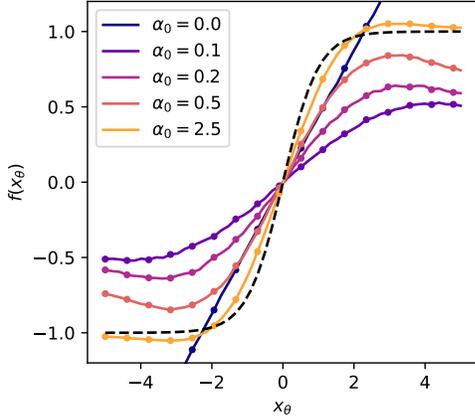
Figure 1. Numerical estimation of the function $f(x_\theta) = \mathbb{E}_x[f_{W^{(1)}, \hat{a}_\lambda} | \theta^\top x/\sqrt{d} = x_\theta]$ implemented by the trained network (1) in the direction spanned by the weights θ of the target function. The activations are $\sigma = \sigma_\star = \tanh$, and simulations were run in dimensions $d = p = 2000$, for a learning rate $\eta = 2.5p$, and a readout regularization $\lambda = 0.01$. The readout was trained with $n_1 = 2d$ samples. Different colors corresponds to different sample complexities $\alpha_0 \equiv n_0/d$ used to implement the gradient step on the first layer weights, with $\alpha_0 = 0$ corresponding to not implementing the step.

$\mathbb{E}_x[f_{W^{(1)}, \hat{a})\lambda}(x) | \theta^\top x/\sqrt{d} = x_\theta]$, along the direction of the target θ, implemented by the network $f_{W^{(1)}, a}$ trained with a single gradient step (eq. (3)). Varying the amounts of data $n_0 = \alpha_0 d$ used in the first gradient step, the function $f(x_\theta)$ moves from a linear approximation of $\sigma_\star$ in the RF limit ($\alpha_0 = 0$), to an accurate non-linear one ($\alpha_0 = 2.5$).

Learning a non-linear approximation of $\sigma_\star$ in the high-dimensional proportional regime therefore requires learning features. (Ba et al., 2022) have proven that with $\eta = \Theta_d(1)$, the GEP holds even after a few gradient steps, corroborating a fact empirically observed by (Loureiro et al., 2021b). Indeed, they have shown that $\eta = \Theta_d(d)$ is *sufficient* to go beyond a linear approximation of $f_\star$ in this regime. (Moniri et al., 2023) considered intermediate scalings of step-size $\eta = \Theta_d(d^s)$ for $1/2 < s < 1$, which allows the network to fit target functions along θ having finite degree, providing a precise characterization of the train and test errors. In this intermediate regime, the feature matrix can be approximated through a finite number of spikes corresponding to increasing degree of functions along θ. Instead, we consider the full scaling of $\eta = \Theta_d(d)$, where such a finite-spike approximation is insufficient and the network can fit arbitrary functions along θ, and provide an approximate characterization through the lens of sRF models. (Dandi et al., 2023) proved that even if the target depends on multiple directions (multi-index model), only a (non-linear) function of a *single direction* θ can be learned with a single gradient step and $\eta = \Theta_d(d)$. This observation justifies the focus on

single-index functions (2) on the regime of interest.

Further related works – Random features were first introduced as a computationally efficient approximation to kernel methods (Rahimi & Recht, 2007). Recently, they have enjoyed renewed interest also as models of two-layer neural networks in the lazy regime. Tight asymptotics for the random features model have been derived by (Goldt et al., 2020; 2022; Gerace et al., 2020; Mei & Montanari, 2022; Hu & Lu, 2022; Dhifallah & Lu, 2020) in the two-layer case, and were extended to deep networks in (Schröder et al., 2023; Schröder et al., 2024; Bosch et al., 2023) in the deep case. Importantly, with the exception of (Gerace et al., 2020) who considered rotationally invariant weights and (Zavatone-Veth & Pehlevan, 2023) for the case of deep *linear* random features, all these works assumed unstructured weights. In sharp contrast, gradient-trained neural networks have fundamentally structured weights. In the present manuscript, we consider such a case, by analyzing sRF models, for which the weights are given by a bulk random matrix plus a rank-one spike.

Feature learning regime – Perturbative feature learning corrections to the large-width lazy regime have been extensively studied in the literature (Yaida, 2020; Hanin & Nica, 2020; Dyer & Gur-Ari, 2020; Seroussi et al., 2023; Naveh & Ringel, 2021). Our work radically contrasts with this line, since we account for feature learning in the first step, *non-perturbatively* (note the gradient in (3) has a norm comparable with the initial weights). Beyond the lazy regime, a major recent development has been the understanding that the training dynamics of two-layer neural networks with small learning rates can be mapped to a Wasserstein gradient flow, known as the *mean-field regime* (Mei et al., 2018; Chizat & Bach, 2018; Rotskoff & Vanden-Eijnden, 2022; Sirignano & Spiliopoulos, 2020; Bordelon & Pehlevan, 2024). Over the past few years, this flow was investigated under different classes of generative data models, such as staircase functions (Abbe et al., 2021; 2022; 2023), single-index (Berthier et al., 2023; Arnaboldi et al., 2023a) and multi-index models (Arnaboldi et al., 2023b), symmetric targets (Hajjar & Chizat, 2022) and Gaussian mixture models (Refinetti et al., 2021a; Ben Arous et al., 2022).

3. Main Technical Results

Our main technical results are a tight asymptotic characterization of the test error achieved by sRFs (defined in 6), which we argue provide good models for two-layer networks trained with a single large gradient step followed by a ridge regression on the readout weights. We provide compelling numerical support that this theoretical characterization captures very closely the learning curves of two-layer networks trained following the protocol (3), and that sRF thus provide a valuable analytical playground

to understand the learning and behaviour of the latter. These results are enabled through the expression of the parameters of an equivalent sRF model in terms of the training parameters of the original network in subsection 3.1, which can in turn be mapped to an equivalent Gaussian model in subsection 3.2. Subsection 3.3 finally states the tight asymptotic characterization of the test error.

3.1. Asymptotics of the first layer weights after one (large) gradient step

The first step is to derive an explicit asymptotic expression for the hidden-layer weights $W^{(1)}$ after one (large) gradient step. In the following, we show that the learning problem introduced in Section 2 can be modelled by a *spiked Random Features* model, which we first define.

Definition 3.1 (sRF model). We define a *spiked Random Features* (sRF) model with bulk variance c, spike strength r as the two-layer neural network

$$g_{F,\boldsymbol{a}}(x) = \frac{1}{\sqrt{p}} \boldsymbol{a}^\top \sigma\left(Fx\right) \tag{6}$$

with trainable readout $\boldsymbol{a}$ and frozen random first layer weights:

$$F = W + r \frac{\boldsymbol{u}\boldsymbol{v}^\top}{\sqrt{d}}. \tag{7}$$

where W is a random matrix with rows independently sampled from $\mathbb{S}^{d-1}(\sqrt{c})$, and $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{S}^{d-1}(\sqrt{d})$. We further say that a sRF has alignment γ with $\boldsymbol{\theta}$ when $\boldsymbol{v}$ is uniformly sampled uniformly sampled among vectors with norm $\sqrt{d}$ satisfying $\boldsymbol{v}^\top \boldsymbol{\theta}/d = \gamma$.

As discussed above, after a large gradient step, the first layer weights $W^{(1)}$ take the form (Ba et al., 2022; Dandi et al., 2023):

$$W^{(1)} = W^{(0)} + \Delta + r \frac{\check{\boldsymbol{u}}\check{\boldsymbol{v}}^\top}{\sqrt{d}}, \tag{8}$$

where the precise expressions of $\Delta, \check{\boldsymbol{u}}, \check{\boldsymbol{v}}$ are detailed in Appendix A. It has thus developed spikes $\check{\boldsymbol{u}}$ and $\check{\boldsymbol{v}}$ that correlated to the target weights $\boldsymbol{\theta}$, while the bulk, originally $W^{(0)}$, also gets modified by an additional term Δ and displays as a consequence a rescaled variance. These changes make it reasonable to model the two-layer neural network with a sRF with matching parameters c, r, γ. The expression of these parameters depend on the specifications $\tilde{\eta}, \alpha_0, \beta, \sigma, \sigma_\star$ of the network and training protocol. Expliciting these expressions is the object of the following result.

Result 3.1 (Equivalent sRF model). *Consider two-layer networks with first-layer weights trained with a single gradient step of learning rate η from initial conditions $\boldsymbol{a}^{(0)} = \mathbf{1}_p/\sqrt{p}$ and $\boldsymbol{w}_i^{(0)}$ with unit norm $\|\boldsymbol{w}_i^{(0)}\| = 1$ and*

weak correlation $\boldsymbol{w}_i^{(0)} \cdot \boldsymbol{w}_j^{(0)} = O_d(1/\sqrt{d})$ for $i \neq j$ (eq. (3)). In the asymptotic limit $n_0, n, d, p \to \infty$, with $\alpha_0 = n_0/d, \alpha = n/d, \beta = p/d, \tilde{\eta} = \eta/d = \Theta_d(1)$, the quantities $c = 1/pd\|W_0 + \Delta\|^2$, $r = \|\check{\boldsymbol{u}}\|\|\check{\boldsymbol{v}}\|/\sqrt{pd}$, $\gamma = \boldsymbol{\theta}^\top \check{\boldsymbol{v}}/\|\check{\boldsymbol{v}}\|\|\boldsymbol{\theta}\|$ (8) concentrated to the values

$$c = 1 + \frac{\tilde{\eta}^2 \check{h}_1^2 h_2^\star}{\alpha_0 \beta^2} \tag{9}$$

$$r = \frac{\tilde{\eta} h_1}{\beta} \left(\frac{h_2^\star}{\alpha_0} + h_1^{\star 2} \right)^{1/2} \tag{10}$$

$$\gamma = \frac{h_1^\star}{\left(\frac{h_2^\star}{\alpha_0} + h_1^{\star 2} \right)^{1/2}} \tag{11}$$

where:

$$h_1 = \mathbb{E}_z[z\sigma(z)], \qquad h_1^\star = \mathbb{E}_z[z\sigma_\star(z)],$$
$$\check{h}_1^2 = \mathbb{E}_z[(\sigma'(z) - h_1)^2], \quad h_2^\star = \mathbb{E}_z[\sigma_\star(z)^2], \tag{12}$$

with $z \sim \mathcal{N}(0,1)$. These parameters c, r, γ specify an equivalent sRF, which we take as a model of the network after one gradient step.

The derivation of Result (3.1) is detailed in Appendix A. One may wonder what are the sources of approximation when modeling two-layer networks after a gradient step (3) by a sRF 6. In considering the equivalent sRF, one is effectively modeling the bulk term Δ in (8) by a random matrix with independent rows sampled from the sphere, with matching norm. This approximation is reasonable in view of the results of (Dandi et al., 2023) (Lemma 12), who rigorously establish the asymptotic near-orthogonality of the rows of Δ, and that they have vanishing overlap with $\boldsymbol{\theta}$. We are, on the other hand, ignoring possible non-trivial structures in the row covariance of Δ. However, as we shall discuss later, for all probed usual activation functions and setups, the sRF provides a quantitatively close approximation. Let us further stress that beyond being approximations of two-layer networks after a gradient step, sRF models constitute an important and natural generalization of the celebrated and extensively probed RF models (Rahimi & Recht, 2007; Gerace et al., 2020; Mei & Montanari, 2022; Hu & Lu, 2022), which crucially allow to learn *non-linear* functions – where RFs can only express linear functions. They hence are also models of standalone interest. Finally, note that above we have assumed a uniform initialization for the readout layer. This can be relaxed in the equivalence above, for instance for $\boldsymbol{a}^{(0)}$ taking a finite number of values, leading to a finite-rank term instead of a single spike. The emergence or usefulness of low-rank structures in network weights has been observed and discussed in numerous empirical and theoretical works, e.g. (Hu et al., 2021; Allen-Zhu & Li, 2022; Wang et al., 2022).

3.2. Conditional Gaussian equivalence

The sharp characterization of the test performance, which we state in Result 3.3 in the following subsection, is enabled by further mapping the sRF model to an exactly solvable (conditional) Gaussian model. We adapt the rigorous result in (Dandi et al., 2023) (Theorem **4**, second point) by constructing explicitly the equivalent stochastic feature map, that we believe is of independent interest.

Result 3.2 (Conditional Gaussian Equivalence). *Consider the sRF model with weights $F = W + r\boldsymbol{u}\boldsymbol{v}^\top/\sqrt{d}$, with $\boldsymbol{u} = \mathbf{1}_p$ and parameters c, r, γ, and the corresponding feature map given by*

$$\boldsymbol{\varphi}(\boldsymbol{x}) = \sigma\left(F\boldsymbol{x}\right) \tag{13}$$

Define the equivalent stochastic feature map

$$\boldsymbol{\varphi}^g(\boldsymbol{x}) \overset{d}{=} \mu_0(\kappa)\mathbf{1}_p + \mu_1(\kappa)W\boldsymbol{x} + \mu_2(\kappa)\mathcal{N}(0, \mathbb{I}_p), \tag{14}$$

where $\kappa \equiv \boldsymbol{v}^\top\boldsymbol{x}/\sqrt{d}$. We introduced the coefficients $\mu_0(\kappa), \mu_1(\kappa), \mu_2(\kappa)$ defined as

$$
\begin{aligned}
\mu_0(\kappa) &= \mathbb{E}_z\, \sigma(z + r\kappa), \\
\mu_1(\kappa) &= \frac{1}{c}\,\mathbb{E}_z\, z\sigma(z + r\kappa), \\
\mu_2(\kappa) &= \sqrt{\mathbb{E}_z\, \sigma^2(z + r\kappa) - c(\mu_1(\kappa))^2 - (\mu_0(\kappa))^2},
\end{aligned}
\tag{15}
$$

with expectations bearing over $z \sim \mathcal{N}(0, c)$. The test error ϵ_g achieved by ridge regression

$$\hat{\boldsymbol{a}}_\lambda = \operatorname*{argmin}_{\boldsymbol{a}\in\mathbb{R}^p} \frac{1}{2}\sum_{\mu=1}^{n_1}\left(y^\mu - \frac{1}{\sqrt{p}}\boldsymbol{a}^\top\phi(\boldsymbol{x})\right)^2 + \frac{\lambda}{2}\|\boldsymbol{a}\|^2. \tag{16}$$

is asymptotically identical for $\phi = \boldsymbol{\varphi}$ and $\phi = \boldsymbol{\varphi}^g$.

Result 3.2 extends the similar linearizations provided e.g. in (Goldt et al., 2020; Hu & Lu, 2022; Cui et al., 2023) for unstructured RFs to sRFs (7). Informally, the quantity κ in the stochastic feature map (14) represents the projection of the input on the spike defining the sRF $\kappa = \boldsymbol{x}^\top\boldsymbol{v}/\sqrt{d}$. The equivalent network $1/\sqrt{d}\boldsymbol{a}^\top\boldsymbol{\varphi}^g(x)$ obtained by replacing $\boldsymbol{\varphi}$ by the equivalent $\boldsymbol{\varphi}^g$ is a linear combination of terms such as $\mu_0(\kappa), \mu_1(\kappa)\kappa, \mu_1(\kappa)\boldsymbol{a}^\top W(\Pi^\perp\boldsymbol{x})$, plus noise. On an intuitive level, this makes it apparent that sRFs can thus express *non-linear* functions of the component κ along the spike $\boldsymbol{v}$, but only linear functions of the component $\Pi^\perp\boldsymbol{x}$ orthogonal thereto. This feature learning correction to the linear regime is visually exemplified in Fig. 1. In the next subsection, we make this discussion more quantitative by providing a tight asymptotic characterization of the test error achieved by two-layer networks trained with a single large gradient step followed by a ridge regression on the readout weights.

3.3. Tight asymptotic characterization of the test error

Finally, we leverage on the sequential mappings of Result 3.1 and Result 3.2 to offer sharp asymptotic guarantees on the test error achieved after training the readout weights using eq. (4).

Assumption 3.2. Denote by $\{\boldsymbol{e}_i\}_{i=1}^p$ ($\{\boldsymbol{f}_i\}_{i=1}^d$) the left (resp. right) singular vectors of W. We further note $\{\lambda_i^\ell\}_{i=1}^p$ the squared singular values of W. The squared singular values $\{\lambda_i^\ell\}_{i=1}^p$ and the projection of the teacher vector $\boldsymbol{\theta}$ and the spike v on the eigenvectors $\{\boldsymbol{f}_i^\top v\}_{i,\ell}, \{\boldsymbol{f}_i^\top\Pi^\perp\boldsymbol{\theta}\}_{i,\ell}$ are assumed to admit a well-defined joint distribution ν as $d \to \infty$.

$$\frac{1}{p}\sum_{i=1}^{\min(p,d)}\delta\left(\lambda_i - \varrho\right)\delta\left(\boldsymbol{f}_i^\top v - \tau\right)\delta\left(\boldsymbol{f}_i^\top\Pi^\perp\boldsymbol{\theta} - \pi\right)\xrightarrow{d\to\infty}\nu(\varrho, \tau, \pi).$$

Result 3.3 (Test error asymptotics). *Consider the ERM problem associated with the training of the readout weights $\boldsymbol{a}$ of the equivalent sRF model defined by Result 3.1 (namely replacing the network $f_{W^{(1)}\boldsymbol{a}}$ by the equivalent sRF $g_{F,\boldsymbol{a}}$ in (4)), and assume 3.2 to hold. Define $\Pi^\perp \equiv \mathbb{I}_d - vv^\top/d$ the projection to the subspace orthogonal to the spike $\boldsymbol{v}$. In the asymptotic limit $d, p, n \to \infty$ with $\alpha = n/d, \beta = p/d = \Theta_d(1)$, the summary statistics*

$$
\begin{aligned}
q_1 &= \frac{\hat{\boldsymbol{a}}^\top W\Pi^\perp W^\top\hat{\boldsymbol{a}}}{p}, & q_2 &= \frac{\hat{\boldsymbol{a}}^\top\hat{\boldsymbol{a}}}{p}, \\
m &= \frac{\mathbf{1}_p^\top\hat{\boldsymbol{a}}}{\sqrt{p}}, & \zeta &= \frac{\hat{\boldsymbol{a}}^\top W\boldsymbol{v}}{\sqrt{dp}}, \\
\psi &= \frac{\hat{\boldsymbol{a}}^\top W\Pi^\perp\boldsymbol{\theta}}{\sqrt{dp}}, & \rho^2 &= \frac{\boldsymbol{\theta}^\top\Pi^\perp\boldsymbol{\theta}}{d}
\end{aligned}
\tag{17}
$$

concentrate in probability to the solutions of the system of equations

$$
\begin{cases}
q_1 = \int d\nu(\varrho, \tau, \pi)\varrho\frac{\left(\hat{q}_1\varrho + \hat{q}_2 + \hat{\zeta}^2\varrho\tau^2 + \hat{\psi}^2\varrho\pi^2\right)}{\left(\lambda + \hat{V}_1\varrho + \hat{V}_2\right)^2} \\
\quad -\beta\hat{\zeta}^2\frac{I(\hat{V}_1, \hat{V}_2)^2}{\left(1 - \beta\hat{V}_1 I(\hat{V}_1, \hat{V}_2)\right)^2} \\
\quad -\hat{\zeta}^2\frac{\int d\nu(\varrho,\tau,\pi)\frac{\tau^2\varrho^2}{(\lambda+\hat{V}_1\varrho+\hat{V}_2)^2}\left[\left(1-\beta\hat{V}_1 I(\hat{V}_1,\hat{V}_2)\right)^2 - 1\right]}{\left(1-\beta\hat{V}_1 I(\hat{V}_1,\hat{V}_2)\right)^2} \\
q_2 = \int d\nu(\varrho, \tau, \pi)\frac{\left(\hat{q}_1\varrho + \hat{q}_2 + \hat{\zeta}^2\varrho\tau^2 + \hat{\psi}^2\varrho\pi^2\right)}{\left(\lambda + \hat{V}_1\varrho + \hat{V}_2\right)^2} \\
\quad -\hat{\zeta}^2\int d\nu(\varrho,\tau,\pi)\frac{\tau^2\varrho}{(\lambda+\hat{V}_1\varrho+\hat{V}_2)^2}\left[1 - \frac{1}{\left(1-\beta\hat{V}_1 I(\hat{V}_1,\hat{V}_2)\right)^2}\right] \\
V_1 = \int d\nu(\varrho, \tau, \pi)\varrho\frac{1}{\lambda+\hat{V}_1\varrho+\hat{V}_2} \\
V_2 = \int d\nu(\varrho, \tau, \pi)\frac{1}{\lambda+\hat{V}_1\varrho+\hat{V}_2} \\
m = \frac{1}{\mathbb{E}_\kappa\left[\frac{\mu_0(\kappa)^2}{1+V(\kappa)}\right]}\mathbb{E}_{\kappa,y}\left[\frac{\mu_0(\kappa)(\sigma_\star(\kappa,y)-\mu_1(\kappa)\kappa\zeta)}{1+V(\kappa)}\right] \\
\zeta = \hat{\zeta}\sqrt{\beta}\int d\nu(\varrho, \tau, \pi)\varrho\tau^2\frac{1}{\lambda+\hat{V}_1\varrho+\hat{V}_2} \\
\quad +\beta^{3/2}\hat{\zeta}\hat{V}_1\frac{I(\hat{V}_1,\hat{V}_2)^2}{1-\beta\hat{V}_1 I(\hat{V}_1,\hat{V}_2)} \\
\psi = \hat{\psi}\sqrt{\beta}\int d\nu(\varrho, \tau, \pi)\varrho\pi^2\frac{1}{\lambda+\hat{V}_1\varrho+\hat{V}_2}
\end{cases}
\tag{18}
$$

$$
\begin{cases}
\hat{V}_1 = \frac{\alpha}{\beta} \mathbb{E}_\kappa \frac{\rho \mu_1(\kappa)^2}{1+V(\kappa)} \\[4pt]
\hat{q}_1 = \frac{\alpha}{\beta} \mathbb{E}_{\kappa,y} \mu_1(\kappa)^2 \frac{b(\kappa,y)^2+\rho q(\kappa)-\mu_1(\kappa)^2\psi^2}{\left(1+V(\kappa)\right)^2} \\[4pt]
\hat{V}_2 = \frac{\alpha}{\beta} \mathbb{E}_\kappa \frac{\rho \mu_2(\kappa)^2}{1+V(\kappa)} \\[4pt]
\hat{q}_2 = \frac{\alpha}{\beta} \mathbb{E}_{\kappa,y} \mu_2(\kappa)^2 \frac{b(\kappa,y)^2+\rho q(\kappa)-\mu_1(\kappa)^2\psi^2}{\left(1+V(\kappa)\right)^2} \\[4pt]
\hat{\zeta} = \frac{\alpha}{\sqrt{\beta}} \mathbb{E}_{\kappa,y} \kappa \mu_1(\kappa) \frac{b(\kappa,y)}{1+V(\kappa)} \\[4pt]
\hat{\psi} = \frac{\alpha}{\sqrt{\beta}} \mathbb{E}_{\kappa,y} \frac{y\mu_1(\kappa)b(\kappa,y)+\psi\mu_1(\kappa)^2}{1+V(\kappa)}
\end{cases}
\tag{19}
$$

and

$$
\rho^2 = 1 - \gamma^2. \tag{20}
$$

We introduced the shorthands

$$
b(y,\kappa) \equiv \sigma_\star(\gamma\kappa + \sqrt{1-\gamma^2}y) - \mu_0(\kappa)m
$$
$$
- \kappa\mu_1(\kappa)\zeta - \mu_1(\kappa)\psi y \tag{21}
$$
$$
V(\kappa) \equiv \mu_1(\kappa)^2 V_1 + \mu_2(\kappa)^2 V_2 \tag{22}
$$
$$
q(\kappa) \equiv \mu_1(\kappa)^2 q_1 + \mu_2(\kappa)^2 q_2, \tag{23}
$$
$$
I(\hat{V}_1, \hat{V}_2) = \int d\nu(\varrho,\tau,\pi) \frac{\tau^2 \varrho}{\hat{V}_1\varrho + \hat{V}_2 + \lambda} \tag{24}
$$

In (18), the expectations over κ, y bear over standard Gaussian variables. We remind that the quantities r, c, γ are characterized in Result 3.1.

Finally, the test error (5) admits the sharp characterization

$$
\epsilon_g = \mathbb{E}_{\kappa,y}\Bigg[\Bigg(\sigma_\star(\gamma\kappa+\sqrt{1-\gamma^2}y) - \mu_0(\kappa)m - \mu_1(\kappa)\kappa\zeta
$$
$$
- \frac{\mu_1(\kappa)\psi}{\sqrt{\rho}}y\Bigg)^2 + q(\kappa) - \frac{\mu_1(\kappa)^2\psi^2}{\rho}\Bigg] \tag{25}
$$

where expectations bear over standard Gaussian variables.

Result 3.3 thus rephrases the high-dimensional learning problem (4) in terms of a finite set of scalar summary statistics which can be efficiently evaluated, yielding excellent agreement with finite d numerical simulations, see Figs. 2-6. We emphasize that while the solid lines indicate the theoretical characterization of the test error achieved by the sRF, the numerical experiments were performed for the *original two-layer network setting*. The good agreement thus further supports that sRFs constitute good models of the latter. While we state Result 3.3 for the square loss and ℓ_2 regularization for clarity, we provide a sharp characterization for any generic convex loss ℓ in Appendix C.

As we discuss in further detail in Appendix C, the derivation of the result using the replica method is sizeably more involved than that for unspiked RF models, as reported in (Gerace et al., 2020). Unlike the classic Gaussian equivalence which holds for vanilla RFs, Result 3.3 is conditioned on the variable κ, which needs to be handled specifically

in the computation, borrowing ideas from the superstatistical approach of (Adomaityte et al., 2023a), see also the discussion between equations (94) and (95) in Appendix C. Another difficulty lies in the treatment of the spike, absent in classic random features, which necessitates the tracking and non-standard asymptotic control of more summary statistics. Finally, while we have used the (non-rigorous) replica method to derive these equations, an interesting avenue of future research is to provide a rigorous probabilistic proof. A possible way to address the problem is to apply Gordon's Gaussian min-max inequalities (Gordon, 1988; Thrampoulidis et al., 2014; Stojnic, 2013a;b)), – and extend the approach of (Mignacco et al., 2020) for binary Gaussian mixtures to the infinite mixture the equivalent that the feature distribution (14) is formally equivalent to–; and generalize the approach used for pure random features (Hu & Lu, 2022; Loureiro et al., 2021a). This is, however, beyond the scope of this manuscript.

Finally, our results also allow to sharply characterize the Stieltjes transform (and thus the spectrum) of the empirical features covariance $1/n\varphi(X)^T\varphi(X)$, denoting $X \in \mathbb{R}^{n\times d}$ the matrix with rows $\{x^\mu\}_{\mu=1}^n$. This analysis is presented in detail in Appendix F, and reveals an intimate connection between the Stieltjes transform and the summary statistic V_2 in (116). This result complements the spectral analysis of (Moniri et al., 2023) for step sizes $\eta = \Theta_d(d^s)$ for $1/2 < s < 1$. In a close setting, (Wang et al., 2022) cover the $s = 1/2$ case, analyzing the conjugate kernel of deep network features for spiked data. Finally, the works of (Guionnet et al., 2023; Feldman, 2023) analyze the spectrum of related non-linear spiked Wigner matrices.

4. Discussion of Main Results

While the self-consistent equations in Result 3.3 might appear cumbersome, they offer valuable insight into the mechanism behind feature learning in two-layer neural networks trained with gradient descent. In this section, we discuss and highlight some of these insights.

4.1. Spiked Random Features vs Random Features

The asymptotic characterization 3.3 encompasses, as a special case, usual RFs (when setting the spike strength r to zero). More precisely, for zero spike strength $r = 0$ in (7), sRFs coincide with RFs, as the coefficients (15) lose their κ dependence, reducing to usual Hermite coefficients. The equivalent feature map φ^g then reduces to the Gaussian equivalent feature map employed in e.g. (Goldt et al., 2020; Hu & Lu, 2022; Schröder et al., 2023). Importantly, while the equivalent feature map for unspiked RFs is *linear* in the input x, the equivalent feature map φ^g (14) of Result 3.2 is *non-linear* in the component κ of the input x.
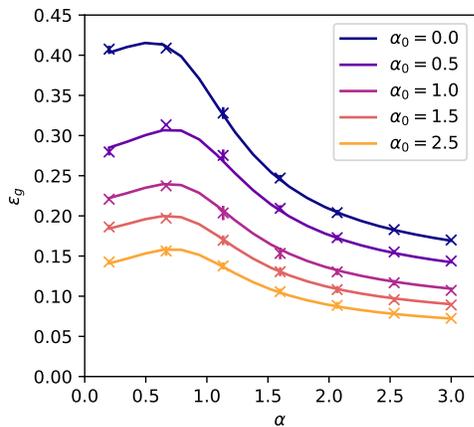
Figure 2. Crosses: numerical evaluation of the test error achieved by a two-layer network whose first layer has been trained following the protocol detailed in section 2, with learning rate $\tilde{\eta} = 1$, readout regularization $\lambda = 0.01$ and activation $\sigma = \tanh$. The target is a single index model with tanh activation. Numerical experiments were performed in $d = 3000$. All points were averaged over 5 instances. Different colours represent different initial sample complexities $\alpha_0 = n_0/d$ used for the first gradient step. Solid lines: theoretical characterization of Result 3.3 for the equivalent sRFs.

Figure 3. Crosses: Test error achieved by a two-layer network after a single gradient step with learning rate $\tilde{\eta} = 3$, readout regularization $\lambda = 0.1$ and activation $\sigma = \tanh$. The target is a single index model with sine activation. Numerical experiments were performed in $d = 2000$. All points were averaged over 5 instances. Different colours represent different initial sample complexities $\alpha_0 = n_0/d$ used for the first gradient step, with $\alpha_0 = 0$ corresponding to an untrained first layer. Solid lines: theoretical characterization of Result 3.3 for the equivalent sRFs. The dashed black line represents the lowest achievable MSE for kernel/linear methods, namely $h_2^\star - (h_1^\star)^2$ (Ba et al., 2022).

sRF models, which provide good models for two-layer networks after a single large gradient step, offer an ideal playground to test the intertwined influence of the spike/target correlation and the test performance of the model. Fig. 5 presents the learning curves of (s)RFs for varying spike strengths r. As is intuitive, larger spikes allow the sRF to more easily express non-linear function in the direction of the spike, and thus lead to relatively smaller test errors. Furthermore, note that even rather small spike strengths $r = 0.2$ already yield test errors which are sizably lower than vanilla (unspiked) RFs ($r = 0$), hinting at the qualitative difference between sRF and RF models discussed in section 2 and Fig. 1. The plot shows a compelling agreement between the theoretical predictions (continuous line) and numerical simulations. Fig. 6 represents the theoretical closed-form expression for the test error (25) for different values of the spike/target alignment γ, for a single-index target $\text{sign}(\boldsymbol{\theta}^\top \boldsymbol{x})$, with good agreement with numerical experiments. Higher alignments γ lead to overall lower test errors escaping the linear curse of Gaussian models.

4.2. Beating kernels in a single step

A key parameter in our formulas is $\gamma = \boldsymbol{v}^\top \boldsymbol{\theta}/d \in [0, 1]$, the correlation between the effective gradient spike and the target weights. From its asymptotic expression eq. 9 in Result 3.1, this is an increasing function of $\alpha_0 = n_0/d$, the number of samples in the first step. As shown in Fig. 2 for $\sigma = \sigma_\star = \tanh$, larger sample complexities in the repre-

sentation learning step α_0 allow for better feature learning when implementing a gradient step on the first layer, enabling a lower test error after the readout layer is retrained. As expected, the lowest error is achieved when $\alpha_0 \to \infty$, in which case the spike $\boldsymbol{v}$ is perfectly aligned with the target weights $\boldsymbol{\theta}$ and $\gamma = 1$, as can be seen from (9).

Fig. 3 presents similar curves for another target activation $\sigma^\star = \sin$, including the test error achieved by the network at initialization ($\alpha_0 = 0$), which corresponds to a usual RF model. The latter is well above the lowest Mean Squared Error (MSE) achievable by a linear estimator (plotted as a dashed black line), namely the projection of the teacher function on Hermite polynomials $\|\mathbb{P}_{>1} f_\star\|_2^2$. This performance corresponds to the kernel one when the number of samples scales proportionally with the input dimension (Ghorbani et al., 2020); using the notations of Result 3.1 the best linear MSE is readily written as $h_2^\star - (h_1^\star)^2$. Note that this value also lower-bounds the MSE achieved by the NTK, see Proposition 1 of (Ba et al., 2022). In sharp contrast, networks with trained first layers ($\alpha_0 > 0$) can learn non-linear functions of the inputs and outperform this baseline.

4.3. What can be learned with a single step?

While training a two-layer network with large gradient steps allows it to escape the linear limitations of, e.g. RF models, it generically only learns the target up to small test
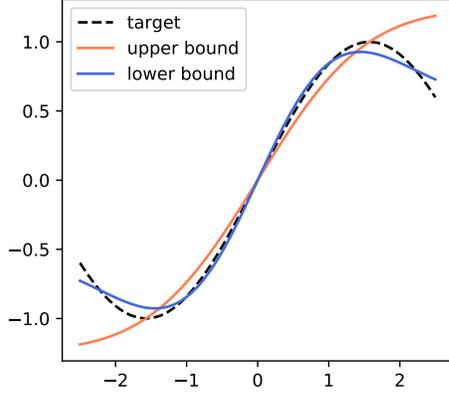
Figure 4. ($c = 1, r = 0.9$ and $\gamma = 1$) Illustration of the functions realizing the upper bound (26) (orange) and lower bound (27) (blue), for $\sigma = \tanh$, for a target $\sigma_\star = \sin$ (dashed black).

Figure 5. Test error for a sRF with $\sigma = \sin$ activation, learning from a single-index model $\text{sign}(\boldsymbol{\theta}^\top \boldsymbol{x}/\sqrt{d})$, with regularization $\lambda = 0.1$. Solid lines: theoretical characterization of Result 3.3. Crosses: numerical simulations in dimensions $d = p = 2000$. Each point is averaged over 10 instances of the problem. Different colours correspond to different spike strengths r (7), with $r = 0$ corresponding to the vanilla RF model.

error, yet not perfectly. Further insight can be extracted from the theoretical characterization of the corresponding sRF. In fact, a closer examination of the sharp asymptotic expression (25) for ϵ_g in Result 3.3 reveals that even for large initial batches $\alpha_0 \to \infty$ (and thus perfect spike/target alignments $\gamma = 1$), at fixed learning rate strength $\tilde{\eta}$ and sample complexity $\alpha > 0$, the test error optimized over the regularization λ is upper bounded as

$$\inf_{\lambda \geq 0} \epsilon_g \leq \inf_{b_1} \mathbb{E}_\kappa \left[\sigma_\star(\kappa) - b_1 \mu_0(\kappa) \right]^2, \tag{26}$$

and lower bounded by:

$$\inf_{\lambda \geq 0} \epsilon_g \geq \inf_{b_1, b_2} \mathbb{E}_\kappa \left[\sigma_\star(\kappa) - b_1 \mu_0(\kappa) - b_2 \mu_1(\kappa)\kappa \right]^2. \tag{27}$$

The upper bound (26) is the equivalent of Lemma 6 of (Ba et al., 2022) in our case of uniform readout initialization $\boldsymbol{a}^{(0)} = \mathbf{1}_p/\sqrt{p}$, and is achieved for $\lambda \to \infty$, as we show in App. E. The lower bound (27), however, shows that the test error cannot be lower than the Gaussian-weighted L^2 distance between the target function $\sigma_\star$ and $\text{span}(\mu_0, \tilde{\mu}_1)$, where $\tilde{\mu}_1(\kappa) \equiv \kappa\mu_1(\kappa)$, and the best approximation would be reached for the projection of the target thereon. Fig. 4 illustrates what functions realize the upper (26) and lower bounds (27), and how they compare with the target $\sigma_\star$. Finally, note that in the vanilla RF limit $r = 0$, the functions $\mu_0(\kappa), \mu_1(\kappa)$ reduce to constants independent of κ, constraining the class of functions that can be learned to that of linear functions.

Finally, let us mention that while this discussion was made at fixed learning rate $\tilde{\eta}$ for clarity, the latter is in practice a tunable hyper-parameter, and the functions μ_0, μ_1 depend thereupon via the spike strength r (15). One can thus refine – and lower– the bounds (26) and (27) over the functions μ_0, μ_1, which can take values in the realizable set $\mathcal{M} =$

$\{(\mu_0(r), \mu_1(r)\}_{r \geq 0}$ (emphasizing the dependence on r via equation (75)) as $\tilde{\eta}$ is varied:

$$\inf_{\lambda \geq 0, \tilde{\eta} \geq 0} \epsilon_g \leq \inf_{b_1 \in \mathbb{R}, \nu_0 \in \mathcal{M}} \mathbb{E}_\kappa \left[\sigma_\star(\kappa) - b_1 \nu_0(\kappa) \right]^2, \tag{28}$$

$$\inf_{\lambda \geq 0, \tilde{\eta} \geq 0} \epsilon_g \geq \inf_{\substack{b_1, b_2 \in \mathbb{R}, \\ (\nu_0, \nu_1) \in \mathcal{M}}} \mathbb{E}_\kappa \left[\sigma_\star(\kappa) - b_1 \nu_0(\kappa) - b_2 \nu_1(\kappa)\kappa \right]^2.$$

In other words, by tuning the learning rate $\tilde{\eta}$ – and thus the spike strength r– one gains the freedom to choose the "best" subspace $\text{span}(\mu_0(r), \mu_1(r))$, i.e. the one allowing to approximate the target $\sigma_\star$ best. Finally, as discussed in (Ba et al., 2022), observe that for $\sigma = \sigma_\star = \text{erf}$, the upper bound in (28) is zero provided one tunes the learning rate to $\tilde{\eta} = \sqrt{3\beta}/h_1$, and perfect learning is therefore achievable. Let us emphasize that this discussion formally holds for the equivalent sRF model; we however expect it to hold, at least qualitatively, also for the original two-layer networks.

4.4. More variability means better feature learning

The discussion of subsection 4.3 thus affords an insightful perspective on the learning of sRF models in terms of approximating the target activation $\sigma_\star$ in a two-dimensional functional space. Interestingly, introducing variability in the readout layer at initialization leads to an even richer functional basis, and hence greater expressivity of the network. As we further discuss in Appendix D, when $\boldsymbol{a}^{(0)}$ is no longer proportional to $\mathbf{1}_p$, but is rather initialized from a distribution over a finite vocabulary V of size $|V| > 1$–e.g. $V = \{-1, 0, +1\}$– the equivalent feature

map of the associated sRF takes the form:

$$\boldsymbol{\varphi}^g(\boldsymbol{x}) = \begin{pmatrix} \mu_0(u_1\kappa) \\ \mu_0(u_2\kappa) \\ \vdots \\ \mu_0(u_p\kappa) \end{pmatrix} + \begin{pmatrix} \mu_1(u_1\kappa) \\ \mu_1(u_2\kappa) \\ \vdots \\ \mu_1(u_p\kappa) \end{pmatrix} \odot W\boldsymbol{x} + \begin{pmatrix} \mu_2(u_1\kappa) \\ \mu_2(u_2\kappa) \\ \vdots \\ \mu_2(u_p\kappa) \end{pmatrix} \odot \boldsymbol{\xi}$$

$$(29)$$

where $\odot$ denotes element-wise multiplication, $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbb{I}_p)$, and $\boldsymbol{u} \not\propto \mathbf{1}_p$ has entries which can now take a finite number of different values. The coefficients of the equivalent map (29) are thus neuron-dependent and thus afford a richer functional basis $\{\mu_0(\omega\cdot), \tilde{\mu}_1(\omega\cdot)\}_{\omega \in V}$, thereby allowing the network to express a larger class of functions. As a matter of fact, the functional space spanned by these functions is generically of dimension $2|V|$ for non-uniform readout initializations $\boldsymbol{a}^{(0)}$, compared to just 2 in the uniform readout case. A sharp asymptotic characterization of the test error for the case of non-uniform readout initialization with finite vocabulary is further provided in Appendix D.

We briefly discuss the limiting case of interest $\lambda, \alpha_0, \tilde{\eta} \to \infty$, for which the equivalent feature map $\boldsymbol{\varphi}^g(\boldsymbol{x})$ (29) reduces to its first term $(\mu_0(u_i\kappa))_{i=1}^p$. Further, observe that $\mu_0(u_i\kappa)$ can be viewed as a one-dimensional neuron acting on the one-dimensional input variable κ with a random weight u_i. Standard results on approximation errors for random feature mappings of finite-dimensional inputs (Rahimi & Recht, 2007) imply that a large class of functions can be approximated from the network features $\boldsymbol{\varphi}^g(\boldsymbol{x})$, provided the vocabulary size $|V|$ is large enough. Similar random features approximations have been leveraged in (Ba et al., 2022; 2024; Damian et al., 2022). The equivalent feature map (29) provides an intuitive picture on how such random feature mappings of low-dimensional inputs can naturally emerge in the setting of the learning of a two-layer network.

Finally, note that the bounds (26) and (27) can be readily generalised to non-uniform readout initializations, provided one replaces in the discussion of subsection 4.3 the two-dimensional functional basis $\mathrm{span}(\mu_0, \tilde{\mu}_1)$ in the case of uniform initialization by the richer functional space $\mathrm{span}(\{\mu_0(\omega\cdot), \tilde{\mu}_1(\omega\cdot)\}_{\omega \in V})$ for non-uniform initialization. For instance, the lower bound (27) thus involves in the non-uniform case the distance between the target $\sigma_\star$ and $\mathrm{span}(\{\mu_0(\omega\cdot), \tilde{\mu}_1(\omega\cdot)\}_{\omega \in V})$.

Conclusion

We provided a tight asymptotic description of the learning of spiked random features models, which we show provides a quantitative approximation for a two-layer neural network after training its first layer with a large, single, gradient step, in the limit where the number of samples, the hid-
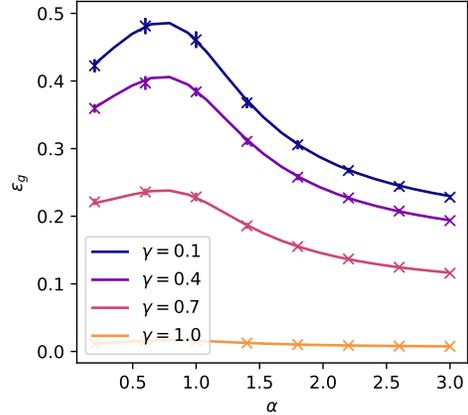
Figure 6. Test error for a sRF with $\sigma = \tanh$ activation, learning from a single-index model $\tanh(\boldsymbol{\theta}^\top \boldsymbol{x}/\sqrt{d})$, with regularization $\lambda = 0.1$. Solid lines: theoretical characterization of Result 3.3. Crosses : numerical simulations in dimensions $d = p = 2000$. Each point is averaged over 10 instances of the problem. Different colours correspond to different overlaps $\gamma \equiv \boldsymbol{\theta}^\top \boldsymbol{v}/d$ between the target weights $\boldsymbol{\theta}$ and the spike $\boldsymbol{v}$.

den layer width, and the input dimension are proportionally large. Our results sharply characterize the corresponding feature map, and how it achieves a test error which non-perturbatively improves over the kernel regime. Crucially, the sRF can efficiently approximate non-linear functions –beyond polynomials– in the direction of the gradient – sizeably improving upon RFs (modeling the network at initialization), which can only express linear functions. We further discuss bounds for the test error and which functions are learnable after a single gradient step. Finally, extensive numerical support is provided to illustrate our findings.

We believe the present work opens exciting research avenues, paving the way towards a tight theoretical understanding of feature learning in gradient-trained networks. Prominent among these research directions is the extension of our results to readout initialization with generic (not necessarily finite) support. to a finite number of gradient steps, and ultimately fully trained networks.

Acknowledgements

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abbe, E., Boix-Adsera, E., Brennan, M. S., Bresler, G., and Nagaraj, D. The staircase property: How hierarchical structure can guide deep learning. *Advances in Neural Information Processing Systems*, 34:26989–27002, 2021.

Abbe, E., Adsera, E. B., and Misiakiewicz, T. 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*, pp. 4782–4887. PMLR, 2022.

Abbe, E., Adsera, E. B., and Misiakiewicz, T. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 2552–2623. PMLR, 2023.

Adlam, B., Lee, J., Padhy, S., Nado, Z., and Snoek, J. Kernel regression with infinite-width neural networks on millions of examples. *arXiv preprint arXiv:2303.05420*, 2023.

Adomaityte, U., Defilippis, L., Loureiro, B., and Sicuro, G. High-dimensional robust regression under heavy-tailed data: Asymptotics and universality. *arXiv preprint arXiv:2309.16476*, 2023a.

Adomaityte, U., Sicuro, G., and Vivo, P. Classification of heavy-tailed features in high dimensions: a superstatistical approach. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023b.

Allen-Zhu, Z. and Li, Y. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 977–988. IEEE, 2022.

Arnaboldi, L., Krzakala, F., Loureiro, B., and Stephan, L. Escaping mediocrity: how two-layer networks learn hard single-index models with sgd. *arXiv preprint arXiv:2305.18502*, 2023a.

Arnaboldi, L., Stephan, L., Krzakala, F., and Loureiro, B. From high-dimensional & mean-field dynamics to dimensionless odes: A unifying approach to sgd in two-layers networks. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 1199–1227, 2023b.

Aubin, B., Krzakala, F., Lu, Y., and Zdeborová, L. Generalization error in high-dimensional perceptrons: Approaching bayes error with convex optimization. *Advances in Neural Information Processing Systems*, 33:12199–12210, 2020.

Ba, J., Erdogdu, M. A., Suzuki, T., Wang, Z., Wu, D., and Yang, G. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. In *Advances in Neural Information Processing Systems*, volume 35, pp. 37932–37946, 2022.

Ba, J., Erdogdu, M. A., Suzuki, T., Wang, Z., and Wu, D. Learning in the presence of low-dimensional structure: a spiked random matrix perspective. *Advances in Neural Information Processing Systems*, 36, 2024.

Ben Arous, G., Gheissari, R., and Jagannath, A. High-dimensional limit theorems for sgd: Effective dynamics and critical scaling. *Advances in Neural Information Processing Systems*, 35:25349–25362, 2022.

Berthier, R., Montanari, A., and Zhou, K. Learning time-scales in two-layers neural networks. *arXiv preprint arXiv:2303.00055*, 2023.

Bordelon, B. and Pehlevan, C. Dynamics of finite width kernel and prediction fluctuations in mean field neural networks. *Advances in Neural Information Processing Systems*, 36, 2024.

Bosch, D., Panahi, A., and Hassibi, B. Precise asymptotic analysis of deep random feature models. In *Proceedings of Thirty Sixth Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 4132–4179, 12–15 Jul 2023.

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

Chizat, L., Oyallon, E., and Bach, F. On Lazy Training in Differentiable Programming. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Cui, H. and Zdeborová, L. High-dimensional asymptotics of denoising autoencoders. *Advances in Neural Information Processing Systems*, 36, 2024.

Cui, H., Krzakala, F., and Zdeborova, L. Bayes-optimal learning of deep random networks of extensive-width. In *International Conference on Machine Learning*, pp. 6468–6521. PMLR, 2023.

Damian, A., Lee, J., and Soltanolkotabi, M. Neural networks can learn representations with gradient descent. In *Proceedings of Thirty Fifth Conference on Learning Theory*, pp. 5413–5452, 2022.

Dandi, Y., Krzakala, F., Loureiro, B., Pesce, L., and Stephan, L. How two-layer neural networks learn, one (giant) step at a time. In *NeurIPS 2023 Workshop on Mathematics of Modern Machine Learning*, 2023.

Dandi, Y., Stephan, L., Krzakala, F., Loureiro, B., and Zdeborová, L. Universality laws for gaussian mixtures in generalized linear models. *Advances in Neural Information Processing Systems*, 36, 2024.

Dhifallah, O. and Lu, Y. M. A precise performance analysis of learning with random features. *arXiv preprint arXiv:2008.11904*, 2020.

Dyer, E. and Gur-Ari, G. Asymptotics of wide networks from feynman diagrams. In *International Conference on Learning Representations*, 2020.

Feldman, M. J. Spectral properties of elementwise-transformed spiked matrices. *arXiv preprint arXiv:2311.02040*, 2023.

Gerace, F., Loureiro, B., Krzakala, F., Mezard, M., and Zdeborova, L. Generalisation error in learning with random features and the hidden manifold model. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3452–3462, 2020.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. Limitations of lazy training of two-layers neural network. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Ghorbani, B., Mei, S., Misiakiewicz, T., and Montanari, A. When do neural networks outperform kernel methods? In *Advances in Neural Information Processing Systems*, volume 33, pp. 14820–14830, 2020.

Goldt, S., Mézard, M., Krzakala, F., and Zdeborová, L. Modeling the influence of data structure on learning in neural networks: The hidden manifold model. *Physical Review X*, 10(4):041044, 2020.

Goldt, S., Loureiro, B., Reeves, G., Krzakala, F., Mezard, M., and Zdeborova, L. The gaussian equivalence of generative models for learning with shallow neural networks. In *Proceedings of the 2nd Mathematical and Scientific Machine Learning Conference*, pp. 426–471, 2022.

Gordon, Y. On milman's inequality and random subspaces which escape through a mesh in rn. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 1986–87*, pp. 84–106, 1988.

Guionnet, A., Ko, J., Krzakala, F., Mergny, P., and Zdeborová, L. Spectral phase transitions in non-linear wigner spiked models. *arXiv preprint arXiv:2310.14055*, 2023.

Hajjar, K. and Chizat, L. On the symmetries in the dynamics of wide two-layer neural networks. *arXiv preprint arXiv:2211.08771*, 2022.

Hanin, B. and Nica, M. Finite depth and width corrections to the neural tangent kernel. In *International Conference on Learning Representations*, 2020.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

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

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Lee, J., Sohl-dickstein, J., Pennington, J., Novak, R., Schoenholz, S., and Bahri, Y. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018.

Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mezard, M., and Zdeborová, L. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151, 2021a.

Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mezard, M., and Zdeborová, L. Learning curves of generic features maps for realistic datasets with a teacher-student model. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18137–18151, 2021b.

Loureiro, B., Sicuro, G., Gerbelot, C., Pacco, A., Krzakala, F., and Zdeborová, L. Learning gaussian mixtures with generalized linear models: Precise asymptotics in high-dimensions. *Advances in Neural Information Processing Systems*, 34:10144–10157, 2021c.

Mei, S. and Montanari, A. 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.

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

Mei, S., Misiakiewicz, T., and Montanari, A. Generalization error of random feature and kernel methods: Hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022.

Mignacco, F., Krzakala, F., Lu, Y., Urbani, P., and Zdeborova, L. The role of regularization in classification of high-dimensional noisy gaussian mixture. In *International conference on machine learning*, pp. 6874–6883, 2020.

Moniri, B., Lee, D., Hassani, H., and Dobriban, E. A theory of non-linear feature learning with one gradient step in two-layer neural networks. *arXiv preprint arXiv:2310.07891*, 2023.

Montanari, A. and Saeed, B. N. Universality of empirical risk minimization. In *Conference on Learning Theory*, pp. 4310–4312. PMLR, 2022.

Mousavi-Hosseini, A., Wu, D., Suzuki, T., and Erdogdu, M. A. Gradient-based feature learning under structured data. *Advances in Neural Information Processing Systems*, 36:71449–71485, 2023.

Naveh, G. and Ringel, Z. A self consistent theory of gaussian processes captures feature learning effects in finite cnns. In *Advances in Neural Information Processing Systems*, volume 34, pp. 21352–21364, 2021.

Neal, R. *Bayesian Learning for Neural Networks*. Lecture Notes in Statistics. 1996. ISBN 9780387947242.

Parisi, G. Toward a mean field theory for spin glasses. *Physics Letters A*, 73(3):203–205, 1979.

Parisi, G. Order parameter for spin-glasses. *Physical Review Letters*, 50(24):1946, 1983.

Pesce, L., Krzakala, F., Loureiro, B., and Stephan, L. Are Gaussian data all you need? The extents and limits of universality in high-dimensional generalized linear estimation. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 27680–27708, 2023.

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

Refinetti, M., Goldt, S., Krzakala, F., and Zdeborova, L. Classifying high-dimensional gaussian mixtures: Where kernel methods fail and neural networks succeed. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8936–8947, 2021a.

Refinetti, M., Goldt, S., Krzakala, F., and Zdeborová, L. Classifying high-dimensional gaussian mixtures: Where kernel methods fail and neural networks succeed. In *International Conference on Machine Learning*, pp. 8936–8947. PMLR, 2021b.

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

Schröder, D., Cui, H., Dmitriev, D., and Loureiro, B. Deterministic equivalent and error universality of deep random features learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 30285–30320, 2023.

Schröder, D., Dmitriev, D., Cui, H., and Loureiro, B. Asymptotics of learning with deep structured (random) features. *arXiv preprint arXiv:2402.13999*, 2024.

Seroussi, I., Naveh, G., and Ringel, Z. Separation of scales and a thermodynamic description of feature learning in some cnns. *Nature Communications*, 14(1):908, 2023.

Shi, Z., Wei, J., and Liang, Y. A theoretical analysis on feature learning in neural networks: Emergence from inputs and advantage over fixed features. *arXiv preprint arXiv:2206.01717*, 2022.

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

Stojnic, M. Meshes that trap random subspaces. *arXiv preprint arXiv:1304.0003*, 2013a.

Stojnic, M. Upper-bounding l1-optimization weak thresholds. *arXiv preprint arXiv:1303.7289*, 2013b.

Thrampoulidis, C., Oymak, S., and Hassibi, B. The gaussian min-max theorem in the presence of convexity. *arXiv preprint arXiv:1408.4837*, 2014.

Wang, Z., Engel, A., Sarwate, A., Dumitriu, I., and Chiang, T. Spectral evolution and invariance in linear-width neural networks. *arXiv preprint arXiv:2211.06506*, 2022.

Xiao, L., Hu, H., Misiakiewicz, T., Lu, Y., and Pennington, J. Precise learning curves and higher-order scalings for dot-product kernel regression. *Advances in Neural Information Processing Systems*, 35:4558–4570, 2022.

Yaida, S. Non-Gaussian processes and neural networks at finite widths. In *Proceedings of The First Mathematical and Scientific Machine Learning Conference*, volume 107 of *Proceedings of Machine Learning Research*, pp. 165–192, 20–24 Jul 2020.

Zavatone-Veth, J. and Pehlevan, C. Learning curves for deep structured gaussian feature models. *Advances in Neural Information Processing Systems*, 36:42866–42897, 2023.

A. Derivation of Result 3.1

This Appendix details the derivation of Result 3.1, which provides tight asymptotic formulae (9) for the parameters of the sRF associated to a two-layer neural network trained with a large gradient step.

A.1. First layer weights after one gradient step –

After one large gradient step, the weights $W^{(1)}$ can be decomposed as a bulk plus a spike in the following way:

$$W^{(1)} = W^{(0)} + \Delta + \frac{\check{\boldsymbol{u}}\check{\boldsymbol{v}}^\top}{\sqrt{p}}, \tag{30}$$

where

$$\check{\boldsymbol{u}} = \eta h_1 \boldsymbol{a}^{(0)}, \tag{31}$$

$$\check{\boldsymbol{v}} = \frac{1}{n_0} \sum_{\mu=1}^{n_0} f_\star(\boldsymbol{x}^\mu)\boldsymbol{x}^\mu, \tag{32}$$

and

$$\Delta = W^{(0)} + \frac{\eta}{n_0\sqrt{p}}\mathrm{diag}(\boldsymbol{a}^{(0)}) \left(\sum_{\mu=1}^{n_0} \check{\sigma}(W^{(0)}\boldsymbol{x}^\mu)f_\star(\boldsymbol{x}^\mu)\boldsymbol{x}^\mu \right). \tag{33}$$

We noted h_1 the first Hermite coefficient of the student activation σ and $\check{\sigma}(\cdot) = \sigma'(\cdot) - h_1$. As discussed in (Ba et al., 2022; Dandi et al., 2024), the bulk Δ behaves as an effective noise which rescales the initial weights $W^{(0)}$; we refer to Sec. **B.6** of (Dandi et al., 2023) for additional rigorous characterization of the bulk Δ. To connect to the convention used below (7) of normalized $\boldsymbol{u} = \mathbf{1}_p$, $\|\boldsymbol{v}\| = \sqrt{d}$, and a single bulk W, one thus needs to assess the asymptotic limits of $r = \|\check{\boldsymbol{u}}\|\|\check{\boldsymbol{v}}\|/\sqrt{pd}$, $c = 1/\sqrt{d}\|W^{(0)} + \Delta\|$ and $\gamma = \check{\boldsymbol{v}}^\top\boldsymbol{\theta}/\|\check{\boldsymbol{v}}\|$. This program is carried out in the following subsection.

A.2. Spike scale

We first focus on the scale of the spike r. We remind here the convention to assume normalized spike vectors, i.e. $\|\boldsymbol{u}\| = \|\boldsymbol{v}\| = \sqrt{d}$. Comparing to (31), it follows that

$$r = \frac{1}{\sqrt{pd}} \left\| \eta h_1 a^{(0)} \right\| \times \left\| \frac{1}{n_0} \sum_{\mu=1}^{n_0} f_\star(\boldsymbol{x}^\mu)\boldsymbol{x}^\mu \right\|. \tag{34}$$

Remembering that $\boldsymbol{a}^{(0)} = \mathbf{1}_p/\sqrt{p}$, it follows that

$$\left\| \eta h_1 \boldsymbol{a}^{(0)} \right\| = d\tilde{\eta}h_1. \tag{35}$$

We now turn to the second term. As a mean of independent vectors, it concentrates in probability. We thus compute the expectation of the squared norm:

$$\mathbb{E}_{\mathcal{D}_0} \left\| \frac{1}{n_0} \sum_{\mu=1}^{n_0} f_\star(\boldsymbol{x}^\mu)\boldsymbol{x}^\mu \right\|^2 = \frac{1}{n_0} \sum_{i=1}^{d} \mathbb{E}_{\boldsymbol{x}}[f_\star(\boldsymbol{x})^2 x_i^2] + \frac{(n_0-1)}{n_0} \sum_{i=1}^{d} \mathbb{E}_x[f_\star(\boldsymbol{x})x_i]^2 \tag{36}$$

Note that $\boldsymbol{\theta}^\top\boldsymbol{x}/\sqrt{d}$ and x_i have asymptotically vanishing correlation. To leading order, therefore

$$\mathbb{E}_{\boldsymbol{x}}[f_\star(\boldsymbol{x})^2 x_i^2] = \mathbb{E}_z[\sigma_\star(z)^2] + O_d(1/\sqrt{d}) = h_2^\star \tag{37}$$

$$\mathbb{E}_{\boldsymbol{x}}[f_\star(\boldsymbol{x})x_i] = \mathbb{E}_z[\sigma_\star'(z)]\frac{\theta_i}{\sqrt{d}} = h_1^\star \frac{\theta_i}{\sqrt{d}} \tag{38}$$

Therefore

$$\mathbb{E}_{\mathcal{D}_0} \left\| \frac{1}{n_0} \sum_{\mu=1}^{n_0} f_\star(\boldsymbol{x}^\mu)\boldsymbol{x}^\mu \right\|^2 = \frac{dh_2^\star}{n_0} + \frac{(n_0-1)h_1^{\star 2}}{n_0} = \frac{h_2^\star}{\alpha_0} + h_1^{\star 2}. \tag{39}$$

Thus

$$r = \frac{\tilde{\eta} h_1}{\sqrt{\beta}} \left(\frac{h_2^\star}{\alpha_0} + h_1^{\star 2} \right)^{1/2}. \tag{40}$$

A.3. Spike/target alignment

We now turn to the spike/target alignment, defined as the overlap between the (normalized) target weights θ and spike v.

$$\gamma = \frac{\boldsymbol{\theta}^\top \boldsymbol{v} \sqrt{d}}{d\|\boldsymbol{v}\|} = \frac{\frac{1}{n_0} \sum_{\mu=1}^{n_0} \sigma_\star(z^\mu) z^\mu}{\left\| \frac{1}{n_0} \sum_{\mu=1}^{n_0} f_\star(x^\mu) x^\mu \right\|}. \tag{41}$$

We noted $z^\mu \equiv \theta^\top \boldsymbol{x}^\mu / \sqrt{d}$. The numerator converges by the law of large numbers to $h_1^\star$, while the numerator was computed in the previous subsection. All in all, we reach

$$\gamma = \frac{h_1^\star}{\left(\frac{h_2^\star}{\alpha_0} + h_1^{\star 2} \right)^{1/2}} \tag{42}$$

A.4. Bulk norm

By symmetry, all rows of the bulk Δ (33) asymptotically share the same norm c_Δ. The norm of a row w of $W = W^{(0)} + \Delta$ thus has asymptotic norm

$$\|w_i\| = \sqrt{1 + c_\Delta^2 + 2c_{W\Delta}} \tag{43}$$

where

$$c_\Delta = \left\| \frac{\eta}{n_0 p} \left(\sum_{\mu=1}^{n_0} \check{\sigma}(w^{(0)\top} x^\mu) f_\star(x^\mu) x^\mu \right) \right\| \tag{44}$$

$$c_{W\Delta} = \frac{\eta}{n_0 p} \left(\sum_{\mu=1}^{n_0} \check{\sigma}(w^{(0)\top} x^\mu) f_\star(x^\mu) w^{(0)\top} x^\mu \right), \tag{45}$$

where with a slight abuse of notation we dropped the row index i, since all rows are equivalent.

$$\mathbb{E}_{\mathcal{D}_0}[c_\Delta^2] = \frac{\eta^2}{n_0 p^2} \sum_{i=1}^d \mathbb{E}_x[\check{\sigma}(w^{(0)\top} x)^2 \sigma_\star(\theta^\top x/\sqrt{d})^2 x_i^2] + \frac{(n_0-1)\eta^2}{n_0 p^2} \sum_{i=1}^d \mathbb{E}_x[\check{\sigma}(w^{(0)\top} x)\sigma_\star(\theta^\top x/\sqrt{d})x_i]^2 \tag{46}$$

Because $w^{(0)\top} x, \theta^\top x/\sqrt{d}, x_i$ all have asymptotically vanishing $\Theta_d(1/\sqrt{d})$ correlations, the first summand reads to leading order

$$\mathbb{E}_x[\check{\sigma}(w^{(0)\top} x)^2 \sigma_\star(\theta^\top x/\sqrt{d})^2 x_i^2] = \check{h}_1^2 h_2^\star \tag{47}$$

The second summand reads, using Stein's lemma

$$\mathbb{E}_x[\check{\sigma}(w^{(0)\top} x)\sigma_\star(\theta^\top x/\sqrt{d})x_i] = w_i^{(0)}\mathbb{E}_x[\check{\sigma}'(w^{(0)\top} x)\sigma_\star(\theta^\top x/\sqrt{d})] + \frac{\theta_i}{\sqrt{d}}\mathbb{E}_x[\check{\sigma}(w^{(0)\top} x)\sigma_\star'(\theta^\top x/\sqrt{d})]$$

$$= \Theta_d(1/d) \tag{48}$$

14

We used

$$\mathbb{E}_z[\check{\sigma}(z)] = \mathbb{E}_z[\check{\sigma}'(z)] = 0 \tag{49}$$

respectively by definition of $\check{\sigma}$ and because σ is odd. Thus

$$\mathbb{E}_{\mathcal{D}_0}[c_\Delta^2] = \frac{\eta^2 \check{h}_1^2 h_2 d}{n_0 p^2} + \Theta_d(1/d) = \frac{\tilde{\eta}^2 \check{h}_1^2 h_2^\star}{\alpha_0 \beta^2} + o_d(1). \tag{50}$$

The cross-term $c_{\Delta W}$ can similarly be computed as

$$c_{W\Delta} = \frac{\eta}{p} \mathbb{E}_x[\check{\sigma}(w^{(0)\top}x)\sigma_\star(\theta^\top x/\sqrt{d})w^{(0)\top}x] \tag{51}$$

Again because all variables in the expectation are weakly correlated, and because the average of $\check{\sigma}$ is zero, we have that $c_{W\Delta}$ is vanishing to leading order.

$$c_{W\Delta} = o_d(1) \tag{52}$$

Finally, we thus have

$$c = 1 + \frac{\tilde{\eta}^2 \check{h}_1^2 h_2^\star}{\alpha_0 \beta^2} \tag{53}$$

A.5. Summary of the mapping

In summary, we have derived the asymptotic formulae for the equivalent sRF:

$$c = 1 + \frac{\tilde{\eta}^2 \check{h}_1^2 h_2^\star}{\alpha_0 \beta^2} \tag{54}$$

$$r = \frac{\tilde{\eta} h_1}{\sqrt{\beta}} \left(\frac{h_2^\star}{\alpha_0} + h_1^{\star 2} \right)^{1/2}, \tag{55}$$

$$\gamma = \frac{h_1^\star}{\left(\frac{h_2^\star}{\alpha_0} + h_1^{\star 2} \right)^{1/2}}, \tag{56}$$

which recovers equation (9) of Result 3.1.

B. Derivation of Result 3.2

In this section we detail the heuristic derivation of Result 3.2, which shows the asymptotic equivalence of the sRF feature map (13) with the conditional feature map (14). This mapping is at the core of the exact asymptotic description detailed in the next appendix.

We first remark that, since the data is Gaussian, $x \sim \mathcal{N}(0, \mathbb{I}_d)$, we can decompose it as

$$x = \kappa + x^\perp, \tag{57}$$

where $\kappa \equiv v^\top x/\sqrt{d} \sim \mathcal{N}(0,1)$ and we noted $x^\perp \equiv \Pi^\perp x$ the component orthogonal to the spike, with Gaussian statistics $x^\perp \sim \mathcal{N}(0, \Pi^\perp)$. The sRF feature map (13) thus reads

$$\varphi(x) = \sigma \left(\kappa 1_p + \kappa \frac{Wv}{\sqrt{d}} + Wx^\perp \right) \tag{58}$$

Conditioning on the projection κ along the spike, this corresponds to a RF feature map with random biases $\kappa 1_p + \kappa^{Wv}/\sqrt{d}$. One can compute the population mean and covariance of φ_x along the exact same lines as e.g. (Cui et al., 2023), leading to

$$\mathbb{E}_x[\varphi(x)_i] = \tilde{\kappa}_i^0. \tag{59}$$

$$\text{Cov}_x[\varphi(x)_i, \varphi(x)_j] = \tilde{\kappa}_i^1 \tilde{\kappa}_j^1 (W)_i^\top \Pi^\perp (W)_j + \delta_{ij}(\kappa_i^*)^2. \tag{60}$$

where

$$\tilde{\kappa}_i^0 \equiv \mathbb{E}_z[\sigma(\kappa + (W)_i^\top v/\sqrt{d} + z)] \tag{61}$$

$$\tilde{\kappa}_i^1 \equiv \mathbb{E}_z[z\sigma(\kappa + (W)_i^\top v/\sqrt{d} + z)] \tag{62}$$

$$\tilde{\kappa}_i^* = \sqrt{\mathbb{E}_z[\sigma(\kappa + (W)_i^\top v/\sqrt{d} + z)^2] - (\tilde{\kappa}_i^1)^2 - (\tilde{\kappa}_i^0)^2} \tag{63}$$

In other words, the features $\varphi(x)$ *share the same second order statistics* as the noisy features

$$\varphi^g(x)_i \equiv \tilde{\kappa}_i^0 + \tilde{\kappa}_i^1 (W)_i^\top x^\perp + \tilde{\kappa}_i^* \xi_i \tag{64}$$

where $\xi_i \sim \mathcal{N}(0,1)$ is a stochastic noise. Note that in the definition of the $\kappa_i^{0,1,*}$ coefficients (61), $(W)_i^\top v/\sqrt{d} = \Theta_d(1/\sqrt{d})$, so one can expand them as

$$\tilde{\kappa}_i^0 \equiv \underbrace{\mathbb{E}_z[\sigma(\kappa + z)]}_{\mu_0(\kappa)} + (W)_i^\top v/\sqrt{d}\mu_1(\kappa) \tag{65}$$

$$\tilde{\kappa}_i^1 \equiv \underbrace{\mathbb{E}_z[z\sigma(\kappa + z)]}_{\mu_1(\kappa)} + (W)_i^\top v/\sqrt{d}\underbrace{\mathbb{E}_z[z\sigma'(\kappa + z)]}_{\equiv\mu_3(\kappa)} \tag{66}$$

$$\tilde{\kappa}_i^* = \underbrace{\sqrt{\mathbb{E}_z[\sigma(\kappa + z)^2] - (\kappa^1)^2 - (\kappa^0)^2}}_{\mu_2(\kappa)} + O_d(1/\sqrt{d}) \tag{67}$$

So

$$\varphi^g(x)_i \equiv \mu_0(\kappa) + \mu_1(\kappa)(W)_i^\top x + \tag{68}$$

$$+ \mu_3(\kappa) \times \frac{(W)_i^\top v}{\sqrt{d}} \times (W)_i x^\perp + \mu_2(\kappa)\xi_i + \xi_i\Theta_d(1/\sqrt{d}) \tag{69}$$

Finally, note that the $\mu_3(\kappa)$ term is subleading compared to the second term, and should be asymptotically unconsequential for the learning, and can thus be safely neglected. By the same token, the last term is a subleading correction to the noise term and can be neglected. Therefore, one finally reaches

$$\varphi^g(x) \equiv \mu_0(\kappa)1_p + \mu_1(\kappa)W^\top x + \mu_2(\kappa)\xi \tag{70}$$

which recovers (14) in Result 3.2.

C. Derivation of Result 3.3

In this Appendix, we detail the derivation of Result 3.3. The derivation leverages the replica method from statistical physics (Parisi, 1979; 1983) in its replica-symmetric formulation, in conjunction with Result 3.2. On a technical level, the derivation builds upon the formal equivalence between the distribution of the equivalent features $\varphi^g(x)$ (14) with a mixture of Gaussians with infinitely many $\kappa-$indexed clusters, similarly to the superstatistical approach of (Adomaityte et al., 2023a;b) in another setting.

First observe that for any test function $\phi(\hat{a})$ of the minimizer $\hat{a}$ of (4),

$$\phi(\hat{a}) = \lim_{\omega \to \infty} \mathbb{E}_\mathcal{D} \frac{1}{Z} \int d\boldsymbol{a}\phi(\boldsymbol{a})e^{-\omega\mathcal{R}_\lambda[\boldsymbol{a}]}, \tag{71}$$

where we remind that $R[\boldsymbol{a}]$ denotes the empirical risk (4), and we denoted

$$Z \equiv \int d\boldsymbol{a}e^{-\omega\mathcal{R}_\lambda[\boldsymbol{a}]} \tag{72}$$

the normalization factor, also known as the *partition function* in statistical physics. It is therefore important to the generating function associated with the measure (71), namely $\mathbb{E}\ln Z$. Such computations can be addressed using the *replica* method from statistical physics (Parisi, 1979; 1983), leveraging the identity

$$\ln Z = \lim_{s \to 0} \frac{Z^s - 1}{s}. \tag{73}$$

The backbone of the derivation thus lies in the computation of $\mathbb{E}Z^s$.

C.1. Reminder of setting and notations

Because of Result 3.2, we take as a starting point that one can assume the sRF features $z = \varphi(x)$ (13) are distributed as

$$
\begin{aligned}
z &= \mu_0(\kappa)\mathbf{1}_p + \mu_1(\kappa)Wx + \mu_2(\kappa)\xi \\
&= \mu_0(\kappa)\mathbf{1}_p + {}^1\!/\!\sqrt{d}\mu_1(\kappa)\kappa W v + \mu_1(\kappa)W\Pi^\perp x + \mu_2(\kappa)\xi
\end{aligned} \tag{74}
$$

where $\kappa \equiv \frac{v^\top z}{\sqrt{d}} \sim \mathcal{N}(0,1), u = \mathbf{1}_d, x \sim \mathcal{N}(0,\mathbb{I}_d), \xi \sim \mathcal{N}(0,\mathbb{I}_p)$. We remind that

$$
\mu_0(\kappa) = \mathbb{E}_w\, \sigma(w + r\kappa) \tag{75}
$$

$$
\mu_1(\kappa) = \frac{1}{c}\, \mathbb{E}_w\, w\sigma(w + r\kappa) \tag{76}
$$

$$
\mu_2(\kappa) = \sqrt{\mathbb{E}_w\, \sigma^2(w + r\kappa) - c(\mu_1(\kappa))^2 - (\mu_0(\kappa))^2}, \tag{77}
$$

where expectations bear over $w \sim \mathcal{N}(0,c)$. Finally, we consider a generic target function

$$
f_\star(x) = \sigma_\star(\kappa, \theta^\top \Pi^\perp x/\sqrt{d}). \tag{78}
$$

Note that single-index models considered in the main text constitute a special case of this family of functions, which we consider here for generality. We look at the empirical risk minimization over a of

$$
R[a] = \sum_{\mu=1}^n \ell(f_\star(x^\mu), a^\top z^\mu/\sqrt{p}) + g(a), \tag{79}
$$

for generic convex loss ℓ and ℓ_2 regularizer $g(\cdot) = {}^\lambda\!/\!_2\|\cdot\|^2$. We shall provide a detailed derivation in this generic case and only specialize to the particular case of square loss addressed in Result 3.3 at the end of the derivation.

C.2. Replica computation

The replicated partition function reads

$$
Z^s = \int \prod_{r=1}^s da_r e^{-\omega g(a_r)} \left[\mathbb{E}_x \underbrace{\prod_{r=1}^s e^{-\omega\ell(f_\star(x), a_r^\top z/\sqrt{p})}}_{(\star)} \right]^n \tag{80}
$$

We now expand $(\star)$.

$$
(\star) = \mathbb{E}_{\kappa, x^\perp, \xi} \prod_{r=1}^s e^{-\omega\ell(\sigma_\star(\kappa,\lambda_\star), \mu_0(\kappa)m_r + \mu_1(\kappa)\kappa\zeta_r + \mu_1(\kappa)g_r + \mu_2(\kappa)h_r)} \tag{81}
$$

We defined the summary statistics

$$
m_r = \frac{a_r^\top \mathbf{1}_p}{\sqrt{p}}, \qquad \zeta_r = \frac{a_r^\top W v}{\sqrt{dp}}, \qquad q_{rr'}^1 = \frac{a_r^\top \Omega a_{r'}}{p}, \qquad q_{rr'}^2 = \frac{a_r^\top a_{r'}}{p}, \tag{82}
$$

and local fields

$$
g_r = \frac{a_r^\top W\Pi^\perp x}{\sqrt{p}}, \qquad h_r = \frac{a_r^\top \xi}{\sqrt{p}}, \qquad \lambda_\star = \frac{\theta^\top \Pi^\perp x}{\sqrt{d}} \tag{83}
$$

where

$$
\Omega = W\Pi^\perp W^\top. \tag{84}
$$

One can then rewrite $(\star)$ more compactly as

$$(\star) = \mathbb{E}_\kappa \left[\mathbb{E}_{\lambda_\star, \{\lambda_r\}_{r=1}^s} e^{-\omega \sum\limits_{r=1}^s \ell(\sigma_\star(\kappa,\lambda_\star),\mu_0(\kappa)m_r+\mu_1(\kappa)\kappa\zeta_r+\lambda_r)} \right] \tag{85}$$

where the fields $\lambda_\star, \lambda_r$ are Gaussian with statistics

$$\langle \lambda_r \lambda_{r'} \rangle = \mu_1(\kappa)^2 q^1_{rr'} + \mu_2(\kappa)^2 q^2_{rr'}$$

$$\langle \lambda_\star^2 \rangle = \frac{\boldsymbol{\theta}^\top \Pi^\perp \boldsymbol{\theta}}{d} \equiv \rho$$

$$\langle \lambda_r \lambda_\star \rangle = \mu_1(\kappa) \frac{\boldsymbol{a}_r^\top W \Pi^\perp \boldsymbol{\theta}}{\sqrt{pd}} \equiv \mu_1(\kappa)\psi_r \tag{86}$$

Introducing Dirac's deltas in their Fourier form to enforce the definitions of the summary statistics $m_r, \zeta_r, q^{1,2}_{rr'}, \psi_r$, the replicated partition function then reads

$$Z^s = \int \prod_{r=1}^s dm_r d\hat{m}_r d\zeta_r d\hat{\zeta}_r d\psi_r d\hat{\psi}_r \prod_{r,r'} dq^1_{rr'} d\hat{q}^1_{rr'} dq^2_{rr'} d\hat{q}^2_{rr'} \underbrace{e^{-\sqrt{p}\sum_r m_r \hat{m}_r - \sqrt{dp}\sum_r(\zeta_r\hat{\zeta}_r+\psi_r\hat{\psi}_r) - p\sum_{a\leq b}(q^1_{rr'}\hat{q}^1_{rr'}+q^2_{rr'}\hat{q}^2_{rr'})}}_{e^{s\omega p \Psi_t}}$$

$$\underbrace{\int \prod_r d\boldsymbol{a}_r e^{-\omega \sum_r g(\boldsymbol{a}_r) + \sum\limits_{r=1}^s \boldsymbol{a}_r^\top(\hat{m}_r \mathbf{1}_p + \hat{\zeta}_r W\boldsymbol{v} + \hat{\psi}_r W \Pi^\perp \boldsymbol{\theta}) + \sum_{r\leq r'}(\hat{q}^1_{rr'}\boldsymbol{w}_r^\top \Omega \boldsymbol{w}_{r'} + \hat{q}^2_{rr'}\boldsymbol{w}_r^\top \boldsymbol{w}_b)}}_{e^{s\omega p \Psi_w}}$$

$$\underbrace{\left[\mathbb{E}_\kappa \left(\mathbb{E}_{\lambda_\star, \{\lambda_r\}_{r=1}^s} e^{-\omega \sum\limits_{r=1}^s \ell(\sigma_\star(\kappa,\lambda_\star),\mu_0(\kappa)m^a+\mu_1(\kappa)\kappa\zeta^a+\lambda_r)} \right) \right]^n}_{e^{\alpha/\beta s\omega p \Psi_y}} \tag{87}$$

We will refer to these three terms respectively as the trace, entropic, and energetic terms. Note that all terms in the exponents are a priori $\Theta_d(d)$, thus large. It is therefore possible to compute the integral over the summary statistics $m_r, \zeta_r, q^{1,2}_{rr'}, \psi_r$ and respective conjugate variables using Laplace's method, thereby reducing the problem to finding the extremum of $s\Psi_t + s\Psi_w + \alpha s\Psi_y$.

C.3. Replica symmetric ansatz

The extremization remains challenging on several levels, notably because it involves $s^2 + 5s$ variables, and the replica method requires to take the limit $s \to 0$. To make the problem tractable, we will seek the extremum assuming it has the *replica-symmetric* form

$$\forall 1 \leq r, r' \leq s, \; q^\iota_{rr'} = \delta_{rr'}(r^\iota - q^\iota) + q^\iota$$

$$\forall 1 \leq r \leq s, \; m_r = m$$

$$\forall 1 \leq r \leq s, \; \zeta_r = \zeta$$

$$\forall 1 \leq r \leq s, \; \psi_r = \psi \tag{88}$$

for $\iota = 1, 2$, and similarly for the hat variables:

$$\forall 1 \leq r, r' \leq s, \; \hat{q}^\iota_{ab} = \delta_{ab}(-\frac{1}{2}\hat{r}^\iota - \hat{q}^\iota) + \hat{q}^\iota$$

$$\forall 1 \leq r \leq s, \; \hat{m}_r = \hat{m}$$

$$\forall 1 \leq r \leq s, \; \hat{\zeta}_r = \hat{\zeta}$$

$$\forall 1 \leq r \leq s, \; \hat{\psi}_r = \hat{\psi} \tag{89}$$

Like in e.g. (Loureiro et al., 2021a), we also introduce the variance order parameters

$$V_\iota = r_\iota - q_\iota, \qquad\qquad \hat{V}_\iota = \hat{r}_\iota + \hat{q}_\iota. \tag{90}$$

C.4. Computing the free energy

Entropic term The computation of the entropic term follows from the derivation of (Aubin et al., 2020), with but minor changes. The entropic term thus reads,

$$p\omega\Psi_w = \mathbb{E}_{\boldsymbol{\xi}} \ln \int d\boldsymbol{a} e^{-\omega g(\boldsymbol{a}) - \frac{1}{2}\boldsymbol{a}^\top(\hat{V}_1\Omega + \hat{V}_2\mathbb{I}_p)\boldsymbol{a} + \boldsymbol{a}^\top\left(\hat{m}\mathbf{1}_p + \hat{\zeta}W\boldsymbol{v} + \hat{\psi}W\Pi^\perp\boldsymbol{\zeta} + (\hat{q}_1\Omega + \hat{q}_2\mathbb{I}_p)^{1/2}\boldsymbol{\xi}\right)} \tag{91}$$

In the $\omega \to \infty$ limit, rescaling the conjugate statistics as

$$\hat{q}_\iota \to \omega^2\hat{q}_\iota, \qquad\qquad \hat{m}, \hat{V}_\iota, \hat{\zeta}, \hat{\psi} \to \omega\hat{m}, \omega\hat{V}_\iota, \omega\hat{\zeta}, \omega\hat{\psi}, \tag{92}$$

the entropic potential can finally be rewritten in the $\omega \to \infty$ limit as

$$\Psi_w = -\mathbb{E}_{\boldsymbol{\xi}}\mathcal{M}_g(\boldsymbol{\xi}) \tag{93}$$

where we define the Moreau envelope

$$\mathcal{M}_g(\boldsymbol{\xi}) = \inf_{\boldsymbol{a}} \left\{ g(\boldsymbol{a}) + \frac{1}{2}\boldsymbol{a}^\top(\hat{V}_1\Omega + \hat{V}_2\mathbb{I}_p)\boldsymbol{a} - \boldsymbol{a}^\top\left(\hat{m}\mathbf{1}_p + \hat{\zeta}W\boldsymbol{v} + \hat{\psi}W\Pi^\perp\boldsymbol{\theta} + (\hat{q}_1\Omega + \hat{q}_2\mathbb{I}_p)^{1/2}\boldsymbol{\xi}\right) \right\} \tag{94}$$

For the case $g = \lambda/2\|\cdot\|^2$, this simplifies to

$$\Psi_w = \frac{1}{2p} \mathrm{Tr}\left[\left(\hat{m}^2\mathbf{1}_p\mathbf{1}_p^\top + \hat{\zeta}^2 W\boldsymbol{v}\boldsymbol{v}^\top W^\top + \hat{\psi}^2 W\Pi^\perp\boldsymbol{\theta}\boldsymbol{\theta}^\top\Pi^\perp W^\top + \hat{q}_1\Omega + \hat{q}_2\mathbb{I}_d\right) \left(\hat{V}_1\Omega + \hat{V}_2\mathbb{I}_d + \lambda\mathbb{I}_d\right)^{-1} \right]. \tag{95}$$

We neglected cross-terms of the form $\mathbf{1}_p\boldsymbol{v}^\top W^\top, \mathbf{1}_p\boldsymbol{\theta}^\top\Pi^\top W^\top, W\boldsymbol{v}\boldsymbol{\theta}^\top\Pi^\top W^\top$ in the numerator, which are expected to be asymptotically subleading. For example, one expects

$$\frac{1}{p} \mathrm{Tr}\left[\mathbf{1}_p\boldsymbol{v}^\top W^\top \left(\hat{V}_1\Omega + \hat{V}_2\mathbb{I}_d + \lambda\mathbb{I}_d\right)^{-1} \right] = \frac{1}{p}\boldsymbol{v}^\top W^\top \left(S^{-1} + \frac{\hat{V}_1}{d}\frac{S^{-1}W\boldsymbol{v}\boldsymbol{v}^\top W^\top S^{-1}}{1 - \frac{\hat{V}_1}{d}\mathrm{Tr}[W\boldsymbol{v}\boldsymbol{v}^\top W^\top S^{-1}]} \right) \mathbf{1}_p$$

$$= O_d(1/\sqrt{d}), \tag{96}$$

where we introduced the shorthand $S = \hat{V}_1WW^\top + \hat{V}_2\mathbb{I}_p + \lambda\mathbb{I}_p$ and used the Sherman-Morrison lemma. The last scaling follows from the fact that since the vectors $W\boldsymbol{v}, \mathbf{1}_p$ are weakly correlated, their scalar product –with the uncorrelated matrix S^{-1} in the middle– should be $\boldsymbol{v}^\top W^\top S^{-1}\mathbf{1}_p = O_d(1/\sqrt{d})$. A similar reasoning holds for the other cross-terms.

Energetic term The energetic term is the same as for a Gaussian mixture computation as in e.g. (Mignacco et al., 2020; Loureiro et al., 2021c; Pesce et al., 2023; Cui & Zdeborová, 2024) for Gaussian mixture distributions, with the $\mathbb{E}_\kappa$ playing the role of the sum over clusters. In other words, the equivalent feature map (14) is formally equivalent to an *infinite mixture* of Gaussian clusters indexed by the projection κ on the spike $\boldsymbol{v}$. Note that the formal equivalence to infinite Gaussian mixtures was already studied in (Adomaityte et al., 2023a;b), in the context of learning from heavy-tailed data. Exploiting this formal equivalence, the expectation over $\{\lambda_a\}$ can be carried out and give a Moreau envelope in the $\omega \to \infty$ limit, yielding

$$\Psi_y = -\mathbb{E}_{\kappa,\xi,y}\mathcal{M}(\xi,\kappa,y) \tag{97}$$

where $\xi \sim \mathcal{N}(0,1)$ and

$$\mathcal{M}(\xi,\kappa,y) = \inf_x \left\{ \frac{1}{2V(\kappa)}(x - \mu_1(\kappa)\psi y - \sqrt{\rho q(\kappa) - \mu_1(\kappa)^2\psi^2}\xi)^2 + \ell(\sigma_\star(\kappa,y), \mu_0(\kappa)m + \mu_1(\kappa)\kappa\zeta + x) \right\} \tag{98}$$

we noted

$$q(\kappa) = \mu_1(\kappa)^2 q_1 + \mu_2(\kappa)^2 q_2, \qquad\qquad V(\kappa) = \mu_1(\kappa)^2 V_1 + \mu_2(\kappa)^2 V_2. \tag{99}$$

In (97), all expectations bear over zero-mean, unit variance Gaussian variables. Note that $q(\kappa), V(\kappa)$ are the exact equivalents of q_k, V_k in e.g. (Loureiro et al., 2021c) or (Cui & Zdeborová, 2024), i.e. the overlaps associated to cluster index κ. As previously mentioned, the equivalent map (14) formally coincides with an infinite Gaussian mixture, with κ indexing the cluster.

Trace term Note that in contrast to e.g. (Mignacco et al., 2020; Loureiro et al., 2021c; Cui & Zdeborová, 2024), in the trace term Ψ_t, the $\hat{m}m$ term is multiplied by $\sqrt{d}$ instead of d. This is due to the fact that the cluster centroids, whose role is here played by $u = 1_p$, is of norm $\sqrt{d}$, whereas (Mignacco et al., 2020)(Loureiro et al., 2021c) consider Gaussian clusters separated by centroids of norm $\Theta_d(1)$. The computation of the trace term however follows identical steps as e.g. (Loureiro et al., 2021a), yielding

$$
\begin{aligned}
\Psi_t &= -\frac{1}{2}(\hat{V}_1 q_1 - \hat{q}_1 V_1) - \frac{1}{2}(\hat{V}_2 q_2 - \hat{q}_2 V_2) + \frac{1}{\sqrt{d}} m\hat{m} + \frac{1}{\sqrt{\beta}}\psi\hat{\psi} + \frac{1}{\sqrt{\beta}}\zeta\hat{\zeta} \\
&= -\frac{1}{2}(\hat{V}_1 q_1 - \hat{q}_1 V_1) - \frac{1}{2}(\hat{V}_2 q_2 - \hat{q}_2 V_2) + \frac{1}{\sqrt{\beta}}\psi\hat{\psi} + \frac{1}{\sqrt{\beta}}\zeta\hat{\zeta}
\end{aligned}
\tag{100}
$$

Comparing to (Loureiro et al., 2021c), the $\hat{m}m$ term in the free energy is thus vanishing asymptotically.

Free energy Putting everything together, the free energy is

$$
\begin{aligned}
f = \operatorname*{extr}_{q_\iota, V_\iota, m, \psi, \zeta, \hat{q}_\iota, \hat{V}_\iota, \hat{m}, \hat{\psi}, \hat{\zeta}} &\left\{ -\frac{1}{2}(\hat{V}_1 q_1 - \hat{q}_1 V_1) - \frac{1}{2}(\hat{V}_2 q_2 - \hat{q}_2 V_2) + \frac{1}{\sqrt{\beta}}\psi\hat{\psi} + \frac{1}{\sqrt{\beta}}\zeta\hat{\zeta} \right. \\
&- \frac{1}{2p}\operatorname{Tr}\left[\left(\hat{m}^2 1_p 1_p^\top + \hat{\zeta}^2 W vv^\top W^\top + \hat{\psi}^2 W\Pi^\perp \boldsymbol{\theta}\boldsymbol{\theta}^\top \Pi^\perp W^\top + \hat{q}_1 \Omega + \hat{q}_2 \mathbb{I}_d\right)\left(\hat{V}_1 \Omega + \hat{V}_2 \mathbb{I}_d + \lambda \mathbb{I}_d\right)^{-1}\right] \\
&\left. + \frac{\alpha}{\beta}\mathbb{E}_{\kappa, \xi}\mathcal{M}(\xi, \kappa) \right\}
\end{aligned}
\tag{101}
$$

The free energy can be simplified, noting that the zero-gradient (saddle-point) condition on $\hat{m}$ can be written as

$$
0 \overset{!}{=} \hat{m} \operatorname{Tr}\left[1_p 1_p^\top \left(\hat{V}_1 \Omega + \hat{V}_2 \mathbb{I}_d + \lambda \mathbb{I}_d\right)^{-1}\right],
\tag{102}
$$

which imposes $\hat{m} = 0$. The free energy now simplifies to

$$
\begin{aligned}
f = \operatorname*{extr}_{q_\iota, V_\iota, m, \psi, \zeta, \hat{q}_\iota, \hat{V}_\iota, \hat{\psi}, \hat{\zeta}} &\left\{ -\frac{1}{2}(\hat{V}_1 q_1 - \hat{q}_1 V_1) - \frac{1}{2}(\hat{V}_2 q_2 - \hat{q}_2 V_2) + \frac{1}{\sqrt{\beta}}\psi\hat{\psi} + \frac{1}{\sqrt{\beta}}\zeta\hat{\zeta} \right. \\
&- \frac{1}{2p}\operatorname{Tr}\left[\left(\hat{\zeta}^2 W vv^\top W^\top + \hat{\psi}^2 W\Pi^\perp \boldsymbol{\theta}\boldsymbol{\theta}^\top \Pi^\perp W^\top + \hat{q}_1 \Omega + \hat{q}_2 \mathbb{I}_d\right)\left(\hat{V}_1 \Omega + \hat{V}_2 \mathbb{I}_d + \lambda \mathbb{I}_d\right)^{-1}\right] \\
&\left. + \frac{\alpha}{\beta}\mathbb{E}_{\kappa, \xi}\mathcal{M}(\xi, \kappa) \right\}
\end{aligned}
\tag{103}
$$

Let us massage the expression in the trace further, by explicitly replacing Ω by its definition, namely $\Omega = WW^\top - {}^1/dW vv^\top W^\top$. The term in the trace then becomes

$$
(a) \equiv \frac{1}{p}\operatorname{Tr}\left[\left((\hat{\zeta}^2 - \hat{q}_1/d)W vv^\top W^\top + \hat{\psi}^2 W\Pi^\perp \boldsymbol{\theta}\boldsymbol{\theta}^\top \Pi^\perp W^\top + \hat{q}_1 WW^\top + \hat{q}_2 \mathbb{I}_d\right)\left(\hat{V}_1 WW^\top - \hat{V}_1/dW vv^\top W + \hat{V}_2 \mathbb{I}_d + \lambda \mathbb{I}_d\right)^{-1}\right]
\tag{104}
$$

The $\hat{q}_1/d$ in the numerator can be safely neglected as a small correction to $\hat{\zeta}^2$. On the other hand, one has to take care of the spike in the denominator using the Sherman-Morrison lemma. Using the shorthand

$$
S \equiv \hat{V}_1 WW^\top + \hat{V}_2 \mathbb{I}_p + \lambda \mathbb{I}_p
\tag{105}
$$

one can write the denominator as

$$
(S - \hat{V}_1/dW vv^\top W^\top)^{-1} = S^{-1} + \frac{\hat{V}_1}{d}\frac{S^{-1}W vv^\top W^\top S^{-1}}{1 - \frac{\hat{V}_1}{d}\operatorname{Tr}\left[W vv^\top W^\top S^{-1}\right]}.
\tag{106}
$$

The trace term (a) thus decomposes as

$$
\begin{aligned}
(a) =& \frac{1}{p} \operatorname{Tr}\left[\left(\hat{\zeta}^2 W v v^\top W^\top + \hat{\psi}^2 W \Pi^\perp \boldsymbol{\theta}\boldsymbol{\theta}^\top \Pi^\perp W^\top + \hat{q}_1 W W^\top + \hat{q}_2 \mathbb{I}_p\right) S^{-1}\right] \\
&+ \frac{\hat{V}_1}{pd} \underbrace{\frac{\operatorname{Tr}\left[\left(\hat{\zeta}^2 W v v^\top W^\top + \hat{\psi}^2 W \Pi^\perp \boldsymbol{\theta}\boldsymbol{\theta}^\top \Pi^\perp W^\top + \hat{q}_1 W W^\top + \hat{q}_2 \mathbb{I}_p\right) S^{-1} W v v^\top W^\top S^{-1}\right]}{1 - \frac{\hat{V}_1}{d}\operatorname{Tr}\left[W v v^\top W^\top S^{-1}\right]}}_{(b)}
\end{aligned}
\tag{107}
$$

We turn to the simplification of the numerator of (b), sequentially examining each of the three terms composing it. First,

$$
\hat{\psi}^2 W \Pi^\perp \boldsymbol{\theta}\boldsymbol{\theta}^\top \Pi^\perp W^\top S^{-1} W v v^\top W^\top S^{-1} = \hat{\psi}^2 \left((\Pi^\perp \theta)^\top W^\top S^{-1} W v\right)^2 = O_d(d),
\tag{108}
$$

where we used that $\Pi^\perp \theta, v$ are orthogonal, so the squared scalar product of these vectors, with the uncorrelated matrix in between, should be asymptotically at most $O_d(\sqrt{d})$. Turning to the second term:

$$
\hat{q}_1 W W^\top S^{-1} W v v^\top W^\top S^{-1} = \hat{q}_1 \|W^\top S^{-1} W v\|^2 = O_d(d).
\tag{109}
$$

Finally

$$
\hat{\zeta}^2 W v v^\top W^\top S^{-1} W v v^\top W^\top S^{-1} = \hat{\zeta}^2 \left(v^\top W^\top S^{-1} W v\right)^2 = \Theta_d(d^2).
\tag{110}
$$

In short, only the third term yields a $\Theta_d(1)$ contribution in (b), i.e.

$$
(b) = o_d(1) + \frac{\hat{\zeta}^2 \hat{V}_1}{pd} \frac{\operatorname{Tr}\left[W v v^\top W^\top S^{-1}\right]^2}{1 - \frac{\hat{V}_1}{d}\operatorname{Tr}\left[W v v^\top W^\top S^{-1}\right]}
\tag{111}
$$

Finally, the (a) term can be written be written in fully asymptotic form, by introducing the limiting distribution $\nu(\varrho, \tau\pi)$ of Assumption 3.2:

$$
(a) = \int d\nu(\varrho, \tau, \pi) \frac{\hat{\zeta}^2 \tau^2 \varrho + \hat{\psi}^2 \varrho \pi^2 + \hat{q}_1 \varrho + \hat{q}_2}{\hat{V}_1 \varrho + \hat{V}_2 + \lambda} + \beta \hat{\zeta}^2 \hat{V}_1 \frac{\left(\int d\nu(\varrho, \tau, \pi) \frac{\tau^2 \varrho}{\hat{V}_1 \varrho + \hat{V}_2 + \lambda}\right)^2}{1 - \beta \hat{V}_1 \int d\nu(\varrho, \tau, \pi) \frac{\tau^2 \varrho}{\hat{V}_1 \varrho + \hat{V}_2 + \lambda}},
\tag{112}
$$

thereby yielding the expression for the free energy

$$
\begin{aligned}
f = \operatorname*{extr}_{q_\iota, V_\iota, m, \psi, \zeta, \hat{q}_\iota, \hat{V}_\iota, \hat{\psi}, \hat{\zeta}} \Bigg\{ & -\frac{1}{2}(\hat{V}_1 q_1 - \hat{q}_1 V_1) - \frac{1}{2}(\hat{V}_2 q_2 - \hat{q}_2 V_2) + \frac{1}{\sqrt{\beta}}\psi\hat{\psi} + \frac{1}{\sqrt{\beta}}\zeta\hat{\zeta} + \frac{\alpha}{\beta}\mathbb{E}_{\kappa,\xi}\mathcal{M}(\xi, \kappa) \\
& -\frac{1}{2}\int d\nu(\varrho, \tau, \pi) \frac{\hat{\zeta}^2 \tau^2 \varrho + \hat{\psi}^2 \varrho \pi^2 + \hat{q}_1 \varrho + \hat{q}_2}{\hat{V}_1 \varrho + \hat{V}_2 + \lambda} - \beta \frac{\hat{\zeta}^2 \hat{V}_1}{2} \frac{\left(\int d\nu(\varrho, \tau, \pi) \frac{\tau^2 \varrho}{\hat{V}_1 \varrho + \hat{V}_2 + \lambda}\right)^2}{1 - \beta \hat{V}_1 \int d\nu(\varrho, \tau, \pi) \frac{\tau^2 \varrho}{\hat{V}_1 \varrho + \hat{V}_2 + \lambda}} \Bigg\}
\end{aligned}
\tag{113}
$$

Comments on the limiting density ν   Let us briefly comment on the form of the limiting distribution ν. To build intuition, let us further relax in this paragraph only the strict normalization on the norm of the rows of W, and assume the rows are independently sampled from $\mathcal{N}(0, c\mathbb{I}_d)$. Finally, by the same token, let us assume $\boldsymbol{v} \sim \mathcal{N}(0, \mathbb{I}_d)$ and $\boldsymbol{\theta} \overset{d}{=} \gamma\boldsymbol{v} + \mathcal{N}(0, (1-\gamma^2)\mathbb{I}_d)$, so that they are Gaussian vectors with standard deviation $\sqrt{d}$ and average overlap γ. Note that relaxing the spherical constraint on all vectors, and allowing them to be sampled from Gaussian distributions with matching statistics, is expected to lead to an asymptotically equivalent problem, and is only considered in this paragraph to ease the discussion. In this setting, it is straightforward to see that the measure ν decomposes as the simple product measure

$$
\nu = c \otimes \mu_{\mathrm{MP}}^\beta \times \check{\nu}
\tag{114}
$$

where μ_{MP}^β denotes the Marcenko-Pastur density with aspect ratio β, and the notation $a \otimes \mu$ denotes the pushforward $(a \otimes \mu)(\lambda) = \mu(a\lambda)$. We denoted $\check{\nu}$ the joint distribution of the random variables $x, y - \gamma x$ for $x, y \sim \mathcal{N}(0, 1)$ independently.

C.5. Saddle-point equations

We explicit the saddle-point equations associated to the optimization problem (113). Define $x^\star$ as the minimizer of the Moreau envelope, and the proximal

$$\text{prox}(y, \kappa, \xi) = \frac{1}{V(\kappa)}(x^\star - \mu_1(\kappa)\psi y - \sqrt{\rho q(\kappa) - \psi^2 \mu_1(\kappa)^2}\xi). \tag{115}$$

The saddle-point equations read

$$
\begin{cases}
q_1 = \int d\nu(\varrho, \tau, \pi)\varrho\frac{(\hat{q}_1\varrho + \hat{q}_2 + \hat{\zeta}^2\varrho\tau^2 + \hat{\psi}^2\varrho\pi^2)}{(\lambda + \hat{V}_1\varrho + \hat{V}_2)^2} \\
\quad -\beta\hat{\zeta}^2\frac{\left(\int d\nu(\varrho,\tau,\pi)\frac{\tau^2\varrho}{\hat{V}_1\varrho + \hat{V}_2 + \lambda}\right)^2 + \frac{1}{\beta}\int d\nu(\varrho,\tau,\pi)\frac{\tau^2\varrho^2}{(\lambda + \hat{V}_1\varrho + \hat{V}_2)^2}\left[\left(1 - \beta\hat{V}_1\int d\nu(\varrho,\tau,\pi)\frac{\tau^2\varrho}{\hat{V}_1\varrho + \hat{V}_2 + \lambda}\right)^2 - 1\right]}{\left(1 - \beta\hat{V}_1\int d\nu(\varrho,\tau,\pi)\frac{\tau^2\varrho}{\hat{V}_1\varrho + \hat{V}_2 + \lambda}\right)^2} \\
q_2 = \int d\nu(\varrho, \tau, \pi)\frac{(\hat{q}_1\varrho + \hat{q}_2 + \hat{\zeta}^2\varrho\tau^2 + \hat{\psi}^2\varrho\pi^2)}{(\lambda + \hat{V}_1\varrho + \hat{V}_2)^2} - \hat{\zeta}^2\int d\nu(\varrho,\tau,\pi)\frac{\tau^2\varrho}{(\lambda + \hat{V}_1\varrho + \hat{V}_2)^2}\left[1 - \frac{1}{\left(1 - \beta\hat{V}_1\int d\nu(\varrho,\tau,\pi)\frac{\tau^2\varrho}{\hat{V}_1\varrho + \hat{V}_2 + \lambda}\right)^2}\right] \\
V_1 = \int d\nu(\varrho, \tau, \pi)\varrho\frac{1}{\lambda + \hat{V}_1\varrho + \hat{V}_2} \\
V_2 = \int d\nu(\varrho, \tau, \pi)\frac{1}{\lambda + \hat{V}_1\varrho + \hat{V}_2} \\
m = \inf_m \mathbb{E}_{\kappa,y,\xi}\mathcal{M} \\
\zeta = \hat{\zeta}\sqrt{\beta}\int d\nu(\varrho,\tau,\pi)\varrho\tau^2\frac{1}{\lambda + \hat{V}_1\varrho + \hat{V}_2} + \beta^{3/2}\hat{\zeta}\hat{V}_1\frac{\left(\int d\nu(\varrho,\tau,\pi)\frac{\tau^2\varrho}{\hat{V}_1\varrho + \hat{V}_2 + \lambda}\right)^2}{1 - \beta\hat{V}_1\int d\nu(\varrho,\tau,\pi)\frac{\tau^2\varrho}{\hat{V}_1\varrho + \hat{V}_2 + \lambda}} \\
\psi = \hat{\psi}\sqrt{\beta}\int d\nu(\varrho,\tau,\pi)\varrho\pi^2\frac{1}{\lambda + \hat{V}_1\varrho + \hat{V}_2}
\end{cases}
\tag{116}
$$

$$
\begin{cases}
\hat{V}_1 = -\frac{\alpha}{\beta}\mathbb{E}_{\kappa,y,\xi}\frac{\rho\mu_1(\kappa)^2}{\sqrt{\rho q(\kappa) - \psi^2\mu_1(\kappa)^2}}\text{prox}(y,\kappa,\xi) \\
\hat{q}_1 = \frac{\alpha}{\beta}\mathbb{E}_{\kappa,y,\xi}\mu_1(\kappa)^2\text{prox}(y,\kappa,\xi)^2 \\
\hat{V}_2 = -\frac{\alpha}{\beta}\mathbb{E}_{\kappa,y,\xi}\frac{\rho\mu_2(\kappa)^2}{\sqrt{\rho q(\kappa) - \psi^2\mu_1(\kappa)^2}}\text{prox}(y,\kappa,\xi) \\
\hat{q}_2 = \frac{\alpha}{\beta}\mathbb{E}_{\kappa,y,\xi}\mu_2(\kappa)^2\text{prox}(y,\kappa,\xi)^2 \\
\hat{\zeta} = -\frac{\alpha}{\sqrt{\beta}}\mathbb{E}_{\kappa,y,\xi}\kappa\mu_1(\kappa)\ell'(\sigma_\star(\kappa,y), \mu_0(\kappa)m + \mu_1(\kappa)\kappa\zeta + x^\star) \\
\hat{\psi} = \frac{\alpha}{\sqrt{\beta}}\mathbb{E}_{\kappa,y,\xi}\left(\mu_1(\kappa)y - \frac{\psi\mu_1(\kappa)^2}{\rho q(\kappa) - \psi^2\mu_1(\kappa)^2}\xi\right)\text{prox}(y,\kappa,\xi)
\end{cases}
\tag{117}
$$

C.6. Saddle point equations for the square loss

These expressions simplify for the square loss. Define the minimizer $x^\star$ of the Moreau envelope for $\ell(y, z) = 1/2(y-z)^2$:

$$x^\star(\kappa, \xi) = \frac{V(\kappa)\left(\sigma_\star(\kappa,y) - \mu_0(\kappa)m - \mu_1(\kappa)\kappa\zeta - \mu_1(\kappa)\psi\right) + \sqrt{\rho q(\kappa) - \mu_1(\kappa)^2\psi^2}\xi + \mu_1(\kappa)\psi y}{1 + V(\kappa)}. \tag{118}$$

The saddle point for m admits a compact closed form expression. Writing the saddle point equation for m indeed imposes that

$$0 = \partial_m\mathbb{E}_{\kappa,\xi}\mathcal{M}(\xi, \kappa)$$
$$= \mathbb{E}_{\kappa,\xi}\left[\mu_0(\kappa)\left(f_\star(\kappa) - \mu_0(\kappa)m - \frac{V(\kappa)f^\star(\kappa) - V(\kappa)\mu_0(\kappa)m + \sqrt{q(\kappa)}\xi}{1 + V(\kappa)}\right)\right], \tag{119}$$

from which it follows that

$$m\mathbb{E}_\kappa\left[\frac{\mu_0(\kappa)^2}{1 + V(\kappa)}\right] = \mathbb{E}_\kappa\left[\frac{\mu_0(\kappa)f_\star(\kappa)}{1 + V(\kappa)}\right], \tag{120}$$

i.e.

$$m = \frac{\mathbb{E}_\kappa \left[\frac{\mu_0(\kappa) f_\star(\kappa)}{1+V(\kappa)} \right]}{\mathbb{E}_\kappa \left[\frac{\mu_0(\kappa)^2}{1+V(\kappa)} \right]} \tag{121}$$

The saddle-point equations thus read

$$
\begin{cases}
q_1 = \int d\nu(\varrho,\tau,\pi) \varrho \frac{\left(\hat{q}_1 \varrho + \hat{q}_2 + \hat\zeta^2 \varrho\tau^2 + \hat\psi^2 \varrho\pi^2\right)}{\left(\lambda + \hat{V}_1 \varrho + \hat{V}_2\right)^2} \\
\qquad - \beta\hat\zeta^2 \frac{\left(\int d\nu(\varrho,\tau,\pi) \frac{\tau^2 \varrho}{\hat{V}_1 \varrho + \hat{V}_2 + \lambda}\right)^2 + \frac{1}{\beta} \int d\nu(\varrho,\tau,\pi) \frac{\tau^2 \varrho^2}{(\lambda + \hat{V}_1 \varrho + \hat{V}_2)^2} \left[\left(1 - \beta\hat{V}_1 \int d\nu(\varrho,\tau,\pi) \frac{\tau^2 \varrho}{\hat{V}_1 \varrho + \hat{V}_2 + \lambda}\right)^2 - 1\right]}{\left(1 - \beta\hat{V}_1 \int d\nu(\varrho,\tau,\pi) \frac{\tau^2 \varrho}{\hat{V}_1 \varrho + \hat{V}_2 + \lambda}\right)^2} \\
q_2 = \int d\nu(\varrho,\tau,\pi) \frac{\left(\hat{q}_1 \varrho + \hat{q}_2 + \hat\zeta^2 \varrho\tau^2 + \hat\psi^2 \varrho\pi^2\right)}{\left(\lambda + \hat{V}_1 \varrho + \hat{V}_2\right)^2} - \hat\zeta^2 \int d\nu(\varrho,\tau,\pi) \frac{\tau^2 \varrho}{(\lambda + \hat{V}_1 \varrho + \hat{V}_2)^2} \left[1 - \frac{1}{\left(1 - \beta\hat{V}_1 \int d\nu(\varrho,\tau,\pi) \frac{\tau^2 \varrho}{\hat{V}_1 \varrho + \hat{V}_2 + \lambda}\right)^2}\right] \\
V_1 = \int d\nu(\varrho,\tau,\pi) \varrho \frac{1}{\lambda + \hat{V}_1 \varrho + \hat{V}_2} \\
V_2 = \int d\nu(\varrho,\tau,\pi) \frac{1}{\lambda + \hat{V}_1 \varrho + \hat{V}_2} \\
m = \frac{1}{\mathbb{E}_\kappa \left[\frac{\mu_0(\kappa)^2}{1+V(\kappa)}\right]} \mathbb{E}_{\kappa,y} \left[\frac{\mu_0(\kappa)(\sigma_\star(\kappa,y) - \mu_1(\kappa)\kappa\zeta)}{1+V(\kappa)}\right] \\
\zeta = \hat\zeta \sqrt{\beta} \int d\nu(\varrho,\tau,\pi) \varrho\tau^2 \frac{1}{\lambda + \hat{V}_1 \varrho + \hat{V}_2} + \beta^{3/2} \hat\zeta \hat{V}_1 \frac{\left(\int d\nu(\varrho,\tau,\pi) \frac{\tau^2 \varrho}{\hat{V}_1 \varrho + \hat{V}_2 + \lambda}\right)^2}{1 - \beta\hat{V}_1 \int d\nu(\varrho,\tau,\pi) \frac{\tau^2 \varrho}{\hat{V}_1 \varrho + \hat{V}_2 + \lambda}} \\
\psi = \hat\psi \sqrt{\beta} \int d\nu(\varrho,\tau,\pi) \varrho\pi^2 \frac{1}{\lambda + \hat{V}_1 \varrho + \hat{V}_2}
\end{cases}
\tag{122}
$$

$$
\begin{cases}
\hat{V}_1 = \frac{\alpha}{\beta} \mathbb{E}_\kappa \frac{\rho\mu_1(\kappa)^2}{1+V(\kappa)} \\
\hat{q}_1 = \frac{\alpha}{\beta} \mathbb{E}_{\kappa,y} \mu_1(\kappa)^2 \frac{b(\kappa,y)^2 + q(\kappa) - \mu_1(\kappa)^2 \psi^2}{\left(1+V(\kappa)\right)^2} \\
\hat{V}_2 = \frac{\alpha}{\beta} \mathbb{E}_\kappa \frac{\rho\mu_2(\kappa)^2}{1+V(\kappa)} \\
\hat{q}_2 = \frac{\alpha}{\beta} \mathbb{E}_{\kappa,y} \mu_2(\kappa)^2 \frac{b(\kappa,y)^2 + q(\kappa) - \mu_1(\kappa)^2 \psi^2}{\left(1+V(\kappa)\right)^2} \\
\hat\zeta = \frac{\alpha}{\sqrt{\beta}} \mathbb{E}_{\kappa,y} \kappa\mu_1(\kappa) \frac{b(\kappa,y)}{1+V(\kappa)} \\
\hat\psi = \frac{\alpha}{\sqrt{\beta}} \sqrt{\beta} \mathbb{E}_{\kappa,y} \frac{y\mu_1(\kappa)b(\kappa,y) + \psi\mu_1(\kappa)^2}{1+V(\kappa)}
\end{cases}
\quad , \tag{123}
$$

which recovers equations (116) from Result 3.3.

C.7. Test error

We conclude this Appendix by deriving the sharp asymptotic characterization for the test error. Using once more Result 3.2, the test error reads

$$
\begin{aligned}
\epsilon_g &= \mathbb{E}_x \left(\sigma_\star(\kappa, \boldsymbol\theta^\top \Pi^\perp x / \sqrt{d}) - \hat{a}^\top \varphi^g(x) / \sqrt{p}\right)^2 \\
&= \mathbb{E}_{x,\xi} \left(\sigma_\star(\kappa, \boldsymbol\theta^\top \Pi^\perp x / \sqrt{d}) - \mu_0(\kappa) \frac{\hat{a}^\top 1_p}{\sqrt{p}} - \mu_1(\kappa)\kappa \frac{\hat{a}^\top W v}{\sqrt{dp}} - \mu_1(\kappa) \frac{\hat{a}^\top W \Pi^\perp x}{\sqrt{p}} - \mu_2(\kappa)\xi\right)^2 \\
&= \mathbb{E}_{\kappa,y,z} \left(\sigma_\star(\kappa,y) - \mu_0(\kappa)m - \mu_1(\kappa)\zeta - z\right)^2
\end{aligned}
\tag{124}
$$

with

$$
y, z \sim \mathcal{N}\left[0, \begin{pmatrix} \rho & \mu_1(\kappa)\psi \\ \mu_1(\kappa)\psi & q(\kappa) \end{pmatrix}\right]. \tag{125}
$$

23

The expression thus simplifies to

$$\epsilon_g = \mathbb{E}_{\kappa,y} \left[\left(\sigma_\star(\kappa, y) - \mu_0(\kappa)m - \mu_1(\kappa)\kappa\zeta - \frac{\mu_1(\kappa)\psi}{\sqrt{\rho}} y \right)^2 + q(\kappa) - \frac{\mu_1(\kappa)^2\psi^2}{\rho} \right] \tag{126}$$

where all expectations now bear on unit variance zero mean Gaussian variables.

C.8. Additional discussion

Before closing the present Appendix, we provide in this subsection further discussion of the results and highlight possible extensions.

Bias– Let us briefly sketch in this subsection how bias terms may be incorporated into the analysis. Consider the two-layer neural network

$$f_{W,\boldsymbol{a},b}(\boldsymbol{x}) = \frac{1}{\sqrt{p}} \sum_{i=1}^{p} a_i \sigma(\boldsymbol{w}_i^\top \boldsymbol{x} + b), \tag{127}$$

where $b \in \mathbb{R}$ is a bias that acts uniformly on all hidden-layer neurons. Then, conditional on b, the equivalent feature map still is of the form (14), with coefficients

$$
\begin{aligned}
\mu_0(\kappa, b) &= \mathbb{E}_z \, \sigma(z + r\kappa + b) \\
\mu_1(\kappa, b) &= \frac{1}{c} \mathbb{E}_z \, z\sigma(z + r\kappa + b) \\
\mu_2(\kappa, b) &= \sqrt{\mathbb{E}_z \, \sigma^2(z + r\kappa + b) - c(\mu_1(\kappa))^2 - (\mu_0(\kappa))^2}.
\end{aligned}
\tag{128}
$$

The analysis of the present Appendix still goes through, by treating b as an additional network parameter to be replicated over. In fact, the only difference in introducing b appears at the level of the Moreau envelope $\mathcal{M}$, in which all the $\mu_{0,1,2}(\kappa)$ coefficients are replaced by $\mu_{0,1,2}(\kappa + b)$. Finally, the corresponding saddle-point equation on b should impose $\partial_b \mathcal{M} = 0$, thereby imposing the value of the learned bias $\hat{b}$. This allows to improve the lower-bound (27) as

$$\epsilon_g \geq \inf_{b_1,b_2,b} \mathbb{E}_\kappa \left[\left(\sigma_\star(\kappa) - \mu_0(\kappa, b)b_1 - \mu_1(\kappa, b)\kappa b_2 \right)^2 \right]. \tag{129}$$

In other words, the bias allows an extra degree of freedom in the functional basis expressible by the network, thus allowing additional expressivity. We have addressed the case of uniform bias. While the case of generic non-uniform bias is unfortunately not presently in the reach of the techniques of this manuscript, the case where the components are untrained, but can take finitely many values, can be analyzed along similar lines as the case of non-uniform u. We briefly discuss this in Appendix D.

Larger rank spikes– Let us now discuss the case where instead of a spike $1_p v^\top / \sqrt{d}$, we have a larger rank term $1_{p \times k} v^\top / \sqrt{d}$, with $k = \Theta_d(1)$ and where $v \in \mathbb{R}^{k \times d}$ is now a low-rank matrix instead of a vector. The equivalence of result 3.2 can be readily adapted as follows. Naming $v_1, ..., v_k \in \mathbb{R}^d$ the k columns of v, and conditioning on $\kappa_1, ..., \kappa_k$ the k projections of the data along these directions, the equivalent feature map (14) retains the same expression, with the coefficient $\mu'_{0,1,2}(\kappa_1, ..., \kappa_k) = \mu_{0,1,2}(\kappa_1 + ... + \kappa_k)$ now becoming multivariate.

Noisy target– Finally, let us mention how one may adapt the analysis to a noisy target function $f_\star(x) + \mathcal{N}(0, 1)$. The only point where this intervenes is during the evaluation of the energetic potential Ψ_y, which is identical to e.g. (Loureiro et al., 2021b), Appendix A. The only effect is to change the distribution of the variable y in the expectation over the Moreau envelope, and we refer the interested reader to e.g. (Loureiro et al., 2021b) for a full expression.

Effect of network width– What is the effect of the network width, as measured by β? On the one hand, large networks result in smaller spike strength r, as can be observed from Result 3.1. On the other hand, it also allows for more overparametrization. The asymptotic characterization of Result 3.3 allows to probe this question. Fig. 7 shows that larger networks tend to achieve lower overall test error.
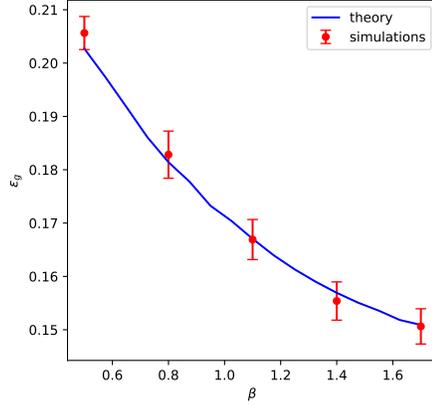
Figure 7. $\sigma = \tanh, \sigma_\star = \sin, \alpha_0 = 1.5, \alpha = 1.2, \lambda = 0.1$ Test error as a function of the network width β, as predicted by the theoretical characterization of Result 3.3 (blue) or measure in numerical simulations in $d = 5000$ (red); error bars represent one standard deviation over 20 trials.

D. Non-Uniform Readout Initialization

D.1. Setting

In this Appendix, we briefly discuss the case where the readout layer is not initialized as a vector with equal components, $a^{(0)} = 1_p/\sqrt{p}$, but the case where the components of $a^{(0)}$ can take values in a *finite* vocabulary V, for instance $\{-1, 0, +1\}$. We further assume the fraction of all letters are asymptotically finite (i.e. the number of occurences of a letter $\sigma \in V$ in $a^{(0)}$, divided by d, tends to a well-defined finite value as $d \to \infty$). For simplicity, we provide a sketch of the derivation directly for the sRF model (7), and do not explicit for conciseness the mapping from the gradient-trained network thereto, which follows identical steps from the mapping detailed in Appendix A. More precisely, we consider directly a sRF with weights

$$W + \frac{uv^\top}{\sqrt{d}}. \tag{130}$$

In contrast to the main text, we thus do not assume $u \propto 1_p$. We further also allow the norm of the rows of W to take different values $c(u_i)^{1/2}$ depending on the value taken by the component of u with the corresponding index. This assumption originates from the form of the Δ matrix (33), which exhibits such row-wise norm variations matching that of $a^{(0)}$ (thus u in the equivalent sRF). Finally, for simplicity and not to complicate the formulae, we assume perfect alignment with the target $\gamma = 1$, and assume aspect ratio $p = d$, i.e. $\beta = 1$. The generic case $0 \le \gamma \le 1$ can be derived following identical steps to Appendix C.

D.2. Sketch of derivation

For non-uniform u, an equivalent feature map can be obtained along the same lines as for 3.2, as detailed in Appendix B. The sRF feature map has an asymptotically equivalent test error as the equivalent feature map

$$\varphi^g(x) = \sum_{\sigma \in V} \mu_0^\sigma(\kappa) 1_\sigma + \mu_1^\sigma(\kappa) \Pi_\sigma W x + \mu_2^\sigma(\kappa) \Pi_\sigma \xi \tag{131}$$

where

$$1_\sigma = (\delta_{u_i,\sigma})_{i=1}^d \qquad\qquad \Pi_\sigma = \mathrm{diag}(1_\sigma) \in \mathbb{R}^{d \times d} \tag{132}$$

25

and ($z \sim \mathcal{N}(0,1)$ in the following)

$$\mu_0^\tau(\kappa) = \mathbb{E}_{z \sim \mathcal{N}(0,c(\tau))}[\sigma(z + \tau\kappa)] \tag{133}$$

$$\mu_1^\tau(\kappa) = \frac{1}{c(\tau)}\mathbb{E}_{z \sim \mathcal{N}(0,c(\tau))}[z\sigma(z + \tau\kappa)] \tag{134}$$

$$\mu_0^\tau(\kappa) = \sqrt{\mathbb{E}_z[\sigma(z+\tau\kappa)^2] - \mu_0^\tau(\kappa)^2 - c(\tau)\mu_1^\tau(\kappa)^2} \tag{135}$$

Note $\Pi_\sigma \Pi_\tau = \delta_{\sigma\tau}\Pi_\sigma$. Building on this equivalent map, the test error can by asymptotically characterized using the replica method, following the same lines as Appendix C. Like in Appendix C, we need to compute

$$(a) = \mathbb{E}_{z,\xi}\prod_{a=1}^{s} e^{-\omega\ell(\sigma_\star(\kappa),\sum_\sigma \mu_0^\sigma(\kappa)m_\sigma^a + \mu_1^\sigma(\kappa)\kappa\zeta_\sigma^a + \mu_1^\sigma(\kappa)g_\sigma^a + \mu_2^\sigma(\kappa)h_\sigma^a)} \tag{136}$$

We defined the order parameters and local fields

$$m_\sigma^a = \frac{a_a^\top 1_\sigma}{\sqrt{d}} \qquad \zeta_\sigma^a = \frac{a_a^\top \Pi_\sigma W v}{d} \qquad g_\sigma^a = \frac{a_a^\top \Pi_\sigma W \Pi^\perp x}{\sqrt{d}} \qquad h_\sigma^a = \frac{a_a^\top \Pi_\sigma \xi}{\sqrt{d}} \tag{137}$$

Again, let us introduce the self-overlaps

$$(q_{\sigma\tau}^1)_{ab} = \frac{a_a^\top \Omega_{\sigma\tau} a_b}{d} \qquad\qquad (q_\sigma^2)_{ab} = \frac{a_a^\top \Pi_\sigma a_b}{d} \tag{138}$$

where we denoted

$$\Omega_{\sigma\tau} = \Pi_\sigma W \Pi W^\top \Pi_\tau \tag{139}$$

One can then rewrite (a) more compactly as

$$(a) = \mathbb{E}_\kappa\left[\mathbb{E}_{\{\lambda_a\}_{a=1}^s} e^{-\omega\sum_{a=1}^s \ell(\sigma_\star(\kappa),m^a(\kappa)+\lambda_a)}\right] \tag{140}$$

where the fields λ_a are Gaussian with covariance

$$\langle\lambda_a\lambda_b\rangle = q_{ab}(\kappa) \tag{141}$$

where

$$q(\kappa)_{ab} = \sum_{\sigma\tau}\mu_1^\sigma(\kappa)\mu_1^\tau(\kappa)q_{\sigma\tau,ab}^1 + \sum_\sigma \mu_2^\sigma(\kappa)^2 q_{\sigma,ab}^2 \tag{142}$$

$$m^a(\kappa) = \sum_{\sigma\in V}(\mu_0^\sigma(\kappa)m_\sigma^a(\kappa) + \mu_1^\sigma(\kappa)\kappa\zeta_\sigma^a(\kappa)) \tag{143}$$

We again use the RS ansatz form

$$\begin{aligned}
&\forall\sigma,\tau\in V, \ \forall 1\le r,r'\le s, \ (q_{\sigma\tau}^\iota)_{rr'} = \delta_{rr'}(r_{\sigma\tau}^\iota - q_{\sigma\tau}^\iota) + q_{\sigma\tau}^\iota \\
&\forall\sigma\in V, \ \forall 1\le r\le s, \ (m_\sigma)_r = m_\sigma \\
&\forall\sigma\in V, \ \forall 1\le r\le s, \ (\zeta_\sigma)_r = \zeta_\sigma
\end{aligned} \tag{144}$$

for $\iota = 1,2$, and similarly for the hat variables:

$$\begin{aligned}
&\forall\sigma,\tau\in V, \ \forall 1\le r,r'\le s, \ (\hat{q}_{\sigma\tau}^\iota)_{rr'} = \delta_{ab}(-\frac{1}{2}\hat{r}_{\sigma\tau}^\iota - \hat{q}_{\sigma\tau}^\iota) + \hat{q}_{\sigma\tau}^\iota \\
&\forall\sigma\in V, \ \forall 1\le r\le s, \ (\hat{m}_\sigma)_r = \hat{m}_\sigma \\
&\forall\sigma\in V, \ \forall 1\le r\le s, \ (\hat{\zeta}_\sigma)_r = \hat{\zeta}_\sigma
\end{aligned} \tag{145}$$

Finally, denoting

$$q(\kappa) = \sum_{\sigma\tau} \mu_1^\sigma(\kappa)\mu_1^\tau(\kappa)q_{\sigma\tau}^1 + \sum_\sigma \mu_2^\sigma(\kappa)^2 q_\sigma^2 \tag{146}$$

$$V(\kappa) = \sum_{\sigma\tau} \mu_1^\sigma(\kappa)\mu_1^\tau(\kappa)V_{\sigma\tau}^1 + \sum_\sigma \mu_2^\sigma(\kappa)^2 V_\sigma^2 \tag{147}$$

one reaches along identical lines as Appendix C the free energy

$$f = -\frac{1}{2}\sum_{\sigma\tau}(\hat{V}_{\sigma\tau}^1 q_{\sigma\tau}^1 - \hat{q}_{\sigma\tau}^1 V_{\sigma\tau}^1) - \frac{1}{2}\sum_\sigma(\hat{V}_\sigma^2 q_\sigma^2 - \hat{q}_\sigma^2 V_\sigma^2) + \sum_\sigma \zeta_\sigma \hat{\zeta}_\sigma$$
$$-\frac{1}{2d}\mathrm{Tr}\left[\left(\sum_{\sigma\leq\tau}\hat{q}_{\sigma\tau}^1\Omega_{\sigma\tau} + \sum_{\sigma\in V}\hat{q}_\sigma^2\Pi_\sigma + \sum_\sigma \hat{\zeta}_\sigma^2\Pi_\sigma W v v^\top W^\top \Pi_\sigma\right)\left(\sum_{\sigma\leq\tau}\hat{V}_{\sigma\tau}^1\Omega_{\sigma\tau} + \sum_{\sigma\in V}\hat{V}_\sigma^2\Pi_\sigma + \lambda\mathbb{I}_d\right)^{-1}\right]$$
$$+ \alpha\mathbb{E}_{\kappa,\xi}\mathcal{M}(\xi,\kappa) \tag{148}$$

where

$$\Omega_{\sigma\tau} = \frac{1}{2}\Pi_\sigma W\Pi W^\top\Pi_\tau + \frac{1}{2}\Pi_\tau W\Pi W^\top\Pi_\sigma. \tag{149}$$

The expression of the Moreau is almost unchanged compared to Appendix C:

$$\mathcal{M}(\xi,\kappa) = \inf_x \frac{1}{2V(\kappa)}(x - \sqrt{q(\kappa)}\xi)^2 + \ell(f^\star(\kappa), m(\kappa) + x), \tag{150}$$

where we noted

$$m(\kappa) = \sum_{\sigma\in V}\mu_0^\sigma(\kappa)m_\sigma + \mu_1^\sigma(\kappa)\kappa\zeta_\sigma. \tag{151}$$

The saddle point equations read, for the square loss $\ell(y,z) = 1/2(y-z)^2$:

27

$$
\begin{cases}
q^1_{\sigma\tau} = \frac{1}{d}\operatorname{Tr}\left[\Omega_{\sigma\tau}\left(\sum_{\sigma\leq\tau}\hat{V}^1_{\sigma\tau}\Omega_{\sigma\tau} + \sum_{\sigma\in V}\hat{V}^2_{\sigma}\Pi_{\sigma} + \lambda\mathbb{I}_d\right)^{-2}\left(\sum_{\sigma\leq\tau}\hat{q}^1_{\sigma\tau}\Omega_{\sigma\tau} + \sum_{\sigma\in V}\hat{q}^2_{\sigma}\Pi_{\sigma} + \sum_{\sigma}\hat{\zeta}^2_{\sigma}\Pi_{\sigma}Wvv^\top W^\top\Pi_{\sigma}\right)\right] \\[2mm]
q^2_{\sigma} = \frac{1}{d}\operatorname{Tr}\left[\Pi_{\sigma}\left(\sum_{\sigma\leq\tau}\hat{V}^1_{\sigma\tau}\Omega_{\sigma\tau} + \sum_{\sigma\in V}\hat{V}^2_{\sigma}\Pi_{\sigma} + \lambda\mathbb{I}_d\right)^{-2}\left(\sum_{\sigma\leq\tau}\hat{q}^1_{\sigma\tau}\Omega_{\sigma\tau} + \sum_{\sigma\in V}\hat{q}^2_{\sigma}\Pi_{\sigma} + \sum_{\sigma}\hat{\zeta}^2_{\sigma}\Pi_{\sigma}Wvv^\top W^\top\Pi_{\sigma}\right)\right] \\[2mm]
V^1_{\sigma\tau} = \frac{1}{d}\operatorname{Tr}\left[\left(\sum_{\sigma\leq\tau}\hat{V}^1_{\sigma\tau}\Omega_{\sigma\tau} + \sum_{\sigma\in V}\hat{V}^2_{\sigma}\Pi_{\sigma} + \lambda\mathbb{I}_d\right)^{-1}\Omega_{\sigma\tau}\right] \\[2mm]
V^2_{\sigma} = \frac{1}{d}\operatorname{Tr}\left[\left(\sum_{\sigma\leq\tau}\hat{V}^1_{\sigma\tau}\Omega_{\sigma\tau} + \sum_{\sigma\in V}\hat{V}^2_{\sigma}\Pi_{\sigma} + \lambda\mathbb{I}_d\right)^{-1}\Pi_{\sigma}\right] \\[2mm]
\zeta_{\sigma} = \frac{\hat{\zeta}_{\sigma}}{d}\operatorname{Tr}\left[\left(\sum_{\sigma\leq\tau}\hat{V}^1_{\sigma\tau}\Omega_{\sigma\tau} + \sum_{\sigma\in V}\hat{V}^2_{\sigma}\Pi_{\sigma} + \lambda\mathbb{I}_d\right)^{-1}\Pi_{\sigma}Wvv^\top W^\top\Pi_{\sigma}\right] \\[2mm]
m_{\sigma} = \frac{1}{\mathbb{E}_\kappa\left[\frac{\mu^\sigma_0(\kappa)^2}{1+V(\kappa)}\right]}\mathbb{E}_\kappa\left[\frac{\mu^\sigma_0(\kappa)\left(f^\star(\kappa) - \sum_{\tau\neq\sigma}\mu^\tau_0(\kappa)m_\tau - \sum_{\tau}\mu^\tau_1(\kappa)\kappa\zeta_\tau\right)}{1+V(\kappa)}\right]
\end{cases}
\tag{152}
$$

$$
\begin{cases}
\hat{V}^1_{\sigma\tau} = \alpha\mathbb{E}_\kappa\frac{\mu^\sigma_1(\kappa)\mu^\tau_1(\kappa)}{1+V(\kappa)} \\[2mm]
\hat{q}^1_{\sigma\tau} = \alpha\mathbb{E}_\kappa\mu^\sigma_1(\kappa)\mu^\tau_1(\kappa)\frac{(f^\star(\kappa)-m(\kappa))^2+q(\kappa)}{\left(1+V(\kappa)\right)^2} \\[2mm]
\hat{V}^2_{\sigma} = \alpha\mathbb{E}_\kappa\frac{\mu^\sigma_2(\kappa)^2}{1+V(\kappa)} \\[2mm]
\hat{q}^2_{\sigma} = \alpha\mathbb{E}_\kappa\mu^\sigma_2(\kappa)^2\frac{(f^\star(\kappa)-m(\kappa))^2+q(\kappa)}{\left(1+V(\kappa)\right)^2} \\[2mm]
\hat{\zeta}_{\sigma} = \alpha\mathbb{E}_\kappa\mu^\sigma_1(\kappa)\kappa\frac{(f^\star(\kappa)-m(\kappa))^2+q(\kappa)}{\left(1+V(\kappa)\right)^2}
\end{cases}
\tag{153}
$$

Importantly, these equations are conjectured to be asymptotically exact, but note that they are not fully asymptotic, in the sense that they still involve traces of large-dimensional matrices. Finally, the test error admits the compact asymptotic expression

$$
\begin{aligned}
\epsilon_g &= \mathbb{E}_\kappa\left[\left(\sigma_\star(\kappa) - \sum_{\sigma\in V}\left[\mu^\sigma_0(\kappa)m_\sigma + \mu^\sigma_1(\kappa)\kappa\zeta_\sigma\right]\right)^2 + q(\kappa)\right] \\
&= \mathbb{E}_\kappa\left[\left(\sigma_\star(\kappa) - m(\kappa)\right)^2 + q(\kappa)\right]
\end{aligned}
\tag{154}
$$

D.3. The special case $W = \Pi^\perp$

We briefly discuss a special case of technical interest, for which the saddle-point equations admit a fully asymptotic formulation, namely when $W = \Pi^\perp$. Then

$$
\Omega_{\sigma\tau} \approx \delta_{\sigma\tau}\Pi_\sigma
\tag{155}
$$
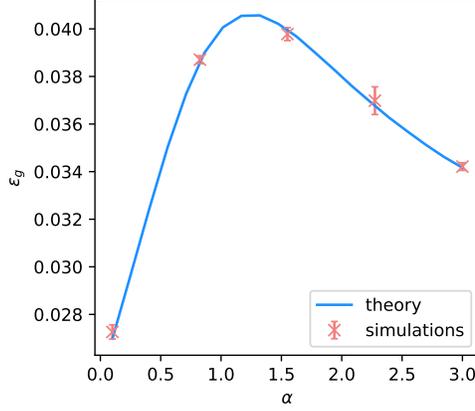
Figure 8. Test error for a sRF with $c = 1, r = 1, \gamma = 1$, and ReLU activation for non-uniform u in the special case $W = \Pi^\perp$, when learning from a single-index model with $\sigma_\star =$ReLU. A regularization $\lambda = 0.1$ is used to train the readout. The components of u were drawn as independent Rademacher variables, $u_i \sim \text{Rad}(1/2) = 1/2\delta_{\cdot,1} + 1/2\delta_{\cdot,-1}$. Solid line: theoretical prediction of (156), for $W = \Pi^\perp$. Crosses represent numerical experiments in dimension $d = 2000$, averaged over 10 runs.

and the saddle point equations reduce to

$$
\begin{cases}
q_\sigma = n_\sigma \dfrac{\hat{q}_\sigma}{\left(\hat{V}_\sigma + \lambda\right)^2} \\
V_\sigma = n_\sigma \dfrac{1}{\hat{V}_\sigma + \lambda} \\
m_\sigma = \dfrac{1}{\mathbb{E}_\kappa\left[\frac{\mu_0^\sigma(\kappa)^2}{1+V(\kappa)}\right]} \mathbb{E}_\kappa\left[\dfrac{\mu_0^\sigma(\kappa)(f^\star(\kappa) - \sum_{\tau \neq \sigma} \mu_0^\tau(\kappa)m_\tau)}{1+V(\kappa)}\right]
\end{cases}
\tag{156}
$$

$$
\begin{cases}
\hat{V}_\sigma = \alpha\mathbb{E}_\kappa \dfrac{\mu_1^\sigma(\kappa)^2 + \mu_2^\sigma(\kappa)^2}{1+V(\kappa)} \\
\hat{q}_\sigma = \alpha\mathbb{E}_\kappa(\mu_1^\sigma(\kappa)^2 + \mu_2^\sigma(\kappa)^2)\dfrac{(f^\star(\kappa) - m(\kappa))^2 + q(\kappa)}{\left(1+V(\kappa)\right)^2}
\end{cases}
\tag{157}
$$

Where n_σ is the fraction of components of u taking value σ. Fig. 8 represents the theoretical prediction of (156), and displays good agreement with the corresponding numerical experiments.

Note that the formulae (156) are further amenable to being rewritten in functional form, denoting $\nu(x)$ the probability density of the components of u. One can then conjecture in this limit the following limiting equations

$$
\begin{cases}
q(x) = \nu(x)\dfrac{\hat{q}(x)}{\left(\hat{V}(x) + \lambda\right)^2} \\
V(x) = \nu(x)\dfrac{1}{\hat{V}(x) + \lambda} \\
m(x) = \dfrac{1}{\mathbb{E}_\kappa\left[\frac{\mu_0(x,\kappa)^2}{1+V(\kappa)}\right]} \mathbb{E}_\kappa\left[\dfrac{\mu_0(x,\kappa)(f^\star(\kappa) - \int dz\nu(z)\mu_0(z,\kappa)m(z))}{1+V(\kappa)}\right]
\end{cases}
\tag{158}
$$

$$
\begin{cases}
\hat{V}(x) = \alpha\mathbb{E}_\kappa \dfrac{(\mu_2(x,\kappa)^2 + \mu_1(x,\kappa)^2)}{1+V(\kappa)} \\
\hat{q}(x) = \alpha\mathbb{E}_\kappa(\mu_2(x,\kappa)^2 + \mu_1(x,\kappa)^2)\dfrac{(f^\star(\kappa) - m(\kappa))^2}{\left(1+V(\kappa)\right)^2}
\end{cases}
\tag{159}
$$

With the notations

$$
q(\kappa) = \int dx\nu(x)(\mu_1(x,\kappa)^2 + \mu_2(x,\kappa)^2)q(x)
\tag{160}
$$

$$
V(\kappa) = \int dx\nu(x)(\mu_1(x,\kappa)^2 + \mu_2(x,\kappa)^2)V(x).
\tag{161}
$$

29
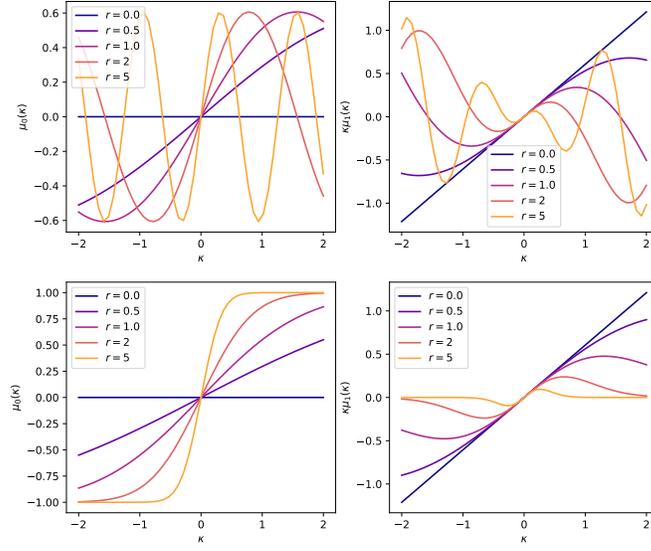
Figure 9. Vizualization of the functions $\mu_0(r), \tilde{\mu}_1(r)$ for $\sigma = \tanh$ (bottom) and $\sigma = \sin$ (top), for various parameters r, in the perfectly aligned case $\gamma = 1, c = 1$, as can be achieved for $\alpha_0 \to \infty$, see also discussion below equation (27) in the main text. We remind that the spike strength r cna be tuned via the learning rate $\tilde{\eta}$. The functions $\mu_0(r), \tilde{\mu}_1(r)$ provide the basis of functions that can be used to approximate the target $\sigma_\star$.

As discussed in the main text, the exploration of the case of non-uniform u is of interest, as it affords a richer functional basis to learn the target $\sigma_\star$ (see also discussion below (27)). This study falls well beyond the scope of the present work, and is left for future work.

A brief comment on incorporating bias   As mentioned in Appendix C, we can accommodate the case of non-uniform quenched bias in the analysis, following identical lines as the one presented above, when the components of the bias can take values in a finite vocabulary V_b. Note that the initial step of the computation involves partitioning the indices $1 \le i \le d$ as a function of the values of u_i, regrouping is corresponding to the same values of $u_i = \sigma$ into the vectors 1_σ and matrices Ω_σ. If non-uniform biases are present, identical conditioning should be done, this time regrouping terms corresponding to the same *joint* value σ, b of u_i, b_i. One implications is that instead of having $\mu_{0,1,2}^\sigma(\kappa)$ as above, one needs to consider $\mu_{0,1,2}^{\sigma,b}(\kappa)$. The analysis proceeds identically. At the level of the network expressivity, quenched but non-uniform biases allow the network to express functions from an even richer basis $\{\mu_0(\omega \cdot + b), \tilde{\mu}_1(\omega + b \cdot)\}_{(b,\omega) \in \mathcal{V}}$, where $\mathcal{V} \subset V_b \times V$ correspond to all realized pairs b, ω by (b_i, u_i).

E. Derivation of the Bounds (26) and (27)

We detail – and illustrate– in this Appendix the derivation of the bounds (26) and (27) on the test error. Again, for definiteness, we consider the case of perfect spike/target alignment $\gamma = 1$ (as can be achieved e.g. by using a large amount of data $\alpha_0 \to \infty$ in the first step).

E.1. Derivation of the upper bound (26)

We first focus on the upper bound (26), showing that it can be attained for $\lambda \to \infty$. In this limit, equations (116) imply that all order parameters but m are vanishing $q_1 = q_2 = V_1 = V_2 = \zeta = \psi = 0$. Then the expression for m simplifies as

$$m = \frac{\mathbb{E}_\kappa[\mu_0(\kappa)\sigma_\star(\kappa)]}{\mathbb{E}_\kappa[\mu_o(\kappa)^2]} = \frac{\langle \mu_0, \sigma_\star \rangle}{\langle \mu_0, \mu_0 \rangle}, \tag{162}$$

where for two univariate functions f, g we denoted the scalar product with respect to the Gaussian measure:

$$\langle f, g \rangle \equiv \int \frac{1}{\sqrt{2\pi}} e^{-x^2/2} f(x) g(x). \tag{163}$$

In other terms, m is proportional to the projection of the target function $\sigma_\star$ on $\mathrm{span}(\mu_0)$. Note that m is can also be rewritten as

$$m = \underset{b_1}{\arg\inf} \|\sigma_\star - b_1 \mu_0\|^2, \tag{164}$$

where the norm is with respect to the scalar product (163). Finally, from (25), the test error in the $\lambda \to \infty$ limit admits the compact expression

$$\lim_{\lambda \to \infty} \epsilon_g = \mathbb{E}_\kappa[(\sigma_\star(\kappa) - m\mu_0(\kappa))^2] = \inf_{b_1} \|\sigma_\star - b_1 \mu_0\|^2. \tag{165}$$

Finally, it follows that

$$\epsilon_g \underset{\lambda \leq 0}{\leq} \lim_{\lambda \to \infty} \epsilon_g = \inf_{b_1} \|\sigma_\star - b_1 \mu_0\|^2, \tag{166}$$

which recovers the upper bound (26).

E.2. Derivation of the lower bound (27)

We now turn to the lower bound (27). First observe that by the definitions of q_1, ψ, ρ (17), it follows from the Cauchy-Schwarz inequality that

$$\psi \leq \sqrt{q_1 \rho}. \tag{167}$$

Therefore, for all κ

$$q(\kappa) - \frac{\mu_1(\kappa)^2 \psi^2}{\rho} = \mu_1(\kappa)^2 \frac{\rho q_1 - \psi^2}{\rho} + \mu_2(\kappa)^2 q_2 \geq 0 \tag{168}$$

as a consequence, the test error (25) is lower bounded as

$$
\begin{aligned}
\epsilon_g = \mathbb{E}_\kappa &\left[\left(\sigma_\star(\kappa) - \mu_0(\kappa)m - \mu_1(\kappa)\kappa\zeta \right)^2 + q(\kappa) - \frac{\mu_1(\kappa)^2 \psi^2}{\rho} \right] \\
&\geq \mathbb{E}_\kappa \left[\left(\sigma_\star(\kappa) - \mu_0(\kappa)m - \mu_1(\kappa)\kappa\zeta \right)^2 \right] \\
&\geq \inf_{b_1, b_2} \mathbb{E}_\kappa \left[\left(\sigma_\star(\kappa) - \mu_0(\kappa)b_1 - \mu_1(\kappa)\kappa b_2 \right)^2 \right],
\end{aligned}
\tag{169}
$$

which recovers the lower bound (27). The functions corresponding to the upper (resp. lower) bound (see (26), (27)) are represented for a $\sigma_\star = \sin$ target in Fig. 4.

E.3. Derivation of bounds (28)

Since the bounds (26) and (27) hold for all learning rates $\tilde{\eta}$, the bounds (28) follow immediately from taking the minimum over $\tilde{\eta}$ on the right-hand side and left-hand side of (26) and (27). The functions $\{\mu_0(r), \tilde{\mu}_1(r)\}_{r \geq 0}$, which form the pool of functions which can be used to approximate the target function, are represented for $\sigma = \sin, \tanh$ in Fig. 9.

F. Spectrum of The Features Empirical Covariance Matrix

In this section, we derive a tight asymptotic expression for the spectrum of the empirical covariance of the features $\{\varphi(\boldsymbol{x}^\mu)\}_{\mu=1}^n$. Before starting the derivation, let us first remind that the *population* covariance is given by the conditional

Gaussian equivalence 3.2 as

$$\mathbb{E}_{\boldsymbol{x}}[\varphi(\boldsymbol{x})\varphi(\boldsymbol{x})^\top] = \mathbb{E}_\kappa\left[\left(\mu_0(\kappa)\mathbf{1}_p + \frac{1}{\sqrt{d}}\kappa\mu_1(\kappa)W\boldsymbol{v}\right)^{\otimes 2} + \mu_1(\kappa)^2 W\Pi^\perp W^\top + \mu_2(\kappa)^2\mathbb{I}_p\right] \tag{170}$$

where $\cdot^{\otimes 2}$ denotes the outer product of a vector with itself.

We now turn to the characterization of the empirical covariance. Let us denote by $\Psi \in \mathbb{R}^{n\times p}$ the matrix with rows $\{\varphi(\boldsymbol{x}^\mu)\}_{\mu=1}^n$, and the empirical covariance $\Sigma = 1/p\Psi^\top\Psi$. Note that up to zero eigenvalues, this matrix has the same spectrum as the conjugate kernel $1/p\Psi\Psi^\top$. In order to access the spectrum of Σ, we aim to compute the Stieltjes transform

$$s(z) = \mathbb{E}_{\{\boldsymbol{x}^\mu\}_{\mu=1}^n} \mathrm{Tr}\left[(\Sigma + z\mathbb{I}_p)^{-1}\right] \tag{171}$$

for some complex number z. This computation can be achieved again using the replica method, first observing the identity

$$s(z) = -2\partial_z\mathbb{E}_{\{\boldsymbol{x}^\mu\}_{\mu=1}^n} \frac{1}{2}\ln\det\left[2\pi(\Sigma + z\mathbb{I}_p)^{-1}\right] \equiv 2\partial_z\Phi(z). \tag{172}$$

We introduced the free entropy $\Phi(z)$, which can be computed as

$$\Phi(z) = \mathbb{E}_{\{\boldsymbol{x}^\mu\}_{\mu=1}^n} \ln \underbrace{\int d\boldsymbol{w}e^{-\frac{1}{2}\boldsymbol{w}^\top(\Sigma+z\mathbb{I}_p)\boldsymbol{w}}}_{\equiv Z(z)}. \tag{173}$$

Once more, the generative function $\mathbb{E}_{\{\boldsymbol{x}^\mu\}_{\mu=1}^n} \ln Z(z)$ can be evaluated using the replica method, leveraging the identity

$$\mathbb{E}_{\{\boldsymbol{x}^\mu\}_{\mu=1}^n} \ln Z(z) = \lim_{s\to 0}\frac{\ln\left(\mathbb{E}_v Z(z)^s\right) - 1}{s}. \tag{174}$$

The problem thus reduces to the computation of the replicated partition function $\mathbb{E}_{\boldsymbol{x}} Z(z)^s$

$$\mathbb{E}_{\{\boldsymbol{x}^\mu\}_{\mu=1}^n} Z(z)^s = \int\prod_{a=1}^s d\boldsymbol{w}_a e^{-\frac{z}{2}\sum_a\|\boldsymbol{w}_a\|^2}\prod_\mu\left[\mathbb{E}_{\boldsymbol{x}}e^{-\frac{1}{2}\sum_a(\boldsymbol{w}_a^\top\varphi(\boldsymbol{x})/\sqrt{p})^2}\right]$$

$$= \int\prod_{a=1}^s d\boldsymbol{w}_a e^{-\frac{z}{2}\sum_a\|\boldsymbol{w}_a\|^2}\left[\mathbb{E}_\kappa\mathbb{E}_\lambda e^{-\frac{1}{2}\sum_a(\mu_0(\kappa)m_a+\mu_1(\kappa)\kappa\zeta_a+\lambda_a)^2}\right]^n. \tag{175}$$

Like in Appendix C, we introduced the order parameters

$$m_a \equiv \frac{\boldsymbol{w}_a^\top\mathbf{1}_p}{\sqrt{p}}, \qquad\qquad \zeta_a \equiv \frac{\boldsymbol{w}_a^\top W\boldsymbol{v}}{\sqrt{dp}}. \tag{176}$$

We also introduced the local fields $\{\lambda_a\}_a$, which are Gaussian with zero mean and joint statistics

$$\mathbb{E}[\lambda_a\lambda_b] = \mu_1(\kappa)^2 q_{ab}^1 + \mu_2(\kappa)q_{ab}^2, \tag{177}$$

denoting the summary statistics

$$q_{ab}^1 \equiv \frac{\boldsymbol{w}_a^\top\Omega\boldsymbol{w}}{p}, \qquad\qquad q_{ab}^2 \equiv \frac{\boldsymbol{w}_a^\top\boldsymbol{w}}{p}, \tag{178}$$

where we remind $\Omega = W\Pi^\perp W^\top$. The replicated partition function can thus be written as

$$Z^s = \int \prod_{a=1}^{s} dm_a d\hat{m}_a d\zeta_a d\hat{\zeta}_a \prod_{a,b} dq_{ab}^1 d\hat{q}_{ab}^1 dq_{ab}^2 d\hat{q}_{ab}^2 \underbrace{e^{-\sqrt{p}\sum_r m_a \hat{m}_a - \sqrt{dp}\sum_a(\zeta_a \hat{\zeta}_a) - p\sum_{a\leq b}(q_{ab}^1 \hat{q}_{ab}^1 + q_{ab}^2 \hat{q}_{ab}^2)}}_{e^{sp\Psi_t}}$$

$$\underbrace{\int \prod_a d\boldsymbol{w}_a e^{-\frac{z}{2}\sum_a \|\boldsymbol{w}_a\|^2} e^{+\sum_{a=1}^{s} \boldsymbol{w}_a^\top(\hat{m}_a \mathbf{1}_p + \hat{\zeta}_a W\boldsymbol{v}) + \sum_{a\leq b}(\hat{q}_{ab}^1 \boldsymbol{w}_a^\top \Omega \boldsymbol{w}_b + \hat{q}_{ab}^2 \boldsymbol{w}_a^\top \boldsymbol{w}_b)}}_{e^{sp\Psi_w}}$$

$$\underbrace{\left[\mathbb{E}_\kappa\left(\mathbb{E}_{\{\lambda_a\}_{a=1}^s} e^{-\frac{1}{2}\sum_{a=1}^{s}(\mu_0(\kappa)m_a + \mu_1(\kappa)\kappa\zeta_a + \lambda_a)^2}\right)\right]^n}_{e^{\alpha/\beta sp\Psi_y}}, \tag{179}$$

an expression bearing great similarity with that reached in Appendix C. In fact, formally, this expression is equivalent to the finite-temperature ($\omega = 1$) version of the one discussed in Appendix C, in the particular case of a regularization $\lambda = z$, and with a loss function $\ell(y, z) = \ell(z) = z^2$. Identical computational steps can then be followed to reach a tight characterization of the free entropy. Again, we assume a RS ansatz

$$\begin{aligned}
&\forall 1 \leq a, b \leq s, \quad q_{ab}^1 = \delta_{ab}V_1 \\
&\forall 1 \leq a, b \leq s, \quad q_{ab}^2 = \delta_{ab}V_2 \\
&\forall 1 \leq a \leq s, \quad \zeta_a = \zeta \\
&\forall 1 \leq a \leq s, \quad m_a = m \\
&\forall 1 \leq a, b \leq s, \quad \hat{q}_{ab}^1 = -\delta_{ab}\hat{V}_1/2 \\
&\forall 1 \leq a, b \leq s, \quad \hat{q}_{ab}^2 = -\delta_{ab}\hat{V}_2/2 \\
&\forall 1 \leq a \leq s, \quad \hat{\zeta}_a = \hat{\zeta} \\
&\forall 1 \leq a \leq s, \quad \hat{m}_a = \hat{m}.
\end{aligned} \tag{180} \tag{181} \tag{182}$$

Following identical steps as Appendix C, one finally reaches

$$\begin{aligned}
\Phi(z) = \underset{V_1, V_2, m, \zeta, \hat{V}_1, \hat{V}_2, \hat{m}, \hat{\zeta}}{\text{extr}} &\frac{\hat{V}_1 V_1 + \hat{V}_2 V_2}{2} - \frac{m\hat{m}}{\sqrt{p}} - \frac{\zeta\hat{\zeta}}{\sqrt{\beta}} - \frac{1}{2}\ln\det\left[z\mathbb{I}_p + \hat{V}_1\Omega + \hat{V}_2\mathbb{I}_p\right] \\
&+ \frac{1}{2}\text{Tr}\left[(z\mathbb{I}_p + \hat{V}_1\Omega + \hat{V}_2\mathbb{I}_p)^{-1}\left(\hat{m}^2 \mathbf{1}_p \mathbf{1}_p^\top + \hat{\zeta}^2 W\boldsymbol{v}\boldsymbol{v}^\top W^\top\right)\right] \\
&- \frac{\alpha}{\beta}\frac{1}{2}\mathbb{E}_\kappa\left[\frac{(\mu_0(\kappa)m + \mu_1(\kappa)\kappa\zeta)^2}{1 + V(\kappa)}\right] - \frac{\alpha}{\beta}\frac{1}{2}\mathbb{E}_\kappa\ln(1 + V(\kappa))
\end{aligned} \tag{183}$$

Expunding the extremization conditions with respect to $m, \hat{m}, \zeta, \hat{\zeta}$ imposes

$$\begin{cases}
0 = \hat{m}\,\text{Tr}\left[(z\mathbb{I}_p + \hat{V}_1\Omega + \hat{V}_2\mathbb{I}_p)^{-1}\mathbf{1}_p \mathbf{1}_p^\top\right] \\
\zeta = \hat{\zeta}\,\text{Tr}\left[(z\mathbb{I}_p + \hat{V}_1\Omega + \hat{V}_2\mathbb{I}_p)^{-1}W\boldsymbol{v}\boldsymbol{v}^\top W^\top\right] \\
m = -\frac{\mathbb{E}_\kappa\left[\frac{\mu_1(\kappa)\kappa\mu_0(\kappa)}{1+V(\kappa)}\right]}{\mathbb{E}_\kappa\left[\frac{\mu_0(\kappa)^2}{1+V(\kappa)}\right]}\zeta \\
\hat{\zeta} = \mathbb{E}_\kappa\left[\frac{(\mu_0(\kappa)m + \mu_1(\kappa)\kappa\zeta)\mu_1(\kappa)\kappa}{1+V(\kappa)}\right]
\end{cases} \tag{184}$$
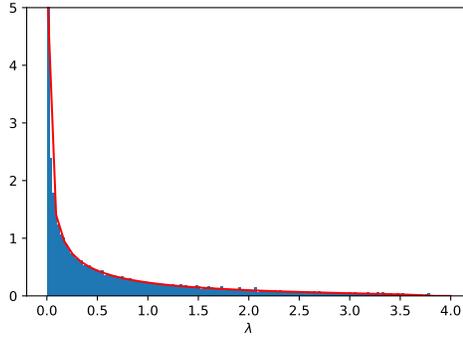
Figure 10. Spectrum of the empirical features covariance for $\alpha = 1, \beta = 1, \sigma = \text{sign}$. Blue: numerical estimation in $d = 2000$. Red: theoretical characterization (187). There is a single spike $\lambda_{\text{spk.}} = \Theta_d(d)$ in addition to the bulk, which is not represented in this figure.

Combining the last two equations imposes

$$\hat{\zeta} = \mathbb{E}_\kappa \left[\frac{\left(-\mu_0(\kappa) \frac{\mathbb{E}_\kappa\left[\frac{\mu_1(\kappa)\kappa\mu_0(\kappa)}{1+V(\kappa)} \right]}{\mathbb{E}_\kappa\left[\frac{\mu_0(\kappa)^2}{1+V(\kappa)} \right]} + \mu_1(\kappa)\kappa \right) \mu_1(\kappa)\kappa}{1 + V(\kappa)} \right] \zeta \tag{185}$$

from which it generically follows that $\zeta = 0$, which also implies $m = \hat{m} = \hat{\zeta} = 0$. Finally, the extremization conditions with respect to $V_{1,2}, \hat{V}_{1,2}$ yield the simple equations

$$\begin{cases} V_2 = \text{Tr}\left[(z\mathbb{I}_p + \hat{V}_1\Omega + \hat{V}_2\mathbb{I}_p)^{-1} \right] \\ V_1 = Tr[(z\mathbb{I}_p + \hat{V}_1\Omega + \hat{V}_2\mathbb{I}_p)^{-1}\Omega] \\ \hat{V}_1 = \frac{\alpha}{\beta}\mathbb{E}_\kappa\left[\frac{\mu_1(\kappa)^2}{1+V(\kappa)} \right] \\ \hat{V}_2 = \frac{\alpha}{\beta}\mathbb{E}_\kappa\left[\frac{\mu_2(\kappa)^2}{1+V(\kappa)} \right] \end{cases} \tag{186}$$

Note how these equations correspond exactly to the ones of Result 3.3, replacing the regularization λ by the Stieltjes argument z. Finally, we observe that $\partial_z\Phi(z) = 1/2V_2$, which implies that the Stieltjes transform $s(z)$ is compactly given by

$$s(z) = V_2, \tag{187}$$

with V_2 a solution of the system of equations above. The spectrum can then be extracted from the Stieltjes transform, and matches well with numerical simulations on sRF features, see Fig. 10.

Spike The previous study allowed to characterize the asymptotic bulk eigenvalue distribution of the empirical features covariance. Note that there is a single spike $\lambda_{\text{spk.}} = \Theta_d(d)$ not captured by the analysis, which relies on a Stieltjes argument $z = \Theta_d(1)$. It is however reasonable to expect this spike eigenvalue to coincide with that of the population covariance. This eigenvalue should then be asymptotically equal to

$$\frac{\lambda_{\text{spk.}}}{d} = \alpha\mathbb{E}_{\kappa\sim\mathcal{N}(0,1)}[c_0(\kappa)^2]. \tag{188}$$