CHALMERS
UNIVERSITY OF TECHNOLOGY

Precise Asymptotic Analysis of Deep Random Feature Models

Precise Asymptotic Analysis of Deep Random Feature Models

David Bosch DAVIDBOS@CHALMERS.SE
Ashkan Panahi ASHKAN.PANAHI@CHALMERS.SE
Department of Data Science and AI, Computer Science and Engineering, Chalmers University of Technology

Babak Hassibi HASSIBI@CALTECH.EDU
Department of Electrical Engineering, California Institute of Technology

Editors: Gergely Neu and Lorenzo Rosasco

Abstract

We provide exact asymptotic expressions for the performance of regression by an $L-$layer deep random feature (RF) model, where the input is mapped through multiple random embedding and non-linear activation functions. For this purpose, we establish two key steps: First, we prove a novel universality result for RF models and deterministic data, by which we demonstrate that a deep random feature model is equivalent to a deep linear Gaussian model that matches it in the first and second moments, at each layer. Second, we make use of the convex Gaussian Min-Max theorem multiple times to obtain the exact behavior of deep RF models. We further characterize the variation of the eigendistribution in different layers of the equivalent Gaussian model, demonstrating that depth has a tangible effect on model performance despite the fact that only the last layer of the model is being trained.

Keywords: Asymptotic Analysis, Universality, Random Features Model, Convex Gaussian Min Max Theorem, Learning Curves

1. Introduction

Recent experimental and theoretical results (Zhang et al., 2021; Belkin et al., 2019) have demonstrated that the classical understanding of overparameterized machine learning (ML) models requires further examination. One model that has been studied extensively is the random features (RF) model (Rahimi and Recht, 2007), which is closely related to overparameterized neural networks (Daniely et al., 2016; Daniely, 2017; Jacot et al., 2018; Liu et al., 2021; Bach, 2017). In this paper, we examine an extension of the RF model, which we call the deep RF (DRF) model, being equivalent to a deep NN, but only trained in the output layer. We consider the asymptotic regime, where the number of data points, model parameters, and input dimension grow infinite at constant ratio (Belkin et al., 2020; Hastie et al., 2019; Bartlett et al., 2020b,a) and give exact expressions that characterize the deep RF model in terms of training and generalization error.

Our analysis consists of two key steps. First, we prove universality, i.e. we demonstrate that the DRF model is asymptotically equivalent to a deep Gaussian surrogate model, matching the original model in the first and second moments, at each layer (Panahi and Hassibi, 2017; Oymak and Tropp, 2018). Universality for the 1-layer RF model has previously been proven, e.g. in (Hu and Lu, 2022). We make use of a different proof technique to extend these results to arbitrary many layers and introduce a new Gaussian surrogate model for DRF. This universality result alleviates the general difficulty of analyzing RF or DRF models, as the non Gaussian features are in general not amenable to stardard analysis techniques such as comparison theorem (Gordon, 1985; Thrampoulidis et al., 2014), Gaussian widths (Chandrasekaran et al., 2012) or replica methods (Mézard et al., 1987).

Having established universality, we then make use of the Convex Gaussian Min Max Theorem (CGMT) (Thrampoulidis et al., 2014) to study DRFs. This theorem allows us to consider an alternative optimization problem with the same asymptotic statistics, and is a popular tool in the analysis of the asymptotic regime (Bosch et al., 2021; Chang et al., 2020; Dhifallah and Lu, 2020; Thrampoulidis et al., 2015; Loureiro et al., 2021b; Bosch et al., 2022). We make use of a recursive application of the CGMT (Bosch et al., 2022) to obtain asymptotic expressions for square loss functions with arbitrary convex regularization for L-layer DRF models.

2. Related Works

The random features (RF) (Rahimi and Recht, 2007) model has been extensively examined in the asymptotic regime, under a multitude of conditions. For an incomplete list see (Hastie et al., 2019; Mei and Montanari, 2019; Montanari et al., 2019; Goldt et al., 2022, 2020; Gerace et al., 2020; Dhifallah and Lu, 2020; Ghorbani et al., 2021; Bosch et al., 2022). In the case of ridge regression (Louart et al., 2018; Mei and Montanari, 2019) exact expression for the training and generalization error can be established. In other cases, exact analysis is difficult. It was observed by many authors (Mei and Montanari, 2019; Hastie et al., 2019; Goldt et al., 2020; Gerace et al., 2020; Goldt et al., 2022) that a Gaussian surrogate model that matched the first and second moments had asymptotically equivalent statistics. A concrete proof of RF universality is given in Hu and Lu (2022). We utilize Lindeberg's approach (Lindeberg, 1922) to demonstrate universality of DRF. This approach has been used to prove universality results in many other optimization problems (Korada and Montanari, 2011; Panahi and Hassibi, 2017; Montanari and Nguyen, 2017; Oymak and Tropp, 2018; Abbasi et al., 2019). Hu and Lu (2022) prove a central limit theorem between random features and their Gaussian equivalent features as a key step in demonstrating universality. We make use of a different proof technique, by instead considering the problem in a dual space, where we may directly bound the difference between the leave one out iterates.

Beside RF, universality has been demonstrated for many other models (Goldt et al., 2022; Seddik et al., 2020; Dhifallah and Lu, 2021; Loureiro et al., 2021a; Gerace et al., 2022). Recently, Montanari and Saeed (2022) gave a proof for the universality of empirical risk minimization for not necessarily convex loss and regularization functions. Their result also assume that a central limit theorem similar to (Hu and Lu, 2022) holds.

Subject to Gaussian features, the CGMT (Gordon, 1985; Thrampoulidis et al., 2014) is a powerful tool in the determination of the asymptotic performance (Loureiro et al., 2021b; Dhifallah and Lu, 2020; Thrampoulidis et al., 2015; Chang et al., 2020; Bosch et al., 2021, 2022). The CGMT determines an alternative, asymptotically equivalent optimization problem in statistical properties. In the case of correlated features, such as in the RF or DRF model, the alternative optimization still remains intractable. This issue is resolved in (Bosch et al., 2022) by applying the CGMT twice. Relying on the particular structure of the DRF covariance matrices, we extend the method of (Bosch et al., 2022) where the CGMT is applied recursively to determine a nested scalar optimization that is asymptotically equivalent to the DRF model.

The covariance matrices for the Gaussian surrogate model that we obtain are similar in structure to the kernel matrices given in Lee et al. (2017). The authors demonstrate an exact equivalence between an infinitely wide deep NN and a Gaussian Process with covariance kernels that are recursively defined in a similar manner to the ones discussed in this paper. However, (Lee et al., 2017) consider networks of fixed size but infinite width, while we consider the asymptotic regime,

where the number of data points and the input dimensions grow as well, hence maintaining a relatively narrower network. Furthermore, covariance matrices of the DRF model can be expressed recursively, with the recursion depth determined by the number of layers. (Fan and Wang, 2020) similarly analyze a the recursive structure of the covariance matrix for the conjugate kernel and Neural tangent kernel by mean of free probability theory. The results for the deep random features case was extended by (Schröder et al., 2023) around the same time as the initial submission of this paper. We similarly use a free probability argument to analyze our obtained recursion. However, we compute recursion for the closed form asymptotic equivalent expression for the population covariance instead of dealing with the distribution of the population covariance directly as in (Fan and Wang, 2020; Schröder et al., 2023).

2.1. Paper Outline

In section 3, we introduce the DRF problem and its Gaussian surrogate, and express the necessary assumptions for our results to hold. In section 4, we prove the main universality theorem of this paper. Our proof takes two steps, first proving universality of a single layer, and subsequently using an inductive argument to extend this result to a full DRF problem. In section 5, we give an alterative scalar optimization problem derived by means of the CGMT, that is asymptotically equivalent to the DRF problem subject to square loss and arbitrary, strongly convex regularization. We demonstrate experimentally the veracity of the determined expressions.

3. Setup and Assumptions

3.1. Random Feature Model and Preliminaries

We consider a supervised learning setup with a dataset $\mathcal{D} = \{(\mathbf{x}_k, y_k) \in \mathbb{R}^d \times \mathbb{R}\}_{k=1}^n$. To find a relationship between the data points $\mathbf{x}_k$ and the labels y_k, we consider a function of the following form

$$Y_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{1}{\sqrt{p}} \boldsymbol{\theta}^T \mathcal{F}(\mathbf{x}), \qquad \boldsymbol{\theta} \in \mathbb{R}^p, \tag{1}$$

where $\mathcal{F} : \mathbb{R}^d \to \mathbb{R}^p$ is a given mapping of the data, called a *feature map*. We note that $Y_{\boldsymbol{\theta}}$ is dependent upon the choice of the vector $\boldsymbol{\theta}$ by a linear relation. This shows the main advantage of (1): while $Y_{\boldsymbol{\theta}}$ can represent nonlinear functions, selecting $\boldsymbol{\theta}$ amounts to a linear regression task. To find the optimal value of $\boldsymbol{\theta}$, denote $\mathbf{f}_i = \mathcal{F}(\mathbf{x}_i)$ and take $\boldsymbol{F}$ as a matrix with $\mathbf{f}_i$ as columns. We consider the empirical risk minimization framework and the following optimization problem:

$$\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{F}) = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \ell\left(\frac{1}{\sqrt{p}} \boldsymbol{\theta}^T \mathbf{f}_i, y_i\right) + R(\boldsymbol{\theta}). \tag{2}$$

Here, $\ell(\cdot, \cdot)$ is a loss function, and $R(\boldsymbol{\theta})$ is a regularization function. To measure the performance of $\hat{\boldsymbol{\theta}}$ we make use of two common metric for supervised learning, that being the training error $\mathcal{E}_{train}(\boldsymbol{F})$, i.e the optimal value in (2), and the generalization error

$$\mathcal{E}_{gen}(\boldsymbol{F}) = \mathbb{E}\left[\ell\left(\frac{1}{\sqrt{p}} \hat{\boldsymbol{\theta}}^T \mathcal{F}(\mathbf{x}_{new}), y_{new}\right)\right], \tag{3}$$

where the expectation is taken over $(\mathbf{x}_{new}, y_{new})$, a new datapoint drawn from the same distribution as the dataset $\mathcal{D}$.

The main purpose of this paper is to obtain exact asymptotic expressions for the supervised learning metrics $\mathcal{E}_{train}, \mathcal{E}_{gen}$ and other properties of $\hat{\boldsymbol{\theta}}$, when the feature map $\mathcal{F}$ is a deep random feature, generalizing the *random features maps* (Rahimi and Recht, 2007). To define the deep features, we remind the (shallow) random features are given by $\phi(\mathbf{x}, \mathbf{w}_j) := \sigma(\mathbf{w}_j^T \mathbf{x})$, for $j = 1, 2, \ldots, p$, where σ is an activation function, and $\mathbf{w}_j \sim \mathcal{N}(0, \frac{1}{d}\mathbf{I}_d)$ are a set of random weights. In vector form, we express these relations as $\boldsymbol{\phi}(\mathbf{x}, \mathbf{W}) = (\sigma(\mathbf{w}_j^T \mathbf{x}))_{j=1}^p$, where the matrix $\mathbf{W}$ has rows $\mathbf{w}_j$. Then, the deep random features are given through the following recursion: For $p_0 = d, p_1, p_2, \ldots, p_L \in \mathbb{N}$, we define the matrices $\mathbf{W}^{(l)} \in \mathbb{R}^{p_l \times p_{l-1}}$ for $l = 1, \ldots, L$, each having independent rows $\mathbf{w}_j^{(l)} \sim \mathcal{N}(0, \frac{1}{p_{l-1}}\mathbf{I})$. Letting $\mathbf{x}^{(0)} := \mathbf{x}$ we define

$$\mathbf{x}^{(l)} = \boldsymbol{\phi}(\mathbf{x}^{(l-1)}, \mathbf{W}^{(l)}) = (\sigma(\mathbf{w}_j^{(l)T} \mathbf{x}^{(l-1)}))_{j=1}^{p_l}, \qquad l = 1, \ldots, L, \tag{4}$$

3.2. Necessary Assumptions

Our results rely on the following assumptions:

A1 For some universal positive constants μ, M, the regularization function R is μ-strongly convex and M-smooth with M-bounded third derivative in tensor (operator) norm. Moreover $\|\nabla R(\mathbf{0})\| \leq C$.

A2 ℓ is a $\frac{1}{Cn}-$ strongly convex function in the first argument and its third derivative with respect to the first argument is bounded by Cn for some constant C. Moreover, there exists a vector $\boldsymbol{\alpha} = (\alpha_k)$ called *isolated predictions* satisfying: $\alpha_k \in \arg\min_\alpha \ell(\alpha, y_k)$, and $\|\boldsymbol{\alpha}\|_2 \leq C\sqrt{n}$ for a fixed constant C.

A3 The activation function σ is an odd function applied element wise, with bounded derivatives. Furthermore, let g_1, g_2 be Gaussian variables distributed as

$$\begin{bmatrix} g_1 \\ g_2 \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \alpha_1 & \rho \\ \rho & \alpha_2 \end{bmatrix}\right). \tag{5}$$

Let the functions $\eta_1(\alpha_1, \alpha_2, \rho) = \mathbb{E}[\sigma(g_1)\sigma(g_2)]$ and $\eta_2(\alpha_1) = \mathbb{E}[\sigma^2(g_1)]$. Then η_1, η_2 should be thrice differentiable at $\alpha_1 = \alpha_2 = 1$ and $\rho = 0$

A4 The dimensions of the number of data points n, the size of the input $p_0 = d$ and the size of subsequent layers p_l, where $l = 1, \ldots, L$ all grow to infinity at fixed ratios. We denote this by $n \sim p_0 \sim \cdots \sim p_L$, where $a \sim b$ is defined to mean that $\frac{a}{b} \xrightarrow[a,b\to\infty]{} C$ for some constant C.

A5 For each layer, l, the weight matrix $\mathbf{W}^{(l)} \in \mathbb{R}^{p_l \times p_{l-1}} = [\mathbf{w}_1^{(l)}\ \mathbf{w}_2^{(l)}\ \cdots\ \mathbf{w}_{p_l}^{(l)}]^T$ are independent Gaussian variables $\mathbf{w}_i^{(l)} \overset{i.i.d}{\sim} \mathcal{N}(0, \frac{1}{p_{l-1}}\mathbf{I}_{p_{l-1}})$ for $1 \leq i \leq p_l$. Furthermore, $\mathbf{W}^{(l)}$ are independent of the input variables $\mathbf{x}$.

Remark 1 *For assumption 2, we note that the strong convexity assumption on the loss function becomes less restrictive as n grows. In the asymptotic limit, the strong convexity is no longer a*

significant requirement. Furthermore, the isolated prediction vectors exist and the condition is satisfied immediately if the loss function is minimized at the labels, i.e. it is minimized at the point $\ell(y_k, y_k) < \infty$.

Remark 2 *For assumption 3, we note that the condition holds for the majority of activation function used in practice including* $\tanh$ *and the error function. Furthermore, if oddness is dropped, the assumption on the functions* η_1 *and* η_2 *are additionally satisfied for functions like ReLU, sigmoids, and Gaussian activations. However, we require oddness.*

Finally we impose a condition upon the input vectors $\mathbf{x}_i$:

Definition 3 *Let* $d \sim n$*, we call a set* $\{\mathbf{x}_k \in \mathbb{R}^d\}_{k=1}^n$ *regular if*

1. *Letting* $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_n]$*, there is a constant* $c < \infty$ *such that* $\frac{1}{\sqrt{n}} \|\mathbf{X}\|_{\text{op}} < c$
2. *It holds that*

$$\max_{i,j} \left| \frac{1}{d} \mathbf{x}_i^T \mathbf{x}_j - \delta_{ij} \right| \leq \frac{\text{polylog } n}{\sqrt{n}}, \tag{6}$$

where δ_{ij} *is the Kronecker delta.*

Note that the first condition for regularity is trivially satisfied for finite n, however the condition must also hold for a fixed c in the asymptotic limit. Further, note that regularity is exhibited by $\mathbf{x}$ being Gaussian with high probability.

4. Universality

In the case of a single layer, it has been proven (Hu and Lu, 2022) that the following Gaussian feature map has asymptotically equivalent statistics to the random features given in section 3

$$\tilde{\phi}(\mathbf{x}, \mathbf{W}) = \rho_1 \mathbf{W}\mathbf{x} + \rho_2 \mathbf{g}, \tag{7}$$

where $\mathbf{g}$ is a standard normal vector, and ρ_1, ρ_2 are constants depending only on the activation function, given by $\rho_1 = \mathbb{E}[\sigma'(z)],\ \rho_2 = \sqrt{\mathbb{E}[\sigma^2(z)] - \rho_0^2 + \rho_1^2},\ z \sim \mathcal{N}(0,1)$. Similarly we define a *deep Gaussian equivalent feature map*, recursively. We define $\boldsymbol{\gamma}^{(0)} = \mathbf{x}$, and then define

$$\boldsymbol{\gamma}^{(l)} = \tilde{\phi}_l(\boldsymbol{\gamma}^{(l-1)}, \mathbf{W}^{(l)}) := \rho_{1,l} \mathbf{W}^{(l)} \boldsymbol{\gamma}^{(l-1)} + \rho_{2,l} \mathbf{g}^{(l)}, \qquad l = 1, \ldots, L, \tag{8}$$

where $\mathbf{g}^{(l)}$ is an independent standard normal vector of dimension p_l, and the constants $\rho_{1,l}, \rho_{2,l}$ are recursively defined as $\rho_{1,l} = \mathbb{E}[\sigma'(\alpha_{l-1} z)],\ \rho_{2,l} = \sqrt{\mathbb{E}[\sigma^2(\alpha_{l-1} z)] - \alpha_{l-1}^2 \rho_{1,l}^2},\ z \sim \mathcal{N}(0,1)$. Here, α_l are constants given by the following recursive definition: $\alpha_0 = 1,\ \alpha_l = \sqrt{\rho_{1,l}^2 \alpha_{l-1}^2 + \rho_{2,l}^2}$. Now, we consider the following two feature mappings for an input vector $\mathbf{x}^{(0)}$:

$$\mathcal{F}(\mathbf{x}^{(0)}) = \mathbf{x}^{(l)}, \qquad \mathcal{G}(\mathbf{x}^{(0)}) = \boldsymbol{\gamma}^{(l)}, \tag{9}$$

where $\mathbf{x}^{(l)}$ and $\boldsymbol{\gamma}^{(l)}$ are given in (4) and (8), respectively.

4.1. Revisiting Universality of a Single Layer

The proof of universality of deep random features is a specific application of a universality theorem for a single layer, which we derive in this section. This result is more general than the previous studies such as Hu and Lu (2022). In the subsequent section, we shall demonstrate how the universality of deep random features follows from these results.

Let $\phi_j : \mathbb{R}^d \times \Omega_j \to \mathbb{R}$ for $j = 1, 2, \ldots, p$ be a random feature map, where Ω_j is a sample space equipped with an arbitrary probability measure, such that $\phi_j(\cdot, \omega)$ for any $\omega \in \Omega_j$ is a particular realization of the feature map. Let $\Omega = \Omega_1 \times \Omega_2 \times \cdots \Omega_p$ be a product space equipped with the product measure, and let $\boldsymbol{\phi} : \mathbb{R}^d \times \Omega \to \mathbb{R}^p$ represent the vector of random features, such that $\boldsymbol{\phi}(\mathbf{x}, \boldsymbol{\omega}) = (\phi_j(\mathbf{x}, \omega_j))_j$, where $\boldsymbol{\omega} = (\omega_j) \in \Omega$ is a realization.

Next we consider a $m \times p$ matrix $\mathbf{D}$ with columns $\mathbf{d}_j \in \mathbb{R}^m$, which we call a *synthesis dictionary*. We define the re-represented random feature vectors $\mathbf{f} : \mathbb{R}^d \times \Omega \to \mathbb{R}^m$ given by

$$\mathbf{f}(\mathbf{x}, \boldsymbol{\omega}) := \sum_{j=1}^{p} \mathbf{d}_j \phi_j(\mathbf{x}, \omega_j) = \mathbf{D}\boldsymbol{\phi}(\mathbf{x}, \omega). \tag{10}$$

We note that if $m = p$ we can choose $\mathbf{D} = \mathbf{I}_p$ and retain the original set of random features. However, re-representing the features is necessary for the proof of the deep random features case. We will drop the argument $\boldsymbol{\omega}$ when there is no risk of confusion, and denote $\boldsymbol{\phi}(\mathbf{x})$, $\mathbf{f}(\mathbf{x})$ as the random features and their re-representation. Similarly let $\boldsymbol{\phi}_k = \boldsymbol{\phi}(\mathbf{x}_k)$ and $\mathbf{f}_k = \mathbf{f}(\mathbf{x}_k)$ for $k = 1, \ldots, n$ which are random vectors. Finally, let $\boldsymbol{\Phi}$ and $\boldsymbol{F}$ be the matrices with $(\boldsymbol{\phi}_k)$, $(\mathbf{f}_k)$ as columns. We assume that the random features are centered:

$$\mathbb{E}_{\boldsymbol{\omega}}[\boldsymbol{\phi}(\mathbf{x}_k, \boldsymbol{\omega})] = \mathbf{0} \qquad k = 1, 2, \ldots, n. \tag{11}$$

We further define the data kernel matrices $\boldsymbol{K}_j = (K_{j,kl})_{kl}$ where $\boldsymbol{K}_j$ is the covariance matrix of the jth row of $\boldsymbol{\Phi}$, given by

$$K_{j,kl} = \mathbb{E}_{\omega_j}[\phi(\mathbf{x}_k, \omega_j)\phi(\mathbf{x}_l, \omega_j)]. \tag{12}$$

Next, we introduce a $p \times n$ Gaussian matrix $\boldsymbol{\Gamma}$ with independent rows, and where the jth row is distributed by $\mathcal{N}(0, \boldsymbol{K}'_j)$. We note that if $\boldsymbol{K}_j = \boldsymbol{K}'_j$ that $\boldsymbol{\Phi}$ and $\boldsymbol{\Gamma}$ have the same first and second moments amongst their elements. We then define $\mathbf{G} = \mathbf{D}\boldsymbol{\Gamma}$ and let $\mathbf{g}_k$ be the kth column of $\mathbf{G}$.

Before stating the main theorem for this section we state the conditions on the dataset and matrices $\boldsymbol{K}_j$ and $\mathbf{D}$ that must hold. We shall show in the next section that these conditions are satisfied in the case of deep random features. We remind the reader of the definition of a sub-Gaussian vector:

Definition 4 *We say that a random vector $\mathbf{u} = (u_k) \in \mathbb{R}^n$ is $\tau-$sub-Gaussian if for any unit vector $\mathbf{a} = (a_k) \in \mathbb{R}^n$ the variable $A = \mathbf{a}^T\mathbf{u}$ is sub-Gaussian with parameter τ, i.e. $\mathbb{E}\left[e^{\lambda A}\right] \leq e^{\frac{\tau^2\lambda^2}{2}}$ for all $\lambda \in \mathbb{R}$.*

We state the following requisite conditions:

- B1 There exists a positive constant C such that for all j, it holds that $\|\boldsymbol{K}_j\|_{\text{op}} \leq C$ and the jth random feature vector $\boldsymbol{\phi}^j = \{\phi(\mathbf{x}_k, \omega_j)\}_k$ is C-sub-Gaussian
- B2 There exists a positive constant C such that $\|\mathbf{D}\|_{\text{op}} \leq C$.

These assumptions must hold for all values of n, d, p, m and must continue to hold when they grow asymptotically. Subject to these conditions we state the following theorem that demonstrates universality.

Theorem 5 *Suppose that assumptions A1, A2, B1 and B2 hold, and that $n \sim p \sim m$. Then,*

1. *For any real function ψ with bounded first, second, and third derivatives, there exists a constant $c < \infty$ such that*

$$|\mathbb{E}\psi(\mathcal{E}_{train}(\boldsymbol{F})) - \mathbb{E}\psi(\mathcal{E}_{train}(\mathbf{G}))| \leq \frac{c}{n}\sum_{j=1}^{p}\left\|\boldsymbol{K}_j - \boldsymbol{K}_j'\right\|_{\text{op}} + \frac{c}{\sqrt{n}}. \tag{13}$$

2. *Let $\hat{\boldsymbol{\theta}}_F$ and $\hat{\boldsymbol{\theta}}_G$ be the optimal points for the optimization (2) for $\boldsymbol{F}$ and $\mathbf{G}$ respectively. Take any bounded function $h : \mathbb{R}^m \to \mathbb{R}$ with bounded $\nabla h(\mathbf{0})$, second and third derivatives (in tensor norm), where the bounds are constant in n, p, m. There exists a constant $c < \infty$ such that*

$$\left|\mathbb{E}h\left(\hat{\boldsymbol{\theta}}_F\right) - \mathbb{E}h\left(\hat{\boldsymbol{\theta}}_G\right)\right| \leq \frac{c}{n}\sum_{j=1}^{p}\left\|\boldsymbol{K}_j - \boldsymbol{K}_j'\right\|_{\text{op}} + \frac{c}{\sqrt{n}}. \tag{14}$$

4.1.1. PROOF SKETCH

The proof is based on an application of Lindeberg's argument with respect to the random features $\mathbf{f}$ in a dual space. We consider the optimization problem given in (2) for some generic map $\mathbf{Z}$ and note that by means of a splitting argument it may be expressed as

$$\begin{aligned}
\mathcal{E}_{train}(\mathbf{Z}) &= \min_{\boldsymbol{\theta}\in\mathbb{R}^p}\frac{1}{n}\sum_{k=1}^{n}\ell(\mathbf{z}_k^T\boldsymbol{\theta}, y_k) + R(\boldsymbol{\theta}) \\
&= \min_{\boldsymbol{\theta}\in\mathbb{R}^p,\boldsymbol{\alpha}\in\mathbb{R}^n}\max_{\mathbf{d}\in\mathbb{R}^n}\frac{1}{n}\left(\sum_{k=1}^{n}\ell(\alpha_k, y_k) + d_k(\alpha_k - \mathbf{z}_k^T\boldsymbol{\theta})\right) + R(\boldsymbol{\theta}) \\
&= -\min_{\mathbf{d}\in\mathbb{R}^n}\frac{1}{n}\sum_{k=1}^{n}\ell^*(-d_k, y_k) + R^*\left(\frac{1}{n}\mathbf{Z}\mathbf{d}\right)
\end{aligned} \tag{15}$$

where ℓ^* and R^* are the Legendre transforms of ℓ and R respectively. We note that by assumption A1 that R^* is $\frac{1}{M}-$ strongly convex and $\frac{1}{\mu}$-smooth. We then proceed in defining a series of $\mathbf{Z}_r$ such that $\mathbf{Z}_0 = \boldsymbol{\Phi}$ and $\mathbf{Z}_p = \boldsymbol{\Gamma}$. We show that the difference of the optimal value in the dual space between $\psi(\mathcal{E}_{train}(\mathbf{Z}_r))$ and $\psi(\mathcal{E}_{train}(\mathbf{Z}_{r+1}))$ is bounded by the sum of a $O(\frac{1}{n^{3/2}})$ term and the difference in operator norm between $\frac{c}{n}\left\|\boldsymbol{K}_r - \boldsymbol{K}_r'\right\|_{op}$, which allows us to bound the total difference as in the given result.

For part two, we note that $R_\epsilon(\boldsymbol{\theta}) = R(\boldsymbol{\theta}) \pm \epsilon h(\boldsymbol{\theta})$ remains strongly convex for sufficiently small values of $\epsilon > 0$. As such, part 1 of the theorem holds for these cases. By bounding the difference in the values of $\mathcal{E}_{train}$ at $\epsilon > 0$ and at $\epsilon = 0$ the bound on $h(\boldsymbol{\theta})$ may be obtained. The proof is given in full in appendix B.

4.2. Multiple Layers

In this section, we apply the results of the previous section to prove universality for DRF. We shall consider the deep random features as given in eq (9). The proof of the equivalence relies on fixing all layers, except a single one, and demonstrating that the individual layer may be replaced by their Gaussian equivalent. This relies on an intermediate result, given in the following theorem, stating that the regularity of a dataset, as defined in definition 3 is preserved under random feature mappings.

Theorem 6 *Suppose that the set $\{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^n$ is regular and assumption 3 holds. Then define $\mathbf{z}_i = \sigma(\mathbf{W}\mathbf{x}_i)$ where $\mathbf{W}$ is a $p \times d$ matrix and has independent rows distributed by $\mathcal{N}(0, \frac{1}{d}\mathbf{I})$. Then with probability higher than $1 - n^{-10}$ the set $\{\mathbf{z}_i\}_{i=1}^n$ is regular*[1].

The main consequence of this theorem is that for L layers where $n \sim p_0 \sim p_1 \sim \cdots \sim p_L$, with a probability converging to 1, all dataset $\mathbf{X}^{(l)} = \{\mathbf{x}_i^{(l)}\}_{i=1}^n$ for $l = 1, \ldots, L$ are regular[2]. Now, we can state the main result of this section, which demonstrates a slightly more generic version of universality for an l-layered deep random feature model.

Theorem 7 *Suppose that $n \sim p_0 \sim \cdots \sim p_l$ and take $q = \mathcal{O}(n)$ and let assumption A1-A5 hold. For a fixed final layer l, define noise appended features $\tilde{\mathbf{x}}_i^{(l)}, \tilde{\boldsymbol{\gamma}}_i^{(l)}$ as*

$$\tilde{\mathbf{x}}_i^{(l)} = \begin{bmatrix} \mathbf{x}_i^{(l)} \\ \mathbf{v}_i^{(l)} \end{bmatrix} \qquad \tilde{\boldsymbol{\gamma}}_i^{(l)} = \begin{bmatrix} \boldsymbol{\gamma}_i^{(l)} \\ \mathbf{v}_i^{(l)} \end{bmatrix} \tag{16}$$

where $\mathbf{v}_i^{(l)} \in \mathbb{R}^q$ are independent standard Gaussian vectors. Take a $m \times (q + p_l)$ dictionary $\mathbf{D}$, where $\|\mathbf{D}\|_{\text{op}} < c$ for some constant $c < \infty$ and define the re-represented features

$$\mathbf{f}_i = \mathbf{D}\tilde{\mathbf{x}}_i, \quad \mathbf{g}_i = \mathbf{D}\tilde{\boldsymbol{\gamma}}_i \tag{17}$$

and let $\boldsymbol{F} = [\mathbf{f}_1 \ \cdots \ \mathbf{f}_n]$ and let $\mathbf{G} = [\mathbf{g}_1 \ \cdots \ \mathbf{g}_n]$ be their matrix representations. Then under the assumption that $\mathbf{X}$ is regular,

1. *For any real function ψ with bounded first, second, and third derivatives, there exists a constant $c < \infty$ such that*

$$|\mathbb{E}\psi(\mathcal{E}_{train}(\boldsymbol{F})) - \mathbb{E}\psi(\mathcal{E}_{train}(\mathbf{G}))| \leq \frac{\text{polylog } n}{\sqrt{n}} \tag{18}$$

2. *Let $\hat{\boldsymbol{\theta}}_F$ and $\hat{\boldsymbol{\theta}}_G$ be the optimal solution of problem (2) for $\boldsymbol{F}$ and $\mathbf{G}$. Then for any bounded function $h : \mathbb{R}^{p_L} \to \mathbb{R}$ with bounded $\nabla h(\mathbf{0})$, second and third derivatives (in tensor norm), where the bounds are constant in n, m, p_i for $0 \leq i \leq l$. There exists a constant $c < \infty$ such that*

$$\left|\mathbb{E}h\left(\hat{\boldsymbol{\theta}}_F\right) - \mathbb{E}h\left(\hat{\boldsymbol{\theta}}_G\right)\right| \leq \frac{\text{polylog } n}{\sqrt{n}} \tag{19}$$

Universality of the DRF problem follows directly from this theorem by choosing the final Lth layer, $q = 0$ and, hence adding no additional noise and $\mathbf{D} = \mathbf{I}_{p_L}$ such that no re-representation appears.

1. The exponent of n is arbitrary and can be replaced by any other number
2. Here we assume that the numbers of layers L is fixed, but it is simple to show that the argument also holds for $L = \text{poly}(n)$

4.2.1. PROOF SKETCH

The proof proceeds by means of induction. For the case that $l = 0$, ie a zero layer network the proof is immediate as $\mathbf{x}^0 = \boldsymbol{\gamma}^{(0)}$. Assuming that the induction hypothesis holds for a layer $l-1$ we may consider layer l.

We make use of an intermediate results which may be found in the appendix. In theorem 15 we show that if the data set $\mathbf{x}^{(l)}$ is regular then covariance matrices of $\mathbf{x}^{(l)}$ and $\boldsymbol{\gamma}^{(l)}$ are bounded by $\frac{c\text{polylog } n}{\sqrt{n}}$ for some constant c. Then, the proof proceeds in two steps: First, we consider an intermediate vector

$$\bar{\boldsymbol{\gamma}}_i^{(l)} = \begin{bmatrix} \rho_{1,l}\mathbf{W}^{(l)}\mathbf{x}_i^{(l-1)} + \rho_{2,l}\mathbf{h}_i^{(l)} \\ \mathbf{v}_i^{(l)} \end{bmatrix}. \tag{20}$$

We bound the performance difference ($\mathcal{E}_{train}$) between $\mathbf{x}^{(l)}$ and $\bar{\boldsymbol{\gamma}}^{(l)}$ by theorem 7. Second, we observe that the difference in performance between $\bar{\boldsymbol{\gamma}}^{(l)}$ and $\boldsymbol{\gamma}^{(l)}$ depends only on the difference between $\mathbf{x}^{(l-1)}$ and $\boldsymbol{\gamma}^{(l-1)}$. As such, we may make use of the induction hypothesis to bound this difference. The full proof is given in appendix C.

5. CGMT Analysis

Thanks to the universality results, we only require to analyze the deep Gaussian features $\boldsymbol{\gamma}_L$. Here, we present this analysis in one particular case where ℓ is the square loss, and the regularization function is generic. Additionally, we need to impose a model for the relationship between the labels $\mathbf{y}$ and the input variables $\mathbf{x}^{(0)}$, which we specifically assume to be independent standard normal vectors. For this we make the following definition

$$y_i = \mathbf{x}_i^{(L)T}\boldsymbol{\theta}^* + \nu_i, \tag{21}$$

where $\boldsymbol{\theta}^* \in \mathbb{R}^{p_L}$ is the "true" relationship between the data and the parameters, $\nu_i \sim \mathcal{N}(0, \sigma_{\boldsymbol{\nu}}^2\mathbf{I})$ is noise, and $\mathbf{x}^{(L)}$ is defined in (4). We let $\boldsymbol{\nu} = (\nu_i)_i$ and let $\mathbf{X}^{(L)} = [\mathbf{x}_1^{(L)}\ \mathbf{x}_2^{(L)}\ \cdots \mathbf{x}_n^{(L)}]$. Then, we consider the following optimization problem

$$P_1 = \min_{\boldsymbol{\theta}} \frac{1}{2n}\left\|\mathbf{y} - \mathbf{X}^{(L)}\boldsymbol{\theta}\right\|_2^2 + R(\boldsymbol{\theta}) = \min_{\mathbf{e}} \frac{1}{2n}\left\|\boldsymbol{\nu} - \mathbf{X}^{(L)}\mathbf{e}\right\|_2^2 + R(\boldsymbol{\theta}^* + \mathbf{e}), \tag{22}$$

where $\mathbf{e} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$ and the optimal solutions are denoted by $\hat{\boldsymbol{\theta}}_1, \hat{\mathbf{e}}_1$. We similarly consider the Gaussian equivalent model defined in eq (8). In this case, the data is generated by

$$\tilde{\mathbf{y}}_i = \boldsymbol{\gamma}_i^{(L)T}\boldsymbol{\theta}^* + \nu_i. \tag{23}$$

Again, we let $\tilde{\mathbf{X}}^{(L)} = [\boldsymbol{\gamma}_1^{(L)}\ \boldsymbol{\gamma}_2^{(L)}\ \cdots \boldsymbol{\gamma}_n^{(L)}]$ and define the Gaussian equivalent optimization problem as

$$P_2 = \min_{\boldsymbol{\theta}} \frac{1}{2n}\left\|\tilde{\mathbf{y}} - \tilde{\mathbf{X}}^{(L)}\boldsymbol{\theta}\right\|_2^2 + R(\boldsymbol{\theta}) = \min_{\mathbf{e}} \frac{1}{2n}\left\|\boldsymbol{\nu} - \tilde{\mathbf{X}}^{(L)}\mathbf{e}\right\|_2^2 + R(\boldsymbol{\theta}^* + \mathbf{e}) \tag{24}$$

with corresponding optimal solutions $\hat{\boldsymbol{\theta}}_2, \hat{\mathbf{e}}_2$. By applying theorem 7 to $\boldsymbol{\theta} \to \mathbf{e}$ and $\mathbf{y} \to \boldsymbol{\nu}$, we establish that the statistics of P_1 and P_2 become weakly similar in the sense of their distributions.

Furthermore, for this particular choice of the relationship between the data and the labels the generalization error for the problem P_1 may be expressed as

$$\mathcal{E}_{gen}(\mathbf{e}) = \sigma_{\boldsymbol{\nu}^2} + \mathbf{e}^T \mathbb{E}[\tilde{\mathbf{x}}_{new}^{(L)T} \tilde{\mathbf{x}}_{new}^{(L)}]\mathbf{e}. \tag{25}$$

This function satisfies the conditions on the function $h(\mathbf{e})$. As such in this case, the generalization error is also universal.

For problem P_2, as the matrix $\tilde{\mathbf{X}}$ is Gaussian, it may be analyzed by the CGMT, (see appendix theorem 12), which gives an asymptotic equivalence to a second alternative problem is follows:

Theorem 8 *Let $n \sim p_0 \sim \cdots \sim p_L$ and let assumptions 1-5 hold true. Consider the following optimization problem*

$$P_3 = \max_{\beta>0} \min_{q} T_L + \max_{\xi_L>0,\chi_L>0} \min_{t_L>0,k_L>0} T_{L-1} + \min_{\mathbf{e}} \frac{a}{2pl} \|\mathbf{e}\|^2 + \frac{b}{p_L} \mathbf{e}^T \mathbf{g} + R(\boldsymbol{\theta} + \boldsymbol{\theta}^*) + \max_{\xi_{L-1}>0,\chi_{L-1}>0} \min_{t_{L-1}>0,k_{L-1}>0} \cdots \max_{\xi_0\geq 0,\chi_0\geq 0} \min_{t_0>0,k_0>0} \sum_{i=1}^{L-2} T_l(\mathbf{e}) \tag{26}$$

Where T_L is a function of β, q; T_{L-1}, a, b are functions of $\beta, q, \xi_L, \chi_L, t_L, k_L$; $\mathbf{g} \in \mathbb{R}^{p_L}$ is a standard normal and T_l are functions of $\mathbf{e}, \beta, q, \xi_i, \chi_i, t_i, k_i$ for $L \geq i \geq l$. The exact expressions for the functions a, b, T_i are complicated and are given in the appendix equation (178).

Then,

1. *Then the values of P_2 and P_3 become close, in sense that if P_3 converges to come value c then P_2 will converge to the same value.*
2. *Let $\hat{\boldsymbol{\theta}}_3$ be the optimal point of P_3. Then for any bounded function $h : \mathbb{R}^{p_L} \to \mathbb{R}$ with bounded second and third derivatives (in tensor norm), where the bounds are constant in n, p_i. for $0 \leq i \leq L$, then*

$$\Pr\left(|h(\hat{\mathbf{e}}_2) - h(\hat{\mathbf{e}}_3)| > \epsilon\right) \to 0 \quad \text{as} \quad n, p_0, \ldots, p_L \to \infty \tag{27}$$

A proof of this theorem may be found in the Appendix Section D. Furthermore, if all layers, except the input have the same dimension p the CGMT result can be simplified substantially, these results may be seen in theorem 22. It can be clearly seen that by the triangle inequality and the results of theorem 7 that P_3 and P_1 will similarly asymptotically become weakly similar; as will $h(\hat{\boldsymbol{\theta}}_3)$ and $h(\hat{\boldsymbol{\theta}}_1)$.

5.1. Experimental Results

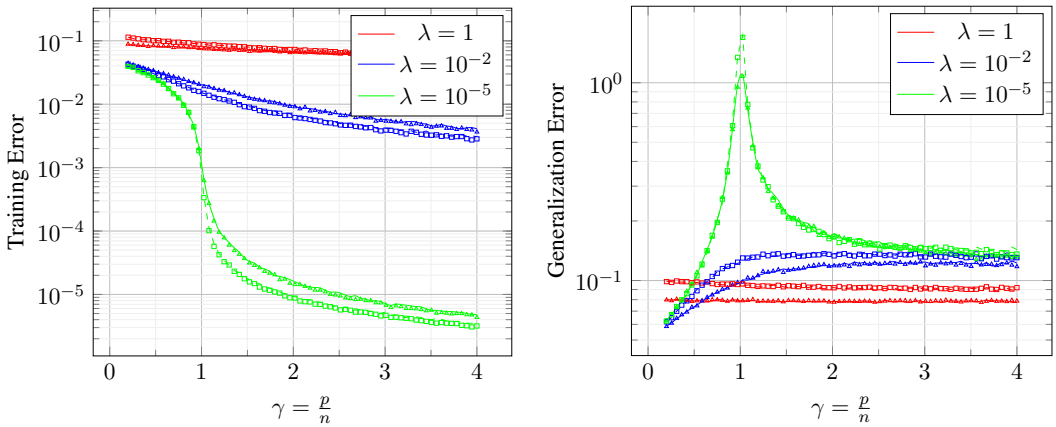
We now demonstrate the validity of our results experimentally. We consider two regularization functions that satisfy assumption A1: the ℓ_2^2 regularization and elastic net regularization, where $R(\boldsymbol{\theta}) = \lambda_1 \|\boldsymbol{\theta}\|_1 + \frac{1}{2}\lambda_2 \|\boldsymbol{\theta}\|_2^2$.

We consider standard Gaussian input of dimension d and examine a 2-Layer RF model where both layers are of dimension p and a 1 layer RF model with hidden layer of dimension p. The ratio $\frac{n}{d}$ was fixed to 1.5 for all experiments. The activation function was chosen to be $\tanh$.

In figure 1 we show the training and generalization error for ℓ_2^2 regularization for 3 different regularization values as a function of the ratio $\frac{p}{n}$. We note that in the 1-Layer case $\frac{p}{n}$ is a measure of

the under or overparameterization of the network. This relationship does not hold in the two layer case, however as may be seen from the figure this ratio is still useful in comparing the two models. In figure 1 the solid line represents the 2-layer case and the dashed line represents the 1-layer case. The triangles are our theoretical predictions for 2-layers, and squares similarly for 1-layer. For the Elastic net case we fix λ_2 to be 10^{-5} and vary only λ_1 these results are similarly shown in figure 2.

We note that in both types of regularization functions, for all values of $\frac{p}{n}$, the 2-layer deep RF model has consistently lower generalization error. With respect to training error the two layer case only outperforms 1-layer at large values of regularization. This suggests that even when training of the layer is not performed there can be a benefit to a deeper embedding of the input data.



(*a*) Training Error (*b*) Generalization Error

Figure 1: Comparison of 1-Layer and 2-Layer RFs, with square loss function, ℓ_2^2 regularization with regularization strength λ. Solid lines represent 2 layer and dashed lines 1-Layer. Triangles are the CGMT results for 2-layers and squares for 1-layer

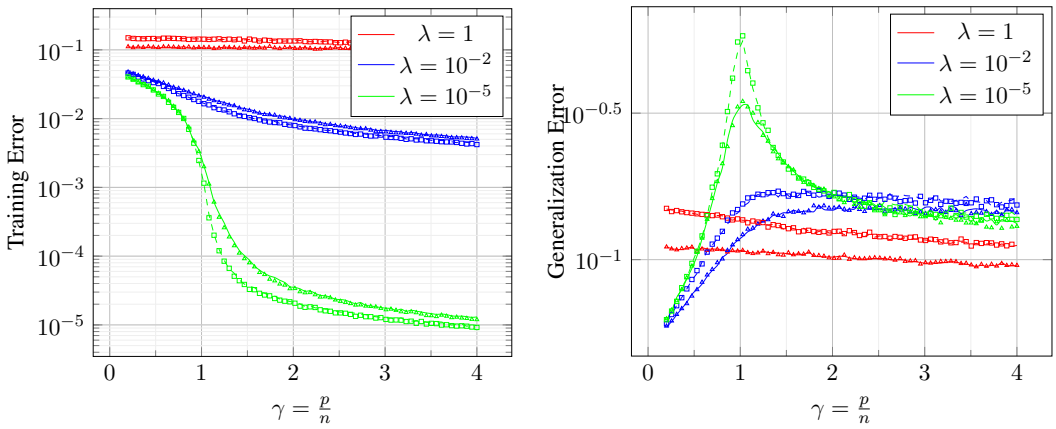
5.2. Eigendistribution of the Covariance Matrix

In the CGMT analysis performed above, where the input data is Gaussian, the Gaussian equivalent features $\gamma^{(L)}$ are distributed as $\mathcal{N}(\mathbf{0}, \mathbf{R}^{(L)})$ where $\mathbf{R}^{(L)}$ is a covariance matrix defined recursively as

$$\mathbf{R}^{(0)} = \mathbf{I} \qquad \mathbf{R}^{(l)} = \rho_{1,l}^2 \mathbf{W}^{(l)} \mathbf{R}^{(l-1)} \mathbf{W}^{(l)T} + \rho_{2,l}^2 \mathbf{I}_{p_l}, \tag{28}$$

where each $\mathbf{W}^{(l)}$ has rows $\mathbf{w}_j^{(l)} \sim \mathcal{N}(\mathbf{0}, \frac{1}{p_{l-1}} \mathbf{I}_{p_{l-1}})$. In the case of ridge regression of linear models, or any rotationally invariant setup, the optimal value is directly dependent upon the eigenvalues of the covariance matrix. As the covariance matrix is random we consider its eigendistribution, the marginal probability distribution over the eigenvalues. We note that what we examine here is the recursively defined covariance matrix of the Gaussian equivalent features, an analysis of the recursively defined covariance of the original distribution is considered in Fan and Wang (2020) and Schröder et al. (2023),

We note that the type of recursion for $\mathbf{R}^{(l)}$ is a form of a Lyapanov recursion, which has been studied in the literature (Vakili, 2011; Emery et al., 2007). We denote the eigendistribu-



(*a*) Training Error (*b*) Generalization Error

Figure 2: Comparison of 1-Layer and 2-Layer RFs, with square loss function and $\ell_1 + \ell_2^2$ regularization with regularization strength λ for the ℓ_1 term and fixed ℓ_2 regularization strength. Solid lines represent 2 layer and dashed lines are 1-Layer. Triangles are the CGMT result for 2-layers and squares for 1-layer

tion of the matrix $\mathbf{R}^{(l)}$ as $f_{\mathbf{R}^{(l)}}(\lambda)$ for eigenvalues λ. In the case of $l = 1$, the matrix $\mathbf{R}^{(1)}$ is a scaled Wishart matrix plus an identity, whose eigendistribution is given by a shifted version of the Marchenko–Pastur distribution. In figure 5.2 we consider the empirical eigendistribution of $\mathbf{R}^{(2)}$, corresponding to the two layer case studied above. We choose $p_0 = 1000$ and $p_2 = 1500$ fixing the input and output dimensions of the layers, and vary the size of the hidden layer p_1. We note as the size of the hidden layer grows the more concentrated the eigendistribution become around zero, while decreasing it results in a more flat structure. In the case of ridge regression, the decreased in the support of the eigenvalues could represent in an increase in model uncertainty at large sizes of the hidden layers.

We also examine the eigendistribution analytically. We make use of the Stieltjes transform $S_{\mathbf{R}^{(l)}}(z)$ of the distribution $f_{\mathbf{R}^{(l)}}$. This transform and its inverse are give by

$$S_{\mathbf{R}^{(l)}}(z) = \int \frac{f_{\mathbf{R}^{(l)}}(\lambda)}{\lambda - z} \mathrm{d}\lambda \qquad f_{\mathbf{R}^{(l)}}(\lambda) = \frac{1}{\pi} \lim_{\omega \to 0^+} \mathrm{Im}[S(\lambda + i\omega)] \tag{29}$$

where i is the imaginary unit, and z is complex. We can demonstrate that the Stieltjes transform of the matrices $\mathbf{R}^{(l)}$ follows the following recursion.

Theorem 9 *Let $\beta_l = \frac{p_l}{p_{l-1}}$, then the Stieltjes transform $S_l(z)$ of $\mathbf{R}^{(l)}$ in (28) is given recursively by*

$$S_{l+1}(z) = \frac{1}{\rho_{1,l+1}^2} \Omega_l \left(\frac{z - \rho_{2,l+1}^2}{\rho_{1,l+1}^2} \right) \tag{30}$$

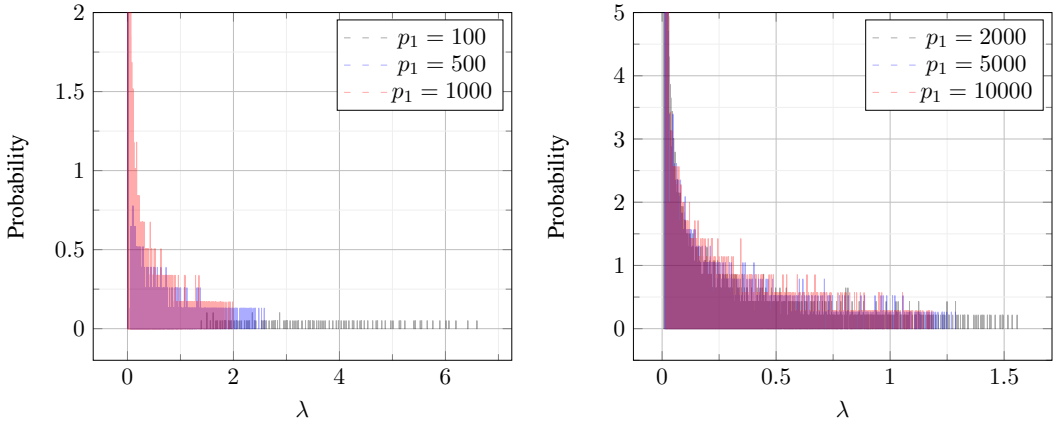
$$\Omega_l(z) = \frac{1}{1 - \beta - \beta z \Omega_l(z)} S_l \left(\frac{z}{1 - \beta - \beta z \Omega_l(z)} \right) \tag{31}$$

Where Ω_0 is the Stieltjes transform of a Wishart matrix, given by

$$\Omega_0 = \frac{1 - \beta_1 - z + \sqrt{z^2 - 2(\beta_1 + 1)z + (\beta_1 - 1)^2}}{2\beta_1 z} \tag{32}$$

Proof The proof is given in appendix E. ■

The recursive definitions given are difficult to compute empirically, as such we will leave visualizing these results to future work. However the recursive structure suggests that there exists a limiting distribution over the eigenvalues in the limit of infinite depth characterized by the different ratio in size between the various layers.



(*a*) $p_1 \leq p_0$ (*b*) $p_1 > p_0$

Figure 3: Empirical Eigendistribution of $\mathbf{R}^{(l)}$ for various sizes p_1 of the 1st hidden layer

6. Conclusion

In this paper, we prove an asymptotic equivalence between deep random feature models and linear Gaussian models with respect to the training and generalization error. As a result of this universality, we can study a Gaussian equivalent model to the DRF model, in the asymptotic limit. We use this fact to provide an exact asymptotic analysis by means of the convex Gaussian min max theorem for an L-layer deep random feature model with Gaussian inputs. We further demonstrate that depth has an effect on training and generalization error both experimentally and by studying the eigendistribution of the Gaussian equivalent model's Covariance matrix.

References

Ehsan Abbasi, Fariborz Salehi, and Babak Hassibi. Universality in learning from linear measurements. *Advances in Neural Information Processing Systems*, 32, 2019.

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *The Journal of Machine Learning Research*, 18(1):714–751, 2017.

Richard Baraniuk, Mark Davenport, Ronald DeVore, and Michael Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020a. ISSN 0027-8424. doi: 10.1073/pnas.1907378117. URL https://www.pnas.org/content/117/48/30063.

Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *arxiv:1906.11300*, 2020b.

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

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

David Bosch, Ashkan Panahi, and Ayca Özcelikkale. Double descent in feature selection: Revisiting lasso and basis pursuit. In *International Conference on Machine Learning (ICML) 2021 Workshop on Overparameterization: Pitfalls & Opportunities*, 2021.

David Bosch, Ashkan Panahi, Ayca Özcelikkale, and Devdatt Dubhash. Double descent in random feature models: Precise asymptotic analysis for general convex regularization, 2022. URL https://arxiv.org/abs/2204.02678.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. The convex geometry of linear inverse problems. *Foundations of Computational mathematics*, 12:805–849, 2012.

Xiangyu Chang, Yingcong Li, Samet Oymak, and Christos Thrampoulidis. Provable benefits of overparameterization in model compression: From double descent to pruning neural networks. *arXiv preprint arXiv:2012.08749*, 2020.

Amit Daniely. Sgd learns the conjugate kernel class of the network. *Advances in Neural Information Processing Systems*, 30, 2017.

Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. *Advances in neural information processing systems*, 29, 2016.

Oussama Dhifallah and Yue Lu. On the inherent regularization effects of noise injection during training. In *International Conference on Machine Learning*, pages 2665–2675. PMLR, 2021.

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

Michel Emery, Arkadi Nemirovski, and Dan Voiculescu. *Lectures on Probability Theory and Statistics: Ecole D'Ete de Probabilites de Saint-Flour XXVIII-1998*. Springer, 2007.

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

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

Federica Gerace, Florent Krzakala, Bruno Loureiro, Ludovic Stephan, and Lenka Zdeborová. Gaussian universality of linear classifiers with random labels in high-dimension. *arXiv preprint arXiv:2205.13303*, 2022.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2):1029 – 1054, 2021. doi: 10.1214/20-AOS1990. URL https://doi.org/10.1214/20-AOS1990.

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

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pages 426–471. PMLR, 2022.

Yehoram Gordon. Some inequalities for gaussian processes and applications. *Israel Journal of Mathematics*, 50(4):265–289, 1985.

Yehoram Gordon. On milman's inequality and random subspaces which escape through a mesh in r n. In *Geometric aspects of functional analysis*, pages 84–106. Springer, 1988.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

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

Arthur Jacot, Clément Hongler, and Franck Gabriel. Neural tangent kernel: Convergence and generalization in neural networks. In *NeurIPS*, pages 8580–8589, 2018.

Sham Kakade, Shai Shalev-Shwartz, Ambuj Tewari, et al. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. *Unpublished Manuscript, http://ttic. uchicago. edu/shai/papers/KakadeShalevTewari09. pdf*, 2(1):35, 2009.

Satish Babu Korada and Andrea Montanari. Applications of the lindeberg principle in communications and statistical learning. *IEEE transactions on information theory*, 57(4):2440–2450, 2011.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. *arXiv preprint arXiv:1711.00165*, 2017.

Jarl Waldemar Lindeberg. Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 15(1):211–225, 1922.

Fanghui Liu, Xiaolin Huang, Yudong Chen, and Johan AK Suykens. Random features for kernel approximation: A survey on algorithms, theory, and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7128–7148, 2021.

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

Bruno Loureiro, Cédric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Capturing the learning curves of generic features maps for realistic data sets with a teacher-student model. *arXiv preprint arXiv:2102.08127*, 2021a.

Bruno Loureiro, Cédric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model, 2021b. URL https://arxiv.org/abs/2102.08127.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 2019. doi: https://doi.org/10.1002/cpa.22008. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.22008.

Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro. *Spin glass theory and beyond: An Introduction to the Replica Method and Its Applications*, volume 9. World Scientific Publishing Company, 1987.

Andrea Montanari and Phan-Minh Nguyen. Universality of the elastic net error. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 2338–2342. IEEE, 2017.

Andrea Montanari and Basil N Saeed. Universality of empirical risk minimization. In *Conference on Learning Theory*, pages 4310–4312. PMLR, 2022.

Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

Samet Oymak and Joel A Tropp. Universality laws for randomized dimension reduction, with applications. *Information and Inference: A Journal of the IMA*, 7(3):337–446, 2018.

Ashkan Panahi and Babak Hassibi. A universal analysis of large-scale regularized least squares solutions. In *NIPS*, pages 3384–3393, 2017.

Omiros Papaspiliopoulos. High-dimensional probability: An introduction with applications in data science, 2020.

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

Dominik Schröder, Hugo Cui, Daniil Dmitriev, and Bruno Loureiro. Deterministic equivalent and error universality of deep random features learning, 2023. URL https://arxiv.org/abs/2302.00401.

Mohamed El Amine Seddik, Cosme Louart, Mohamed Tamaazousti, and Romain Couillet. Random matrix theory proves that deep learning representations of gan-data behave as gaussian mixtures. In *International Conference on Machine Learning*, pages 8573–8582. PMLR, 2020.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. The Gaussian min-max theorem in the Presence of Convexity. *arXiv e-prints*, art. arXiv:1408.4837, August 2014.

Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*, pages 1683–1709. PMLR, 2015.

Ali Vakili. *Random Matrix Recursions in Estimation, Control, and Adaptive Filtering*. PhD thesis, California Institute of Technology, 2011. URL https://resolver.caltech.edu/CaltechTHESIS:06022011-214438378.

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

Dan Voiculescu. Limit laws for random matrices and free products. *Inventiones mathematicae*, 104 (1):201–220, 1991.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Appendix A. Technical Lemmas and Theorem

In this section we give a number of Lemmas and Theorems that will be used in the proofs below.

In the following lemma we demonstrate that passing the input through an activation function with Gaussian weights result in a subgaussian random variable under mild assumptions.

Lemma 10 *Consider* $\mathbf{X} = [\mathbf{x}_1\ \mathbf{x}_2\ \cdots\ \mathbf{x}_n]$, *where* $\mathbf{x}_i \in \mathbb{R}^d$ *and define* $r = \|\mathbf{X}\|_{\text{op}}$. *Suppose that the derivative* σ' *of the activation function* σ *is bounded, i.e.* $\|\sigma'\|_\infty \leq \tau$. *Let* $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \frac{1}{d}\mathbf{I})$. *Then, the random vector* $(\sigma(\mathbf{x}_k^T\mathbf{w}))_k$ *is* $\frac{\tau r}{\sqrt{d}}-$*sub-Gaussian*

Proof Take a unit vector $\mathbf{a} = (a_j) \in \mathbb{R}^n$. We show that $\mathbf{A}(\mathbf{w}) := \sum_{k=1}^n \mathbf{a}_k\sigma(\mathbf{w}^T\mathbf{x}_k)$ is sub-Gaussian with parameter $\tau r/\sqrt{d}$. For this, we show that the function $A(\mathbf{w})$ is $\tau r-$Lipschitz continuous, which implies the desired result (see (Boucheron et al., 2013)). For this, observe that

$$\nabla A = \sum_{k=1}^n \mathbf{x}_k\sigma'(\mathbf{w}^T\mathbf{x}_k)a_k = \mathbf{X}\boldsymbol{\sigma}, \tag{33}$$

where $\boldsymbol{\sigma} = (\sigma'(\mathbf{w}^T\mathbf{x}_k)a_k)_k$ and hence by assumption $\|\boldsymbol{\sigma}\| \leq \tau$. We conclude that

$$\|\nabla A\| \leq \|\mathbf{X}\|_{\text{op}} \|\boldsymbol{\sigma}\| \leq \tau r. \tag{34}$$

This concludes the proof. ■

Here we give a lemma that gives a high probability bound on the norm of a random matrix.

Lemma 11 *Consider a $p \times n$ random matrix $\mathbf{S}$ where each row is independent and τ-sub-Gaussian. Moreover, the covariance of each row is bounded by τ in operator norm. Then, there exists constants c_0, κ only depending on τ such that for any $c > c_0$ the following holds:*

$$\Pr\left[\|\mathbf{S}\| > c(\sqrt{p} + \sqrt{n})\right] \leq e^{-\kappa c n}. \tag{35}$$

Proof The proof is based on the standard $\epsilon-$net argument. Hence we do not give it here. See, for example, (Baraniuk et al., 2008) for a similar proof. ■

Next for completeness we state the Convex Gaussian Min Max Theorem (Gordon, 1985, 1988; Thrampoulidis et al., 2014). We make heavy use of this theorem in the proof of theorem 8.

Theorem 12 (Convex Gaussin Min Max Theorem (CGMT)) *Let $\mathbf{G} \in \mathbb{R}^{n\times m}$, $\mathbf{g} \in \mathbb{R}^m$, and $\mathbf{h} \in \mathbb{R}^n$ be independent of each other and have entries distributed according to $\mathcal{N}(0,1)$. Let $\mathcal{S}_1 \subset \mathbb{R}^n$ and $\mathcal{S}_2 \subset \mathbb{R}^m$ be non empty compact sets. Let $f(\cdot,\cdot)$ be a continuous function on $\mathcal{S}_1 \times \mathcal{S}_2$. We define the primary and alternative optimization problems as follows:*

$$P(\mathbf{G}) := \quad \min_{\mathbf{x}\in\mathcal{S}_1} \max_{\mathbf{y}\in\mathcal{S}_2} \mathbf{x}^T\mathbf{G}\mathbf{y} + f(\mathbf{x},\mathbf{y}) \tag{36}$$

$$A(\mathbf{g},\mathbf{h}) := \quad \min_{\mathbf{x}\in\mathcal{S}_1} \max_{\mathbf{y}\in\mathcal{S}_2} \|\mathbf{x}\|_2 \mathbf{g}^T\mathbf{y} + \|\mathbf{y}\|_2 \mathbf{h}^T\mathbf{x} + f(\mathbf{x},\mathbf{y}), \tag{37}$$

Then for any $c_1 \in \mathbb{R}$ we have that

$$\Pr(P(\mathbf{G}) < c_1) \leq 2\Pr(A(\mathbf{g},\mathbf{h}) \leq c_1). \tag{38}$$

Under the further assumption that $\mathcal{S}_1$ and $\mathcal{S}_2$ are convex sets, and f is concave-convex on $\mathcal{S}_1 \times \mathcal{S}_2$ then for all $c_2 \in \mathbb{R}$ we have that

$$\Pr(P(\mathbf{G}) > c_2) \leq 2\Pr(A(\mathbf{g},\mathbf{h}) \geq c_2). \tag{39}$$

We note this theorem demonstrates that if $A(\mathbf{g},\mathbf{h})$ concentrates on a particular value c, ie

$$\Pr(|A(\mathbf{g},\mathbf{h}) - c| > \epsilon) \xrightarrow[n,m\to\infty]{P} 0, \qquad \forall \epsilon > 0 \tag{40}$$

then $P(\mathbf{G})$ will concentrate on the same limit.

Appendix B. Proof of Theorem 5

Our proof is based on an application of Lindeberg's argument to the sequence of features ϕ_j for $j = 1, \dots, p$. We will adopt the following notation for this section. For a matrix $\mathbf{A}$ we denote its ith row by means of superscript $\mathbf{a}^i$ and its jth column by means of subscript $\mathbf{a}_j$.

For simplicity, for any $m \times n$ matrix $\mathbf{Z}$ with columns $(\mathbf{z}_k)$ we define

$$L(\mathbf{z}) = \min_{\boldsymbol{\theta} \in \mathbb{R}^m} \frac{1}{n} \sum_{k=1}^{n} \ell(\mathbf{z}_k^T \boldsymbol{\theta}, y_k) + R(\boldsymbol{\theta}). \tag{41}$$

By means of a splitting technique, we may express this as

$$\begin{aligned} L(\mathbf{Z}) &= \min_{\boldsymbol{\theta} \in \mathbb{R}^m, \boldsymbol{\alpha} \in \mathbb{R}^n} \max_{\mathbf{d} \in \mathbb{R}^n} \frac{1}{n} \left(\sum_{k=1}^{n} \ell(\alpha_k, y_k) + d_k(\alpha_k - \mathbf{z}_k^T \boldsymbol{\theta}) \right) + R(\boldsymbol{\theta}) \\ &= - \min_{\mathbf{d} \in \mathbb{R}^n} \underbrace{\frac{1}{n} \sum_{k=1}^{n} \ell^*(-d_k, y_k) + R^*(\frac{1}{n} \mathbf{Z}\mathbf{d})}_{\Lambda(\mathbf{d}, \mathbf{Z})} \end{aligned} \tag{42}$$

where ℓ^*, R^* are the Legendre transforms of ℓ and R respectively. We note that $L(\boldsymbol{F}) = \mathcal{E}_{train}(\boldsymbol{F})$ and $L(\mathbf{G}) = \mathcal{E}_{train}(\mathbf{G})$. Furthermore, we define $\mathbf{Z}_r$ for $r = 0, 1, \dots, m$ as

$$\mathbf{Z}_r = \sum_{j=1}^{r} \mathbf{d}_j \boldsymbol{\gamma}^j + \sum_{j=r+1}^{p} \mathbf{d}_j \boldsymbol{\phi}^j, \tag{43}$$

where $\boldsymbol{\gamma}^j, \boldsymbol{\phi}^j$ are the jth row of $\boldsymbol{\Phi}$ and $\boldsymbol{\Gamma}$ respectively. We note that, $\mathbf{d}_j \boldsymbol{\gamma}^j$ and $\mathbf{d}_j \boldsymbol{\phi}^j$ are outer (tensor) products, resulting in matrices. As a result $\mathbf{Z}_0 = \boldsymbol{F}$ and $\mathbf{Z}_m = \mathbf{G}$. We have that

$$|\mathbb{E}\psi(L(\boldsymbol{F})) - \mathbb{E}\psi(L(\mathbf{G}))| \leq \sum_{r=1}^{m} |\mathbb{E}\psi(L(\mathbf{Z}_r)) - \mathbb{E}\psi(L(\mathbf{Z}_{r-1}))| \,. \tag{44}$$

Now , for $r = 1, 2, \dots, m$ for any vector $\mathbf{u} \in \mathbb{R}^n$, define

$$\mathbf{Z}_{-r}(\mathbf{u}) = \sum_{j=1}^{r-1} \mathbf{d}_j \boldsymbol{\gamma}^j + \mathbf{d}_r \mathbf{u}^T + \sum_{j=r+1}^{p} \mathbf{d}_j \boldsymbol{\phi}^j. \tag{45}$$

We note that $\mathbf{Z}_r = \mathbf{Z}_{-r}(\boldsymbol{\gamma}^r)$ and that $\mathbf{Z}_{r-1} = \mathbf{Z}_{-r}(\boldsymbol{\phi}^r)$, as such

$$\begin{aligned} &\mathbb{E}\psi(L(\mathbf{Z}_r)) - \mathbb{E}\psi(L(\mathbf{Z}_{r-1})) = \\ &[\mathbb{E}\psi(L(\mathbf{Z}_{-r}(\boldsymbol{\gamma}^r))) - \mathbb{E}\psi(L(\mathbf{Z}_{-r}(\mathbf{0})))] - [\mathbb{E}\psi(L(\mathbf{Z}_{-r}(\boldsymbol{\phi}^r))) - \mathbb{E}\psi(L(\mathbf{Z}_{-r}(\mathbf{0})))] \,. \end{aligned} \tag{46}$$

We now define $\hat{\mathbf{d}}_r$ and $\hat{\mathbf{d}}_{-r}$ as the minimal solutions of $\Lambda(\mathbf{d}, \mathbf{Z}_r)$ and $\Lambda(\mathbf{d}, \mathbf{Z}_{-r}(\mathbf{0}))$ respectively. We note that $\boldsymbol{\gamma}^r, \boldsymbol{\phi}^r$ are $\tau-$sub-Gaussian and independent of $\mathbf{Z}_{-r}(\mathbf{0})$. Hence, we examine the following term:

$$\mathbb{E}\psi(L(\mathbf{Z}_{-r}(\mathbf{u}))) - \mathbb{E}\psi(L(\mathbf{Z}_{-r}(\mathbf{0}))), \tag{47}$$

for a generic τ-sub-Gaussian independent random vector $\mathbf{u}$.

We recall that $\mathbf{R}$ is μ-strongly convex and M-smooth, we have that R^* is $\frac{1}{M}-$strongly convex and $\frac{1}{\mu}$ smooth for any $\mathbf{Z}$, (Kakade et al., 2009)[theorem 6]. The optimal solution $\hat{\mathbf{d}}$ is therefore uniquely identified by the first order optimiality condition

$$\mathbf{h}(\mathbf{d},\mathbf{Z}) := \nabla_{\mathbf{d}}\Lambda(\mathbf{d},\mathbf{Z}) = \frac{1}{n}\boldsymbol{\zeta}(\mathbf{d}) + \frac{1}{n}\mathbf{Z}^T\nabla R^*\left(\frac{1}{n}\mathbf{Z}\mathbf{d}\right) = \mathbf{0} \tag{48}$$

where $\boldsymbol{\zeta}(\mathbf{d})$ is the vector of values $(\ell(d_k, y_k))_k$ with ℓ' being the partial derivative of ℓ^* with respect to the first argument. In particular, $\mathbf{h}(\hat{\mathbf{d}}_{-r}, \mathbf{Z}_{-r}(\mathbf{0})) = \mathbf{0}$. We can therefore conclude that for every $\mathbf{u} = (u_k)_{k=1}^n$ that

$$\mathbf{h}(\hat{\mathbf{d}}_{-r}, \mathbf{Z}_{-r}(\mathbf{u})) = \frac{1}{n}\mathbf{u}\mathbf{d}_r^T\nabla R^*\left(\frac{1}{n}\mathbf{Z}_{-r}\hat{\mathbf{d}}_{-r}\right) + \frac{1}{n}\mathbf{Z}_{-r}^T(\mathbf{u})\left(\nabla R^*\left(\frac{1}{n}\mathbf{Z}_{-r}(\mathbf{u})\hat{\mathbf{d}}_{-r}\right) - \nabla R^*\left(\frac{1}{n}\mathbf{Z}_{-r}(\mathbf{0})\hat{\mathbf{d}}_{-r}\right)\right). \tag{49}$$

Where we have used the fact that

$$\mathbf{Z}_{-r}(\mathbf{u}) = \mathbf{Z}_{-r}(\mathbf{0}) + \mathbf{d}_r\mathbf{u}^T. \tag{50}$$

We can further conclude that

$$\frac{1}{n}\mathbf{Z}_{-r}(\mathbf{u})\hat{\mathbf{d}}_{-r} = \frac{1}{n}\mathbf{Z}_{-r}(\mathbf{0})\hat{\mathbf{d}}_{-r} + \mathbf{d}_r\frac{\mathbf{u}^T\hat{\mathbf{d}}_{-r}}{n}. \tag{51}$$

B.1. Bounding the terms in 49

Now, we introduce a series of bounds and approximations on the terms involved in 49. For ease of notation, we introduce the following:

Definition 13 *We say than an expression including the parameter c holds with high probability (w.h.p) if there are constants c_0, κ such that for any $c > c_0$, the expression holds with probability higher than $1 - \kappa e^{-\kappa c n}$. We also denote $C := \mathrm{poly}(c)$.*

Recall that we have assumed that $\mathbf{u}$ is a τ-sub-Gaussian vector. We now note that all the matrices $\mathbf{Z}_r, \mathbf{Z}_{-r}(\mathbf{0})$ and $\mathbf{Z}_{-r}(\mathbf{u})$ can be expressed as $\mathbf{Z} = \mathbf{D}\mathbf{S}$ where each row of $\mathbf{S}$ is independent an associated with either a random feature, a replaced Gaussian feature, or $\mathbf{u}$. Hence, for $p \sim n$, by assumption A1 and lemma 11, we have that $\|\mathbf{S}\|_2 \leq C(\sqrt{p} + \sqrt{n}) \leq C\sqrt{n}$ holds with high probability, and by the conditions on $\mathbf{D}$ assumed for the theorem the matrices $\mathbf{Z}_r, \mathbf{Z}_{-r}(\mathbf{0})$ and $\mathbf{Z}_{-r}(\mathbf{u})$ are also bounded in operator norm by $C\sqrt{n}$ with high probability.

Next we note by assumption A2 that $\|\boldsymbol{\zeta}(\mathbf{0})\| \leq C\sqrt{n}$ and by assumption A1, that $\nabla R^*(\mathbf{0}) = \mathcal{O}(1)$. Moreover, as R^* is $\frac{1}{M}-$strongly convex, we obtain that

$$\left\|\hat{\mathbf{d}}_{-r}\right\| \leq M\left\|\mathbf{h}(\mathbf{0}, \mathbf{Z}_{-r}(\mathbf{0}))\right\| \leq \frac{C}{\sqrt{n}}. \tag{52}$$

By the $\frac{1}{\mu}-$ smoothness of R^*, we also obtain that:

$$\left\|\nabla R^*\left(\frac{1}{n}\mathbf{Z}_{-r}(\mathbf{0})\hat{\mathbf{d}}_{-r}\right)\right\| \leq \frac{1}{\mu n}\left\|\mathbf{Z}_{-r}(\mathbf{0})\hat{\mathbf{d}}_{-r}\right\| \leq \frac{C}{n} \quad \text{w.h.p} \tag{53}$$

and

$$\left\|\nabla R^*\left(\frac{1}{n}\mathbf{Z}_{-r}(\mathbf{u})\hat{\mathbf{d}}_{-r}\right)-\nabla R^*\left(\frac{1}{n}\mathbf{Z}_{-r}(\mathbf{0}\hat{\mathbf{d}}_{-r})\right)\right\| \le \frac{1}{\mu n}\|\mathbf{d}_r\|\left|\mathbf{u}^T\hat{\mathbf{d}}_{-r}\right| \le \frac{C}{n\sqrt{n}}\left|\mathbf{u}^T\hat{\mathbf{d}}_{-r}\right|. \quad (54)$$

Recalling that $\mathbf{u}$ is τ-sub-Gaussian, hence:

$$\Pr\left[\left|\mathbf{u}^T\hat{\mathbf{d}}_{-r}\right| > c\sqrt{n}\left\|\hat{\mathbf{d}}_{-r}\right\|\right] \le e^{-kcn}, \quad (55)$$

where κ only depends on τ. From this we conclude that,

$$\left\|\nabla R^*\left(\frac{1}{n}\mathbf{Z}_{-r}(\mathbf{u})\hat{\mathbf{d}}_{-r}\right)-\nabla R^*\left(\frac{1}{n}\mathbf{Z}_{-r}(\mathbf{0})\hat{\mathbf{d}}_{-r}\right)\right\| \le \frac{C}{n\sqrt{n}} \quad \text{w.h.p.} \quad (56)$$

Finally, applying lemma 11 to $\mathbf{S} = \mathbf{u}^T$ (with $p = 1$) shows that $\|\mathbf{u}\| \le C\sqrt{n}$ with high probability. As such we can make the following conclusion about (49):

$$\left\|\mathbf{h}(\hat{\mathbf{d}}_{-r}, \mathbf{Z}_{-r}(\mathbf{u})) - \frac{\delta_r}{n}\mathbf{u}\right\| \le \frac{C}{n^{3/2}} \quad \text{w.h.p} \quad (57)$$

where $\delta_r = \mathbf{d}_r^T\nabla R^*\left(\frac{1}{n}\mathbf{Z}_{-r}(\mathbf{0})\hat{\mathbf{d}}_{-r}\right)$. Hence, $|\delta_r| \le C$ w.h.p and

$$\left\|\mathbf{h}(\hat{\mathbf{d}}_r, \mathbf{Z}_{-r}(\mathbf{u}))\right\| \le \frac{C}{\sqrt{n}} \quad \text{w.h.p.} \quad (58)$$

B.2. Approximating $L(\mathbf{Z}_{-r}(\mathbf{u}))$

We now denote $\mathbf{J}_r = \frac{\partial \mathbf{h}}{\partial \mathbf{d}}(\hat{\mathbf{d}}_{-r}, \mathbf{Z}_{-r}(\mathbf{0}))$ and introduce the following point:

$$\hat{\mathbf{d}}_{+r}(\mathbf{u}) = \hat{\mathbf{d}}_{-r} - \frac{\delta_r}{n}\mathbf{J}_r^{-1}\mathbf{u}. \quad (59)$$

We note that $\frac{\delta_r}{n}\mathbf{u} + \mathbf{J}_r(\hat{\mathbf{d}}_{+r}(\mathbf{u}) - \hat{\mathbf{d}}_{-r}) = \mathbf{0}$ and by strong convexity that $\mathbf{J}_r \succeq \frac{1}{M}L$. Furthermore, by the assumption on the third derivatives,

$$\begin{aligned}
&\left\|\mathbf{h}(\hat{\mathbf{d}}_{+r}, \mathbf{Z}_{-r}(\mathbf{u}))\right\| \\
&= \left\|\hat{\mathbf{d}}_{+r}, \mathbf{Z}_{-r}(\mathbf{u}) - \frac{\delta_r}{n} - \mathbf{J}_r\left(\hat{\mathbf{d}}_{+r}(\mathbf{u}) - \hat{\mathbf{d}}_{-r}\right)\right\| \\
&\le \left\|\mathbf{h}(\hat{\mathbf{d}}_{+r}, \mathbf{Z}_{-r}(\mathbf{u})) - \mathbf{h}(\hat{\mathbf{d}}_{+r}, \mathbf{Z}_{-r}(\mathbf{u})) - \mathbf{J}_r\left(\hat{\mathbf{d}}_{+r}(\mathbf{u}) - \hat{\mathbf{d}}_{-r}\right)\right\| + \frac{C}{n^{3/2}} \\
&\le C\left\|\hat{\mathbf{d}}_{+r}(\mathbf{u}) - \hat{\mathbf{d}}_{-r}\right\|^2 + \frac{C}{n^{3/2}} = \frac{C\delta_r^2}{n^2}\left\|\mathbf{J}_r^{-1}\mathbf{u}\right\|^2 + \frac{C}{n^{3/2}} \le \frac{C}{n} \quad \text{w.h.p.}
\end{aligned} \quad (60)$$

Finally, from strong convexity, we conclude that

$$0 \le \Lambda(\hat{\mathbf{d}}_{+r}(\mathbf{u}), \mathbf{Z}_{-r}(\mathbf{u})) + L(\mathbf{Z}_{-r}(\mathbf{u})) \le \frac{M}{2}\left\|\mathbf{h}(\hat{\mathbf{d}}_{+r}(\mathbf{u}), \mathbf{Z}_{-r}(\mathbf{u}))\right\|^2 \le \frac{C}{n^2} \quad \text{w.h.p.} \quad (61)$$

On the other hand, we note that

$$\frac{1}{n}\mathbf{Z}_{-r}(\mathbf{u})\hat{\mathbf{d}}_{+r} = \frac{1}{n}\mathbf{Z}_{-r}(\mathbf{0})\mathbf{d}_{-r} + \frac{1}{n}\mathbf{d}_r\mathbf{u}^T\mathbf{d}_{-r} - \frac{\delta_r}{n^2}\mathbf{Z}_{-r}(\mathbf{0})\mathbf{J}_r^{-1}\mathbf{u} - \frac{\delta_r}{n^2}\mathbf{d}_r\mathbf{u}^T\mathbf{J}_r^{-1}\mathbf{u}. \quad (62)$$

We now define

$$B_r(\mathbf{u}) := \boldsymbol{\eta}_r^T \left[\frac{1}{n} \mathbf{d}_r \mathbf{u}^T \mathbf{d}_{-r} - \frac{\delta_r}{n^2} \mathbf{u}^T \mathbf{J}_r^{-1} \mathbf{u} \right] + \frac{1}{2n^2} \mathbf{d}_r^T \mathbf{H}_r \mathbf{d}_r \left(\mathbf{u}^T \mathbf{d}_{-r} \right)^2 + \frac{\delta_r^2}{2n^3} \mathbf{u}^T \mathbf{J}_r^{-1} \boldsymbol{\Lambda}_r \mathbf{J}_r^{-1} \mathbf{u}, \quad (63)$$

where $\boldsymbol{\eta}_r$ and $\mathbf{H}_r$ are the gradient and Hessian of $\mathbf{R}^*$ respectively at $\frac{1}{n}\mathbf{Z}_{-r}(\mathbf{0})\mathbf{d}_{-r}$ and $\boldsymbol{\Lambda}_r$ is the diagonal matrix of elements $(\ell''(d_k, y_k))$ where ℓ'' is the second derivative of ℓ^* with respect to the first argument. From the previous bounds we conclude that

$$\left| \Lambda(\hat{\mathbf{d}}_{+r}(\mathbf{u}), \mathbf{Z}_{-r}(\mathbf{u})) + L(\mathbf{Z}_{-r}(\mathbf{0})) - B_r(\mathbf{u}) \right| \leq \frac{C}{n^{3/2}} \quad \text{w.h.p} \quad (64)$$

from which we find that

$$|L(\mathbf{Z}_{-r}(\mathbf{u})) - L(\mathbf{Z}_{-r}(\mathbf{0})) - B_r(\mathbf{u})| \leq \frac{C}{n^{3/2}} \quad \text{w.h.p.} \quad (65)$$

Hence, by the bounded derivatives of ψ we have:

$$|\psi(L(\mathbf{Z}_{-r}(\mathbf{u}))) - \psi(L(\mathbf{Z}_{-r}(\mathbf{0}))) - B_r(\mathbf{u})| \leq \frac{C}{n^{3/2}} \quad \text{w.h.p.} \quad (66)$$

From the mean value theorem, we have that

$$\begin{aligned} &|\psi(L(\mathbf{Z}_{-r}(\mathbf{0}))) + B_r(\mathbf{u}) - \psi(L(\mathbf{Z}_{-r}(\mathbf{0}))) \\ &-\psi'(L(\mathbf{Z}_{-r}(\mathbf{0})))B_r(\mathbf{u}) - \frac{1}{2}\psi''(L(\mathbf{Z}_{-r}(\mathbf{0})))B_r^2(\mathbf{u}) \Big| \leq C \left| B_r(\mathbf{u}) \right|^3 \end{aligned} \quad (67)$$

Again, making use of the previous bounds, under the product measure Ω we observe that

$$\left| |B_r(\mathbf{u})|^2 - \frac{(\boldsymbol{\eta}^T \mathbf{d}_r)^2 (\mathbf{u}^T \mathbf{d}_{-r})^2}{n^2} \right| \leq \frac{C}{n^{3/2}}, \qquad |B_r(\mathbf{u})| \leq \frac{C}{\sqrt{n}} \quad \text{w.h.p} \quad (68)$$

and hence

$$\begin{aligned} &|\psi(L(\mathbf{Z}_{-r}(\mathbf{0}))) + B_r(\mathbf{u}) - \psi(L(\mathbf{Z}_{-r}(\mathbf{0}))) \\ &-\psi'(L(\mathbf{Z}_{-r}(\mathbf{0})))B_r(\mathbf{u}) - \frac{1}{2}\psi''(L(\mathbf{Z}_{-r}(\mathbf{0}))) \frac{(\boldsymbol{\eta}^T \mathbf{d}_r)^2 (\mathbf{u}^T \mathbf{d}_{-r})^2}{n^2} \Big| \leq \frac{C}{n^{3/2}} \quad \text{w.h.p.} \end{aligned} \quad (69)$$

Combining all of the steps together, we obtain that

$$\begin{aligned} &|\psi(L(\mathbf{Z}_{-r}(\mathbf{u}))) - \psi(L(\mathbf{Z}_{-r}(\mathbf{0}))) \\ &-\psi'(L(\mathbf{Z}_{-r}(\mathbf{0})))B_r(\mathbf{u}) - \frac{1}{2}\psi''(L(\mathbf{Z}_{-r}(\mathbf{0}))) \frac{(\boldsymbol{\eta}^T \mathbf{d}_r)^2 (\mathbf{u}^T \mathbf{d}_{-r})^2}{n^2} \Big| \leq \frac{C}{n^{3/2}} \quad \text{w.h.p} \end{aligned} \quad (70)$$

B.3. Bounding the Increments of (44) and Final Steps

We now employ the following observation:

Lemma 14 *Suppose that A is a non-negative random variable such that $A \leq C$ w.h.p with $C = \text{poly}(c)$. There exists a universal constant c_1 such that $\mathbb{E}[A] \leq c_1$.*

Proof Note that the assumptions imply that there exist universal constants c_0, κ such that for $c > c_0$

$$\Pr[A > C] \leq \kappa e^{-\kappa n c} \tag{71}$$

Note that $C = \text{poly}(c) \leq (\alpha c)^\beta$ for some constants $\alpha, \beta > 0$. Hence for $C > C_0 := (\alpha c_0)^\beta$, we have

$$\Pr[A > C] \leq \kappa e^{-\frac{\kappa}{\alpha} n C^{\frac{1}{\beta}}}. \tag{72}$$

As such, by making use of Tonelli's theorem we have that

$$\mathbb{E}[A] = \int_0^\infty \Pr[A > C] dC \leq C_0 + \kappa \int_{C_0}^\infty e^{-\frac{\kappa}{\alpha} n C^{\frac{1}{\beta}}}. \tag{73}$$

It is simple to check that the right hand side is bounded by a universal constant. ■

According to lemma 14 we have that

$$\begin{aligned} &|\mathbb{E}\psi(L(\mathbf{Z}_{-r}(\mathbf{u}))) - \mathbb{E}\psi(L(\mathbf{Z}_{-r}(\mathbf{0}))) \\ &-\mathbb{E}\psi'(L(\mathbf{Z}_{-r}(\mathbf{0})))B_r(\mathbf{u}) - \frac{1}{2}\mathbb{E}\psi''(L(\mathbf{Z}_{-r}(\mathbf{0})))\frac{(\boldsymbol{\eta}^T\mathbf{d}_r)^2(\mathbf{u}^T\mathbf{d}_{-r})^2}{n^2}\Big| \leq \frac{c_1}{n^{3/2}}, \end{aligned} \tag{74}$$

for some universal constant c_1. Now we note that each expectation can be carried out y first conditioning on $\mathbf{Z}_{-r}(\mathbf{0})$ and then taking the expectation with respect to it. Accordingly, we denote $\mathbb{E}_{\mathbf{u}} := \mathbb{E}[\cdot|\mathbf{Z}_{-r}(\mathbf{0})]$ as this expectation is only over $\mathbf{u}$, which is independent of $\mathbf{Z}_{-r}(\mathbf{0})$. Furthermore, we repeat the above bound for $\mathbf{u} = \boldsymbol{\gamma}^r$ and $\mathbf{u} = \boldsymbol{\phi}^r$, from which we obtain:

$$\begin{aligned} &\Big|\mathbb{E}\psi(L(\mathbf{Z}_r)) - \mathbb{E}\psi(L(\mathbf{Z}_{r-1})) - \mathbb{E}\psi'(L(\mathbf{Z}_{-r}(\mathbf{0})))[\mathbb{E}_{\mathbf{u}}B_r(\boldsymbol{\gamma}^r) - \mathbb{E}_{\mathbf{u}}B_r(\boldsymbol{\phi}^r)] \\ &-\frac{1}{2}\mathbb{E}\psi''(L(\mathbf{Z}_{-r}(\mathbf{0})))\frac{(\boldsymbol{\eta}^T\mathbf{d}_r)^2}{n^2}\left[\mathbb{E}_{\mathbf{u}}(\mathbf{d}_{-r}^T\boldsymbol{\phi}^r)^2 - \mathbb{E}_{\mathbf{u}}(\mathbf{d}_{-r}^T\boldsymbol{\gamma}^r)^2\right]\Big| \leq \frac{2c_1}{n^{3/2}}. \end{aligned} \tag{75}$$

Making use of the bounds on the derivatives of ψ, we obtain:

$$\begin{aligned} |\mathbb{E}\psi(L(\mathbf{Z}_r)) - \mathbb{E}\psi(L(\mathbf{Z}_{r-1}))| &\leq c\mathbb{E}\left|\mathbb{E}_{\mathbf{u}}B_r(\boldsymbol{\gamma}^r) - \mathbb{E}_{\mathbf{u}}B_r(\boldsymbol{\phi}^r)\right| + \\ &\frac{c}{2}\mathbb{E}\left|\frac{(\boldsymbol{\eta}^T\mathbf{d}_r)^2}{n^2}[\mathbb{E}_{\mathbf{u}}(\mathbf{d}_{-r}^T\boldsymbol{\phi}^r)^2 - \mathbb{E}_{\mathbf{u}}(\mathbf{d}_{-r}^T\boldsymbol{\gamma}^r)]\right| + \frac{2c_1}{n^{3/2}}. \end{aligned} \tag{76}$$

By the previous bounds, it is straightforward to see that

$$\mathbb{E}\left|\mathbb{E}_{\mathbf{u}}B_r(\boldsymbol{\gamma}^r) - \mathbb{E}_{\mathbf{u}}B_r(\boldsymbol{\phi}^r)\right| \leq \frac{C}{n}\left\|\boldsymbol{K}_r - \boldsymbol{K}_r'\right\|_{\text{op}} \quad \text{w.h.p,} \tag{77}$$

and

$$\frac{(\boldsymbol{\eta}^T\mathbf{d}_r)^2}{n^2}[\mathbb{E}_{\mathbf{u}}(\mathbf{d}_{-r}^T\boldsymbol{\phi}^r)^2 - \mathbb{E}_{\mathbf{u}}(\mathbf{d}_{-r}^T\boldsymbol{\gamma}^r)] \leq \frac{C}{n}\left\|\boldsymbol{K}_r - \boldsymbol{K}_r'\right\|_{\text{op}} \quad \text{w.h.p.} \tag{78}$$

Hence by lemma 14 and (44) we conclude that there exists a universal constant c_1 such that

$$|\mathbb{E}\psi(L(\boldsymbol{F})) - \mathbb{E}\psi(L(\mathbf{G}))| \leq \frac{c_1}{n}\sum_{r=1}^p \left\|\boldsymbol{K}_r - \boldsymbol{K}_r'\right\|_{\text{op}} + \frac{c_1}{\sqrt{n}} \tag{79}$$

This concludes the proof of part 1 of the theorem

B.4. Proof of part 2

For $\epsilon \in \mathbb{R}$, we define $R_\epsilon := R + \epsilon h$. Define $L_\epsilon(F), L_\epsilon(G)$ as the optimal values with R_ϵ and note that for all ϵ

$$\epsilon h(\hat{\boldsymbol{\theta}}_F) \geq L_\epsilon(F) - L(F), \tag{80}$$

and

$$\epsilon h(\hat{\boldsymbol{\theta}}_G) \geq L_\epsilon(G) - L(G). \tag{81}$$

We then note that for sufficiently (but finitely) small ϵ the conditions of the theorem are satisfied. Choose $\epsilon > 0$ such that both ϵ and $-\epsilon$ statisfy these conditions. Then we have

$$\frac{L_\epsilon(F) - L(F) + L_{-\epsilon}(G) + L(G)}{\epsilon} \leq h(\hat{\boldsymbol{\theta}}_F) - h(\hat{\boldsymbol{\theta}}_G) \leq \frac{L(F) - L_{-\epsilon}(F) - L_\epsilon(F) + L(G)}{\epsilon}. \tag{82}$$

Taking the expectation, and making use of the results of part 1, with $\psi(x) = x$ we conclude that

$$\left|\mathbb{E}h(\hat{\boldsymbol{\theta}}_F) - \mathbb{E}h(\hat{\boldsymbol{\theta}}_G)\right| \leq \mathbb{E}\left[\frac{2L(F) - L_{-\epsilon}(F) - L_\epsilon(F)}{\epsilon}\right] + \frac{c}{n\epsilon}\sum_{r=1}^{p}\left\|\boldsymbol{K}_r - \boldsymbol{K}_r'\right\| + \frac{c}{\epsilon\sqrt{n}}. \tag{83}$$

we can now choose $\epsilon = \frac{1}{n^{1/4}}$, from this we can see that the latter two terms go to zero in the limit of large n. For the first term we note that when ϵ grows small that

$$\frac{L(F) - L_\epsilon(F)}{\epsilon} + \frac{L(F) - L_{-\epsilon}(F)}{\epsilon} \to 0. \tag{84}$$

Moreover, $-\frac{L_{-\epsilon}(F) - L(F)}{\epsilon} - \frac{L_\epsilon(F) - L(F)}{\epsilon}$ is bounded by twice the bound h. Then, we may invoke the dominated convergence theorem and conclude that

$$\mathbb{E}\left[-\frac{L_\epsilon(G) - L(G)}{\epsilon} - \frac{L_\epsilon(F) - L(F)}{\epsilon}\right] \to 0. \tag{85}$$

Which concludes the proof.

Appendix C. Proof of Theorems 6 and 7

The proof of these theorem relies on two intermediate results, we shall prove both of these first. Firstly consider the following theorem:

Theorem 15 *Assume that σ is odd and that assumption A4 holds. Take*

$$\mu := \sup_{i,j}\left|\frac{\mathbf{x}_i^T\mathbf{x}_j}{d} - \delta_{ij}\right|, \tag{86}$$

and $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \frac{1}{d}\mathbf{I})$. Consider the random vector $\boldsymbol{\phi} = (\sigma(\mathbf{x}_k^T\mathbf{w}))_k$ and denote its covariance matrix by $\boldsymbol{K}$. Then,

$$\left\|\boldsymbol{K} - \left(\frac{\rho_1^2}{d}\mathbf{X}^T\mathbf{X} + \rho_2^2\mathbf{I}\right)\right\|_{\text{op}} \leq c\left(\mu^3 n + \mu + \mu\frac{\|\mathbf{X}\|_{\text{op}}^2}{d}\right), \tag{87}$$

where c is a universal constant.

Proof Note that $K_{ij} = \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{w}^T\mathbf{x}_i)\sigma(\mathbf{w}^T\mathbf{x}_j)]$. For $i \neq j$, we have that $K_{ij} = \eta_1\left(\frac{\|\mathbf{x}_i\|^2}{d}, \frac{\|\mathbf{x}_j\|^2}{d}, \frac{\mathbf{x}_i^T\mathbf{x}_j}{d}\right)$ where η_1 and for $i = j$, we have that $K_{ii} = \eta_2\left(\frac{\|\mathbf{x}_i\|^2}{d}\right)$. Where η_1 and η_2 are defined in assumption A4. Note that by oddness of the activation function

$$\eta_1(1,1,0) = 0, \qquad \eta_2(1) = \mathbb{E}\sigma^2(g) \tag{88}$$

$$\nabla\eta_1(1,1,0) = (0,0,\mathbb{E}[g\sigma(g)]^2), \tag{89}$$

where g is a standard normal. We also note that the hessian of η_1

$$H_{\eta_1}(1,1,0) = \begin{bmatrix} 0 & 0 & -\mathbb{E}[g\sigma(g)]^2 \\ 0 & 0 & 0 \\ -\mathbb{E}[g\sigma(g)]^2 & 0 & 0 \end{bmatrix}. \tag{90}$$

Then, by the mean value theorem and assumption A4 we have that

$$|K_{ij} - K'_{ij}| \leq \begin{cases} c\mu^3 & i \neq j \\ c\mu & i = j \end{cases}, \tag{91}$$

where

$$K'_{ij} = \begin{cases} \frac{\mathbf{x}_i^T\mathbf{x}_j}{d}[\mathbb{E}[g\sigma(g)]]^2\left(1 - \frac{\|\mathbf{x}_i\|^2}{d}\right) & i \neq j \\ \mathbb{E}[\sigma^2(g)] & i = j \end{cases}. \tag{92}$$

From this we conclude that

$$\left\|\boldsymbol{K} - \boldsymbol{K}'\right\|_{\text{op}} \leq c(\mu^3 n + \mu). \tag{93}$$

It can also straightforwardly be checked that $\left\|\boldsymbol{K}' - \left(\frac{\rho_1^2}{d}\mathbf{X}^T\mathbf{X} + \rho_2^2\mathbf{I}\right)\right\|_{\text{op}} \leq c\mu\frac{\|\mathbf{X}\|_{op}^2}{d}$, from which the desired result can be obtained. ■

The second intermediate result is shown in the following theorem.

Theorem 16 *Suppose that σ is odd with bounded derivatives and assumption A4 holds. Moreover, the set $\{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^n$ satisfies:*

$$\sup_{i,j}\left|\frac{\mathbf{x}_i^T\mathbf{x}_j}{d} - \delta_{ij}\right| \leq \frac{\text{polylog } n}{\sqrt{n}}. \tag{94}$$

Define $\mathbf{z}_i = \sigma(\mathbf{W}\mathbf{x}_i)$ where $\mathbf{W} \in \mathbb{R}^{p\times d}$ has independent row distributed by $\mathcal{N}(0, \frac{1}{d}\mathbf{I})$. Then, with a probability higher than $1 - n^{-10}$ it holds that[3]:

$$\sup_{i,j}\left|\frac{\mathbf{z}_i^T\mathbf{z}_j}{p} - \delta_{ij}\right| \leq \frac{\text{polylog } n}{\sqrt{n}}. \tag{95}$$

3. The exponent is arbitrary and can be replaced by any other number

Proof We note that

$$\frac{1}{p}\mathbf{z}_i^T\mathbf{z}_j = \frac{1}{p}\sum_r \sigma(\mathbf{w}_r^T\mathbf{x}_i)\sigma(\mathbf{w}_r^T\mathbf{x}_j), \tag{96}$$

and by the assumptions $\sigma(\mathbf{w}_r^T\mathbf{x}_i)\sigma(\mathbf{w}_r^T\mathbf{x}_j)$ are i.i.d and sub-exponential. Hence, there exists a constant c such that for every $t = o(\sqrt{n})$:

$$\Pr\left[\left|\frac{1}{p}\mathbf{z}_i^T\mathbf{z}_j - \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{w}^T\mathbf{x}_i)\sigma(\mathbf{w}^T\mathbf{x}_j)]\right| > \frac{t}{\sqrt{n}}\right] \leq 2e^{-ct}. \tag{97}$$

In particular, we may take $t = c\log n$ for a sufficiently large c, which by the union bound leads to

$$\sup_{i,j}\left|\frac{1}{p}\mathbf{z}_i^T\mathbf{z}_j - \mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{w}^T\mathbf{x}_i)\sigma(\mathbf{w}^T\mathbf{x}_j)]\right| < \frac{c\log n}{\sqrt{n}} \tag{98}$$

with the desired probability. On the other hand $\mathbb{E}_{\mathbf{w}}[\sigma(\mathbf{w}^T\mathbf{x}_i)\sigma(\mathbf{w}^T\mathbf{x}_j)]$ equals either $\eta_1\left(\frac{\|\mathbf{x}_i\|^2}{d}, \frac{\|\mathbf{x}_j\|^2}{d}, \frac{\mathbf{x}_i^T\mathbf{x}_j}{d}\right)$ for $i \neq j$ or $\eta_2\left(\frac{\|\mathbf{x}_i\|^2}{d}\right)$ for $i = j$. Then by assumption A4 the result holds. ■

C.1. Proof of theorem 6

To prove the theorem we need to show two properties hold. Firstly,

$$\max_{i,j}\left|\frac{\mathbf{z}_i^T\mathbf{z}_j}{p} - \delta_{ij}\right| \leq \frac{\text{polylog } n}{\sqrt{n}}. \tag{99}$$

This has been shown by theorem 16. Next, we need to show that

$$\|\mathbf{Z}\| \leq c\sqrt{n}. \tag{100}$$

For this we note that the rows of $\mathbf{Z}$ are independent. Moreover, by lemma 10 and the assumptions, each row is $c-$sub-Gaussian for a constant c. Finally, by theorem 15, we have that

$$\|\boldsymbol{K}\|_{\text{op}} \leq \frac{\text{polylog } n}{\sqrt{n}} + \left\|\frac{\rho_1^2}{d}\mathbf{X}\mathbf{X}^T + \rho_2^2\mathbf{I}\right\|_{\text{op}} \leq c. \tag{101}$$

Then by lemma 11 the result follows.

C.2. Proof of theorem 7

The proof is by induction. For $l = 0$, the claim is trivially holds. For a given l, note that $\mathbf{x}_i^{(l)} = \phi(\mathbf{x}_i^{(l-1)}, \mathbf{W}^{(l)})$. Furthermore, $\{\mathbf{x}_i^{(l-1)}\}_{i=1}^n$ is regular with a probability higher than $1 - n^{-10}$ and hence by lemma 10, each row of $\tilde{\mathbf{X}}^{(l)} = [\tilde{\mathbf{x}}_i\ \tilde{\mathbf{x}}_2\ \cdots \tilde{\mathbf{x}}_n]$ is $c-$sub-Gaussian. Moreover, by theorem 15, we have

$$\|\boldsymbol{K}\|_{\text{op}} \leq \frac{\text{polylog } n}{\sqrt{n}} + \left\|\frac{\rho_1^2}{d}\mathbf{X}^T\mathbf{X} + \rho_2^2\mathbf{I}\right\|_{\text{op}} \leq c \tag{102}$$

and hence by assumption the first condition for Theorem 5 holds true with a probability higher than $1 - n^{-10}$. As a result, defining

$$\tilde{\mathbf{g}}_i^{(l)'} = \begin{bmatrix} \rho_1 \mathbf{W}^{(l)} \mathbf{x}_i^{(l-1)} + \rho_2 \mathbf{h}_i^{(l)} \\ \mathbf{v}_i^{(l)} \end{bmatrix}, \tag{103}$$

then theorem 5 holds for $\{\tilde{\mathbf{x}}_i^{(l)}\}_{i=1}^n$ and $\{\tilde{\mathbf{g}}_i^{(l)'}\}_{i=1}^n$. Denoting the optimal value and the optimal point for the latter by $L', \hat{\boldsymbol{\theta}}'$, we note that

$$\left|\mathbb{E}_{\mathbf{W}^{(l)}}[\psi(L(\boldsymbol{F}))] - \mathbb{E}_{\mathbf{W}^{(l)}}[\psi(L')]\right| \leq \frac{\text{polylog } n}{\sqrt{n}} \tag{104}$$

with probability $1 - n^{-10}$. Note that $L(\boldsymbol{F})$ and L' are bounded, hence:

$$\left|\mathbb{E}[\psi(L(\boldsymbol{F}))] - \mathbb{E}[\psi(L')]\right| \leq \frac{\text{polylog } n}{\sqrt{n}}, \tag{105}$$

Where $\mathbb{E}[\psi(L')] = \mathbb{E}[\mathbb{E}[\psi(L;)|\mathbf{W}^{(l)}]]$. On the other hand,

$$\mathbf{D}\tilde{\mathbf{g}}_i^{(l)'} = \underbrace{\mathbf{D}\begin{bmatrix} \rho_1 \mathbf{W}^{(l)} & \rho_2 \mathbf{I} & \\ 0 & 0 & \mathbf{I} \end{bmatrix}}_{\mathbf{D}'} \begin{bmatrix} \mathbf{x}_i^{(l-1)} \\ \mathbf{h}_i^{(l)} \\ \mathbf{v}_i^{(l)} \end{bmatrix}. \tag{106}$$

Now we observe that with probability higher than $1 - e^{-cn}$ it holds that $\|\mathbf{D}'\| \leq C$ and hence we may invoke the induction hypothesis for layer $l-1$ with $\mathbf{D}'$ and $\mathbf{v}_i^{(l-1)} = [\mathbf{h}_i^{(l)} \; \mathbf{v}_i^{(l)}]$ to conclude that

$$\left|\mathbb{E}[\psi(L(\mathbf{G}))|\mathbf{W}^{(l)}] - \mathbb{E}[\psi(L')|\mathbf{W}^{(l)}]\right| \leq \frac{\text{polylog } n}{\sqrt{n}}, \tag{107}$$

with a probability higher than $1 - e^{-cn}$. Again using the fact that the optimal value is bounded, we conclude that

$$\left|\mathbb{E}[\psi(L(\mathbf{G}))] - \mathbb{E}[\psi(L')]\right| \leq \frac{\text{polylog } n}{\sqrt{n}}. \tag{108}$$

Which concludes the claim for part 1. Part 2 is proven exactly by the same argument.

Appendix D. Proof of Theorem 8

To prove Theorem 8, our goal is to make use of the CGMT (theorem 12) to obtain an alternative optimization problem to (24). Upon simplification we note that this problem relies entirely upon $\mathbf{R}^{(L)}$ and note that is can once again be expressed as another CGMT style optimization. Applying the CGMT again results in a problem dependent upon $\mathbf{R}^{(L-1)}$. Repeating the processes iteratively eventually results in the alternative optimization problem given in (26). We adopt the same process for a recursive CGMT solution as in (Bosch et al., 2022), and follow the direction of their proof.

To begin this processes we first recall the definition of problem P_2 given in (24). We fix $\mathbf{W}^{(1)}, \ldots, \mathbf{W}^{(L)}$ and make a change of variables. Recalling the definition of $\mathbf{y}$, given in (23), we introduce the error vector $\mathbf{e} = \boldsymbol{\theta} - \boldsymbol{\theta}^*$:

$$P_2 = \min_{\mathbf{e} \in \mathbb{R}^{p_L}} \frac{1}{2n} \left\| \boldsymbol{\nu} - \frac{1}{\sqrt{p_L}} \tilde{\mathbf{X}}^{(L)} \mathbf{e} \right\|_2^2 + R(\mathbf{e} + \boldsymbol{\theta}^*). \tag{109}$$

We now recall that the rows $\tilde{\mathbf{x}}_i^{(L)}$ of $\tilde{\mathbf{X}}^{(L)}$ are i.i.d normally distributed with covariance $\mathbf{R}^{(L)}$. As such we can express $\tilde{\mathbf{X}}^{(L)} = \mathbf{U}^{(L)} \left(\mathbf{R}^{(L)}\right)^{1/2}$ where $\mathbf{U}^{(L)} \in \mathbb{R}^{n \times p_L}$ and has i.i.d normal Gaussian entries and $\mathbf{R}^{(L)}$ is given by

$$\mathbf{R}^{(0)} = \mathbf{I} \qquad \mathbf{R}^{(l)} = \rho_{1,l}^2 \mathbf{W}^{(l)} \mathbf{R}^{(l-1)} \mathbf{W}^{(l)T} + \rho_{2,l}^2 \mathbf{I} \quad 1 \leq l \leq L. \tag{110}$$

For the sake of notational simplicity we will express $\left(\mathbf{R}^{(L)}\right)^{1/2}$ as $\mathbf{R}^{(L)/2}$ when there is no chance of confusion.

Next we make use of the Legendre transform of the 2-norm. We obtain

$$P_2 = \min_{\mathbf{e} \in \mathbb{R}^{p_L}} \max_{\boldsymbol{\lambda} \in \mathbb{R}^n} \frac{1}{n} \boldsymbol{\lambda}^T \boldsymbol{\nu} - \frac{1}{n\sqrt{p_L}} \boldsymbol{\lambda}^T \mathbf{U}^{(L)} \mathbf{R}^{(L)/2} \mathbf{e} - \frac{1}{2n} \|\boldsymbol{\lambda}\|_2^2 + R(\mathbf{e} + \boldsymbol{\theta}^*). \tag{111}$$

We note that the problem is now in the correct form to apply the CGMT. However, the CGMT requires that the optimizations over $\mathbf{e}$ and $\boldsymbol{\lambda}$ are over compact and convex sets. In the subsequent lemmas we show that we can restrict the problem to compact and convex subsets of $\mathbb{R}^{p_L}$ and $\mathbb{R}^n$.

Firstly, we show that $\mathbf{R}^{(l)}$ for all $0 \leq l \leq L$ can be bounded above by a constant in operator norm with high probability.

Lemma 17 *Let $\mathbf{R}^{(l)}$ be defined as in (110), then for each $0 \leq l \leq L$ there exists a constant $C_{\mathbf{R}^{(l)}}$ such that*

$$\Pr\left(\left\| \mathbf{R}^{(l)} \right\|_2 < C_{\mathbf{R}^{(l)}} \right) \geq 1 - \sum_{j=1}^{l} 2e^{-cp_l}. \tag{112}$$

For some universal constant $c > 0$. By $\|\cdot\|_2$ we mean the spectral norm.

Proof The proof is by induction. For $\mathbf{R}^{(0)} = \mathbf{I}$ it is clear that $\left\|\mathbf{R}^{(0)}\right\|_2 = 1$. Now assume that the following event holds

$$\left\{ \left\| \mathbf{R}^{(l-1)} \right\|_2 \leq C_{\mathbf{R}^{(l-1)}} \right\}, \tag{113}$$

then by the definition of $\mathbf{R}^{(l)}$ we have that

$$\begin{aligned} \left\| \mathbf{R}^{(l)} \right\|_2 = \left\| \rho_{1,l}^2 \mathbf{W}^{(l)} \mathbf{R}^{(l-1)} \mathbf{W}^{(l)T} + \rho_{2,l}^2 \mathbf{I} \right\| &\leq \rho_{1,l}^2 \left\| \mathbf{W}^{(l)} \right\|_2^2 \left\| \mathbf{R}^{(l-1)} \right\|_2 + \rho_{2,l}^2 \\ &\leq \rho_{1,l}^2 C_{\mathbf{R}^{(l-1)}} \left\| \mathbf{W}^{(l)} \right\|_2^2 + \rho_{2,l}^2 \end{aligned} \tag{114}$$

Now we recall that the elements of $\mathbf{W}^{(l)}$ are i.i.d normally distributed with variance $\frac{1}{p_{l-1}}$. Standard results from Random matrix theory (see for example (Vershynin, 2018)[corollary 7.3.3]) demonstrate that

$$\Pr\left(\left\|\mathbf{W}^{(l)}\right\|_2 \geq 1+\sqrt{p_l/p_{l-1}}+t\right) \leq 2e^{cp_{l-1}t^2}. \tag{115}$$

We choose $t=\sqrt{p_l/p_{l-1}}$ from which we obtain

$$\Pr\left(\left\|\mathbf{W}^{(l)}\right\|_2 \geq 1+2\sqrt{p_l/p_{l-1}}\right) \leq 2e^{cp_l}. \tag{116}$$

As such we can choose

$$\boldsymbol{C}_{\mathbf{R}^{(l)}} = \rho_{1,l}^2 C_{\mathbf{R}^{(l-1)}}(1+2\sqrt{p_l/p_{l-1}})^2+\rho_{2,l}^2 \tag{117}$$

Now we note that the probability of the event (113) hols true with probability

$$\Pr\left(\left\|\mathbf{W}^{(1)}\right\|_2 < 1+2\sqrt{p_1/p_0}, \cdots, \left\|\mathbf{W}^{(l-1)}\right\|_2 < 1+2\sqrt{p_{l-1}/p_{l-2}}\right) \geq 1-\sum_{j}^{l-1} 2e^{cp_j} \tag{118}$$

where we have made use of the union bound. As such we can say that with high probability $\left\|\mathbf{R}^{(l)}\right\|_2$ is bounded. ■

Next, we show that the optimizations over $\mathbf{e}$ and $\boldsymbol{\lambda}$ can be restricted to compact sets

Lemma 18 *Consider the following two optimization problems, which correspond to the problem P_2 and the alternative problem after applying the CGMT:*

$$P_{2,1} = \min_{\mathbf{e}\in\mathbb{R}^{p_L}} \max_{\boldsymbol{\lambda}\in\mathbb{R}^n} \frac{1}{n}\boldsymbol{\lambda}^T\mathbf{u} - \frac{1}{n\sqrt{p_L}}\boldsymbol{\lambda}^T\mathbf{U}^{(L)}\mathbf{R}^{(L)/2}\mathbf{e} - \frac{1}{2n}\|\boldsymbol{\lambda}\|_2^2 + R(\mathbf{e}+\boldsymbol{\theta}^*), \tag{119}$$

$$P_{2,2} = \min_{\mathbf{e}\in\mathbb{R}^{p_L}} \max_{\boldsymbol{\lambda}\in\mathbb{R}^n} \frac{1}{n}\boldsymbol{\lambda}^T\mathbf{u} - \frac{1}{n\sqrt{p_L}}\|\boldsymbol{\lambda}\|_2\,\mathbf{g}^T\mathbf{R}^{(L)/2}\mathbf{e} - \frac{1}{n\sqrt{p_L}}\left\|\mathbf{R}^{(L)/2}\mathbf{e}\right\|_2\mathbf{h}^T\boldsymbol{\lambda} - \frac{1}{2n}\|\boldsymbol{\lambda}\|_2^2 + R(\mathbf{e}+\boldsymbol{\theta}^*). \tag{120}$$

where $\mathbf{g}\in\mathbb{R}^{p_L}, \mathbf{h}\in\mathbb{R}^n$ are standard normal vectors. We define $\hat{\mathbf{e}}_1$ and $\hat{\mathbf{e}}_2$ to be the optimal solutions of $P_{2,1}$ and $P_{2,2}$ respectively. Furthermore, let $\hat{\boldsymbol{\lambda}}_1(\mathbf{e}), \hat{\boldsymbol{\lambda}}_2(\mathbf{e})$ be the optimal solutions of the inner optimization of $P_{2,1}$ and $P_{2,2}$ respectively as functions of $\mathbf{e}$. Let R be μ-strongly convex and let $\|\nabla R(\boldsymbol{\theta}^*)\| = \mathcal{O}(\sqrt{p_L})$. Then there exist positive constants $C_{\mathbf{e}}$ and $C_{\boldsymbol{\lambda}}$ that depend only on μ such that

- *The solutions $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2$ are*

$$\lim_{p_L\to\infty}\Pr\left(\max\{\|\hat{\mathbf{e}}_1\|, \|\hat{\mathbf{e}}\|_2\} \leq C_{\mathbf{e}}\sqrt{p_L}\right) = 1 \tag{121}$$

- *and*

$$\lim_{n\to\infty}\Pr\left(\sup_{\mathbf{e}:\|\mathbf{e}\|\leq C_{\mathbf{e}}\sqrt{m}} \max\{\left\|\hat{\boldsymbol{\lambda}}_1\right\|, \left\|\hat{\boldsymbol{\lambda}}_2\right\|\} \leq C_{\boldsymbol{\lambda}}\sqrt{n}\right) = 1 \tag{122}$$

Proof We recall that R is μ strongly convex, and we let the function $B(\mathbf{e}) = R(\mathbf{e} + \boldsymbol{\theta}^*)$. Solving for $\boldsymbol{\lambda}$ in both optimizations, we may expressed the resultant optimization over $\mathbf{e}$ as

$$\min_{\mathbf{e}} F_i(\mathbf{e}) \qquad i = 1, 2. \tag{123}$$

Such that $F_i(\mathbf{e})$ is the optimal value over the parameter $\boldsymbol{\lambda}$. Next, we note if we set $\boldsymbol{\lambda} = \mathbf{0}$, both optimizations yield $F_i(\mathbf{e}) \geq R(\mathbf{e})$. Then we note that

$$B(\mathbf{e}) \geq B(\mathbf{0}) + \mathbf{d}^T\mathbf{e} + \mu \left\|\mathbf{e}\right\|_2^2, \tag{124}$$

from the strong convexity of R, where $\mathbf{d} = \nabla B(\mathbf{0}) = \nabla R(\boldsymbol{\theta}^*)$. We note that by assumption $\|\mathbf{d}\| = \mathcal{O}(\sqrt{p_L})$.

For the first optimization P_1, we note that

$$F(\mathbf{0}) = B(\mathbf{0}) + \frac{1}{2n}\left\|\boldsymbol{\nu}\right\|_2^2. \tag{125}$$

From this we note that for the optimal solution $\hat{\mathbf{e}}$ we have

$$B(\mathbf{0}) + \frac{1}{2n}\left\|\boldsymbol{\nu}\right\|_2^2 = F(\mathbf{0}) \geq F(\hat{\mathbf{e}}_1) \geq R(\mathbf{0}) + \mathbf{d}^T\hat{\mathbf{e}}_1 + \mu\left\|\hat{\mathbf{e}}\right\|_2^2, \tag{126}$$

from which we obtain

$$\mu\left\|\hat{\mathbf{e}}_1 + \frac{1}{\mu}\mathbf{d}\right\| \leq \frac{1}{2n}\left\|\boldsymbol{\nu}\right\|_2^2 + \frac{1}{4\mu}\left\|\mathbf{d}\right\|_2^2. \tag{127}$$

As such

$$\left\|\hat{\mathbf{e}}_1\right\|_2 \leq \left\|\frac{1}{\mu}\mathbf{d}\right\|_2 + \sqrt{\frac{1}{2n\mu}\left\|\boldsymbol{\nu}\right\|_2^2 + \frac{1}{\mu^2}\left\|\mathbf{d}\right\|_2^2} \tag{128}$$

We recall that from standard random matrix theory (Papaspiliopoulos, 2020)[Theorem 2.8.1] we know that $\|\boldsymbol{\nu}\|_2^2 \leq cn$ for some n with high probability. We may therefore observe that exists a constant $C_{\mathbf{e}_1}$ such that

$$\lim_{p_L \to \infty} \Pr(\left\|\hat{\mathbf{e}}_1\right\|_2 \geq C_{\mathbf{e}_1}\sqrt{p_L}) = 0. \tag{129}$$

We can now consider problem (119). We make use of the same strategy in this case. We note that, when we let $\beta = \|\boldsymbol{\lambda}\|$, the optimization over $\boldsymbol{\lambda}$ with fixed norm can be solved to obtain:

$$F(\mathbf{e}) = \max_{\beta \geq 0} \frac{\beta}{n}\left\|\boldsymbol{\nu} - \frac{1}{\sqrt{p_L}}\left\|\mathbf{R}^{(L)/2}\mathbf{e}\right\|_2 \mathbf{g}\right\| - \frac{\beta}{n\sqrt{m}}\mathbf{h}^T\mathbf{R}^{(L)/2}\mathbf{e} - \frac{\beta^2}{2nm} + B(\mathbf{e}). \tag{130}$$

We note that this optimization is constrained to the set $\beta \geq 0$, as such dropping the constraint can only increase the optimal value. Dropping the constrains results in a quadratic optimizations which may be solved. We obtain the following inequality

$$F(\mathbf{e}) \leq B(\mathbf{e}) + \frac{1}{2n}\left(\left\|\boldsymbol{\nu} - \frac{1}{\sqrt{p_L}}\left\|\mathbf{R}^{(L)/2}\mathbf{e}\right\|_2 \mathbf{g}\right\|_2 - \frac{\beta}{\sqrt{m}}\mathbf{h}^T\mathbf{R}^{(L)/2}\mathbf{e}\right)^2, \tag{131}$$

and in particular

$$F(\mathbf{0}) \leq B(\mathbf{0}) + \frac{1}{2n} \|\mathbf{e}\|_2^2 . \tag{132}$$

Now making use of the same inequality as in equation (126) from which we may find that

$$\|\hat{\mathbf{e}}_2\|_2 \leq \left\| \frac{1}{\mu} \mathbf{d} \right\|_2 + \sqrt{\frac{1}{2n\mu} \|\boldsymbol{\nu}\|_2^2 + \frac{1}{\mu^2} \|\mathbf{d}\|_2^2}. \tag{133}$$

As such we can demonstrate that

$$\lim_{p_L \to \infty} \Pr(\|\hat{\mathbf{e}}_2\|_2 \geq C_{\mathbf{e}_2} \sqrt{p_L}) = 0. \tag{134}$$

We let $C_{\mathbf{e}} = \max(C_{\mathbf{e}_1}, C_{\mathbf{e}_2})$, and we make use of this constant to define $A_{\mathbf{e}} = \{\mathbf{e} \in \mathbb{R}^{p_L} | \ \|\mathbf{e}\|_2 \leq C_{\mathbf{e}} \sqrt{m}\}$

Making use of the optimality condition of the inner optimization in equation (119), we see that

$$\hat{\boldsymbol{\lambda}}_1(\mathbf{e}) = \boldsymbol{\nu} - \frac{1}{\sqrt{m}} \mathbf{U} \mathbf{R}^{(L)/2} \mathbf{e}. \tag{135}$$

As such, for all $\mathbf{e} \in A_{\mathbf{e}}$

$$\left\| \hat{\boldsymbol{\lambda}}_1(\mathbf{e}) \right\|_2 \leq \|\boldsymbol{\nu}\|_2 + \left\| \frac{1}{\sqrt{m}} \mathbf{U} \mathbf{R}^{(L)/2} \right\|_2 \|\mathbf{e}\|_2 \leq \|\boldsymbol{\nu}\|_2 + \left\| \frac{1}{\sqrt{m}} \mathbf{U} \right\|_2 \left\| \mathbf{R}^{(L)/2} \right\|_2 \|\mathbf{e}\|_2 . \tag{136}$$

We can then note by lemma 17 that $\left\|\mathbf{R}^{(L)/2}\right\|_2$ is bounded. Furthermore, by standard random matrix theory results we can conclude that $\left\| \frac{1}{\sqrt{m}} \mathbf{U} \right\|_2 < C$ for some constant C with high probability. Then, using the same arguments as above, we can conclude that t here must exist a constant $C_{\boldsymbol{\lambda}_1}$ such that for all $\mathbf{e} \in A_{\mathbf{e}}$:

$$\lim_{n \to \infty} \Pr \left(\sup_{\mathbf{e} \in A_{\mathbf{e}}} \left\| \hat{\boldsymbol{\lambda}}_1(\mathbf{e}) \right\|_2 \geq C_{\boldsymbol{\lambda}_1} \sqrt{n} \right) = 0 \tag{137}$$

Finally, consider the optimality condition over β of problem 120 we see that for all $\mathbf{e} \in A_{\mathbf{e}}$ that

$$\begin{aligned} \hat{\beta} = \left\| \hat{\boldsymbol{\lambda}}_2(\mathbf{e}) \right\|_2 &= \left\| \boldsymbol{\nu} - \frac{1}{\sqrt{m}} \left\| \mathbf{R}^{(L)/2} \mathbf{e} \right\|_2 \mathbf{g} \right\|_2 - \frac{1}{\sqrt{m}} \mathbf{R}^{(L)/2} \mathbf{h} \\ &\leq \|\boldsymbol{\nu}\|_2 + \frac{1}{\sqrt{m}} \|\mathbf{g}\|_2 \left\| \mathbf{R}(L)/2 \right\|_2 \|\mathbf{e}\|_2 + \frac{1}{\sqrt{m}} \left\| \mathbf{R}^{(L)/2} \right\|_2 \|\mathbf{h}\|_2 \end{aligned} \tag{138}$$

With high probability we note that $\|\boldsymbol{\nu}\|_2 < C\sqrt{n}$, $\|\mathbf{g}\|_2 < C\sqrt{n}$ and $\|\mathbf{h}\| < C\sqrt{p_L}$. As such we can find a constant $C_{\boldsymbol{\lambda}_2}$ with

$$\lim_{n \to \infty} \Pr \left(\sup_{\mathbf{e} \in A_{\mathbf{e}}} \left\| \hat{\boldsymbol{\lambda}}_2(\mathbf{e}) \right\|_2 \geq C_{\boldsymbol{\lambda}_2} \sqrt{n} \right) = 0. \tag{139}$$

Choosing $C_{\boldsymbol{\lambda}} = \max(C_{\boldsymbol{\lambda}_1}, C_{\boldsymbol{\lambda}_2})$, the proof is complete. ■

Making use of this lemma we can define the sets $S_1 = \{\mathbf{e}| \ \|\mathbf{e}\| \leq C_{\mathbf{e}}\sqrt{m}\}$ and $S_2 = \{\boldsymbol{\lambda}| \ \|\boldsymbol{\lambda}\| \leq C_{\boldsymbol{\lambda}}\sqrt{n}\}$ and note that these sets are compact and convex. We can with high probability restrict ourselves to the problem

$$P_2' = \min_{\mathbf{e}\in S_1}\max_{\boldsymbol{\lambda}\in S_2} \frac{1}{n}\boldsymbol{\lambda}^T\boldsymbol{\nu} - \frac{1}{n\sqrt{p_L}}\boldsymbol{\lambda}^T\mathbf{U}^{(L)}\mathbf{R}^{(L)/2}\mathbf{e} - \frac{1}{2n}\|\boldsymbol{\lambda}\|_2^2 + R(\mathbf{e}+\boldsymbol{\theta}^*) \tag{140}$$

and note that the optimal value of P_2' will be close that of P_2. We now satisfy the conditions for applying the CMGT. Applying it we obtain the following problem:

$$A_2 = \min_{\mathbf{e}\in S_1}\max_{\boldsymbol{\lambda}\in S_2} \frac{1}{n}\boldsymbol{\lambda}^T\boldsymbol{\nu} - \frac{1}{n\sqrt{p_L}}\|\boldsymbol{\lambda}\|_2\,\mathbf{g}^T\mathbf{R}^{(L)/2}\mathbf{e} - \frac{1}{n\sqrt{p_L}}\left\|\mathbf{R}^{(L)/2}\mathbf{e}\right\|_2\mathbf{h}^T\boldsymbol{\lambda} - \frac{1}{2n}\|\boldsymbol{\lambda}\|_2^2 + R(\mathbf{e}+\boldsymbol{\theta}^*). \tag{141}$$

Where $\mathbf{g}\in\mathbb{R}^{p_L}, \mathbf{h}\in\mathbb{R}^n$ have elements that are i.i.d standard normals. By theorem 12 we know that the optimal values of A_2 and P_2' will be asymptotically equal if A_2 converges to a finite value. Next we let $\beta = \frac{1}{\sqrt{n}}\|\boldsymbol{\lambda}\|$. We note that $0\leq\beta\leq\beta_{max}$, where $\beta_{max}\in\mathbb{R}$ is some constant, whose value can be chosen arbitrarily larger than $C_{\boldsymbol{\lambda}}$. We can now solve the optimization over the vector $\boldsymbol{\lambda}$ fixing its length to β. We obtain

$$A_2 = \min_{\mathbf{e}\in S_1}\max_{0\leq\beta\leq\beta_{max}} \beta\left\|\frac{1}{\sqrt{n}}\boldsymbol{\nu} - \frac{1}{\sqrt{np_L}}\left\|\mathbf{R}^{(L)/2}\mathbf{e}\right\|\mathbf{h}\right\|_2 - \frac{\beta}{\sqrt{np_L}}\mathbf{g}^T\mathbf{R}^{(L)/2}\mathbf{e} - \frac{\beta^2}{2} + R(\mathbf{e}+\boldsymbol{\theta}^*). \tag{142}$$

Now we note that the first term in the $2-$norm concentrates as n grows large. We prove this in the following lemma

Lemma 19 *Let A be given by*

$$A(\mathbf{e},\beta) = \beta\left\|\frac{1}{\sqrt{n}}\boldsymbol{\nu} - \frac{1}{\sqrt{np_L}}\left\|\mathbf{R}^{(L)/2}\mathbf{e}\right\|\mathbf{h}\right\| - \frac{\beta}{\sqrt{np_L}}\mathbf{g}^T\mathbf{R}^{(L)/2}\mathbf{e} - \frac{\beta^2}{2} + R(\mathbf{e}+\boldsymbol{\theta}^*). \tag{143}$$

Let $\tilde{A}(\mathbf{e},\beta)$ be given by

$$\tilde{A}(\mathbf{e},\beta) = \beta\sqrt{\sigma_{\boldsymbol{\nu}}^2 + \frac{1}{p_L}\mathbf{e}^T\mathbf{R}^{(L)}\mathbf{e}} - \frac{\beta}{\sqrt{np_L}}\mathbf{g}^T\mathbf{R}^{(L)/2}\mathbf{e} - \frac{\beta^2}{2} + R(\mathbf{e}+\boldsymbol{\theta}^*). \tag{144}$$

Then, there exists positive constants C, c such that for any $\epsilon>0$:

$$\Pr\left(\sup_{\mathbf{e}\in S_1, 0\leq\beta\leq\beta_{max}} |A(\mathbf{e},\beta) - \tilde{A}(\mathbf{e},\beta)| \geq \epsilon\right) \leq Ce^{-cn\epsilon} \tag{145}$$

Proof We note that $A(\mathbf{e},\beta)$ can be expressed as

$$A = \beta\sqrt{\frac{1}{n}\|\boldsymbol{\nu}\|_2^2 + \frac{1}{np_L}\left\|\mathbf{R}^{(L)/2}\mathbf{e}\right\|_2^2\|\mathbf{h}\|_2^2 - \frac{2}{n\sqrt{p_L}}\left\|\mathbf{R}^{(L)/2}\mathbf{e}\right\|_2\boldsymbol{\nu}^T\mathbf{h}} - \frac{\beta}{\sqrt{np_L}}\mathbf{g}^T\mathbf{R}^{(L)/2}\mathbf{e} - \frac{\beta^2}{2} + R(\mathbf{e}+\boldsymbol{\theta}^*) \tag{146}$$

Or equivalently

$$A = \beta \left[\left(\frac{1}{n} \left\| \boldsymbol{\nu} \right\|_2^2 - \sigma_{\boldsymbol{\nu}}^2 \right) + \sigma_{\boldsymbol{\nu}}^2 + \frac{1}{p_L} \left\| \mathbf{R}^{(L)/2} \mathbf{e} \right\|_2^2 \left(\frac{1}{n} \left\| \mathbf{h} \right\|_2^2 - \right) + \frac{1}{p_L} \left\| \mathbf{R}^{(L)/2} \mathbf{e} \right\|_2^2 \right.$$
$$\left. - \frac{2}{\sqrt{p_L}} \left\| \mathbf{R}^{(L)/2} \mathbf{e} \right\|_2 \frac{\boldsymbol{\nu}^T \mathbf{h}}{n} \right]^{1/2} - \frac{\beta}{\sqrt{n p_L}} \mathbf{g}^T \mathbf{R}^{(L)/2} \mathbf{e} - \frac{\beta^2}{2} + R(\mathbf{e} + \boldsymbol{\theta}^*)$$
$$\leq \bar{A} + \beta \sqrt{\delta} \leq \bar{A} + \beta_{max} \sqrt{\bar{\delta}} \tag{147}$$

where

$$\delta = \left(\frac{1}{n} \left\| \boldsymbol{\nu} \right\|_2^2 - \sigma_{\boldsymbol{\nu}}^2 \right) + \frac{1}{p_L} \left\| \mathbf{R}^{(L)/2} \mathbf{e} \right\|_2^2 \left(\frac{1}{n} \left\| \mathbf{h} \right\|_2^2 - \right) - \frac{2}{\sqrt{p_L}} \left\| \mathbf{R}^{(L)/2} \mathbf{e} \right\|_2 \frac{\boldsymbol{\nu}^T \mathbf{h}}{n}$$
$$\leq \left(\frac{1}{n} \left\| \boldsymbol{\nu} \right\|_2^2 - \sigma_{\boldsymbol{\nu}}^2 \right) + C_{\mathbf{e}}^2 C_{\mathbf{R}^{(L)}} \left(\frac{1}{n} \left\| \mathbf{h} \right\|_2^2 - \right) - 2 \sqrt{C_{\mathbf{R}^{(L)}}} C_{\mathbf{e}} \left| \frac{\boldsymbol{\nu}^T \mathbf{h}}{n} \right| \stackrel{def}{=} \bar{\delta}. \tag{148}$$

From the lemmas above we note that $C_{\mathbf{R}^{(L)}}$ and $C_{\mathbf{e}}$ are universal constants. Furthermore, it can be readily observed that $\Pr(|\bar{\delta}| \geq \epsilon) \leq C e^{-cn\epsilon}$ for some constants $C, c > 0$. As such, we see that

$$\Pr \left(\sup_{\mathbf{e} \in S_1, 0 \leq \beta \leq \beta_{max}} |A(\mathbf{e}, \beta) - \bar{A}(\mathbf{e}, \beta)| \geq \epsilon \right) \leq$$
$$\Pr \left(\sup_{\mathbf{e} \in S_1, 0 \leq \beta \leq \beta_{max}} |\delta \beta| \geq \epsilon \right) \leq \Pr \left(|\beta_{max} \bar{\delta}| \geq \epsilon \right) \leq C e^{-cn\epsilon} \tag{149}$$

For some constants $C, c > 0$. ■

By means of this lemma we can, with high probability, consider the following problem

$$\bar{A}_2 = \min_{\mathbf{e} \in S_1} \max_{0 \leq \beta \leq \beta_{max}} \beta \sqrt{\sigma_{\boldsymbol{\nu}}^2 + \frac{1}{p_L} \mathbf{e}^T \mathbf{R}^{(L)} \mathbf{e}} - \frac{\beta}{\sqrt{n p_L}} \mathbf{g}^T \mathbf{R}^{(L)/2} \mathbf{e} - \frac{\beta^2}{2} + R(\mathbf{e} + \boldsymbol{\theta}^*). \tag{150}$$

We now note that this optimization problem is convex in $\mathbf{e}$ and concave in β. Furthermore, both optimizations are over convex sets. As such we can interchange the order of min and max

$$\bar{A}_2 = \max_{0 \leq \beta \leq \beta_{max}} \min_{\mathbf{e} \in S_1} \beta \sqrt{\sigma_{\boldsymbol{\nu}}^2 + \frac{1}{p_L} \mathbf{e}^T \mathbf{R}^{(L)} \mathbf{e}} - \frac{\beta}{\sqrt{n p_L}} \mathbf{g}^T \mathbf{R}^{(L)/2} \mathbf{e} - \frac{\beta^2}{2} + R(\mathbf{e} + \boldsymbol{\theta}^*). \tag{151}$$

Now we make use of the "square root trick", which notes that for any scalar $c > 0$ we can express $\sqrt{c} = \min_{q > 0} \frac{q}{2} + \frac{c}{2q}$. Using this technique we obtain:

$$\bar{A}_2 = \max_{0 \leq \beta \leq \beta_{max}} \min_{q_{min} < q \leq q_{\max}} \frac{\beta \sigma_{\boldsymbol{\nu}}^2}{2q} + \frac{\beta q}{2} - \frac{\beta^2}{2}$$
$$+ \min_{\mathbf{e} \in S_1} \frac{\beta}{2 q p_L} \mathbf{e}^T \mathbf{R}^{(L)} \mathbf{e} - \frac{\beta}{\sqrt{n p_L}} \mathbf{g}^T \mathbf{R}^{(L)/2} \mathbf{e} + R(\mathbf{e} + \boldsymbol{\theta}^*). \tag{152}$$

Where we have interchanged the order of the two minimizations, and have noted that q can be both upper bounded and lower bounded, by $q_{min} = \sigma_{\boldsymbol{\nu}}$, achieved when $\mathbf{e} = 0$ and $q_{max} > \sqrt{\sigma_{\boldsymbol{\nu}}^2 + C_{\mathbf{e}}^2 C_{\mathbf{R}^{(L)}}}$.

We now fix the values of β and q and focus only on the inner optimization over $\mathbf{e}$. We shall discuss the outer optimizations below. We define

$$D^{(L)} = D_2^{(L)}(\beta, q) = \min_{\mathbf{e}\in S_1} \frac{c_L}{2p_L}\mathbf{e}^T\mathbf{R}^{(L)}\mathbf{e} - \frac{d_L}{p_L}\mathbf{g}^T\mathbf{R}^{(L)/2}\mathbf{e} + R(\mathbf{e}+\boldsymbol{\theta}^*). \tag{153}$$

$$c_L = \frac{\beta}{q} \qquad d_L = \beta\sqrt{\frac{m}{n}} \qquad T_L(\beta, q) = \frac{\beta\sigma_{\boldsymbol{\nu}}^2}{2q} + \frac{\beta q}{2} - \frac{\beta^2}{2} \tag{154}$$

such that

$$A_2 = \max_{\beta}\min_{q} T_L(\beta, q) + D_2^{(L)}(\beta, q). \tag{155}$$

We shall focus on $D^{(L)}$ for fixed β, q. We shall now demonstrate that studying $D^{(L)}$ it maybe expressed as another min max problem. Applying the CGMT recursively to the inner problem and simplifying results in a new problem.

First we recall the definition of $\mathbf{R}^{(L)}$ and further note that for a Gaussian $\mathbf{g}$ that

$$\begin{aligned}\mathbf{R}^{(l)/2}\mathbf{g} = \tilde{\mathbf{g}} \sim \mathcal{N}(\mathbf{0}, \mathbf{R}^{(l)}) = \mathcal{N}(\mathbf{0}, \frac{\rho_{1,l}^2}{p_{l-1}}\mathbf{W}^{(l)}\mathbf{R}^{(l-1)/2}\mathbf{W}^{(l)T} + \rho_{2,l}^2\mathbf{I}_{p_l}) \\ = \rho_{1,l}\mathbf{W}^{(l)}\mathbf{R}^{(l-1)/2}\mathbf{g}_1 + \rho_{2,l}\mathbf{g}_2 \qquad \mathbf{g}_1 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{p_{l-1}}), \mathbf{g}_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{p_l}).\end{aligned} \tag{156}$$

We can now substitute in this definition. We obtain:

$$\begin{aligned}\min_{\mathbf{e}\in S_1^{(l)}} \frac{c_L\rho_{1,L}^2}{2p_Lp_{L-1}}\mathbf{e}^T\mathbf{W}^{(L)}\mathbf{R}^{(L-1)}\mathbf{W}^{(L)T}\mathbf{e} + \frac{d_L\rho_{1,L}}{p_L\sqrt{p_{L-1}}}\mathbf{g}_1^T\mathbf{R}^{(L-1)/2}\mathbf{W}^{(L)T}\mathbf{e} + \frac{c_L\rho_{2,L}^2}{2p_L}\|\mathbf{e}\|^2 \\ + \frac{d_L\rho_{2,L}}{p_L}\mathbf{g}_2^T\mathbf{e} + R(\mathbf{e}+\boldsymbol{\theta}^*),\end{aligned} \tag{157}$$

Where $\mathbf{g}_1 \in \mathbb{R}^{p_{L-1}}, \mathbf{g}_2 \in \mathbb{R}^{p_L}$ are standard normal vectors. We then complete the square over the vector $\mathbf{R}^{(L-1)/2}\mathbf{W}^{(L)T}\mathbf{e}$, we obtain

$$\begin{aligned}\min_{\mathbf{e}\in S_1^{(l)}} \frac{c_L\rho_{1,L}^2}{2p_Lp_{L-1}}\left\|\mathbf{R}^{(L-1)/2}\mathbf{W}^{(L)T}\mathbf{e} + \frac{d_L\sqrt{p_L}}{c_L\rho_{1,L}}\mathbf{g}_1^T\right\|^2 - \frac{d_L^2}{2c_Lp_L}\|\mathbf{g}_1\|^2 \\ + \frac{c_L\rho_{2,L}^2}{2p_L}\|\mathbf{e}\|^2 + \frac{d_L\rho_{2,L}}{p_L}\mathbf{g}_2^T\mathbf{e} + R(\mathbf{e}+\boldsymbol{\theta}^*).\end{aligned} \tag{158}$$

We can then introduce a new variable $\mathbf{s} \in \mathbb{R}^{p_{L-1}}$ and take the Legendre transform of the 2-norm to create a min-max problem

$$\begin{aligned}\min_{\mathbf{e}\in S_1^{(l)}} \max_{\mathbf{s}} \frac{c_L\rho_{1,L}^2}{p_Lp_{L-1}}\mathbf{s}^T\mathbf{R}^{(L-1)/2}\mathbf{W}^{(L)T}\mathbf{e} + \frac{d_L\rho_{1,L}}{p_L\sqrt{p_{L-1}}}\mathbf{s}^T\mathbf{g}_1 - \frac{c_L\rho_{1,L}^2}{2p_Lp_{L-1}}\|\mathbf{s}\|^2 - \frac{d_L^2}{2c_Lp_L}\|\mathbf{g}_1\|^2 \\ + \frac{c_L\rho_{2,L}^2}{2p_L}\|\mathbf{e}\|^2 + \frac{d_L\rho_{2,L}}{p_L}\mathbf{g}_2^T\mathbf{e} + R(\mathbf{e}+\boldsymbol{\theta}^*)\end{aligned} \tag{159}$$

We note that $\mathbf{W}^{(L)}$ is a Random Matrix with i.i.d standard normal entries, as such we if we can restrict the problem over $\mathbf{s}$ to a compact and convex set we may make use of the CGMT theorem. We show that we make this restriction in Lemma 20. As such we can consider the following problem:

$$\min_{\mathbf{e}\in S_1^{(l)}} \max_{\mathbf{s}\in S_2^{(l)}} \frac{c_L\rho_{1,L}^2}{p_Lp_{L-1}}\mathbf{s}^T\mathbf{R}^{(L-1)/2}\mathbf{W}^{(L)T}\mathbf{e} + \frac{d_L\rho_{1,L}}{p_L\sqrt{p_{L-1}}}\mathbf{s}^T\mathbf{g}_1 - \frac{c_L\rho_{1,L}^2}{2p_Lp_{L-1}}\|\mathbf{s}\|^2 - \frac{d_L^2}{2c_Lp_L}\|\mathbf{g}_1\|^2 + \frac{c_L\rho_{2,L}^2}{2p_L}\|\mathbf{e}\|^2 + \frac{d_L\rho_{2,L}}{p_L}\mathbf{g}_2^T\mathbf{e} + R(\mathbf{e}+\boldsymbol{\theta}^*), \tag{160}$$

where the set $S_2^{(l)} = \{\mathbf{s}\in\mathbb{R}^{p_{L-1}} |\ \|\mathbf{s}\| \le C_{\mathbf{s}}\sqrt{p_Lp_{L-1}}\}$ where $C_{\mathbf{s}}$ is a postive constant. We can then apply the CGMT to obtain the following problem

$$\min_{\mathbf{e}\in S_1^{(l)}} \max_{\mathbf{s}\in S_2^{(l)}} \frac{c_L\rho_{1,L}^2}{p_Lp_{L-1}}\left\|\mathbf{R}^{(L-1)}\mathbf{s}\right\|\mathbf{e}^T\mathbf{g}_3 + \frac{c_L\rho_{1,L}^2}{p_Lp_{L-1}}\|\mathbf{e}\|\mathbf{g}_4^T\mathbf{R}^{(L-1)/2}\mathbf{s} + \frac{d_L\rho_{1,L}}{p_L\sqrt{p_{L-1}}}\mathbf{s}^T\mathbf{g}_1 - \frac{c_L\rho_{1,L}^2}{2p_Lp_{L-1}}\|\mathbf{s}\|^2 - \frac{d_L^2}{2c_Lp_L}\|\mathbf{g}_1\|^2 + \frac{c_L\rho_{2,L}^2}{2p_L}\|\mathbf{e}\|^2 + \frac{d_L\rho_{2,L}}{p_L}\mathbf{g}_2^T\mathbf{e} + R(\mathbf{e}+\boldsymbol{\theta}^*) \tag{161}$$

where $\mathbf{g}_3\in\mathbb{R}^{p_L}$ and $\mathbf{g}_4\in\mathbb{R}^{p_{L-1}}$ are standard normal vectors. We introduce a new variable $\mathbf{v} = \mathbf{R}^{(L-1)/2}\mathbf{s}$ and note that $\mathbf{v}$ can be restricted to a compact set, due to the bounds on $\mathbf{R}^{(L-1)}$ and $\mathbf{s}$. We can denote this set $S_3^{(l)} = \{\mathbf{v}\in\mathbb{R}^{p_{L-1}} |\ \|\mathbf{v}\| \le C_{\mathbf{v}}\sqrt{p_Lp_{L-1}}\}$ where $C_{\mathbf{v}}$ is a positive constant. We then reintroduce this constrain with a Lagrange multiplier $\rho_{1,L}^2\boldsymbol{\mu}/p_{L-1}\sqrt{p_L}\in\mathbb{R}^{L-1}$. We obtain

$$\min_{\mathbf{e}\in S_1^{(l)},\boldsymbol{\mu}} \max_{\mathbf{s}\in S_2^{(l)},\mathbf{v}\in S_3^{(l)}} \frac{c_L\rho_{1,L}^2}{p_Lp_{L-1}}\|\mathbf{v}\|\mathbf{e}^T\mathbf{g}_3 + \frac{c_L\rho_{1,L}^2}{p_Lp_{L-1}}\|\mathbf{e}\|\mathbf{g}_4^T\mathbf{v} + \frac{d_L\rho_{1,L}}{p_L\sqrt{p_{L-1}}}\mathbf{s}^T\mathbf{g}_1 - \frac{c_L\rho_{1,L}^2}{2p_Lp_{L-1}}\|\mathbf{s}\|^2 - \frac{d_L^2}{2c_Lp_L}\|\mathbf{g}_1\|^2 + \frac{c_L\rho_{2,L}^2}{2p_L}\|\mathbf{e}\|^2 + \frac{d_L\rho_{2,L}}{p_L}\mathbf{g}_2^T\mathbf{e} + R(\mathbf{e}+\boldsymbol{\theta}^*) + \frac{\rho_{1,L}^2}{p_{L-1}\sqrt{p_L}}\boldsymbol{\mu}^T\mathbf{v} - \frac{\rho_{1,L}^2}{p_{L-1}\sqrt{p_L}}\boldsymbol{\mu}^T\mathbf{R}^{(L-1)/2}\mathbf{s} \tag{162}$$

We then let $\xi_L = \frac{\rho_{1,L}}{\sqrt{p_Lp_{L-1}}}\|\mathbf{s}\|$ and $\chi_L = \frac{\rho_{1,L}}{\sqrt{p_Lp_{L-1}}}\|\mathbf{v}\|$ and solve the optimizations over $\mathbf{s}$ and $\mathbf{v}$. We obtain the following problem:

$$\min_{\mathbf{e}\in S_1^{(l)},\boldsymbol{\mu}} \max_{0\le\xi_L\le\xi_{L,max},0\le\chi_L\le\chi_{L,max}} \frac{c_L\rho_{1,L}\chi}{\sqrt{p_Lp_{L-1}}}\mathbf{e}^T\mathbf{g}_3 + \chi\left\|\frac{c_L\rho_{1,L}}{\sqrt{p_Lp_{L-1}}}\|\mathbf{e}\|\mathbf{g}_4 + \frac{\rho_{1,l}}{\sqrt{p_{L-1}}}\boldsymbol{\mu}\right\| - \frac{c_L\xi^2}{2} + \xi\left\|\frac{d_L\rho_{1,L}}{\sqrt{p_L}}\mathbf{g}_1 - \frac{\rho_{1,L}}{\sqrt{p_{L-1}}}\mathbf{R}^{(L-1)/2}\boldsymbol{\mu}\right\| - \frac{d_L^2}{2c_Lp_L}\|\mathbf{g}_1\|^2 + \frac{c_L\rho_{2,L}^2}{2p_L}\|\mathbf{e}\|^2 + \frac{d_L\rho_{2,L}}{p_L}\mathbf{g}_2^T\mathbf{e} + R(\mathbf{e}+\boldsymbol{\theta}^*) \tag{163}$$

We interchange the order of the min and max terms and then make use of the square root trick to get rid of the two norms. We introduce two new variables t_L and k_L:

$$\max_{0\leq\xi_L\leq\xi_{L,max},0\leq\chi_L\leq\chi_{L,max}}\min_{0\leq t_l\leq t_{L,min},0\leq k_L\leq k_{L,max}}\min_{\mathbf{e}\in S_1^{(l)},\boldsymbol{\mu}}$$
$$\frac{c_L\rho_{1,L}\chi}{\sqrt{p_Lp_{L-1}}}\mathbf{e}^T\mathbf{g}_3+\frac{\chi_Lk_L}{2}+\frac{\chi_Lc_L^2\rho_{1,L}^2}{2k_Lp_Lp_{L-1}}\|\mathbf{e}\|^2\|\mathbf{g}_4\|^2+\frac{\chi_Lc_L\rho_{1,L}^2}{2k_Lp_{L-1}\sqrt{p_L}}\|\mathbf{e}\|\mathbf{g}_4^T\boldsymbol{\mu}+\frac{\chi_L\rho_{1,l}^2}{2k_Lp_{L-1}}\|\boldsymbol{\mu}\|^2$$
$$-\frac{c_L\xi^2}{2}+\frac{\xi_Lt_L}{2}+\frac{\xi_Ld_L^2\rho_{1,L}^2}{2t_Lp_L}\|\mathbf{g}_1\|-\frac{\xi_Ld_L\rho_{1,L}^2}{\sqrt{2t_Lp_Lp_{L-1}}}\mathbf{g}_1\mathbf{R}^{(l-1)/2}\boldsymbol{\mu}-\frac{\xi_L\rho_{1,L}^2}{2t_Lp_{L-1}}\boldsymbol{\mu}^T\mathbf{R}^{(L-1)}\boldsymbol{\mu}$$
$$-\frac{d_L^2}{2c_Lp_L}\|\mathbf{g}_1\|^2+\frac{c_L\rho_{2,L}^2}{2p_L}\|\mathbf{e}\|^2+\frac{d_L\rho_{2,L}}{p_L}\mathbf{g}_2^T\mathbf{e}+R(\mathbf{e}+\boldsymbol{\theta}^*)\quad(164)$$

Using the same arguments as in lemma 19 it can be seen that the problem concentrates on:

$$\max_{0\leq\xi_L\leq\xi_{L,max},0\leq\chi_L\leq\chi_{L,max}}\min_{0\leq t_l\leq t_{L,min},0\leq k_L\leq k_{L,max}}\min_{\mathbf{e}\in S_1^{(l)},\boldsymbol{\mu}}$$
$$\frac{c_L\rho_{1,L}\chi}{\sqrt{p_Lp_{L-1}}}\mathbf{e}^T\mathbf{g}_3+\frac{\chi_Lk_L}{2}+\frac{\chi_Lc_L^2\rho_{1,L}^2}{2k_Lp_L}\|\mathbf{e}\|^2+\frac{\chi_Lc_L\rho_{1,L}^2}{2k_Lp_{L-1}\sqrt{p_L}}\|\mathbf{e}\|\mathbf{g}_4^T\boldsymbol{\mu}+\frac{\chi_L\rho_{1,l}^2}{2k_Lp_{L-1}}\|\boldsymbol{\mu}\|^2$$
$$-\frac{c_L\xi^2}{2}+\frac{\xi_Lt_L}{2}+\frac{\xi_Ld_L^2\rho_{1,L}^2p_{L-1}}{2t_Lp_L}-\frac{\xi_Ld_L\rho_{1,L}^2}{2t_L\sqrt{p_Lp_{L-1}}}\mathbf{g}_1\mathbf{R}^{(l-1)/2}\boldsymbol{\mu}+\frac{\xi_L\rho_{1,L}^2}{2t_Lp_{L-1}}\boldsymbol{\mu}^T\mathbf{R}^{(L-1)}\boldsymbol{\mu}$$
$$-\frac{d_L^2p_{L-1}}{2c_Lp_L}+\frac{c_L\rho_{2,L}^2}{2p_L}\|\mathbf{e}\|^2+\frac{d_L\rho_{2,L}}{p_L}\mathbf{g}_2^T\mathbf{e}+R(\mathbf{e}+\boldsymbol{\theta}^*)\quad(165)$$

We now let

$$T_{L-1}=-\frac{d_L^2p_{L-1}}{2c_Lp_L}-\frac{c_L\xi^2}{2}+\frac{\xi_Lt_L}{2}+\frac{\xi_Ld_L^2\rho_{1,L}^2p_{L-1}}{2t_Lp_L}+\frac{\chi_Lk_L}{2}\quad(166)$$

$$a=\frac{\chi_Lc_L^2\rho_{1,L}^2}{k_L}+c_L\rho_{2,L}^2\qquad b=\sqrt{\frac{c_L^2\rho_{1,L}^2\chi^2p_L}{p_{L-1}}+d_L^2\rho_{2,L}^2}\quad(167)$$

$$c_{L-1}=\frac{\chi_L\rho_{1,l}^2}{k_l}\qquad d_{L-1}=\frac{\chi_Lc_L\rho_{1,L}^2}{k_L}\qquad\bar{c}=\frac{\xi_L\rho_{1,L}^2}{t_L}\qquad\bar{d}=\frac{\xi_l\rho_{1,L}^2}{t_L}\quad(168)$$

as such we can obtain:

$$\max_{0\leq\xi_L\leq\xi_{L,max},0\leq\chi_L\leq\chi_{L,max}}\min_{0\leq t_l\leq t_{L,min},0\leq k_L\leq k_{L,max}}T_{L-1}+\min_{\mathbf{e}\in S_1^{(l)},\boldsymbol{\mu}}$$
$$\frac{a}{2p_L}\|\mathbf{e}\|^2+\frac{b}{p_L}\mathbf{e}^T\mathbf{g}_1+\frac{c_{L-1}}{2p_{L-1}}\|\boldsymbol{\mu}\|^2+\frac{d_{L-1}}{p_{L-1}}\frac{\|\mathbf{e}\|}{\sqrt{p_{L-1}}}\mathbf{g}_2^T\boldsymbol{\mu}+\frac{\bar{c}}{2p_{L-1}}\boldsymbol{\mu}^T\mathbf{R}^{(L-1)}\boldsymbol{\mu}$$
$$+\frac{\bar{d}}{p_L}\mathbf{g}_3^T\mathbf{R}^{(L-1)/2}\boldsymbol{\mu}+R(\mathbf{e}+\boldsymbol{\theta}^*)\quad(169)$$

Where $\mathbf{g}_1\in\mathbb{R}^{p_L},\mathbf{g}_2,\mathbf{g}_3\in\mathbb{R}^{p_{L-1}}$ are standard normal vectors. We now fix all parameters of the optimization except for $\boldsymbol{\mu}$ and focus specifically on the last four terms terms. We shall note that this

can once again be expressed as a min-max optimization amenable to the CGMT. However at this point we enter a recursive structure. We demonstrate in Lemma 21 that a problem of the form

$$\max_{\boldsymbol{\mu}} \frac{\gamma_1}{2p_{l-1}} \|\boldsymbol{\mu}\|^2 + \frac{\gamma_2}{p_{l-1}} \mathbf{g}_2^T \boldsymbol{\mu} + \frac{\gamma_3}{2p_{l-1}} \boldsymbol{\mu}^T \mathbf{R}^{(l-1)} \boldsymbol{\mu} + \frac{\gamma_4}{p_l} \mathbf{g}_3^T \mathbf{R}^{(l-1)/2} \boldsymbol{\mu} \tag{170}$$

With generic constants $\boldsymbol{\gamma}_i$ $(i = 1, \dots 4)$ can be expressed by means of the CGMT as:

$$\begin{aligned} &\max_{0\le\xi_l\le\xi_{l,max}, 0\le\chi_l\le\chi_{l,max}} \min_{0\le t_l\le t_{l,max}, 0\le k_l\le k_{l,max}} T_l \\ &+ \min_{\boldsymbol{\mu}} \frac{\bar{\gamma}_1}{2p_l} \|\boldsymbol{\mu}\|^2 + \frac{\bar{\gamma}_2}{p_l} \mathbf{g}_2^T \boldsymbol{\mu} + \frac{\bar{\gamma}_3}{2p_l} \boldsymbol{\mu}^T \mathbf{R}^{(l-1)} \boldsymbol{\mu} + \frac{\bar{\gamma}_4}{p_l} \mathbf{g}_3^T \mathbf{R}^{(l)/2} \boldsymbol{\mu} \end{aligned} \tag{171}$$

Where

$$T_l = \frac{\chi_l k_l}{2} - \frac{\gamma_3 \xi_l^2}{2} + \frac{\xi_l t_l}{2} + \frac{\xi_l \gamma_4^2 p_{l-1}}{2 t_l p_l} - \frac{\gamma_4^2 p_{l-1}}{2\gamma_3 p_l}$$
$$- \left(\gamma_1 + \frac{\gamma_3^2 \rho_{1,l}^2 \chi_l}{k_l} + \gamma_3 \rho_{2,l}^2 \right)^{-1} \left(\gamma_4^2 \rho_{2,l}^2 + \frac{\gamma_3^2 \rho_{1,l}^2 \chi_l^2 p_l}{p_{l-1}} + \gamma_2 \right) \tag{172}$$

$$\bar{\gamma_1} = \frac{\xi_l \rho_{1,l}^2}{k_l} - \left(\gamma_1 + \frac{\gamma_3^2 \rho_{1,l}^2 \chi_l}{k_l} + \gamma_3 \rho_{2,l}^2 \right)^{-1} \frac{\gamma_3^2 \rho_{1,l}^4 \xi_l^2}{2k_l^2} \tag{173}$$

$$\bar{\gamma_2} = - \left(\gamma_1 + \frac{\gamma_3^2 \rho_{1,l}^2 \chi_l}{k_l} + \gamma_3 \rho_{2,l}^2 \right)^{-1} \left(\gamma_4^2 \rho_{2,l}^2 + \frac{\gamma_3^2 \rho_{1,l}^2 \chi_l^2 p_l}{p_{l-1}} + \gamma_2 \right)^{1/2} \frac{\gamma_3 \rho_{1,l}^2 \chi_l}{2k_l} \tag{174}$$

$$\bar{\gamma_3} = \frac{\xi_l \rho_{1,l}^2}{t_l} \qquad \bar{\gamma}_4 = \frac{\xi_l \gamma_4 \rho_{1,l} \sqrt{p_l}}{2 t_l \sqrt{p_l - 1}} \tag{175}$$

We can also note that the termination of the recursion is given by the optimization problem where $\mathbf{R}^{(0)} = \mathbf{I}$, in this case

$$\begin{aligned} &\min_{\boldsymbol{\mu}} \frac{\bar{\gamma}_1}{2p_l} \|\boldsymbol{\mu}\|^2 + \frac{\bar{\gamma}_2}{p_l} \mathbf{g}_2^T \boldsymbol{\mu} + \frac{\bar{\gamma}_3}{2p_l} \boldsymbol{\mu}^T \boldsymbol{\mu} + \frac{\bar{\gamma}_4}{p_l} \mathbf{g}_3^T \boldsymbol{\mu} \\ &= - \frac{\bar{\gamma}_2^2 + \bar{\gamma}_4^2}{\bar{\gamma}_1 + \bar{\gamma}_3} \overset{def}{=} F_0 \end{aligned} \tag{176}$$

As such we can express the final result for the $L-$layer deep RF model as being given by

$$\begin{aligned} &\max_{\beta>0} \min_{q} T_L + \max_{\xi_L>0, \chi_L>0} \min_{t_L>0, k_L>0} T_{L-1} + \min_{\mathbf{e}} \frac{a}{2pl} \|\mathbf{e}\| + \frac{b}{p_L} \mathbf{e}^T \mathbf{g}_1 + + R(\boldsymbol{\theta} + \boldsymbol{\theta}^*) + \\ &\max_{\xi_{L-1}>0, \chi_{L-1}>0} \min_{t_{L-1}>0, k_{L-1}>0} \cdots \max_{\xi_0\ge 0, \chi_0\ge 0} \min_{t_0>0, k_0>0} \sum_{i=1}^{L-2} T_l(\mathbf{e}) \end{aligned} \tag{177}$$

Where

$$T_L = \frac{\beta q}{2} + \frac{\beta \sigma_{\boldsymbol{\nu}^2}}{2q} - \frac{\beta^2}{2} \tag{178}$$

$$T_{L-1} = -\frac{d_L^2 p_{L-1}}{2c_L p_L} - \frac{c_L \xi^2}{2} + \frac{\xi_L t_L}{2} + \frac{\xi_L d_L^2 \rho_{1,L}^2 p_{L-1}}{2t_L p_L} + \frac{\chi_L k_L}{2} \tag{179}$$

$$T_l = \frac{\chi_l k_l}{2} - \frac{\gamma_3 \xi_l^2}{2} + \frac{\xi_l t_l}{2} + \frac{\xi_l \gamma_4^2 p_{l-1}}{2t_l p_l} - \frac{\gamma_4^2 p_{l-1}}{2\gamma_3 p_l}$$
$$- \left(\gamma_1 + \frac{c_l^2 \rho_{1,l}^2 \chi_l}{k_l} + \gamma_3 \rho_{2,l}^2 \right)^{-1} \left(d_l^2 \rho_{2,l}^2 + \frac{c_l^2 \rho_{1,l}^2 \chi_l^2 p_l}{p_{l-1}} + \bar{d} \right) \qquad 1 \le l \le L-2 \tag{180}$$

$$T_0 = \frac{d_0^2 + \bar{d}_0^2}{c_0 + \bar{c}_0} \tag{181}$$

$$a = \frac{\chi_L c_L^2 \rho_{1,L}^2}{k_L} + c_L \rho_{2,L}^2 \qquad b = \sqrt{\frac{c_L^2 \rho_{1,L}^2 \chi^2 p_L}{p_{L-1}} + d_L^2 \rho_{2,L}^2} \tag{182}$$

and the constants $c_i, d_i, \bar{c}_i, \bar{d}_i$ are given by

$$c_L = \frac{\beta}{q} \qquad d_\beta \sqrt{\frac{p_L}{n}} c_{L-1} = \frac{\chi_L \rho_{1,l}^2}{k_l} \qquad d_{L-1} = \frac{\chi_L c_L \rho_{1,L}^2}{k_L} \frac{\|\mathbf{e}\|}{\sqrt{p_{l-1}}} \tag{183}$$

$$\bar{c}_L = \frac{\xi_L \rho_{1,L}^2}{t_L} \qquad \bar{d}_L = \frac{\xi_l \rho_{1,L}^2}{t_L} \tag{184}$$

$$c_l = \frac{\xi_l \rho_{1,l}^2}{k_l} - \left(c_{l+1} + \frac{\bar{c}_{l+1}^2 \rho_{1,l}^2 \chi_l}{k_l} + \bar{c}_{l+1} \rho_{2,l}^2 \right)^{-1} \frac{\bar{c}_l^2 \rho_{1,l}^4 \xi_l^2}{2k_l^2} \tag{185}$$

$$d_l = -\left(c_{l+1} + \frac{\bar{c}_{l+1}^2 \rho_{1,l}^2 \chi_l}{k_l} + \bar{c}_{l+1} \rho_{2,l}^2 \right)^{-1} \left(\bar{d}_{l+1}^2 \rho_{2,l}^2 + \frac{\bar{c}_{l+1}^2 \rho_{1,l}^2 \chi_l^2 p_l}{p_{l-1}} + d_{l+1} \right)^{1/2} \frac{\bar{c}_{l+1} \rho_{1,l}^2 \chi_l}{2k_l} \tag{186}$$

$$\bar{c}_l = \frac{\xi_l \rho_{1,l}^2}{t_l} \qquad \bar{d}_l = \frac{\xi_l \bar{d}_{l+1} \rho_{1,l} \sqrt{p_l}}{2t_l \sqrt{p_{l-1}}} \tag{187}$$

For the final step of the proof we note that for each successive application of the CGMT we froze all previous values of β, q as well as ξ_l, χ_l, t_l, k_l for $l \le L$. By the properties of the CGMT we know that for these fixed values we have pointwise convergence. However, we wish to demonstrate uniform convergence for the properties that we are interested in. This however is simple to see in this case.

There are two problems we need to consider. We need to show that Eq (169) converges uniformly to (159) For each value of β, q and that for each problem (171) converges uniformly to (170). We can see that all optimization variablse $\beta, q, \xi_l, \chi_l, t_l, k_l$ exist in bounded regions. For example $\beta \in [0, \beta_{max}]$. Our goal is to show that each problem is Lipschitz continuous on these regions with some Lipschitz constant K. As each problem is strongly convex it has a unique solution, and all are continuously differentiable on the existing region. As such to show Lipschitz continuity one has to show that each of the partial derivatives is bounded, calculation is tedious but can be completed readily. By bounding the derivatives we can show that all problems are Lipschitz. Uniform convergence can then be demonstrated by means of a simple ϵ-net argument. For an application of

this to a recursive CGMT problem, see Bosch et al. (2022)[Appendix B]. This completes the proof of part 1 of the theorem.

D.1. Proof of Part 2 of the Theorem

The proof of part 2 is the same as the proof of part 2 of theorem 5 given in Appendix B. A Regularization function $R_\epsilon(\mathbf{e}+\boldsymbol{\theta}^*) = R(\mathbf{e}+\boldsymbol{\theta}^*) \pm \epsilon h(\mathbf{e}+\boldsymbol{\theta}^*)$ with ϵ chosen sufficiently small for R_ϵ to remain strongly convex. As such the first part of the theorem holds. Then by bounding the difference and making use of the bounds on $h(\mathbf{e})$ the proof can be obtained.

D.2. Auxiliary Lemmas

Lemma 20 *Consider the following two problems given in equations (159), (161) that correspond to a problem and the alternative problem given by the CGMT*

$$P_1 = \min_{\mathbf{e}\in S_1^{(l)}} \max_{\mathbf{s}\in S_2^{(l)}} \frac{c_L\rho_{1,L}^2}{p_L p_{L-1}}\mathbf{s}^T\mathbf{R}^{(L-1)/2}\mathbf{W}^{(L)T}\mathbf{e} + \frac{d_L\rho_{1,L}}{p_L\sqrt{p_{L-1}}}\mathbf{s}^T\mathbf{g}_1 - \frac{c_L\rho_{1,L}^2}{2p_Lp_{L-1}}\|\mathbf{s}\|^2$$
$$-\frac{d_L^2}{2c_Lp_L}\|\mathbf{g}_1\|^2 + \frac{c_L\rho_{2,L}^2}{2p_L}\|\mathbf{e}\|^2 + \frac{d_L\rho_{2,L}}{p_L}\mathbf{g}_2^T\mathbf{e} + R(\mathbf{e}+\boldsymbol{\theta}^*) \quad (188)$$

$$P_2 = \min_{\mathbf{e}\in S_1^{(l)}} \max_{\mathbf{s}\in S_2^{(l)}} \frac{c_L\rho_{1,L}^2}{p_Lp_{L-1}}\left\|\mathbf{R}^{(L-1)}\mathbf{s}\right\|\mathbf{e}^T\mathbf{g}_3 + \frac{c_L\rho_{1,L}^2}{p_Lp_{L-1}}\|\mathbf{e}\|\,\mathbf{g}_4^T\mathbf{R}^{(L-1)/2}\mathbf{s} + \frac{d_L\rho_{1,L}}{p_L\sqrt{p_{L-1}}}\mathbf{s}^T\mathbf{g}_1$$
$$-\frac{c_L\rho_{1,L}^2}{2p_Lp_{L-1}}\|\mathbf{s}\|^2 - \frac{d_L^2}{2c_Lp_L}\|\mathbf{g}_1\|^2 + \frac{c_L\rho_{2,L}^2}{2p_L}\|\mathbf{e}\|^2 + \frac{d_L\rho_{2,L}}{p_L}\mathbf{g}_2^T\mathbf{e} + R(\mathbf{e}+\boldsymbol{\theta}^*) \quad (189)$$

Where $\mathbf{g}_i$ are standard normal vectors. Denote $\hat{\mathbf{e}}_1$ and $\hat{\mathbf{e}}_2$ as the optimal points of the two problems and let $\hat{s}_1(\mathbf{e})$ and $\hat{s}_2(\mathbf{e})$ be the optimal points of the inner optimizations as functions of a fixed $\mathbf{e}$. Recall that R is $\mu-$strongly convex and that $\|\nabla R(\mathbf{0})\| = \mathcal{O}(\sqrt{p_L})$. Then there exists positive constants $C_\mathbf{e}$ and $C_\mathbf{s}$ depending only on μ such that

$$\lim_{p_L\to\infty}\Pr\left(\|\hat{\mathbf{e}}_i\|_2 \le C_\mathbf{e}\sqrt{m}\right) = 1 \qquad i=1,2 \quad (190)$$

and

$$\lim_{p_L\to\infty}\Pr\left(\sup_{\mathbf{e}|\|\mathbf{e}\|\le C_\mathbf{e}\sqrt{m}}\|\hat{\mathbf{s}}_i(\mathbf{e})\| \le C_\mathbf{s}\sqrt{p_Lp_{L-1}}\right) = 1 \qquad i=1,2 \quad (191)$$

Proof We note that $\mathbf{e}$ in problem P_1 is already bounded to a compact set. For both optimizations, we solve the inner optimization over $\mathbf{s}$ and denote this solution as

$$\min_\mathbf{e} F_i(\mathbf{e}) \qquad i=1,2 \quad (192)$$

Such that F_i is the optimal value over $\mathbf{s}$. When we set $\mathbf{s}=0$ in both optimizations we see that

$$F(\mathbf{e}) \ge T(\mathbf{e}) \overset{def}{=} -\frac{d_L^2}{2c_Lp_{L-1}}\|\mathbf{g}_1\|_2^2 + \frac{c_L\rho_{2,L}^2}{2p_L}\|\mathbf{e}\|_2^2 + \frac{d_L\rho_{2,L}}{p_L}\mathbf{g}_2^T\mathbf{e} + R(\mathbf{e}+\boldsymbol{\theta}^*) \quad (193)$$

We can note readily that $T(\mathbf{e})$ is ν-strongly convex, for some constant ν with respect to $\mathbf{e}$. We see that

$$T(\mathbf{e}) \geq T(\mathbf{0}) + \mathbf{d}^T\mathbf{e} + \frac{\nu}{2}\left\|\mathbf{e}\right\|_2^2 \tag{194}$$

where $\mathbf{d} = \nabla T(\mathbf{0})$. By assumption we note that $\mathbf{d} = \mathcal{O}(\sqrt{p_L})$. For problem p_2 we now note the following:

$$\begin{aligned} F_2(\mathbf{e}) &= \max_{\mathbf{s}} \frac{c_L\rho_{1,L}^2}{p_l p_{L-1}} \left\|\mathbf{R}^{(L-1)/2}\mathbf{s}\right\| \mathbf{g}_3^T\mathbf{e} + \frac{c_L\rho_{1,L}^2}{p_L p_{L-1}} \left\|\mathbf{e}\right\|_2 \mathbf{g}_4^T\mathbf{R}^{(L-1)/2}\mathbf{s} + \frac{d_L\rho_{1,L}}{p_L\sqrt{p_{L-1}}}\mathbf{s}^T\mathbf{g}_1 \\ &\qquad - \frac{c_L\rho_{1,L^2}}{2p_L p_{L-1}}\left\|\mathbf{s}\right\|_2^2 + T(\mathbf{e}) \\ &\leq \max_{\mathbf{s}} \frac{c_L\rho_{1,L}^2}{p_l p_{L-1}} \left\|\mathbf{R}^{(L-1)/2}\right\| \left\|\mathbf{s}\right\| \mathbf{g}_3^T\mathbf{e} + \frac{c_L\rho_{1,L}^2}{p_L p_{L-1}} \left\|\mathbf{e}\right\|_2 \mathbf{g}_4^T\mathbf{R}^{(L-1)/2}\mathbf{s} + \frac{d_L\rho_{1,L}}{p_L\sqrt{p_{L-1}}}\mathbf{s}^T\mathbf{g}_1 \\ &\qquad - \frac{c_L\rho_{1,L^2}}{2p_L p_{L-1}}\left\|\mathbf{s}\right\|_2^2 + T(\mathbf{e}) \end{aligned} \tag{195}$$

Then letting $\xi = \left\|\mathbf{s}\right\|$ the optimization over $\mathbf{s}$ may be solved to find that

$$\begin{aligned} F_2(\mathbf{e}) &\leq \max_{\xi>0} \frac{c_L\rho_{1,L}^2\xi}{p_l p_{L-1}} \left\|\mathbf{R}^{(L-1)/2}\right\| \mathbf{g}_3^T\mathbf{e} + \xi\left\|\frac{c_L\rho_{1,L}^2}{p_L p_{L-1}} \left\|\mathbf{e}\right\|_2 \mathbf{R}^{(L-1)/2}\mathbf{g}_4 + \frac{d_L\rho_{1,L}}{p_L\sqrt{p_{L-1}}}\mathbf{g}_1\right\| \\ &\qquad - \frac{c_L\rho_{1,L^2}\xi^2}{2p_L p_{L-1}} + T(\mathbf{e}) \end{aligned} \tag{196}$$

We now note that this value will only be increased if the constraint over ξ is dropped, as such

$$\begin{aligned} F_2(\mathbf{e}) &\leq \max_{\xi} \frac{c_L\rho_{1,L}^2\xi}{p_l p_{L-1}} \left\|\mathbf{R}^{(L-1)/2}\right\| \mathbf{g}_3^T\mathbf{e} + \xi\left\|\frac{c_L\rho_{1,L}^2}{p_L p_{L-1}} \left\|\mathbf{e}\right\|_2 \mathbf{R}^{(L-1)/2}\mathbf{g}_4 + \frac{d_L\rho_{1,L}}{p_L\sqrt{p_{L-1}}}\mathbf{g}_1\right\| \\ &\qquad - \frac{c_L\rho_{1,L^2}\xi^2}{2p_L p_{L-1}} + T(\mathbf{e}) \end{aligned} \tag{197}$$

solving this optimization we see that

$$F_2(\mathbf{0}) \leq \frac{d_L^2}{c_L p_L}\left\|\mathbf{g}_1^2\right\|^2 + T(\mathbf{0}) \tag{198}$$

As such we can see that

$$\frac{d_L^2}{c_L p_L}\left\|\mathbf{g}_1^2\right\|^2 + T(\mathbf{0}) \geq F(\mathbf{0}) \geq F(\hat{\mathbf{e}}) \geq T(\mathbf{0}) + \mathbf{d}^T\mathbf{e} + \frac{\nu}{2}\left\|\mathbf{e}\right\| \tag{199}$$

Hence,

$$\frac{\nu}{2}\left\|\mathbf{e} + \frac{1}{\nu}\mathbf{d}\right\|^2 \leq \frac{1}{\nu}\left\|\mathbf{d}\right\|_2^2 + \frac{d_L^2}{c_L p_L}\left\|\mathbf{g}_1^2\right\|^2 \tag{200}$$

and as such

$$\|\mathbf{e}\|_2 \leq \frac{1}{\nu}\|\mathbf{d}\|_2 + \sqrt{\frac{2}{\nu}\|\mathbf{d}\|_2^2 + \frac{d_L^2}{c_L p_L}\left\|\mathbf{g}_1^2\right\|^2} \tag{201}$$

We recall that with high probability $\left\|\mathbf{g}_1^2\right\| < C\sqrt{p_{L-1}}$. Recalling the assumptions on $\mathbf{d}$ and that all contants $p_0 \cdots, p_L$ grow at constant ratios we see that there must exist a constant $C_{\mathbf{e}}$ such that

$$\Pr(\|\hat{\mathbf{e}}\| > C_{\mathbf{e}}\sqrt{m}) \to 0 \tag{202}$$

We now consider the bounds on $\mathbf{s}$. For problem P_1 we can note from the optimality condition over $\mathbf{s}$ that

$$\hat{\mathbf{s}}_1(\mathbf{e}) = \mathbf{R}^{(L-1)/2}\mathbf{W}^{(l)T}\mathbf{e} + \frac{d_l\sqrt{p_{L-1}}}{c_1\rho_{1,l}}\mathbf{g} \tag{203}$$

As such for all $\mathbf{e} \in S_1^{(l)}$ we can see that

$$\|\hat{\mathbf{s}}(\mathbf{e})\|_2 \leq \left\|\mathbf{R}^{(L-1)/2}\right\|_2 \|\mathbf{W}\| \|\mathbf{e}\|_2 + \frac{d_l\sqrt{p_{l-1}}}{c_1\rho_{1,L}}\|\mathbf{g}\| \tag{204}$$

From Standard results we know that $\left\|\mathbf{W}^{(l)}\right\|_2 < C\sqrt{p_{L-1}}$ and that $\|\mathbf{g}\|_2 \leq C\sqrt{p_{L-1}}$. Using the bounds on $\mathbf{e}$ and $\mathbf{R}^{(l)}$ we and recalling that $p_L \sim p_{L-1}$ we note that there exists a constant $C_{\mathbf{s}_1}$ exists.

Now noting that $\hat{\xi}$ is an upper bound for $\|\hat{\mathbf{s}}_2\|$ in problem for problem P_2 we can note from its optimality condition that

$$\begin{aligned}\|\hat{\mathbf{s}}_2(\mathbf{e})\| \leq \hat{\xi} &= \left\|\mathbf{R}^{(L-1)/2}\right\| \mathbf{g}_3^T\mathbf{e} + \left\|\|\mathbf{e}\|\mathbf{R}^{(L-1)/2}\mathbf{g}_4 + \frac{d_L\sqrt{p_{L-1}}}{\rho_{1,l}c_L}\mathbf{g}_1\right\|_2 \\ &\leq \left\|\mathbf{R}^{(L-1)/2}\right\| \|\mathbf{g}_3\| \|\mathbf{e}\| + \|\mathbf{e}\| \left\|\mathbf{R}^{(L-1)/2}\right\| \|\mathbf{g}_4\| + \frac{d_L\sqrt{p_{L-1}}}{\rho_{1,l}c_L}\|\mathbf{g}_1\|\end{aligned} \tag{205}$$

Which making use of the bounds used above we can once again determine that there exists a constant $C_{\mathbf{s}_2}$. Choosing $C_{\mathbf{s}}$ to be the maximum of $\boldsymbol{C}_{\mathbf{s}_1}, C_{\mathbf{s}_2}$ we can then construct the set $S_2^{(l)} = \{\mathbf{s} \in \mathbb{R}^{p_{L-1}} |\ \|\mathbf{s}\| \leq C_{\mathbf{s}}\sqrt{p_{L-1}p_L}\}$ ■

Lemma 21 *Consider the following optimization problem given in* (170)

$$\min_{\boldsymbol{\mu}} \frac{\gamma_1}{2p_l}\|\boldsymbol{\mu}\|^2 + \frac{\gamma_2}{p_l}\mathbf{g}_2^T\boldsymbol{\mu} + \frac{\gamma_3}{2p_l}\boldsymbol{\mu}^T\mathbf{R}^{(l-1)}\boldsymbol{\mu} + \frac{\gamma_4}{p_l}\mathbf{g}_3^T\mathbf{R}^{(l)/2}\boldsymbol{\mu} \tag{206}$$

This problem is asymptotically equivalent to the following problem

$$\begin{aligned}&\max_{0\leq\xi_l\leq\xi_{l,max},0\leq\chi_l\leq\chi_{l,max}} \min_{0\leq t_l\leq t_{l,max},0\leq k_l\leq k_{l,max}} T_l \\ &+ \min_{\boldsymbol{\eta}} \frac{\bar{\gamma}_1}{2p_l}\|\boldsymbol{\mu}\|^2 + \frac{\bar{\gamma}_2}{p_l}\mathbf{g}_2^T\boldsymbol{\mu} + \frac{\bar{\gamma}_3}{2p_l}\boldsymbol{\mu}^T\mathbf{R}^{(l-1)}\boldsymbol{\mu} + \frac{\bar{\gamma}_4}{p_l}\mathbf{g}_3^T\mathbf{R}^{(l)/2}\boldsymbol{\mu}\end{aligned} \tag{207}$$

Where

$$T_l = \frac{\chi_l k_l}{2} - \frac{\gamma_3 \xi_l^2}{2} + \frac{\xi_l t_l}{2} + \frac{\xi_l \gamma_4^2 p_{l-1}}{2 t_l p_l} - \frac{\gamma_4^2 p_{l-1}}{2\gamma_3 p_l} - \left(\gamma_1 + \frac{\gamma_3^2 \rho_{1,l}^2 \chi_l}{k_l} + \gamma_3 \rho_{2,l}^2\right)^{-1} \left(\gamma_4^2 \rho_{2,l}^2 + \frac{\gamma_3^2 \rho_{1,l}^2 \chi_l^2 p_l}{p_{l-1}} + \gamma_2\right) \tag{208}$$

$$\bar{\gamma_1} = \frac{\xi_l \rho_{1,l}^2}{k_l} - \left(\gamma_1 + \frac{\gamma_3^2 \rho_{1,l}^2 \chi_l}{k_l} + \gamma_3 \rho_{2,l}^2\right)^{-1} \frac{\gamma_3^2 \rho_{1,l}^4 \xi_l^2}{2 k_l^2} \tag{209}$$

$$\bar{\gamma_2} = -\left(\gamma_1 + \frac{\gamma_3^2 \rho_{1,l}^2 \chi_l}{k_l} + \gamma_3 \rho_{2,l}^2\right)^{-1} \left(\gamma_4^2 \rho_{2,l}^2 + \frac{\gamma_3^2 \rho_{1,l}^2 \chi_l^2 p_l}{p_{l-1}} + \gamma_2\right)^{1/2} \frac{\gamma_3 \rho_{1,l}^2 \chi_l}{2 k_l} \tag{210}$$

$$\bar{\gamma_3} = \frac{\xi_l \rho_{1,l}^2}{t_l} \qquad \bar{\gamma_4} = \frac{\xi_l \gamma_4 \rho_{1,l} \sqrt{p_l}}{2 t_l \sqrt{p_{l-1}}} \tag{211}$$

Proof We first substitute in the value of $\mathbf{R}^{(l)}$. From this we obtain

$$\min_{\boldsymbol{\mu}} \frac{\gamma_1}{2p_l} \|\boldsymbol{\mu}\|^2 + \frac{\gamma_2}{p_l} \mathbf{g}_2^T \boldsymbol{\mu} + \frac{\gamma_3 \rho_{1,l}^2}{2 p_l p_{l-1}} \boldsymbol{\mu}^T \mathbf{W}^{(l)} \mathbf{R}^{(l-1)} \mathbf{W}^{(l)} \boldsymbol{\mu} + \frac{\gamma_3 \rho_{2,l}^2}{2p_l} \|\boldsymbol{\mu}\|^2 + \frac{\gamma_4 \rho_{1,l}}{p_l \sqrt{p_l}} \mathbf{g}_2^T \mathbf{R}^{(l)/2} \boldsymbol{\mu} + \frac{\gamma_4 \rho 2, l}{p_l} \mathbf{g}_3^T \boldsymbol{\mu}. \tag{212}$$

We then complete the square over the vector $\mathbf{R}^{(l-1)/2} \mathbf{W}^{(l)T} \boldsymbol{\mu}$ from which we obtain

$$\min_{\boldsymbol{\mu}} \frac{\gamma_3 \rho_{1,l}^2}{2 p_l p_{l-1}} \left\| \mathbf{R}^{(l-1)/2} \mathbf{W}^{(l)T} \boldsymbol{\mu} + \frac{\gamma_4 \sqrt{p_{l-1}}}{\gamma_3 \rho_{1,l}} \mathbf{g}_2 \right\|^2 - \frac{\gamma_4^2}{2\gamma_3 p_l} \|\mathbf{g}_2\| \frac{\gamma_1}{2p_l} \|\boldsymbol{\mu}\|^2 + \frac{\gamma_2}{p_l} \mathbf{g}_2^T \boldsymbol{\mu} + \frac{\gamma_3 \rho_{2,l}^2}{2p_l} \|\boldsymbol{\mu}\|^2 + \frac{\gamma_4 \rho_{2,l}}{p_l} \mathbf{g}_3^T \boldsymbol{\mu}. \tag{213}$$

We then take the Legendre transform of the 2-norm and introduce a new variable $\mathbf{s}$

$$\min_{\boldsymbol{\mu}} \max_{\mathbf{s}} \frac{\gamma_3 \rho_{1,l}^2}{p_l p_{l-1}} \mathbf{s}^T \mathbf{R}^{(l-1)/2} \mathbf{W}^{(l)T} \boldsymbol{\mu} + \frac{\gamma_4 \rho_{1,l}}{p_l \sqrt{p_{l-1}}} \mathbf{s}^T \mathbf{g}_2 - \frac{\gamma_3 \rho_{1,l}^2}{2 p_l p_{l-1}} \|\mathbf{s}\|^2 - \frac{\gamma_4^2}{2\gamma_3 p_l} \|\mathbf{g}_2\| \frac{\gamma_1}{2p_l} \|\boldsymbol{\mu}\|^2 + \frac{\gamma_2}{p_l} \mathbf{g}_2^T \boldsymbol{\mu} + \frac{\gamma_3 \rho_{2,l}^2}{2p_l} \|\boldsymbol{\mu}\|^2 + \frac{\gamma_4 \rho_{2,l}}{p_l} \mathbf{g}_3^T \boldsymbol{\mu} \tag{214}$$

Using the same argument as lemmas 18 and 20 we can show that these problems can be bounded to compact sets $\mathbf{S}_1^{(l)}, \mathbf{S}_2^{(l)}$. As such we can consider the problem

$$\min_{\boldsymbol{\mu} \in S_1^{(l)}} \max_{\mathbf{s} \in S_2^{(l)}} \frac{\gamma_3 \rho_{1,l}^2}{p_l p_{l-1}} \mathbf{s}^T \mathbf{R}^{(l-1)/2} \mathbf{W}^{(l)T} \boldsymbol{\mu} + \frac{\gamma_4 \rho_{1,l}}{p_l \sqrt{p_{l-1}}} \mathbf{s}^T \mathbf{g}_2 - \frac{\gamma_3 \rho_{1,l}^2}{2 p_l p_{l-1}} \|\mathbf{s}\|^2 - \frac{\gamma_4^2}{2\gamma_3 p_l} \|\mathbf{g}_2\| \frac{\gamma_1}{2p_l} \|\boldsymbol{\mu}\|^2 + \frac{\gamma_2}{p_l} \mathbf{g}_2^T \boldsymbol{\mu} + \frac{\gamma_3 \rho_{2,l}^2}{2p_l} \|\boldsymbol{\mu}\|^2 + \frac{\gamma_4 \rho_{2,l}}{p_l} \mathbf{g}_3^T \boldsymbol{\mu}. \tag{215}$$

We can now apply the CGMT obtaining:

$$\min_{\boldsymbol{\mu}\in S_1^{(l)}} \max_{\mathbf{s}\in S_2^{(l)}} \frac{\gamma_3\rho_{1,l}^2}{p_l p_{l-1}} \|\boldsymbol{\mu}\| \mathbf{s}^T\mathbf{R}^{(l-1)/2}\mathbf{g}_4 + \frac{\gamma_3\rho_{1,l}^2}{p_l p_{l-1}} \left\|\mathbf{R}^{(l-1)}\mathbf{s}\right\| \boldsymbol{\mu}^T\mathbf{g}_5 + \frac{\gamma_4\rho_{1,l}}{p_l\sqrt{p_{l-1}}}\mathbf{s}^T\mathbf{g}_2$$
$$-\frac{\gamma_3\rho_{1,l}^2}{2p_l p_{l-1}}\|\mathbf{s}\|^2 - \frac{\gamma_4^2}{2\gamma_3 p_l}\|\mathbf{g}_2\| \frac{\gamma_1}{2p_l}\|\boldsymbol{\mu}\|^2 + \frac{\gamma_2}{p_l}\mathbf{g}_2^T\boldsymbol{\mu} + \frac{\gamma_3\rho_{2,l}^2}{2p_l}\|\boldsymbol{\mu}\|^2 + \frac{\gamma_4\rho_{2,l}}{p_l}\mathbf{g}_3^T\boldsymbol{\mu}. \tag{216}$$

We introduce a new variable $\mathbf{v} = \mathbf{R}^{(l-1)/2}\mathbf{s}$ and note that it can be restricted to compact and convex set by means of the bounds on $\mathbf{s}$ and $\mathbf{R}^{(l-1)}$. We reintroduce the constraint with a Lagrange multiplier $\frac{\rho_{1,l}}{\sqrt{p_l p_{l-1}}}\boldsymbol{\eta}$

$$\min_{\boldsymbol{\mu}\in S_1^{(l)}} \max_{\mathbf{s}\in S_2^{(l)},\mathbf{v}\in S_3^{(l)}} \frac{\gamma_3\rho_{1,l}^2}{p_l p_{l-1}} \|\boldsymbol{\mu}\| \mathbf{v}^T\mathbf{g}_4 + \frac{\gamma_3\rho_{1,l}^2}{p_l p_{l-1}} \|\mathbf{v}\| \boldsymbol{\mu}^T\mathbf{g}_5 + \frac{\gamma_4\rho_{1,l}}{p_l\sqrt{p_{l-1}}}\mathbf{s}^T\mathbf{g}_2$$
$$-\frac{\gamma_3\rho_{1,l}^2}{2p_l p_{l-1}}\|\mathbf{s}\|^2 - \frac{\gamma_4^2}{2\gamma_3 p_l}\|\mathbf{g}_2\| + \frac{\gamma_1}{2p_l}\|\boldsymbol{\mu}\|^2 + \frac{\gamma_2}{p_l}\mathbf{g}_2^T\boldsymbol{\mu} + \frac{\gamma_3\rho_{2,l}^2}{2p_l}\|\boldsymbol{\mu}\|^2 + \frac{\gamma_4\rho_{2,l}}{p_l}\mathbf{g}_3^T\boldsymbol{\mu}$$
$$+\frac{\rho_{1,l}}{\sqrt{p_l}p_{l-1}}\boldsymbol{\eta}^T\mathbf{v} - \frac{\rho_{1,l}}{\sqrt{p_l}p_{l-1}}\boldsymbol{\eta}^T\mathbf{R}^{(l-1)/2}\mathbf{s}. \tag{217}$$

We then let $\xi_l = \rho_{1,l}\|\mathbf{s}\|/\sqrt{p_l p_{l-1}}$ and let $\chi_l = \rho_{1,l}\|\mathbf{v}\|/\sqrt{p_l p_{l-1}}$ and solve the optimizations over $\mathbf{s}$ and $\mathbf{v}$, from which we obtain:

$$\min_{\boldsymbol{\mu}\in S_1^{(l)}} \max_{0\le\xi_l\le\xi_{l,max},0\le\chi_l\le\chi_{l,max}} \frac{\gamma_3\rho_{1,l}\chi_l}{\sqrt{p_l p_{l-1}}}\mathbf{g}_5^T\boldsymbol{\mu} + \chi_l\left\|\frac{\gamma_3\rho_{1,l}}{\sqrt{p_l p_{l-1}}}\|\boldsymbol{\mu}\|\mathbf{g}_4 + \frac{\rho_{1,l}}{\sqrt{p_{l-1}}}\boldsymbol{\eta}\right\|$$
$$-\frac{\gamma_3\xi^2}{2} + \xi_l\left\|\frac{\gamma_4}{\sqrt{p_l}}\mathbf{g}_2 - \frac{\rho_{1,l}}{\sqrt{p_{l-1}}}\mathbf{R}^{(l-1)/2}\boldsymbol{\eta}\right\|$$
$$-\frac{\gamma_4^2}{2\gamma_3 p_l}\|\mathbf{g}_2\| + \frac{\gamma_1}{2p_l}\|\boldsymbol{\mu}\|^2 + \frac{\gamma_2}{p_l}\mathbf{g}_2^T\boldsymbol{\mu} + \frac{\gamma_3\rho_{2,l}^2}{2p_l}\|\boldsymbol{\mu}\|^2 + \frac{\gamma_4\rho_{2,l}}{p_l}\mathbf{g}_3^T\boldsymbol{\mu}. \tag{218}$$

We interchange the order of the min and max and then make use of the square root trick twice introducing new variables t_l, k_l. We obtain

$$\max_{0\le\xi_l\le\xi_{l,max},0\le\chi_l\le\chi_{l,max}} \min_{0\le t_l\le t_{l,max},0\le k_l\le k_{l,max}} \min_{\boldsymbol{\mu}\in S_1^{(l)},\boldsymbol{\eta}}$$
$$\frac{\gamma_3\rho_{1,l}\chi_l}{\sqrt{p_l p_{l-1}}}\mathbf{g}_5^T\boldsymbol{\mu} + \frac{\chi_l k_l}{2} + \frac{\gamma_3^2\rho_{1,l}^2\chi_l}{2k_l p_l p_{l-1}}\|\boldsymbol{\mu}\|^2\|\mathbf{g}_4\|^2 + \frac{\gamma_3\rho_{1,l}^2\chi_l}{2k_l p_{l-1}\sqrt{p_l}}\|\boldsymbol{\mu}\|\mathbf{g}_4^T\boldsymbol{\eta} + \frac{\chi_l\rho_{1,l}^2}{2k_l p_{l-1}}\|\boldsymbol{\eta}\|^2$$
$$-\frac{\gamma_3\xi^2}{2} + \frac{\xi_l t_l}{2} + \frac{\xi_l\gamma_4^2}{2t_l p_l}\|\mathbf{g}_2\|^2 - \frac{\xi_l\gamma_4\rho_{1,l}}{2t_l\sqrt{p_l p_{l-1}}}\mathbf{g}_2^T\mathbf{R}^{(l-1)/2}\boldsymbol{\eta} + \frac{\xi_l\rho_{1,l}^2}{2t_l p_{l-1}}\boldsymbol{\eta}^T\mathbf{R}^{(l-1)}\boldsymbol{\eta}$$
$$-\frac{\gamma_4^2}{2\gamma_3 p_l}\|\mathbf{g}_2\| + \frac{\gamma_1}{2p_l}\|\boldsymbol{\mu}\|^2 + \frac{\gamma_2}{p_l}\mathbf{g}_2^T\boldsymbol{\mu} + \frac{\gamma_3\rho_{2,l}^2}{2p_l}\|\boldsymbol{\mu}\|^2 + \frac{\gamma_4\rho_{2,l}}{p_l}\mathbf{g}_3^T\boldsymbol{\mu}. \tag{219}$$

Now let $\boldsymbol{\alpha} = \|\boldsymbol{\mu}\| / \sqrt{p_l}$ and solve over $\boldsymbol{\mu}$, from this we obtain

$$\begin{aligned}
&\max_{0\leq\xi_l\leq\xi_{l,max},0\leq\chi_l\leq\chi_{l,max}} \min_{0\leq t_l\leq t_{l,max},0\leq k_l\leq k_{l,max}} \min_{\alpha\leq\alpha_{max},\boldsymbol{\eta}} \\
&\frac{\chi_l k_l}{2} + \frac{\gamma_3^2\rho_{1,l}^2\chi_l\alpha^2}{2k_l p_{l-1}}\|\mathbf{g}_4\|^2 + \frac{\gamma_3\rho_{1,l}^2\chi_l\alpha}{2k_l p_{l-1}}\mathbf{g}_4^T\boldsymbol{\eta} + \frac{\chi_l\rho_{1,l}^2}{2k_l p_{l-1}}\|\boldsymbol{\eta}\|^2 \\
&-\frac{\gamma_3\xi^2}{2} + \frac{\xi_l t_l}{2} + \frac{\xi_l\gamma_4^2}{2t_l p_l}\|\mathbf{g}_2\|^2 - \frac{\xi_l\gamma_4\rho_{1,l}}{2t_l\sqrt{p_l p_{l-1}}}\mathbf{g}_2^T\mathbf{R}^{(l-1)/2}\boldsymbol{\eta} + \frac{\xi_l\rho_{1,l}^2}{2t_l p_{l-1}}\boldsymbol{\eta}^T\mathbf{R}^{(l-1)}\boldsymbol{\eta} \\
&-\frac{\gamma_4^2}{2\gamma_3 p_l}\|\mathbf{g}_2\| + \frac{\gamma_1\alpha^2}{2} + +\frac{\gamma_3\rho_{2,l}^2\alpha^2}{2} + \alpha\left\|\frac{\gamma_4\rho_{2,l}}{\sqrt{p_l}}\mathbf{g}_3 + \frac{\gamma_3\rho_{1,l}\chi_l}{\sqrt{p_{l-1}}}\mathbf{g}_5 + \frac{\gamma_2}{\sqrt{p_l}}\mathbf{g}_2\right\|. \qquad (220)
\end{aligned}$$

This now using the same arguments as lemma 19 this problem concentrates on:

$$\begin{aligned}
&\max_{0\leq\xi_l\leq\xi_{l,max},0\leq\chi_l\leq\chi_{l,max}} \min_{0\leq t_l\leq t_{l,max},0\leq k_l\leq k_{l,max}} \min_{\alpha\leq\alpha_{max},\boldsymbol{\eta}} \\
&\frac{\chi_l k_l}{2} + \frac{\gamma_3^2\rho_{1,l}^2\chi_l\alpha^2}{2k_l} + \frac{\gamma_3\rho_{1,l}^2\chi_l\alpha}{2k_l p_{l-1}}\mathbf{g}_4^T\boldsymbol{\eta} + \frac{\chi_l\rho_{1,l}^2}{2k_l p_{l-1}}\|\boldsymbol{\eta}\|^2 \\
&-\frac{\gamma_3\xi^2}{2} + \frac{\xi_l t_l}{2} + \frac{\xi_l\gamma_4^2 p_{l-1}}{2t_l p_l} - \frac{\xi_l\gamma_4\rho_{1,l}}{2t_l\sqrt{p_l p_{l-1}}}\mathbf{g}_2^T\mathbf{R}^{(l-1)/2}\boldsymbol{\eta} + \frac{\xi_l\rho_{1,l}^2}{2t_l p_{l-1}}\boldsymbol{\eta}^T\mathbf{R}^{(l-1)}\boldsymbol{\eta} \\
&-\frac{\gamma_4^2 p_{l-1}}{2\gamma_3 p_l} + \frac{\gamma_1\alpha^2}{2} + \frac{\gamma_3\rho_{2,l}^2\alpha^2}{2} + \alpha\left(\gamma_4^2\rho_{2,l}^2 + \frac{\gamma_3^2\rho_{1,l}^2\chi_l^2 p_l}{p_{l-1}} + \gamma_2\right)^{1/2}. \qquad (221)
\end{aligned}$$

Examining just the optimization over α we see that this may be solved explicitly:

$$\min_{\alpha}\left(\frac{\gamma_1}{2} + \frac{\gamma_3^2\rho_{1,l}^2\chi_l}{2k_l} + \frac{\gamma_3\rho_{2,l}^2}{2}\right)\alpha^2 + \left(\left(\gamma_4^2\rho_{2,l}^2 + \frac{\gamma_3^2\rho_{1,l}^2\chi_l^2 p_l}{p_{l-1}} + \gamma_2\right)^{1/2} + \frac{\gamma_3\rho_{1,l}^2\chi_l}{2k_l p_{l-1}}\mathbf{g}_4^T\boldsymbol{\eta}\right)\alpha \qquad (222)$$

Which has optimal value

$$\begin{aligned}
&-\left(\gamma_1 + \frac{\gamma_3^2\rho_{1,l}^2\chi_l}{k_l} + \gamma_3\rho_{2,l}^2\right)^{-1}\left(\gamma_4^2\rho_{2,l}^2 + \frac{\gamma_3^2\rho_{1,l}^2\chi_l^2 p_l}{p_{l-1}} + \gamma_2\right. \\
&\left.+\left(\gamma_4^2\rho_{2,l}^2 + \frac{\gamma_3^2\rho_{1,l}^2\chi_l^2 p_l}{p_{l-1}} + \gamma_2\right)^{1/2}\frac{\gamma_3\rho_{1,l}^2\chi_l}{2k_l p_{l-1}}\mathbf{g}_4^T\boldsymbol{\eta} + \frac{\gamma_3^2\rho_{1,l}^4\xi_l^2}{4k_l^2 2p_{l-1}}\|\boldsymbol{\eta}\|^2\right) \qquad (223)
\end{aligned}$$

As such we can collect all of the terms together. Making the following definitions:

$$T_l = \frac{\chi_l k_l}{2} - \frac{\gamma_3 \xi_l^2}{2} + \frac{\xi_l t_l}{2} + \frac{\xi_l \gamma_4^2 p_{l-1}}{2 t_l p_l} - \frac{\gamma_4^2 p_{l-1}}{2\gamma_3 p_l}$$
$$- \left(\gamma_1 + \frac{\gamma_3^2 \rho_{1,l}^2 \chi_l}{k_l} + \gamma_3 \rho_{2,l}^2\right)^{-1} \left(\gamma_4^2 \rho_{2,l}^2 + \frac{\gamma_3^2 \rho_{1,l}^2 \chi_l^2 p_l}{p_{l-1}} + \gamma_2\right) \tag{224}$$

$$\bar{\gamma_1} = \frac{\xi_l \rho_{1,l}^2}{k_l} - \left(\gamma_1 + \frac{\gamma_3^2 \rho_{1,l}^2 \chi_l}{k_l} + \gamma_3 \rho_{2,l}^2\right)^{-1} \frac{\gamma_3^2 \rho_{1,l}^4 \xi_l^2}{2 k_l^2} \tag{225}$$

$$\bar{\gamma_2} = -\left(\gamma_1 + \frac{\gamma_3^2 \rho_{1,l}^2 \chi_l}{k_l} + \gamma_3 \rho_{2,l}^2\right)^{-1} \left(\gamma_4^2 \rho_{2,l}^2 + \frac{\gamma_3^2 \rho_{1,l}^2 \chi_l^2 p_l}{p_{l-1}} + \gamma_2\right)^{1/2} \frac{\gamma_3 \rho_{1,l}^2 \chi_l}{2 k_l} \tag{226}$$

$$\bar{\gamma_3} = \frac{\xi_l \rho_{1,l}^2}{t_l} \qquad \bar{\gamma}_4 = \frac{\xi_l \gamma_4 \rho_{1,l} \sqrt{p_l}}{2 t_l \sqrt{p_{l-1}}}. \tag{227}$$

As such we find that the optimization is equal to

$$\max_{0 \le \xi_l \le \xi_{l,max}, 0 \le \chi_l \le \chi_{l,max}} \min_{0 \le t_l \le t_{l,max}, 0 \le k_l \le k_{l,max}} T_l$$
$$+ \min_{\boldsymbol{\eta}} \frac{\bar{\gamma}_1}{2 p_l} \|\boldsymbol{\mu}\|^2 + \frac{\bar{\gamma}_2}{p_l} \mathbf{g}_2^T \boldsymbol{\mu} + \frac{\bar{\gamma}_3}{2 p_l} \boldsymbol{\mu}^T \mathbf{R}^{(l-1)} \boldsymbol{\mu} + \frac{\bar{\gamma}_4}{p_l} \mathbf{g}_3^T \mathbf{R}^{(l)/2} \boldsymbol{\mu}. \tag{228}$$

■

D.3. All Layers of Same Size

Consider the case that the input dimension is d and all subsequent hidden layers are of dimension p. In this case we note that $\mathbf{R}^{(l)} \in \mathbb{R}^{p \times p}$ for all $l > 1$. In this case the recursive application of the CGMT analysis simplifies considerably. The recursion is given in the following lemma.

Theorem 22 *Consider the problem P_2 given in (24) and assume that the layers $p_1 = p_2 = \cdots p_L = p$, ie all layers are of the same size. Let the input dimension be of size p_0 which is not necessarily the same as p. In this case the alternative optimization problem may be given by:*

$$\max_{\beta>0} \min_{q>0} \max_{\xi_1>0} \min_{t_1>0} \cdots \max_{\xi_L>0} \min_{t_L>0} \mathcal{M}_{\frac{p}{C} R(\cdot + \boldsymbol{\theta}^*)} \left(-\frac{D}{C}\mathbf{g}\right) + T_L \tag{229}$$

Where

$$c_0 = \frac{\beta}{q} \qquad d_0 = \beta \sqrt{\frac{n}{p}} \qquad T_0 = \frac{\beta q}{2} + \frac{\beta \sigma_{\boldsymbol{\nu}}^2}{2q} - \frac{\beta^2}{q} \tag{230}$$

$$c_{l+1} = \frac{\xi_l c_l^2 \rho_{1,l}^2}{t_l} \qquad d_{l+1} = c_l^2 \xi_l^2 \rho_{1,l}^2 \frac{p_{L-l-1}}{p_{L-l}} \tag{231}$$

$$C = c_L + \sum_{l=0^{L-1}} \rho_{2,L-l}^2 c_L \qquad D = \sqrt{d_L^2 + \sum_{l=0}^{L-1} \rho_{2,L-l}^2 d_l}$$

$$T_{l+1} = T_l + \frac{d_l^2 \rho_{1,l}^2 \xi_l}{2 t_l} \frac{p_{L-l-1}}{p_{L-l}} - \frac{c_l \xi_l^2}{2} + \frac{\xi_l t_l}{2} - \frac{d_l^2}{2 c_l} \frac{p_{L-l-1}}{p_{L-l}} \tag{232}$$

Note that as $p_1 = p_2 = \cdots p_L = p$ *the value of* $\frac{p_{L-l-1}}{p_{L-l}} = 1$ *except in the case of* $p_0 = d$.

Proof The proof is the same as the one given for the CGMT analysis for layers of different sizes. We therefore do not give it here in full. ■

Appendix E. Lyapunov Recursions

Let $\mathbf{A}$ be a $n \times m$ matrix with random entries. Consider the function $f_A(\lambda)$ with gives the probability distribution, or *eigendistribution*, of the eigenvalues of the matrix A, defined to be

$$f_{\mathbf{A}}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i} \tag{233}$$

where λ_i is the ith eigenvalue of A and δ is the dirac measure .

To analyze this distribution, we may instead consider the Stieltjes transform of the distribution f_A, this transform is defined by

$$S_{\mathbf{A}}(z) = \mathbb{E}\left[\frac{1}{\lambda - z}\right] = \int \frac{f_{\mathbf{A}}(\lambda)}{\lambda - z} d\lambda. \tag{234}$$

Here z is a complex number. The original distribution may be recovered by means of the inverse transform

$$f_{\mathbf{A}}(\lambda) = \lim_{\omega \to 0^+} \frac{1}{\pi} \mathrm{Im}\left[S_{\mathbf{A}}(\lambda + i\omega)\right] \tag{235}$$

where i is the imaginary unit. The Stieltjes transform can also be compute directly from the random matrix $\mathbf{A}$ instead of using equation (234). We give the following lemma

Lemma 23 *The Stieltjes transform of the expected eigendistribution of a Hermitian random* $n \times n$ *matrix* $\mathbf{A}$ *may be expressed as*

$$S_{\mathbf{A}}(z) = \frac{1}{n} \mathbb{E}\, Tr(\mathbf{A} - z\mathbf{I})^{-1} \tag{236}$$

Proof For a proof see Vakili (2011) [lemma 2.3.1] ■

Another transform that we will make use of in our analysis of the recursively defined matrix $\mathbf{R}$ is the S-transform may be expressed in terms of the Stieltjes transform by means of

$$\Sigma_{\mathbf{A}}(z) = \frac{z+1}{z}\left(-\frac{1}{z} S_{\mathbf{A}}\left(\frac{1}{z}\right) - 1\right)^{\{-1\}} = \frac{z+1}{z}\left(\sum_{i=1}^{\infty} m_i z^i\right)^{\{-1\}} \tag{237}$$

Here $\{-1\}$ denotes the functional inverse, and m_i is the ith moment of the distribution $f_{\mathbf{A}}$. The S-transform has two properties that are instrumental for our analysis. Firstly, the S-transform and the Stieltjes transform satisfy the following relation:

$$\Sigma_{\mathbf{A}} = -\frac{1}{z} S_{\mathbf{A}}\left(\frac{1+z}{z \Sigma_{\mathbf{A}}(z)}\right) \tag{238}$$

The second key property of the S-transform relates it how it behaves with respect to matrix product. For this we introduce the following lemma

Lemma 24 *Let $\mathbf{A}, \boldsymbol{B}$ be two non negative unitarily invariant matrices, and let $\boldsymbol{C} = \mathbf{A}\boldsymbol{B}$, then the S transform of the eigendistribution of $\boldsymbol{C}$ satisfies*

$$\Sigma_{\boldsymbol{C}}(z) = \Sigma_{\mathbf{A}}(z)\Sigma_{\boldsymbol{B}}(z) \tag{239}$$

Proof The S-transform is multiplicative for matrix product that are asymptotically free Emery et al. (2007). To see that unitarily invariant matrices are free see Voiculescu (1991). ■

Finally, we note that if $\mathbf{H}$ is a $m \times n$ matrix with element distributed as $\mathcal{N}(0,1)$, then the matrix $\mathbf{A} = \frac{1}{n}\mathbf{H}\mathbf{H}^T$ is a Wishart matrix. We note that the Stieltjes transform of a Wishart matrix is given by the Marcenko-Pastur Law

$$S_{\mathbf{A}}(z) = \frac{1 - \frac{m}{n} - z + \sqrt{z^2 - 2\left(\frac{m}{n}+1\right)z + \left(\frac{m}{n}-1\right)^2}}{\frac{2m}{n}z} \tag{240}$$

and the $S-$transform of a Wishart is given by

$$\Sigma_{\mathbf{A}}(z) = \frac{1}{1 + \frac{m}{n}z} \tag{241}$$

E.1. Analysis of the Covariance Matrix R

In this section we adopt an approach for studying Stieltjes transforms of Lyapanov Recursions of Random matrices discussed by Vakili (2011)[Section 3].

We recall that $\mathbf{R}^{(l)}$ is given by

$$\mathbf{R}^{(l)} = \frac{\rho_{1,l}^2}{p_{l-1}}\mathbf{W}^{(l)}\mathbf{R}^{(l-1)}\mathbf{W}^{(l)T} + (\rho_{2,l})^2\,\mathbf{I}. \tag{242}$$

where we recall that $\mathbf{R}^{(0)} = \mathbf{I}$ and that the rows of $\mathbf{W}^{(l)}$, $\mathbf{w}_i^{(l)}$ are distributed as $\mathcal{N}(0, \mathbf{I})$. We can note that $\mathbf{W}^{(l)T}\mathbf{W}^{(l)}/p_{l-1}$ is a Wishart matrix. We now wish to compute the Stieltjes transform of $\mathbf{R}^{(l)}$. The Stieltjes transform is given by

$$S_{\mathbf{R}^{(l)}}(z) = \frac{1}{p_l}\mathbb{E}\mathrm{Tr}\left(\frac{\rho_{1,l}^2}{p_{l-1}}\mathbf{W}^{(l)}\mathbf{R}^{(l-1)}\mathbf{W}^{(l)T} + (\rho_{2,l}^2 - z)\mathbf{I}\right)^{-1} \tag{243}$$

We now let the matrix $\mathbf{A}^{(l)} = \mathbf{W}^{(l)T}\mathbf{R}^{(l-1)}\mathbf{W}^{(l)T}/p_{l-1}$. We can then note that

$$\begin{aligned} S_{\mathbf{R}^{(l)}}(z) = \frac{1}{p_l}\mathbb{E}\mathrm{Tr}\left(\rho_{1,l}^2\mathbf{A}^{(l)} + (\rho_{2,l}^2 - z)\mathbf{I}\right)^{-1} &= \frac{1}{\rho_{1,l}^2}\frac{1}{p_l}\mathbb{E}\mathrm{Tr}\left(\mathbf{A}^{(l)} + \frac{(\rho_{2,l}^2 - z)}{\rho_{1,l}^2}\mathbf{I}\right)^{-1} \\ &= \frac{1}{\rho_{1,l}^2}S_{\mathbf{A}^{(l)}}\left(\frac{z - \rho_{2,l}^2}{\rho_{1,l}^2}\right) \end{aligned} \tag{244}$$

Our goal is to now find an expression for the Stieltjes transform of $\mathbf{A}^{(l)}$. We note that $\mathbf{W}^{(l)T}\mathbf{R}^{(l-1)}\mathbf{W}^{(l)T}/p_{l-1}$ has the same eigenvalues as $\mathbf{W}^{(l)T}\mathbf{W}^{(l)T}/p_{l-1}\mathbf{R}^{(l-1)}$. we recall that $\mathbf{W}^{(l)T}\mathbf{W}^{(l)T}/p_{l-1}$ is Wishart

and unitarily Invariant, and similarly is $\mathbf{R}^{(l-1)}$. As such we can make use of the properties of $S-$transforms to note that:

$$\Sigma_{\mathbf{A}^{(l)}}(z) = \Sigma_{\mathbf{W}^{(l)T}\mathbf{W}^{(l)T}/p_l}(z)\Sigma_{\mathbf{R}^{(l-1)}}(z). \tag{245}$$

We can then make use of equation (238) to obtain

$$\begin{aligned} S_{\mathbf{A}^{(l)}}\left(\frac{1+z}{z\Sigma_{\mathbf{A}^{(l)}}}(z)\right) &= \Sigma_{\mathbf{W}^{(l)T}\mathbf{W}^{(l)T}/p_{l-}}(z)S_{\mathbf{R}^{(l-1)}}\left(\frac{1+z}{z\Sigma_{\mathbf{R}^{(l-1)}}(z)}\right) \\ &= \Sigma_{\mathbf{W}^{(l)T}\mathbf{W}^{(l)T}/p_{l-1}}(z)S_{\mathbf{R}^{(l-1)}}\left(\frac{1+z}{z\Sigma_{\mathbf{A}^{(l)}}(z)}\Sigma_{\mathbf{W}^{(l)T}\mathbf{W}^{(l)T}/p_{l-1}}(z)\right) \end{aligned} \tag{246}$$

We now let

$$x = \frac{1+z}{z\Sigma_{\mathbf{A}^{(l)}}(z)}, \tag{247}$$

and then note that

$$\begin{aligned} x\Sigma_{\mathbf{A}^{(l)}}(z) = \frac{1+z}{z} \Rightarrow x\left(\frac{-1}{z}\right)S_{\mathbf{A}^{(l)}}(x) &= \frac{1+z}{z} \\ \Rightarrow z = -1 - xS_{\mathbf{A}^{(l)}}(x) & \end{aligned} \tag{248}$$

By substituting in this expression we obtain

$$S_{\mathbf{A}^{(l)}}(x) = \Sigma_{\mathbf{W}^{(l)T}\mathbf{W}^{(l)T}/p_{l-1}}(-1-xS_{\mathbf{A}^{(l)}}(x))S_{\mathbf{R}^{(l-1)}}\left(x\Sigma_{\mathbf{W}^{(l)T}\mathbf{W}^{(l)T}/p_{l-1}}(-1-xS_{\mathbf{A}^{(l)}}(x))\right) \tag{249}$$

Finally, we recall equation (241). Letting $\beta_l = \frac{p_l}{p_{l-1}}$ we use this property to simplify the relation to:

$$S_{\mathbf{A}^{(l)}}(x) = \frac{1}{1-\beta_l-\beta_l x S_{\mathbf{A}^{(l)}}(x)}S_{\mathbf{R}^{(l-1)}}\left(\frac{x}{1-\beta_l-\beta_l x S_{\mathbf{A}^{(l)}}(x)}\right) \tag{250}$$

Finally, letting $\Omega_{l-1}(\cdot) = S_{\mathbf{A}^{(l)}}(\cdot)$. We can conclude that

$$S_{\mathbf{R}^{(l+1)}}(z) = \frac{1}{\rho_{1,l}^2}\Omega_l\left(\frac{z-\rho_{2,l}^2}{\rho_{1,l}^2}\right) \tag{251}$$

$$\Omega_l(z) = \frac{1}{1-\beta_l-z\beta_l\Omega_l(z)}S_{\mathbf{R}^{(l)}}\left(\frac{z}{1-\beta_l-\beta_l z\Omega_l(z)}\right) \tag{252}$$

Which concludes the proof.