On the Inherent Regularization Effects of Noise Injection During Training

Oussama Dhifallah [1] **Yue M. Lu** [1]

Abstract

Randomly perturbing networks during the training process is a commonly used approach to improving generalization performance. In this paper, we present a theoretical study of one particular way of random perturbation, which corresponds to injecting artificial noise to the training data. We provide a precise asymptotic characterization of the training and generalization errors of such randomly perturbed learning problems on a random feature model. Our analysis shows that Gaussian noise injection in the training process is equivalent to introducing a weighted ridge regularization, when the number of noise injections tends to infinity. The explicit form of the regularization is also given. Numerical results corroborate our asymptotic predictions, showing that they are accurate even in moderate problem dimensions. Our theoretical predictions are based on a new correlated Gaussian equivalence conjecture that generalizes recent results in the study of random feature models.

1. Introduction

A popular approach to improving the generalization performance is to randomly perturb the network during the training process (Srivastava et al., 2014; Bishop, 1995; Gulcehre et al., 2016; LeJeune et al., 2020; Kobak et al., 2020). Such random perturbations are widely used as an implicit regularization to the learning problem. One way that random perturbation has been used as a regularization is by injecting it to the input data before starting the learning process (Gong et al., 2020; Rakin et al., 2018; Poole et al., 2014). In this paper, we provide a theoretical analysis of such learning procedure on a random feature model (Rahimi & Recht, 2008) under Gaussian input and perturbation vectors. Our analysis particularly shows that Gaussian noise injection introduces a weighted ridge regularization, asymptotically.

First, we describe the models for our theoretical analysis. We are given a collection of training data $\{(y_i, \boldsymbol{a}_i)\}_{i=1}^n$, where $\boldsymbol{a}_i \in \mathbb{R}^p$ is referred to as the input vector and $y_i \in \mathbb{R}$ is referred to as the label corresponding to $\boldsymbol{a}_i$. In this paper, we shall assume that the labels are generated according to the standard *teacher–student* model, i.e.

$$y_i = \varphi(\boldsymbol{a}_i^\top \boldsymbol{\xi}),\ \forall i \in \{1, \dots, n\}, \tag{1}$$

where $\boldsymbol{\xi} \in \mathbb{R}^p$ is an unknown teacher weight vector, and $\varphi(\cdot)$ is a scalar deterministic or probabilistic function. Here, we use the random feature model (Rahimi & Recht, 2008) to learn the model described in (1). The random feature model considers the following class of functions

$$\mathcal{F}_{\text{RF}}(\boldsymbol{a}) = \Big\{ g_{\boldsymbol{w}}(\boldsymbol{a}) = \boldsymbol{w}^\top \sigma(\boldsymbol{F}^\top \boldsymbol{a}),\ \boldsymbol{w} \in \mathbb{R}^k \Big\}, \tag{2}$$

where $\boldsymbol{a} \in \mathbb{R}^p$ is an input vector, $\boldsymbol{F} \in \mathbb{R}^{p \times k}$ is a random matrix referred to as the *feature matrix*, and $\sigma(\cdot)$ is a scalar function referred to as the *activation function*. This model assumes that $\boldsymbol{F}$ is fixed during the training. Note that the family in (2) can be viewed as a two–layer neural network where the first layer weights are fixed, i.e. $\boldsymbol{F}$ is fixed.

1.1. Learning Formulation

Before starting the learning process, ℓ independent perturbation vectors are injected to each $\boldsymbol{a}_i$. This procedure forms the augmented family $\{\boldsymbol{a}_i + \Delta \boldsymbol{z}_{ij}\}_{j=1}^\ell$ for each $\boldsymbol{a}_i$, where $\{\boldsymbol{z}_{ij}\}_{j=1}^\ell$ are independent random perturbations and $\Delta \geq 0$ denotes the *noise variance*. In this paper, we study the effects of such perturbation method on an average loss and a random feature model. Specifically, we analyze formulations of the following form

$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{w} \in \mathbb{R}^k}{\operatorname{argmin}} \frac{1}{2n\ell} \sum_{i=1}^n \sum_{j=1}^\ell \big(y_i - \boldsymbol{w}^\top \sigma\big(\boldsymbol{F}^\top [\boldsymbol{a}_i + \Delta \boldsymbol{z}_{ij}]\big)\big)^2 + \tfrac{\lambda}{2} \|\boldsymbol{w}\|^2, \tag{3}$$

where $\lambda > 0$ denotes the regularization parameter. Note that the problem in (3) is a standard feature formulation when

[1]O. Dhifallah and Y. M. Lu are with the John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA.. Correspondence to: Oussama Dhifallah <oussama_dhifallah@g.harvard.edu>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

$\Delta = 0$. Then, we refer to (3) as the *noisy formulation*, when $\Delta > 0$ and the *standard formulation*, otherwise.

1.2. Performance Measure

The main objective in this paper is to study the performance of the learning formulation in (3) on unobserved test data. For every test vector $\boldsymbol{a}_{\text{new}} \in \mathbb{R}^p$, the corresponding label $\widehat{y}$ can be predicted using the following (probabilistic) role

$$\widehat{y} = \widehat{\varphi}[\widehat{\boldsymbol{w}}^\top \sigma(\boldsymbol{F}^\top \boldsymbol{a}_{\text{new}})], \tag{4}$$

for some predefined function $\widehat{\varphi}(\cdot)$, where $\widehat{\boldsymbol{w}} \in \mathbb{R}^k$ denotes the optimal solution of the formulation given in (3). To measure the performance of the learning problem in (3) on any unobserved test data $\{(y_{\text{new}}, \boldsymbol{a}_{\text{new}})\}$, we use the *generalization error* defined as follows

$$\mathcal{E}_{\text{test}} = \frac{1}{4^\upsilon}\mathbb{E}\Big[\big(y_{\text{new}} - \widehat{\varphi}(\widehat{\boldsymbol{w}}^\top \sigma(\boldsymbol{F}^\top \boldsymbol{a}_{\text{new}}))\big)^2\Big]. \tag{5}$$

Here, the expectation is taken over the distribution of the unobserved test vector $\boldsymbol{a}_{\text{new}}$ and the (random) functions $\varphi(\cdot)$ and $\widehat{\varphi}(\cdot)$. We take $\upsilon = 0$ for regression problems (e.g. $\varphi(\cdot)$ is the identity function) and $\upsilon = 1$ for binary classification problems (e.g. $\varphi(\cdot)$ is the sign function). In this paper, we assume that the test data is generated according to the same training model introduced in (1). Furthermore, we measure the performance of the formulation in (3) on the training data via the *training error* defined as follows

$$\mathcal{E}_{\text{train}} = \frac{1}{2n\ell}\sum_{i=1}^{n}\sum_{j=1}^{\ell}\big(y_i - \widehat{\boldsymbol{w}}^\top \sigma\big(\boldsymbol{F}^\top[\boldsymbol{a}_i + \Delta \boldsymbol{z}_{ij}]\big)\big)^2.$$

Note that the training error is the optimal cost value of our learning formulation in (3) without regularization.

1.3. Contributions

The contribution of this paper can be summarized as follows:

(C.1) Our first contribution is a *correlated Gaussian equivalence conjecture* (cGEC). Our conjecture considers Gaussian input and perturbation vectors. It states that the learning formulation in (3) is asymptotically equivalent to a simpler optimization problem that can be formulated by replacing the non–linear vectors

$$\boldsymbol{v}_{ij} = \sigma\big(\boldsymbol{F}^\top[\boldsymbol{a}_i + \Delta \boldsymbol{z}_{ij}]\big),$$

with linear vectors with the following form

$$\boldsymbol{q}_{ij} = \mu_0 \boldsymbol{1}_k + \widetilde{\mu}_1 \boldsymbol{F}^\top \boldsymbol{a}_i + \widehat{\mu}_1 \boldsymbol{F}^\top \boldsymbol{z}_{ij} + \mu_2 \boldsymbol{b}_i + \mu_3 \boldsymbol{p}_{ij}.$$

Here, $\{\boldsymbol{b}_i\}_{i=1}^n$ and $\{\boldsymbol{p}_{ij}\}_{i,j=1}^{n,\ell}$ are independent standard Gaussian random vectors and independent of $\{\boldsymbol{a}_i\}_{i=1}^n$ and $\{\boldsymbol{z}_{ij}\}_{i,j=1}^{n,\ell}$. Moreover, the weights μ_0, $\widetilde{\mu}_1$, $\widehat{\mu}_1$, μ_2 and μ_3 depend on $\sigma(\cdot)$ and Δ as follows

$$\begin{cases} \mu_0 = \mathbb{E}[\sigma(x_1)], \widetilde{\mu}_1 = \mathbb{E}[z\sigma(x_1)], \widehat{\mu}_1 = \mathbb{E}[v_1\sigma(x_1)] \\ \mu_2^2 = \mathbb{E}[\sigma(x_1)\sigma(x_2)] - \mu_0^2 - \widetilde{\mu}_1^2 \\ \mu_3^2 = \mathbb{E}[\sigma(x_1)^2] - \mathbb{E}[\sigma(x_1)\sigma(x_2)] - \widehat{\mu}_1^2, \end{cases}$$

where $x_1 = z + \Delta v_1$, $x_2 = z + \Delta v_2$, and z, v_1 and v_2 are independent standard Gaussian random variables. Specifically, the cGEC states that the performance of the formulation:

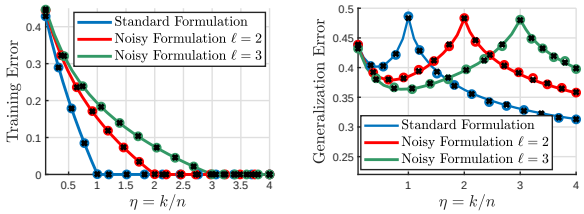
$$\min_{\boldsymbol{w}\in\mathbb{R}^k} \frac{1}{2n\ell}\sum_{i=1}^{n}\sum_{j=1}^{\ell}\big(y_i - \widetilde{\mu}_1 \boldsymbol{w}^\top \boldsymbol{F}^\top \boldsymbol{a}_i - \widehat{\mu}_1 \boldsymbol{w}^\top \boldsymbol{F}^\top \boldsymbol{z}_{ij} - \mu_0 \boldsymbol{w}^\top \boldsymbol{1}_k - \mu_2 \boldsymbol{w}^\top \boldsymbol{b}_i - \mu_3 \boldsymbol{w}^\top \boldsymbol{p}_{ij}\big)^2 + \tfrac{\lambda}{2}\|\boldsymbol{w}\|^2, \tag{6}$$

is asymptotically equivalent to the performance of the noisy formulation. This conjecture is valid in the asymptotic limit (i.e. n, p and k grow to infinity at finite ratios). More details about this equivalence is provided in Section 2. We refer to (6) as the *Gaussian formulation*. The cGEC is verified by presenting multiple simulations in different scenarios.

(C.2) The second contribution is a precise characterization of the training and generalization errors of the noise injection procedure formulated in (3) for Gaussian input and perturbation vectors. The theoretical predictions are obtained using an extended version of the convex Gaussian min-max theorem (CGMT) (Thrampoulidis et al., 2016; 2015). From a purely technical point of view, our analysis technique is novel. Rather than a routine and direct application of the standard CGMT method from previous work, we have developed a new multivariate version of the CGMT that is a significant extension of the existing formulation. Specifically, the standard CGMT method provides precise performance analysis of problems in the following form: $\min_{\boldsymbol{w}\in\mathcal{S}_{\boldsymbol{w}}} \max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}} \boldsymbol{u}^\top \boldsymbol{G}\boldsymbol{w} + \psi(\boldsymbol{w}, \boldsymbol{u})$, where the matrix $\boldsymbol{G}$ has independent standard Gaussian entries and the function $\psi(\cdot,\cdot)$ satisfies convexity assumptions. In our problem, we are dealing with

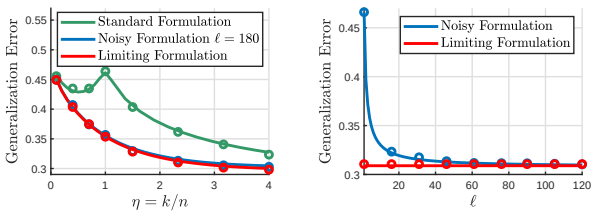
$$\min_{\boldsymbol{w}\in\mathcal{S}_{\boldsymbol{w}}} \max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}} [\boldsymbol{u}_1^\top, \ldots, \boldsymbol{u}_\ell^\top] \begin{bmatrix} \boldsymbol{G}_1 \\ \vdots \\ \boldsymbol{G}_\ell \end{bmatrix} \boldsymbol{w} + \psi(\boldsymbol{w}, \boldsymbol{u}). \tag{7}$$

Here, the matrix $\boldsymbol{G}_j = \boldsymbol{K}\boldsymbol{\Sigma}^{\frac{1}{2}} + \boldsymbol{T}_j\boldsymbol{\Gamma}^{\frac{1}{2}}$, for $j \in \{1, \ldots, \ell\}$, where $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}$ are two different covariance matrices and $\boldsymbol{K}$ and $\{\boldsymbol{T}_j\}_{1\le j\le \ell}$ are all independent standard Gaussian matrices. We can see that every $\boldsymbol{G}_j$ has independent rows. However, any two different matrices $\boldsymbol{G}_i$ and $\boldsymbol{G}_j$ are *dependent* as their constructions share the same matrix $\boldsymbol{K}$. Clearly, the classical CGMT method is not applicable in this case. To our knowledge, our paper provides the first theoretical analysis that can handle such correlation in the input data. We refer to this extended version the *multivariate CGMT*.



(a) (b)

Figure 1. Solid line: Theoretical predictions. Circle: Numerical simulations for (3). Black cross: Numerical simulations for (6). $\varphi(\cdot)$ is the sign function with probability θ of flipping the sign. $\widehat{\varphi}(\cdot)$ and $\sigma(\cdot)$ are the sign function. We set $p = 500$, $\Delta = 0.5$, $\alpha = n/p = 2$, $\theta = 0.1$, $\lambda = 10^{-5}$. $\boldsymbol{F}$ has independent Gaussian components with zero mean and variance $1/p$. The results are averaged over 200 independent Monte Carlo trials.



(a) (b)

Figure 2. Solid line: Theoretical predictions. Circle: Numerical simulations for (3) and (8). $\varphi(\cdot)$, $\widehat{\varphi}(\cdot)$ and $\sigma(\cdot)$ are the sign function. **(a)** $p = 700$, $\ell = 180$, $\alpha = n/p = 1$, $\Delta = 1$ and $\lambda = 10^{-3}$. **(b)** $p = 600$, $\alpha = n/p = 1.5$, $\eta = k/n = 1$, $\Delta = 1$ and $\lambda = 10^{-3}$. $\boldsymbol{F}$ has independent Gaussian components with zero mean and variance $1/p$. The results are averaged over 100 independent Monte Carlo trials.

In Figure 1, we compare our theoretical predictions with the actual performance of the learning problem given in (3). First, note that our asymptotic predictions are in excellent agreement with the actual performance of (3) and its Gaussian formulation given in (6), even for moderate values of p, n and k. This provides a first empirical validation of our results. Figure 1 also study the effects of ℓ on the training and generalization performance. Note that the generalization error follows a double descent curve (Belkin et al., 2018; 2019). Specifically, the generalization error decreases monotonically as a function of the complexity $\eta = k/n$ after reaching a peak known as the interpolation threshold (Belkin et al., 2018; 2019). Figure 1(b) particularly demonstrates that the location of the interpolation threshold depends on the number of noise samples. Specifically, the interpolation threshold peak occurs at ℓ for fixed noise variance $\Delta = 0.5$. Additionally, Figure 1(a) shows that the interpolation threshold occurs when the training error converges to zero. Then, we can see that perturbing the input data with ℓ random noise vectors moves the interpolation threshold from 1 to ℓ and improves the generalization error for complexity η lower than ℓ.

(C.3) The third contribution is a precise analysis of the regularization effects of the considered noise injection procedure. Specifically, we use the asymptotic predictions of the noisy formulation to show that the noise injection model in (3) is equivalent to solving a standard feature formulation with an additional weighted ridge regularization. This theoretical result is valid when the number of noise samples ℓ tends to infinity. In particular, we show that the formulation in (3) is equivalent to solving the problem

$$\min_{\boldsymbol{w}\in\mathbb{R}^k} \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \boldsymbol{w}^\top\widehat{\sigma}(\boldsymbol{F}^\top\boldsymbol{a}_i)\right)^2 + \frac{1}{2}\|\boldsymbol{R}^{\frac{1}{2}}\boldsymbol{w}\|^2 + \tfrac{\lambda}{2}\|\boldsymbol{w}\|^2, \tag{8}$$

when ℓ grows to infinity slower than the dimensions n, p and k. Here, $\widehat{\sigma}(\cdot)$ is a new activation function and $\boldsymbol{R}$ is defined as follows

$$\boldsymbol{R} = \widehat{\mu}_1^2\boldsymbol{F}^\top\boldsymbol{F} + \mu_3^2\boldsymbol{I}_k. \tag{9}$$

Finally, we provide a precise asymptotic characterization of the training and generalization errors corresponding to (8). We refer to this formulation as the *limiting formulation*.

Figure 2 provides another empirical verification of our theoretical predictions since it shows that they are in excellent agreement with the actual performance of (3) and (8). Figure 2(a) shows that the noisy formulation in (3) has approximately the same performance as the formulation in (8) for $\ell = 180$. This is aligned with our theoretical prediction which states that the formulations in (3) and (8) are equivalent when ℓ grows to infinity slower than the dimensions n, p and k. Figure 2(b) illustrates the convergence behavior of the generalization error corresponding to (3) for a fixed value of η. It particularly shows that the noisy formulation has a good *convergence rate*, i.e. the limit is already attained with a moderate value of ℓ. Moreover, we can see from Figure 2(a) that the convergence rate depends on the complexity parameter η.

1.4. Related Work

There has been significant interest in precisely characterizing the performance of the random feature model in recent literature (Mei & Montanari, 2019; Gerace et al., 2020; Dhifallah & Lu, 2020; Hu & Lu, 2020). The ridge regression formulation, (i.e. $\varphi(\cdot)$ is the identity function and $\Delta = 0$ in (3)) is precisely analyzed in (Mei & Montanari, 2019) where the feature matrix is Gaussian. In a subsequent work, (Montanari et al., 2019) uses the CGMT to accurately analyze the maximum-margin linear classifier in the overparametrized

regime. The work in (Gerace et al., 2020) precisely characterizes the performance of the standard formulation, i.e. $\Delta = 0$, for general families of feature matrices and convex loss functions. The results presented in (Gerace et al., 2020) are derived using the non–rigorous replica method (Mezard et al., 1986). The predictions in (Gerace et al., 2020) are rigiourously verified in (Dhifallah & Lu, 2020) using the CGMT. All the previous work consider an unperturbed formulation of the random feature model. In this paper, we study the effects of adding random noise during training. Our analysis is based on an extended version of the CGMT referred to as the multivariate CGMT. The CGMT is first used in (Stojnic, 2013) and further developed in (Thrampoulidis et al., 2016). It extends a Gaussian theorem first introduced in (Gordon, 1988). It relies on (strong) convexity properties to prove an equivalence between two Gaussian processes. It has been successfully applied in the analysis of convex regression (Thrampoulidis et al., 2016; Dhifallah et al., 2018; Dhifallah & Lu, 2020) and convex classification (Salehi et al., 2019; Sifaou et al., 2019; Mignacco et al., 2020; Dhifallah & Lu, 2021) formulations.

There has been significant interest in studying the effects of random noise injection during training (see e.g. (Bishop, 1995; An, 1996; Gulcehre et al., 2016)). In particular, prior literature (Zantedeschi et al., 2017; Kannan et al., 2018) shows that Gaussian noise injection during training improves the robustness of the network. Moreover, several recent papers (Bishop, 1995; Gong et al., 2020) show that such perturbation technique introduces some sort of regularization to the loss function. In particular, the work in (Gong et al., 2020) shows that minimizing the worst–case loss introduces a gradient norm regularization.

Another popular perturbation approach used in regularizing learning models is the *dropout* method (Srivastava et al., 2014; Wei et al., 2020). It consists of perturbing the learning problem by randomly dropping units from the network during the training procedure. In this paper, we precisely analyze the Gaussian noise injection method and we leave the analysis of the dropout technique for future work. Our empirical studies suggest that the dropout method has a better convergence rate as compared to the noisy formulation. Moreover, they suggest that both methods have comparable generalization performance.

2. Gaussian Equivalence Conjecture with an Intuitive Explanation

Consider three independent standard Gaussian random vectors $\boldsymbol{a} \in \mathbb{R}^p$, $\boldsymbol{z}_1 \in \mathbb{R}^p$ and $\boldsymbol{z}_2 \in \mathbb{R}^p$. Moreover, consider the random variables $\nu_1 = \boldsymbol{\xi}^\top \boldsymbol{a}$, ν_2 and ν_3 defined as follows

$$\nu_2 = \boldsymbol{w}^\top \sigma(\boldsymbol{F}^\top[\boldsymbol{a} + \Delta \boldsymbol{z}_1]), \nu_3 = \boldsymbol{w}^\top \sigma(\boldsymbol{F}^\top[\boldsymbol{a} + \Delta \boldsymbol{z}_2]),$$

where $\sigma(\cdot)$, $\boldsymbol{\xi} \in \mathbb{R}^p$ and $\boldsymbol{F} \in \mathbb{R}^{p\times k}$ satisfy some regularity assumptions, and where $\boldsymbol{w} \in \mathbb{R}^k$. Moreover, define the joint probability distribution of ν_1, ν_2 and ν_3 as $\mathbb{P}(\nu_1, \nu_2, \nu_3)$. The cGEC states that the joint distribution $\mathbb{P}(\nu_1, \nu_2, \nu_3)$ is asymptotically Gaussian, i.e. $d(\mathbb{P}(\nu_1, \nu_2, \nu_3), \mathbb{P}(\nu_{g,1}, \nu_{g,2}, \nu_{g,3}))$ converges in probability to zero where $\nu_{g,1}$, $\nu_{g,2}$ and $\nu_{g,3}$ are jointly Gaussian with the same first and second moments of ν_1 , ν_2 and ν_3 and $d(\cdot, \cdot)$ is some probability distance that metrizes the convergence in distribution (e.g maximum-sliced (MS) distance (Kolouri et al., 2019; Goldt et al., 2020a)). To have the same first two moments, the random variables $\nu_{g,1}$, $\nu_{g,2}$ and $\nu_{g,3}$ are selected as follows $\nu_{g,1} = \nu_1$ and

$$\nu_{g,2} = \boldsymbol{w}^\top \big(\mu_0 \mathbf{1}_k + \boldsymbol{F}^\top[\widetilde{\mu}_1 \boldsymbol{a} + \widehat{\mu}_1 \boldsymbol{z}_1] + \mu_2 \boldsymbol{b} + \mu_3 \boldsymbol{p}_1\big),$$
$$\nu_{g,3} = \boldsymbol{w}^\top \big(\mu_0 \mathbf{1}_k + \boldsymbol{F}^\top[\widetilde{\mu}_1 \boldsymbol{a} + \widehat{\mu}_1 \boldsymbol{z}_2] + \mu_2 \boldsymbol{b} + \mu_3 \boldsymbol{p}_2\big),$$

where $\mathbf{1}_k$ represents the all 1 vector with size k. Here, $\boldsymbol{b} \in \mathbb{R}^k$, $\boldsymbol{p}_1 \in \mathbb{R}^k$ and $\boldsymbol{p}_2 \in \mathbb{R}^k$ are three independent standard Gaussian random vectors and they are independent of $\boldsymbol{a}$, $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$. The weights μ_0, $\widetilde{\mu}_1$, $\widehat{\mu}_1$, μ_2 and μ_3 are as defined in Section 1.3.

In the standard setting, i.e. $\Delta = 0$, the cGEC is equivalent to the uniform equivalence theorem (uGET), observed and used in many earlier papers (Montanari et al., 2019; Gerace et al., 2020; Goldt et al., 2020b; Dhifallah & Lu, 2020). Recently, the work in (Hu & Lu, 2020) provided a rigorous proof of the uGET. Specifically, the work in (Hu & Lu, 2020) proves a special case of cGEC when $\Delta = 0$, the feature matrix is Gaussian and the activation functions have bounded first three derivatives. However, similar to previous literature (Goldt et al., 2020b), we conjecture that the cGEC is valid under more general settings. We believe that the analysis in (Hu & Lu, 2020) can be extended to prove the cGEC and we leave the technical details for future work.

Our theoretical results are based on this conjecture. It is thus useful to also provide an intuitive explanation for the plausibility of the cGEC. Assume that $\boldsymbol{f}_i$ is the ith column of $\boldsymbol{F}$. The nonlinear term $I_1 = \sigma(\boldsymbol{f}_i^\top(\boldsymbol{a} + \Delta \boldsymbol{z}_1))$ can be decomposed by projecting on the basis $(1, \boldsymbol{f}_i^\top \boldsymbol{a}, \boldsymbol{f}_i^\top \boldsymbol{z}_1)$, i.e. $I_1 = \mu_0 + \widetilde{\mu}_1 \boldsymbol{f}_i^\top \boldsymbol{a} + \widehat{\mu}_1 \boldsymbol{f}_i^\top \boldsymbol{z}_1 + \sigma_i^\perp$. The term $\sigma_i^\perp$ is selected so that we match the variance of I_1 and the correlation with $I_2 = \sigma(\boldsymbol{f}_i^\top(\boldsymbol{a} + \Delta \boldsymbol{z}_2))$. We note that the cGEC makes sense when the columns of $\boldsymbol{F}$ are independent and have the same norm. These are the regularity assumptions for the feature matrix in (Goldt et al., 2020b). The same intuition also appears in the analysis of the unperturbed random kernel models, in particular, the random feature model (Montanari et al., 2019). In this paper, we suppose that the feature matrix and the activation function satisfy the regularity assumptions in (Goldt et al., 2020b) and conjecture that the Gaussian equivalence is valid for $(\nu_1, \nu_2, \ldots, \nu_\ell)$ for $\ell \geq 1$ and uniformly in $\boldsymbol{w} \in \mathbb{R}^k$. Using the cGEC, the

performance of the formulation in (3) can be characterized by asymptotically analyzing the Gaussian formulation given in (6). We verify this conjecture by performing multiple simulation examples in different settings.

3. Technical Assumptions

In this paper, we precisely characterize the noisy formulation under the following technical assumptions.

Assumption 1 (Gaussian Vectors)**.** *The input vectors* $\{\boldsymbol{a}_i\}_{i=1}^n$ *and the perturbation vectors* $\{\boldsymbol{z}_{ij}\}_{i=1,j=1}^{n,\ell}$ *are known and drawn independently from a standard Gaussian distribution. Without loss of generality, we assume that the hidden vector* $\boldsymbol{\xi} \in \mathbb{R}^p$ *has unit norm. Also, it is independent of the input vectors, the noise vectors and* $\boldsymbol{F}$.

This paper makes specific assumptions about the input/noise vectors distribution. We wish to emphasize that such assumptions are essential for our asymptotic analysis. An interesting future work is to relax the Gaussian assumption by establishing universality properties (e.g. (Oymak & Tropp, 2017; Panahi & Hassibi, 2017)). Our theoretical predictions are valid in the high-dimensional setting where n, p and k grow to infinity at finite ratios.

Assumption 2 (Asymptotic Limit)**.** *The number of samples and the number of hidden neurons satisfy* $n = n(p)$ *and* $k = k(p)$*, respectively. We assume that* $\alpha_p = n(p)/p \to \alpha > 0$ *and* $\eta_p = k(p)/n(p) \to \eta > 0$ *as* $p \to \infty$*. Also, the number of noise injections* ℓ *is independent of* p.

Moreover, we consider the following assumption to ensure that the generalization error defined in (5) concentrates.

Assumption 3 (Generative Model)**.** *The data generating function* $\varphi(\cdot)$ *introduced in (1) is independent of the input vectors, the noise vectors and the feature matrix. Moreover, the following conditions are satisfied.*

(a) $\varphi(\cdot)$ *and* $\widehat{\varphi}(\cdot)$ *are continuous almost everywhere in* $\mathbb{R}$*. For every* $h > 0$ *and* $z \sim \mathcal{N}(0, h)$*, we have* $\mathbb{E}[\varphi^2(z)] < +\infty$*,* $\mathbb{E}[z\varphi(z)] \neq 0$ *and* $0 < \mathbb{E}[\widehat{\varphi}^2(z)] < +\infty$.

(b) *For any* $[c, C]$*, there exists a function* $g(\cdot)$ *such that*

$$\sup_{h,\chi \in [c,C]} |\widehat{\varphi}(\chi + hx)|^2 \leq g(x) \quad \textit{for all } x \in \mathbb{R}.$$

Additionally, the function $g(\cdot)$ *satisfies* $\mathbb{E}[g^2(z)] < +\infty$*, where* $z \sim \mathcal{N}(0, 1)$.

In addition to the assumptions in Section 2, we consider the following regularity conditions for the activation function.

Assumption 4 (Activation Function)**.** *The activation function* $\sigma(\cdot)$ *is independent of the input vectors, the noise vectors and the feature matrix. It also satisfies* $\mathbb{E}[\sigma(z)^2] < +\infty$ *and* $\mathbb{E}[z\sigma(z)] \neq 0$*, where* $z \sim \mathcal{N}(0, 1)$.

In addition to the assumptions discussed in Section 2, we consider a family of feature matrices that satisfy the following assumption to guarantee that the Gaussian formulation converges to a deterministic problem.

Assumption 5 (Feature Matrix)**.** *The SVD decomposition of the feature matrix can be expressed as* $\boldsymbol{F} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}$*, where* $\boldsymbol{U} \in \mathbb{R}^{p\times p}$ *and* $\boldsymbol{V} \in \mathbb{R}^{k\times k}$ *are random orthogonal matrices and* $\boldsymbol{S} \in \mathbb{R}^{p\times k}$ *is a diagonal matrix formed by the singular values of* $\boldsymbol{F}$*. Define the matrix* $\boldsymbol{M}$ *as* $\boldsymbol{M} = \boldsymbol{F}^\top \boldsymbol{F}$.

(a) *We assume that* $\boldsymbol{U}$ *is a Haar-distributed random unitary matrix.*

(b) *We also assume that the empirical distribution of the eigenvalues of the matrix* $\boldsymbol{M}$ *converges weakly to a probability distribution* $\mathbb{P}_\kappa(\cdot)$ *supported in* $[0\ \zeta_{max}]$*, where* $\zeta_{max} > 0$ *is a constant independent of* (p, ℓ).

(c) *We finally assume that* $\mathbb{E}_\kappa[\kappa] > 0$*, where the expectation is taken over the distribution* $\mathbb{P}_\kappa(\cdot)$.

Based on Assumption 2, we also have the following property $\delta_p = k(p)/p \to \delta > 0$ as p grows to infinity. Moreover, the assumption on the feature matrix is used to show that some key quantities in the cost function concentrate in the high dimensional limit. The above assumptions are essential for the technical tools we use. The simulation results in Section 5.4 show the robustness of the phenomenology uncovered by our analysis on real data sets.

4. Precise Analysis of the Noisy Formulation

In this section, we asymptotically analyze the noise injection procedure introduced in (3). Specifically, we provide a precise asymptotic characterization of the training and generalization errors corresponding to (3).

4.1. Precise Asymptotic Analysis

Before stating our technical results, we start with few definitions. Define the following two deterministic functions

$$T_{2,\lambda}(\boldsymbol{t}, \boldsymbol{\tau}) = \frac{\delta}{T_1^2}\mathbb{E}_\kappa\Big[\frac{\kappa}{g_{\kappa,\lambda}(\boldsymbol{t}, \boldsymbol{\tau})}\Big] \Big/ \Big(1 - \frac{\widetilde{\mu}_1^2 t_1 \delta}{\tau_1}\mathbb{E}_\kappa\Big[\frac{\kappa}{g_{\kappa,\lambda}(\boldsymbol{t}, \boldsymbol{\tau})}\Big]\Big)$$

$$T_{3,\lambda}(\boldsymbol{t}, \boldsymbol{\tau}) = \frac{t_1^2}{\ell}\mathbb{E}_\kappa\Big[\frac{\widetilde{\mu}_1^2\kappa + \mu_2^2}{g_{\kappa,\lambda}(\boldsymbol{t}, \boldsymbol{\tau})}\Big] + \frac{t_1^2 + t_2^2}{\ell^2}\mathbb{E}_\kappa\Big[\frac{\widehat{\mu}_1^2\kappa + \mu_3^2}{g_{\kappa,\lambda}(\boldsymbol{t}, \boldsymbol{\tau})}\Big],$$

where the expectations are taken over the probability distribution $\mathbb{P}_\kappa(\cdot)$ defined in Assumption 5 and where $\boldsymbol{t} = [t_1, t_2]^\top$ and $\boldsymbol{\tau} = [\tau_1, \tau_2]^\top$. Here, the function $g_{\kappa,\lambda}(\cdot, \cdot)$ is defined as follows

$$g_{\kappa,\lambda}(\boldsymbol{t}, \boldsymbol{\tau}) = \frac{t_1}{\tau_1}\big(\widetilde{\mu}_1^2\kappa + \mu_2^2\big) + \Big(\frac{t_1}{\tau_1\ell} + \frac{t_2(\ell-1)}{\tau_2\ell}\Big) \times \big(\widehat{\mu}_1^2\kappa + \mu_3^2\big) + \lambda. \tag{10}$$

Furthermore, define the following four-dimensional deterministic optimization problem

$$\inf_{\substack{\tau_1>0\ t_1\geq 0\\ \tau_2>0\ t_2\geq 0}} \max \frac{t_1}{2\tau_1}\big(\gamma_1 - 2\widetilde{\mu}_1 T_1 q^\star_{t,\tau}\gamma_2 + \widetilde{\mu}_1^2 T_1^2 (q^\star_{t,\tau})^2 + \mu_0^2(\vartheta^\star)^2 - 2\mu_0\vartheta^\star\gamma_3\big) + \frac{\tau_1 t_1 + \tau_2 t_2}{2\ell} - \frac{t_1^2+t_2^2}{2\ell} + \frac{(q^\star_{t,\tau})^2}{2T_{2,\lambda}(\boldsymbol{t},\boldsymbol{\tau})} - \frac{\eta T_{3,\lambda}(\boldsymbol{t},\boldsymbol{\tau})}{2}, \tag{11}$$

where the constant $\vartheta^\star$ satisfies $\vartheta^\star = 0$ if $\mu_0 = 0$ and $\vartheta^\star = \gamma_3/\mu_0$ otherwise, and $T_1 = \sqrt{\delta \mathbb{E}_\kappa[\kappa]}$. Here, γ_1, γ_2 and γ_3 depend on the data distribution and are defined as $\gamma_1 = \mathbb{E}[y^2]$, $\gamma_2 = \mathbb{E}[ys]$, $\gamma_3 = \mathbb{E}[y]$, where $y = \varphi(s)$, and s is a standard Gaussian random variable. Note that the problem defined in (11) depends on $q^\star_{t,\tau}$ which is given by

$$q^\star_{t,\tau} = \frac{\gamma_2 t_1 \widetilde{\mu}_1 T_1 T_{2,\lambda}(\boldsymbol{t},\boldsymbol{\tau})}{\tau_1 + t_1 \widetilde{\mu}_1^2 T_1^2 T_{2,\lambda}(\boldsymbol{t},\boldsymbol{\tau})}. \tag{12}$$

Now, we summarize our main theoretical results in the following theorem.

Theorem 1 (Noisy Formulation Characterization)**.** *Suppose that the assumptions in Section 3 are all satisfied and the cGEC introduced in Section 2 is valid. Then, the training error converges in probability as follows*

$$\mathcal{E}_{train} \xrightarrow{p\to+\infty} C^\star(\Delta,\lambda) - \frac{\lambda}{2}\left((q^\star)^2 + h'(\lambda)\right),$$

where $C^\star(\Delta,\lambda)$ is the optimal cost of the deterministic problem in (11). Here, the function $h(\cdot)$ is defined as follows

$$h(\lambda) = -(q^\star)^2\Big(\lambda - \frac{1}{T_{2,\lambda}(\boldsymbol{t}^\star,\boldsymbol{\tau}^\star)}\Big) - \eta T_{3,\lambda}(\boldsymbol{t}^\star,\boldsymbol{\tau}^\star).$$

Moreover, the generalization error defined in (5) converges in probability to a deterministic function as follows

$$\mathcal{E}_{test} \xrightarrow{p\to+\infty} \frac{1}{4^v}\mathbb{E}\left[\left(\varphi(g_1) - \widehat{\varphi}(g_2)\right)^2\right], \tag{13}$$

where g_1 and g_2 have a bivariate Gaussian distribution with mean vector $[0, \mu_{0s}\vartheta^\star]$ and covariance matrix $\boldsymbol{C}$, defined as follows

$$\boldsymbol{C} = \begin{bmatrix} 1 & \mu_{1s}T_1 q^\star \\ \mu_{1s}T_1 q^\star & \mu_{1s}^2\beta^\star + \mu_{2s}^2\left((q^\star)^2 + h'(\lambda)\right)\end{bmatrix}.$$

The constant $\vartheta^\star$ satisfies $\vartheta^\star = 0$ if $\mu_0 = 0$ and $\vartheta^\star = \gamma_3/\mu_0$ otherwise. Here, the constants μ_{0s}, μ_{1s} and μ_{2s} are defined as $\mu_{0s} = \mathbb{E}[\sigma(z)]$, $\mu_{1s} = \mathbb{E}[z\sigma(z)]$ and $\mu_{2s}^2 = \mathbb{E}[\sigma(z)^2] - \mu_{0s}^2 - \mu_{1s}^2$, where z is a standard Gaussian random variable. Additionally, the constant $\beta^\star$ can be computed via the following expression

$$\beta^\star = \frac{1}{V_1+V_3}\Big(V_1 T_1^2 - V_2 - V_4 - \lambda + \frac{1}{T_{2,\lambda}(\boldsymbol{t}^\star,\boldsymbol{\tau}^\star)}\Big)(q^\star)^2 + \frac{\eta T_{3,\lambda}(\boldsymbol{t}^\star,\boldsymbol{\tau}^\star)}{V_1+V_3} - \frac{V_2+V_4+\lambda}{V_1+V_3}h'(\lambda), \tag{14}$$

where the constants V_1, V_2, V_3 and V_4 are defined as follows

$$V_1 = \frac{t_1^\star \widetilde{\mu}_1^2}{\tau_1^\star},\ V_3 = \widehat{\mu}_1^2\Big(\frac{t_1^\star}{\tau_1^\star \ell} + \frac{t_2^\star(\ell-1)}{\tau_2^\star \ell}\Big)$$
$$V_2 = \frac{t_1^\star \mu_2^2}{\tau_1^\star}, V_4 = \mu_3^2\Big(\frac{t_1^\star}{\tau_1^\star \ell} + \frac{t_2^\star(\ell-1)}{\tau_2^\star \ell}\Big).$$

Here, $q^\star = q^\star_{t^\star,\tau^\star}$ is given in (12), $\boldsymbol{t}^\star = [t_1^\star, t_2^\star]^\top$ and $\boldsymbol{\tau}^\star = [\tau_1^\star, \tau_2^\star]^\top$. Moreover, $\{t_1^\star, t_2^\star, \tau_1^\star, \tau_2^\star\}$ denotes the optimal solution of the problem defined in (11). Also, we treat $q^\star$, $\boldsymbol{t}^\star$ and $\boldsymbol{\tau}^\star$ as constants independent of λ when we compute the derivative of the function $h(\cdot)$.

To streamline our presentation, we postpone the proof of Theorem 1 to Section 8. Note that Theorem 1 provides a full asymptotic characterization of the training and generalization errors corresponding to the formulation given in (3). Specifically, it shows that the performance of (3) can be fully characterized after solving a deterministic scalar formulation where the cost function depends on the parameters ℓ and Δ. The theoretical predictions stated in Theorem 1 are valid for any fixed noise variance $\Delta \geq 0$ and number of noise samples $\ell \geq 1$. Additionally, it is valid for a general family of feature matrices, activation functions and generative models satisfying (1). The analysis presented in Section 8 shows that the deterministic problem in (11) is strictly convex-concave. This implies the uniqueness of the optimal solutions of the optimization in (11). Next, we study the properties of the noise injection method in (3) when ℓ grows to infinity slower than n, p and k.

4.2. Noise Regularization Effects

Now, we consider the setting where ℓ grows to infinity slower than the dimensions n, p and k. We use the theoretical predictions stated in Theorem 1 to study the regularization effects of the noise injection method in (3). Our first theoretical result is introduced in the following theorem.

Theorem 2 (Regularization Effects)**.** *Suppose that the assumptions in Theorem 1 are all satisfied. Moreover, define the following formulation*

$$\min_{\boldsymbol{w}\in\mathbb{R}^k} \frac{1}{2n}\sum_{i=1}^n \left(y_i - \boldsymbol{w}^\top \widehat{\sigma}(\boldsymbol{F}^\top \boldsymbol{a}_i)\right)^2 + \frac{1}{2}\|\boldsymbol{R}^{\frac{1}{2}}\boldsymbol{w}\|^2 + \tfrac{\lambda}{2}\|\boldsymbol{w}\|^2. \tag{15}$$

Here, the regularization matrix $\boldsymbol{R}$ is defined as follows

$$\boldsymbol{R} = \widehat{\mu}_1^2 \boldsymbol{F}^\top \boldsymbol{F} + \mu_3^2 \boldsymbol{I}_k, \tag{16}$$

and the new activation function $\widehat{\sigma}(\cdot)$ satisfies the properties

$$\mathbb{E}[\widehat{\sigma}(z)] = \mathbb{E}[\sigma(x_1)],\ \mathbb{E}[z\widehat{\sigma}(z)] = \mathbb{E}[z\sigma(x_1)]$$
$$\mathbb{E}[\widehat{\sigma}(z)^2] = \mathbb{E}[\sigma(x_1)\sigma(x_2)], \tag{17}$$

where $x_1 = z + \Delta v_1$, $x_2 = z + \Delta v_2$ and z, v_1 and v_2 are independent standard Gaussian random variables. Also, define $\widehat{\mathcal{E}}_{train}$ and $\widehat{\mathcal{E}}_{test}$ as the training and generalization errors corresponding to the problem in (15). Then, for any $\zeta > 0$, we have the following convergence results

$$\begin{cases} \lim_{\ell \to +\infty} \lim_{p \to +\infty} \mathbb{P}\Big(|\mathcal{E}_{train} - \widehat{\mathcal{E}}_{train}| < \zeta \Big) = 1 \\ \lim_{\ell \to +\infty} \lim_{p \to +\infty} \mathbb{P}\Big(|\mathcal{E}_{test} - \widehat{\mathcal{E}}_{test}| < \zeta \Big) = 1, \end{cases} \tag{18}$$

where $\mathcal{E}_{test}$ and $\mathcal{E}_{train}$ are the training and generalization errors corresponding to the noisy formulation.

To streamline our presentation, we postpone the proof of Theorem 2 to Section 8. The above theorem shows that the noisy formulation given in (3) is equivalent to a standard formulation with a new activation function and an additional weighted ridge regularization, when $\ell \to +\infty$. It also provides the explicit form of the regularization. This shows that inserting Gaussian noise during the training procedure introduces a regularization that depend on the activation function, the feature matrix and the noise variance. Now, we provide a precise asymptotic characterization of the formulation in (15). Before stating our asymptotic result, we define the following deterministic problem

$$\inf_{\tau_1 > 0} \sup_{t_1 \geq 0} \frac{t_1}{2\tau_1}\big(\gamma_1 - 2\widetilde{\mu}_1 T_1 \widehat{q}^{\star}_{t,\tau}\gamma_2 + \widetilde{\mu}_1^2 T_1^2 (\widehat{q}^{\star}_{t,\tau})^2 + \mu_0^2(\vartheta^{\star})^2 - 2\mu_0\vartheta^{\star}\gamma_3\big) + \frac{\tau_1 t_1}{2} - \frac{t_1^2}{2} + \frac{(\widehat{q}^{\star}_{t,\tau})^2}{2\widehat{T}_{2,\lambda}(t_1,\tau_1)} - \frac{\eta \widehat{T}_{3,\lambda}(t_1,\tau_1)}{2}.$$

Here, the constant $\vartheta^{\star}$ satisfies $\vartheta^{\star} = 0$ if $\mu_0 = 0$ and $\vartheta^{\star} = \gamma_3/\mu_0$ otherwise, and T_1 is defined in Section 4.1. Moreover, the functions $\widehat{q}^{\star}_{t,\tau}$ and $\widehat{T}_{2,\lambda}(\cdot,\cdot)$ are defined as follows

$$\widehat{q}^{\star}_{t,\tau} = \frac{\gamma_2 t_1 \widetilde{\mu}_1 T_1 \widehat{T}_{2,\lambda}(t_1,\tau_1)}{\tau_1 + t_1 \widetilde{\mu}_1^2 T_1^2 \widehat{T}_{2,\lambda}(t_1,\tau_1)}, \text{ and } \widehat{T}_{2,\lambda}(t_1,\tau_1) = \frac{\delta}{T_1^2}\mathbb{E}_{\kappa}\Big[\frac{\kappa}{\widehat{g}_{\kappa,\lambda}(t_1,\tau_1)}\Big] \Big/ \Big(1 - \frac{\widetilde{\mu}_1^2 t_1 \delta}{\tau_1}\mathbb{E}_{\kappa}\Big[\frac{\kappa}{\widehat{g}_{\kappa,\lambda}(t_1,\tau_1)}\Big]\Big).$$

Here, the functions $\widehat{T}_{3,\lambda}(\cdot,\cdot)$ and $\widehat{g}_{\kappa,\lambda}(\cdot,\cdot)$ are defined as follows

$$\widehat{T}_{3,\lambda}(t_1,\tau_1) = t_1^2 \mathbb{E}_{\kappa}\Big[(\widetilde{\mu}_1^2\kappa + \mu_2^2)/\widehat{g}_{\kappa,\lambda}(t_1,\tau_1)\Big],$$
$$\widehat{g}_{\kappa,\lambda}(t_1,\tau_1) = \frac{t_1}{\tau_1}\Big(\widetilde{\mu}_1^2\kappa + \mu_2^2\Big) + \Big(\widehat{\mu}_1^2\kappa + \mu_3^2\Big) + \lambda,$$

where the expectations are taken over the probability distribution $\mathbb{P}_{\kappa}(\cdot)$ defined in Assumption 5. Now, we summarize the asymptotic properties of the limiting formulation in (15) in the following theorem.

Lemma 1 (Limiting Formulation Characterization)**.** *Suppose that the assumptions in Theorem 1 are all satisfied. Then, the training error corresponding to the limiting formulation in (15) converges in probability as follows*

$$\widehat{\mathcal{E}}_{train} \xrightarrow{p \to +\infty} \widehat{C}^{\star}(\Delta,\lambda) - \frac{\lambda}{2}\Big((\widehat{q}^{\star})^2 + \widehat{h}'(\lambda)\Big),$$

where $\widehat{C}^{\star}(\Delta,\lambda)$ is the optimal cost of the deterministic problem in (19). Here, the function $\widehat{h}(\cdot)$ is defined as follows

$$\widehat{h}(\lambda) = -(\widehat{q}^{\star})^2\Big(\lambda - \frac{1}{\widehat{T}_{2,\lambda}(t_1^{\star},\tau_1^{\star})}\Big) - \eta \widehat{T}_{3,\lambda}(t_1^{\star},\tau_1^{\star}).$$

Moreover, the generalization error corresponding to the limiting formulation in (15) converges in probability to a deterministic function as follows

$$\widehat{\mathcal{E}}_{test} \xrightarrow{p \to +\infty} \frac{1}{4^v}\mathbb{E}\Big[\big(\varphi(g_1) - \widehat{\varphi}(g_2)\big)^2\Big], \tag{19}$$

where g_1 and g_2 have a bivariate Gaussian distribution with mean vector $[0, \mu_{0s}\vartheta^{\star}]$ and covariance matrix $\boldsymbol{C}$, defined as follows

$$\boldsymbol{C} = \begin{bmatrix} 1 & \mu_{1s} T_1 \widehat{q}^{\star} \\ \mu_{1s} T_1 \widehat{q}^{\star} & \mu_{1s}^2 \widehat{\beta}^{\star} + \mu_{2s}^2\Big((\widehat{q}^{\star})^2 + \widehat{h}'(\lambda)\Big) \end{bmatrix}.$$

The constant $\vartheta^{\star}$ satisfies $\vartheta^{\star} = 0$ if $\mu_0 = 0$ and $\vartheta^{\star} = \gamma_3/\mu_0$ otherwise. Here, the constants μ_{0s}, μ_{1s} and μ_{2s} are defined as $\mu_{0s} = \mathbb{E}[\sigma(z)]$, $\mu_{1s} = \mathbb{E}[z\sigma(z)]$ and $\mu_{2s}^2 = \mathbb{E}[\sigma(z)^2] - \mu_{0s}^2 - \mu_{1s}^2$, where z is a standard Gaussian random variable. Additionally, the constant $\widehat{\beta}^{\star}$ can be computed via the following expression

$$\widehat{\beta}^{\star} = \frac{1}{V_1 + V_3}\Big(V_1 T_1^2 - V_2 - V_4 - \lambda + \frac{1}{\widehat{T}_{2,\lambda}(t_1^{\star},\tau_1^{\star})}\Big)(\widehat{q}^{\star})^2 + \frac{\eta \widehat{T}_{3,\lambda}(t_1^{\star},\tau_1^{\star})}{V_1 + V_3} - \frac{V_2 + V_4 + \lambda}{V_1 + V_3}\widehat{h}'(\lambda), \tag{20}$$

where the constants V_1, V_2, V_3 and V_4 are defined as follows

$$V_1 = \frac{t_1^{\star}\widetilde{\mu}_1^2}{\tau_1^{\star}},\ V_3 = \widehat{\mu}_1^2,\ V_2 = \frac{t_1^{\star}\mu_2^2}{\tau_1^{\star}}, V_4 = \mu_3^2.$$

Here, $\widehat{q}^{\star} = \widehat{q}^{\star}_{t^{\star},\tau^{\star}}$ is given in (19). Moreover, $\{t_1^{\star}, \tau_1^{\star}\}$ denotes the optimal solution of the problem defined in (19). Also, we treat $\widehat{q}^{\star}$, $t_1^{\star}$ and $\tau_1^{\star}$ as constants independent of λ when we compute the derivative of the function $\widehat{h}(\cdot)$.

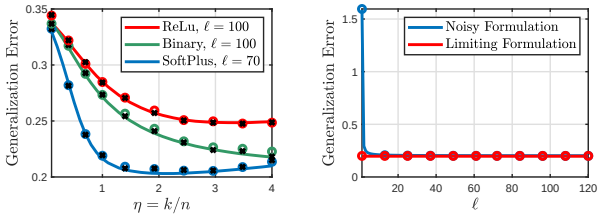
The proof of Lemma 1 is provided in Section 8. The results in Theorem 2 and Lemma 1 are based on the asymptotic predictions stated in Theorem 1. Specifically, we show in Section 8 that the asymptotic problem corresponding to the noisy formulation in (11) converges to the deterministic problem in (19), when ℓ grows to infinity. Then, we show that the deterministic problem in (19) is the asymptotic limit of the formulation in (15) using the CGMT framework. The analysis presented in Section 8 shows that the deterministic problem in (19) is strictly convex-concave. This implies the uniqueness of its optimal solutions. Due to space limitation, we provide the detailed proof of our theoretical results in the supplementary material.

5. Simulation Results

In this part, we provide additional simulation examples to verify our asymptotic results stated in Theorem 1, Theorem 2 and Lemma 1. Our predictions stated in Section 4 are valid for a general family of feature matrices, activation functions and generative models satisfying (1). In this part, we specialize our general results to popular learning models. In particular, we consider two families of feature matrices. We consider feature matrices that can be expressed as $\boldsymbol{F} = d\boldsymbol{V}$, where: **(a)** The scalar d satisfies $d = 1/\sqrt{p}$ and the matrix $\boldsymbol{V}$ has independent standard Gaussian components. We refer to this matrix as the *Gaussian feature matrix*. **(b)** The scalar d satisfies $d = \sqrt{3/p}$ and the matrix $\boldsymbol{V}$ has independent uniformly distributed components in $[-1\ 1]$. We refer to this matrix as the *uniform feature matrix*.
Also, we consider two popular regression and classification models. For the regression model, we assume that $\varphi(\cdot)$ is the ReLu function and $\widehat{\varphi}(\cdot)$ is the identity function. For the classification model, we assume that $\varphi(\cdot)$ is the sign function with possible sign flip with probability θ and $\widehat{\varphi}(\cdot)$ is the sign function.

5.1. Limiting Performance

Our third simulation considers the non–linear regression model. Figure 3 compares the numerical predictions and our predictions stated in Theorem 2 and Lemma 1. This



(a) (b)

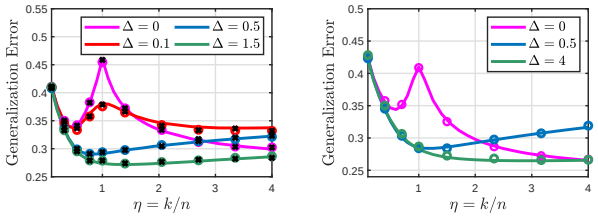
Figure 3. Solid line: Theoretical predictions. Circle: Numerical simulations for (3) in 3(a) and for both (3) and (8) in 3(b). Black cross: Numerical simulations for (8). **(a)** $p = 500$, $\Delta = 0.4$, $\alpha = 1.5$ and $\lambda = 10^{-2}$. **(b)** $p = 500$, $\Delta = 0.6$, $\alpha = 2$, $\lambda = 10^{-4}$, $\eta = 1$ and $\sigma(\cdot)$ is the SoftPlus. Binary denotes the binary step activation. $\boldsymbol{F}$ is the Gaussian feature matrix. The number of Monte Carlo trials is 100.

simulation example first provides an empirical verification of the theoretical predictions in Theorem 2 and Lemma 1. It particularly shows that our predictions are in excellent agreement with the empirical results for (3) and (8). Furthermore, note that the performance of the deterministic formulation given in Lemma 1 is achieved with a moderate number of noise samples, i.e. $\ell = 70$ and $\ell = 100$. This verifies the results stated in Theorem 2 and Lemma 1 and provides an empirical verification of the cGEC introduced in Section 2.

Figure 3(a) further shows that the considered noisy formulation can asymptotically mitigate the double descent in the generalization error for an appropriately selected activation function and fixed noise variance. Specifically, note that the ReLu and binary activation functions lead to a decreasing generalization performance which is not the case for the SoftPlus activation. Figure 3(b) illustrates the convergence behavior of the generalization error corresponding to (3) for the SoftPlus activation and fixed η. It particularly shows that the generalization error of (3) converges to the generalization error of (8) when ℓ grows to infinity. Moreover, note that the limit is already achieved with a small value of ℓ. This verifies the predictions in Theorem 2.

5.2. Impact of the Noise Variance

In this simulation example, we study the effects of the noise variance Δ on the generalization error corresponding to the noisy formulation and the limiting formulation. Here, we consider the binary classification model. Figure 4 compares the numerical predictions and our theoretical predictions stated in Theorem 1, Theorem 2 and Lemma 1. It provides



(a) (b)

Figure 4. Solid line: Theoretical predictions. Circle: Numerical simulations for (3) in 4(a) and for (8) in 4(b). Black cross: Numerical simulations for (6). **(a)** $\boldsymbol{F}$ is the Gaussian feature matrix and $\sigma(\cdot)$ is the tanh activation function. We set $p = 400$, $\ell = 50$, $\alpha = 2$, $\theta = 0.1$ and $\lambda = 10^{-5}$. **(b)** $\boldsymbol{F}$ is the uniform feature matrix and $\sigma(\cdot)$ is the SoftPlus activation. We set $p = 1500$, $\alpha = 1.5$, $\theta = 0$ and $\lambda = 10^{-4}$. The number of Monte Carlo trials is 100.

another empirical verification of our theoretical predictions since our results are in excellent agreement with the actual performance of the considered formulations. It also provides an empirical verification of the cGEC discussed in Section 2. Figure 4(a) studies the effects of the noise variance Δ on the generalization error corresponding to the noisy formulation for fixed ℓ. Note that increasing the noise variance improves the generalization error especially at low η. Figure 4(a) also suggests that an optimized noise variance can reduce the effects of the double descent phenomenon. Now, Figure 4(b) studies the effects of the noise variance Δ on the generalization error corresponding to the limiting formulation. We can see that the generalization error increases after reaching a minimum for $\Delta = 0.5$. For $\Delta = 4$, observe that the generalization error is deceasing. This suggests that

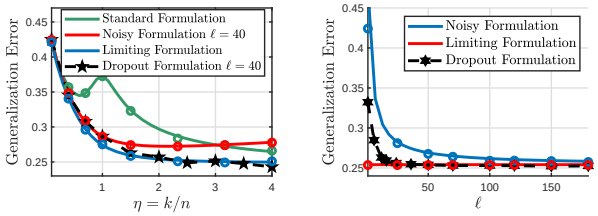
the double descent phenomenon can be mitigated for an appropriately selected noise variance.

5.3. Alternative Formulations

Now, we consider the binary classification model, where $\theta = 0$. We compare the performance of the noisy formulation given in (3) and the dropout technique. In this paper, we consider the following version of the dropout method

$$\min_{\boldsymbol{w}\in\mathbb{R}^k} \frac{1}{2n\ell} \sum_{i=1}^{n}\sum_{j=1}^{\ell} \left(y_i - \boldsymbol{w}^\top \sigma\left(\boldsymbol{D}_{ij}\boldsymbol{F}^\top \boldsymbol{a}_i\right)\right)^2 + \tfrac{\lambda}{2}\|\boldsymbol{w}\|^2,$$

where $\{\boldsymbol{D}_{ij}\}_{i,j}^{n,\ell}$ are diagonal matrices with independent and identically distributed diagonal entries drawn from the distribution, $\mathbb{P}(d=1) = 1-\epsilon$ and $\mathbb{P}(d=0) = \epsilon$, where ϵ denotes the probability of dropping a unit. The above formulation is similar to the one considered in (Srivastava et al., 2014; Wei et al., 2020). In Figure 5, we compare the generalization performance of the noisy and dropout formulations. First, we can notice that our asymptotic results provided in Theorem 1, Theorem 2 and Lemma 1 match with the actual performance of (3) and (15). This gives an empirical verification of our results. Figure 5(a) considers the erf activation function. It first shows that the dropout and noisy formulations have comparable performance at low η. However, we can see that the dropout method provides a largely better performance as compared to the noisy formulation for large values of η. Figure 5(a) also shows that the limiting and dropout formulations have a similar generalization performance. Figure 5(b) studies the convergence properties of both approaches as a function of ℓ. It particularly suggests that the dropout method has a better convergence rate as compared to the noisy formulation. Now, Figures 5(a) and 5(b) suggest that the noisy and dropout formulations have comparable generalization performance when ℓ grows to infinity. We provide more simulation examples in the supplementary material.

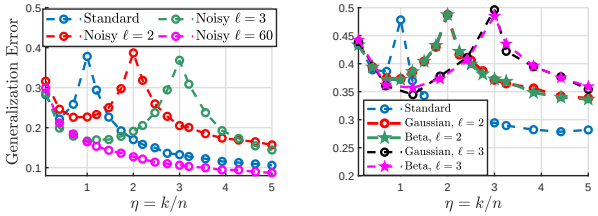


(a) (b)

Figure 5. Solid line: Theoretical predictions. Circle: Numerical simulations. Hexagram: Numerical simulations for the dropout formulation. Erf activation function and we set $p = 600$, $\alpha = 1.4$, $\Delta = 2$, $\lambda = 10^{-3}$ and $\epsilon = 0.3$. **(a)** $\ell = 40$. **(b)** $\eta = 2$. $\boldsymbol{F}$ is the uniform feature matrix. The number of Monte Carlo trials is 35.

5.4. Generalizing the Theoretical Predictions

In Figure 6(a), we consider the *Semeion Handwritten Digit* Data Set downloaded from the "Machine Learning Repository". Figure 6(a) shows that the generalization errors on real data sets exhibit the same *qualitative* behavior as for the Gaussian input vectors. This suggests that the i.i.d. Gaussian assumption can be removed/eased in practice, perhaps by considering a different random ensemble model with a covariance matching the input data set. Note that our results cannot be directly applied to noise distributions other than Gaussian. In principle, we believe that non-Gaussian noise can be treated by appealing to universality arguments (one observed in Figure 6(b) for centered beta/Gaussian noise). Our analysis is only valid for the squared loss, as some of the techniques used to obtain the asymptotic formulation are tailored to the squared loss. We leave the extension to general loss functions as an important future work. We can see from Figure 6(b) that the generalization error shows the same behaviors for beta distributed noise and the LAD loss.



(a) (b)

Figure 6. Numerical simulations. **(a)** ReLu activation and we set $p = 256$, $\alpha = 2$, $\Delta = 0.5$, $\lambda = 10^{-5}$ and $\epsilon = 0.3$. $\ell = 40$. **(b)** The least absolute deviation (LAD) loss , the tanh activation, $p = 150$, $\alpha = 1.2$, $\Delta = 0.8$, $\lambda = 10^{-6}$. $\boldsymbol{F}$ is the Gaussian feature matrix. The number of Monte Carlo trials is 100.

6. Conclusion

In this paper, we precisely analyzed a random perturbation method used to regularize machine learning problems. Specifically, we provided an accurate characterization of the training and generalization errors corresponding to the noisy feature formulation. Our predictions are based on a correlated Gaussian equivalence conjecture and an extended version of the CGMT, referred to as the multivariate CGMT. Moreover, our analysis shows that Gaussian noise injection in the input data has the same effects of a weighted ridge regularization when the number of noise samples grows to infinity. Additionally, it provides the explicit dependence of the introduced regularization on the feature matrix, the activation function and the noise variance. Simulation results validate our predictions and show that inserting noise during training moves the interpolation threshold and can mitigate the double descent phenomenon in the generalization error.

References

An, G. The effects of adding noise during backpropagation training on a generalization performance. *Neural Computation*, 8(3):643–674, 1996.

Belkin, M., Ma, S., and Mandal, S. To understand deep learning we need to understand kernel learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 541–549, 10–15 Jul 2018.

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

Bishop, C. M. Training with noise is equivalent to tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995.

Debbah, M., Hachem, W., Loubaton, P., and de Courville, M. MMSE analysis of certain large isometric random precoded systems. *IEEE Transactions on Information Theory*, 49(5):1293–1311, 2003.

Dhifallah, O. and Lu, Y. M. A Precise Performance Analysis of Learning with Random Features. *arXiv:2008.11904*, 2020.

Dhifallah, O. and Lu, Y. M. Phase Transitions in Transfer Learning for High-Dimensional Perceptrons. *arXiv:2101.01918*, 2021.

Dhifallah, O., Thrampoulidis, C., and Lu, Y. M. Phase retrieval via polytope optimization: Geometry, phase transitions, and new algorithms. *CoRR*, abs/1805.09555, 2018.

Gerace, F., Loureiro, B., Krzakala, F., Mzard, M., and Zdeborov, L. Generalisation error in learning with random features and the hidden manifold model. *arXiv:2002.09339*, 2020.

Goldt, S., Loureiro, B., Reeves, G., Krzakala, F., Mzard, M., and Zdeborov, L. The Gaussian equivalence of generative models for learning with shallow neural networks. *arXiv:2006.14709*, 2020a.

Goldt, S., Mzard, M., Krzakala, F., and Zdeborov, L. Modelling the influence of data structure on learning in neural networks: the hidden manifold model. *arXiv:1909.11500*, 2020b.

Gong, C., Ren, T., Ye, M., and Liu, Q. MaxUp: A Simple Way to Improve Generalization of Neural Network Training. *arXiv:2002.09024*, 2020.

Gordon, Y. On milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$. In Lindenstrauss, J. and Milman, V. D. (eds.), *Geometric Aspects of Functional Analysis*, pp. 84–106, Berlin, Heidelberg, 1988. Springer Berlin Heidelberg. ISBN 978-3-540-39235-4.

Gulcehre, C., Moczulski, M., Denil, M., and Bengio, Y. Noisy activation functions. volume 48 of *Proceedings of Machine Learning Research*, pp. 3059–3068, New York, New York, USA, 20–22 Jun 2016. PMLR.

Hu, H. and Lu, Y. M. Universality Laws for High-Dimensional Learning with Random Features. *arXiv:2009.07669*, 2020.

Kannan, H., Kurakin, A., and Goodfellow, I. Adversarial Logit Pairing. *arXiv:1803.06373*, 2018.

Kobak, D., Lomond, J., and Sanchez, B. Optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *arXiv:1805.10939*, 2020.

Kolouri, S., Nadjahi, K., Simsekli, U., Badeau, R., and Rohde, G. K. Generalized Sliced Wasserstein Distances. *arXiv:1902.00434*, 2019.

LeJeune, D., Javadi, H., and Baraniuk, R. The implicit regularization of ordinary least squares ensembles. volume 108 of *Proceedings of Machine Learning Research*, pp. 3525–3535, Online, 26–28 Aug 2020. PMLR.

Mei, S. and Montanari, A. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv:1908.05355*, 2019.

Mezard, M., Parisi, G., and Virasoro, M. *Spin Glass Theory and Beyond: An Introduction to the Replica Method and Its Applications*, volume 9 of *World Scientific Lecture Notes in Physics*. World Scientific, November 1986.

Mignacco, F., Krzakala, F., Lu, Y. M., and Zdeborov, L. The role of regularization in classification of high-dimensional noisy Gaussian mixture. *arXiv:2002.11544*, 2020.

Montanari, A., Ruan, F., Sohn, Y., and Yan, J. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv:1911.01544*, 2019.

Newey, W. K. and McFadden, D. Chapter 36 large sample estimation and hypothesis testing. volume 4 of *Handbook of Econometrics*, pp. 2111 – 2245. Elsevier, 1994.

Oymak, S. and Tropp, J. A. Universality laws for randomized dimension reduction, with applications. *Information and Inference: A Journal of the IMA*, 7(3):337–446, 11 2017.

Panahi, A. and Hassibi, B. A universal analysis of large-scale regularized least squares solutions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Poole, B., Sohl-Dickstein, J., and Ganguli, S. Analyzing noise in autoencoders and deep networks. *arXiv:1406.1831*, 2014.

Rahimi, A. and Recht, B. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems 20*, pp. 1177–1184. 2008.

Rakin, A. S., He, Z., and Fan, D. Parametric Noise Injection: Trainable Randomness to Improve Deep Neural Network Robustness against Adversarial Attack. *arXiv:1811.09310*, 2018.

Rockafellar, R. T. and Wets, R. J.-B. *Variational Analysis*. Springer-Verlag Berlin Heidelberg, 1998. ISBN 978-3-540-62772-2.

Rudelson, M. and Vershynin, R. Non-asymptotic theory of random matrices: extreme singular values, 2010.

Salehi, F., Abbasi, E., and Hassibi, B. The impact of regularization on high-dimensional logistic regression. In *Advances in Neural Information Processing Systems 32*, pp. 12005–12015. Curran Associates, Inc., 2019.

Schilling, R. L. *Measures, Integrals and Martingales*. Cambridge University Press, 2005. doi: 10.1017/CBO9780511810886.

Sifaou, H., Kammoun, A., and Alouini, M. Phase transition in the hard-margin support vector machines. In *2019 IEEE 8th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pp. 415–419, 2019.

Sion, M. On general minimax theorems. *Pacific J. Math.*, 8 (1):171–176, 1958.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014.

Stojnic, M. A framework to characterize performance of LASSO algorithms. *arXiv:1303.7291*, 2013.

Thrampoulidis, C., Oymak, S., and Hassibi, B. Regularized linear regression: A precise analysis of the estimation error. In *Proceedings of The 28th Conference on Learning Theory*, volume 40, pp. 1683–1709, Paris, France, 03–06 Jul 2015. PMLR.

Thrampoulidis, C., Abbasi, E., and Hassibi, B. Precise error analysis of regularized M-estimators in high-dimensions. *CoRR*, abs/1601.06233, 2016.

Thrampoulidis, C., Xu, W., and Hassibi, B. Symbol error rate performance of box-relaxation decoders in massive MIMO. *IEEE Transactions on Signal Processing*, 66(13): 33773392, Jul 2018. ISSN 1941-0476.

Wei, C., Kakade, S., and Ma, T. The Implicit and Explicit Regularization Effects of Dropout. *arXiv:2002.12915*, 2020.

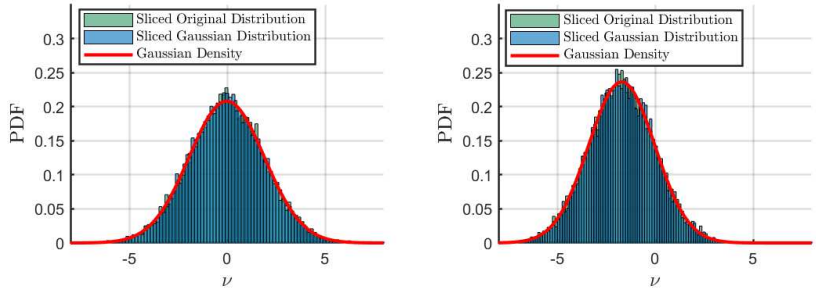
Zantedeschi, V., Nicolae, M.-I., and Rawat, A. Efficient Defenses Against Adversarial Attacks. *arXiv:1707.06728*, 2017.

7. Appendix I: Additional Simulation Examples

In this part, we provide additional simulation examples to support our conclusions in Section 5 and to illustrate the cGEC discussed in Section 2. Furthermore, we consider the family of feature matrices and the regression and binary classification models introduced in Section 5.

7.1. Correlated Gaussian Equivalence

In this simulation example, we consider a particular case of the cGEC to illustrate the Gaussian equivalence. Here, we consider a projected version of the probability density function $\mathbb{P}(\nu_1, \nu_2, \nu_3)$, defined in Section 2. Specifically, we consider the probability density function of the random variable $\nu = \nu_1 + \nu_2 + \nu_3$. Figure 7 considers two different feature matrices



(a) (b)

Figure 7. Sliced original distribution measures the distribution of $\nu = \nu_1 + \nu_2 + \nu_3$. *Sliced Gaussian distribution* measures the distribution of $\nu = \nu_{g,1} + \nu_{g,2} + \nu_{g,3}$. The *Gaussian density* denotes the Gaussian distribution with the corresponding first two moments of ν. $\boldsymbol{F}$ can be expressed as $\boldsymbol{F} = \boldsymbol{V}/\sqrt{p}$. **(a)** $\boldsymbol{V}$ has independent standard Gaussian components and $\sigma(\cdot)$ is the sign function. **(b)** $\boldsymbol{V}$ has independent uniformly distributed components in $[-\sqrt{3}\ \sqrt{3}]$ and $\sigma(\cdot)$ is the ReLu function. We take $p = 10^4$, $\alpha = 1.2$, $\Delta = 1$ and $\eta = 1$.

and activation functions. It also compares the probability density function of the random variable $\nu = \nu_1 + \nu_2 + \nu_3$ with the probability density function of the random variable $\nu_g = \nu_{g,1} + \nu_{g,2} + \nu_{g,3}$. Figure 7 shows that the random variable ν is Gaussian with the same first two moments of ν_g for particular choices of $\boldsymbol{F}$, $\sigma(\cdot)$, $\boldsymbol{\xi}$ and $\boldsymbol{w}$. This provides a particular illustration of the cGEC.

7.2. Impact of the Noise Variance

In this simulation example, we provide an additional simulation example to study the effects of the noise variance Δ on the generalization error corresponding to the noisy formulation. We consider the binary classification model introduced in Section 5. Figure 8 compares the numerical predictions and our theoretical predictions stated in Theorem 1. First, Figure

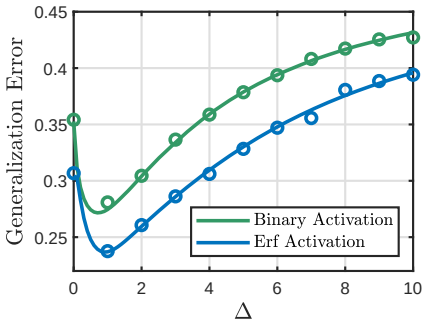


Figure 8. Solid line: Theoretical predictions. Circle: Numerical simulations for (3). $\boldsymbol{F}$ is the uniform feature matrix. We set $p = 600$, $\alpha = 2$, $\theta = 0$, $\ell = 20$, $\eta = 1.5$ and $\lambda = 10^{-4}$. The results are averaged over 100 independent Monte Carlo trials.

8 provides another empirical verification of our theoretical predictions. Specifically, observe that our predictions are in

excellent agreement with the actual performance of the formulation given in (3). This also provides an empirical verification of the cGEC discussed in Section 2. Figure 8 also shows that the generalization error corresponding to the noisy formulation has a unique minimum as a function of Δ for the considered activation functions. This suggests that the optimization problem over the noise variance Δ has a unique solution. This can simplify the design of an efficient optimization scheme of the generalization error in terms of Δ.

7.3. Alternative Formulations

In the this simulation example, we consider the binary classification model introduced in Section 5, where $\theta = 0$. In Figure 9, we compare the performance of the noisy formulation and the dropout formulation for three different activation functions. First, we can notice that our asymptotic results provided in Theorem 1, Theorem 2 and Lemma 1 match with the actual

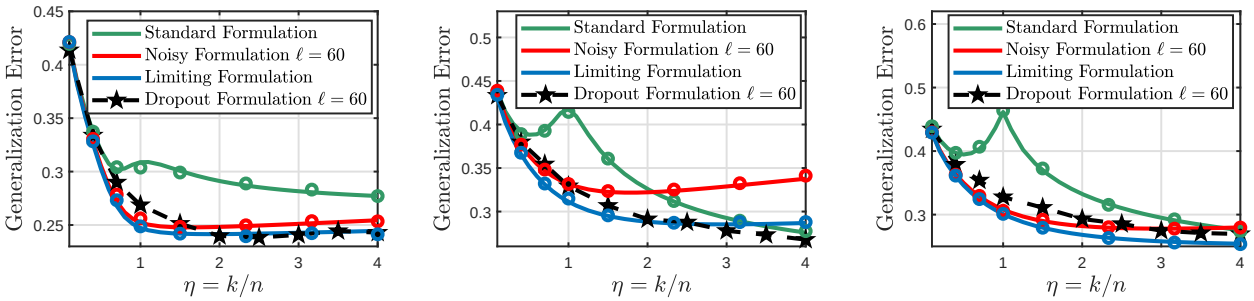


Figure 9. Solid line: Theoretical predictions. Circle: Numerical simulations for (3) and (8). Hexagram: Numerical simulations for the dropout formulation. **(a)** Sigmoid activation and we set $p = 600$, $\alpha = 1.4$, $\Delta = 1.5$, $\lambda = 10^{-4}$, $\ell = 60$ and $\epsilon = 0.3$. **(b)** ReLu activation function and we set the parameters as $p = 600$, $\alpha = 1.4$, $\Delta = 3.5$, $\lambda = 10^{-3}$, $\ell = 60$ and $\epsilon = 0.3$. **(c)** Sign activation function and we set the parameters as $p = 600$, $\alpha = 1.6$, $\Delta = 1.5$, $\lambda = 10^{-4}$, $\ell = 60$ and $\epsilon = 0.4$. The feature matrix $\boldsymbol{F}$ is the uniform feature matrix. The results are averaged over 35 independent Monte Carlo trials.

performance of the noisy formulation and its limiting formulation. This gives another empirical verification of our theoretical predictions. Figure 9 suggests that the noisy and dropout formulations significantly improve the generalization performance of the standard formulation for an appropriately selected activation function, Δ, ℓ and ϵ. Figure 9(a) shows that the noisy and dropout formulations have a similar generalization performance for fixed $\ell = 60$ and for the sigmoid activation function. Moreover, it shows that the dropout formulation approaches the generalization performance of the limiting formulation at high η. Moreover, 9(b) shows that the dropout method provides a largely better performance as compared to the noisy formulation for fixed $\ell = 60$ and for the ReLu activation. It also suggests that the limiting and dropout formulations have a similar generalization performance where the dropout method is better at high η. Now, Figure 9(c) considers the sign activation function and shows that the noisy formulation provides a better generalization performance as compared to the dropout formulation at low η and for a fixed number of noise injections, $\ell = 60$. We can also see that the limiting formulation generalizes better than the dropout method for the considered parameters. This simulation example particularly suggests that the performance of the noisy and dropout formulations depends on the activation function. Moreover, the dropout formulation have a similar generalization performance as compared to the limiting performance for the considered parameters.

7.4. Convergence Behavior

In the last simulation example, we study the convergence behavior of the noisy and dropout formulations for different activation functions. Figure 10 first shows that our theoretical predictions stated in Theorem 1, Theorem 2 and Lemma 1 match with the actual performance of the noisy formulation and its limiting formulation. This gives another empirical verification of our predictions. Figure 10 studies the convergence properties of both approaches as a function of ℓ and for different activation functions. It particularly suggests that the dropout method has a better convergence rate as compared to the noisy formulation. Moreover, Figure 10 suggests that the noisy and dropout formulations have comparable generalization performance when the number of noise injections grows to infinity.

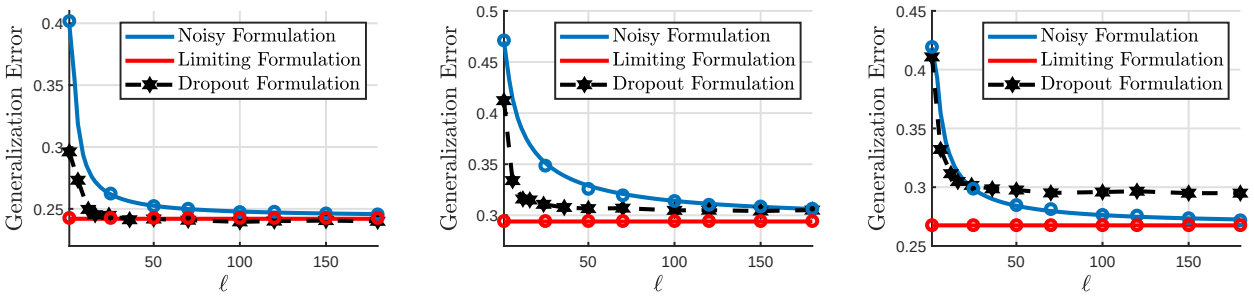


Figure 10. Solid line: Theoretical predictions. Circle: Numerical simulations for (3) and (8). Hexagram: Numerical simulations for the dropout formulation. **(a)** Sigmoid activation function and we set the parameters as $p = 500$, $\alpha = 1.4$, $\Delta = 1.5$, $\lambda = 10^{-4}$, $\eta = 2.5$ and $\epsilon = 0.3$. **(b)** ReLu activation function and we set the parameters as $p = 500$, $\alpha = 1.4$, $\Delta = 3.5$, $\lambda = 10^{-3}$, $\eta = 1.5$ and $\epsilon = 0.3$. **(c)** Sign activation function and we set the parameters as $p = 500$, $\alpha = 1.6$, $\Delta = 1.5$, $\lambda = 10^{-4}$, $\eta = 2$ and $\epsilon = 0.4$. The feature matrix $\boldsymbol{F}$ is the uniform feature matrix. The results are averaged over 35 independent Monte Carlo trials.

8. Appendix II: Analysis of the Noisy Formulation

In this part, we provide a rigorous proof of the predictions stated in Theorem 1, Theorem 2 and Lemma 1. To this end, we suppose that the assumptions considered in Sections 2 and 3 are all satisfied. We derive our theoretical results using an extended version of the CGMT framework which we refer to as the multivariate CGMT.

8.1. Multivariate Convex Gaussian Min–Max Theorem

To derive the asymptotic results stated in Theorem 1, Theorem 2 and Lemma 1, we use an extended version of the CGMT framework introduced in (Thrampoulidis et al., 2016). The CGMT is used to accurately analyze a generally hard primary formulation by introducing an asymptotically equivalent auxiliary optimization problem. In this paper, we consider primary optimization problems of the following form

$$\Phi_k = \min_{\boldsymbol{w}\in\mathcal{S}_{\boldsymbol{w}}} \max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}} \sum_{i=1}^{\ell} \boldsymbol{u}_i \boldsymbol{G}_i \boldsymbol{w}_i + \psi(\boldsymbol{w}, \boldsymbol{u}), \tag{21}$$

where $\boldsymbol{u}_i \in \mathbb{R}^{n_i}$ and $\boldsymbol{w}_i \in \mathbb{R}^{k_i}$ are optimization variables and $\boldsymbol{G}_i \in \mathbb{R}^{n_i \times k_i}$ has independent standard Gaussian random components, for any $i \in \{1 \dots, \ell\}$. Additionally, the vectors $\boldsymbol{w}$ and $\boldsymbol{u}$ are formed by the concatenation of the vectors $\{\boldsymbol{w}_i\}_{i=1}^{\ell}$ and $\{\boldsymbol{u}_i\}_{i=1}^{\ell}$, respectively. We refer to the formulation in (21) as the multivariate primary optimization (multivariate PO). We show that the corresponding multivariate auxiliary optimization (multivariate AO) is given by

$$\phi_k = \min_{\boldsymbol{w}\in\mathcal{S}_{\boldsymbol{w}}} \max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}} \sum_{i=1}^{\ell} \|\boldsymbol{u}_i\| \boldsymbol{g}_i^{\top} \boldsymbol{w}_i + \sum_{i=1}^{\ell} \|\boldsymbol{w}_i\| \boldsymbol{h}_i^{\top} \boldsymbol{u}_i + \psi(\boldsymbol{w}, \boldsymbol{u}), \tag{22}$$

where $\boldsymbol{g}_i \in \mathbb{R}^{k_i}$ and $\boldsymbol{h}_i \in \mathbb{R}^{n_i}$ are independent standard Gaussian random vectors, for any $i \in \{1 \dots, \ell\}$. Here, we assume that $\boldsymbol{G}_i \in \mathbb{R}^{n_i \times k_i}$, $\boldsymbol{g}_i \in \mathbb{R}^{k_i}$ and $\boldsymbol{h}_i \in \mathbb{R}^{n_i}$, are all independent, the feasibility sets $\mathcal{S}_{\boldsymbol{w}} \subset \mathbb{R}^k$ and $\mathcal{S}_{\boldsymbol{u}} \subset \mathbb{R}^n$ are convex and compact, and the function $\psi : \mathbb{R}^k \times \mathbb{R}^n \to \mathbb{R}$ is continuous *convex-concave* on $\mathcal{S}_{\boldsymbol{w}} \times \mathcal{S}_{\boldsymbol{u}}$, where $k = \sum_{i=1}^{\ell} k_i$ and $n = \sum_{i=1}^{\ell} n_i$. Now, we summarize our theoretical result in the following theorem.

Theorem 3 (Multivariate CGMT)**.** *Assume that the above assumptions are all satisfied. For any fixed $\ell \geq 1$ and $\epsilon > 0$, consider an open set $\mathcal{S}_{p,\epsilon}$. Moreover, define the set $\mathcal{S}_{p,\epsilon}^c = \mathcal{S}_{\boldsymbol{w}} \setminus \mathcal{S}_{p,\epsilon}$. Let ϕ_k and ϕ_k^c be the optimal cost values of the multivariate AO formulation in (22) with feasibility sets $\mathcal{S}_{\boldsymbol{w}}$ and $\mathcal{S}_{p,\epsilon}^c$, respectively. Assume that the following properties are all satisfied*

(1) *There exists a constant ϕ such that the optimal cost ϕ_k converges in probability to ϕ as k goes to $+\infty$.*

(2) *There exists a constant ϕ^c such that the optimal cost ϕ_k^c converges in probability to ϕ^c as k goes to $+\infty$, for any fixed $\epsilon > 0$.*

(3) *There exists a positive constant $\zeta > 0$ such that $\phi^c \geq \phi + \zeta$, for any fixed $\epsilon > 0$.*

Then, the following convergence in probability holds

$$|\Phi_k - \phi_k| \overset{k\to+\infty}{\longrightarrow} 0, \text{ and } \mathbb{P}(\widehat{\boldsymbol{w}}_k \in \mathcal{S}_{p,\epsilon}) \overset{k\to\infty}{\longrightarrow} 1,$$

for any fixed $\epsilon > 0$, where Φ_k and $\widehat{\boldsymbol{w}}_k$ are the optimal cost and the optimal solution of the multivariate PO formulation in (21).

The above theorem allows us to analyze the generally easy multivariate AO formulation given in (22) to infer asymptotic properties of the generally hard multivariate PO formulation in (21). The proof of Theorem 3 follows by showing that the formulation in (22) and the following formulation

$$\widehat{\Phi}_k = \min_{\boldsymbol{w}\in\mathcal{S}_{\boldsymbol{w}}} \max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}} \sum_{i=1}^{\ell} \boldsymbol{u}_i \boldsymbol{G}_i \boldsymbol{w}_i + \psi(\boldsymbol{w},\boldsymbol{u}) + \sum_{i=1}^{\ell} \|\boldsymbol{u}_i\| \|\boldsymbol{w}_i\| g_i,$$

satisfy all the assumptions in (Gordon, 1988), where $\{g_i\}_{i=1}^{\ell}$ are independent standard Gaussian random variables. Then, following the same analysis in (Thrampoulidis et al., 2015) and (Thrampoulidis et al., 2016), we can show that for any $\chi \in \mathbb{R}$ and $c > 0$, it holds

$$\mathbb{P}\left(|\Phi_k - \chi| > c\right) \leq 2^{\ell}\, \mathbb{P}\left(|\phi_k - \chi| > c\right). \tag{23}$$

Combining this result with the assumptions of Theorem 3 completes the proof. We omit the detailed proof since it is similar to the analysis in (Thrampoulidis et al., 2015) and (Thrampoulidis et al., 2016). We refer to Theorem 3 as the multivariate convex Gaussian min-max theorem (multivariate CGMT).

Next, we use the multivariate CGMT to rigorously prove the technical results provided in Theorem 1, Theorem 2 and Lemma 1. Our approach is to reformulate the Gaussian formulation in (6) in the form of the multivariate PO problem given in (21). Then, use the multivariate CGMT framework to show that the formulation in (3) is asymptotically equivalent to an easier formulation that can be written in the form of the multivariate AO problem given in (22). The next step is to show that the multivariate AO formulation converges in probability to a deterministic problem that can be expressed in the form of the formulation given in (11).

8.2. Asymptotic Analysis of the Noisy Formulation

In this part, we provide the technical steps to obtain the theoretical results stated in Theorem 1. Specifically, we use the multivariate CGMT framework to precisely analyze the noisy formulation introduced in (3). Next, we suppose that the assumptions introduced in Section 3 are all satisfied.

8.2.1. Formulating the Multivariate Primary Formulation

Based on the cGEC introduced in Section 2, it suffices to precisely analyze the Gaussian formulation in the large system limit. Then, it suffices to analyze the following formulation

$$\begin{aligned}\min_{\boldsymbol{w}\in\mathbb{R}^k} \frac{1}{2n\ell} \sum_{i=1}^{n}\sum_{j=1}^{\ell} \big(y_i - \widetilde{\mu}_1 \boldsymbol{w}^\top \boldsymbol{F}^\top \boldsymbol{a}_i - \widehat{\mu}_1 \boldsymbol{w}^\top \boldsymbol{F}^\top \boldsymbol{z}_{ij} \\ - \mu_0 \boldsymbol{w}^\top \mathbf{1}_k - \mu_2 \boldsymbol{w}^\top \boldsymbol{b}_i - \mu_3 \boldsymbol{w}^\top \boldsymbol{p}_{ij}\big)^2 + \tfrac{\lambda}{2}\|\boldsymbol{w}\|^2.\end{aligned} \tag{24}$$

Note that the formulation in (24) is strongly convex with a strong convexity parameter equals to λ. This means that it has a unique optimal solution. Note that the multivariate CGMT framework assumes that the feasibility sets of the multivariate PO formulation in (21) are compact. The following lemma shows that this assumption is satisfied by our formulation.

Lemma 2 (Primal Compactness)**.** *Assume that $\widehat{\boldsymbol{w}}_p$ is the unique optimal solution of the formulation given in (24). Then, there exist two positive constants $C_w > 0$ and $C_\vartheta > 0$ such that*

$$\mathbb{P}\big(\|\widehat{\boldsymbol{w}}_p\| \leq C_w\big) \overset{p\to\infty}{\longrightarrow} 1,\ \mathbb{P}\big(|\mathbf{1}_k^\top \widehat{\boldsymbol{w}}_p| \leq C_\vartheta\big) \overset{p\to\infty}{\longrightarrow} 1, \tag{25}$$

where the second asymptotic result is valid only when $\mu_0 \neq 0$.

Given that the loss function in (24) is proper and strongly convex, one can use the results in Lemma 1 in (Dhifallah & Lu, 2020) to prove Lemma 2. This asymptotic result follows using Assumptions 3, 4, and 5 and Theorem 2.1 in (Rudelson & Vershynin, 2010). Combining this result with the theoretical result stated in Proposition 1 in (Dhifallah & Lu, 2020), the Gaussian formulation is asymptotically equivalent to the following formulation

$$\min_{\substack{c_w \leq \|\boldsymbol{w}\| \leq C_w \\ |\vartheta| \leq C_\vartheta}} \frac{1}{2n\ell} \sum_{j=1}^{\ell} \|\boldsymbol{y} - [\widetilde{\mu}_1 \boldsymbol{A} + \widehat{\mu}_1 \boldsymbol{Z}_j]\boldsymbol{F}\boldsymbol{w} - \mu_0 \vartheta \mathbf{1}_n - \mu_2 \boldsymbol{B}\boldsymbol{w} - \mu_3 \boldsymbol{P}_j \boldsymbol{w}\|^2 + \tfrac{\lambda}{2}\|\boldsymbol{w}\|^2, \tag{26}$$

where the matrices $\boldsymbol{A} \in \mathbb{R}^{n \times p}$, $\boldsymbol{Z}_j \in \mathbb{R}^{n \times p}$, $\boldsymbol{B} \in \mathbb{R}^{n \times k}$ and $\boldsymbol{P}_j \in \mathbb{R}^{n \times k}$ are formed by the concatenation of the vectors $\boldsymbol{a}_i$, $\boldsymbol{z}_{ij}$, $\boldsymbol{b}_i$ and $\boldsymbol{p}_{ij}$, respectively. Here, the label vector $\boldsymbol{y} \in \mathbb{R}^n$ is formed by the labels $\{y_i\}_{i=1}^n$ and can be expressed as follows $\boldsymbol{y} = \varphi(\boldsymbol{A}\boldsymbol{\xi})$. Based on the analysis in Proposition 2 in (Dhifallah & Lu, 2020), it suffices to precisely analyze the problem in (26) for fixed feasible ϑ. Then, minimize its asymptotic limit over ϑ to infer the asymptotic properties of (26). Next, we start by analyzing the formulation in (26) for fixed feasible ϑ. To express (26) in the form of the multivariate PO introduced in (21), we introduce additional dual optimization variables. Specifically, the formulation in (26) can be equivalently formulated as follows

$$\begin{aligned} \min_{\|\boldsymbol{w}\| \leq C_w} \max_{\boldsymbol{u} \in \mathbb{R}^{\ell n}} & -\frac{\|\boldsymbol{u}\|^2}{2n\ell} + \frac{1}{n\ell} \sum_{j=1}^{\ell} \boldsymbol{u}_j^\top \Big(\mu_0 \vartheta \mathbf{1}_n + \mu_2 \boldsymbol{B}\boldsymbol{w} - \boldsymbol{y} \\ & + [\widetilde{\mu}_1 \boldsymbol{A} + \widehat{\mu}_1 \boldsymbol{Z}_j]\boldsymbol{F}\boldsymbol{w} + \mu_3 \boldsymbol{P}_j \boldsymbol{w} \Big) + \tfrac{\lambda}{2}\|\boldsymbol{w}\|^2, \end{aligned} \tag{27}$$

where the dual optimization vector $\boldsymbol{u} \in \mathbb{R}^{\ell n}$ can be decomposed as follows $\boldsymbol{u}^\top = [\boldsymbol{u}_1^\top, \ldots, \boldsymbol{u}_\ell^\top]^\top$, where $\boldsymbol{u}_i \in \mathbb{R}^n$, for any $1 \leq i \leq \ell$. Note that the optimization problem given in (27) has a unique optimal solution. The multivariate CGMT also assumes that the feasibility set of the maximization problem in (27) is compact. The following lemma shows that this assumption is also satisfied by our formulation.

Lemma 3 (Dual Compactness)**.** *Assume that $\widehat{\boldsymbol{u}}_p$ is the unique optimal solution of the formulation given in (27). Then, there exists a positive constants $C_u > 0$ such that*

$$\mathbb{P}\big(\|\widehat{\boldsymbol{u}}_p\|/\sqrt{n} \leq C_u\big) \overset{p \to \infty}{\longrightarrow} 1. \tag{28}$$

This result can also be proved using similar steps as in Lemma 2 in (Dhifallah & Lu, 2020). Specifically, we can use the result in Proposition 11.3 in (Rockafellar & Wets, 1998) to show the compactness of the optimal dual vector $\widehat{\boldsymbol{u}}_p$. The results in Lemma 2 and Lemma 3 show that the Gaussian formulation is asymptotically equivalent to the following formulation

$$\begin{aligned} \min_{\|\boldsymbol{w}\| \leq C_w} \max_{\frac{\|\boldsymbol{u}\|}{\sqrt{n}} \leq C_u} & -\frac{\|\boldsymbol{u}\|^2}{2n\ell} + \frac{1}{n\ell} \sum_{j=1}^{\ell} \boldsymbol{u}_j^\top \big(\mu_0 \vartheta \mathbf{1}_n + \mu_2 \boldsymbol{B}\boldsymbol{w} - \boldsymbol{y} \\ & + [\widetilde{\mu}_1 \boldsymbol{A} + \widehat{\mu}_1 \boldsymbol{Z}_j]\boldsymbol{F}\boldsymbol{w} + \mu_3 \boldsymbol{P}_j \boldsymbol{w} \big) + \tfrac{\lambda}{2}\|\boldsymbol{w}\|^2. \end{aligned} \tag{29}$$

Next, we focus on precisely analyzing the formulation in (29). Now, note that the label vector $\boldsymbol{y}$ depend on the Gaussian matrix $\boldsymbol{A}$. Then, we decompose $\boldsymbol{A}$ as follows

$$\boldsymbol{A} = \boldsymbol{A}\boldsymbol{P}_{\boldsymbol{\xi}} + \boldsymbol{A}\boldsymbol{P}_{\boldsymbol{\xi}}^{\perp} = \boldsymbol{A}\boldsymbol{\xi}\boldsymbol{\xi}^\top + \boldsymbol{A}\boldsymbol{P}_{\boldsymbol{\xi}}^{\perp}, \tag{30}$$

where $\boldsymbol{P}_{\boldsymbol{\xi}} \in \mathbb{R}^{p \times p}$ denotes the projection matrix onto the space spanned by the vector $\boldsymbol{\xi} \in \mathbb{R}^p$ and $\boldsymbol{P}_{\boldsymbol{\xi}}^{\perp} = \boldsymbol{I}_p - \boldsymbol{\xi}\boldsymbol{\xi}^\top$ denotes the projection matrix onto the orthogonal complement of the space spanned by the vector $\boldsymbol{\xi}$. Note that the random matrix $\boldsymbol{A}\boldsymbol{\xi}\boldsymbol{\xi}^\top$ is independent of the random matrix $\boldsymbol{A}\boldsymbol{P}_{\boldsymbol{\xi}}^{\perp}$. Then, we can express $\boldsymbol{A}$ as follows without changing its statistics

$$\boldsymbol{A} = \boldsymbol{s}\boldsymbol{\xi}^\top + \boldsymbol{A}\boldsymbol{P}_{\boldsymbol{\xi}}^{\perp}, \tag{31}$$

where $\boldsymbol{s} \in \mathbb{R}^n$ has independent standard Gaussian components and the two random quantities $\boldsymbol{s}$ and $\boldsymbol{A}$ are independent. This shows that the optimization problem formulated in (29) is statistically equivalent to the following formulation

$$\begin{aligned} \min_{\|\boldsymbol{w}\| \leq C_w} \max_{\frac{\|\boldsymbol{u}\|}{\sqrt{n}} \leq C_u} & -\frac{\|\boldsymbol{u}\|^2}{2n\ell} + \frac{1}{n\ell} \sum_{j=1}^{\ell} \boldsymbol{u}_j^\top \big(-\boldsymbol{y} + \widetilde{\mu}_1 \boldsymbol{s}\boldsymbol{\xi}^\top \boldsymbol{F}\boldsymbol{w} \\ & + \mu_0 \vartheta \mathbf{1}_n + \boldsymbol{G}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{w} + \boldsymbol{T}_j \boldsymbol{\Gamma}^{\frac{1}{2}}\boldsymbol{w} \big) + \tfrac{\lambda}{2}\|\boldsymbol{w}\|^2, \end{aligned} \tag{32}$$

where $\boldsymbol{G}$ and $\{\boldsymbol{T}_j\}_{j=1}^{\ell}$ are independent matrices with independent and identically distributed standard Gaussian components. Here, $\boldsymbol{G} \in \mathbb{R}^{n\times k}$, $\boldsymbol{T}_j \in \mathbb{R}^{n\times k}$. Moreover, $\boldsymbol{\Sigma} \in \mathbb{R}^{k\times k}$ and $\boldsymbol{\Gamma} \in \mathbb{R}^{k\times k}$ are positive definite matrices and defined as follows

$$\begin{cases} \boldsymbol{\Sigma} = \widetilde{\mu}_1^2 \boldsymbol{F}^\top \boldsymbol{P}_{\boldsymbol{\xi}}^{\perp} \boldsymbol{F} + \mu_2^2 \boldsymbol{I}_k \\ \boldsymbol{\Gamma} = \widehat{\mu}_1^2 \boldsymbol{F}^\top \boldsymbol{F} + \mu_3^2 \boldsymbol{I}_k. \end{cases} \tag{33}$$

The above results show that it suffices to precisely analyze the formulation given in (32). Moreover, note that (32) can be equivalently formulated as follows

$$\begin{aligned} \min_{\|\boldsymbol{w}\|\leq C_w} \max_{\frac{\|\boldsymbol{u}\|}{\sqrt{n}}\leq C_u} & \frac{\mu_0 \vartheta}{n\ell} \boldsymbol{u}^\top \mathbf{1}_{\ell n} - \frac{\|\boldsymbol{u}\|^2}{2n\ell} - \frac{\boldsymbol{u}^\top \widehat{\boldsymbol{y}}}{n\ell} + \frac{\widetilde{\mu}_1 \bar{\boldsymbol{\xi}}^\top \boldsymbol{F}\boldsymbol{w}}{\ell n} \boldsymbol{u}^\top \widehat{\boldsymbol{s}} \\ & + \frac{1}{n\ell} \boldsymbol{u}^\top \boldsymbol{M}^\top \boldsymbol{G} \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{w} + \frac{1}{n\ell} \boldsymbol{u}^\top \boldsymbol{T} \boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{w} + \tfrac{\lambda}{2} \|\boldsymbol{w}\|^2, \end{aligned} \tag{34}$$

where $\mathbf{1}_{\ell n}$ denotes the all one vector of size ℓn, $\widehat{\boldsymbol{y}} \in \mathbb{R}^{\ell n}$ and $\widehat{\boldsymbol{s}} \in \mathbb{R}^{\ell n}$ are formed after performing an ℓ times concatenation of the vectors $\boldsymbol{y}$ and $\boldsymbol{s}$, respectively. Here, the matrices $\boldsymbol{T} \in \mathbb{R}^{n\ell\times k}$ and $\boldsymbol{M} \in \mathbb{R}^{n\times \ell n}$ are given as follows

$$\boldsymbol{T}^\top = [\boldsymbol{T}_1^\top, \boldsymbol{T}_2^\top, \ldots, \boldsymbol{T}_\ell^\top]^\top, \ \boldsymbol{M} = [\boldsymbol{I}_n, \boldsymbol{I}_n, \ldots, \boldsymbol{I}_n].$$

We can notice that the optimization problem formulated in (34) is in the form of the multivariate PO problem given in (21). Therefore, applying the multivariate CGMT, the corresponding multivariate AO problem can be expressed as follows

$$\begin{aligned} \min_{\|\boldsymbol{w}\|\leq C_w} \max_{\frac{\|\boldsymbol{u}\|}{\sqrt{n}}\leq C_u} & - \frac{\|\boldsymbol{u}\|^2}{2n\ell} + \frac{\widetilde{\mu}_1 \bar{\boldsymbol{\xi}}^\top \boldsymbol{F}\boldsymbol{w}}{\ell n} \boldsymbol{u}^\top \widehat{\boldsymbol{s}} + \frac{\mu_0 \vartheta}{n\ell} \boldsymbol{u}^\top \mathbf{1}_{\ell n} \\ & - \frac{\boldsymbol{u}^\top \widehat{\boldsymbol{y}}}{n\ell} + \frac{1}{n\ell} \|\boldsymbol{M}\boldsymbol{u}\| \boldsymbol{g}_1^\top \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{w} + \frac{1}{\ell n} \|\boldsymbol{u}\| \boldsymbol{g}_2^\top \boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{w} \\ & + \frac{\|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{w}\|}{\ell n} \boldsymbol{h}_1^\top \boldsymbol{M}\boldsymbol{u} + \frac{\|\boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{w}\|}{\ell n} \boldsymbol{h}_2^\top \boldsymbol{u} + \frac{\lambda}{2} \|\boldsymbol{w}\|^2, \end{aligned} \tag{35}$$

where the vectors $\boldsymbol{g}_1 \in \mathbb{R}^k$, $\boldsymbol{g}_2 \in \mathbb{R}^k$, $\boldsymbol{h}_1 \in \mathbb{R}^n$ and $\boldsymbol{h}_2 \in \mathbb{R}^{\ell n}$ are independent standard Gaussian random vectors. First, observe that the convexity assumption in Theorem 3 is satisfied by our multivariate PO formulation in (34). Furthermore, note that the compactness assumptions in the multivariate CGMT framework are also satisfied by our primary problem in (34). Then, following the multivariate CGMT framework, we focus on analyzing the multivariate AO formulation introduced in (35). Specifically, the objective is to simplify the multivariate AO problem and study its asymptotic properties.

8.2.2. Simplifying the Multivariate Auxiliary Formulation

In this part, our objective is to simplify the multivariate AO problem given in (35). Specifically, the main objective is to express the formulation in (35) in terms of scalar optimization variables. First, observe that the singular value decomposition (SVD) of the matrix $\boldsymbol{M}$ can be expressed as $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{V} \in \mathbb{R}^{\ell n\times \ell n}$ are two orthogonal matrices and $\boldsymbol{S} \in \mathbb{R}^{n\times \ell n}$ is given by $\boldsymbol{S} = [\sqrt{\ell}\boldsymbol{I}_n \ \ \mathbf{0}_{n\times(\ell-1)n}]$. Therefore, the optimization problem expressed in (35) can be formulated as follows

$$\begin{aligned} \min_{\|\boldsymbol{w}\|\leq C_w} \max_{\frac{\|\boldsymbol{u}\|}{\sqrt{n}}\leq C_u} & \frac{\widetilde{\mu}_1 \bar{\boldsymbol{\xi}}^\top \boldsymbol{F}\boldsymbol{w}}{\ell n} \boldsymbol{u}^\top \boldsymbol{V}^\top \widehat{\boldsymbol{s}} - \frac{\|\boldsymbol{u}\|^2}{2n\ell} + \frac{\mu_0 \vartheta}{n\ell} \boldsymbol{u}^\top \boldsymbol{V}^\top \mathbf{1}_{\ell n} \\ & - \frac{\boldsymbol{u}^\top \boldsymbol{V}^\top \widehat{\boldsymbol{y}}}{n\ell} + \frac{\sqrt{\ell}}{n\ell} \|\boldsymbol{u}_1\| \boldsymbol{g}_1^\top \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{w} + \frac{1}{\ell n} \|\boldsymbol{u}\| \boldsymbol{g}_2^\top \boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{w} \\ & + \frac{\sqrt{\ell}}{\ell n} \|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{w}\| \boldsymbol{h}_1^\top \boldsymbol{u}_1 + \frac{\|\boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{w}\|}{\ell n} \boldsymbol{h}_2^\top \boldsymbol{u} + \tfrac{\lambda}{2} \|\boldsymbol{w}\|^2, \end{aligned} \tag{36}$$

where we perform the change of variable $\boldsymbol{u}_{\text{new}} = \boldsymbol{V}^\top \boldsymbol{u}$, we decompose the new vector as $\boldsymbol{u}_{\text{new}}^\top = [\boldsymbol{u}_1^\top, \ldots, \boldsymbol{u}_\ell^\top]$ and we replace $\boldsymbol{u}_{\text{new}}$ by $\boldsymbol{u}$. Now, we denote by t_1 and t_2 the norms of the independent vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_{-1}^\top = [\boldsymbol{u}_2^\top, \ldots, \boldsymbol{u}_\ell^\top]$, i.e. $t_1 = \|\boldsymbol{u}_1\|$ and $t_2 = \|\boldsymbol{u}_{-1}\|$. Additionally, we decompose the orthogonal matrix $\boldsymbol{V}$ as follows $\boldsymbol{V} = [\boldsymbol{V}_1 \ \ \boldsymbol{V}_2]$, where

$\boldsymbol{V}_1 \in \mathbb{R}^{\ell n \times n}$ and $\boldsymbol{V}_2 \in \mathbb{R}^{\ell n \times (\ell-1)n}$. Define the vector $\boldsymbol{v} \in \mathbb{R}^k$ as $\boldsymbol{v} = \boldsymbol{F}^\top \boldsymbol{\xi}$ and the scalar q as $q = \bar{\boldsymbol{v}}^\top \boldsymbol{w}$, where $\bar{\boldsymbol{v}}$ is defined as follows $\bar{\boldsymbol{v}} = \boldsymbol{v}/\|\boldsymbol{v}\|$. Also, define the scalar $T_{p,1}$ as $T_{p,1} = \|\boldsymbol{v}\|$.

Now, we are ready to further simplify the multivariate AO formulation. The first step is to fix t_1 and t_2 and solve the formulation in (36) over the direction of the independent vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_{-1}$. Specifically, based on the result in Lemma 3, the formulation given in (36) can be simplified as follows

$$\begin{aligned}
&\min_{\|\boldsymbol{w}\| \leq C_w} \max_{\substack{0 \leq t_1/\sqrt{n} \leq C_{t_1} \\ 0 \leq t_2/\sqrt{n} \leq C_{t_2}}} \frac{\sqrt{\ell} t_1}{n\ell} \boldsymbol{g}_1^\top \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{w} + \frac{\sqrt{t_1^2 + t_2^2}}{\ell n} \boldsymbol{g}_2^\top \boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{w} - \frac{t_1^2 + t_2^2}{2n\ell} + \tfrac{\lambda}{2} \|\boldsymbol{w}\|^2 \\
&+ \frac{t_1}{\ell n} \|\sqrt{\ell} \|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{w}\| \boldsymbol{h}_1 + \|\boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{w}\| \widehat{\boldsymbol{h}}_2 - \boldsymbol{V}_1^\top \widehat{\boldsymbol{y}} + \mu_0 \vartheta \boldsymbol{V}_1^\top \boldsymbol{1}_{\ell n} + \widetilde{\mu}_1 T_{p,1} q \boldsymbol{V}_1^\top \widehat{\boldsymbol{s}}\| \\
&+ \frac{t_2}{\ell n} \|\|\boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{w}\| \widetilde{\boldsymbol{h}}_2 - \boldsymbol{V}_2^\top \widehat{\boldsymbol{y}} + \mu_0 \vartheta \boldsymbol{V}_2^\top \boldsymbol{1}_{\ell n} + \widetilde{\mu}_1 T_{p,1} q \boldsymbol{V}_2^\top \widehat{\boldsymbol{s}}\|,
\end{aligned} \tag{37}$$

where we decompose the Gaussian vector $\boldsymbol{h}_2$ as $\boldsymbol{h}_2 = [\widehat{\boldsymbol{h}}_2^\top \ \ \widetilde{\boldsymbol{h}}_2^\top]$, where $\widehat{\boldsymbol{h}}_2 \in \mathbb{R}^n$ and $\widetilde{\boldsymbol{h}}_2 \in \mathbb{R}^{(\ell-1)n}$. Here, C_{t_1} and C_{t_2} are sufficiently large positive constants that ensure the asymptotic result in Lemma 3. Note that it remains to solve over the primal vector $\boldsymbol{w}$ to obtain a scalar formulation of the multivariate AO problem. We continue our analysis by defining the following optimization problem

$$\begin{aligned}
&\min_{\|\boldsymbol{w}\| \leq C_w} \max_{\substack{0 \leq t_1/\sqrt{n} \leq C_{t_1} \\ 0 \leq t_2/\sqrt{n} \leq C_{t_2}}} \frac{\sqrt{\ell} t_1}{n\ell} \boldsymbol{g}_1^\top \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{w} + \frac{\sqrt{t_1^2 + t_2^2}}{\ell n} \boldsymbol{g}_2^\top \boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{w} - \frac{t_1^2 + t_2^2}{2n\ell} + \tfrac{\lambda}{2} \|\boldsymbol{w}\|^2 \\
&+ \frac{t_1}{\ell n} \sqrt{\ell \|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{w}\|^2 \|\boldsymbol{h}_1\|^2 + \|\boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{w}\|^2 \|\widehat{\boldsymbol{h}}_2\|^2 + \|\mu_0 \vartheta \boldsymbol{V}_1^\top \boldsymbol{1}_{\ell n} + \widetilde{\mu}_1 T_{p,1} q \boldsymbol{V}_1^\top \widehat{\boldsymbol{s}} - \boldsymbol{V}_1^\top \widehat{\boldsymbol{y}}\|^2} \\
&+ \frac{t_2}{\ell n} \sqrt{\|\boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{w}\|^2 \|\widetilde{\boldsymbol{h}}_2\|^2 + \|\mu_0 \vartheta \boldsymbol{V}_2^\top \boldsymbol{1}_{\ell n} - \boldsymbol{V}_2^\top \widehat{\boldsymbol{y}} + \widetilde{\mu}_1 T_{p,1} q \boldsymbol{V}_2^\top \widehat{\boldsymbol{s}}\|^2}.
\end{aligned} \tag{38}$$

Note that the difference between the cost functions of the formulations in (37) and (38) are terms that converge in probability to zero. Before showing the asymptotic equivalence between the formulations in (37) and (38), we provide important convexity properties of the optimization problem in (38) as given in the following lemma.

Lemma 4 (Strong–convexity of (38))**.** *Define $\widehat{f}_{p,2}$ as the cost function of the problem in (38). Then, $\widehat{f}_{p,2}$ is strongly convex in the vector $\boldsymbol{w}$ where λ is a strong convexity parameter. Moreover, it is strongly concave in the variables t_1 and t_2 in the feasibility sets where -1 is a strong concavity parameter.*

Proof. The strong convexity can be proved by observing that the cost function of (38) is a positive sum of convex and strongly convex functions in terms of $\boldsymbol{w}$ for fixed feasible t_1 and t_2. Moreover, note that the term $\sqrt{t_1^2 + t_2^2} \boldsymbol{g}_2^\top \boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{w}$ can be replaced with $t_1 \boldsymbol{g}_{21}^\top \boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{w} + t_2 \boldsymbol{g}_{22}^\top \boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{w}$ without changing the statistics of our formulation, where $\boldsymbol{g}_{21}$ and $\boldsymbol{g}_{22}$ are two independent Gaussian vectors. Then, one can see that the cost function of (38) is strongly concave in the variables t_1 and t_2 where -1 is a strong concavity parameter. □

Lemma 4 provides important convexity properties of the optimization problem formulated in (38). These properties are essential to prove the equivalence between (37) and (38) as stated in the following lemma.

Lemma 5 (High-dimensional Equivalence I)**.** *Define $\widehat{\mathcal{S}}_{p,1}^\star$ and $\widehat{\mathcal{S}}_{p,2}^\star$ as the sets of optimal solutions of the minimization problems in (37) and (38), respectively. Moreover, let $\widehat{O}_{p,1}^\star$ and $\widehat{O}_{p,2}^\star$ be the optimal objective values of the optimization problems in (37) and (38), respectively. Then, the following convergence in probability holds*

$$|\widehat{O}_{p,1}^\star - \widehat{O}_{p,2}^\star| \overset{p \to +\infty}{\longrightarrow} 0, \ \text{and} \ \mathbb{D}(\widehat{\mathcal{S}}_{p,1}^\star, \widehat{\mathcal{S}}_{p,2}^\star) \overset{p \to +\infty}{\longrightarrow} 0, \tag{39}$$

where $\mathbb{D}(\mathcal{A}, \mathcal{B})$ denotes the deviation between the sets $\mathcal{A}$ and $\mathcal{B}$ and is defined as $\mathbb{D}(\mathcal{A}, \mathcal{B}) = \sup_{\boldsymbol{x}_1 \in \mathcal{A}} \inf_{\boldsymbol{x}_2 \in \mathcal{B}} \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2$.

The detailed proof of Lemma 5 is deferred to Appendix 10.1. Lemma 5 particularly shows that the optimization problems in (37) and (38) are asymptotically equivalent. Then, it suffices to precisely analyze the formulation in (38). To solve over the

primal vector $\boldsymbol{w}$, we introduce two independent scalar optimization variables τ_1 and τ_2 where they both solve optimization problems of the following form

$$\sqrt{x} = \inf_{\tau>0} \ \frac{\tau}{2} + \frac{x}{2\tau}, \text{ for any } x \geq 0. \tag{40}$$

Here, note that the optimal solution of the problem in (40) can be expressed as $\tau^\star = \sqrt{x}$. Next, we use the identity in (40) to transform the non-smooth square roots in the cost function of the formulation given in (38) to smooth terms. This is an essential step to solve over the primal vector $\boldsymbol{w}$. Specifically, based on the result in Lemma 2, our multivariate AO formulation given in (38) can be expressed as follows

$$\begin{aligned}
&\min_{\|\boldsymbol{w}\|\leq C_w} \max_{\substack{0\leq t_1\leq C_{t_1}\\ 0\leq t_2\leq C_{t_2}}} \inf_{\substack{\tau_1>0\\ \tau_2>0}} \frac{\sqrt{\ell}t_1}{\sqrt{n\ell}}\boldsymbol{g}_1^\top \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{w} + \frac{\sqrt{t_1^2+t_2^2}}{\ell\sqrt{n}}\boldsymbol{g}_2^\top\boldsymbol{\Gamma}^{\frac{1}{2}}\boldsymbol{w} + \frac{\tau_1 t_1}{2\ell} + \frac{\tau_2 t_2}{2\ell} - \frac{t_1^2+t_2^2}{2\ell} + \tfrac{\lambda}{2}\|\boldsymbol{w}\|^2\\
&+ \frac{t_1\|\boldsymbol{h}_1\|^2}{2\tau_1 n}\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{w}\|^2 + \frac{\zeta_{p,t,\tau}}{2n}\|\boldsymbol{\Gamma}^{\frac{1}{2}}\boldsymbol{w}\|^2 + \frac{t_1}{2\tau_1\ell n}\|\mu_0\vartheta\boldsymbol{V}_1^\top\mathbf{1}_{\ell n} - \boldsymbol{V}_1^\top\widehat{\boldsymbol{y}} + \widetilde{\mu}_1 T_{p,1} q\boldsymbol{V}_1^\top\widehat{\boldsymbol{s}}\|^2\\
&+ \frac{t_2}{2\tau_2\ell n}\|\mu_0\vartheta\boldsymbol{V}_2^\top\mathbf{1}_{\ell n} - \boldsymbol{V}_2^\top\widehat{\boldsymbol{y}} + \widetilde{\mu}_1 T_{p,1} q\boldsymbol{V}_2^\top\widehat{\boldsymbol{s}}\|^2,
\end{aligned} \tag{41}$$

where $\zeta_{p,t,\tau} = t_1\|\widehat{\boldsymbol{h}}_2\|^2/(\ell\tau_1) + t_2\|\widetilde{\boldsymbol{h}}_2\|^2/(\ell\tau_2)$. Here, we also perform the change of variable $t_{1,\text{new}} = t_1/\sqrt{n}$ and $t_{2,\text{new}} = t_2/\sqrt{n}$, then, replace $t_{1,\text{new}}$ and $t_{2,\text{new}}$ by t_1 and t_2. Note that the feasibility sets of the optimization variables τ_1 and τ_2 are open unbounded sets. To simplify the analysis, we show that the feasibility sets of the variables τ_1 and τ_2 can be restricted to compact sets with probability going to 1 as p grows to $+\infty$ as stated in the following lemma.

Lemma 6 (Additional Compactness)**.** *There exist positive constants independent of p, $c_{\tau_1} > 0$, $C_{\tau_1} > 0$, $c_{\tau_2} > 0$ and $C_{\tau_2} > 0$, such that the following convergence in probability holds*

$$\mathbb{P}(c_{\tau_1} \leq \widehat{\tau}_1 \leq C_{\tau_1}) \xrightarrow{p\to\infty} 1, \ \mathbb{P}(c_{\tau_2} \leq \widehat{\tau}_2 \leq C_{\tau_2}) \xrightarrow{p\to\infty} 1, \tag{42}$$

where $\widehat{\tau}_1$ and $\widehat{\tau}_2$ are the optimal solutions of the formulation in (41).

The detailed proof of Lemma 6 is provided in Appendix 10.2. Based on Lemmas 4 and 6, the optimization problem given in (41) is asymptotically equivalent to the following problem

$$\begin{aligned}
&\min_{\|\boldsymbol{w}\|\leq C_w} \max_{\substack{0\leq t_1\leq C_{t_1}\\ 0\leq t_2\leq C_{t_2}}} \min_{\substack{c_{\tau_1}\leq\tau_1\leq C_{\tau_1}\\ c_{\tau_2}\leq\tau_2\leq C_{\tau_2}}} \frac{\sqrt{\ell}t_1}{\sqrt{n\ell}}\boldsymbol{g}_1^\top \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{w} + \frac{\sqrt{t_1^2+t_2^2}}{\ell\sqrt{n}}\boldsymbol{g}_2^\top\boldsymbol{\Gamma}^{\frac{1}{2}}\boldsymbol{w} + \frac{\tau_1 t_1}{2\ell} + \frac{\tau_2 t_2}{2\ell} - \frac{t_1^2+t_2^2}{2\ell} + \tfrac{\lambda}{2}\|\boldsymbol{w}\|^2\\
&+ \frac{t_1\|\boldsymbol{h}_1\|^2}{2\tau_1 n}\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{w}\|^2 + \frac{\zeta_{p,t,\tau}}{2n}\|\boldsymbol{\Gamma}^{\frac{1}{2}}\boldsymbol{w}\|^2 + \frac{t_1}{2\tau_1\ell n}\|\mu_0\vartheta\boldsymbol{V}_1^\top\mathbf{1}_{\ell n} - \boldsymbol{V}_1^\top\widehat{\boldsymbol{y}} + \widetilde{\mu}_1 T_{p,1} q\boldsymbol{V}_1^\top\widehat{\boldsymbol{s}}\|^2\\
&+ \frac{t_2}{2\tau_2\ell n}\|\mu_0\vartheta\boldsymbol{V}_2^\top\mathbf{1}_{\ell n} - \boldsymbol{V}_2^\top\widehat{\boldsymbol{y}} + \widetilde{\mu}_1 T_{p,1} q\boldsymbol{V}_2^\top\widehat{\boldsymbol{s}}\|^2.
\end{aligned} \tag{43}$$

Now, we are ready to simplify the formulation in (43) over the optimization vector $\boldsymbol{w}$. We start our analysis by decomposing the optimization vector $\boldsymbol{w} \in \mathbb{R}^k$ as follows

$$\boldsymbol{w} = q\bar{\boldsymbol{v}} + \boldsymbol{B}_{\boldsymbol{v}}^\perp\boldsymbol{r}, \tag{44}$$

where $\boldsymbol{r} \in \mathbb{R}^{k-1}$ and $\boldsymbol{B}_{\boldsymbol{v}}^\perp \in \mathbb{R}^{k\times(k-1)}$ is formed by an orthonormal basis orthogonal to the vector $\boldsymbol{v} \in \mathbb{R}^k$. Based on the result in Lemma 2, one can equivalently formulate the problem in (43) as follows

$$\begin{aligned}
&\min_{\substack{|q|\leq C_q\\ \|\boldsymbol{r}\|\leq C_r}} \max_{\substack{0\leq t_1\leq C_{t_1}\\ 0\leq t_2\leq C_{t_2}}} \min_{\substack{c_{\tau_1}\leq\tau_1\leq C_{\tau_1}\\ c_{\tau_2}\leq\tau_2\leq C_{\tau_2}}} \frac{\sqrt{\ell}t_1}{\sqrt{n\ell}}\boldsymbol{g}_1^\top \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{B}_{\boldsymbol{v}}^\perp\boldsymbol{r} + \frac{\sqrt{t_1^2+t_2^2}}{\ell\sqrt{n}}\boldsymbol{g}_2^\top\boldsymbol{\Gamma}^{\frac{1}{2}}\boldsymbol{B}_{\boldsymbol{v}}^\perp\boldsymbol{r} + \frac{\tau_1 t_1}{2\ell} + \frac{\tau_2 t_2}{2\ell} - \frac{t_1^2+t_2^2}{2\ell}\\
&+ \frac{t_1\|\boldsymbol{h}_1\|^2}{2\tau_1 n}\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{B}_{\boldsymbol{v}}^\perp\boldsymbol{r} + q\boldsymbol{\Sigma}^{\frac{1}{2}}\bar{\boldsymbol{v}}\|^2 + \frac{\zeta_{p,t,\tau}}{2n}\|\boldsymbol{\Gamma}^{\frac{1}{2}}\boldsymbol{B}_{\boldsymbol{v}}^\perp\boldsymbol{r} + q\boldsymbol{\Gamma}^{\frac{1}{2}}\bar{\boldsymbol{v}}\|^2 + \tfrac{\lambda}{2}(q^2+\|\boldsymbol{r}\|^2)\\
&+ \frac{t_1}{2\tau_1\ell n}\|\mu_0\vartheta\boldsymbol{V}_1^\top\mathbf{1}_{\ell n} - \boldsymbol{V}_1^\top\widehat{\boldsymbol{y}} + \widetilde{\mu}_1 T_{p,1} q\boldsymbol{V}_1^\top\widehat{\boldsymbol{s}}\|^2 + \frac{t_2}{2\tau_2\ell n}\|\mu_0\vartheta\boldsymbol{V}_2^\top\mathbf{1}_{\ell n} - \boldsymbol{V}_2^\top\widehat{\boldsymbol{y}} + \widetilde{\mu}_1 T_{p,1} q\boldsymbol{V}_2^\top\widehat{\boldsymbol{s}}\|^2,
\end{aligned} \tag{45}$$

where $C_q > 0$ and $C_r > 0$ are two positive constants selected to satisfy the asymptotic result in Lemma 2. Here, we also drop terms that converge in probability to zero. One way to justify this step is using similar analysis as in Lemma 5. Note that the convexity results in Lemma 4 are still satisfied by the formulation in (45). Specifically, the cost function in (45) is jointly strongly convex in the minimization variables and jointly strongly concave in the maximization variables.

Now, it remains to solve over the optimization vector $\boldsymbol{r} \in \mathbb{R}^{k-1}$. To solve over $\boldsymbol{r}$, we interchange the minimization over $\boldsymbol{r}$ and the maximization over t_1 and t_2. This step is justified using the result in (Sion, 1958). The cost function of the optimization vector $\boldsymbol{r}$ can then be expressed as follows

$$\begin{aligned} g(\boldsymbol{r}) = &\frac{\sqrt{\ell} t_1}{\sqrt{n\ell}} \boldsymbol{g}_1^\top \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{B}_{\boldsymbol{v}}^\perp \boldsymbol{r} + \frac{\sqrt{t_1^2 + t_2^2}}{\ell \sqrt{n}} \boldsymbol{g}_2^\top \boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{B}_{\boldsymbol{v}}^\perp \boldsymbol{r} + \frac{t_1 \|\boldsymbol{h}_1\|^2}{2\tau_1 n} \|\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{B}_{\boldsymbol{v}}^\perp \boldsymbol{r} + q \boldsymbol{\Sigma}^{\frac{1}{2}} \bar{\boldsymbol{v}}\|^2 + \frac{\lambda}{2} \|\boldsymbol{r}\|^2 \\ &+ \frac{\zeta_{p,t,\tau}}{2n} \|\boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{B}_{\boldsymbol{v}}^\perp \boldsymbol{r} + q \boldsymbol{\Gamma}^{\frac{1}{2}} \bar{\boldsymbol{v}}\|^2, \end{aligned} \tag{46}$$

where we ignore the terms independent of $\boldsymbol{r}$. Note that the function $g(\cdot)$ is convex and smooth. Before solving the minimization problem of the function $g(\cdot)$, we define the matrix $\boldsymbol{G} \in \mathbb{R}^{k \times k}$ and the vectors $\boldsymbol{f} \in \mathbb{R}^{k-1}$ and $\boldsymbol{z} \in \mathbb{R}^{k-1}$ as follows

$$\begin{cases} \boldsymbol{G} = \frac{t_1 \|\boldsymbol{h}_1\|^2}{\tau_1 n} \boldsymbol{\Sigma} + \frac{\zeta_{p,t,\tau}}{n} \boldsymbol{\Gamma} \\ \boldsymbol{f} = \frac{t_1 \|\boldsymbol{h}_1\|^2}{\tau_1 n} \bar{\boldsymbol{B}}_{\boldsymbol{v}}^\perp \boldsymbol{\Sigma} \bar{\boldsymbol{v}} + \frac{\zeta_{p,t,\tau}}{n} \bar{\boldsymbol{B}}_{\boldsymbol{v}}^\perp \boldsymbol{\Gamma} \bar{\boldsymbol{v}} \\ \boldsymbol{z} = \frac{\sqrt{\ell} t_1}{\sqrt{n\ell}} \bar{\boldsymbol{B}}_{\boldsymbol{v}}^\perp \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{g}_1 + \frac{\sqrt{t_1^2 + t_2^2}}{\ell \sqrt{n}} \bar{\boldsymbol{B}}_{\boldsymbol{v}}^\perp \boldsymbol{\Gamma}^{\frac{1}{2}} \boldsymbol{g}_2, \end{cases} \tag{47}$$

where $\bar{\boldsymbol{B}}_{\boldsymbol{v}}^\perp = (\boldsymbol{B}_{\boldsymbol{v}}^\perp)^\top$. After computing the derivative of the function $g(\cdot)$ and setting it to zero, the optimal solution of the unconstrained version of minimizing the function $g(\cdot)$ can be expressed as follows

$$\widetilde{\boldsymbol{r}}^\star = -\left[\bar{\boldsymbol{B}}_{\boldsymbol{v}}^\perp \boldsymbol{G} \boldsymbol{B}_{\boldsymbol{v}}^\perp + \lambda \boldsymbol{I}_{k-1}\right]^{-1} [q\boldsymbol{f} + \boldsymbol{z}]. \tag{48}$$

Similar to the analysis in Lemmas 2, 3 and 6, one can show that the norm of the optimal vector $\widetilde{\boldsymbol{r}}^\star$ is bounded. This means that $\widetilde{\boldsymbol{r}}^\star$ is an optimal solution of the formulation in (45). Then, the optimal loss function can be expressed as follows

$$\begin{aligned} g^\star = &-\frac{q^2}{2} \boldsymbol{f}^\top \left[\bar{\boldsymbol{B}}_{\boldsymbol{v}}^\perp \boldsymbol{G} \boldsymbol{B}_{\boldsymbol{v}}^\perp + \lambda \boldsymbol{I}_{k-1}\right]^{-1} \boldsymbol{f} + \frac{q^2}{2} \bar{\boldsymbol{v}}^\top \boldsymbol{G} \bar{\boldsymbol{v}} \\ &- \frac{1}{2} \boldsymbol{z}^\top \left[\bar{\boldsymbol{B}}_{\boldsymbol{v}}^\perp \boldsymbol{G} \boldsymbol{B}_{\boldsymbol{v}}^\perp + \lambda \boldsymbol{I}_{k-1}\right]^{-1} \boldsymbol{z}. \end{aligned} \tag{49}$$

Based on the SVD decomposition of the matrix $\boldsymbol{M}$, it can be checked that the last term in the multivariate AO formulation given in (45) is zero. Then, the formulation given in (45) can be expressed as follows

$$\begin{aligned} \min_{|q| \leq C_q} \max_{\substack{0 \leq t_1 \leq C_{t_1} \\ 0 \leq t_2 \leq C_{t_2}}} \min_{\substack{c_{\tau_1} \leq \tau_1 \leq C_{\tau_1} \\ c_{\tau_2} \leq \tau_2 \leq C_{\tau_2}}} &\frac{\tau_1 t_1}{2\ell} - \frac{t_1^2 + t_2^2}{2\ell} + \frac{\tau_2 t_2}{2\ell} + \frac{\lambda + V_{p,2}(\boldsymbol{t}, \boldsymbol{\tau}) - V_{p,3}(\boldsymbol{t}, \boldsymbol{\tau})}{2} q^2 \\ &+ \frac{t_1}{2\tau_1 n} \|\mu_0 \vartheta \boldsymbol{1}_n - \boldsymbol{y} + \widetilde{\mu}_1 T_{p,1} q \boldsymbol{s}\|^2 - \frac{1}{2} V_{p,4}(\boldsymbol{t}, \boldsymbol{\tau}), \end{aligned} \tag{50}$$

where $\boldsymbol{t} = [t_1, t_2]^\top$ and $\boldsymbol{\tau} = [\tau_1, \tau_2]^\top$. Here, $\boldsymbol{1}_n$ denotes the vector of all one with size n and the functions $V_{p,2}(\cdot, \cdot)$, $V_{p,3}(\cdot, \cdot)$ and $V_{p,4}(\cdot, \cdot)$ depend on the optimization variables and are given by

$$\begin{cases} V_{p,2}(\boldsymbol{t}, \boldsymbol{\tau}) = \bar{\boldsymbol{v}}^\top \boldsymbol{G} \bar{\boldsymbol{v}} \\ V_{p,3}(\boldsymbol{t}, \boldsymbol{\tau}) = \boldsymbol{f}^\top \left[\bar{\boldsymbol{B}}_{\boldsymbol{v}}^\perp \boldsymbol{G} \boldsymbol{B}_{\boldsymbol{v}}^\perp + \lambda \boldsymbol{I}_{k-1}\right]^{-1} \boldsymbol{f} \\ V_{p,4}(\boldsymbol{t}, \boldsymbol{\tau}) = \boldsymbol{z}^\top \left[\bar{\boldsymbol{B}}_{\boldsymbol{v}}^\perp \boldsymbol{G} \boldsymbol{B}_{\boldsymbol{v}}^\perp + \lambda \boldsymbol{I}_{k-1}\right]^{-1} \boldsymbol{z}. \end{cases} \tag{51}$$

Note that we simplified the multivariate AO formulation given in (35) to a scalar optimization problem as given in (50). Then, it remains to study the asymptotic properties of the scalar formulation in (50). We refer to this problem as the *scalar formulation*.

8.2.3. Asymptotic Analysis of the Scalar Formulation

In this part, we study the asymptotic properties of the scalar formulation in (50) corresponding to the multivariate AO problem. Based on Assumption 5 and the result in Proposition 3 in (Debbah et al., 2003), the random variable $T_{p,1}$ converges pointwisely in probability to the scalar T_1 defined as follows

$$T_{p,1} = \sqrt{\boldsymbol{\xi}^\top \boldsymbol{F}\boldsymbol{F}^\top \boldsymbol{\xi}} \xrightarrow{p\to\infty} T_1 = \sqrt{\delta \mathbb{E}_\kappa[\kappa]}, \tag{52}$$

where the expectations are over the probability distribution $\mathbb{P}_\kappa(\cdot)$ defined in Assumption 5. Furthermore, the random function $V_{p,2}(\cdot,\cdot)$ can be expressed as follows

$$\begin{aligned} V_{p,2}(\boldsymbol{t},\boldsymbol{\tau}) &= \frac{1}{T_{p,1}^2}\boldsymbol{\xi}^\top \boldsymbol{F}\Big(\frac{t_1\|\boldsymbol{h}_1\|^2}{\tau_1 n}\boldsymbol{\Sigma} + \frac{\zeta_{p,t,\tau}}{n}\boldsymbol{\Gamma}\Big)\boldsymbol{F}^\top\boldsymbol{\xi} \\ &= \frac{1}{T_{p,1}^2}\boldsymbol{\xi}^\top \boldsymbol{F}\Big[\Big(\frac{t_1\|\boldsymbol{h}_1\|^2\widetilde{\mu}_1^2}{\tau_1 n} + \frac{\zeta_{p,t,\tau}\widehat{\mu}_1^2}{n}\Big)\boldsymbol{F}^\top\boldsymbol{F} - \frac{t_1\|\boldsymbol{h}_1\|^2\widetilde{\mu}_1^2}{\tau_1 n} \\ &\times \boldsymbol{F}^\top\boldsymbol{\xi}\boldsymbol{\xi}^\top\boldsymbol{F} + \Big(\frac{t_1\|\boldsymbol{h}_1\|^2\mu_2^2}{\tau_1 n} + \frac{\zeta_{p,t,\tau}\mu_3^2}{n}\Big)\boldsymbol{I}_k\Big]\boldsymbol{F}^\top\boldsymbol{\xi}, \end{aligned} \tag{53}$$

where $\boldsymbol{t} = [t_1, t_2]^\top$ and $\boldsymbol{\tau} = [\tau_1, \tau_2]^\top$. Then, using the theoretical results in (Debbah et al., 2003) and based on Assumption 5, the random function $V_{p,2}(\cdot,\cdot)$ converges pointwisely in probability as follows

$$\begin{aligned} V_{p,2}(\boldsymbol{t},\boldsymbol{\tau}) \xrightarrow{p\to+\infty} V_2(\boldsymbol{t},\boldsymbol{\tau}) &= \frac{\delta}{T_1^2}\Big(\frac{t_1\widetilde{\mu}_1^2}{\tau_1} + \zeta_{t,\tau}\widehat{\mu}_1^2\Big)\mathbb{E}_\kappa[\kappa^2] - \delta^2\frac{t_1\widetilde{\mu}_1^2}{\tau_1}\mathbb{E}_\kappa[\kappa]^2 \\ &+ \Big(\frac{t_1\mu_2^2}{\tau_1} + \zeta_{t,\tau}\mu_3^2\Big)\delta\mathbb{E}_\kappa[\kappa], \end{aligned} \tag{54}$$

where $\zeta_{t,\tau} = t_1/(\ell\tau_1) + t_2(\ell-1)/(\ell\tau_2)$ and the expectations are over the probability distribution $\mathbb{P}_\kappa(\cdot)$ defined in Assumption 5. The theoretical results in Proposition 3 in (Debbah et al., 2003) also show that the random function $V_{p,4}(\cdot,\cdot)$ satisfies the following asymptotic result

$$V_{p,4}(\boldsymbol{t},\boldsymbol{\tau}) - \widehat{V}_{p,4}(\boldsymbol{t},\boldsymbol{\tau}) \xrightarrow{p\to+\infty} 0, \tag{55}$$

where the random function $\widehat{V}_{p,4}(\cdot,\cdot)$ is defined as follows

$$\begin{aligned} \widehat{V}_{p,4}(\boldsymbol{t},\boldsymbol{\tau}) &= \frac{t_1^2}{n\ell}\mathrm{Tr}\Big[\widehat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\Big[\widehat{\boldsymbol{G}} + \lambda\boldsymbol{I}_{k-1}\Big]^{-1}\widehat{\boldsymbol{\Sigma}}^{\frac{1}{2}}\Big] \\ &+ \frac{t_1^2+t_2^2}{\ell^2 n}\mathrm{Tr}\Big[\boldsymbol{\Gamma}^{\frac{1}{2}}\Big[\widehat{\boldsymbol{G}} + \lambda\boldsymbol{I}_{k-1}\Big]^{-1}\boldsymbol{\Gamma}^{\frac{1}{2}}\Big]. \end{aligned} \tag{56}$$

Here, Tr[.] represents the trace. Additionally, the matrix $\widehat{\boldsymbol{\Sigma}}$ is given as $\widehat{\boldsymbol{\Sigma}} = \widetilde{\mu}_1^2\boldsymbol{F}^\top\boldsymbol{F} + \mu_2^2\boldsymbol{I}_k$ and the matrix $\widehat{\boldsymbol{G}}$ has the following expression $\widehat{\boldsymbol{G}} = t_1\|\boldsymbol{h}_1\|^2/(\tau_1 n)\widehat{\boldsymbol{\Sigma}} + \zeta_{p,t,\tau}/n\boldsymbol{\Gamma}$. Using again Proposition 3 (Debbah et al., 2003) and Assumption 5, we can also see that the random function $V_{p,4}(\cdot,\cdot)$ converges in probability to the function $V_4(\cdot,\cdot)$ defined as follows

$$\begin{aligned} V_4(\boldsymbol{t},\boldsymbol{\tau}) &= \frac{\eta t_1^2}{\ell}\mathbb{E}_\kappa\Big[\frac{\widetilde{\mu}_1^2\kappa + \mu_2^2}{\frac{t_1}{\tau_1}(\widetilde{\mu}_1^2\kappa + \mu_2^2) + \zeta_{t,\tau}(\widehat{\mu}_1^2\kappa + \mu_3^2) + \lambda}\Big] \\ &+ \frac{\eta(t_1^2+t_2^2)}{\ell^2}\mathbb{E}_\kappa\Big[\frac{\widehat{\mu}_1^2\kappa + \mu_3^2}{\frac{t_1}{\tau_1}(\widetilde{\mu}_1^2\kappa + \mu_2^2) + \zeta_{t,\tau}(\widehat{\mu}_1^2\kappa + \mu_3^2) + \lambda}\Big]. \end{aligned}$$

Now, it remains to study the asymptotic properties of the random function $V_{p,3}(\cdot,\cdot)$. Based on the block matrix inversion lemma, it can be checked that the random function $V_{p,3}(\cdot,\cdot)$ satisfies the following

$$V_{p,3}(\boldsymbol{t},\boldsymbol{\tau}) = V_{p,2}(\boldsymbol{t},\boldsymbol{\tau}) + \lambda - \frac{1}{T_{p,2}(\boldsymbol{t},\boldsymbol{\tau})}. \tag{57}$$

Here, the random function $T_{p,2}(\cdot,\cdot)$ is defined as follows

$$T_{p,2}(\boldsymbol{t},\boldsymbol{\tau}) = \bar{\boldsymbol{v}}^{\top}\left[\boldsymbol{G}+\lambda \boldsymbol{I}_k\right]^{-1}\bar{\boldsymbol{v}}. \tag{58}$$

Using the matrix inversion lemma, it can be checked that the random function $T_{p,2}(\cdot,\cdot)$ converges in probability to the function $T_{2,\lambda}(\cdot,\cdot)$ defined as follows

$$T_{2,\lambda}(\boldsymbol{t},\boldsymbol{\tau}) = \frac{\delta}{T_1^2}\mathbb{E}\Big[\frac{\kappa}{g_{\kappa,\lambda}(t,\tau)}\Big]\Big/\Big(1-\frac{\widetilde{\mu}_1^2 t_1 \delta}{\tau_1}\mathbb{E}\Big[\frac{\kappa}{g_{\kappa,\lambda}(t,\tau)}\Big]\Big),$$

where the function $g_{\kappa,\lambda}(\cdot,\cdot)$ is defined in Section 4. Additionally, using the weak law of large numbers (WLLN), we have the following convergence property

$$\frac{1}{n}\|\mu_0\vartheta\mathbf{1}_n - \boldsymbol{y} + \widetilde{\mu}_1 T_{p,1} q\boldsymbol{s}\|^2 \xrightarrow{p\to+\infty} \mu_0^2\vartheta^2 + \gamma_1 + \widetilde{\mu}_1^2 T_1^2 q^2 - 2\widetilde{\mu}_1 T_1 q\gamma_2 - 2\mu_0\vartheta\gamma_3. \tag{59}$$

Here, γ_1, γ_2 and γ_3 depend on the data distribution and are defined as $\gamma_1 = \mathbb{E}[y^2]$, $\gamma_2 = \mathbb{E}[ys]$, $\gamma_3 = \mathbb{E}[y]$, where $y = \varphi(s)$, and s is a standard Gaussian random variable. This proves that the cost function of the following deterministic problem

$$\begin{aligned}\min_{|q|\le C_q}\max_{\substack{0\le t_1\le C_{t_1}\\ 0\le t_2\le C_{t_2}}}\min_{\substack{c_{\tau_1}\le\tau_1\le C_{\tau_1}\\ c_{\tau_2}\le\tau_2\le C_{\tau_2}}} & \frac{\tau_1 t_1+\tau_2 t_2}{2\ell} - \frac{t_1^2+t_2^2}{2\ell} + \frac{q^2}{2T_{2,\lambda}(\boldsymbol{t},\boldsymbol{\tau})} - \frac{\eta T_{3,\lambda}(\boldsymbol{t},\boldsymbol{\tau})}{2}\\ & + \frac{t_1}{2\tau_1}\Big(\gamma_1 - 2\widetilde{\mu}_1 T_1 q\gamma_2 + \widetilde{\mu}_1^2 T_1^2 q^2 + \mu_0^2\vartheta^2 - 2\mu_0\vartheta\gamma_3\Big),\end{aligned} \tag{60}$$

is the converging limit of the cost function of the scalar formulation in (50), where the function $T_{3,\lambda}(\cdot,\cdot)$ is given by $V_4(\boldsymbol{t},\boldsymbol{\tau})/\eta$. Before continuing our analysis, we summarize convexity properties of the cost function of (60) in the following lemma.

Lemma 7 (Strong–convexity of (60))**.** *Define $\widehat{f}$ as the cost function of the problem in (60) defined in the feasibility set. Then, $\widehat{f}$ is jointly strongly convex in the variables $(q,\vartheta,\tau_1,\tau_2)$ for fixed feasible (t_1,t_2). Moreover, it is jointly strongly concave in the variables (t_1,t_2) for fixed feasible $(q,\vartheta,\tau_1,\tau_2)$.*

This result can be proved by observing that the strong convexity parameters in Lemma 4 are independent of p and that the operations performed after Lemma 4 preserve the convexity properties. Another property of the scalar formulation is that its set of optimal solutions concentrates around the set of optimal solutions of the formulation in (60) as summarized in the following lemma.

Lemma 8 (Consistency of the Scalar Formulation)**.** *Define $\tau_{p,1}^\star$, $\tau_{p,2}^\star$, $t_{p,1}^\star$, $t_{p,2}^\star$ and $q_p^\star$ as the optimal solutions of the scalar formulation given in (50). Additionally, define $\tau_1^\star$, $\tau_2^\star$, $t_1^\star$, $t_2^\star$ and $q^\star$ as the optimal solutions of the deterministic optimization problem given in (60). Therefore, the following convergence in probability holds*

$$\begin{aligned}\tau_{p,1}^\star \xrightarrow{p\to+\infty} \tau_1^\star,\ \tau_{p,2}^\star \xrightarrow{p\to+\infty} \tau_2^\star,\ t_{p,1}^\star \xrightarrow{p\to+\infty} t_1^\star,\\ t_{p,2}^\star \xrightarrow{p\to+\infty} t_2^\star,\ q_p^\star \xrightarrow{p\to+\infty} q^\star.\end{aligned} \tag{61}$$

Moreover, define $\vartheta_p^\star$ and $\vartheta^\star$ as the optimal solutions of the minimization problems of (50) and (60) over ϑ in the feasibility set defined in (26). Then, we also have the following convergence in probability

$$\vartheta_p^\star \xrightarrow{p\to+\infty} \vartheta^\star. \tag{62}$$

The convergence result in (61) follows using Theorem 2.1 in (Newey & McFadden, 1994). We can see that all the assumptions in Theorem 2.1 in (Newey & McFadden, 1994) are satisfied by the formulations in (50) and (60). Moreover, the result in (62) follows using Proposition 2 in (Dhifallah & Lu, 2020). The detailed proof is omitted since it follows similar ideas as in Proposition 4 and Proposition 5 in (Dhifallah & Lu, 2020). Based on (Sion, 1958), we can further simplify the formulation in (60) by solving the minimization problem over the variables q and ϑ. Note that the optimal $\vartheta^\star$ satisfies $\vartheta^\star = 0$ if $\mu_0 = 0$ and $\vartheta^\star = \gamma_3/\mu_0$ otherwise. Furthermore, the optimal q denoted by $q_{t,\tau}^\star$ can be expressed as follows

$$q_{t,\tau}^\star = \frac{\gamma_2 t_1 \widetilde{\mu}_1 T_1 T_{2,\lambda}(\boldsymbol{t},\boldsymbol{\tau})}{\tau_1 + t_1\widetilde{\mu}_1^2 T_1^2 T_{2,\lambda}(\boldsymbol{t},\boldsymbol{\tau})}. \tag{63}$$

Observe that the optimal solutions, $\vartheta^\star$ and $q_{t,\tau}^\star$, satisfy the boundedness constraints. Moreover, note that our results are valid for any bounds that satisfy the results in Lemmas 2, 3 and 6. Now that we obtained the asymptotic scalar optimization problem, it remains to study the asymptotic behavior of the training and generalization errors.

8.2.4. Asymptotic Analysis of the Training and Generalization Errors

First, the generalization error is given by

$$\mathcal{E}_{\text{test}} = \frac{1}{4^v}\mathbb{E}\left[\left(\varphi(\boldsymbol{a}_{\text{new}}^\top\boldsymbol{\xi}) - \widehat{\varphi}(\widehat{\boldsymbol{w}}^\top\sigma(\boldsymbol{F}^\top\boldsymbol{a}_{\text{new}}))\right)^2\right],$$

where $\boldsymbol{a}_{\text{new}}$ is an unseen data sample and $\widehat{\boldsymbol{w}}$ is the optimal solution of the noisy formulation. Based on the uniform Gaussian equivalence theorem (uGET), observed and proved in many earlier papers (Hu & Lu, 2020; Montanari et al., 2019; Gerace et al., 2020; Goldt et al., 2020b; Dhifallah & Lu, 2020), the asymptotic properties of the generalization error are equivalent to the asymptotic properties of $\overline{\mathcal{E}}_{\text{test}}$ defined as follows

$$\overline{\mathcal{E}}_{\text{test}} = \frac{1}{4^v}\mathbb{E}\Big[\Big(\varphi(\boldsymbol{a}_{\text{new}}^\top\boldsymbol{\xi}) - \widehat{\varphi}(\mu_{0s}\widehat{\vartheta}_p + \mu_{1s}\widehat{\boldsymbol{w}}^\top\boldsymbol{F}^\top\boldsymbol{a}_{\text{new}} + \mu_{2s}\widehat{\boldsymbol{w}}^\top\boldsymbol{z})\Big)^2\Big].$$

Here, $\widehat{\boldsymbol{w}}$ and $\widehat{\vartheta}_p$ are the optimal solutions of our primary formulation given in (26). The expectation is taken over the distribution of the random vector $\boldsymbol{a}_{\text{new}}$, the random vector $\boldsymbol{z}$ and the possibly random functions $\varphi(\cdot)$ and $\widehat{\varphi}(\cdot)$, where $\boldsymbol{z}$ is independent of $\boldsymbol{a}_{\text{new}}$ and drawn from a standard Gaussian distribution. Moreover, the constants μ_{0s}, μ_{1s} and μ_{2s} are defined as $\mu_{0s} = \mathbb{E}[\sigma(z)]$, $\mu_{1s} = \mathbb{E}[z\sigma(z)]$ and $\mu_{2s}^2 = \mathbb{E}[\sigma(z)^2] - \mu_{0s}^2 - \mu_{1s}^2$, where z is a standard Gaussian random vector. Now, consider the following two random variables

$$g_1 = \boldsymbol{a}_{\text{new}}^\top\boldsymbol{\xi}, \text{ and } g_2 = \mu_{0s}\widehat{\vartheta}_p + \mu_{1s}\widehat{\boldsymbol{w}}^\top\boldsymbol{F}^\top\boldsymbol{a}_{\text{new}} + \mu_{2s}\widehat{\boldsymbol{w}}^\top\boldsymbol{z}.$$

Given the optimal solutions $\widehat{\boldsymbol{w}}$ and $\widehat{\vartheta}_p$, the random variables g_1 and g_2 have a bivariate Gaussian distribution with mean vector $[0, \mu_{0s}\widehat{\vartheta}_p]^\top$ and covariance matrix given by

$$\boldsymbol{C}_p = \begin{bmatrix} \|\boldsymbol{\xi}\|^2 & \mu_{1s}\boldsymbol{\xi}^\top\boldsymbol{F}\widehat{\boldsymbol{w}} \\ \mu_{1s}\boldsymbol{\xi}^\top\boldsymbol{F}\widehat{\boldsymbol{w}} & \mu_{1s}^2\|\boldsymbol{F}\widehat{\boldsymbol{w}}\|^2 + \mu_{2s}^2\|\widehat{\boldsymbol{w}}\|^2 \end{bmatrix}.$$

Define the random variables $\widehat{q}_p^\star$, $\widehat{\beta}_p^\star$ and $\widehat{r}_p^\star$ as follows

$$\widehat{q}_p^\star = \bar{\boldsymbol{v}}^\top\widehat{\boldsymbol{w}},\ \widehat{\beta}_p^\star = \|\boldsymbol{F}\widehat{\boldsymbol{w}}\|^2, \text{ and } \widehat{r}_p^\star = \|\widehat{\boldsymbol{w}}\|^2, \tag{64}$$

where $\bar{\boldsymbol{v}} = \boldsymbol{v}/\|\boldsymbol{v}\|$ and the vector $\boldsymbol{v}$ is defined as $\boldsymbol{v} = \boldsymbol{F}^\top\boldsymbol{\xi}$. Then, the covariance matrix $\boldsymbol{C}_p$ can be expressed as follows

$$\boldsymbol{C}_p = \begin{bmatrix} 1 & \mu_{1s}T_{p,1}\widehat{q}_p^\star \\ \mu_{1s}T_{p,1}\widehat{q}_p^\star & \mu_{1s}^2\widehat{\beta}_p^\star + \mu_{2s}^2\widehat{r}_p^\star \end{bmatrix}. \tag{65}$$

Hence, to study the asymptotic properties of the generalization error, it suffices to study the asymptotic properties of $\widehat{\vartheta}_p^\star$, $\widehat{q}_p^\star$, $\widehat{\beta}_p^\star$ and $\widehat{r}_p^\star$. The following lemma summarizes the asymptotic properties of our primal formulation given in (26).

Lemma 9 (Primal Consistency)**.** *The random variables $\widehat{\vartheta}_p^\star$, $\widehat{q}_p^\star$, $\widehat{\beta}_p^\star$ and $\widehat{r}_p^\star$ converge in probability as follows*

$$\begin{gathered} \widehat{q}_p^\star \xrightarrow{p\to+\infty} q^\star,\ \widehat{\vartheta}_p^\star \xrightarrow{p\to+\infty} \vartheta^\star, \widehat{\beta}_p^\star \xrightarrow{p\to+\infty} \beta^\star \\ \widehat{r}_p^\star \xrightarrow{p\to+\infty} r^\star = (q^\star)^2 + h'(\lambda), \end{gathered} \tag{66}$$

where $q^\star$ and $\vartheta^\star$ are optimal solutions of the deterministic scalar formulation in (60). Moreover, the function $h(\cdot)$ and $\beta^\star$ are defined in Theorem 1.

Proof. Note that the analysis in Section 8.2.2 shows that the scalar formulation given in (50) is a simplified version of the multivariate AO formulation given in (35). Define the random variable $\widetilde{\vartheta}_p^\star$ as the optimal solution of the minimization of the problem (35) over ϑ in the feasibility set defined in (26). Moreover, define the random variables $\widetilde{q}_p^\star$, $\widetilde{\beta}_p^\star$ and $\widetilde{r}_p^\star$ as follows

$$\widetilde{q}_p^\star = \bar{\boldsymbol{v}}^\top\widetilde{\boldsymbol{w}},\ \widetilde{\beta}_p^\star = \|\boldsymbol{F}\widetilde{\boldsymbol{w}}\|^2, \text{ and } \widetilde{r}_p^\star = \|\widetilde{\boldsymbol{w}}\|^2, \tag{67}$$

where $\widetilde{\boldsymbol{w}}$ is the optimal solution of the multivariate AO formulation given in (35). Based on the decomposition in (44), note that $\widetilde{\beta}_p^\star$ satisfies the following expression

$$\begin{aligned}\widetilde{\beta}_p^\star = \|\boldsymbol{F}\widetilde{\boldsymbol{w}}\|^2 = (\widetilde{q}_p^\star)^2 \bar{\boldsymbol{v}}^\top \boldsymbol{F}^\top \boldsymbol{F}\bar{\boldsymbol{v}} + 2\widetilde{q}_p^\star \bar{\boldsymbol{v}}^\top \boldsymbol{F}^\top \boldsymbol{F}\boldsymbol{B}_{\boldsymbol{v}}^\perp \widetilde{\boldsymbol{r}}^\star \\ + (\widetilde{\boldsymbol{r}}^\star)^\top \bar{\boldsymbol{B}}_{\boldsymbol{v}}^\perp \boldsymbol{F}^\top \boldsymbol{F}\boldsymbol{B}_{\boldsymbol{v}}^\perp \widetilde{\boldsymbol{r}}^\star,\end{aligned} \tag{68}$$

where $\widetilde{\boldsymbol{r}}^\star$ is defined in (48) and is the optimal solution of minimizing the function $g(\cdot)$ introduced in (46). Substituting the expression of $\widetilde{\boldsymbol{r}}^\star$ given in (48), performing the same analysis as in Section 8.2.3 and using the convergence result in (61), it can be shown that the random quantity $\widetilde{\beta}_p^\star$ converges in probability to $\beta^\star$ defined in (20). Additionally, observe that $\widetilde{r}_p^\star$ can be expressed as follows

$$\widetilde{r}_p^\star = \|\widetilde{\boldsymbol{w}}\|^2 = (\widetilde{q}_p^\star)^2 + \|\widetilde{\boldsymbol{r}}^\star\|^2. \tag{69}$$

Define the function $h_p : \lambda \to -(\widetilde{q}_p^\star)^2 V_{p,3}(\boldsymbol{t}_p^\star, \boldsymbol{\tau}_p^\star) - V_{p,4}(\boldsymbol{t}_p^\star, \boldsymbol{\tau}_p^\star)$, where the random functions $V_{p,3}(\cdot,\cdot)$ and $V_{p,4}(\cdot,\cdot)$ are defined in (51) and where $\boldsymbol{t}_p^\star = [t_{p,1}^\star, t_{p,2}^\star]^\top$ and $\boldsymbol{\tau}_p^\star = [\tau_{p,1}^\star, \tau_{p,2}^\star]^\top$. Here, $\{t_{p,1}^\star, t_{p,2}^\star, \tau_{p,1}^\star, \tau_{p,2}^\star\}$ are defined in Lemma 8. Given the expression of $\widetilde{\boldsymbol{r}}^\star$ in (48), we can see that $\widetilde{r}_p^\star$ can be expressed as follows

$$\widetilde{r}_p^\star = (\widetilde{q}_p^\star)^2 + h_p'(\lambda), \tag{70}$$

where the optimal solutions are treated as constants independent of λ. Performing the same analysis as in Section 8.2.3 and using the convergence result in (61), it can be shown that the random quantity $\widetilde{r}_p^\star$ converges in probability as follows

$$\widetilde{r}_p^\star \xrightarrow{p\to+\infty} r^\star = (q^\star)^2 + h'(\lambda), \tag{71}$$

where $q^\star$ is the optimal solution of (60) and the function $h(\cdot)$ is defined in Theorem 1. Given that the scalar formulation given in (50) is a simplified version of the multivariate AO formulation and based on Lemma 8, we obtain the following asymptotic properties

$$\widetilde{q}_p^\star \xrightarrow{p\to+\infty} q^\star, \ \widetilde{\vartheta}_p^\star \xrightarrow{p\to+\infty} \vartheta^\star, \tag{72}$$

where $q^\star$ and $\vartheta^\star$ are the optimal solutions of the deterministic scalar formulation in (60). Following a similar analysis as in (Thrampoulidis et al., 2018), we can show that the assumptions in Theorem 3 are all satisfied. The main idea is to define the set $\mathcal{S}_{p,\epsilon}$ introduced in Theorem 3 as

$$\{\boldsymbol{w} : |\|\boldsymbol{F}\boldsymbol{w}\|^2 - \beta^\star| < \epsilon\}, \text{ and } \{\boldsymbol{w} : |\|\boldsymbol{w}\|^2 - r^\star| < \epsilon\}. \tag{73}$$

Then, use the strong convexity properties of the formulation in (45) to prove that the assumptions in Theorem 3 are satisfied. This means that $\widehat{\vartheta}_p^\star$, $\widehat{q}_p^\star$, $\widehat{\beta}_p^\star$ and $\widehat{r}_p^\star$ defined in (64) concentrates around the same values as $\widetilde{\vartheta}_p^\star$, $\widetilde{q}_p^\star$, $\widetilde{\beta}_p^\star$ and $\widetilde{r}_p^\star$ defined in (67). □

Now, to show the convergence of the generalization error in Theorem 1, it suffices to show that $\overline{\mathcal{E}}_{\text{test}}$ is a continuous function in $\widehat{\vartheta}_p^\star$, $\widehat{q}_p^\star$, $\widehat{\beta}_p^\star$ and $\widehat{r}_p^\star$. Based on Assumption 3, the functions $\varphi(\cdot)$ and $\widehat{\varphi}(\cdot)$ are square integrable over Gaussian distributions. Moreover, the optimal solutions $\widehat{\vartheta}_p^\star$, $\widehat{q}_p^\star$, $\widehat{\beta}_p^\star$ and $\widehat{r}_p^\star$ are bounded. Based on Assumption 3 and the continuity under integral sign property (Schilling, 2005), the continuity of $\overline{\mathcal{E}}_{\text{test}}$ follows. These properties lead to the convergence result given in (19) in Theorem 1. Based on the analysis in Lemma 9 and Theorem 3, the optimal cost value of the noisy formulation converges in probability to the optimal cost value of the deterministic formulation in (60). Combining this result with the asymptotic property stated in (71) shows the convergence of the training error stated in Theorem 1.

8.3. Large Number of Noise Injections

Note that the analysis in the previous parts studies the properties of the training and generalization errors corresponding to (3) when p grows to infinity. In this part, we study the properties of the noisy formulation when ℓ grows to infinity slower than the dimensions n, p and k. Note that in this regime and based on the analysis in the previous part, the noisy formulation converges to the deterministic formulation in (60). Then, the objective is to analyze the deterministic formulation in (60)

when ℓ grows to infinity. Here, we note that the bounds on the feasibility sets of the deterministic formulation in (60) depends on ℓ as follows

$$C_{t_1} = \sqrt{\ell}\overline{C}_{t_1},\ C_{t_2} = \sqrt{\ell}\overline{C}_{t_2},\ c_{\tau_1} = \sqrt{\ell}\overline{c}_{\tau_1},$$
$$c_{\tau_2} = \sqrt{\ell}\overline{c}_{\tau_2},\ C_{\tau_1} = \sqrt{\ell}\overline{C}_{\tau_1},\ C_{\tau_2} = \sqrt{\ell}\overline{C}_{\tau_2}. \tag{74}$$

We start our analysis by performing the change of variable $\tau_1 = \tau_1/\sqrt{\ell}$, $\tau_2 = \tau_2/\sqrt{\ell}$, $t_1 = t_1/\sqrt{\ell}$ and $t_2 = t_2/\sqrt{\ell}$. This means that the deterministic scalar optimization problem given in (60) can be expressed as follows

$$\begin{aligned}\max_{\substack{0\leq t_1\leq \overline{C}_{t_1}\\ 0\leq t_2\leq \overline{C}_{t_2}}}\ \min_{\substack{\overline{c}_{\tau_1}\leq \tau_1\leq \overline{C}_{\tau_1}\\ \overline{c}_{\tau_2}\leq \tau_2\leq \overline{C}_{\tau_2}}}\ & \frac{\tau_1 t_1 + \tau_2 t_2}{2} - \frac{t_1^2+t_2^2}{2} + \frac{(q^\star_{t,\tau})^2}{2T_{2,\lambda}(\boldsymbol{t},\boldsymbol{\tau})} - \frac{\eta T_{3,\lambda}(\boldsymbol{t},\boldsymbol{\tau})}{2}\\ &+ \frac{t_1}{2\tau_1}\Big(\gamma_1 - 2\widetilde{\mu}_1 T_1 q^\star_{t,\tau}\gamma_2 + \widetilde{\mu}_1^2 T_1^2 (q^\star_{t,\tau})^2 + \mu_0^2(\vartheta^\star)^2 - 2\mu_0\vartheta^\star\gamma_3\Big),\end{aligned} \tag{75}$$

where $\boldsymbol{t} = [t_1, t_2]^\top$ and $\boldsymbol{\tau} = [\tau_1, \tau_2]^\top$. Here, the functions $T_{2,\lambda}(\cdot,\cdot)$ and $q^\star_{t,\tau}$ are the same as the ones provided in Section 4. Furthermore, the functions $T_{3,\lambda}(\cdot,\cdot)$ and $g_{\kappa,\lambda}(\cdot,\cdot)$ are given as follows

$$\begin{aligned}T_{3,\lambda}(\boldsymbol{t},\boldsymbol{\tau}) &= t_1^2\mathbb{E}\Big[\frac{\widetilde{\mu}_1^2\kappa+\mu_2^2}{g_{\kappa,\lambda}(\boldsymbol{t},\boldsymbol{\tau})}\Big] + \frac{t_1^2+t_2^2}{\ell}\mathbb{E}\Big[\frac{\widehat{\mu}_1^2\kappa+\mu_3^2}{g_{\kappa,\lambda}(\boldsymbol{t},\boldsymbol{\tau})}\Big],\\ g_{\kappa,\lambda}(\boldsymbol{t},\boldsymbol{\tau}) &= \frac{t_1}{\tau_1}\Big(\widetilde{\mu}_1^2\kappa+\mu_2^2\Big) + \Big(\frac{t_1}{\tau_1\ell} + \frac{t_2(\ell-1)}{\tau_2\ell}\Big)\\ &\quad\times\Big(\widehat{\mu}_1^2\kappa+\mu_3^2\Big) + \lambda.\end{aligned}$$

Now, we focus on analyzing the formulation in (75) when the number of noise injections grows to infinity. The following lemma summarizes our main technical results.

Lemma 10 (Large Number of Noise Injections)**.** *When ℓ grows to infinity, the asymptotic limit of the formulation in (75) is obtained by updating the functions $T_{3,\lambda}(\cdot,\cdot)$ and $g_{\kappa,\lambda}(\cdot,\cdot)$ as follows*

$$T_{3,\lambda}(\boldsymbol{t},\boldsymbol{\tau}) = t_1^2\mathbb{E}\Big[\frac{\widetilde{\mu}_1^2\kappa+\mu_2^2}{g_{\kappa,\lambda}(\boldsymbol{t},\boldsymbol{\tau})}\Big],\ g_{\kappa,\lambda}(\boldsymbol{t},\boldsymbol{\tau}) = \frac{t_1}{\tau_1}\Big(\widetilde{\mu}_1^2\kappa+\mu_2^2\Big) + \frac{t_2}{\tau_2}\Big(\widehat{\mu}_1^2\kappa+\mu_3^2\Big) + \lambda.$$

The convergence result in Lemma 10 follows using Theorem 2.1 in (Newey & McFadden, 1994). Specifically, we use the strong convexity property in Lemma 7. Also, we use the pointwise convergence of the cost functions based on Assumptions 4 and 5 and the dominated convergence theorem. This shows that all the assumptions in Theorem 2.1 in (Newey & McFadden, 1994) are satisfied by the formulation in (75) and its asymptotic formulation mentioned in Lemma 10. Next, we refer to the asymptotic limit obtained in Lemma 10 as the *asymptotic deterministic formulation*.

Performing a similar analysis as in Sections 8.2.1, 8.2.2 and 8.2.3, it can be checked that the asymptotic deterministic formulation obtained in Lemma 10 is the asymptotic limit of the following formulation

$$\min_{\boldsymbol{w}\in\mathbb{R}^k}\frac{1}{2n}\sum_{i=1}^{n}\Big(y_i - \boldsymbol{w}^\top\widehat{\sigma}(\boldsymbol{F}^\top\boldsymbol{a}_i)\Big)^2 + \tfrac{1}{2}\|\boldsymbol{R}^{\frac{1}{2}}\boldsymbol{w}\|^2 + \tfrac{\lambda}{2}\|\boldsymbol{w}\|^2. \tag{76}$$

Here, the regularization matrix $\boldsymbol{R}$ is defined as follows

$$\boldsymbol{R} = \widehat{\mu}_1^2\boldsymbol{F}^\top\boldsymbol{F} + \mu_3^2\boldsymbol{I}_k, \tag{77}$$

and the new activation function $\widehat{\sigma}(\cdot)$ satisfies the following properties

$$\begin{aligned}\mathbb{E}[\widehat{\sigma}(z)] = \mathbb{E}[\sigma(x_1)],\ \mathbb{E}[z\widehat{\sigma}(z)] = \mathbb{E}[z\sigma(x_1)]\\ \mathbb{E}[\widehat{\sigma}(z)^2] = \mathbb{E}[\sigma(x_1)\sigma(x_2)],\end{aligned} \tag{78}$$

where $x_1 = z + \Delta v_1$, $x_2 = z + \Delta v_2$ and z, v_1 and v_2 are independent standard Gaussian random variables. Now, note that the norm of any vector $\boldsymbol{x} \in \mathbb{R}^p$ can be expressed as follows

$$\begin{aligned}\frac{1}{2}\|\boldsymbol{x}\|^2 &= \max_{\boldsymbol{u}\in\mathbb{R}^p} -\frac{\|\boldsymbol{u}\|^2}{2} + \boldsymbol{u}^\top \boldsymbol{x} = \max_{t\geq 0} -\frac{t^2}{2} + t\|\boldsymbol{x}\| \\ &= \max_{t\geq 0}\inf_{\tau>0} -\frac{t^2}{2} + \frac{t\tau}{2} + \frac{t}{2\tau}\|\boldsymbol{x}\|^2. \end{aligned} \tag{79}$$

We can see that the optimal solution of the max-min problem in (79), denoted by $t^\star$ and $\tau^\star$, satisfies $t^\star = \tau^\star$. This trick can be used in the CGMT framework to show that the asymptotic limit of the formulation in (76) can also be expressed as follows

$$\begin{aligned}&\max_{0\leq t_1\leq \overline{C}_{t_1}} \min_{\bar{c}_{\tau_1}\leq \tau_1\leq \overline{C}_{\tau_1}} \frac{\tau_1 t_1}{2} - \frac{t_1^2}{2} + \frac{(q^\star_{t,\tau,\infty})^2}{2T_{2,\lambda,\infty}(t_1,\tau_1)} - \frac{\eta T_{3,\lambda,\infty}(t_1,\tau_1)}{2} \\ &+ \frac{t_1}{2\tau_1}\Big(\gamma_1 - 2\widetilde{\mu}_1 T_1 q^\star_{t,\tau,\infty}\gamma_2 + \widetilde{\mu}_1^2 T_1^2 (q^\star_{t,\tau,\infty})^2 + \mu_0^2(\vartheta^\star)^2 - 2\mu_0\vartheta^\star\gamma_3\Big),\end{aligned} \tag{80}$$

where the constant $\vartheta^\star$ satisfies $\vartheta^\star = 0$ if $\mu_0 = 0$ and $\vartheta^\star = \gamma_3/\mu_0$ otherwise. Moreover, $q^\star_{t,\tau,\infty}$ is defined as follows

$$q^\star_{t,\tau,\infty} = \frac{\gamma_2 t_1 \widetilde{\mu}_1 T_1 T_{2,\lambda,\infty}(t_1,\tau_1)}{\tau_1 + t_1\widetilde{\mu}_1^2 T_1^2 T_{2,\lambda,\infty}(t_1,\tau_1)}. \tag{81}$$

Here, the functions $T_{2,\lambda,\infty}(\cdot,\cdot)$, $T_{3,\lambda,\infty}(\cdot,\cdot)$ and $g_{\kappa,\lambda,\infty}(\cdot,\cdot)$ can be expressed as follows

$$\begin{aligned}T_{2,\lambda,\infty}(t_1,\tau_1) &= \frac{\delta}{T_1^2}\mathbb{E}\Big[\frac{\kappa}{g_{\kappa,\lambda,\infty}(t_1,\tau_1)}\Big] \Big/ \Big(1 - \frac{\widetilde{\mu}_1^2 t_1 \delta}{\tau_1}\mathbb{E}\Big[\frac{\kappa}{g_{\kappa,\lambda,\infty}(t_1,\tau_1)}\Big]\Big) \\ T_{3,\lambda,\infty}(t_1,\tau_1) &= t_1^2\mathbb{E}\Big[(\widetilde{\mu}_1^2\kappa + \mu_2^2)/g_{\kappa,\lambda,\infty}(t_1,\tau_1)\Big] \\ g_{\kappa,\lambda,\infty}(t_1,\tau_1) &= \frac{t_1}{\tau_1}\Big(\widetilde{\mu}_1^2\kappa + \mu_2^2\Big) + \Big(\widehat{\mu}_1^2\kappa + \mu_3^2\Big) + \lambda.\end{aligned} \tag{82}$$

The property in (79) can also be used to show that the optimal solution $t_2^\star$ and $\tau_2^\star$ of the asymptotic deterministic formulation obtained in Lemma 10 satisfy $t_2^\star = \tau_2^\star$. This then leads to the formulation in (80).

Now, define the asymptotic training and generalization errors stated in Theorem 1 as $\mathcal{E}_{\text{train},\infty}$ and $\mathcal{E}_{\text{test},\infty}$, respectively. Then, the asymptotic training error converges as follows

$$\mathcal{E}_{\text{train},\infty} \xrightarrow{\ell\to+\infty} C^\star(\Delta,\lambda) - \frac{\lambda}{2}\left((q^\star)^2 + h'_\infty(\lambda)\right), \tag{83}$$

where $C^\star(\Delta,\lambda)$ is the optimal cost of the deterministic problem in (80). Here, the function $h_\infty(\cdot)$ is defined as follows

$$h_\infty(\lambda) = -(q^\star)^2\Big(\lambda - \frac{1}{T_{2,\lambda,\infty}(t_1^\star,\tau_1^\star)}\Big) - \eta T_{3,\lambda,\infty}(t_1^\star,\tau_1^\star).$$

Moreover, the asymptotic generalization error converges as follows

$$\mathcal{E}_{\text{train},\infty} \xrightarrow{\ell\to+\infty} \frac{1}{4^v}\mathbb{E}\left[\left(\varphi(g_1) - \widehat{\varphi}(g_2)\right)^2\right], \tag{84}$$

where g_1 and g_2 have a bivariate Gaussian distribution with mean vector $[0, \mu_{0s}\vartheta^\star]$ and covariance matrix $\boldsymbol{C}$, defined as follows

$$\boldsymbol{C} = \begin{bmatrix} 1 & \mu_{1s}\rho T_1 q^\star \\ \mu_{1s}\rho T_1 q^\star & \mu_{1s}^2\beta_\infty^\star + \mu_{2s}^2\left((q^\star)^2 + h'_\infty(\lambda)\right)\end{bmatrix}.$$

The constant $\vartheta^\star$ satisfies $\vartheta^\star = 0$ if $\mu_0 = 0$ and $\vartheta^\star = \gamma_3/\mu_0$ otherwise. Here, the constants μ_{0s}, μ_{1s} and μ_{2s} are defined as $\mu_{0s} = \mathbb{E}[\sigma(z)]$, $\mu_{1s} = \mathbb{E}[z\sigma(z)]$ and $\mu_{2s}^2 = \mathbb{E}[\sigma(z)^2] - \mu_{0s}^2 - \mu_{1s}^2$, where z is a standard Gaussian random variable. Additionally, the constant $\beta_\infty^\star$ can be computed via the following expression

$$\begin{aligned}\beta_\infty^\star &= \frac{1}{V_1+V_3}\Big(V_1T_1^2 - V_2 - V_4 - \lambda + \frac{1}{T_{2,\lambda,\infty}(t_1^\star,\tau_1^\star)}\Big)(q^\star)^2 \\ &+ \frac{\eta T_{3,\lambda,\infty}(t_1^\star,\tau_1^\star)}{V_1+V_3} - \frac{V_2+V_4+\lambda}{V_1+V_3}h'_\infty(\lambda),\end{aligned} \tag{85}$$

where the constants V_1, V_2, V_3 and V_4 are defined as follows

$$V_1 = \frac{t_1^\star \widetilde{\mu}_1^2}{\tau_1^\star},\ V_3 = \widehat{\mu}_1^2,\ V_2 = \frac{t_1^\star \mu_2^2}{\tau_1^\star},\ V_4 = \mu_3^2.$$

Here, $q^\star = q^\star_{t^\star,\tau^\star,\infty}$ satisfies the expression in (81). Moreover, $t_1^\star$ and $\tau_1^\star$ denote the optimal solution of the problem defined in (80). Also, we treat $q^\star$, $t_1^\star$ and $\tau_1^\star$ as constants independent of λ when we compute the derivative of the function $h_\infty(\cdot)$. The results in (83) and (84) can be proved using a similar analysis as in Section 8.2.4. Performing a similar analysis as in Sections 8.2.1, 8.2.2, 8.2.3 and 8.2.4, it can be checked that the training and generalization errors corresponding to the formulation in (76) converge in probability to the limiting functions obtained in (83) and (84), respectively.

Note that the analysis in this Section is valid for any bounds that satisfy the theoretical results in Lemmas 2, 3 and 6. Moreover, observe that the cost functions of both deterministic problems in (60) and (80) diverge when t_1, t_2, τ_1 or τ_2 grows to infinity or when τ_1 or τ_2 goes to 0. This means that the solution of the unconstrained version of the formulations in (60) and (80) should satisfy the feasibility constraints in (60) and (80). This means that the optimization problems in (60) and (80) can be equivalently formulated as in (11) and (19). This completes the proof of Theorem 1, Theorem 2 and Lemma 1.

9. Appendix: Additional Technical Details

In this part, we provide additional technical details to prove the results stated in Theorem 1, Theorem 2 and Lemma 1. Specifically, we provide a rigorous proof of the theoretical results stated in Lemma 5 and Lemma 6.

9.1. Proof of Lemma 5: High–dimensional Equivalence I

The optimization problems given in (37) and (38) share the same feasibility set $\mathcal{D}$ which we define as follows

$$\mathcal{D} = \{(\boldsymbol{w}, t_1, t_2) : \|\boldsymbol{w}\| \leq C_w, 0 \leq t_1 \leq C_{t_1}, 0 \leq t_2 \leq C_{t_2}\}. \tag{86}$$

Define $\widehat{f}_{p,1}$ as the cost function of the optimization problem given in (37) and define $\widehat{f}_{p,2}$ as the cost function of the optimization problem given in (38). Note that the following inequality $|\sqrt{x} - \sqrt{y}| \leq \sqrt{|x-y|}$ is true for any $x \geq 0$ and $y \geq 0$. Therefore, we have the following inequality

$$\sup_{(\boldsymbol{w},t_1,t_2)\in\mathcal{D}} |\widehat{f}_{p,2}(\boldsymbol{w},t_1,t_2) - \widehat{f}_{p,1}(\boldsymbol{w},t_1,t_2)| \leq \sup_{(\boldsymbol{w},t_1,t_2)\in\mathcal{D}} \Big\{\sqrt{\frac{2t_1^2}{\ell^2 n}|Z_{p,1}|} + \sqrt{\frac{2t_2^2}{\ell^2 n}|Z_{p,2}|}\Big\}. \tag{87}$$

where we perform the change of variable $t_1 = t_1/\sqrt{n}$ and $t_2 = t_2/\sqrt{n}$. Here, $Z_{p,1}$ is defined as follows

$$\begin{aligned} Z_{p,1} = &\sqrt{\ell}\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{w}\|\|\boldsymbol{\Gamma}^{\frac{1}{2}}\boldsymbol{w}\|\boldsymbol{h}_1^\top \widehat{\boldsymbol{h}}_2 + \sqrt{\ell}\|\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{w}\|\boldsymbol{h}_1^\top(-\boldsymbol{V}_1^\top\widehat{\boldsymbol{y}} + \mu_0\vartheta\boldsymbol{V}_1^\top\mathbf{1}_{\ell n} + \widetilde{\mu}_1 T_{p,1} q \boldsymbol{V}_1^\top\widehat{\boldsymbol{s}}) \\ &+ \|\boldsymbol{\Gamma}^{\frac{1}{2}}\boldsymbol{w}\|\widehat{\boldsymbol{h}}_2^\top(-\boldsymbol{V}_1^\top\widehat{\boldsymbol{y}} + \mu_0\vartheta\boldsymbol{V}_1^\top\mathbf{1}_{\ell n} + \widetilde{\mu}_1 T_{p,1} q \boldsymbol{V}_1^\top\widehat{\boldsymbol{s}}), \end{aligned} \tag{88}$$

and $Z_{p,2}$ is defined as follows

$$Z_{p,2} = \|\boldsymbol{\Gamma}^{\frac{1}{2}}\boldsymbol{w}\|\widetilde{\boldsymbol{h}}_2^\top(-\boldsymbol{V}_2^\top\widehat{\boldsymbol{y}} + \mu_0\vartheta\boldsymbol{V}_2^\top\mathbf{1}_{\ell n} + \widetilde{\mu}_1 T_{p,1} q \boldsymbol{V}_2^\top\widehat{\boldsymbol{s}}). \tag{89}$$

Given that the set $\mathcal{D}$ is bounded and based on Assumptions 4 and 5, $Z_{p,1}$ and $Z_{p,2}$ can be bounded by a constant independent of the optimization variables. Combining this with the weak law of large numbers, one can see that the right hand side of (87) converges in probability to zero. Then, we obtain the following convergence in probability

$$\sup_{(\boldsymbol{w},t_1,t_2)\in\mathcal{D}} |\widehat{f}_{p,2}(\boldsymbol{w},t_1,t_2) - \widehat{f}_{p,1}(\boldsymbol{w},t_1,t_2)| \xrightarrow{p\to+\infty} 0. \tag{90}$$

Moreover, the following two properties are true for bounded functions

$$\begin{cases} |\sup_{\boldsymbol{x}} f(\boldsymbol{x}) - \sup_{\boldsymbol{x}} g(\boldsymbol{x})| \leq \sup_{\boldsymbol{x}} |f(\boldsymbol{x}) - g(\boldsymbol{x})| \\ |\inf_{\boldsymbol{x}} f(\boldsymbol{x}) - \inf_{\boldsymbol{x}} g(\boldsymbol{x})| \leq \sup_{\boldsymbol{x}} |f(\boldsymbol{x}) - g(\boldsymbol{x})|. \end{cases} \tag{91}$$

Given that the functions $\widehat{f}_{p,1}$ and $\widehat{f}_{p,2}$ are bounded in the set $\mathcal{D}$ and the result in (90), we get the following convergence in probability

$$|\widehat{O}^{\star}_{p,1} - \widehat{O}^{\star}_{p,2}| \xrightarrow{p\to+\infty} 0, \tag{92}$$

where $\widehat{O}^{\star}_{p,1}$ and $\widehat{O}^{\star}_{p,2}$ are the optimal objective values of the optimization problems given in (37) and (38), respectively. Now, define $\widehat{\mathcal{S}}^{\star}_{p,1}$ and $\widehat{\mathcal{S}}^{\star}_{p,2}$ as the set of optimal solutions of the minimization problems in (37) and (38), respectively. Next, the objective is to show that

$$\mathbb{D}(\widehat{\mathcal{S}}^{\star}_{p,1}, \widehat{\mathcal{S}}^{\star}_{p,2}) \xrightarrow{p} 0. \tag{93}$$

Moreover, define the functions $\widetilde{f}_{p,1}$ and $\widetilde{f}_{p,2}$ as follows

$$\begin{cases} \widetilde{f}_{p,1}(\boldsymbol{w}) = \max\limits_{\substack{0\leq t_1\leq C_{t_1}\\ 0\leq t\leq C_{t_2}}} \widehat{f}_{p,1}(\boldsymbol{w}, t_1, t_2) \\ \widetilde{f}_{p,2}(\boldsymbol{w}) = \max\limits_{\substack{0\leq t_1\leq C_{t_1}\\ 0\leq t\leq C_{t_2}}} \widehat{f}_{p,2}(\boldsymbol{w}, t_1, t_2). \end{cases} \tag{94}$$

Note that the set $\widehat{\mathcal{S}}^{\star}_{p,1}$ is the set of minimizing $\boldsymbol{w}$ of the first function in (94). Based on Lemma 4, the function $\widetilde{f}_{p,2}$ is strongly convex in the feasibility set where λ is a strong convexity parameter. This means that it has a unique minimizer denoted by $\boldsymbol{w}^{\star}_{p,2}$. Now, assume that $\boldsymbol{w}^{\star}_{p,1}$ is a minimizer of the function $\widetilde{f}_{p,1}$. Moreover, assume that there exists $\gamma > 0$ independent of p such that the following convergence holds true

$$\mathbb{P}\Big(\sup_{\boldsymbol{w}^{\star}\in\widehat{\mathcal{S}}^{\star}_{p,1}} \|\boldsymbol{w}^{\star} - \boldsymbol{w}^{\star}_{p,2}\|_2 \geq \gamma \Big) \xrightarrow{p\to\infty} 1. \tag{95}$$

Given the strong convexity of the function $\widetilde{f}_{p,2}$, we have the following inequality

$$\begin{aligned} \widetilde{f}_{p,2}(\beta\boldsymbol{w}_1 + (1-\beta)\boldsymbol{w}_2) \leq \beta\widetilde{f}_{p,2}(\boldsymbol{w}_1) + (1-\beta)\widetilde{f}_{p,2}(\boldsymbol{w}_2) \\ - \frac{\lambda}{2}\beta(1-\beta)\|\boldsymbol{w}_1 - \boldsymbol{w}_2\|_2^2, \end{aligned} \tag{96}$$

where this is valid for any $\beta \in [0,1]$ and feasible $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$. Take $\boldsymbol{w}_1 = \boldsymbol{w}^{\star}_{p,1}$, $\boldsymbol{w}_2 = \boldsymbol{w}^{\star}_{p,2}$ and $\beta = 1/2$. Based on the fact that $\boldsymbol{w}^{\star}_{p,2}$ is a minimizer of the function $\widetilde{f}_{p,2}$, there exists $\gamma > 0$ independent of p such that

$$\mathbb{P}\Big(\sup_{\boldsymbol{w}^{\star}\in\widehat{\mathcal{S}}^{\star}_{p,1}} |\widetilde{f}_{p,2}(\boldsymbol{w}^{\star}_{p,2}) - \widetilde{f}_{p,2}(\boldsymbol{w}^{\star})| \geq \frac{\lambda\gamma^2}{4} \Big) \xrightarrow{p\to\infty} 1. \tag{97}$$

Next, we use the convergence in probability established in (90) and (92) to show that the result in (97) produces a contradiction. To this end, note that the following inequality is always valid

$$|\widetilde{f}_{p,2}(\boldsymbol{w}^{\star}_{p,2}) - \widetilde{f}_{p,2}(\boldsymbol{w}^{\star}_{n,1})| \leq |\widetilde{f}_{p,2}(\boldsymbol{w}^{\star}_{p,2}) - \widetilde{f}_{p,1}(\boldsymbol{w}^{\star}_{p,1})| + |\widetilde{f}_{p,1}(\boldsymbol{w}^{\star}_{p,1}) - \widetilde{f}_{p,2}(\boldsymbol{w}^{\star}_{p,1})|, \tag{98}$$

which means that the following inequality is always true

$$|\widetilde{f}_{p,2}(\boldsymbol{w}^{\star}_{p,2}) - \widetilde{f}_{p,2}(\boldsymbol{w}^{\star}_{p,1})| \leq |O^{\star}_{p,2} - O^{\star}_{p,1}| + \sup_{\|\boldsymbol{w}\|\leq C_w} |\widetilde{f}_{p,1}(\boldsymbol{w}) - \widetilde{f}_{p,2}(\boldsymbol{w})|. \tag{99}$$

Observe that the inequality derived in (99) implies that the following inequality holds true

$$\sup_{\boldsymbol{w}^{\star}\in\widehat{\mathcal{S}}^{\star}_{p,1}} |\widetilde{f}_{p,2}(\boldsymbol{w}^{\star}_{p,2}) - \widetilde{f}_{p,2}(\boldsymbol{w}^{\star})| \leq |O^{\star}_{p,2} - O^{\star}_{p,1}| + \sup_{\|\boldsymbol{w}\|\leq C_w} |\widetilde{f}_{p,1}(\boldsymbol{w}) - \widetilde{f}_{p,2}(\boldsymbol{w})|. \tag{100}$$

Now, based on (90), (91) and (92), the right hand side of (99), converges in probability to zero. This means that the following convergence in probability holds

$$\sup_{\boldsymbol{w}^{\star}\in\widehat{\mathcal{S}}^{\star}_{p,1}} |\widetilde{f}_{p,2}(\boldsymbol{w}^{\star}_{p,2}) - \widetilde{f}_{p,2}(\boldsymbol{w}^{\star})| \xrightarrow{p\to+\infty} 0. \tag{101}$$

This contradicts with the result in (97). This means that for any $\epsilon_1 > 0$ and $\epsilon_2 > 0$, there exists $p_0 \in \mathbb{N}$ such that for any $p \geq p_0$, we have that

$$\mathbb{P}\Big(\sup_{\boldsymbol{w}^\star \in \widehat{\mathcal{S}}^\star_{p,1}} \|\boldsymbol{w}^\star - \boldsymbol{w}^\star_{p,2}\|_2 < \epsilon_1 \Big) \geq 1 - \epsilon_2. \tag{102}$$

This means that the following convergence in probability is true

$$\mathbb{D}(\widehat{\mathcal{S}}^\star_{p,1}, \widehat{\mathcal{S}}^\star_{p,2}) \xrightarrow{p \to +\infty} 0, \tag{103}$$

where $\mathbb{D}(\mathcal{A}, \mathcal{B})$ denotes the deviation between the sets $\mathcal{A}$ and $\mathcal{B}$ and is defined as $\mathbb{D}(\mathcal{A}, \mathcal{B}) = \sup_{\boldsymbol{x}_1 \in \mathcal{A}} \inf_{\boldsymbol{x}_2 \in \mathcal{B}} \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2$. This completes the proof of Lemma 5.

9.2. Proof of Lemma 6: Additional Compactness

We start our prove by analyzing the feasibility sets of the primal formulation in (27). Note that the optimal solution of the formulation given in (27) can be expressed in closed form as follows

$$\widehat{\boldsymbol{w}}_p = \left[\frac{1}{n\ell} \boldsymbol{K}^\top \boldsymbol{K} + \lambda \boldsymbol{I}_k \right]^{-1} \left(\frac{\boldsymbol{K}^\top \bar{\boldsymbol{y}}}{n\ell} \right), \tag{104}$$

for a sufficiently large C_w. Here, the matrix $\boldsymbol{K} \in \mathbb{R}^{n\ell \times k}$ is defined as follows

$$\boldsymbol{K} = \widetilde{\mu}_1 \bar{\boldsymbol{s}} \boldsymbol{\xi}^\top \boldsymbol{F} + \bar{\boldsymbol{G}} \boldsymbol{\Sigma}^{\frac{1}{2}} + \boldsymbol{T} \boldsymbol{\Gamma}^{\frac{1}{2}}. \tag{105}$$

The matrices $\boldsymbol{\Sigma}$ and $\boldsymbol{\Gamma}$ are defined in (33). Moreover, $\bar{\boldsymbol{y}}$, $\bar{\boldsymbol{s}}$ and $\bar{\boldsymbol{G}}$ are formed by performing ℓ times concatenation of $\widetilde{\boldsymbol{y}} = \boldsymbol{y} - \mu_0 \vartheta \mathbf{1}_n$, $\boldsymbol{s}$ and $\boldsymbol{G}$. Here, $\boldsymbol{y} = \varphi(\boldsymbol{s})$ and $\boldsymbol{s}$, $\boldsymbol{G}$ and $\boldsymbol{T}$ have independent standard Gaussian components. Now, based on the results in (Rudelson & Vershynin, 2010) and Assumptions 4 and 5, there exists a positive constant $C_1 > 0$ such that

$$\|\boldsymbol{K}\| / \sqrt{n} \leq C_1, \tag{106}$$

with probability going to 1 as p grows to $+\infty$. Therefore, there exists a positive constant $C_2 > 0$ such that

$$\sigma_{\min}\Big(\big[\frac{1}{n\ell} \boldsymbol{K}^\top \boldsymbol{K} + \lambda \boldsymbol{I}_k \big]^{-1} \Big) \geq \frac{1}{C_2 + \lambda}, \tag{107}$$

where $\sigma_{\min}(\cdot)$ denotes the minimum eigenvalue. Now, observe that

$$\|\boldsymbol{K}^\top \bar{\boldsymbol{y}}\|^2 = \widetilde{\mu}_1^2 (\bar{\boldsymbol{s}}^\top \bar{\boldsymbol{y}})^2 \boldsymbol{\xi}^\top \boldsymbol{F} \boldsymbol{F}^\top \boldsymbol{\xi} + \bar{\boldsymbol{y}}^\top \boldsymbol{B} \boldsymbol{B}^\top \bar{\boldsymbol{y}} + 2 \widetilde{\mu}_1 (\bar{\boldsymbol{s}}^\top \bar{\boldsymbol{y}}) \boldsymbol{\xi}^\top \boldsymbol{F} \boldsymbol{B}^\top \bar{\boldsymbol{y}}, \tag{108}$$

where $\boldsymbol{B} = \bar{\boldsymbol{G}} \boldsymbol{\Sigma}^{\frac{1}{2}} + \boldsymbol{T} \boldsymbol{\Gamma}^{\frac{1}{2}}$. Given that the random quantities $\boldsymbol{s}$, $\boldsymbol{G}$ and $\boldsymbol{T}$ have independent standard Gaussian components, we have the following

$$\frac{1}{\ell n} \boldsymbol{\xi}^\top \boldsymbol{F} \boldsymbol{B}^\top \bar{\boldsymbol{y}} \xrightarrow{p \to +\infty} 0. \tag{109}$$

Moreover, using the weak law of large numbers and Assumptions 3 and 5, we obtain the following asymptotic results

$$\frac{\bar{\boldsymbol{s}}^\top \bar{\boldsymbol{y}}}{\ell n} \xrightarrow{p \to +\infty} \mathbb{E}[z\varphi(z)], \; \boldsymbol{\xi}^\top \boldsymbol{F} \boldsymbol{F}^\top \boldsymbol{\xi} \xrightarrow{p \to +\infty} \delta \mathbb{E}[\kappa]. \tag{110}$$

Combining this with Assumptions 3, 4 and 5, we obtain the following inequality

$$\frac{1}{(\ell n)^2} \|\boldsymbol{K}^\top \bar{\boldsymbol{y}}\|^2 \geq \frac{1}{2} \widetilde{\mu}_1^2 \delta \mathbb{E}[z\varphi(z)]^2 \mathbb{E}[\kappa], \tag{111}$$

valid with probability going to 1 as p grows to infinity. This shows that there exists a positive constant $c_w > 0$ such that

$$\|\widehat{\boldsymbol{w}}_p\| \geq c_w, \tag{112}$$

with probability going to 1 as p grows to infinity. Then, we can apply the multivariate CGMT framework with the additional constraint in (112). Based on this result and Assumption 4, there exists positive constants $c_{\tau_1} > 0$, $C_{\tau_1} > 0$, $c_{\tau_2} > 0$ and $C_{\tau_2} > 0$, such that the following convergence in probability holds

$$\mathbb{P}(c_{\tau_1} \leq \widehat{\tau}_1 \leq C_{\tau_1}) \xrightarrow{n \to \infty} 1, \; \mathbb{P}(c_{\tau_2} \leq \widehat{\tau}_2 \leq C_{\tau_2}) \xrightarrow{n \to \infty} 1, \tag{113}$$

where $\widehat{\tau}_1$and $\widehat{\tau}_2$ are the optimal solutions of the formulation in (41). This completes the proof of Lemma 6.