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:

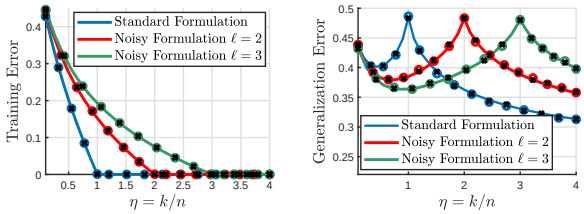
$$\min_{\boldsymbol{w}\in\mathbb{R}^k} \frac{1}{2n\ell}\sum_{i=1}^{n}\sum_{j=1}^{\ell}\big(y_i - \widetilde{\mu}_1 \boldsymbol{w}^\top \boldsymbol{F}^\top \boldsymbol{a}_i - \widehat{\mu}_1 \boldsymbol{w}^\top \boldsymbol{F}^\top \boldsymbol{z}_{ij} - \mu_0 \boldsymbol{w}^\top \boldsymbol{1}_k - \mu_2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

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

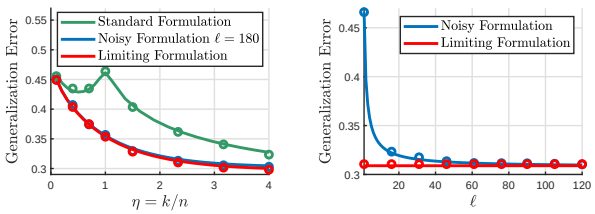
$$\min_{\boldsymbol{w}\in\mathcal{S}_{\boldsymbol{w}}} \max_{\boldsymbol{u}\in\mathcal{S}_{\boldsymbol{u}}} [\boldsymbol{u}_1^\top, \ldots, \boldsymbol{u}_\ell^\top] \begin{bmatrix} \boldsymbol{G}_1 \\ \vdots \\ \boldsymbol{G}_\ell \end{bmatrix} \boldsymbol{w} + \psi(\boldsymbol{w}, \boldsymbol{u}).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

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* $\boldsymbol{F} = \boldsymbol{USV}$*, 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

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

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

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

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

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 the supplementary materiel. 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 the supplementary materiel 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, 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 the supplementary materiel. 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

$$\widehat{\beta}^{\star} = \frac{1}{V_1 + V_3}\Big(V_1T_1^2 - V_2 - V_4 - \lambda + \frac{1}{\widehat{T}_{2,\lambda}(t_1^{\star},\tau_1^{\star})}\Big)(\widehat{q}^{\star})^2 + \frac{\eta\widehat{T}_{3,\lambda}(t_1^{\star},\tau_1^{\star})}{V_1 + V_3} - \frac{V_2 + V_4 + \lambda}{V_1 + V_3}\widehat{h}'(\lambda), \tag{20}$$

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 $\widehat{h}(\cdot)$.

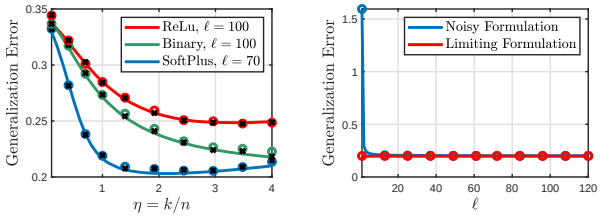
The proof of Lemma 1 is provided in the supplementary materiel. The results in Theorem 2 and Lemma 1 are based on the asymptotic predictions stated in Theorem 1. Specifically, we show in the supplementary materiel that the asymptotic problem corresponding to the noisy formulation in (11) converges to 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 the supplementary materiel shows that the deterministic problem in (19) is strictly convex-concave. This implies the uniqueness of its optimal solutions.

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)

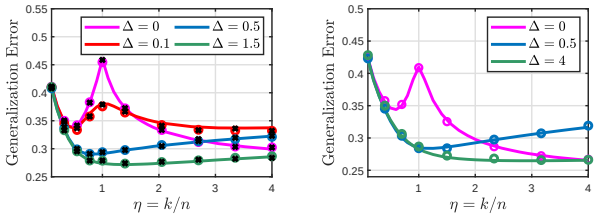
Figure 3. Solid line: Theoretical predictions. Circle: Numerical simulations for (3) in 3(a) and for both (3) and (8) in 3(b). Black cross: Numerical simulations for (8). **(a)** $p = 500$, $\Delta = 0.4$, $\alpha = 1.5$ and $\lambda = 10^{-2}$. **(b)** $p = 500$, $\Delta = 0.6$, $\alpha = 2$, $\lambda = 10^{-4}$, $\eta = 1$ 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

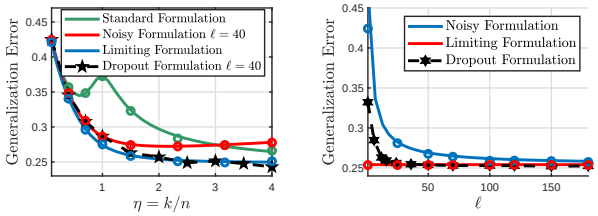
the double descent phenomenon can be mitigated for an appropriately selected noise variance.

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.

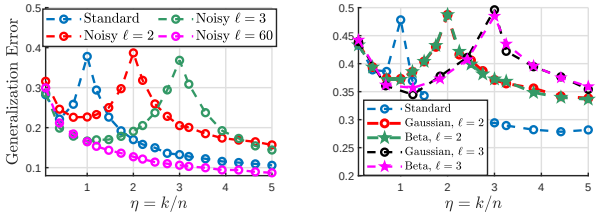


(a) (b)

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.

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.

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.

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.

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.

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.

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.