Regularization Trade-offs with Fake Features

Martin Hellkvist, Ayça Özçelikkale, Anders Ahlén
Department of Electrical Engineering, Uppsala University, Sweden
{Martin.Hellkvist, Ayca.Ozcelikkale, Anders.Ahlen}@angstrom.uu.se

***Abstract*—Recent successes of massively overparameterized models have inspired a new line of work investigating the underlying conditions that enable overparameterized models to generalize well. This paper considers a framework where the possibly overparametrized model includes fake features, i.e., features that are present in the model but not in the data. We present a non-asymptotic high-probability bound on the generalization error of the ridge regression problem under the model misspecification of having fake features. Our high-probability results provide insights into the interplay between the implicit regularization provided by the fake features and the explicit regularization provided by the ridge parameter. Numerical results illustrate the trade-off between the number of fake features and how the optimal ridge parameter may heavily depend on the number of fake features.**

***Index Terms*—Linear systems, inverse problems, interpolation, least-squares methods, robust linear regression**

I. Introduction

Conventional wisdom in statistical learning suggests that the number of training samples should exceed the number of model parameters in order to generalize well to data unseen during training. However, it has recently been highlighted that the generalization error initially decreases with the model size in the underparametrized setting, and then again in the overparametrized setting, hence the phenomenon of *double descent* has been proposed [1], [2].

Double-descent behaviour can be caused by *missing features*, i.e., features present in the data but not in the model [2], [3]. Recently, surprising effects of additional irrelevant features in the model, referred to as *fake features*, i.e., features present in the model but not in the data, have been demonstrated [4]–[7]. In particular, inclusion of fake features has been used to improve the estimation performance [4], [5]. In this paper, we contribute to this line of work by providing high probability results in the finite regime for the generalization error associated with the ridge regression problem and reveal insights into the trade-offs between the implicit regularization provided by the fake features and the explicit ridge regularization.

Observed for a wide range of models [1], the double-descent phenomenon in linear regression has been studied for the finite-dimensional case with Gaussian, subgaussian and random features [2], [8]–[11] and in a Bayesian estimation setting [5], as well as in the asymptotic high-dimensional setting [3]. Extensions have been made to the investigation of optimal ridge parameter values [6], [12], and to the study of fake features [4]–[6]. Trade-offs between explicit regularization and implicit regularization provided by different problem aspects have been investigated, e.g., implicit regularization by asymptotic overparameterization [6] and the equivalence of training noise and Tikhonov regularization in [13].

Model misspecification often lead to double-descent curves [2], [3], [5]. Robust methods under model misspecification have been focused in various works, such as covariance matrix uncertainties in linear minimum mean-square error estimation [14], [15], and robust estimation with missing features [16].

Contributions: In this article, we contribute to the line of work with fake features under ridge-regression. Our main contribution, Theorem 1, presents a high-probability bound for the generalization error of the finite-dimensional ridge regression problem with fake features. This is in contrast to earlier work which do not study regularization [2], [8] or fake features [17], study the asymptotic regime [3], [6], or provide results in terms of expectations over the regressor distribution [5]. Our result in Theorem 1 quantifies the trade-off between the fake features and the regularization parameter, and provides insights into the mechanism behind this trade-off through high-probability bounds on the eigenvalues. Our focus on the i.i.d. Gaussian case allows us to provide clear expressions. Our numerical results quantify how the implicit regularization provided by the fake features may compensate for a small ridge parameter in certain scenarios.

II. Problem Statement

A. Data generation:

The data comes from the following linear underlying system,

$$\boldsymbol{y} = \tilde{\boldsymbol{A}}\tilde{\boldsymbol{x}} + \boldsymbol{v} = \boldsymbol{A}_S\boldsymbol{x}_S + \boldsymbol{A}_C\boldsymbol{x}_C + \boldsymbol{v}, \tag{1}$$

where $\boldsymbol{y} = [y_1, \cdots, y_n]^{\mathrm{T}} \in \mathbb{R}^{n\times 1}$ is the vector of outputs/observations, $\tilde{\boldsymbol{x}} \in \mathbb{R}^{\tilde{p}\times 1}$ is the unknowns of interest and $\boldsymbol{v} = [v_1, \cdots, v_n] \in \mathbb{R}^{n\times 1}$ is the vector of noise, with $v_i \sim \mathcal{N}(0, \sigma_v^2)$, $\forall i$, $\sigma_v \geq 0$. The feature matrix $\tilde{\boldsymbol{A}} \in \mathbb{R}^{n\times\tilde{p}}$ is composed of the matrices $\boldsymbol{A}_S \in \mathbb{R}^{n\times p_S}$ and $\boldsymbol{A}_C \in \mathbb{R}^{n\times p_C}$, as

$$\tilde{\boldsymbol{A}} = [\boldsymbol{A}_S, \boldsymbol{A}_C], \tag{2}$$

with $\tilde{p} = p_S + p_C$. The matrices $\boldsymbol{A}_S$ and $\boldsymbol{A}_C$ consist of identically and independently distributed (i.i.d.) standard Gaussian entries $\sim \mathcal{N}(0,1)$, which are uncorrelated with the noise $\boldsymbol{v}$. The vector of unknowns $\tilde{\boldsymbol{x}}$ is composed of the components $\boldsymbol{x}_S \in \mathbb{R}^{p_S\times 1}$ and $\boldsymbol{x}_C \in \mathbb{R}^{p_C\times 1}$, such that

$$\tilde{\boldsymbol{x}} = [\boldsymbol{x}_S^{\mathrm{T}}, \boldsymbol{x}_C^{\mathrm{T}}]^{\mathrm{T}}. \tag{3}$$

B. Misspecified model:

While the data is generated by the underlying system in (1), the estimation is performed based on the following misspecified model,

$$\boldsymbol{y} = \bar{\boldsymbol{A}}\bar{\boldsymbol{x}} + \boldsymbol{v} = \boldsymbol{A}_F\boldsymbol{x}_F + \boldsymbol{A}_S\boldsymbol{x}_S + \boldsymbol{v}, \tag{4}$$

where $\bar{\boldsymbol{A}} \in \mathbb{R}^{n\times\bar{p}}$ is composed by

$$\bar{\boldsymbol{A}} = [\boldsymbol{A}_F, \boldsymbol{A}_S] \tag{5}$$

with $\bar{p} = p_F + p_S$. The matrix $\boldsymbol{A}_F \in \mathbb{R}^{n\times p_F}$ has random i.i.d. standard Gaussian entries, statistically independent of $\boldsymbol{A}_S$ and $\boldsymbol{A}_C$. The vector $\bar{\boldsymbol{x}}$ is correspondingly composed as $\bar{\boldsymbol{x}} = [\boldsymbol{x}_F^{\mathrm{T}}, \boldsymbol{x}_S^{\mathrm{T}}] \in \mathbb{R}^{\bar{p}\times 1}$, where $\boldsymbol{x}_F \in \mathbb{R}^{p_F\times 1}$.

We refer to the features in $\boldsymbol{A}_F$, $\boldsymbol{A}_S$ and $\boldsymbol{A}_C$, as follows:

- The features in $\boldsymbol{A}_F$ are included in the misspecified model in (4), but are irrelevant to the output variable $\boldsymbol{y}$, i.e., the data in (1), hence we refer to $\boldsymbol{A}_F$ as **fake features**.
- The features $\boldsymbol{A}_S$ are present both in the data generated by (1) and the misspecified model in (4), hence we refer to them as **included underlying features**.
- The features in $\boldsymbol{A}_C$, which are relevant to the data in $\boldsymbol{y}$, are missing from the misspecified model in (4). Hence we refer to the features $\boldsymbol{A}_C$ as **missing features**.

We employ the notation

$$\boldsymbol{A} = [\boldsymbol{A}_F, \boldsymbol{A}_S, \boldsymbol{A}_C] \in \mathbb{R}^{n\times p} \tag{6}$$

to refer to the full set of features, and correspondingly for the full set of unknowns,

$$\boldsymbol{x} = [\boldsymbol{x}_F^{\mathrm{T}}, \boldsymbol{x}_S^{\mathrm{T}}, \boldsymbol{x}_C^{\mathrm{T}}]^{\mathrm{T}} \in \mathbb{R}^{p\times 1}, \tag{7}$$

where $p = p_F + p_S + p_C$.

With the misspecified model in (4), we estimate $\boldsymbol{x}_F$ and $\boldsymbol{x}_S$ and we obtain the prediction of $\boldsymbol{y}$ as

$$\hat{\boldsymbol{y}} = \boldsymbol{A}_F\hat{\boldsymbol{x}}_F + \boldsymbol{A}_S\hat{\boldsymbol{x}}_S = \bar{\boldsymbol{A}}\hat{\bar{\boldsymbol{x}}} \in \mathbb{R}^{n\times 1}. \tag{8}$$

Recall that $\bar{\boldsymbol{x}} = [\boldsymbol{x}_F^{\mathrm{T}}, \boldsymbol{x}_S^{\mathrm{T}}]$. We obtain the estimate $\hat{\bar{\boldsymbol{x}}}$ by solving the following problem,

$$\begin{aligned}\hat{\bar{\boldsymbol{x}}} &= \arg\min_{\hat{\bar{\boldsymbol{x}}}} \|\boldsymbol{y} - (\boldsymbol{A}_F\hat{\boldsymbol{x}}_F + \boldsymbol{A}_S\hat{\boldsymbol{x}}_S)\|^2 + \lambda\|\hat{\bar{\boldsymbol{x}}}\|^2 && (9)\\ &= \arg\min_{\hat{\bar{\boldsymbol{x}}}} \left\|\boldsymbol{y} - \bar{\boldsymbol{A}}\hat{\bar{\boldsymbol{x}}}\right\|^2 + \lambda\|\hat{\bar{\boldsymbol{x}}}\|^2, && (10)\end{aligned}$$

where $\lambda \geq 0$ is the regularization parameter. Here, (9) with $\lambda > 0$ corresponds to the ridge regression problem whose solution is given by

$$\hat{\bar{\boldsymbol{x}}} = \left(\bar{\boldsymbol{A}}^{\mathrm{T}}\bar{\boldsymbol{A}} + \lambda\boldsymbol{I}_p\right)^{-1}\bar{\boldsymbol{A}}^{\mathrm{T}}\boldsymbol{y} = \bar{\boldsymbol{A}}^{\mathrm{T}}\left(\bar{\boldsymbol{A}}\bar{\boldsymbol{A}}^{\mathrm{T}} + \lambda\boldsymbol{I}_n\right)^{-1}\boldsymbol{y}. \tag{11}$$

If $\lambda = 0$, we consider the minimum ℓ_2-norm solution of (9),

$$\hat{\bar{\boldsymbol{x}}} = \bar{\boldsymbol{A}}^{+}\boldsymbol{y}, \tag{12}$$

where $(\cdot)^{+}$ denotes the Moore-Penrose pseudoinverse. The estimate obtained by solving (9) can be decomposed as

$$\hat{\bar{\boldsymbol{x}}} = \left[\hat{\boldsymbol{x}}_F^{\mathrm{T}}\ \hat{\boldsymbol{x}}_S^{\mathrm{T}}\right]^{\mathrm{T}}. \tag{13}$$

Using $\hat{\bar{\boldsymbol{x}}}$, we obtain the estimate for $\boldsymbol{x} = [\boldsymbol{x}_F^{\mathrm{T}}, \boldsymbol{x}_S^{\mathrm{T}}, \boldsymbol{x}_C^{\mathrm{T}}]^{\mathrm{T}}$ as follows,

$$\hat{\boldsymbol{x}} = \begin{bmatrix}\hat{\bar{\boldsymbol{x}}}\\ \hat{\boldsymbol{x}}_C\end{bmatrix} = \begin{bmatrix}\hat{\boldsymbol{x}}_F\\ \hat{\boldsymbol{x}}_S\\ \hat{\boldsymbol{x}}_C\end{bmatrix} = \begin{bmatrix}\hat{\boldsymbol{x}}_F\\ \hat{\boldsymbol{x}}_S\\ \boldsymbol{0}\end{bmatrix}, \tag{14}$$

where the estimate for the missing features is set to zero, i.e., $\hat{\boldsymbol{x}}_C = \boldsymbol{0}$, as $\boldsymbol{A}_C$ does not appear in the misspecified model (4).

C. Generalization Error:

Suppose that we have obtained an estimate $\hat{\boldsymbol{x}}$ as in (14). A new unseen sample $(y_*, \boldsymbol{a}_*)$ comes where $\boldsymbol{a}_* = [\boldsymbol{a}_{F*}^{\mathrm{T}}, \boldsymbol{a}_{S*}^{\mathrm{T}}, \boldsymbol{a}_{C*}^{\mathrm{T}}]^{\mathrm{T}}$. Hence,

$$y_* = \boldsymbol{a}_{S*}^{\mathrm{T}}\boldsymbol{x}_S + \boldsymbol{a}_{C*}^{\mathrm{T}}\boldsymbol{x}_C + v_* \in \mathbb{R}^{1\times 1}, \tag{15}$$

where $\boldsymbol{a}_{F*}^{\mathrm{T}} \in \mathbb{R}^{1\times p_F}$, $\boldsymbol{a}_{S*}^{\mathrm{T}} \in \mathbb{R}^{1\times p_S}$, and $\boldsymbol{a}_{C*}^{\mathrm{T}} \in \mathbb{R}^{1\times p_C}$ are i.i.d. with the rows of $\boldsymbol{A}_F$, $\boldsymbol{A}_S$ and $\boldsymbol{A}_C$ respectively, and $v_* \in \mathbb{R}^{1\times 1}$ is i.i.d. with the noise samples in $\boldsymbol{v}$. The corresponding prediction using $\hat{\boldsymbol{x}}$ is

$$\hat{y}_* = \boldsymbol{a}_{F*}^{\mathrm{T}}\hat{\boldsymbol{x}}_F + \boldsymbol{a}_{S*}^{\mathrm{T}}\hat{\boldsymbol{x}}_S. \tag{16}$$

The *generalization error* is given by

$$\begin{aligned}J_y &= \mathop{\mathbb{E}}_{y_*,\boldsymbol{a}_*}\left[(y_* - \hat{y}_*)^2\right] && (17)\\ &= \mathop{\mathbb{E}}_{y_*,\boldsymbol{a}_*}\left[(\boldsymbol{a}_{S*}^{\mathrm{T}}\boldsymbol{x}_S + \boldsymbol{a}_{C*}^{\mathrm{T}}\boldsymbol{x}_C + v_* - \boldsymbol{a}_{F*}^{\mathrm{T}}\hat{\boldsymbol{x}}_F - \boldsymbol{a}_{S*}^{\mathrm{T}}\hat{\boldsymbol{x}}_S)^2\right] && (18)\\ &= \mathop{\mathbb{E}}_{y_*,\boldsymbol{a}_*}\left[\left([\boldsymbol{a}_{F*}^{\mathrm{T}}, \boldsymbol{a}_{S*}^{\mathrm{T}}, \boldsymbol{a}_{C*}^{\mathrm{T}}]\left(\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{x}_S\\ \boldsymbol{x}_C\end{bmatrix} - \begin{bmatrix}\hat{\boldsymbol{x}}_F\\ \hat{\boldsymbol{x}}_S\\ \boldsymbol{0}\end{bmatrix}\right) + v_*\right)^2\right] && (19)\\ &= \left\|\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{x}_S\\ \boldsymbol{x}_C\end{bmatrix} - \begin{bmatrix}\hat{\boldsymbol{x}}_F\\ \hat{\boldsymbol{x}}_S\\ \boldsymbol{0}\end{bmatrix}\right\|^2 + \sigma_v^2 && (20)\end{aligned}$$

We note that the generalization error consists of the respective errors in the components of $\boldsymbol{x}$ that correspond to the fake features $\boldsymbol{A}_F$, the included underlying features $\boldsymbol{A}_S$ and the missing features $\boldsymbol{A}_C$.

Remark 1. (Interpolation with fake features) *Recall that $\bar{\boldsymbol{A}} = [\boldsymbol{A}_F, \boldsymbol{A}_S] \in \mathbb{R}^{n\times\bar{p}}$, and that $\bar{p} = p_F + p_S$, hence the estimate $\hat{\bar{\boldsymbol{x}}} = \bar{\boldsymbol{A}}^{+}\boldsymbol{y}$ in* (12) *is created using the fake features in $\boldsymbol{A}_F \in \mathbb{R}^{n\times p_F}$ and included underlying features $\boldsymbol{A}_S \in \mathbb{R}^{n\times p_S}$. If $n < \bar{p}$, then $\bar{\boldsymbol{A}}\bar{\boldsymbol{A}}^{\mathrm{T}}$ is full rank with probability one (since entries of $\bar{\boldsymbol{A}}$ are standard Gaussian i.i.d.), and the estimate $\hat{\boldsymbol{y}} \in \mathbb{R}^{n\times 1}$ of the data $\boldsymbol{y}$ is*

$$\hat{\boldsymbol{y}} = \boldsymbol{A}_F\hat{\boldsymbol{x}}_F + \boldsymbol{A}_S\hat{\boldsymbol{x}}_S = \bar{\boldsymbol{A}}\hat{\bar{\boldsymbol{x}}} = \bar{\boldsymbol{A}}\bar{\boldsymbol{A}}^{+}\boldsymbol{y} = \boldsymbol{y}, \tag{21}$$

hence the training data is interpolated for $n < \bar{p}$, even when there are fake features in the misspecified model. Furthermore, we note that even if the misspecified model consists purely of fake features, i.e., if $p_S = 0$, and $n < p_F$, then we still have $\hat{\boldsymbol{y}} = \boldsymbol{y}$. Hence, we still obtain interpolation without using any of the underlying features $\boldsymbol{A}_S$ and $\boldsymbol{A}_C$ in the estimation process. We refer to the point where $n = \bar{p}$ as the interpolation threshold.

III. Generalization Error Bound

In this section, we give our main result of the paper, which is a high-probability bound on the generalization error J_y in the finite-dimensional regime for the ridge regression problem with $\lambda > 0$. Note that here we analyze the generalization error

J_y in high probability with respect to training data whereas J_y itself is an average over test data.

Theorem 1. *Let the regularization parameter be nonzero, i.e.,* $\lambda > 0$, *and* $t_1, t_2 \geq 0$, $r_{\max} = \max(n, \bar{p})$, $r_{\min} = \min(n, \bar{p})$, *and*

$$f_g = \frac{(\sqrt{n} + \sqrt{\bar{p}} + t_2)^2}{((\sqrt{r_{\max}} - \sqrt{r_{\min}} - t_2)_+^2 + \lambda)^2}, \qquad (22)$$

where $(\cdot)_+ = \max(\cdot, 0)$, $(\cdot)_+^2 = ((\cdot)_+)^2$ *and*

$$\bar{f}_g = \begin{cases} \dfrac{\lambda^2}{((\sqrt{n} - \sqrt{\bar{p}} - t_2)_+^2 + \lambda)^2}, & \text{if } n \geq \bar{p}, \qquad (23a) \\ 1, & \text{if } n < \bar{p}, \qquad (23b) \end{cases}$$

then the following holds for the generalization error in (20),

$$\begin{aligned} \mathbb{P}\Big(J_y < \|\boldsymbol{x}_S\|^2 \bar{f}_g & \\ + \left(\|\boldsymbol{x}_C\|^2 + \sigma_v^2\right) f_g \left(r_{\min} + 2\sqrt{r_{\min} t_1} + 2t_1\right) & \qquad (24) \\ + \left(\|\boldsymbol{x}_C\|^2 + \sigma_v^2\right) \Big) > 1 - e^{-t_1} - 2e^{-t_2^2/2}. & \end{aligned}$$

Proof: See Appendix A.

Note that if $t_2 \geq \sqrt{r_{\max}} - \sqrt{r_{\min}}$, then the denominators in (22) and (23a) reduces to λ^2.

In Theorem 1, both the upper bound on J_y and the probability that the upper bound holds depend on t_1 and t_2. Hence, by varying t_1 and t_2, one obtains a series of upper bounds and associated probabilities.

From Theorem 1, we observe the following:

1) In order to avoid a very high value in the generalization error at the interpolation threshold $n = \bar{p}$ (Remark 1), the ridge parameter λ needs to be large enough. Otherwise, the probability parameter t_2 cannot be large enough to guarantee the bound in (24) holds without making the bound very large due to the denominators being too small in (22) and (23a).

2) In addition to the explicit ridge regularization, the fake features in $\boldsymbol{A}_F$ have a regularizing effect on the error bound. Suppose that $n \approx p_S$ and λ is very small, hence the problem without fake features is close to the interpolation threshold at $n = p_S$, and the bound in (24) is very large. If there are enough fake features, then the actual problem dimensions will be far away from the threshold $n \approx \bar{p}$, hence the bound will take on smaller values. Nevertheless, if the regularization parameter λ is large enough, then the bound takes on small values regardless of the presence of fake features.

Remark 2. *Theorem 1 quantifies the trade-off between the number of fake features and the ridge parameter using a high-probability bound. In contrast to the works that study the regression problem without regularization [2], [8] or regularization in the asymptotic high-dimensional regime [3], [6], in terms of expectation over the regressor distribution [5], here we provide high-probability bounds that consider the presence of both fake features and regularization.*

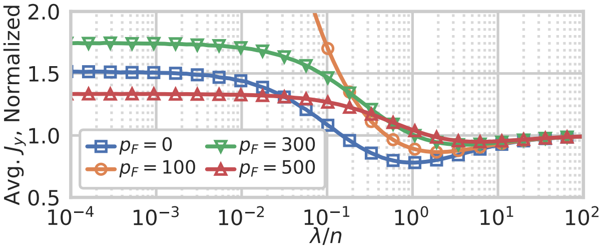

Fig. 1: The empirical average of the generalization error J_y versus the ridge parameter λ.

IV. Numerical Results

A. Details of the Numerical Simulations

In the following simulations we compute empirical averages for the generalization error. We now describe how we obtain these averages for a given set of problem dimensions n, p_F, p_S and p_C, the fixed power ratio coefficient r_S, and the noise level σ_v^2, and total power P of the underlying unknowns.

We generate the underlying unknowns $\boldsymbol{x}_S$ and $\boldsymbol{x}_C$ as $\boldsymbol{x}_S = \sqrt{r_S \frac{P}{p_S}} \mathbf{1} \in \mathbb{R}^{p_S \times 1}$, $\boldsymbol{x}_C = \sqrt{(1 - r_S)\frac{P}{p_C}} \mathbf{1} \in \mathbb{R}^{p_C \times 1}$, where $\mathbf{1}$ denotes a vector of ones with appropriate dimensions. For the test data, we have $n_{test} = 20000$ samples. We generate $M = 100$ realizations of the training feature matrices $\boldsymbol{A}_F$, $\boldsymbol{A}_S$, $\boldsymbol{A}_C$, as well as corresponding test feature matrices $\boldsymbol{A}_{F,test} \in \mathbb{R}^{n_{test} \times p_F}$, $\boldsymbol{A}_{S,test} \in \mathbb{R}^{n_{test} \times p_S}$ and $\boldsymbol{A}_{C,test} \in \mathbb{R}^{n_{test} \times p_C}$. The feature matrices are all i.i.d. standard Gaussian matrices. For each of these M sets we generate M noise vectors $\boldsymbol{v} \in \mathbb{R}^{n \times 1}$ and $\boldsymbol{v}_{test} \in \mathbb{R}^{n_{test} \times 1}$, with standard Gaussian entries, scaled with σ_v. We generate the corresponding training and test data as $\boldsymbol{y} = \boldsymbol{A}_S \boldsymbol{x}_S + \boldsymbol{A}_C \boldsymbol{x}_C + \boldsymbol{v}$, $\boldsymbol{y}_{test} = \boldsymbol{A}_{S,test} \boldsymbol{x}_S + \boldsymbol{A}_{C,test} \boldsymbol{x}_C + \boldsymbol{v}_{test}$. We then compute $\hat{\bar{\boldsymbol{x}}}$ as the solution to (9), i.e., $\hat{\bar{\boldsymbol{x}}} = \bar{\boldsymbol{A}}^{\mathrm{T}}(\bar{\boldsymbol{A}}\bar{\boldsymbol{A}}^{\mathrm{T}} + \lambda \boldsymbol{I}_n)^{+}\boldsymbol{y}$. which corresponds to the minimum-norm solution for $\lambda = 0$. The predictions of the test data is computed as $\hat{\boldsymbol{y}}_{test} = \boldsymbol{A}_{F,test} \hat{\boldsymbol{x}}_F + \boldsymbol{A}_{S,test} \hat{\boldsymbol{x}}_S$, and the corresponding error instance as $J_y = \|\boldsymbol{y}_{test} - \hat{\boldsymbol{y}}_{test}\|^2 - \sigma_v^2$, which is then averaged over the M sets of noise vectors, and then as well over the M sets of feature matrices. We have $n = 200$, and the number of included and missing features is $p_S = p_C = 100$, $\sigma_v = 10$, the signal power in $\tilde{\boldsymbol{x}}$ is $\|\tilde{\boldsymbol{x}}\|^2 = 200$, and ratio of the power in the included underlying unknowns $\boldsymbol{x}_S$ is $\frac{\|\boldsymbol{x}_S\|^2}{\|\tilde{\boldsymbol{x}}\|^2} = r_S = 0.5$.

B. Trade-offs between the regularization parameter λ and the number of fake features

We investigate the effect that ridge regularization has on the problem under the presence of the fake features in $\boldsymbol{A}_F$ by plotting the average generalization error J_y in Figure 1 and 2, obtained via simulation of the problem in (12). In Figure 1, we plot the empirical average generalization error versus the ridge parameter λ, for varying number of fake features p_F. In Figure 2 we plot the error versus p_F, for varying values of λ. The shaded areas in Figure 2 indicate the standard deviations.

These figures support the following conclusions: *i)* It is possible to decrease the error by increasing the number of fake

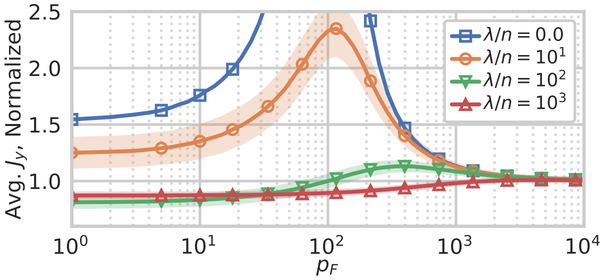

Fig. 2: The empirical average of the generalization error J_y (solid lines) $+/-$ one standard deviation (shaded areas) versus the number of fake features p_F.

features. *ii)* The best choice of λ depends on the number of fake features in the model. The effect of fake features can be interpreted as the fake features providing implicit regularization to the problem; and that the regularization provided by fake features can be used to compensate for low levels of λ, i.e., low levels of explicit regularization, for some scenarios.

For *i)*, see Figure 2 for small values of λ, i.e., $\lambda/n \in \{0, 10\}$: Here the lowest error over all p_F is achieved by increasing p_F to $p_F > 10^3$, rather than having p_F small. Hence, having a large number of fake features p_F can compensate for having a small ridge parameter λ by providing implicit regularization to the problem. For *ii)*, the plot in Figure 1 illustrates that the best choices of λ, i.e., the locations of the local minima of the respective curves, increase as the number of fake features p_F is increased, as well the values of these minima. Hence, if the explicit regularization parameter λ is large enough, then the problem without fake features has enough regularization, and the smallest possible number of fake features p_F gives the lowest error. These observations illustrate that a large number of fake features may compensate for having too little explicit regularization, but if there is enough explicit regularization, then a higher number of fake features may increase the error.

As shown in Theorem 1, λ should be large enough in order to bound the generalization error J_y around the interpolation threshold with high probability. We observe this effect in Figure 2, where the large enough values of λ, i.e., $\lambda/n \in \{10^2, 10^3\}$, dampen the peak in error around the interpolation threshold, that is otherwise seen for the smaller values of λ. Furthermore, we note that if $\lambda = 0$, then the standard deviation is extremely large around the interpolation threshold of $p_F = 100$, and if λ increases, then the standard deviation decreases. In general, higher values of λ decrease the standard deviation, i.e., the variation around the mean value. Similarly, increasing p_F decreases the variance, e.g., compare $\lambda/n = 0$ curve for $p_F \approx 0$ and $\approx 10^3$. This again suggests that p_F can have a regularizing effect, similar to the ridge parameter λ.

V. Conclusions

We provide a non-asymptotic high-probability bound for the generalization error of the ridge regression solution when an arbitrary number of fake features are present. This result reveals analytical insights on the interplay between the implicit regularization provided by the fake features and the explicit regularization provided by the ridge regularization.

We have considered linear models with isotropic Gaussian features. Extensions into non-linear models with more general feature covariance structures and other regularization frameworks are considered important research directions.

Appendix

A. Proof of Theorem 1

With $\hat{\hat{\boldsymbol{x}}} = \bar{\boldsymbol{W}}\boldsymbol{y}$, we denote the estimator as

$$\bar{\boldsymbol{W}} = \bar{\boldsymbol{A}}^{\mathrm{T}}(\bar{\boldsymbol{A}}\bar{\boldsymbol{A}}^{\mathrm{T}} + \lambda \boldsymbol{I}_n)^{-1}. \tag{25}$$

We denote the full singular value decomposition of $\bar{\boldsymbol{A}}$ by

$$\bar{\boldsymbol{A}} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}}, \tag{26}$$

where $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ and $\boldsymbol{V} \in \mathbb{R}^{\bar{p}\times\bar{p}}$ are orthogonal matrices, and the diagonal matrix $\boldsymbol{S} \in \mathbb{R}^{n\times\bar{p}}$ contains the singular values s_i of $\bar{\boldsymbol{A}}$, $i = 1, \ldots, \min(n, \bar{p})$. We let $s_i = 0$ if $i > \min(n, \bar{p})$.

From (20), we have

$$J_y = \left\| \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{x}_S \end{bmatrix} - \begin{bmatrix} \hat{\boldsymbol{x}}_F \\ \hat{\boldsymbol{x}}_S \end{bmatrix} \right\|^2 + \|\boldsymbol{x}_C\|^2 + \sigma_v^2 \tag{27}$$

$$= \left\| \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{x}_S \end{bmatrix} - \bar{\boldsymbol{W}}\left(\boldsymbol{A}_S\boldsymbol{x}_S + \boldsymbol{A}_C\boldsymbol{x}_C + \boldsymbol{v}\right) \right\|^2 + \|\boldsymbol{x}_C\|^2 + \sigma_v^2 \tag{28}$$

$$= \left\| \left(\boldsymbol{I}_{\bar{p}} - \bar{\boldsymbol{W}}\bar{\boldsymbol{A}}\right) \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{x}_S \end{bmatrix} - \bar{\boldsymbol{W}}\boldsymbol{z} \right\|^2 + \omega_z^2. \tag{29}$$

Here we introduced the vector $\boldsymbol{z} = \boldsymbol{A}_C\boldsymbol{x}_C + \boldsymbol{v}^{\mathrm{T}} \in \mathbb{R}^{n\times 1}$, with the entries $[z_1, \cdots, z_n]$ which are i.i.d. random variables with $\mathcal{N}(0, \omega_z^2)$, and $\omega_z^2 = \|\boldsymbol{x}_C\|^2 + \sigma_v^2$.

Using the triangle inequality (for two vectors $\boldsymbol{v}$, $\boldsymbol{w}$, $\|\boldsymbol{v} - \boldsymbol{w}\|^2 \leq 2\|\boldsymbol{v}\|^2 + 2\|\boldsymbol{w}\|^2$), as well as the submultiplicativity of the ℓ_2-norm, we have $J_y \leq 2\left\|\boldsymbol{I}_{\bar{p}} - \bar{\boldsymbol{W}}\bar{\boldsymbol{A}}\right\|^2 \|\boldsymbol{x}_S\|^2 + 2\left\|\bar{\boldsymbol{W}}\boldsymbol{z}\right\|^2 + \omega_z^2$. We continue by plugging in (25) and (26),

$$\begin{aligned} J_y \leq\ & 2\left\|\boldsymbol{I}_{\bar{p}} - \boldsymbol{S}^{\mathrm{T}}(\boldsymbol{S}\boldsymbol{S}^{\mathrm{T}} + \lambda\boldsymbol{I}_n)^{-1}\boldsymbol{S}\right\|^2 \|\boldsymbol{x}_S\|^2 \\ & + 2\left\|\boldsymbol{S}^{\mathrm{T}}(\boldsymbol{S}\boldsymbol{S}^{\mathrm{T}} + \lambda\boldsymbol{I}_n)^{-1}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{z}\right\|^2 + \omega_z^2 \end{aligned} \tag{30}$$

$$\begin{aligned} \sim\ & 2\left\|\boldsymbol{I}_{\bar{p}} - \boldsymbol{S}^{\mathrm{T}}(\boldsymbol{S}\boldsymbol{S}^{\mathrm{T}} + \lambda\boldsymbol{I}_n)^{-1}\boldsymbol{S}\right\|^2 \|\boldsymbol{x}_S\|^2 \\ & + 2\left\|\boldsymbol{S}^{\mathrm{T}}(\boldsymbol{S}\boldsymbol{S}^{\mathrm{T}} + \lambda\boldsymbol{I}_n)^{-1}\boldsymbol{z}\right\|^2 + \omega_z^2 \end{aligned} \tag{31}$$

where we used the unitary invariance of the norm, and $\boldsymbol{V}^{\mathrm{T}}\boldsymbol{z} \sim \boldsymbol{z}$ due to the rotational invariance of the distribution of $\boldsymbol{z}$.

We continue by utilizing the diagonal structure of $\boldsymbol{S}$,

$$\begin{aligned} J_y \leq\ & 2\left\|\boldsymbol{I}_{\bar{p}} - \mathrm{diag}\left(\tfrac{s_i^2}{s_i^2+\lambda}\right)\right\|^2 \|\boldsymbol{x}_S\|^2 \\ & + 2\left\|\left[\tfrac{s_1}{s_1^2+\lambda}z_1, \cdots, \tfrac{s_{r_{\min}}}{s_{r_{\min}}^2+\lambda}z_{r_{\min}}\right]^{\mathrm{T}}\right\|^2 + \omega_z^2 \end{aligned} \tag{32}$$

$$= 2\left\|\mathrm{diag}\left(\tfrac{\lambda^2}{(s_i^2+\lambda)^2}\right)\right\| \|\boldsymbol{x}_S\|^2 + 2\sum_{i=1}^{r_{\min}} g_i\, z_i^2 + \omega_z^2 \tag{33}$$

where $i = 1, \ldots, \bar{p}$ in the first term and $s_i = 0$ if $i > r_{\min} \triangleq \min(n, \bar{p})$, and where we have introduced the coefficients

$$g_i = \tfrac{s_i^2}{(s_i^2+\lambda)^2},\ \ i = 1, \ldots, r_{\min}. \tag{34}$$

We now focus on the second term of (33). The following corollary can be derived from [18, Lemma 1]:

Corollary 1. *Let z_i, $i = 1, \ldots, r$, be i.i.d. with $z_i \sim \mathcal{N}(0, \omega_z^2)$, and let $\boldsymbol{g} = [g_1, \cdots, g_r]^{\mathrm{T}} \in \mathbb{R}^{r \times 1}$, with $g_i > 0$, $\forall i$, and $t > 0$. Consider the event $E = \left\{\sum_{i=1}^{r} g_i z_i^2 < \omega_z^2 \left(\sum_{i=1}^{r} g_i + 2\|\boldsymbol{g}\|\sqrt{t} + 2\|\boldsymbol{g}\|_\infty t\right)\right\}$, where $\|\boldsymbol{g}\|_\infty = \sup_{i=1,\ldots,r} g_i$. Then, $\mathbb{P}(E) \geq 1 - e^{-t}$.*

With $t_1 > 0$ and g_i and z_i as in (33), we denote the event

$$E_1 = \left\{\sum_{i=1}^{r_{\min}} g_i z_i^2 < \omega_z^2 \left(\sum_{i=1}^{r_{\min}} g_i + 2\|\boldsymbol{g}\|\sqrt{t_1} + 2\|\boldsymbol{g}\|_\infty t_1\right)\right\}, \quad (35)$$

where $\boldsymbol{g} = [g_1, \cdots, g_{r_{\min}}]^{\mathrm{T}} \in \mathbb{R}^{r_{\min} \times 1}$, and from Corollary 1, we have that

$$\mathbb{P}(E_1) > 1 - e^{-t_1}. \quad (36)$$

We note that the variables g_i in (34) are random over the singular values s_i of $\bar{\boldsymbol{A}} \in \mathbb{R}^{n \times \bar{p}}$, and we continue by upper bound these g_i with a high-probability bound based on the distribution of s_i. We begin by noting that for each g_i,

$$g_i \leq \tfrac{s_{\max}^2}{(s_{\min}^2 + \lambda)^2}, i = 1, \ldots, r_{\min}. \quad (37)$$

We denote the event E_{2a} to bound the singular values as

$$E_{2a} = \left\{\sqrt{r_{\max}} - \sqrt{r_{\min}} - t_2 \leq s_{\min} \leq s_{\max} \leq \sqrt{n} + \sqrt{\bar{p}} + t_2\right\}, \quad (38)$$

where $s_{\min}$ and $s_{\max}$ denotes the smallest and the largest singular values of $\bar{\boldsymbol{A}}$, respectively, and $r_{\max} = \max(n, \bar{p})$, and $r_{\min} = \min(n, \bar{p})$, as defined previously. Using [19, eqn. (2.3)], we have that for any $t_2 \geq 0$,

$$\mathbb{P}(E_{2a}) \geq 1 - 2e^{-t_2^2/2}. \quad (39)$$

We will use this probability bound later in the proof to find the desired probability bound on J_y.

We now define f_g by plugging in the lower and upper bounds of (38) into the bound in (37),

$$f_g = \tfrac{(\sqrt{n} + \sqrt{\bar{p}} + t_2)^2}{((\sqrt{r_{\max}} - \sqrt{r_{\min}} - t_2)^2 + \lambda)^2}. \quad (40)$$

We now define the event E_2 using (37) and (40),

$$E_2 = \{g_i \leq f_g\}, \quad (41)$$

where $E_{2a} \Rightarrow E_2$. We combining the events E_1 in (35) and E_2 in (41), to obtain the event E_3 as

$$E_3 = \left\{\sum_{i=1}^{r_{\min}} g_i z_i^2 < \omega_z^2 f_g \left(r_{\min} + 2\sqrt{r_{\min} t_1} + 2t_1\right)\right\}, \quad (42)$$

where $E_1 \cap E_2 \Rightarrow E_3$. We now continue with the leading term of (33), which is bounded as $\left\|\operatorname{diag}\left(\frac{\lambda^2}{(s_i^2 + \lambda)^2}\right)\right\| \leq \frac{\lambda^2}{(s_{\min}^2 + \lambda)^2}$. where $i = 1, \ldots, \bar{p}$. We recall that if $i > \min(n, \bar{p})$ then $s_i = 0$. Hence if $n < \bar{p}$, we define E_4 as follows

$$E_4 = \left\{\left\|\operatorname{diag}\left(\tfrac{\lambda^2}{(s_i^2 + \lambda)^2}\right)\right\| = 1\right\}. \quad (43)$$

If instead $n \geq \bar{p}$, then we define

$$E_4 = \left\{\left\|\operatorname{diag}\left(\tfrac{\lambda^2}{(s_i^2 + \lambda)^2}\right)\right\| \leq \tfrac{\lambda^2}{((\sqrt{n} - \sqrt{\bar{p}} - t_2)^2 + \lambda)^2}\right\} \quad (44)$$

and note that $E_{2a} \Rightarrow E_4$. We combine (43), (44) and (42) with the bound on J_y in (33) to obtain $E_5 = \Big\{J_y < \|\boldsymbol{x}_S\|^2 \bar{f}_g + \left(\|\boldsymbol{x}_C\|^2 + \sigma_v^2\right)\left(f_g\left(r_{\min} + 2\sqrt{r_{\min} t_1} + 2t_1\right) + 1\right)\Big\}$. with $t_1, t_2 \geq 0$, f_g as in (40), and where

$$\bar{f}_g = \begin{cases} \frac{\lambda^2}{((\sqrt{n} - \sqrt{\bar{p}} - t_2)^2 + \lambda)^2} & \text{if } n \geq \bar{p}, \quad (45a) \\ 1 & \text{if } n < \bar{p}. \quad (45b) \end{cases}$$

We note that i) E_1 is independent from E_2 and E_4, ii) $E_{2a} \Rightarrow E_2$ and $E_{2a} \Rightarrow E_4$, hence $E_{2a} \Rightarrow E_2 \cap E_4$, and if we denote $E_{24} = E_2 \cap E_4$, then by (39) we can write

$$\mathbb{P}(E_{24}) \geq \mathbb{P}(E_{2a}) \geq 1 - 2e^{-t_2^2/2}, \quad \mathbb{P}(E_{24}^c) \leq 2e^{-t_2^2/2}. \quad (46)$$

By (36) we have that $\mathbb{P}(E_1^c) \leq e^{-t_1}$. Furthermore, we have

$$\begin{aligned} \mathbb{P}(E_5) &\geq \mathbb{P}(E_3 \cap E_4) \geq \mathbb{P}(E_1 \cap E_2 \cap E_4) & (47) \\ &= \mathbb{P}(E_1 \cap E_{24}) = 1 - \mathbb{P}(E_1^c \cup E_{24}^c) & (48) \\ &\geq 1 - \mathbb{P}(E_1^c) - \mathbb{P}(E_{24}^c) \geq 1 - e^{-t_1} - 2e^{-t_2^2/2}, & (49) \end{aligned}$$

where we have used the union bound to obtain $\mathbb{P}(E_1^c \cup E_{24}^c) \leq \mathbb{P}(E_1^c) + \mathbb{P}(E_{24}^c) \leq e^{-t_1} + 2e^{-t_2^2/2}$. This concludes the proof.

REFERENCES

[1] M. Belkin, D. Hsu, S. Ma, and S. Mandal, "Reconciling modern machine-learning practice and the classical bias–variance trade-off," *Proc. of the Nat. Acad. of Sciences*, vol. 116, no. 32, 2019.

[2] M. Belkin, D. Hsu, and J. Xu, "Two models of double descent for weak features," *SIAM J. Math. Data Sci.*, vol. 2, no. 4, pp. 1167–1180, 2020.

[3] T. Hastie, A. Montanari, S. Rosset, and R. J. Tibshirani, "Surprises in high-dimensional ridgeless least squares interpolation," *arXiv:1903.08560*, Dec. 2020.

[4] L. Chen, Y. Min, M. Belkin, and A. Karbasi, "Multiple descent: Design your own generalization curve," in *Advances in Neural Information Processing Systems*, 2021.

[5] M. Hellkvist, A. Özçelikkale, and A. Ahlén, "Estimation under model misspecification with fake features," *IEEE Transactions on Signal Processing*, vol. 71, pp. 47–60, 2023.

[6] D. Kobak, J. Lomond, and B. Sanchez, "The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization," *Journal of Machine Learning Research*, vol. 21, no. 169, pp. 1–16, 2020.

[7] P. Rao, "Some notes on misspecification in multiple regressions," *The American Statistician*, vol. 25, no. 5, pp. 37–39, 1971.

[8] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler, "Benign overfitting in linear regression," *Proc. of the Nat. Acad. of Sciences*, vol. 117, no. 48, pp. 30 063–30 070, 2020.

[9] S. D'Ascoli, M. Refinetti, G. Biroli, and F. Krzakala, "Double trouble in double descent: Bias and variance(s) in the lazy regime," in *Proceedings of the 37th International Conference on Machine Learning*, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 2280–2290.

[10] D. Holzmüller, "On the universality of the double descent peak in ridgeless regression," in *Intern. Conf. on Learn. Representations*, 2021.

[11] Z. Liao, R. Couillet, and M. W. Mahoney, "A random matrix analysis of random fourier features: beyond the gaussian kernel, a precise phase transition, and the corresponding double descent," *J. of Stat. Mechanics: Theory and Experiment*, vol. 2021, no. 12, p. 124006, dec 2021.

[12] P. Nakkiran, P. Venkat, S. M. Kakade, and T. Ma, "Optimal regularization can mitigate double descent," in *Int. Conf. Learn. Representations*, 2021.

[13] C. M. Bishop, "Training with noise is equivalent to Tikhonov regularization," *Neural Computation*, vol. 7, no. 1, pp. 108–116, 1995.

[14] D. Lederman and J. Tabrikian, "Constrained MMSE estimator for distribution mismatch compensation," in *IEEE Workshop on Sens. Array and Multichannel Process.*, 2006, pp. 439–443.

[15] R. Mittelman and E. L. Miller, "Robust estimation of a random parameter in a Gaussian linear model with joint eigenvalue and elementwise covariance uncertainties," *IEEE Trans. on Signal Process.*, vol. 58, no. 3, pp. 1001–1011, 2010.

[16] X. Liu, D. Zachariah, and P. Stoica, "Robust prediction when features are missing," *IEEE Signal Process. Letters*, vol. 27, p. 720–724, 2020.

[17] A. Tsigler and P. L. Bartlett, "Benign overfitting in ridge regression," *arXiv:2009.14286*, 2020.

[18] B. Laurent and P. Massart, "Adaptive estimation of a quadratic functional by model selection," *Annals of Statistics*, vol. 28, pp. 1302–1338, 2000.

[19] M. Rudelson and R. Vershynin, "Non-asymptotic theory of random matrices: Extreme singular values," *Proc. of the Int. Congress of Mathematicians*, pp. 1576–1602, 2010.