Model Collapse Demystified: The Case of Regression

Anonymous Authors[1]

Abstract

In the era of large language and image generation models, the phenomenon of "model collapse" refers to the situation whereby as a model is trained recursively on data generated from previous generations of itself over time, its performance degrades until the model eventually becomes completely useless, i.e the model collapses. In this work, we study this phenomenon in the simplified setting of kernel regression and obtain results which show a clear crossover between where the model can cope with AI generated data, and a regime where the model's performance completely collapses. Under polynomial decaying spectral and source conditions, we obtain modified scaling laws which exhibit new crossover phenomena from fast to slow rates. We also propose a simple strategy based on adaptive regularization to mitigate model collapse. Our theoretical results are validated with experiments.

1. Introduction

Model collapse describes the situation where the performance of large language models (LLMs) or large image generators degrade as more and more AI-generated data becomes present in their training dataset (Shumailov et al., 2023). Indeed, in the early stages of the generative AI evolution (e.g the ChatGPT-xyz series of models), there is emerging evidence suggesting that retraining a generative AI model on its own outputs can lead to various anomalies in the model's later outputs.

This phenomenon has been particularly observed in LLMs, where retraining on their generated content introduces irreparable defects, resulting in what is known as "model collapse", the production of nonsensical or gibberish output (Shumailov et al., 2023; Bohacek & Farid, 2023). Though several recent works demonstrate facets of this phenomenon *empirically* in various settings (Hataya et al., 2023; Martínez et al., 2023a;b; Bohacek & Farid, 2023; Briesch et al., 2023; Guo et al., 2023), a theoretical understanding is still missing.

In this work, we initiate a theoretical study of model collapse in the setting of high-dimensional supervised-learning with kernel regression. Kernel methods are popular in machine learning because, despite their simplicity, they define a framework powerful enough to exhibit non-linear features, while remaining in the convex optimization domain. While popular in their own right, kernel methods have made a recent spectacular comeback as proxies for neural networks in different regimes (Belkin et al., 2018), for instance in the infinite-width limit (Neal, 1996; Williams, 1996; Jacot et al., 2018; Lee et al., 2018) or in the lazy regime of training (Chizat et al., 2019). Caponnetto & de Vito (2007) characterize the power-law generalization error of regularized least-squares kernel algorithms. The role of optimization can also be taken into account in this setting (Nitanda & Suzuki, 2021). In the nonparametric literature, for example Schmidt-Hieber (2017) and Suzuki (2019) derived the test error scaling of deep neural network in fitting certain target functions and Bordelon et al. (2020) analyze spectral dependence. More recently, scaling laws have been shown for kernel models under the Gaussian design, e.g. in Spigler et al. (2020); Cui et al. (2021; 2022) for regression and Cui et al. (2023) for classification. Rahimi & Recht (2008); Rudi & Rosasco (2017); Maloney et al. (2022) study scaling laws for the random feature model in the context of regression.

Summary of Main Contributions. We study the Gaussian design where the input x is sampled from a multivariate zero-mean Gaussian and labels y are determined by a linear ground truth function with independent label noise as $y = x^T w_0 + \epsilon$. At each generation step, an approximation to w_0 is learned from the data, and used to generate new, "fake"/synthetic labels for the next generation. Our main findings can be summarized as follows:

(1) Exact Characterization of Test Error. In Section 4 (Theorem 4.2), we obtain analytic formulae for test error under the influence of training data with fake / synthesized labels. For n-fold iteration of data-generation, this formula writes

$$E_{test} = E_{test}^{clean} + n \times Scaling, \tag{1}$$

where E_{test}^{clean} is the usual test error of the model trained on clean data (not AI-generated). The term $Scaling$ precisely highlights the effects of all the relevant problems parameters: feature covariance matrix, sample size, strength of data-generator, label noise level in clean data distribution, label noise level in fake data distribution, etc.

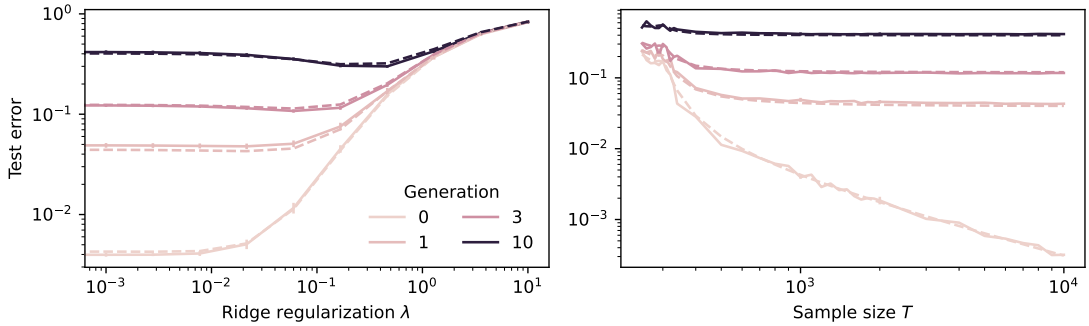


Figure 1. **Demystifying model collapse in ridge regression (isotropic covariance spectrum).** We show the evolution of test error for different sample size (T), different levels of ridge-regularization (λ), and training data from different generations (n) of fake data. The setup is: $\Sigma = I_d$ (i.e isotropic features), $d = 300$, $T_0 = 600$ (sample size on which each generation of the fake data generator is built), $\sigma = 0.1$ (label noise in true data distribution), and $\sigma_0 = 0.2$ (label noise for each generation of the fake data generator). **Left plot** is for $T = 1000$ and different values of λ. Notice the U-shape of the curves for large values of n, indicating the existence of a sweet spot (optimal regularization parameter). **Right plot** is for $\lambda = 10^{-3}$ and different values of T. Error bars correspond to uncertainty induced by the data-generating process, over different runs. The broken lines correspond to the theoretical result established in Theorem 4.1.

A direct consequence of this multiplicative degradation is that as the number of generations becomes large, the effect of re-synthesizing will make learning impossible.

(2) Modified Scaling Laws. In the case of power-law spectra, which is ubiquitous in machine learning Caponnetto & de Vito (2007); Spigler et al. (2020); Cui et al. (2022); Liang & Rakhlin (2020), we obtain in Section 5 (see Theorem 5.2) precise scaling laws which clearly highlight quantitatively the negative effect of training on fake data.

Further exploiting our analytic estimates, we obtain (Corollary 5.3) the optimal ridge regularization parameter as a function of all the problem parameters (sample size, spectral exponents, strength of fake data-generator, etc.). This new regularization parameter corresponds to a correction of the the value proposed in the classical theory on clean data Cui et al. (2022), and highlights a novel crossover phenomenon where for an appropriate tuning of the regularization parameter, the effect of training on fake data is a degradation of the fast error rate in the noiseless regime Cui et al. (2022); Caponnetto & de Vito (2007) to a much slower error rate which depends on the amount of true data on which the fake data-generator was trained in the first place. On the other hand, a choice of regularization which is optimal for the classical setting (training on real data), might lead to catastrophic failure: the test error diverges.

Apart from the above contributions, we hope the arguments and techniques used to derive our results will find broader use in the community. We have employed tools from Random Matrix Theory (RMT), pushing forward the envelope of prior work. We hope our technical contributions are of independent interest and lend themselves to generalization.

2. Review of Literature

Model Collapse. Current LLMs Devlin et al. (2018); Liu et al. (2019); Brown et al. (2020); Touvron et al. (2023), including GPT-4 Achiam et al. (2023), were trained on predominantly human-generated text; similarly, diffusion models like DALL-E Ramesh et al. (2021), Stable Diffusion Rombach et al. (2022), Midjourney Midjourney (2023) are trained on web-scale image datasets. Their training corpora already potentially exhaust all the available clean data on the internet. A growing number of synthetic data generated with these increasingly popular models starts to populate the web, often indistinguishable from "real" data. Recent works call attention to the potential dramatic deterioration in the resulting models, an effect referred to as *"model collapse"* Shumailov et al. (2023). Several recent works demonstrate facets of this phenomenon *empirically* in various settings Hataya et al. (2023); Martínez et al. (2023a;b); Bohacek & Farid (2023); Briesch et al. (2023); Guo et al. (2023). Theoretically, a few works are emerging to analyze the effect of iterative training on self-generated (or mixed) data. Taori & Hashimoto (2023) call attention to the bias amplification of iterative "data-feedback" loops, which has also been observed in the recommendations systems literature where feedback loops create echo chambers, and can be viewed as one form of model collapse. Shumailov et al. (2023) attribute model collapse to two mechanisms: a finite sampling bias leading to more and more peaked distributions and function approximation errors, and analyze the (single) Gaussian case. In the context of vision models, Alemohammad et al. (2023) analyze *"self-consuming loops"* by introducing a sampling bias that narrows the variance of the data at each generation, and provide theoretical analysis for the Gaussian model. Bertrand et al. (2023) explore scenarios involving a mix of clean data, representative of the true distribution,

and synthesized data from previous iterations of the generator. Their analysis reveals that if the data mix consists exclusively of synthesized data, the generative process is likely to degenerate over time, leading to what they describe as a 'clueless generator'. Conversely, they found that when the proportion of clean data in the mix is sufficiently high, the generator, under certain technical conditions, retains the capability to learn and accurately reflect the true data distribution. Note that such a compounding effect of synthesized data is already reminiscent of our decomposition (1).

Self-Distillation. Importantly, the fake data generation process which is responsible for model collapse should not be confused with self-distillation as formulated in Mobahi et al. (2020) for example. Unlike model collapse, the data generation process in self-distillation actually helps performance of the downstream model. The model has access to training labels from the true data distribution, but decides to fit a model on this data, and then use its outputs as the new labels, iterating this process possibly over severable steps. Thus, self-distillation has control over the data generating process, which is carefully optimized for the next stage training. In the setting of model collapse, there is no control over the data generation process, since it constitutes synthesized data which typically comes from the wide web.

Kernel Ridge Regression with Gaussian Design. This model has been studied by a vast body of works, in particular because it allows to analyze an important trade-off: the relative decay of the eigenvalues of the kernel (*capacity*) and the coefficients of the target function in feature space (*source*). Sizeable effort has been dedicated to characterize the influence on the decay rate of the test error as a function of these two relative decays (aka *power laws*) (Caponnetto & de Vito, 2007; Pillaud-Vivien et al., 2018; Berthier et al., 2020; Richards et al., 2021; Spigler et al., 2020; Cui et al., 2022; 2023). In Section 5 we extend these efforts, in particular based on works of Cui et al. (2021; 2022) which has given a full characterization of all regimes and test error decay that can be observed at the interplay of noise and regularization, characterizing a crossover transition of rates in the noisy setting. Our work uncovers fascinating new effects as a result of iterative training on synthetic data.

3. Theoretical Setup

We now present a setup which is simple enough to be analytically tractable, but rich enough to exhibit a wide range of regimes to illustrate a range of new phenomena that emerge with *model collapse*, described in Section 1 and Section 2.

Notations. This manuscript will make use of the following standard notations. The set of integers from 1 through d is denoted $[d]$. Given a variable z (which can be the input dimension d or the sample size T, etc.) the notation $f(z) \lesssim g(z)$ means that $f(z) \leq Cg(z)$ for sufficiently large z and an absolute constant C, while $f(z) \asymp g(z)$ means $f(z) \lesssim g(z) \lesssim f(z)$. For example, $1 + z^2 \asymp \max(1, z^2)$. Further, $f(z) \simeq g(z)$ means $f(z) = (1 + o(1))g(z)$, where $o(1)$ stands for a quantity which tends to zero in the limit $z \to \infty$. Finally, $\|u\|_A := \sqrt{u^\top A u}$ defines the Mahalanobis norm induced by a positive-definite matrix A. We denote with $X^\dagger$ the Moore-Penrose pseudo-inverse of X.

Parameters. d denotes the data dimension and n is the number of generations of the synthetic data generation process. Σ and σ pertain to the data generation, defined in (2. Parameters T and T_0 denote dataset size, and, together with generation-dependent label noise σ_m, linear labellers $\hat{w}_m$ and regularization λ_0 are defined in box (4).

Data Distribution & Fake Data-Generation Process. Consider the a distribution P_{Σ,w_0,σ^2} on $\mathbb{R}^d \times \mathbb{R}$ given by

$$\begin{aligned} \textbf{(Input)} \quad & x \sim N(0, \Sigma), \\ \textbf{(Noise)} \quad & \epsilon \sim N(0, \sigma^2), \text{ independent of } x \\ \textbf{(Output / Label)} \quad & y = x^\top w_0 + \epsilon. \end{aligned} \tag{2}$$

The positive integer d is the input-dimension, the vector $w_0 \in \mathbb{R}^d$ defines the ground-truth labelling function $x \mapsto x^\top w_0$, the matrix $\Sigma \in \mathbb{R}^{d\times d}$ is the covariance structure of the inputs. The scalar σ^2 is the level of label noise. To begin, we consider the linear case for clarity. We shall discuss the kernel setting at the end of this section. Thus, in classical linear regression, we are given a sample $(X, Y) \equiv \{(x_1, y_1), \ldots, (x_T, y_T)\}$ of size T from P_{Σ,w_0,σ^2} and we seek a linear model $\widehat{w}^{pred} \in \mathbb{R}^d$ with small test error $E_{test}(\widehat{w}^{pred})$, defined by

$$\begin{aligned} E_{test}(\widehat{w}^{pred}) &:= \mathbb{E}_{X,Y}\, \mathbb{E}_{x,y}[(x^\top \widehat{w}^{pred} - y)^2] - \sigma^2 \\ &= \mathbb{E}_{X,Y}[\|\widehat{w}^{pred} - w_0\|_\Sigma^2], \end{aligned} \tag{3}$$

where $(x, y) \sim P_{\Sigma,w_0,\sigma^2}$ is a random clean test point.

In our setup for studying model collapse, the design matrix $X \in \mathbb{R}^{T\times d}$ stays the same, but the vector of labels $Y \in \mathbb{R}^T$ is generated by an iterative relabelling process, where each generation of the model serves as the labeller for the data for the next generation.

The mental picture is as follows: each generation $\widehat{w}_m$ ($m \in \{1, 2, \ldots, n\}$) can be seen as a proxy for a specific version of ChatGPT, for example. The sample size T_0 used to create the fake labelling functions $\widehat{w}_m$ is a proxy for the strength of the fake data-generator thus constructed. Other works which have considered model collapse under such a self-looping training process include (Shumailov et al., 2023; Alemohammad et al., 2023; Bertrand et al., 2023).

This process is described below.

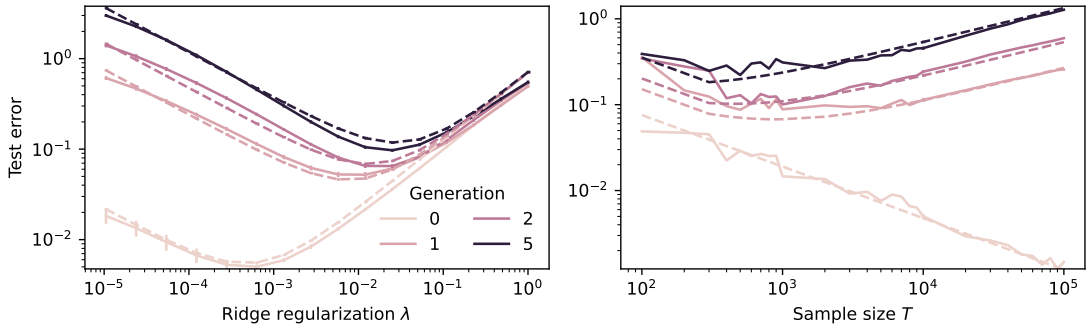


Figure 2. **Demystifying model collapse in ridge regression (power-law covariance spectrum)**. The setup is: $d = 300$, $\sigma = \sigma_0 = 1$, $\Sigma = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$, where $\lambda_k \propto k^{-2}$. **Left plot** corresponds to $T = 10,000$ and **Right plot** corresponds to adaptive regularization $\lambda = T^{-\ell_{crit}}$, where $\lambda = \lambda(T)$ as proposed in Cui et al. (2022). See Section 6 for details. The broken curves are as predicted by our Theorem 5.2. Though the exponent $\ell = \ell_{crit}$ is optimal in classical case, it isn't in the setup of model collapse. In fact here, the test error diverges with sample size T. Our theory proposes a corrected value of this exponent which gracefully adapts to synthesized data.

Construction of Fake / Synthesized Data Generator.

$$P_{\Sigma, w_0, \sigma^2} \to P_{\Sigma, \widehat{w}_1, \sigma_0^2} \to \dots \to P_{\Sigma, \widehat{w}_n, \sigma_{n-1}^2}, \qquad (4)$$

where $n \in \mathbb{N}$ is the number of generations and for each generation index $m \in \{1, 2, \dots, n\}$, the **fake data** labelling function $\widehat{w}_m$ is the least-squares estimator fitted on an iid dataset $\mathcal{D}_{m-1}$ of size T_0 from $P_{\Sigma, \widehat{w}_{m-1}, \sigma_0^2}$. In vector form

- $E_{m-1} \sim N(0, \sigma_{m-1}^2 I_{T_{m-1}})$,
- $\widehat{Y}_{m-1} = X_{m-1} \widehat{w}_{m-1} + E_{m-1}$,
- $\widehat{w}_m =$ ridge regression on (X_{m-1}, Y_{m-1}), with regularization parameter λ_0,

where X_{m-1} and E_{m-1} are independent, as well as independent of all previous generations.

The regularization parameter λ_0 of this synthetic data-generation process is as follows:

- If $n \geq 2$ (i.e more than one generation), we take $\lambda_0 = 0$, leading to ordinary least squares regression (OLS).
- If $n = 1$, we set $\lambda_0 \geq 0$ as arbitrary.

The above process is completely determined by the triplet (n, σ_0^2, T_0). We do not require $\sigma_0^2 = \sigma^2$. Note that the distribution $N(0, \Sigma)$ of the inputs stays the same all through the above process; only the conditional distribution of the labels is changed.

Finally, note that this setup is not self-distillation (as defined in Mobahi et al. (2020) for example), as the downstream model has no control whatsoever on the generation of the synthetic / fake labels.

The Downstream Model: Ridge Regression. For a regularization parameter $\lambda \geq 0$, let $\widehat{w}_n^{pred} = \widehat{w}_{n, T_0, \sigma_0^2, T, \lambda}^{pred} \in \mathbb{R}^d$ be the ridge predictor constructed from and iid sample $\{(x_1, y_1), \dots, (x_T, y_T)\}$ of size T from the n-fold fake data distribution $P_{\Sigma, \widehat{w}_n, \sigma_0^2}$, i.e

$$\widehat{w}_n^{pred} = \begin{cases} X^\dagger Y, & \text{if } \lambda = 0, \\ R X^\top Y / T, & \text{otherwise,} \end{cases} \qquad (5)$$

where $X = (x_1, \dots, x_T) \in \mathbb{R}^{T \times d}$ is the design matrix, $Y := (y_1, \dots, y_T) \in \mathbb{R}^T$ is the vector of labels, and $\widehat{\Sigma} := X^\top X / T \in \mathbb{R}^{d \times d}$ is the sample covariance matrix, and $R = R(\lambda) := (\widehat{\Sigma} + \lambda I_d)^{-1}$.

We are interested in the dynamics of the test error $E_{test}(\widehat{w}_n^{pred})$ (according to formula (3)) of this linear model. Note that the evaluation of the model is done on the true data distribution $P_{\Sigma, w_0, \sigma^2}$, even though the model is trained on the fake data distribution $P_{\Sigma, \widehat{w}_n, \sigma_0^2}$. Note that for $n = 0$, $E_{test}^{clean} := E_{test}(\widehat{w}_n^{pred})$ corresponds to the usual test error when the downstream model is trained on clean data.

A Note on Extension to Kernel Methods Though we present our results in the case of linear regression in $\mathbb{R}^d$ for clarity, they can be rewritten in equivalent form in the kernel setting. Indeed, as in Caponnetto & de Vito (2007); Simon et al. (2021); Cui et al. (2022); Liang & Rakhlin (2020), it suffices to replace x with a feature map induced by a kernel K, namely $\psi(x) := K_x \in \mathcal{H}_K$. Here, $\mathcal{H}_K$ is the reproducing kernel Hilbert space (RKHS) induced by K. In the data distribution (2), we must now replace the Gaussian marginal distribution condition $x \sim N(0, \Sigma)$ with $\psi(x) \sim N(0, \Sigma)$.

The ground-truth labeling linear function in (2) is now just a general function $f_0 \in L^2$. The ridge predictor (5) is then given by (*Representer Theorem*) $\widehat{f}_n^{pred}(x) := K(X,x)^\top \widehat{c}_n$, with $\widehat{c}_n = (G + \lambda T I_d)^{-1} Y \in \mathbb{R}^n$, where $K(X) := (K_{x_1}, \dots, K_{x_T})$, and $G = K(X)K(X)^\top = K(X,X) = (K(x_i,x_j))_{1\le i,j\le T} \in \mathbb{R}^{n\times n}$ is the Gram matrix.

4. Exact Test Error Characterization

In this section we establish generic analytic formulae for the test error of the downstream model $\widehat{w}_n^{pred}$ (5) trained on n-fold fake data-generation as outlined in Section 3. All proofs of this section are relegated to Appendix A.

4.1. Warm-up: Unregularized Case

For a start, let us first consider the case of ridgeless regression (corresponding to $\lambda = 0$ in Equation (5)), which amounts to OLS.

Theorem 4.1. *For an n-fold fake data generation process with $T_0 \ge d+2$ samples, the test error for the linear predictor $\widehat{w}_n^{pred}$ given in* (5) *learned on $T \ge d+2$ samples, with $\lambda = 0$ (i.e unregularized), is given by*

$$E_{test}(\widehat{w}_n^{pred}) \simeq \frac{\sigma^2\phi}{1-\phi} + \frac{n\sigma_0^2\phi_0}{1-\phi_0}, \text{ with } \phi = \frac{d}{T},\ \phi_0 = \frac{d}{T_0}.$$

The first term $E_{test}(\widehat{w}_0^{pred}) \simeq \sigma^2\phi/(1-\phi)$ corresponds to the usual error when the downstream model is fitted on clean data (see Hastie et al. (2022), for example). The additional term $n\sigma_0^2\phi_0/(1-\phi_0)$, proportional to the number of generations n, is responsible for model collapse.

Low-Dimensional Regime. In the low-dimensional problem (fixed d), Theorem 4.1 already predicts a change of scaling law from $E_{test}(\widehat{w}_0^{pred}) \asymp \sigma^2 T^{-1}$ to $E_{test}(\widehat{w}_n^{pred}) \asymp \sigma^2 T^{-1} + n\sigma_0^2 T_0^{-1}$. Thus, as the sample size T is scaled up, the test error eventually plateaus and does not vanish. This phenomenon is clearly visible in Figure 1. In Section 5, we shall establish a similar picture in a high-dimensional regime.

Regularize ? Note that the test error of the null predictor $w_{null} = 0$ is $E_{test}(w_{null}) = \|w_0\|_\Sigma^2$, and so

$$\frac{E_{test}(\widehat{w}_n^{pred})}{E_{test}(w_{null})} = \frac{1}{\mathrm{SNR}}\frac{\phi}{1-\phi} + \frac{n}{\mathrm{SNR}_0}\frac{\phi_0}{1-\phi_0},$$

where $\mathrm{SNR} := \|w_0\|_\Sigma^2/\sigma^2$ and $\mathrm{SNR}_0 := \|w_0\|_\Sigma^2/\sigma_0^2$. We deduce that if $n \gg \mathrm{SNR}_0/(1/\phi_0 - 1)$, then the learned model is already much worse than the null predictor! This suggests that a good strategy for mitigating the negative effects on learning on AI-generated data is regularization at an appropriate level, as illustrated in Figure 3.

4.2. Main Result I: A General Formula for Test Error

We now consider the case of general ridge penalty $\lambda > 0$, and drop the requirement $T \ge d+2$. We still make the simplifying assumption that $T_0 \ge d+2$, i.e each generation of the fake data-generator is based on enough examples from the previous generation.

Theorem 4.2. *For an n-fold fake data generation process, the test error of a ridge predictor $\widehat{w}_n^{pred}$ based on a sample of size $T \ge 1$ with regularization parameter λ is given by*

$$\begin{aligned} E_{test}(\widehat{w}_n^{pred}) &= \widetilde{Bias} + Var + \Delta_1 + \Delta_2 + n\sigma_0^2\rho, \\ \widetilde{Bias} &= \mathbb{E}\left[\|SRP_0w_0 - P_0w_0\|_\Sigma^2\right], \\ \Delta_1 &:= \mathbb{E}\left[\|\Delta w_0\|_\Sigma^2\right], \\ \Delta_2 &:= 2\mathbb{E}\left[\Delta w_0^\top \Sigma(I-SR)w_0\right], \\ \rho &:= \mathbb{E}\ \mathrm{tr}\, R_0SR\Sigma RS) \end{aligned} \tag{6}$$

where $P_0 := R_0S_0 \in \mathbb{R}^{d\times d}$ with $S_0 = X_0^\top X_0/T_0$ and $R_0 := \lim_{a\to\lambda_0}(S_0 + aI)^{-1}$; $\Delta w_0 := w_0 - P_0w_0 \in \mathbb{R}^d$ and Var is as given in formula (13)*.*

In particular, if $T_0 \ge d+2$ (under-parametrized fake data-generator) and $\lambda_0 = 0$, then $\Delta_1 = \Delta_2 = 0$, $\widetilde{Bias} = Bias$ (as given in formula (12)*), and*

$$\begin{aligned} E_{test}(\widehat{w}_n^{pred}) &\simeq E_{test}^{clean} + n\sigma_0^2\rho, \\ E_{test}^{clean} &= Bias + Var, \\ \rho &= \frac{1}{T_0 - d - 1}\mathbb{E}\,\mathrm{tr}(\Sigma^{-1}SR\Sigma SR). \end{aligned}$$

In the second part of the theorem, the term $E_{test}^{clean} := Bias + Var$ corresponds to the usual test error when the downstream model is trained on real (not fake) data, for which well-known formulae exist in a variety of scenarios (see Proposition 4.4). Even in this special case, what is new is the term $n\sigma_0^2\rho$, where n is the number of generations. This result is of the promised form (1), with $\Delta = \sigma_0^2\rho$. This additional term means that there is competition between usual test error E_{test}^{clean} and the additional term induced by the fake labeling process. Understanding the interaction of these two terms is key to demystifying the origins of model collapse.

Low-Dimensional Limit. Observe that if d is fixed and $T \to \infty$, then $S \to \Sigma$ (e.g. weakly), and so for $T_0 \ge d+2$, we have

$$\rho \simeq \frac{\mathrm{tr}\,\Sigma^2(\Sigma + \lambda I_d)^{-2}}{T_0 - d} = \frac{\mathrm{df}_2(\lambda)}{T_0 - d},$$

where for any $\lambda \ge 0$ and $m \in \mathbb{N}_\star$, the mth order "degrees of freedom" of the covariance matrix Σ is $\mathrm{df}_m(\lambda)$, defined by

$$\mathrm{df}_m(\lambda) = \mathrm{df}_m(\lambda;\Sigma) := \mathrm{tr}\,\Sigma^m(\Sigma + \lambda I_d)^{-m}.$$

Note that $\mathrm{df}_m(\lambda) \leq d$ always. In the high-dimensional setting (where d is no longer fixed), the precise analysis of ρ will be carried out via random matrix theory (RMT) tools.

4.3. High-Dimensional Regimes

In order to analyze the trace term ρ appearing in (6), we need some tools from RMT, and ultimately obtain analytic formulae for $E_{test}(\widehat{w}_n^{pred})$ in Theorem 4.2. Such tools have been used extensively to analyze anisotropic ridge regression (Richards et al., 2021; Hastie et al., 2022; Bach, 2023). A standard reference on RMT as a whole is (Bai & Silverstein, 2010).

Random Matrix Equivalents. For any sample size $T \geq 1$, define an increasing function $\lambda \to \kappa(\lambda, T)$ implicitly by

$$\kappa(\lambda, T) - \lambda = \kappa(\lambda, T) \cdot \mathrm{df}_1(\kappa(\lambda, T))/T\,. \tag{7}$$

The effect of ridge regularization at level $\lambda \geq 0$ is to improve the condition of the empirical covariance matrix S, what the κ-function does is translate this into regularization on Σ at level $\kappa(\lambda, T)$, so as control the capacity of the former, i.e the "effective dimension" of the underlying problem. Quantitatively, there is an equivalence of the form

$$\mathrm{df}_1(\lambda; \widehat{\Sigma}) \approx \mathrm{df}_1(\kappa(\lambda, T); \Sigma).$$

Roughly speaking, RMT is the business of formalizing such relationship and derivatives (w.r.t λ) thereof.

Example: Isotropic Covariance. For example, note that $\mathrm{df}_m(t) \equiv d/(1+t)^m$ (polynomial decay) in the isotropic case where $\Sigma = I_d$. Consequently, we have

$$\kappa(\lambda, T) - \lambda = \phi\kappa(\lambda, T)/(1 + \kappa(\lambda, T)), \text{ with } \phi := d/T.$$

In this case, it is easy to obtain the following well-known formula for $\kappa = \kappa(\lambda, T)$

$$\kappa = \frac{1}{2}\left(\lambda + \overline{\phi} + \sqrt{(\lambda + \overline{\phi})^2 + 4\lambda}\right), \text{ with } \overline{\phi} := \phi - 1, \tag{8}$$

which is reminiscent of the celebrated Marchenko-Pastur law (Marčenko & Pastur, 1967).

We will temporarily work under the following standard assumption

Assumption 4.3. $\Sigma = \Sigma_d$ (i.e a sequence of covariance matrix indexed by the dimensionality d) has spectrum bounded away from 0 uniformly w.r.t d. Moreover, the empirical spectral distribution $\sum_{j=1}^{d} \delta_{\lambda_j}$ of Σ converges in the limit $d \to \infty$, to a compactly-supported distribution μ on $\mathbb{R}^+$.

Furthermore, we shall work in the following so-called proportionate asymptotic scaling which is a standard analysis based on random matrix theory (RMT)

$$T, d \to \infty, \quad d/T \to \phi \in (0, \infty). \tag{9}$$

Later in Section 5 when we consider power-law spectra, this scaling will be extended to account for the more realistic case where d and T are allowed to be polynomial in one order,

Polynomial Scaling Regime.

$$T, d \to \infty, \quad d^{1/C} \lesssim T \lesssim d^C, \tag{10}$$

for some absolute constant $C \geq 1$.

Such non-proportionate settings are covered by the theory developed in Knowles & Yin (2017); Wei et al. (2022).

Bias-Variance Decomposition. With everything now in place, let us recall for later use, the following classical bias-variance decomposition for ridge regression (for example, see Richards et al. (2021); Hastie et al. (2022); Bach (2023))

Proposition 4.4. *Under Assumption 4.3, in the limit* (9) *the test error of a ridge predictor $w(\lambda)$ based on T iid samples from the true data distribution $P_{\Sigma, w_0, \sigma^2}$ is given by*

$$E_{test}(w(\lambda)) = \mathbb{E}\,\|w(\lambda) - w_0\|_\Sigma^2 \simeq Bias + Var, \tag{11}$$

$$Bias \simeq \frac{\kappa^2 w_0^\top \Sigma(\Sigma + \kappa I)^{-2} w_0}{1 - \mathrm{df}_2(\kappa)/T}, \tag{12}$$

$$Var \simeq \frac{\sigma^2 \mathrm{df}_2(\kappa)}{T} \cdot \frac{1}{1 - \mathrm{df}_2(\kappa)/T}, \tag{13}$$

where $\kappa = \kappa(\lambda, T)$ is as given in (7)*.*

In particular, in the isotropic case where $\Sigma = I_d$, we have

$$E_{test}(w(\lambda)) \simeq \frac{\kappa^2 \|w_0\|_2^2 + \sigma^2\phi}{(1+\kappa)^2 - \phi},$$

where $\kappa = \kappa(\lambda, T)$ is as given in (8)*.*

4.4. Main Result II: Analytic Formula for Test Error

The following result gives the test error for the downstream ridge predictor $\widehat{w}_n^{pred}$ defined in (5), in the context of fake training data, and will be heavily exploited later to obtain precise estimates in different regimes.

Theorem 4.5. *Suppose Assumption 4.3 is in order. For an n-fold fake data-generation process, the test error of a ridge predictor $\widehat{w}_n^{pred}$ based on a sample of size T with regularization parameter λ in the polynomial scaling regime $T, T_0, d \gg 1$ s.t. $\log T, \log T_0 \asymp \log d$, is given by*

$$E_{test}(\widehat{w}_n^{pred}) \simeq \widetilde{Bias} + Var + \Delta_1 + \Delta_2 + n\sigma_0^2\rho,$$

where Var is as given in (13) *and the other terms are de-*

fined by

$$\begin{aligned}
\widetilde{Bias} &= \kappa^2 \frac{w_0^\top \Sigma^3(\Sigma+\kappa_0 I)^{-2}(\Sigma+\kappa I)^{-1} w_0}{1-\mathrm{df}_2(\kappa)/T} \\
&\quad + \kappa^2 \frac{\kappa_0^2 w_0^\top \Sigma(\Sigma+\kappa_0 I)^{-2} w_0}{1-\mathrm{df}_2(\kappa)/T} \\
&\quad + \frac{\kappa^2 \,\mathrm{tr}\,\Sigma^2(\Sigma+\kappa_0 I)^{-2}(\Sigma+\kappa I)^{-1}}{T_0-\mathrm{df}_2(\kappa_0)}, \\
\Delta_1 &= \frac{\kappa_0^2 w_0^\top \Sigma(\Sigma+\kappa_0 I)^{-1} w_0}{1-\mathrm{df}_2(\kappa_0)/T_0}, \\
\Delta_2 &= \kappa_0 \kappa \cdot w_0^\top \Sigma(\Sigma+\kappa_0)^{-1}(\Sigma+\kappa I_d)^{-1} w_0, \\
\rho &= \frac{\mathrm{tr}\,\Sigma^4(\Sigma+\kappa_0 I)^{-2}(\Sigma+\kappa I)^{-2}}{T_0-\mathrm{df}_2(\kappa_0)} \\
&\quad + \frac{\kappa^2 \,\mathrm{tr}\,\Sigma^2(\Sigma+\kappa_0 I)^{-2}(\Sigma+\kappa I)^{-2}}{T_0-\mathrm{df}_2(\kappa_0)} \cdot \frac{\mathrm{df}_2(\kappa)}{T-\mathrm{df}_2(\kappa)},
\end{aligned}$$

with $\kappa = \kappa(\lambda, T)$ and $\kappa_0 := \kappa(0, T_0)$ are as given in (7)*.*

4.5. Under-Parametrized Fake Data-Generator

Observe that if $d \le T_0$, then $\kappa_0 = 0$ in Theorem 4.2, leading to $\widetilde{Bias} = Bias$ (given as in formula (12)), $\Delta_1 = \Delta_2 = 0$, and

$$\rho = \frac{\mathrm{df}_2(\kappa)}{T_0 - d} + \frac{\kappa^2 \,\mathrm{tr}\,(\Sigma+\kappa I)^{-2}}{T_0 - d} \frac{\mathrm{df}_2(\kappa)}{T - \mathrm{df}_2(\kappa)}, \tag{14}$$

where κ is as in Theorem 4.2. We deduce the following corollary.

Corollary 4.6. *Consider the setting of Theorem 4.2. If $\phi_0 \le 1$ additionally, then in the limit $T, T_0, d \to \infty$ such that $d/T \to \phi$ and $d/T_0 \to \phi_0$ with $\phi_0 \le 1$, it holds that*

$$E_{test}(\widehat{w}_n^{pred}) \simeq Bias + Var + n\sigma^2\rho,$$

where $Bias$ and Var are as given in formulae (12) *and* (13) *respectively, and ρ is as given in* (14)*.*

In the special case of isotropic features, it holds that

$$\begin{aligned}
Bias + Var &\simeq \frac{\kappa^2 \|w_0\|_2^2 + \sigma^2\phi}{(1+\kappa)^2 - \phi}, \\
\rho &\simeq \frac{\phi_0}{1-\phi_0}\left(\frac{1}{(1+\kappa)^2} + \frac{1}{(1+\kappa)^2}\frac{\phi\kappa^2}{(1+\kappa)^2-\phi}\right),
\end{aligned}$$

where $\kappa = \kappa(\lambda, T)$ is as given in (8)*.*

Such a result, empirically illustrated in Figures 1 and 2, gives us the needed analytical handle for understanding n-fold model collapse in terms of all problem hyperparameters (covariance spectrum, regularization, label-noise level, etc.).

5. The Case of Heavy Tails (Power Law)

Now, consider a variant of the distribution (2), in the setting considered in (Caponnetto & de Vito, 2007; Richards et al., 2021; Simon et al., 2021; Cui et al., 2022), for $d \to \infty$. Let

$$\Sigma = \lambda_1 v_1 v_1^\top + \ldots + \lambda_d v_d v_d^\top$$

be the spectral decomposition of Σ, with eigenvalues $\lambda_1 \ge \ldots \ge \lambda_d \ge 0$ and eigenvectors $v_1, \ldots, v_d \in \mathbb{R}^d$. Define coefficients $c_j := w_0^\top v_j$, i.e the projection of w_0 along the jth eigendirection of Σ.

We shall work under the following spectral conditions

$$\begin{aligned}
&\textbf{(Capacity Condition)}\ \lambda_j \asymp j^{-\beta} \text{ for all } j \in [d], \\
&\textbf{(Source Condition)}\ \|\Sigma^{1/2-r} w_0\| = O(1),
\end{aligned} \tag{15}$$

where $\beta > 1$ and $r \ge 0$. The parameter r measures the amount of dispersion of w_0 relative to the spectrum of Σ; a large value of r means w_0 is concentrated only along a few important eigendirections directions (i.e the learning problem is easy). For later convenience, define $\underline{r}$ and δ by

$$\underline{r} := \min(r, 1), \quad \delta := 1 + \beta(2r - 1).$$

As noted in (Cui et al., 2022), the above source condition is satisfied if $c_j \asymp j^{-\delta/2}$ for all $j \in [d]$.

As in (Cui et al., 2022), consider adaptive ridge regularization strength of the form

$$\lambda = \lambda(T) \asymp T^{-\ell}, \tag{16}$$

for fixed $\ell \ge 0$. The case where $\ell = 0$ corresponds to non-adaptive regularization; otherwise, the level of regularization decays polynomially with the sample size T. Define

$$\ell_{crit} := \beta/(1 + 2\beta\underline{r}), \tag{17}$$

In (Cui et al., 2022) KRR under normal circumstances (corresponding to $n = 0$, i.e no fake data) was considered and it was shown that this value for the regularization exponent in (16) is minimax-optimal for normal test error in the noisy regime, namely $E_{test}(\widehat{w}_0^{pred}) \asymp T^{-c}$, where

$$c := \frac{2\beta\underline{r}}{2\beta\underline{r} + 1} \in (0, 1).$$

This represents a crossover from the noiseless regime where it was shown that the test error scales like $E_{test}(\widehat{w}_0^{pred}) \asymp T^{-2\beta\underline{r}}$, a much faster rate. We shall show that the picture drastically changes in the context of training on fake data considered in this manuscript for the purpose of understanding model collapse (Shumailov et al., 2023).

Remark 5.1. Unlike Cui et al. (2022) which considered the proportionate scaling limit (9) for input dimension d and sample size T, we shall consider the more general (and more realistic) polynomial scaling limit (10), and invoke the tools of so-called *anisotropic local RMT* developed in Knowles & Yin (2017) to compute deterministic equivalents for quantities involving the spectra of random matrices.

5.1. Main Result III: A "Collapsed" Scaling Law

The following result shows that model collapse is a modification of usual scaling laws induced by fake data. All proofs of this section can be found in Appendix B.

Theorem 5.2. *Consider n-fold fake-data generation with sample size $T_0 \geq d+2$ and set $\phi_0 := d/T_0 \in (0,1)$. For a ridge predictor $\widehat{w}_n^{pred}$ given in* (5) *based on a fake data sample of size T, with regularization parameter $\lambda = \lambda(T)$ tuned adaptively as in* (16) *with exponent $0 \leq \ell < \beta$, the test error satisfies the following scaling law in the limit* (10)

$$E_{test}(\widehat{w}_n^{pred}) \asymp \max(\sigma^2, T^{1-2\underline{r}\ell-\ell/\beta}) \cdot T^{-(1-\ell/\beta)} + n\frac{\sigma_0^2}{1-\phi_0} \max\left(T/T_0, \phi_0\right) \cdot T^{-(1-\ell/\beta)}.$$

5.2. Optimal Regularization for Mitigating Collapse

Let us provide an instructive interpretation of the result.

Noiseless Regime. Suppose $\sigma = 0$ (or equivalently, exponentially small in T) and $\phi_0 \in (0,1)$ is fixed, and consider a number of generations n such that $n\sigma_0^2 \asymp T^a$, where $0 \leq a \leq 1 - \ell/\beta \leq 1$. Note that $a = 0$ corresponds to a constant number of generations. Also take $T_0 = T^b$, for some constant $b \in (0, \infty)$. According to Theorem 5.2, if we want to balance out the model-collapsing negative effect of training on fake data, we should chose ℓ so as to balance the second term $n(T/T_0)T^{-(1-\ell/\beta)} = T^{-(b-\ell/\beta-a)}$ and the first term $T^{-2\ell\underline{r}}$. This gives the following result.

Corollary 5.3. *In the setting of Theorem 5.2 with $T_0 \asymp T^b$ and $n \asymp T^b$, the optimal exponent of the ridge regularization parameter in* (16) *is $\ell = \ell_\star$, where*

$$\ell_\star = \min((b-a)\ell_{crit}, \beta), \tag{18}$$

and ℓ_{crit} is as in (17)*, with corresponding optimal test error*

$$\inf_{\ell \geq 0} E_{test}(\widehat{w}_n^{pred}) \asymp E_{test}(\widehat{w}_n^{pred})\big|_{\ell=\ell_\star} \asymp T^{-(b-a)c}.$$

Observe that when $(b-a)c < 2\beta\underline{r}$, which is the case when $n = O(1)$, $r \geq 1$ and $b \leq a+1$, this corresponds to the condition $T \gtrsim T_0$). The above result represents a crossover from the fast rate $E_{test}(\widehat{w}_0^{pred}) \asymp T^{-2\beta\underline{r}}$ in the case of training on clean data (Cui et al., 2022), to a much slower rate $E_{test}(\widehat{w}_n^{pred}) \asymp T^{-(b-a)c}$, attained by the adaptive regularization $\lambda \asymp T^{-\ell_\star}$, which is optimal in this setting. Furthermore, in this setting if we still use $\lambda \asymp T^{-\ell_{crit}}$ as proposed in (Cui et al., 2022) in the clean data setting, Corollary 5.3 predicts that $E_{test}(\widehat{w}_n^{pred}) \gtrsim T^{-(b-\ell_{crit}/\beta-a)} = T^{-(c+b-a-1)}$, which diverges to infinity if $b \geq a+1-c$. This is a catastrophic form of model collapse, and is empirically illustrated in Figures 2 and 3.

Noisy Regime. Now fix $\sigma^2 \neq 0$ and $\phi_0 \in (0,1)$. In this regime, Theorem 5.2 predicts that consistency (i.e $E_{test}(\widehat{w}_n^{pred}) \overset{T\to\infty}{\longrightarrow} 0$) is only possible if $\ell \leq \ell_\star$. First consider values of ℓ for which the variance is smaller than the bias, i.e i.e $0 \leq \ell \leq \ell_{crit}$. We get

$$E_{test}(\widehat{w}_n^{pred}) \asymp T^{-2\ell\underline{r}} + T^{-(b-a-\ell/\beta)},$$

which is minimized by taking $\ell = \min(\ell_\star, \ell_{crit})$. For other values of ℓ, the variance dominates and we have

$$E_{test}(\widehat{w}_n^{pred}) \asymp T^{-(1-\ell/\beta)} + T^{-(b-\ell/\beta-a)} \asymp T^{-(\gamma-\ell/\beta)},$$

where $\gamma := \min(1, b-a)$. This is minimized by taking $\ell = \ell_{crit}$, leading to $E_{test}(\widehat{w}_n^{pred}) \asymp T^{-(\gamma-1/(2\beta\underline{r}+1))}$. This tends to zero with $T \to \infty$ only if $b > a + 1/(2\beta\underline{r}+1)$.

6. Experiments

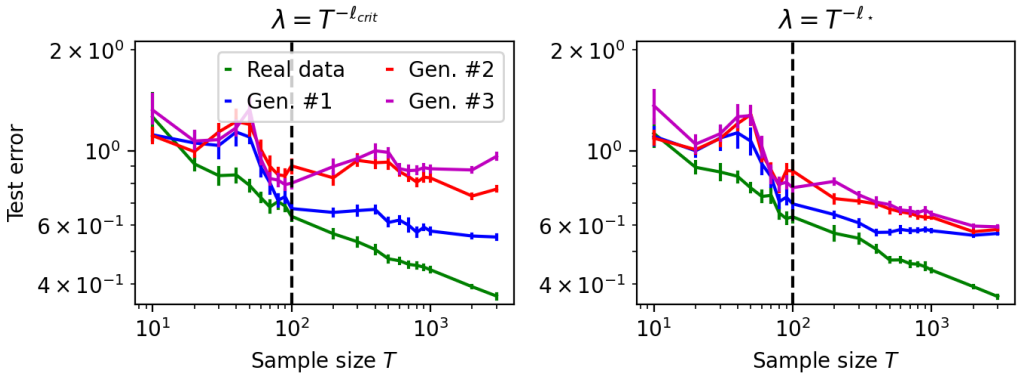
We performed the following experiments on both simulated and real data to empirically validate our theoretical results.

6.1. Simulated Data

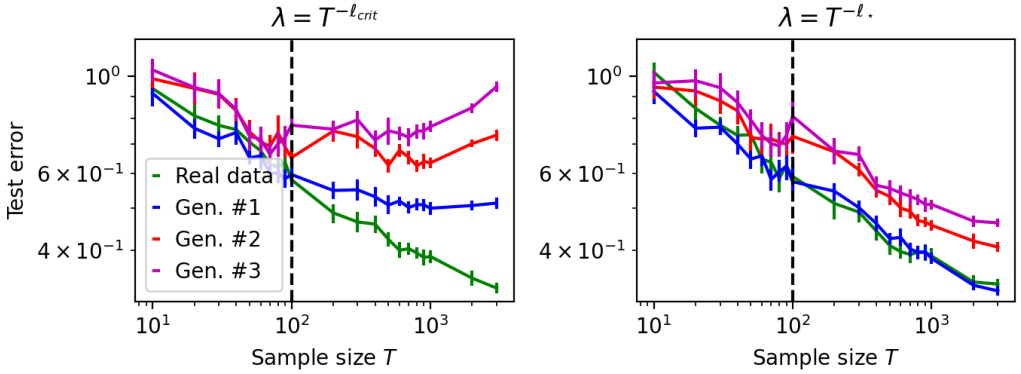
We consider ordinary / linear ridge regression in $\mathbb{R}^d$, for $d = 300$ and different structures for the covariance matrix Σ of the inputs: isotropic (i.e $\Sigma = I_d$) and power-law (15), with $(\beta, r) = (2, 0.375)$. For each value of n (the generation index), the fake data-generator is constructed according to the process described in (4), with $\sigma = 0.1$ and $\sigma_0 = 0.2$ (results are similar for other nonzero values of σ and σ_0). Then, for different values of T (between 1 and $1000,000$), a sample of size T is drawn from this fake data-generator and then a downstream ridge model (5) is fitted. The test set consists of $100,000$ clean pairs (x, y) form the true data distribution $P_{\Sigma, w_0, \sigma^2}$. This experiment is repeated 10 times to generate error bars. The results for the isotropic setting are shown in Figure 1 and the results for the power-law setting are shown in Figure 2.

6.2. Real Data: Kernel Ridge Regression on MNIST

As in Cui et al. (2022); Wei et al. (2022) we consider a distribution on MNIST, a popular dataset in the ML community. The classification dataset contains $60,000$ training and $10,000$ test data points (handwritten), with labels from 0 to 9 inclusive. Like in Cui et al. (2022), we convert the labels into real numbers (i.e a regression problem) as follows: $y =$ label mod $2 +$ noise , where the variance of the noise is $\sigma^2 = 1$ (for simplicity, we also set $\sigma_0^2 = 1$). The test set consists of $10,000$ pairs (x, y), with the labels y constructed as described in the previous sentence. The fake data used for training is generated as in the previous experiment, but via kernel ridge regression (instead of least squares) with the RBF kernel (bandwidth = 10^{-4}) and the polynomial kernel (degree = 5, bandwidth = 10^{-3}). Note that it was empirically shown in Cui et al. (2022) that these datasets verify (15) with $(\beta, r) \approx (1.65, 0.097)$ in the case of the



(a) RBF kernel (bandwidth = 10^{-4})



(b) Polynomial kernel (degree = 5, bandwidth = 10^{-3})

Figure 3. **Demystifying model collapse in kernel ridge regression on MNIST.** with adaptive regularization $T^{-\ell}$ for different values of the exponent $\ell \geq 0$ (see Section 6 for full experimental setup). **Top row:** RBF kernel. **Bottom row:** polynomial kernel. In each plot, we show test error curves as a function of sample size T, from different generations (n) of fake data. The broken vertical line corresponds to $T = T_0$, where T_0 is the number of samples (from the true data distribution) which was used to train the label faker. The value of the exponent regularization $\ell = \ell_\star$ (broken curves) is the optimal value in the presence of iterative data relabeling, while $\ell = \ell_{crit}$ (solid curves) corresponds to the optimal value without iterative re-labelling (i.e $n = 0$) proposed in Cui et al. (2022) (see (17)). Specifically, we take $\ell_\star = (b-a)\ell_{cirt} = b\ell_{crit}$, where $b = \log T_0 / \log T$ (so that $T_0 = T^b$), as proposed in Theorem 5.2, formula (18). Notice how the effect of fake data makes the test error become non decreasing in sample size T. This is effectively a collapse of the learned model.

aforementioned RBF kernel, and $(\beta, r) \approx (1.2, 0.15)$ in the case of the polynomial kernel. Then, for different values of T (between 1 and 1000), a sample of size T is drawn from this fake data-generator and then a downstream kernel ridge model is fitted. Each of these experiments are repeated 10 times to generate error bars (due to different realizations of label noise). The results are shown in Figure 3.

7. Concluding Remarks

As we navigate the "synthetic data age", our findings signal a departure from traditional test error rates (e.g neural scaling laws), introducing novel challenges and phenomena with the integration of synthetic data from preceding AI models into training sets. Our work provides a solid analytical handle for demystifying the model collapse phenomenon as a modification of usual scaling laws caused by fake / synthesized training data.

A direct consequence of our multiplicative degradation result is that, over time (i.e as the number of generations becomes large), the effect of large language models (like ChatGPT) in the wild will be a pollution of the web to the extent that learning will be impossible. This will likely increase the value and cost of clean / non-AI-generated data.

On the practical side, our analysis reveals that AI-generated data alters the optimal regularization for downstream models. Drawing from the insight that regularization mirrors early stopping (Ali et al., 2019), our study suggests that models trained on mixed real and AI-generated data may initially improve but later decline in performance (model collapse), necessitating early detection of this inflection point. This observation prompts a re-evaluation of current training approaches and underscores the complexity of model optimization in the era of synthetic data.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Alemohammad, S., Casco-Rodriguez, J., Luzi, L., Humayun, A. I., Babaei, H., LeJeune, D., Siahkoohi, A., and Baraniuk, R. G. Self-consuming generative models go mad. *arXiv preprint arxiv:2307.01850*, 2023.

Ali, A., Kolter, J. Z., and Tibshirani, R. J. A continuous-time view of early stopping for least squares regression. In Chaudhuri, K. and Sugiyama, M. (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1370–1378. PMLR, 16–18 Apr 2019.

Bach, F. High-dimensional analysis of double descent for linear regression with random projections, 2023.

Bai, Z. and Silverstein, J. W. J. W. *Spectral analysis of large dimensional random matrices*. Springer series in statistics. Springer, New York ;, 2nd ed. edition, 2010. ISBN 9781441906601.

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 of *Proceedings of Machine Learning Research*, pp. 541–549. PMLR, 2018.

Berthier, R., Bach, F. R., and Gaillard, P. Tight nonparametric convergence rates for stochastic gradient descent under the noiseless linear model. *CoRR*, abs/2006.08212, 2020. URL `https://arxiv.org/abs/2006.08212`.

Bertrand, Q., Bose, A. J., Duplessis, A., Jiralerspong, M., and Gidel, G. On the stability of iterative retraining of generative models on their own data. *arXiv preprint arxiv:2310.00429*, 2023.

Bohacek, M. and Farid, H. Nepotistically trained generative-ai models collapse, 2023.

Bordelon, B., Canatar, A., and Pehlevan, C. Spectrum dependent learning curves in kernel regression and wide neural networks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1024–1034. PMLR, 2020.

Briesch, M., Sobania, D., and Rothlauf, F. Large language models suffer from their own output: An analysis of the self-consuming training loop, 2023.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020.

Caponnetto, A. and de Vito, E. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

Cui, H., Loureiro, B., Krzakala, F., and Zdeborova, L. Generalization error rates in kernel regression: The crossover from the noiseless to noisy regime. In Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, 2021.

Cui, H., Loureiro, B., Krzakala, F., and Zdeborová, L. Generalization error rates in kernel regression: the crossover from the noiseless to noisy regime. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114004, nov 2022.

Cui, H., Loureiro, B., Krzakala, F., and Zdeborová, L. Error scaling laws for kernel classification under source and capacity conditions. *Machine Learning: Science and Technology*, 4(3):035033, August 2023. ISSN 2632-2153.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Guo, Y., Shang, G., Vazirgiannis, M., and Clavel, C. The curious decline of linguistic diversity: Training language models on synthetic text, 2023.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2), 2022.

Hataya, R., Bao, H., and Arai, H. Will large-scale generative models corrupt future datasets? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20555–20565, October 2023.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In Bengio, S., Wallach, H., Larochelle, H., Grauman,

K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Knowles, A. and Yin, J. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169(1): 257–352, 2017.

Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., and Sohl-Dickstein, J. Deep neural networks as gaussian processes. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Liang, T. and Rakhlin, A. Just interpolate: Kernel "Ridgeless" regression can generalize. *The Annals of Statistics*, 48(3), 2020.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Maloney, A., Roberts, D. A., and Sully, J. A solvable model of neural scaling laws, 2022.

Martínez, G., Watson, L., Reviriego, P., Hernández, J. A., Juarez, M., and Sarkar, R. Combining generative artificial intelligence (ai) and the internet: Heading towards evolution or degradation? *arXiv preprint arxiv: 2303.01255*, 2023a.

Martínez, G., Watson, L., Reviriego, P., Hernández, J. A., Juarez, M., and Sarkar, R. Towards understanding the interplay of generative artificial intelligence and the internet. *arXiv preprint arxiv: 2306.06130*, 2023b.

Marčenko, V. and Pastur, L. Distribution of eigenvalues for some sets of random matrices. *Math USSR Sb*, 1: 457–483, 01 1967.

Midjourney. Midjourney ai, 2023. URL `https://www.midjourney.com/`.

Mobahi, H., Farajtabar, M., and Bartlett, P. Self-distillation amplifies regularization in Hilbert space. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3351–3361. Curran Associates, Inc., 2020.

Neal, R. M. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pp. 29–53. Springer, New York, 1996.

Nitanda, A. and Suzuki, T. Optimal rates for averaged stochastic gradient descent under neural tangent kernel regime. In *International Conference on Learning Representations*, 2021.

Pillaud-Vivien, L., Rudi, A., and Bach, F. R. Statistical optimality of stochastic gradient descent on hard learning problems through multiple passes. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018*, pp. 8125–8135, 2018.

Rahimi, A. and Recht, B. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2008.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., and Sutskever, I. Zero-shot text-to-image generation. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8821–8831. PMLR, 18–24 Jul 2021.

Richards, D., Mourtada, J., and Rosasco, L. Asymptotics of ridge(less) regression under general source condition. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*. PMLR, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.

Rudi, A. and Rosasco, L. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*. Curran Associates Inc., 2017. ISBN 9781510860964.

Schmidt-Hieber, J. Nonparametric regression using deep neural networks with relu activation function. *Annals of Statistics*, 48, 08 2017.

Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., and Anderson, R. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arxiv:2305.17493*, 2023.

Simon, J. B., Dickens, M., and DeWeese, M. R. Neural tangent kernel eigenvalues accurately predict generalization. 2021.

Spigler, S., Geiger, M., and Wyart, M. Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12), 2020.

Suzuki, T. Adaptivity of deep reLU network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019.

Taori, R. and Hashimoto, T. B. Data feedback loops: model-driven amplification of dataset biases. ICML'23. JMLR.org, 2023.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Wei, A., Hu, W., and Steinhardt, J. More than a toy: Random matrix models predict how real-world neural representations generalize. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*. PMLR, 2022.

Williams, C. Computing with infinite networks. In Mozer, M., Jordan, M., and Petsche, T. (eds.), *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996.

A. Exact Characterization of Test Error Under Model Collapse

A.1. Proof of Theorem 4.1 (Rigeless Regression)

The proof is by induction on the number of generations n of fake data. For $n = 0$, we have

$$\begin{aligned} E_{test}(\widehat{w}_0^{pred}) &= \mathbb{E}\,\|\widehat{w}_0^{pred} - w_0\|_\Sigma^2 = \mathbb{E}\,\|\widehat{w}_0^{pred} - \widehat{w}_0\|_2^2 = \mathbb{E}\|(X_0^\top X_0)^{-1}X_0^\top E_{n-1}\|_2^2 \\ &= \sigma^2\mathbb{E}\mathrm{tr}((X_0^\top X_0)^{-1}) = \sigma^2\frac{d}{T-d-1} \simeq \frac{\sigma^2\phi}{1-\phi}, \end{aligned} \tag{19}$$

where $\phi := d/T \in (0,1)$ and the last step has made use of Lemma A.1 below. This is a well-known result for the test error of linear regression in the under-parametrized regime, without any AI pollution (fake / synthesized training data).

Analogously, for $n = 1$ one computes the test error after the first generation of fake data as follows

$$\begin{aligned} E_{test}(\widehat{w}_1^{pred}) &= \mathbb{E}\|\widehat{w}_1^{pred} - w_0\|_\Sigma^2 = \mathbb{E}\|\widehat{w}_1^{pred} - \widehat{w}_0\|_2^2 = \mathbb{E}\|\widehat{w}_1^{pred} - \widehat{w}_1 + \widehat{w}_1 - \widehat{w}_0\|_2^2 \\ &= \mathbb{E}\|(X_0^\top X_0)^{-1}X_0E_0 + \widehat{w}_0^{pred} - w_0\|_2^2 = \mathbb{E}\,\|w_0 - \widehat{w}_0^{pred}\|_2^2 + \mathbb{E}\,\|(X_0^\top X_0)^{-1}X_0^\top E_0\|_2^2 \\ &= E_{test}(\widehat{w}_0^{pred}) + \frac{\sigma_0^2 d}{T_0-d-1} \simeq \frac{\sigma^2\phi}{1-\phi} + \frac{\sigma_0^2\phi_0}{1-\phi_0}, \end{aligned}$$

where $\phi_0 = d/T_0 \in (0,1)$. Continuing the induction on n, we obtain the result. □

Lemma A.1. *Let X_0 be an $T_0 \times d$ random matrix with iid rows from $N(0,\Sigma)$. If $T_0 \geq d+1$, then the empirical covariance matrix $\widehat{\Sigma}_0 := X_0^\top X_0/T_0$ is invertible a.s and*

$$\mathbb{E}\,[\widehat{\Sigma}_0^{-1}] = \frac{T_0}{T_0-d-1}\Sigma^{-1} \simeq \frac{1}{1-\phi_0}\Sigma^{-1},$$

where $\phi_0 := d/T_0$.

A.2. Proof of Theorem 4.2 (Ridge Regression + General Covariance)

The case $n = 0$ is tautological, it corresponds to ridge regression on clean data. The proof for the cases $n \geq 1$ is by induction. Let X a random $T \times d$ matrix from $N(0,\Sigma)$ and let E be a T-dimensional Gaussian vector independent of X, with iid entries from $N(0,\sigma^2)$. The final predictor fitted on the nth generation of fake data is $\widehat{w}_n^{pred} = RX^\top Y/T \in \mathbb{R}^d$, where $S = \widehat{\Sigma} := X^\top X/T \in \mathbb{R}^{d\times d}$, $R = (S+\lambda I)^{-1} \in \mathbb{R}^{d\times d}$, and $Y = X\widehat{w}_n + E \in \mathbb{R}^T$, where

$$\widehat{w}_n := R_{n-1}X_{n-1}^\top\widehat{Y}_{n-1}/T_{n-1} = P_{n-1}\widehat{w}_{n-1} + R_{n-1}X_{n-1}^\top E_{n-1}/T_{n-1}, \tag{20}$$

where $P_m := R_mS_m$, and $\overline{E}_{n-1} := \sum_{i=0}^{n-1}E_i$. Thus, we have

$$\begin{aligned} \widehat{w}_n^{pred} &= RX^\top Y/T = RS\widehat{w}_{n-1} + RSR_{n-1}X_{n-1}^\top E_{n-1}/T_{n-1} + RX^\top E/T \\ &\overset{D}{=} u_{n-1} + RSR_{n-1}X_{n-1}^\top E_{n-1}/T_{n-1}, \end{aligned} \tag{21}$$

where $u_{n-1} := RSP_0w_0 + RX^\top E/T \in \mathbb{R}^d$. Since u_{n-1} and $RSX_0^\dagger\overline{E}_{n-1}$ are independent by construction, we deduce

$$\begin{aligned} E_{test}(\widehat{w}_n^{pred}) &= \mathbb{E}\,\|\widehat{w}_n^{pred} - w_0\|_\Sigma^2 = \mathbb{E}\,\|u_{n-1} - w_0\|_\Sigma^2 + \mathbb{E}\,\|RSR_{n-1}X_{n-1}^\top E_{n-1}/T_{n-1}\|_\Sigma^2 \\ &= \mathbb{E}\,\|u_{n-1} - P_{n-1}\widehat{w}_{n-1}\|_\Sigma^2 + \mathbb{E}\,\|\Delta w_{n-1}\|_\Sigma^2 + 2\mathbb{E}\,[\Delta w_{n-1}^\top\Sigma(I-RS)\widehat{w}_{n-1}] \\ &\quad + \mathbb{E}\,\|RSR_{n-1}X_{n-1}^\top E_{n-1}/T_{n-1}\|_\Sigma^2, \end{aligned} \tag{22}$$

and the result follows. □

A.3. Proof Of Theorem 4.5

Third Term. The 3rd term in (6) is nothing but the bias error term when we regress Y on X_0 via OLS, i.e

$$\mathbb{E}\,[\|\Delta w_0\|_\Sigma^2] = \mathbb{E}\,[\|P_0w_0 - w_0\|_\Sigma^2] \simeq \frac{\kappa_0^2 w_0^\top\Sigma(\Sigma+\kappa_0 I)^{-1}w_0}{1-\mathrm{df}_2(\kappa_0)/T_0}. \tag{23}$$

First Term. The 1st term $\widetilde{Bias}$ in (6) is just the bias term of the test error for ridge regression fitted on a clean dataset from the distribution $P_{\Sigma,P_0w_0,\sigma^2}$, and evaluated on the same distribution,

$$\widetilde{Bias} := \mathbb{E}\left[\|RSP_0w_0 - P_0w_0\|_\Sigma^2\right] \simeq \frac{\kappa^2\mathbb{E}\left[(P_0w_0)^\top\Sigma(\Sigma+\kappa I)^{-1}P_0w_0\right]}{1-\mathrm{df}_2(\kappa)/T}. \tag{24}$$

Now, define $d\times d$ positive semidefinite (psd) matrices $A := w_0w_0^\top$ and $B := \Sigma(\Sigma+\kappa I)^{-1}$. For $\lambda_0 > 0$, define $R_0 := (S_0+\lambda I_0)^{-1}$ where $S_0 := X_0^\top X_0/T_0$. Noting that $P_0 = \lim_{\lambda_0\to 0^+} S_0R_0$, the expectation of the RHS of the above display can be computed like so

$$\begin{aligned}
&\mathbb{E}\left[(P_0w_0)^\top\Sigma(\Sigma+\kappa I)^{-1}P_0w_0\right] = \lim_{\lambda_0\to 0^+}\mathbb{E}\left[\mathrm{tr}(AS_0R_0BS_0R_0)\right]\\
&\simeq \mathrm{tr}(A\Sigma(\Sigma+\kappa_0 I)^{-1}B\Sigma(\Sigma+\kappa_0 I)^{-1}) + \kappa_0^2\mathrm{tr}(A(\Sigma+\kappa_0 I)^{-2}\Sigma)\cdot\frac{\mathrm{tr}(B(\Sigma+\kappa_0 I)^{-2}\Sigma)}{T_0-\mathrm{df}_2(\kappa_0)}\\
&= w_0^\top\Sigma^3(\Sigma+\kappa_0 I)^{-2}(\Sigma+\kappa I)^{-1}w_0 + \kappa_0^2 w_0^\top\Sigma(\Sigma+\kappa_0 I)^{-2}w_0\cdot\frac{\mathrm{tr}(\Sigma^2(\Sigma+\kappa_0 I)^{-2}(\Sigma+\kappa I)^{-1})}{T_0-\mathrm{df}_2(\kappa_0)}.
\end{aligned}$$

Here are some interesting special cases.

- (Under-parametrized Faker) If $\phi_0 \le 1$, then $\kappa_0 = 0$ and so
$$\mathbb{E}\left[(P_0w_0)^\top\Sigma(\Sigma+\kappa I)^{-1}P_0w_0\right] \simeq w_0^\top\Sigma(\Sigma+\kappa I)^{-1}w_0,$$
which is the usual bias term in ridge regression on clean data.
- (Over-Parametrized Faker) If $\phi_0 > 1$, then κ_0 is defined implicitly by $\mathrm{df}_1(\kappa_0) = T_0$. For isotropic covariance, we have $\kappa_0 = \phi_0 - 1$ and so
$$\begin{aligned}
(1+\kappa)\times\mathbb{E}\left[(P_0w_0)^\top\Sigma(\Sigma+\kappa I)^{-1}P_0w_0\right] &\simeq \phi_0^{-2}\|w_0\|_2^2 + \frac{\phi_0^{-4}(\phi_0-1)^2\|w_0\|_2^2 d}{T_0 - d/\phi_0^2}\\
&= \phi_0^{-2}\|w_0\|_2^2\left(1+\frac{(\phi_0-1)^2 d}{\phi_0^2 T_0 - d}\right)\\
&\simeq \phi_0^{-2}\|w_0\|_2^2\left(1+\frac{(\phi_0-1)^2}{\phi_0-1}\right) = \frac{\|w_0\|_2^2}{\phi_0}.
\end{aligned} \tag{25}$$

Fourth Term. We can rewrite

$$\begin{aligned}
\Delta_3/2 &:= \mathbb{E}\left[\Delta w_0^\top\Sigma(I-RS)w_0\right]\\
&= w_0^\top\Sigma w_0 - \mathbb{E}\left[w_0^\top P_0\Sigma w_0\right] + \mathbb{E}\left[w_0^\top\Sigma RSw_0\right] - \mathbb{E}\left[w_0^\top\Sigma P_0RSw_0\right].
\end{aligned} \tag{26}$$

Thanks to Propositions 1 and 2 of Bach (2023), observe that for any "spectrally nice A", we have

$$\mathbb{E}\ \mathrm{tr}\,AP_0 \simeq \mathrm{tr}\,A\Sigma(\Sigma+\kappa_0 I)^{-1} \text{ and } \mathbb{E}\ \mathrm{tr}\,ARS \simeq \mathrm{tr}\,A(\Sigma+\kappa I)^{-1}\Sigma. \tag{27}$$

We deduce that

$$\begin{aligned}
\mathbb{E}\, w_0^\top P_0\Sigma w_0 = \mathbb{E}\,\Sigma w_0w_0^\top P_0 &\simeq \mathrm{tr}\,\Sigma w_0w_0^\top\Sigma(\Sigma+\kappa_0 I)^{-1}\\
&= w_0^\top\Sigma^2(\Sigma+\kappa_0 I)^{-1}w_0, &(28)\\
\mathbb{E}\, w_0^\top\Sigma RSw_0 = \mathbb{E}\ \mathrm{tr}\, w_0w_0^\top\Sigma RS &\simeq \mathrm{tr}\, w_0w_0^\top\Sigma(\Sigma+\kappa I)^{-1}\Sigma\\
&= w_0^\top\Sigma^2(\Sigma+\kappa I)^{-1}w_0, &(29)\\
\mathbb{E}\, w_0^\top P_0\Sigma RSw_0 = \mathbb{E}\ \mathrm{tr}\,\Sigma RSw_0w_0^\top P_0 &\simeq \mathbb{E}\ \mathrm{tr}\,\Sigma RSw_0w_0^\top\Sigma(\Sigma+\kappa_0 I)^{-1}\\
&= \mathbb{E}\ \mathrm{tr}\, w_0w_0^\top\Sigma(\Sigma+\kappa_0 I)^{-1}\Sigma RS\\
&\simeq \mathrm{tr}\, w_0w_0^\top\Sigma(\Sigma+\kappa_0 I)^{-1}\Sigma(\Sigma+\kappa I)^{-1}\Sigma\\
&= w_0^\top\Sigma(\Sigma+\kappa_0 I)^{-1}\Sigma^2(\Sigma+\kappa I)^{-1}w_0. &(30)
\end{aligned}$$

Observing that $I - \Sigma(\Sigma-zI)^{-1} \equiv z(\Sigma-zI)^{-1}$, we deduce that

$$\begin{aligned}
\mathbb{E}\,\Delta w_0^\top\Sigma(I-RS)w_0 &\simeq \kappa_0 w_0^\top\Sigma(\Sigma+\kappa_0 I)^{-1}\left(I-\Sigma(\Sigma+\kappa I)^{-1}\right)w_0\\
&= \kappa_0\kappa\cdot w_0^\top\Sigma(\Sigma+\kappa_0)^{-1}(\Sigma+\kappa I_d)^{-1}w_0.
\end{aligned} \tag{31}$$

Fifth Term. Define a $d \times d$ random psd matrix $H := SR\Sigma SR$. Observe that the 5th term in (6) then writes

$$\begin{aligned}\rho &= \mathbb{E}\left[\operatorname{tr} X_0^\dagger (X_0^\dagger)^\top H \mid H\right] = \mathbb{E}\left[\operatorname{tr} X_0^\top (X_0 X_0^\top)^{-2} X_0 H \mid H\right] \\ &= \lim_{\lambda_0 \to 0+} \frac{1}{T_0} \frac{\partial}{\partial \lambda_0} \mathbb{E}\left[\operatorname{tr} X_0^\top (X_0 X_0^\top + \lambda_0 T_0 I)^{-1} X_0 H \mid H\right].\end{aligned} \tag{32}$$

Now, one computes

$$\begin{aligned}\operatorname{tr} X_0^\top (X_0 X_0^\top + \lambda_0 T_0 I)^{-1} X_0 H &\stackrel{D}{=} \operatorname{tr} \Sigma^{1/2} Z_0^\top (Z_0 \Sigma Z_0^\top + \lambda_0 T_0 I)^{-1} Z_0 \Sigma^{1/2} H \\ &= \operatorname{tr} A Z_0^\top (Z_0 \Sigma Z_0^\top + \lambda_0 T_0 I)^{-1} Z_0\end{aligned}$$

where $A := \Sigma^{1/2} H \Sigma^{1/2}$. We deduce that

$$\mathbb{E}\left[\operatorname{tr} X_0^\top (X_0 X_0^\top + \lambda_0 T_0 I)^{-1} X_0 H \mid H\right] \simeq \operatorname{tr} A(\Sigma + \kappa(\lambda_0, T_0) I)^{-1} = \operatorname{tr} H(\Sigma + \kappa_0 I)^{-1}\Sigma, \tag{33}$$

where $\kappa_0 := \kappa(\lambda_0, T_0)$. Differentiating w.r.T λ_0 and letting this parameter tend to zero from above gives

$$\begin{aligned}\rho = \mathbb{E}\left[\operatorname{tr} X_0^\dagger (X_0^\dagger)^\top H \mid H\right] &= -\frac{1}{T_0} \frac{\partial}{\partial \lambda_0} \mathbb{E}\left[\operatorname{tr} X_0^\top (X_0 X_0^\top + \lambda_0 T_0 I)^{-1} X_0 H \mid H\right] \\ &\simeq -\frac{1}{T_0} \frac{\partial \kappa_0}{\partial \lambda_0} \operatorname{tr} H(\Sigma + \kappa_0 I)^{-1}\Sigma \\ &\simeq \frac{\operatorname{tr} H(\Sigma + \kappa_0 I)^{-2}\Sigma}{T_0 - \mathrm{df}_2(\kappa_0)},\end{aligned} \tag{34}$$

where we have made use of Lemma C.2. We deduce that

Proposition A.2. *In the limit $d, T, T_0 \to \infty$ such that $d/T \to \phi$ and $d/T_0 \to \phi_0$, it holds for any $\lambda > 0$ that*

$$\rho = \frac{\operatorname{tr}((\Sigma + \kappa_0 I)^{-2}\Sigma^2(\Sigma + \kappa I)^{-2}\Sigma^2)}{T_0 - \mathrm{df}_2(\kappa_0)} + \frac{\kappa^2 \operatorname{tr}((\Sigma + \kappa_0 I)^{-2}\Sigma^2(\Sigma + \kappa I)^{-2})}{T_0 - \mathrm{df}_2(\kappa_0)} \cdot \frac{\mathrm{df}_2(\kappa)}{T - \mathrm{df}_2(\kappa)}, \tag{35}$$

where $\kappa_0 := \kappa(\lambda_0, T_0)$ and $\kappa = \kappa(\lambda, T)$.

This result completes the proof of Theorem 4.2. □

We now proof the corollary. For the first part, we know from Theorem 4.2 that

$$E_{test}(\widehat{w}_n^{pred}) = E_{test}(\widehat{w}_0^{pred}) + n\sigma_0^2\rho, \text{ with} \tag{36}$$

$$\rho := \frac{\mathbb{E}\operatorname{tr}(\Sigma^{-1}S(S + \lambda I)^{-1}\Sigma(S + \lambda I)^{-1}S)}{T_0 - d}. \tag{37}$$

The $E_{test}(\widehat{w}_0^{pred})$ term is taken care of by Proposition 4.4, since this corresponds to generalization error on clean training data. For the ρ term, we use Proposition 1 of (Bach, 2023) with $A = \Sigma^{-1}$ and $B = \Sigma$ to get

$$\begin{aligned}\rho &\simeq \frac{\operatorname{tr}((\Sigma + \kappa I)^{-2}\Sigma^2)}{T_0 - d} + \frac{\kappa^2 \operatorname{tr}((\Sigma + \kappa I)^{-2})}{T_0 - d} \frac{\operatorname{tr}((\Sigma + \kappa I)^{-2}\Sigma^2)}{T - \mathrm{df}_2(\kappa)} \\ &= \frac{\mathrm{df}_2(\kappa)}{T_0 - d} + \frac{\kappa^2 \operatorname{tr}((\Sigma + \kappa I)^{-2})}{T_0 - d} \frac{\mathrm{df}_2(\kappa)}{T - \mathrm{df}_2(\kappa)},\end{aligned}$$

which proves the first part of the result.

For the second part, note that $\mathrm{df}_2(\kappa) = d/(1 + \kappa)^2$ when $\Sigma = I$, (8) holds, and so

$$\begin{aligned}(1 - 1/\phi_0)\rho &\simeq \frac{1}{(1 + \kappa)^2} + \frac{\kappa^2}{(1 + \kappa)^4} \frac{d}{T - d/(1 + \kappa)^2} \\ &\simeq \frac{1}{(1 + \kappa)^2} + \frac{\kappa^2}{(1 + \kappa)^4} \frac{\phi}{1 - \phi/(1 + \kappa)^2} \\ &= \frac{1}{(1 + \kappa)^2} + \frac{1}{(1 + \kappa)^2} \frac{\phi\kappa^2}{(1 + \kappa)^2 - \phi},\end{aligned}$$

and the result follows. □

A.4. A Note on Proposition 4.4

The second part of the result follows from the first as we now see. Indeed, $w_0^\top \Sigma(\Sigma+\kappa I)^{-2} w_0 = r^2/(1+\kappa)^2$, $\mathrm{df}_2(\kappa) = d/(1+\kappa)^2$ and so we deduce from the first part that

$$Var \simeq \sigma^2\phi \frac{1}{(1+\kappa)^2}\frac{1}{1-\phi/(1+\kappa)^2} = \frac{\sigma^2\phi}{(1+\kappa)^2-\phi},$$
$$Bias \simeq \kappa^2\|w_0\|_2^2 \frac{1}{(1+\kappa)^2}\frac{1}{1-\phi/(1+\kappa)^2} = \frac{\kappa^2\|w_0\|_2^2}{(1+\kappa)^2-\phi},$$

from which the result follows. □

B. Power-Law Regime

B.1. Proof of Theorem 5.2

Let us pretend that (14) continues to hold even though Assumption 4.3 is clearly violated. Then, we need to analyze the quantity

$$\rho \simeq \frac{\mathrm{df}_2(\kappa(\lambda))}{T_0-d} + \frac{\kappa(\lambda)^2 \mathrm{tr}\left((\Sigma+\kappa(\lambda)I_d)^{-2}\right)}{T_0-d}\cdot\frac{\mathrm{df}_2(\kappa(\lambda))}{T-\mathrm{df}_2(\kappa(\lambda))}. \tag{38}$$

Now, for small λ, $\kappa := \kappa(\lambda)$ is small and one can computes

$$\mathrm{df}_m(\kappa) \asymp \sum_i \frac{\lambda_i^m}{(\lambda_i+\kappa)^m} = \kappa^{-m}\sum_i \frac{\lambda_i^m}{(1+\kappa^{-1}\lambda_i)^m} \asymp \kappa^{-m}\kappa^{(m-1/\beta)} = \kappa^{-1/\beta}, \tag{39}$$

where we have used Lemma C.1 with $D=\kappa^{-1}$ and $n=m$ in the last step. On the other hand, we can use some the results of Appendix A (Section 3) of (Cui et al., 2022) to do the following. It can be shown (see aforementioned paper) that

- If $\ell > \beta$, then $\kappa \asymp T^{-\beta}$, and so $\mathrm{df}_m(\kappa) \asymp T$ for all $m \geq 1$.
- If $\ell < \beta$, then $\kappa \asymp \lambda \asymp T^{-\ell}$, and so $\mathrm{df}_m(\kappa) \asymp T^{\ell/\beta} = o(T)$ for all $m \geq 1$.

For $\ell < \beta$, plugging this into (14) gives

$$\begin{aligned}\rho &\asymp \frac{T^{\ell/\beta}}{T_0-d} + \frac{d}{T_0-d}\frac{T^{\ell/\beta}}{T-T^{\ell/\beta}} \asymp T_0^{-1}T^{\ell/\beta} + \frac{\phi_0}{1-\phi_0}T^{-(1-\ell/\beta)} \\ &\asymp \frac{1}{1-\phi_0}\max\left(T/T_0, \phi_0\right)T^{-(1-\ell/\beta)},\end{aligned} \tag{40}$$

where $\phi_0 := d/T_0$. Combining our Theorem 4.2 with (45), we get the claimed result. □

B.2. Representation of Clean Test Error

We make a small digression to present the following curiosity: with a slight leap of faith, the main results of (Cui et al., 2022) can be obtained in a few lines from the tools developed in (Bach, 2023), namely Proposition 4.4. This is significant, because the computations in (Cui et al., 2022) were done via methods of statistical physics (replica trick), while (Bach, 2023) is based on RMT.

Indeed, for regularization parameter $\lambda \asymp T^{-\ell}$ given in (16), we have $\kappa = \kappa(\lambda) \simeq \lambda$. Thus

$$\kappa \asymp T^{-\ell},\ \mathrm{df}_2(\kappa) \asymp \kappa^{-1/\beta} \asymp T^{\ell/\beta}. \tag{41}$$

Now, since $\lambda_i \asymp i^{-\beta}$ (capacity condition) and $(w_0^\top v_i)^2 = c_i^2 \asymp i^{-\delta}$ (source condition), we deduce

$$\begin{aligned}\kappa^2 w_0^\top \Sigma(\Sigma+\kappa I)^{-2} w_0 &\asymp w_0^\top\left(\sum_i \frac{\lambda_i}{(\lambda_i+\kappa^{-1}\lambda_i)^2} v_i v_i^\top\right) w_0 = \sum_i \frac{c_i^2\lambda_i}{(\lambda_i+\kappa^{-1}\lambda_i)^2} \\ &= \sum_i \frac{c_i^2\lambda_i}{(\lambda_i+\kappa^{-1}\lambda_i)^2} \asymp \sum_i \frac{\lambda_i^{1+\delta/\beta}}{(\lambda_i+\kappa^{-1}\lambda_i)^2} \asymp \kappa^{-\gamma} \asymp T^{-\ell\gamma},\end{aligned} \tag{42}$$

where $\gamma = \min(2, 1+\delta/\beta-1/\beta) = \min(2, 2r) = 2\underline{r}$, with $\underline{r} := \min(r, 1)$. The exponent is so because $\delta = 1+\beta(2r-1)$, and so $\delta/\beta = 1/\beta + 2r - 1$ by construction. The estimation of the last sum in (42) is thanks to Lemma C.1 applied with $D = \kappa^{-1}$, $n = 1+\delta/\beta$, and $m = 2$. Therefore, invoking Proposition 4.4 gives

$$Bias \simeq \frac{\kappa^2 w_0^\top \Sigma(\Sigma+\kappa)^{-2} w_0}{1-\mathrm{df}_2(\kappa)/T} \asymp \frac{T^{\ell\gamma}}{1-T^{-(1-\ell/\beta)}} \asymp T^{-\ell\gamma} = T^{-2\ell\underline{r}} \tag{43}$$

$$Var \simeq \sigma^2 \frac{\mathrm{df}_2(\kappa)}{T} \cdot \frac{1}{1-\mathrm{df}_2(\kappa)/T} \asymp \sigma^2 \frac{T^{\ell/\beta}}{T} \frac{1}{1-o(1)} \asymp \sigma^2 T^{-(1-\ell/\beta)}. \tag{44}$$

We deduce the scaling law

$$E_{test} \simeq Bias + Var \asymp T^{-2\ell\underline{r}} + \sigma^2 T^{-(1-\ell/\beta)} \asymp \max(\sigma^2, T^{1-2\ell\underline{r}-\ell/\beta}) T^{-(1-\ell/\beta)}, \tag{45}$$

which is precisely the main result of (Cui et al., 2022).

Low-Noise Regime. In the low noise regime where $\sigma^2 = O(T^{-2\beta\underline{r}})$, one may take $\ell = \beta$; the variance is then much smaller than the bias, and one has the fast rate

$$E_{test} \asymp T^{-2\beta\underline{r}}. \tag{46}$$

High-Noise Regime. Now, consider the case where $\sigma^2 = \Theta(1)$. Setting $2\ell\underline{r} = 1-\ell/\beta$ to balance out the bias and variance gives $\ell = \ell_{crit}$, where

$$\ell_{crit} := \frac{\beta}{2\beta\underline{r}+1} \in (0, \beta). \tag{47}$$

With this value of the exponent ℓ, we get the error rate

$$E_{test} \asymp T^{-2\ell_{crit}\cdot\underline{r}} = T^{-c}, \text{ with } c := \frac{2\beta\underline{r}}{2\beta\underline{r}+1}, \tag{48}$$

which is precisely the main result of (Cui et al., 2022), known to be minimax optimal (de Vito (Caponnetto & de Vito, 2007), etc.) !

C. Auxiliary Lemmas

C.1. Power-Law Computations

Lemma C.1. *Let the sequence $(\lambda_k)_{k\geq 1}$ of positive numbers be such that $\lambda_k \asymp k^{-\beta}$ for some constant $\beta > 0$, and let $m, n \geq 0$ with $n\beta > 1$. Then, for $D \gg 1$, it holds that*

$$\sum_{k=1}^{\infty} \frac{\lambda_k^n}{(1+D\lambda_k)^m} \asymp D^{-c} \begin{cases} \log D, & \text{if } m = n-1/\beta, \\ 1, & \text{else,} \end{cases} \tag{49}$$

where $c := \min(m, n-1/\beta) \geq 0$.

Proof. First observe that

$$\begin{aligned} \lambda_k^n/(1+D\lambda_k)^m &\asymp \lambda_k^n \min(1, (D\lambda_k)^{-m}) \\ &= \begin{cases} \lambda_k^n = k^{-n\beta}, & \text{if } D\lambda_k < 1, \text{ i.e if } k > D^{1/\beta}, \\ D^{-m}\lambda_k^{-(m-n)} = D^{-m}k^{(m-n)\beta}, & \text{else.} \end{cases} \end{aligned}$$

We deduce that

$$\sum_{k=1}^{\infty} \frac{\lambda_k^n}{(1+D\lambda_k)^m} \asymp D^{-m} \sum_{1\leq k\leq D^{1/\beta}} k^{(m-n)\beta} + \sum_{k>D^{1/\beta}} k^{-n\beta}. \tag{50}$$

By comparing with the corresponding integral, one can write the first sum in (50) as

$$\begin{aligned} D^{-m}\sum_{1\le k\le D^{1/\beta}} k^{(m-n)\beta} &\asymp D^{-m}\int_1^{D^{1/\beta}} u^{(m-n)\beta}\mathrm{d}u \\ &\asymp D^{-m}\begin{cases}(D^{1/\beta})^{1+(m-n)\beta}=D^{-(n-1/\beta)}, & \text{if } n-1/\beta<m,\\ \log D, & \text{if } m=n-1/\beta,\\ 1, & \text{else.}\end{cases} \\ &= \begin{cases} D^{-(n-1/\beta)}, & \text{if } n-1/\beta<m,\\ D^{-m}\log D, & \text{if } m=n-1/\beta,\\ D^{-m}, & \text{else.}\end{cases} \\ &= D^{-c}\begin{cases}\log D, & \text{if } m=n-1/\beta,\\ 1, & \text{else,}\end{cases} \end{aligned}$$

where $c\ge 0$ is as given in the lemma.

Analogously, one can write the second sum in (50) as

$$\sum_{k>D^{1/\beta}} k^{-n\beta} \asymp \int_{D^{1/\beta}}^{\infty} u^{-n\beta}\mathrm{d}u \asymp (D^{1/\beta})^{1-n\beta} = D^{-(n-1/\beta)},$$

and the result follows upon putting things together. □

Lemma C.2. *For $\kappa=\kappa(\lambda,T)$ defined as in (7), it holds that*

$$\frac{\partial\kappa}{\partial\lambda}=\frac{1}{1-\mathrm{df}_2(\kappa)/T}\ge 1. \tag{51}$$

Thus, perhaps more conveniently, this lemma allows us to rewrite

$$Bias = w_0^\top\Sigma(\Sigma+\kappa I)^{-2}w_0\frac{\partial\kappa}{\partial\lambda}, \tag{52}$$

$$Var = \sigma^2\frac{\mathrm{df}_2(\kappa)}{T}\frac{\partial\kappa}{\partial\lambda}. \tag{53}$$

The RHS of (52) is usually referred to as the omniscient risk Hastie et al. (2022); Cui et al. (2021); Wei et al. (2022).

Proof of Lemma C.2. By definition of κ, we know that

$$\kappa-\lambda=\kappa\mathrm{df}_1(\kappa)/T=\kappa\,\mathrm{tr}\,\Sigma(\Sigma+\kappa I)^{-1}/T.$$

Differentiating w.r.t λ gives

$$\kappa'-1=\kappa'(\mathrm{tr}\,\Sigma(\Sigma+\kappa I)^{-1}-\kappa\,\mathrm{tr}\,\Sigma(\Sigma+\kappa)^{-2})/T=\kappa'\,\mathrm{tr}\,\Sigma^2(\Sigma+\kappa I)^{-2}/T=\kappa'\mathrm{df}_2(\kappa)/T,$$

and the result follows upon rearranging. Note that we have used the identity

$$I-\kappa(\Sigma+\kappa I)^{-1}=\Sigma(\Sigma+\kappa I)^{-1},$$

to rewrite $\Sigma(\Sigma+\kappa I)^{-1}-\kappa\Sigma(\Sigma+\kappa I)^{-2}=\Sigma^2(\Sigma+\kappa I)^{-2}$. □