Proceedings of Machine Learning Research vol 283:1–21, 2025 7th Annual Conference on Learning for Dynamics and Control

Asymptotics of Linear Regression with Linearly Dependent Data

Behrad Moniri BEMONIRI@SEAS.UPENN.EDU
Hamed Hassani HASSANI@SEAS.UPENN.EDU
University of Pennsylvania, Philadelphia, PA, USA.

Editors: A. Abate, L. Balzano, N. Ozay, D. Panagou

Abstract

In this paper we study the asymptotics of linear regression in settings where the covariates exhibit a linear dependency structure, departing from the standard assumption of independence. We model the covariates as a non-Gaussian stochastic process with spatio-temporal covariance and analyze the performance of ridge regression in the high-dimensional proportional regime, where the number of samples and feature dimensions grow proportionally. A Gaussian universality theorem is proven, demonstrating that the asymptotics are invariant under replacing the non-Gaussian covariates with Gaussian vectors preserving mean and covariance, for which tools from random matrix theory can be used to derive precise characterizations of the estimation error. The estimation error is characterized by a fixed-point equation involving the spectral properties of the spatio-temporal covariance matrices, enabling efficient computation. We then study optimal regularization, overparameterization, and the double descent phenomenon in the context of dependent data. Simulations validate our theoretical predictions, shedding light on how dependencies influence estimation error and the choice of regularization parameters.

Keywords: statistics with dependent data, ridge regression, high-dimensional asymptotics.

1. Introduction

Linear regression is one of the most widely used techniques in machine learning and statistics. Given independent training samples, there are many tools available to study its performance under a set of relatively general conditions (see e.g., Hsu et al. (2012); Oliveira (2016); Mourtada (2022), etc.). In particular, the asymptotic behavior of the estimation error in this case has been fully characterized in the high-dimensional proportional regime where the number of samples and the dimension tend to infinity with a proportional rate (see e.g., Dobriban and Wager (2018); Wu and Xu (2020); Richards et al. (2021); Hastie et al. (2022), etc.).

In contrast, much less is known when the independence assumption on training samples is lifted. The problem of linear regression with dependent data has a wide range of applications from system identification in linear dynamical systems to time-series analysis (see e.g., Ljung (1999); Kailath et al. (2000); Verhaegen and Verdult (2007); Chiuso and Pillonetto (2019); Tsiamis et al. (2023), etc.). Prior work has examined fairly general dependency structures, often providing finite-sample concentration inequalities on the estimation error of (ridgeless) linear regression without deriving precise formulas (see, e.g., Simchowitz et al. (2018); Faradonbeh et al. (2018); Sarkar and Rakhlin (2019); Nagaraj et al. (2020); Ziemann et al. (2024)).

Very recently, Luo et al. (2024); Atanasov et al. (2024) studied high-dimensional asymptotics of ridge regression with non-i.i.d. data under various assumptions on the covariates. Luo et al. (2024) assumed that the covariates follow a right-rotationally invariant covariate distribution. Atanasov et al. (2024) considered the case where the covariates are generated according a *Gaussian* stochastic process with *spatio-temporal* *covariance*. In these papers, precise formulas for the asymptotic training and test errors were derived, and a generalized cross validation estimators (GCV) was proposed to estimate the test error of the model.

In this paper, we first extend these results by considering the case where the covariates are generated according to a general stochastic process with a spatio-temporal covariance; lifting the Gaussianity assumption of prior work. We formally prove that a *Gaussian Universality* property holds and the estimation error remains invariant under changing the covariates to Gaussian vectors while keeping their mean and covariance structure intact. Further, we provide novel precise formulas for the limiting estimation error using the linear pencil method and operator-valued free probability (Mingo and Speicher, 2017; Far et al., 2006). We then use these results to conduct a phenomenological study. In particular, we study the optimal ridge regularization, the effect of regularization and overparameterization, and also the double descent phenomenon in the setting with dependent data. We show that although the optimal ridge regularization is independent of the dependency structure, the dependence among covariates affects the degree of overparameterization in which the peak of the double descent and the phase transition from underparameterization to overparameterization happens.

2. Preliminaries

In this paper, we consider the problem of linear regression with *linearly* dependent data. We assume that we observe n samples of an input d-dimensional time-series $\{x_i\}_{i=1}^n$, and an output real-valued time-series $\{y_i\}_{i=1}^n$ from the following linear model

$$y_i = \langle x_i, \beta_\star \rangle + \varepsilon_i, \quad \text{for} \quad i = 1, 2, \dots, n, \tag{1}$$

where $\beta_\star \in \mathbb{R}^d$ is an unknown linear map, and $\{\varepsilon_i\}_{i=1}^n$ is an additive noise independent of the input time-series $\{x_i\}_{i=1}^n$ with $\varepsilon_i \overset{\text{i.i.d.}}{\sim}$ sub-gaussian$(0, \sigma_\varepsilon^2)$. The goal is to use the samples $\{(x_i, y_i)\}_{i\in[n]}$ to recover the unknown vector $\beta_\star$. To estimate $\beta_\star$, we will use the ridge estimator with parameter $\lambda > 0$, given by

$$\hat{\beta} := \arg\min_{\beta\in\mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle x_i, \beta \rangle)^2 + \lambda \|\beta\|_2^2. \tag{2}$$

To measure the performance of the ridge regularized linear regression estimator, we focus on the estimation error defined as $R(\hat{\beta}) := \|\hat{\beta} - \beta_\star\|_2^2$. We assume that $\{x_i\}_{i=1}^n$ follows a stochastic process with *Spatio-Temporal Covariance*, with a temporal covariance structure that is the same in all dimensions. Such dependency structures have been previously been studied in the literature. See e.g., Nakakita and Imaizumi (2022); Chen et al. (2021); Kyriakidis and Journel (1999). Formally, we let $X := [\, x_1 \,|\, \cdots \,|\, x_n \,]^\top \in \mathbb{R}^{n\times d}$ and make the following assumption on X.

Condition 1 (Covariate Structure) *We assume that $X = AZB$ in which $Z \in \mathbb{R}^{n\times d}$ is a random matrix with i.i.d. entries drawn from a distribution $\mathcal{D}_z$, and that $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{d\times d}$ are deterministic matrices controlling the temporal and spatial dependencies of the covariates.*

In the random matrix theory literature, this model is also known as the bi-correlated model or the separable covariance model (see e.g., Burda et al. (2005); Lixin (2006); Karoui (2009); Paul and Silverstein (2009); Couillet et al. (2011); Couillet and Hachem (2014)). We will present some examples of problems satisfying this condition in Section 4.1. Defining the output vector $y := [\, y_1, \dots, y_n \,]^\top \in \mathbb{R}^n$ and the noise vector $\varepsilon = [\, \varepsilon_1, \dots, \varepsilon_n \,]^\top \in \mathbb{R}^n$, we can use equation 1 to write $y = X\beta_\star + \varepsilon$. With this matrix notation, the estimator $\hat{\beta}$ is given by $\hat{\beta} = (X^\top X + \lambda n I_d)^{-1} X^\top y$. We study this linear regression problem in the following asymptotic regime. The standard linear regression problem without dependencies has been studied extensively in the literature in a similar high-dimensional proportional regime (see e.g., Tulino and Verdú (2004); Dobriban and Wager (2018); Wu and Xu (2020); Richards et al. (2021); Hastie et al. (2022), etc.).

Condition 2 *The sample size n and the dimension d tend to infinity at a proportional rate; i.e., $d, n \to \infty$, with $d/n \to \gamma$, where $\gamma > 0$ is a constant. Throughout the paper, we will denote this limit by* $\lim_{d/n\to\gamma}$.

Consistent with a line of prior work (e.g., Dobriban and Wager (2018); Adlam and Pennington (2020b,a); Lee et al. (2023)), we further make the following *random-coefficient* assumption on the ground truth vector $\beta_\star$. We assume that the effect strength of each feature is drawn independently at random.

Condition 3 (Random Coefficient) *The coefficient $\beta_\star \in \mathbb{R}^d$ is random with $\beta_\star \sim \mathsf{N}\left(0, \frac{\alpha^2}{d} I_d\right)$.*

Note that based on this assumption, the covariates $\{x_i\}$ do not carry any information about $\beta_\star$. Extending the results for the case where the regression coefficients are allowed to be anisotropic is an interesting problem, and we leave it as future work. See e.g., Mel and Pennington (2021); Wu and Xu (2020); Hastie et al. (2022).

Throughout the paper, we study the convergence of random variables in probability. We omit explicit mention of this mode of convergence when clear from the context.

2.1. Related Work

The study of linear regression with dependent data has attracted significant interest, particularly in the control theory community. The typical approach involves deriving non-asymptotic upper and lower bounds on the estimation error for linear regression under various dependency structures. For example, Nagaraj et al. (2020) examine this problem (without regularization) under a realizability condition, assuming data from an exponentially ergodic Markov chain. Additionally, Simchowitz et al. (2018) showed that if we assume both realizability and that the noise forms a martingale difference sequence, then dependent linear regression is no harder than its independent counterpart. Meanwhile, Ziemann et al. (2024) established upper bounds for random design linear regression with dependent (β-mixing) data, without imposing realizability assumptions. Another closely related problem, parameter identification in autoregressive models, has also been widely studied (see Tsiamis et al. (2023) for a recent survey).

Studying problems with non-i.i.d. data is a recent trend in high-dimensional statistics. For example, Bigot et al. (2024) study the ridge regression in the high-dimensional regime with independent but non-identically distributed samples. Zhang et al. (2024) studies GLMs with correlated designs, and Zhang and Mondelli (2024) study the problem of matrix denoising with correlated noise.

Most closely related to the current paper is the results of Nakakita and Imaizumi (2022) and the very recent results of Atanasov et al. (2024) and Luo et al. (2024). Nakakita and Imaizumi (2022) derive bounds for linear regression with a spatio-temporal covariance structure similar to the one we consider. Their analysis investigates the influence of the spectral properties of temporal covariance matrix on these bounds. Atanasov et al. (2024) consider the problem of ridge regression trained on a dataset with a spatio-temporal dependency structure, but for Gaussian covariates. They analyze the precise asymptotics of in-sample and out-sample risks of the ridge estimator and propose a modified generalized cross validation to estimate the out-sample risk, given the in-sample risk. Luo et al. (2024) solve a similar problem with the assumption that the design matrix $\mathbf{X}$ is right-rotationally invariant. Although this setting does not restrict the covariates to be Gaussian, it cannot cover settings with general spatial covariance.

3. Main Results

In this section, we state the main results of the paper that characterize the high-dimensional limits of the estimation error $R(\hat\beta) = \|\hat\beta - \beta_\star\|_2^2$. Using the fact that $y = X\beta_\star + \varepsilon$, the estimator $\hat\beta$ can be written as

$$\hat\beta = \beta_\star - \lambda n(X^\top X + \lambda n I_d)^{-1}\beta_\star + (X^\top X + \lambda n I_d)^{-1}X^\top\varepsilon.$$

Thus, plugging this into the expression for $R(\hat{\beta})$, the estimation error $R(\hat{\beta})$ satisfies

$$R(\hat{\beta}) = \underbrace{\varepsilon^\top X(X^\top X + \lambda n I_d)^{-2} X^\top \varepsilon}_{:=\mathcal{V}} + \underbrace{\lambda^2 n^2 \beta_\star^\top (X^\top X + \lambda n I_d)^{-2} \beta_\star}_{:=\mathcal{B}} - 2\lambda n \beta_\star^\top (X^\top X + \lambda n I_d)^{-2} X^\top \varepsilon.$$

This is the bias-variance decomposition for $R(\hat{\beta})$ where $\mathcal{V}$ is the variance and $\mathcal{B}$ is the bias. Using the Hanson-Wright inequality (Rudelson and Vershynin, 2013), and the fact that ε is independent of all other randomness in the problem, the third term in the above sum converges to zero and we have $|R(\hat{\beta}) - \bar{\mathcal{B}} - \bar{\mathcal{V}}| \to 0$ in which the values $\bar{\mathcal{V}} := \mathbb{E}_\varepsilon[\mathcal{V}]$ and $\bar{\mathcal{B}} := \mathbb{E}_{\beta_\star}[\mathcal{B}]$ are given by

$$\bar{\mathcal{V}} := \frac{\sigma_\varepsilon^2}{n} \operatorname{trace}\left[X(X^\top X + \lambda n I_d)^{-2} X^\top\right], \quad \text{and} \quad \bar{\mathcal{B}} = \frac{\alpha^2 \lambda^2 n^2}{d} \operatorname{trace}\left[(X^\top X + \lambda n I_d)^{-2}\right].$$

Next, we define the functional $m : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, as

$$m(\lambda; \gamma) := \lim_{d/n \to \gamma} \frac{1}{d} \operatorname{tr}\left((B^\top Z^\top A^\top A Z B + \lambda I_d)^{-1}\right). \tag{3}$$

In the subsequent sections, we show that this limit exists, and we will fully characterize it. Recalling that under Condition 2, $X = AZB$, we have

$$\left|\bar{\mathcal{B}} + \alpha^2 \lambda^2 \frac{\partial m(\lambda;\gamma)}{\partial \lambda}\right| \to 0, \text{ and } \left|\bar{\mathcal{V}} - \gamma \sigma_\varepsilon^2 m(\lambda;\gamma) - \gamma \sigma_\varepsilon^2 \lambda \frac{\partial m(\lambda;\gamma)}{\partial \lambda}\right| \to 0. \tag{4}$$

Thus, the estimation error $R(\hat{\beta})$, variance $\bar{\mathcal{V}}$, and bias $\bar{\mathcal{B}}$ are fully characterized by $m(\lambda; \gamma)$.

3.1. Universality

Under Condition 1, the covariates are given by $X = AZB$ in which $Z \in \mathbb{R}^{n \times d}$ is a random matrix with i.i.d. entries drawn from a distribution $\mathcal{D}_z$, and $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{d \times d}$ are deterministic matrices. In this section, we show that for a broad class of distributions $\mathcal{D}_z$, under the high-dimensional setting of Condition 2, the limiting value of the estimation error $R(\hat{\beta})$ is invariant under replacing the distribution $\mathcal{D}_z$ with a Gaussian distribution with the same mean and variance as $\mathcal{D}_z$. This statement is formalized in the following theorem.

Theorem 4 (Gaussian Universality) *Let $G = [g_1, \ldots, g_n]^\top \in \mathbb{R}^{n \times d}$ have entries drawn i.i.d. from Gaussian distribution with the same mean and variance as $\mathcal{D}_z$, and let $\tilde{X} = AGB$ and $\tilde{y} = \tilde{X}\beta_\star + \varepsilon$. Under Conditions 1-3, for the ridge estimator given the dataset $\tilde{X}, \tilde{y}$, i.e., $\tilde{\beta} = (\tilde{X}^\top \tilde{X} + \lambda n I_d)^{-1} \tilde{X}^\top \tilde{y}$, we have*

$$|R(\hat{\beta}) - R(\tilde{\beta})| \to 0$$

in probability, under the condition that $\mathcal{D}_z$ is sub-Gaussian.

This theorem shows that under mild conditions, the limiting value of the estimation error depends on the first two moments of $\mathcal{D}_z$ and not on the other properties of the distribution.

Proof Sketch. Without loss of generality, we assume that $\mathcal{D}_z$ has unit variance. To prove this theorem, given equation 4, it is enough to show that $m(\lambda; \gamma)$ is invariant under the change of $\mathcal{D}_z$ to the corresponding Gaussian. For this, we use a Lindeberg exchange argument (Lindeberg, 1922; Korada and Montanari, 2011; Abbasi et al., 2019) where we replace the rows of Z with corresponding Gaussian vectors one at a time and show that the error incurred by these replacements are negligible. Let $G = [g_1, \ldots, g_n]^\top \in \mathbb{R}^{n \times d}$ be a matrix with entries drawn i.i.d. from Gaussian distribution with the same mean and variance as $\mathcal{D}_z$. For any $i \in [n]$ we define the matrix M_i as

$$M_i := \sum_{k=1}^{i-1} e_k g_k^\top + \sum_{k=i+1}^{n} e_k z_k^\top,$$

where $e_1, \dots, e_n$ are the standard basis of $\mathbb{R}^n$. The matrix will be used to interpolate between Z and G. Also, we define $Z_i := M_i + e_i g_i^\top$. Note that $G = Z_n$ and $Z = Z_0$. Denoting $\Omega_A = A^\top A$, we can write

$$\begin{aligned}\Delta &:= \left|\operatorname{trace}\left(B^\top G^\top \Omega_A G B + \lambda n I_d\right)^{-1} - \operatorname{trace}\left(B^\top Z^\top \Omega_A Z B + \lambda n I_d\right)^{-1}\right| \\ &\leq \sum_{i=1}^{n} \left|\operatorname{trace}\left(B^\top Z_i^\top \Omega_A Z_i B + \lambda n I_d\right)^{-1} - \operatorname{trace}\left(B^\top Z_{i-1}^\top \Omega_A Z_{i-1} B + \lambda n I_d\right)^{-1}\right|.\end{aligned}$$

To complete the proof of the theorem, it is enough to show that each term in the above sum is $o(1/n)$. For this, we use that fact that $Z_i = M_i + e_i g_i^\top$ and $Z_{i-1} = M_i + e_i z_i^\top$ and the apply the Woodbury matrix identity (alongside some elementary algebra) to arrive at $\Delta \leq \sum_{i=1}^n \left|f_i(g_i) - f_i(z_i)\right|$, in which for any $u \in \mathbb{R}^d$, $f_i(u)$ is given by

$$f_i(u) := \operatorname{trace}\left(\begin{bmatrix} u^\top B\, S_i\, B^\top u & 1 + u^\top B\, S_i\, B^\top v_i \\ 1 + v_i^\top B\, S_i\, B^\top u & v_i^\top B\, S_i\, B^\top v_i - e_i^\top \Omega_A e_i \end{bmatrix}^{-1} \begin{bmatrix} u^\top B\, S_i^2\, B^\top u & u^\top B\, S_i^2\, B^\top v_i \\ u^\top B\, S_i^2\, B^\top v_i & v_i^\top B\, S_i^2\, B^\top v_i \end{bmatrix}\right),$$

where $S_i := \left(B^\top M_i^\top\ \Omega_A\ M_i B + \lambda n I_d\right)^{-1}$ and $v_i := M_i^\top \Omega_A e_i \in \mathbb{R}^d$. Given that $\mathcal{D}_z$ is sub-Gaussian and using the Hanson-Wright inequality and noting that by construction $g_i, z_i \perp\!\!\!\perp S_i$, for both $u = g_i$ and $u = z_i$ we have $u^\top B\, S_i B^\top\, u = \operatorname{tr}(B\, S_i B^\top) + O(1/\sqrt{n})$, and $v_i^\top B\, S_i B^\top u = O(1/\sqrt{n})$. Similarly, $u^\top B\, S_i^2 B^\top\, u = \operatorname{tr}(B\, S_i^2 B^\top) + O(n^{-3/2})$, and $v_i^\top B\, S_i^2 B^\top u = O(n^{-3/2})$. Also, using a simple order-wise argument $v_i^\top S_i\, v_i = O(1)$. This shows that given $\omega_i = O(1)$, we *get* $|f_i(g_i) - f_i(z_i)| = O(n^{-3/2}) = o(1/n)$, proving the theorem. The detailed proof can be found in Section A. ■

3.2. Precise Asymptotics

In this section, we provide a precise formula for $\lim_{d/n \to \gamma} m(\lambda; \gamma)$ and use it to fully characterize the estimation error $R(\hat{\beta})$ in the high-dimensional setting of Condition 2. For this, we assume that the empirical spectral distributions of $A^\top A$ and $B^\top B$ converge to limiting distributions μ_A and μ_B respectively and then write $m(\lambda; \gamma)$ in terms of the spectral properties of these limiting distributions.

Condition 5 (Empirical Spectral Distribution) *Let $\lambda_1, \dots, \lambda_n$ be the eigenvalues of $A^\top A \in \mathbb{R}^{n \times n}$ and $\tilde{\lambda}_1, \dots, \tilde{\lambda}_d$ be the eigenvalues of $B^\top B \in \mathbb{R}^{d \times d}$. We assume that as $n, d \to \infty$, we have*

$$\frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i} \implies \mu_A, \quad \textit{and} \quad \frac{1}{d}\sum_{i=1}^{d} \delta_{\tilde{\lambda}_i} \implies \mu_B,$$

where μ_A and μ_B are two probability measures over $\mathbb{R}$, and $\implies$ denotes weak convergence.

This is a mild technical condition and is standard in high-dimensional statistics. Under this condition, in the next theorem we provide a precise formulae for $\lim_{d/n \to \gamma} m(\lambda; \gamma)$, which we will later use to characterize the limiting value of the estimation error $R(\hat{\beta})$.

Theorem 6 *Assume that the conditions of Theorem 4 hold. For any $\lambda, \gamma > 0$, $m(\lambda;\gamma)$ converges in probability to $\bar{m}(\lambda;\gamma)$ that is given by $\bar{m}(\lambda;\gamma) = \kappa\, m_B(\kappa)/\lambda$ where κ is the solution of the following nonlinear equation*

$$\lambda m_A\left(\frac{\lambda}{\gamma\kappa - \gamma\kappa^2\, m_B(\kappa)}\right) + \gamma\kappa\,(\kappa\, m_B(\kappa) - 1)\,(1 - \gamma + \gamma\kappa\, m_B(\kappa)) = 0, \tag{5}$$

where for any $z \in \mathbb{R}$, $m_A(z)$ and $m_B(z)$ are defined as

$$m_A(z) := \int_{\mathbb{R}} \frac{1}{z+x} d\mu_A(x), \quad \textit{and} \quad m_B(z) := \int_{\mathbb{R}} \frac{1}{z+x} d\mu_B(x).$$

This theorem shows that although the limiting expression for $\bar{m}(\lambda;\gamma)$ does not have a closed form, it can be written in terms of a single scalar κ that is the solution to a scalar fixed-point equation which can be solved very efficiently using fixed-point iteration. In contrast, directly using $R(\hat{\beta}) = \|\hat{\beta} - \beta_\star\|_2^2$ requires inversion and multiplication of large-dimensional matrices.

Proof Sketch. Given Theorem 4, without loss of generality we assume that $\mathcal{D}_z$ is Gaussian. To find the limiting value, we construct linear pencils and use the theory of operator-valued free probability to derive the limit of $m(\lambda;\gamma)$. This technique has been used to study random feature models (see e.g., Adlam and Pennington (2020b); Adlam et al. (2022); Tripuraneni et al. (2021); Mel and Pennington (2021); Lee et al. (2023); Dohmatob et al. (2024), etc.). See (Bodin, 2024, Part 1) for an overview of the linear pencil technique in random matrix theory.

A linear pencil is a block-matrix whose blocks are matrices with known spectral properties, and one of the blocks of its inverse is equal to $m(\lambda;\gamma)$. Let $H_0 := \left(B^\top Z^\top \Omega_A\, Z\, B/\lambda n + I_d\right)^{-1}$, where $m(\lambda;\gamma) = \frac{1}{\lambda d}\operatorname{tr}(H_0)$. We can linearize the problem by introducing auxiliary random matrices H_1, H_2, H_3, H_4 to write

$$\underbrace{\begin{bmatrix} I & 0 & 0 & 0 & B^\top \\ -B & I & 0 & 0 & 0 \\ 0 & -Z/\sqrt{\lambda n} & I & 0 & 0 \\ 0 & 0 & -\Omega_A & I & 0 \\ 0 & 0 & 0 & -Z^\top/\sqrt{\lambda n} & I \end{bmatrix}}_{:=L} \begin{bmatrix} H_0 \\ H_1 \\ H_2 \\ H_3 \\ H_4 \end{bmatrix} = \begin{bmatrix} I \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}. \tag{6}$$

This shows that the matrix H_0 is equal to the top left block of L^{-1}. Hence, the matrix L is a suitable linear pencil for our problem. The matrix L can be written as $L = I - L_Z - L_{A,B}$, in which

$$L_Z := \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & Z/\sqrt{\lambda n} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & Z^\top/\sqrt{\lambda n} & 0 \end{bmatrix}, \quad \text{and} \quad L_{A,B} := \begin{bmatrix} 0 & 0 & 0 & 0 & -B^\top \\ B & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \Omega_A & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

We next define the matrix $\mathcal{G} := (\mathrm{id} \otimes \mathbb{E}\bar{\mathrm{tr}})(L^{-1}) = (\mathrm{id} \otimes \mathbb{E}\bar{\mathrm{tr}})\left((I - L_Z - L_{A,B})^{-1}\right) \in \mathbb{R}^{5\times 5}$ where $(\mathrm{id} \otimes \mathbb{E}\bar{\mathrm{tr}})$ is the normalized block-trace; i.e., replacing each block by its trace divided by the size of the block. This matrix can be seen as the operator-valued Cauchy transform (Mingo and Speicher, 2017, Definition 9.10) of $L_Z + L_{A,B}$ evaluated at I; i.e., $\mathcal{G}_{L_Z + L_{A,B}}(I)$. The Cauchy transform of sum of random matrices has been studied extensively in the literature. A common tool for compute this is the R-transform introduced by Voiculescu (1986, 2006) which enables the characterization of the spectrum of a sum of asymptotically freely independent random matrices. Because $L_{A,B}$ is a deterministic matrix, the matrices L_Z and $L_{A,B}$ are asymptotically freely independent (Mingo and Speicher, 2017, Theorem 4.8) which enables us to use the subordination property (Mingo and Speicher, 2017, Equation 9.21) and write $\mathcal{G} = \mathcal{G}_{L_Z+L_{A,B}}(I)$ as

$$\mathcal{G} = (\mathrm{id} \otimes \mathbb{E}\bar{\mathrm{tr}}) \left((I - \mathcal{R}_{L_Z}(\mathcal{G}) - L_{A,B})^{-1} \right), \tag{7}$$

where $\mathcal{R}_{L_Z}(\mathcal{G})$ is the operator-valued R-transform (Mingo and Speicher, 2017, Definition 9.10) of the matrix L_Z, evaluated at $\mathcal{G}$. If we can write the R-transform $\mathcal{R}_{L_Z}(\mathcal{G})$ as a function of $\mathcal{G}$, equation 7 will give a fixed-point equation for $\mathcal{G}$. Since the matrix L_Z consists of i.i.d. Gaussian blocks, we can use (Far et al., 2006, Theorem 4) to compute $\mathcal{R}_{L_Z}(\mathcal{G})$ in terms of $\mathcal{G}$. With this, we find that $\mathcal{R}_{L_Z}(\mathcal{G})$ is a block matrix and its non-zero blocks are $[\mathcal{R}_{L_Z}(\mathcal{G})]_{3,4} = \frac{\gamma}{\lambda}[\mathcal{G}]_{2,5}$, and $[\mathcal{R}_{L_Z}(\mathcal{G})]_{5,2} = \frac{1}{\lambda}[\mathcal{G}]_{4,3}$. Note that by construction, we have $[\mathcal{G}]_{1,1} = \frac{1}{d}\,\mathrm{tr}(H_0) = \lambda m(\lambda;\gamma)$. Plugging this R-transform into equation 7, and looking at the $(1,1)$, $(4,3)$, and $(2,5)$ blocks of both sides, we get the following three equalities

$$\lambda m(\lambda;\gamma) = [\mathcal{G}]_{1,1} = \frac{\lambda}{d}\,\mathrm{tr}\left((\lambda I + [\mathcal{G}]_{4,3}\,\Omega_B)^{-1}\right), \quad [\mathcal{G}]_{4,3} = \frac{\lambda}{n}\,\mathrm{tr}\left(A^\top A(\lambda I - \gamma\,[\mathcal{G}]_{2,5}\,\Omega_A)^{-1}\right), \quad \text{and}$$

$$[\mathcal{G}]_{2,5} = -\frac{\lambda}{d}\,\mathrm{tr}\left(B^\top B(\lambda I + [\mathcal{G}]_{4,3}\,\Omega_B)^{-1}\right).$$

Defining $\kappa := \lambda/[\mathcal{G}]_{4,3}$ and simplifying this system of three equations, we arrive $m(\lambda;\gamma) = \kappa m_B(\kappa)/\lambda$, and equation 5 which concludes the proof. The full proof can be found in Section B. ■

The limiting Stieltjes transform $\bar{m}(\lambda;\gamma)$ has previously been studied transform and is shown to be charactereized by a system of equations (see e.g., Lixin (2006, Theorem 1.2.1)). However, in this paper, by leveraging the power machinery of operator-valued free probability, we provide a novel and simpler characterization of $\bar{m}(\lambda;\gamma)$, through a much simpler proof.

This theorem can be used alongside equation 8 to find a formula for the limiting bias, variance, and estimation error in terms of the implicit variable κ as follows.

Corollary 7 *Under the conditions of Theorem 6, with probability $1-o(1)$, we have*

$$\bar{\mathcal{B}} \to -\alpha^2\lambda^2\frac{\partial \bar{m}}{\partial \lambda}, \quad \bar{\mathcal{V}} \to \gamma\sigma_\varepsilon^2\kappa\, m_B(\kappa)/\lambda + \gamma\sigma_\varepsilon^2\lambda\frac{\partial \bar{m}}{\partial \lambda}, \quad \textit{and} \quad |R(\hat{\beta}) - \bar{\mathcal{B}} - \bar{\mathcal{V}}| \to 0, \tag{8}$$

in which κ and $\bar{m}$ is defined in Theorem 6 and $\frac{\partial \bar{m}}{\partial \lambda}$ can be derived by implicit differentiation of equation 5. The formula for $\frac{\partial \bar{m}}{\partial \lambda}$ in terms of κ can be found in Section C.

3.3. Optimal Regularization

Despite the lack of a closed form solution for the limiting estimation error, the optimal value for the ridge parameter λ can still be derived, even in the case where the covariates are dependent.

Proposition 8 *Let the data be generated according to equation 1 and $\hat{\beta}$ be the solution to equation 2 with regularization parameter $\lambda \in \mathbb{R}$. Assume that Condition 3 holds. The optimal value of the ridge regularization parameter is given by*

$$\lambda_\star := \arg\min_{\lambda>0}\left(\lim_{d/n\to\gamma} R(\hat{\beta})\right) = \sigma_\varepsilon^2\gamma/\alpha^2.$$

The case where the covariates are i.i.d. was studied by Dobriban and Wager (2018) where it was shown that the optimal value of λ is again equal to $\sigma_\varepsilon^2\gamma/\alpha^2$. Hence, the optimal λ does *not* depend on the dependency structure of the covariates.

Proof. Similar to Dobriban and Wager (2018), we note that in the finite d, n setting and under the assumption that $\varepsilon \sim \mathsf{N}(0, \sigma_\varepsilon^2 I_n)$ and $\beta_\star \sim \mathsf{N}(0, \alpha^2 I_d/d)$, the ridge regression of equation 2 with $\lambda_n = \sigma_\varepsilon^2 d/\alpha^2 n$ is the Bayes-optimal estimator for $\beta_\star$. Using (Dobriban and Wager, 2018, Lemma 6.1), $\lambda_\star = \lim_{d/n\to\gamma} \lambda_n = \sigma_\varepsilon^2 \gamma/\alpha^2$ is the optimal ridge regularization under the high-dimensional limit of Condition 2.

4. Special Cases

In this section, we study special cases of Corollary 7 where the fixed point equation of equation 5 simplifies. Although equation 5 is a complicated fixed point equation, in special cases it can be used to recover prior work that study special cases of this problem.

Case 1: $B^\top B = I_d$ and $A^\top A = I_n$ In this case, for any $z \in \mathbb{R}$ we have $m_B(z) = m_A(z) = (1+z)^{-1}$. Using the fixed-point equation 5, we find that $\bar{m}(\lambda;\gamma)$ is the solution of $\lambda\gamma\bar{m}^2(\lambda;\gamma) + (1+\lambda-\gamma)\,\bar{m}(\lambda;\gamma) - 1 = 0$, which corresponds to the standard formula for the Stieltjes transform of the Marchenko-Pastur law (Bai and Silverstein, 2010). This recovers the formula for $\bar{m}$ in the case where samples are i.i.d. and have a covariance matrix equal to identity (Tulino and Verdú, 2004).

Case 2: $A^\top A = I_n$ In this case, for any $z \in \mathbb{R}$ we have $m_A(z) = (1+z)^{-1}$. Plugging this into the fixed-point equations in equation 5, we find that $\bar{m}(\lambda;\gamma) = \kappa m_B(\kappa)/\lambda$ where κ satisfies $\gamma\,\kappa^2\, m_B(\kappa) + \kappa\,(1-\gamma) - \lambda = 0$. In this case, as expected, we recover the formula for the i.i.d. case (but with a general covariance matrix) from Dobriban and Wager (2018); Hastie et al. (2022); Wu and Xu (2020).

Case 3: $B^\top B = I_d$ In this case, for any $z \in \mathbb{R}$ we have $m_B(z) = (1+z)^{-1}$. Plugging this into the fixed-point equations in equation 5, we find that $\bar{m}(\lambda;\gamma)$ is the solution to the following equation

$$\lambda\gamma^2\,\bar{m}^2(\lambda;\gamma) + \gamma\,(1-\gamma)\,\bar{m}(\lambda;\gamma) - m_A\left(\frac{1}{\gamma\,\bar{m}(\lambda;\gamma)}\right) = 0.$$

4.1. Examples

The linear dependency structure we assume for the covariates in Condition 1 is fairly general. To illustrate this, we present two examples of linear regression problems that fit within the framework of this paper. First, given $\omega, \omega_0, \ldots, \omega_q \in \mathbb{R}$ we define the following two Toeplitz matrices (all other entries are zero)

$$A_{\mathrm{AR}} = \begin{bmatrix} \omega_0 & & & & \\ \vdots & \omega_0 & & & \\ \omega_q & \vdots & \ddots & & \\ & \omega_q & \ddots & \omega_0 & \\ & & \omega_q & \cdots & \omega_0 \end{bmatrix} \in \mathbb{R}^{n\times n}, \quad \text{and} \quad A_{\mathrm{R}} = \begin{bmatrix} 1 & & & & & \\ 0 & \omega & & & & \\ & 1-\omega & 1 & & & \\ & & 0 & \omega & & \\ & & & 1-\omega & \ddots & \\ & & & & & \ddots \end{bmatrix} \in \mathbb{R}^{n\times n}. \tag{9}$$

Example 1 (Autoregressive): Let $z_1, \ldots, z_n \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, I_d)$, and $\{x_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$ be generated according from the following autoregressive model with order $q \in \mathbb{N}$, given by

$$x_t = \sum_{i=0}^{q} \omega_i z_{t-i}, \quad \text{and} \quad y_t = \langle x_t, \beta_\star\rangle + \varepsilon_i,$$

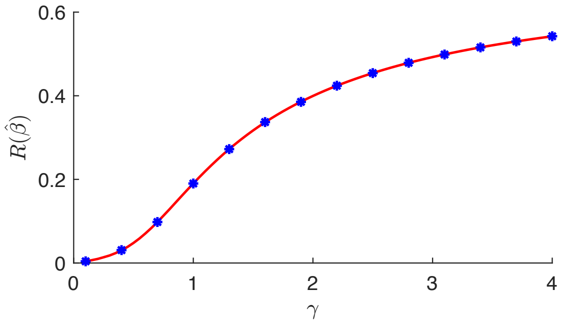

Figure 1: The estimation error $R(\hat{\beta})$ as a function of γ with $\mu_A = 1/3\,\delta_1 + 1/3\,\delta_2 + 1/3\,\delta_3$, $\mu_B = 1/2\,\delta_1 + 1/2\,\delta_2$, $\sigma_\varepsilon = 0.2$, and $\alpha = 0.7$. We set $\lambda = 0.03$ (Left) and $\lambda = \gamma$ (Right). In these plots, the solid lines are the theoretical predictions using Corollary 7 and the dots are numerical predictions, averaged over 50 trials. Error bars are negligible and cannot be viewed.

where $\omega_0, \ldots, \omega_q \in \mathbb{R}$ are real coefficients. This data generation process satisfies Condition 1 with $B = I_d$, and A equal to the Toeplitz matrix A_{AR} from equation 9.

The limiting spectral distribution of such Toeplitz converges and the limiting distribution is characterized by the celebrated Szegő theorem (see e.g, Grenander and Szego (1958); Tyrtyshnikov (1996); Gray (1972)). In particular, it is known this example satisfies Condition 5 and for any $z \in \mathbb{R}$, we have

$$m_{A_{\mathrm{AR}}}(z) = \frac{1}{2\pi}\int_0^{2\pi} \frac{1}{z + |f|^2(\lambda)}\, d\lambda, \quad \text{in which} \quad f(\lambda) := \sum_{k=0}^{q} \omega_k e^{ik\lambda}.$$

Hence, Corollary 7 fully characterizes the bias, variance, and estimation error for this example.

Example 2 (Redundancy): In this example, we will present a dependency structure that exhibit a controllable amount of redundancy, and we will use this setting for some experiments in the next section. Let $\omega \in [0, 1]$ be a real number and $z_1, \ldots, z_n \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0, I_d)$. Assume that the covariates are given by $x_i = z_i$ if i is odd and $x_i = \omega z_i + (1-\omega) z_{i-1}$ if i is even and that y_i are generated similar to Example 1. The case where $\omega = 1$ corresponds to the case with i.i.d. $\mathsf{N}(0, I_d)$ covariates, and the case with $\omega = 0$ corresponds to the case where each covariate is repeated twice. This data generation process satisfies Condition 1 with $B = I_d$, and A equal to A_{R} from equation 9.

5. Experiments and Simulations

Theory vs. Simulation. We set $\mu_A = 1/3\,\delta_1 + 1/3\,\delta_2 + 1/3\,\delta_3$, $\mu_B = 1/2\,\delta_1 + 1/2\,\delta_2$, $\sigma_\varepsilon = 0.2$, and $\alpha = 0.7$. In this experiment, we consider the estimation error once with a fixed $\lambda = 0.03$, and once with $\lambda = \gamma$. In these plots, we use Corollary 7 to plot the limiting $R(\hat{\beta})$ as a function of $\gamma \in [0, 4]$. For each setting, we also numerically simulate results and average over 50 trials. The results can be found in Figure 1. We observe that the theoretical predictions of the paper match perfectly with the numerical simulations.

Effect of Dependence. We consider the setting of Example 2 in Section 4.1 and set $\sigma_\varepsilon = 1, \alpha = 1$, and $\lambda = 0.05$. In Figure 4, we plot $R(\hat{\beta})$, variance, and bias, once with $\omega = 0.2$ (Right) and once for $\omega = 0.8$ (Left).

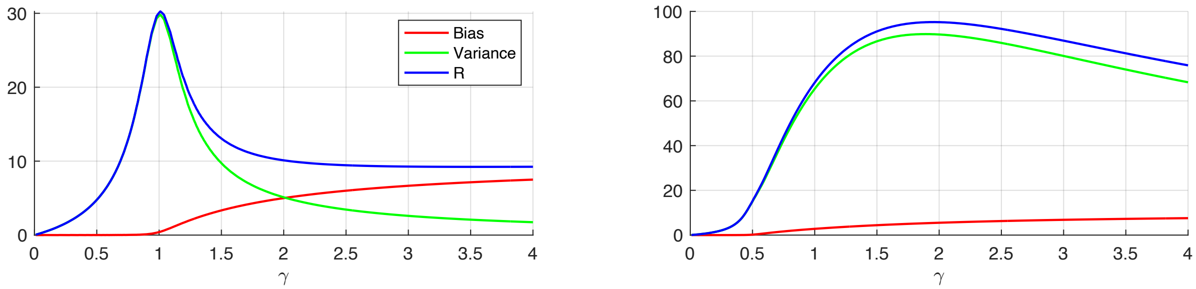

Figure 4: The estimation error, bias, and variance of Example 2 in Section 4.1 with $\sigma_\varepsilon = 1, \alpha = 1$, and $\gamma = 0.2$ as a function of γ for $\omega = 0.8$ (Left) and $\omega = 0.2$ (Right).

We observe that in the high redundancy case where $\omega = 0.2$, the bias starts increasing around $\gamma = 0.5$ whereas in the low redundancy case with $\omega = 0.8$, the bias is almost zero for $\gamma < 1$, changing the critical γ after which the model is overparameterized. This is due to the change of the number of effective samples. The peak of $R(\hat{\beta})$ which corresponds to double descent (Belkin et al., 2019; Hastie et al., 2022) also shifts as we change the amount of redundancy. In Figure 2, we consider the same example but set $\lambda = \lambda_\star$, and $\gamma = 2$, and plot $R(\hat{\beta})$ as a function of ω. We see that the error of the optimally-tuned ridge estimator is decreasing as the redundancy is decreased.

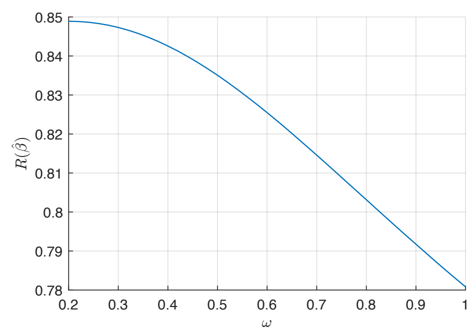

Figure 2: The estimation error as a function of ω for Example 2 in Section 4.1 as a function of ω with $\sigma_\varepsilon = 1, \alpha = 1$, $\gamma = 2$ and $\lambda = \lambda_\star$.

Effect of Regularization. In Figure 3, we again consider the setting of Example 2 in Section 4.1 and set $\sigma_\varepsilon = 1, \alpha = 1$, and $\gamma = 0.2$. We use Corollary 7 to plot the estimation error as a function of $\log(\lambda)$. In this plot, different curves correspond to different values of $\omega \in [0, 1]$. The dashed vertical line corresponds to $\log(\lambda) = \log(\lambda_\star)$. We observed that, as predicted by Proposition 8, the minimum risk is attained at $\lambda = \lambda_\star$ and the optimal ridge regularize does not depend on the matrix A and is the same for all values of ω.

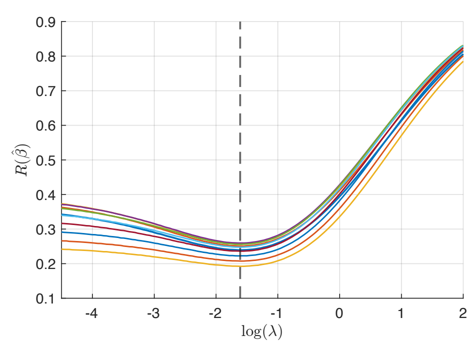

Figure 3: The estimation error as a function of λ in the same setting of Figure 2 with $\sigma_\varepsilon = 1, \alpha = 1$, and $\gamma = 0.2$.

6. Conclusions

This paper presents a detailed asymptotic analysis of ridge regularized linear regression with linearly dependent data in the high-dimensional proportional regime. By employing random matrix theory and establishing Gaussian universality, we derived precise characterizations of the estimation error and its dependence on spatio-temporal covariance structures. Our results provide new insights into optimal regularization, overparameterization, and the double descent phenomenon in settings with dependent data. Theoretical results closely align with simulations, demonstrating the validity of our findings.

7. Acknowledgments

B.M. would like to thank Ingvar Ziemann, Thomas T.C.K. Zhang, and Bruce D. Lee for stimulating discussions on the topic of statistical estimation with dependent data, during the early phases of this project. B.M. is supported by a gift from AWS AI to Penn Engineering's ASSET Center for Trustworthy AI. The work of B.M. and H.H. is supported by The Institute for Learning-enabled Optimization at Scale (TILOS), under award number NSF-CCF-2112665, and the NSF CAREER award CIF-1943064.

References

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

Ben Adlam and Jeffrey Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, 2020a.

Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. In *Advances in Neural Information Processing Systems*, 2020b.

Ben Adlam, Jake A Levinson, and Jeffrey Pennington. A random matrix perspective on mixtures of nonlinearities in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, 2022.

Alexander Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Risk and cross validation in ridge regression with correlated samples. *arXiv preprint arXiv:2408.04607*, 2024.

Zhidong Bai and Jack W Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*, volume 20. Springer, 2010.

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

Jérémie Bigot, Issa-Mbenard Dabo, and Camille Male. High-dimensional analysis of ridge regression for non-identically distributed data with a variance profile. *arXiv preprint arXiv:2403.20200*, 2024.

Antoine Philippe Michel Bodin. Random matrix methods for high-dimensional machine learning models. Technical report, EPFL, 2024.

Zdzisław Burda, Jerzy Jurkiewicz, and Bartłomiej Wacław. Spectral moments of correlated wishart matrices. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 71(2):026111, 2005.

Wanfang Chen, Marc G Genton, and Ying Sun. Space-time covariance structures and models. *Annual Review of Statistics and Its Application*, 8(1):191–215, 2021.

Alessandro Chiuso and Gianluigi Pillonetto. System identification: A machine learning perspective. *Annual Review of Control, Robotics, and Autonomous Systems*, 2(1):281–304, 2019.

Romain Couillet and Walid Hachem. Analysis of the limiting spectral measure of large random matrices of the separable covariance type. *Random Matrices: Theory and Applications*, 3(04):1450016, 2014.

Romain Couillet, Mérouane Debbah, and Jack W Silverstein. A deterministic equivalent for the analysis of correlated mimo multiple access channels. *IEEE Transactions on Information Theory*, 57(6):3493–3514, 2011.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. Strong model collapse. *arXiv preprint arXiv:2410.04840*, 2024.

Reza Rashidi Far, Tamer Oraby, Wlodzimierz Bryc, and Roland Speicher. Spectra of large block matrices. *arXiv preprint cs/0610045*, 2006.

Mohamad Kazem Shirani Faradonbeh, Ambuj Tewari, and George Michailidis. Finite time identification in unstable linear systems. *Automatica*, 96:342–353, 2018.

Robert Gray. On the asymptotic eigenvalue distribution of toeplitz matrices. *IEEE Transactions on Information Theory*, 18(6):725–730, 1972.

Ulf Grenander and Gabor Szego. *Toeplitz Forms and Their Applications*. University of California Press, 1958.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, 50(2):949–986, 2022.

Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, pages 9–1. JMLR Workshop and Conference Proceedings, 2012.

Thomas Kailath, Ali H Sayed, and Babak Hassibi. *Linear estimation*. Prentice Hall, 2000.

Noureddine El Karoui. Concentration of measure and spectra of random matrices: Applications to correlation matrices, elliptical distributions and beyond. *The Annals of Applied Probability*, 19(6), 2009.

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

Phaedon C Kyriakidis and André G Journel. Geostatistical space–time models: a review. *Mathematical geology*, 31:651–684, 1999.

Donghwan Lee, Behrad Moniri, Xinmeng Huang, Edgar Dobriban, and Hamed Hassani. Demystifying disagreement-on-the-line in high dimensions. In *International Conference on Machine Learning*, 2023.

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

Zhang Lixin. *Spectral analysis of large dimensional random matrices*. PhD thesis, Ph. D. thesis, National University of Singapore, Singapore, 2006.

Lennart Ljung. *System identification: theory for the user*. Prentice Hall, 1999.

Kevin Luo, Yufan Li, and Pragya Sur. Roti-gcv: Generalized cross-validation for right-rotationally invariant data. *arXiv preprint arXiv:2406.11666*, 2024.

Gabriel Mel and Jeffrey Pennington. Anisotropic random feature regression in high dimensions. In *International Conference on Learning Representations*, 2021.

James A Mingo and Roland Speicher. *Free probability and random matrices*, volume 35. Springer, 2017.

Jaouad Mourtada. Exact minimax risk for linear least squares, and the lower tail of sample covariance matrices. *The Annals of Statistics*, 50(4):2157–2178, 2022.

Dheeraj Nagaraj, Xian Wu, Guy Bresler, Prateek Jain, and Praneeth Netrapalli. Least squares regression with markovian data: Fundamental limits and algorithms. In *Advances in Neural Information Processing systems*, volume 33, 2020.

Shogo Nakakita and Masaaki Imaizumi. Benign overfitting in time series linear model with over-parameterization. *arXiv preprint arXiv:2204.08369*, 2022.

Roberto Imbuzeiro Oliveira. The lower tail of random quadratic forms with applications to ordinary least squares. *Probability Theory and Related Fields*, 166:1175–1194, 2016.

Debashis Paul and Jack W Silverstein. No eigenvalues outside the support of the limiting empirical spectral distribution of a separable covariance matrix. *Journal of Multivariate Analysis*, 100(1):37–57, 2009.

Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. Asymptotics of ridge (less) regression under general source condition. In *International Conference on Artificial Intelligence and Statistics*, 2021.

Mark Rudelson and Roman Vershynin. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18:1–9, 2013.

Tuhin Sarkar and Alexander Rakhlin. Near optimal finite time identification of arbitrary linear dynamical systems. In *International Conference on Machine Learning*, 2019.

Max Simchowitz, Horia Mania, Stephen Tu, Michael I Jordan, and Benjamin Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Conference On Learning Theory*, pages 439–473, 2018.

Nilesh Tripuraneni, Ben Adlam, and Jeffrey Pennington. Overparameterization improves robustness to covariate shift in high dimensions. In *Advances in Neural Information Processing Systems*, 2021.

Anastasios Tsiamis, Ingvar Ziemann, Nikolai Matni, and George J Pappas. Statistical learning theory for control: A finite-sample perspective. *IEEE Control Systems Magazine*, 43(6):67–97, 2023.

Antonio M Tulino and Sergio Verdú. Random matrix theory and wireless communications. *Communications and Information Theory*, 1(1):1–182, 2004.

Evgenij E Tyrtyshnikov. A unifying approach to some old and new theorems on distribution and clustering. *Linear algebra and its applications*, 232:1–43, 1996.

Michel Verhaegen and Vincent Verdult. *Filtering and system identification: a least squares approach*. Cambridge university press, 2007.

Dan Voiculescu. Addition of certain non-commuting random variables. *Journal of Functional Analysis*, 66 (3):323–346, 1986.

Dan Voiculescu. Symmetries of some reduced free product c^*-algebras. In *Operator Algebras and their Connections with Topology and Ergodic Theory: Proceedings of the OATE Conference held in Buşteni, Romania, Aug. 29–Sept. 9, 1983*, pages 556–588. Springer, 2006.

Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. In *Advances in Neural Information Processing Systems*, 2020.

Yihan Zhang and Marco Mondelli. Matrix denoising with doubly heteroscedastic noise: Fundamental limits and optimal spectral methods. *arXiv preprint arXiv:2405.13912*, 2024.

Yihan Zhang, Hong Chang Ji, Ramji Venkataramanan, and Marco Mondelli. Spectral estimators for structured generalized linear models via approximate message passing. In *Conference on Learning Theory*, 2024.

Ingvar Ziemann, Stephen Tu, George J Pappas, and Nikolai Matni. The noise level in linear regression with dependent data. In *Advances in Neural Information Processing Systems*, 2024.