AM254 Final Project

Ishaan Sinha and Armie Rysmakhanov

December 2025

1 Abstract

Given recent advances in Natural Language Processing, it is natural to wonder whether we see similar scaling laws apply to denoising models. Namely, if we expand the size of the model and increase the size of the training dataset, do we expect the loss function of our model to drastically reduce. To explore this question, we study linear denoising models using techniques like Continuous Gaussian Min-Max Theorem and the Cavity Method. We also derive the optimal linear estimators and explore the performance of other good estimators.

2 Introduction

Deep neural networks have achieved remarkable success in a variety of image–processing tasks, including image sharpening, super-resolution, and denoising. In particular, denoising methods have become foundational for modern generative models, where denoising serves as the core mechanism for sampling from complex image distributions [11, 2, 8]. A variety of architectures have been explored for these tasks, most notably convolutional neural networks (CNNs) [12] and Vision Transformer–based models [1]. Despite empirical progress, theoretical understanding of these models has lagged, and understanding of how performance scales with model size, dataset size, and architectural choice remains limited.

In contrast, Language Modeling has benefited from extensive study of scaling laws, demonstrating substantial and predictable improvements with increases in model capacity, dataset size, and training compute [4, 3]. State-of-the-art language models now contain hundreds of billions of parameters, trained on trillions of tokens, whereas high-performing image denoisers typically rely on datasets on the order of 10^4–10^5 images, several orders of magnitude smaller. This discrepancy raises fundamental questions about the role of scaling in image denoising and about whether similar scaling laws might apply.

In this work, we investigate these questions through theoretical and empirical lenses. Our contributions:

- We apply tools from high-dimensional optimization, particularly the Convex Gaussian Min–Max Theorem (CGMT) [10, 9], to derive theoretical scaling laws for image denoising under simplified but analytically tractable models.
- Drawing on statistical physics, we use the cavity method and related ideas from spin-glass theory [7, 6] to characterize the asymptotic behavior of large denoising models.
- We empirically validate our theoretical predictions through controlled simulations, examining performance across varying model sizes, dataset sizes, and noise regimes.

Our analysis reveals that, beyond a certain capacity threshold, scaling model size yields diminishing or even negligible returns for denoising accuracy. These findings suggest that architectural or algorithmic innovations may ultimately be more impactful than brute-force scaling for this class of problems.

3 Problem Setup and Background

3.1 Problem

We consider the following denoising problem to analyze the accuracy of a regularized linear denoiser in the asymptotic limit. Starting with some $d < n$ dimensional subspace of $\mathbb{R}^n$ that the signal comes from, we parametrize this space with an orthonormal basis $U \in \mathbb{R}^{n\times d}$ and draw samples approximately uniformly from the subspace, by taking $x = Uc$, where $c \sim \mathcal{N}(0, I_d)$. The observed noisy measurement is $y = x + z$ where z is Gaussian noise parametrized by $z \sim \mathcal{N}(0, \sigma_z^2 I_n)$. Given samples $\{(x_1, y_1), \ldots, (x_N, y_N)\}$, we wish to train our denoiser $W \in \mathbb{R}^{n\times n}$ that minimizes the regularized sqaure loss:

$$W^* = \arg\min l(W),\ l(W) = \frac{1}{N}\sum_{i=1}^{N} ||Wy_i - x_i||_2^2 + \lambda||W||_F^2$$

where λ is the regularization parameter. We measure the test error by

$$R(W) = \frac{1}{d}\mathbb{E}||Wy - x||_2^2$$

and focus on the asymptotic limit where we maintain $N/n \to \alpha$ and $n/d \to \beta$.

3.2 Optimal Linear Estimator

We now derive the optimal linear estimator, ie the estimator that minimizes the risk $R(W)$ given that U is apriori known.

We write the minimization objective as:

$$l(W) = \mathbb{E}\|Wy - x\|_2^2 + \lambda\|W\|_F^2 = \mathbb{E}[y^\top W^\top W y] - 2\,\mathbb{E}[x^\top W y] + \mathbb{E}\|x\|_2^2 + \lambda||W||_F^2$$

Letting the covariance matrices be $\Sigma_{yy} = \mathbb{E}[yy^\top]$ and $\Sigma_{xy} = \mathbb{E}[xy^\top]$ we get

$$l(W) = \operatorname{tr}(W\Sigma_{yy}W^\top) - 2\operatorname{tr}(W\Sigma_{xy}^\top) + \operatorname{tr}(\mathbb{E}[xx^\top]) + \lambda \operatorname{tr}(WW^\top).$$

Removing the constant term $\operatorname{tr}(\mathbb{E}[xx^\top])$ we differentiate to get

$$\nabla_W l(W) = 2W\Sigma_{yy} - 2\Sigma_{xy} + 2\lambda W$$

Setting the gradient to zero gives

$$W(\Sigma_{yy} + \lambda I_n) = \Sigma_{xy}.$$

Assuming $\Sigma_{yy} + \lambda I_n$ is invertible, we get:

$$W^\star = \Sigma_{xy}\left(\Sigma_{yy} + \lambda I_n\right)^{-1}$$

Recalling that $x = Uc$ with $U^\top U = I_d$ we have

$$\Sigma_{xy} = \mathbb{E}[x(x+z)^\top] = UU^\top, \qquad \Sigma_{yy} = \mathbb{E}[(x+z)(x+z)^\top] = UU^\top + \sigma_z^2 I_n$$

using the covariance matrices of c and z. Notice that $P = UU^T$ is an orthogonal projector onto $\operatorname{span}(U)$, since it has eigenvalue 1 on $\operatorname{span}(U)$ and 0 on $\operatorname{span}(U)^\top$. Therefore,

$$W^* = P\big(P + (\sigma_z^2 + \lambda)I_n\big)^{-1}$$

From the Woodbury formula, we get

$$\big(P+(\sigma_z^2+\lambda)I_n\big)^{-1} \;=\; \frac{1}{1+\sigma_z^2+\lambda}\,P + \frac{1}{\sigma_z^2+\lambda}\,(I_n - P) \implies W^* = \frac{1}{1+\sigma_z^2+\lambda}\,UU^\top$$

which is an operator that projects onto the subspace of the signal with a shrinkage factor of

$$\alpha = \frac{1}{1+\sigma_z^2+\lambda}$$

For $W^* = \alpha P$, and since $Px = x$, the risk of this optimal linear estimator is given by

$$W^\star y - x \;=\; (\alpha - 1)x + \alpha P z$$

By the independence of x and z

$$\mathbb{E}\|W^\star y - x\|_2^2 = \mathbb{E}\|(\alpha-1)x\|_2^2 + \mathbb{E}\|\alpha P z\|_2^2.$$

Since $\mathbb{E}\|x\|_2^2 = \mathbb{E}\|c\|_2^2 = d$ and $\mathbb{E}\|Pz\|_2^2 = \operatorname{tr}(P\,\mathbb{E}[zz^\top]) = \sigma_z^2 d$

$$R(W^\star) = \frac{1}{d}\mathbb{E}\|W^\star y - x\|_2^2 = (\alpha-1)^2 + \alpha^2\sigma_z^2 = \frac{(\sigma_z^2+\lambda)^2 + \sigma_z^2}{(1+\sigma_z^2+\lambda)^2}$$

3.3 Approximations

Since U is not known apriori, [5] suggest using a Principal Component Analysis on the given $Y = [y_1, \ldots, y_N]$ and estimating the subspace U as the d leading singular vectors of the empirical covariance matrix $YY^T \in \mathbb{R}^{n \times n}$. We write this a $\hat{U} \in \mathbb{R}^{n \times d}$. Denoting the estimator W_{PCA} by the same form as the derived W^* estimator, but using our empirical $\hat{U}$, [5] has shown that as long as the number of training examples N is large compared to $d + n\sigma_z^2$, the risk of the PCA estimator is close to the risk of the optimal estimator.

The authors in [5] also suggest training a neural network end to end by applying gradient descent to the empirical risk, regularizing using early-stopping.

4 Theory

4.1 Cavity Method

For a fixed index i, define the leave one out objective

$$\ell_{(i)}(W) = \frac{1}{N} \sum_{j \neq i} \|Wy_j - x_j\|_2^2 + \lambda \|W\|_F^2, \tag{1}$$

so that

$$\ell(W) = \ell_{(i)}(W) + \frac{1}{N} \|Wy_i - x_i\|_2^2. \tag{2}$$

Let

$$W_{(i)} = \arg\min_W \ell_{(i)}(W), \qquad W^* = \arg\min_W \ell(W), \tag{3}$$

and write the cavity displacement as $W^* = W_{(i)} + \Delta W_i$.

Since $\nabla \ell(W^*) = 0$, we have

$$0 = \nabla \ell_{(i)}(W_{(i)} + \Delta W_i) + \frac{1}{N} \nabla \|(W_{(i)} + \Delta W_i) y_i - x_i\|_2^2. \tag{4}$$

We now Taylor expand the first term at $W_{(i)}$:

$$\nabla \ell_{(i)}(W_{(i)} + \Delta W_i) = \nabla \ell_{(i)}(W_{(i)}) + \nabla^2 \ell_{(i)}(W_{(i)})[\Delta W_i] + o(\|\Delta W_i\|) \tag{5}$$

Since $W_{(i)}$ is the minimizer of $\ell_{(i)}$, $\nabla \ell_{(i)}(W_{(i)}) = 0$.

Next, expand the gradient:

$$\frac{1}{N} \nabla \|(W_{(i)} + \Delta W_i) y_i - x_i\|_2^2 = \frac{2}{N} \big((W_{(i)} + \Delta W_i) y_i - x_i\big) y_i^\top \tag{6}$$

$$= \frac{2}{N} (W_{(i)} y_i - x_i) y_i^\top + \frac{2}{N} (\Delta W_i y_i) y_i^\top \tag{7}$$

Therefore,

$$\nabla^2 \ell_{(i)}(W_{(i)})[\Delta W_i] = -\frac{2}{N} (W_{(i)} y_i - x_i) y_i^\top - \frac{2}{N} (\Delta W_i y_i) y_i^\top \tag{8}$$

We can also check that the Hessian is given by

$$\nabla^2 \ell_{(i)}(W_{(i)})[\Delta W_i] = \frac{2}{N} \sum_{j \neq i} \Delta W_i y_j y_j^\top + 2\lambda \Delta W_i \tag{9}$$

Next, notice that the $\frac{2}{N}(\Delta W_i y_i) y_i^\top$ is of order $\frac{1}{N}$ so can be ignored to give the cavity equation:

$$\left(\frac{1}{N} \sum_{j \neq i} y_j y_j^\top + \lambda I_n \right) \Delta W_i = -\frac{1}{N} (W_{(i)} y_i - x_i) y_i^\top$$

$$\implies \Delta W_i = -\frac{1}{N} (W_{(i)} y_i - x_i) y_i^\top \left(\frac{1}{N} \sum_{j \neq i} y_j y_j^\top + \lambda I_n \right)^{-1}$$

Since the difference between $W_{(i)}$ and W^* is itself $O(1/N)$, we may replace $W_{(i)}$ by W^*:

$$\Delta W_i = -\frac{1}{N} (W^* y_i - x_i) y_i^\top \left(\frac{1}{N} \sum_{j \neq i} y_j y_j^\top + \lambda I_n \right)^{-1} + o\Big(\frac{1}{N}\Big) \tag{10}$$

Next, We define the leave-one-out residual vector h_i and the empirical training residual vector r_i as:

$$h_i = W_{(i)} y_i - x_i, \qquad r_i = W^* y_i - x_i. \tag{11}$$

In the asymptotic limit where $N, n, d \to \infty$, the test error $R(W)$ concentrates and can be expressed as the average of the squared leave-one-out residuals:

$$R(W) \approx \frac{1}{d} \sum_{i=1}^{N} \|h_i\|_2^2. \tag{12}$$

In the large system limit, the resolvent matrix $G_{(i)}$ can be replaced by the full resolvent $G = (\frac{1}{N} Y Y^\top + \lambda I_n)^{-1}$, giving

$$r_i = (W_{(i)} + \Delta W_i) y_i - x_i \tag{13}$$

$$= (W_{(i)} y_i - x_i) + \Delta W_i y_i \tag{14}$$

$$= h_i - \frac{1}{N} r_i (y_i^\top G y_i). \tag{15}$$

$$h_i = r_i \left(1 + \frac{1}{N} y_i^\top G y_i \right). \tag{16}$$

Letting $\nu = \lim_{N \to \infty} \frac{1}{N} \mathrm{Tr}(G \Sigma_y)$, where $\Sigma_y = \mathbb{E}[y y^\top]$, we get that $h_i \approx (1+\nu) r_i$. Therefore, letting $L_{\text{train}} = \frac{1}{N} \sum \|r_i\|^2$, we get

$$R(W) = \frac{1}{d} \sum_{i=1}^{N} \|(1+\nu) r_i\|_2^2 = \frac{N}{d} (1+\nu)^2 \left(\frac{1}{N} \sum_{i=1}^{N} \|r_i\|_2^2 \right) = \alpha \beta (1+\nu)^2 L_{\text{train}}(W^*) \tag{17}$$

4.2 Convex Gaussian Min-Max Theorem

4.2.1 Φ_H - hard problem

The training set is $(x_j, y_j)_{j=1}^N$ with

$$c_j \sim N(0, I_d), \quad z_j \sim N(0, I_n), \quad x_j = Uc_j, \quad y_j = Uc_j + \sigma z_j, \quad \sigma > 0.$$

Assume $U \in \mathbb{R}^{n\times d}$ is column orthogonal, i.e. $U^\top U = I_d$. Write

$$C = [c_1, \dots, c_N] \in \mathbb{R}^{d\times N}, \quad Z = [z_1, \dots, z_N] \in \mathbb{R}^{n\times N}, \quad X = UC, \quad Y = UC + \sigma Z.$$

First rewrite the problem:

$$\begin{aligned}
\ell(W) &= \frac{1}{N}\sum_{i=1}^N \|Wy_i - x_i\|_2^2 + \lambda\|W\|_F^2 \\
&= \frac{1}{N}\|WY - X\|_F^2 + \lambda\|W\|_F^2 \\
&= \sum_{i=1}^n \Big(\frac{1}{N}\|W_iY - X_i\|_2^2 + \lambda\|W_i\|_2^2\Big), \quad W_i, X_i \text{ are rows} \\
&= \sum_{i=1}^n \Big(\frac{1}{N}\|(W_iY - X_i)^\top\|_2^2 + \lambda\|W_i^\top\|_2^2\Big) \\
&= \sum_{i=1}^n \Big(\frac{1}{N}\|Y^\top W_i^\top - X_i^\top\|_2^2 + \lambda\|W_i^\top\|_2^2\Big).
\end{aligned}$$

As we see, there are n independent minimization problem here, one for each W_i. Let $W_i^\top$ be w and $X_i^\top$ be x.

We have $X = UC$, so the i-th row of X is $e_i^\top UC$. Taking transpose,

$$x = X_i^\top = (e_i^\top UC)^\top = C^\top U^\top e_i.$$

Define

$$k := U^\top e_i \in \mathbb{R}^d.$$

Then

$$x = C^\top k.$$

Use the formula from class:

$$\|s\|_2^2 = \max_{u\in\mathbb{R}^N} \Big(2u^\top s - \|u\|_2^2\Big).$$

Apply it with $s = Y^\top w - x$:

$$\begin{aligned}
\Phi_H &= \min_w \ \frac{1}{N}\max_u \Big(2u^\top(Y^\top w - x) - \|u\|^2\Big) + \lambda\|w\|^2 \\
&= \min_w \ \max_u \ \frac{2}{N}u^\top(Y^\top w - x) - \frac{1}{N}\|u\|^2 + \lambda\|w\|^2.
\end{aligned}$$

4.2.2 Φ_E - easy problem

Because $U^\top U = I_d$, $\mathrm{col}(U)$ is a d-dimensional subspace of $\mathbb{R}^n$. Decompose

$$w = Ua + b, \quad a := U^\top w \in \mathbb{R}^d, \quad b \perp \mathrm{col}(U),$$

such that $U^\top w = a$ and $\|w\|_2^2 = \|a\|_2^2 + \|b\|_2^2$. Using $Y = UC + \sigma Z$,

$$\begin{aligned} Y^\top w - x &= (UC + \sigma Z)^\top w - C^\top k \\ &= C^\top (U^\top w - k) + \sigma Z^\top w \\ &= C^\top (a - k) + \sigma Z^\top (Ua + b). \end{aligned}$$

Focus on the second term for now. Let $Q = [U \;\; U_\perp] \in \mathbb{R}^{n \times n}$ be orthogonal, where $U_\perp$ is any orthonormal basis for the orthogonal complement of $\mathrm{col}(U)$. Using the rotation trick, $Q^\top Z \overset{d}{=} Z$.

Write $Q^\top Z = [Z_1 \quad Z_2]^\top$ with $Z_1 \in \mathbb{R}^{d \times N}$, $Z_2 \in \mathbb{R}^{(n-d) \times N}$ independent and iid $N(0,1)$. Then $Z^\top U = Z_1^\top$. Then

$$Z^\top (Ua + b) \overset{d}{=} Z_1^\top a + Z_2^\top b,$$

with Z_1, Z_2 independent of each other and of C.

Define

$$A := C^\top \in \mathbb{R}^{N \times d}, \quad B := Z_1^\top \in \mathbb{R}^{N \times d}, \quad G := Z_2^\top \in \mathbb{R}^{N \times (n-d)},$$

so that A, B, G have iid $N(0,1)$ entries and are mutually independent. Then

$$Y^\top w - x \overset{d}{=} A(a - k) + \sigma Ba + \sigma Gb = (A + \sigma B)a + \sigma Gb - Ak.$$

Let $M := A + \sigma B$. Since $(A_{k\ell}, M_{k\ell})$ is jointly Gaussian with $\mathrm{Var}(A_{k\ell}) = 1$, $\mathrm{Var}(M_{k\ell}) = 1 + \sigma^2$, and $\mathrm{Cov}(A_{k\ell}, M_{k\ell}) = 1$, so $\rho = 1/(1 + \sigma^2)$ and

$$A = \rho M + \sqrt{1 - \rho} A_0 = \frac{1}{1 + \sigma^2} M + \frac{\sigma}{\sqrt{1 + \sigma^2}} A_0,$$

where A_0 has iid $N(0,1)$ entries and is independent of M. Set

$$a' := a - \frac{1}{1 + \sigma^2} k, \quad x_0 := A_0 k.$$

Note that $x_0 \sim N(0, \|k\|_2^2 I_N)$ and is independent of M, G. Then

$$Y^\top w - x \overset{d}{=} Ma' + \sigma Gb - \frac{\sigma}{\sqrt{1 + \sigma^2}} x_0.$$

Since M has iid $N(0, 1 + \sigma^2)$ entries, write $M = \sqrt{1 + \sigma^2} G_1$ with G_1 iid $N(0,1)$ and independent of G, x_0, so, given k,

$$Y^\top w - x \overset{d}{=} \sqrt{1 + \sigma^2} G_1 a' + \sigma Gb - \frac{\sigma}{\sqrt{1 + \sigma^2}} x_0, \quad G_1 \perp G \perp x_0.$$

Note that

$$\|w\|_2^2 = \|Ua+b\|_2^2 = \|a\|_2^2 + \|b\|_2^2 = \|a'\|_2^2 + \|b\|_2^2 + \frac{\|k\|_2^2}{(1+\sigma^2)^2} + \frac{2}{1+\sigma^2} a'^\top k.$$

Therefore

$$\begin{aligned}\Phi_H \overset{d}{=} \min_{a',b} \max_{u} \frac{2}{N} u^\top \Big(\sqrt{1+\sigma^2} G_1 a' + \sigma G b - \frac{\sigma}{\sqrt{1+\sigma^2}} x_0 \Big) - \frac{1}{N}\|u\|_2^2 \\ + \lambda(\|a'\|_2^2 + \|b\|_2^2) + \lambda \frac{\|k\|_2^2}{(1+\sigma^2)^2} + \frac{2\lambda}{1+\sigma^2} a'^\top k.\end{aligned}$$

Now we will apply CGMT. Before that, we have to collect the matrices into a single matrix with iid standard Gaussian entries. Let

$$\bar{G} := [G_1 \quad G] \in \mathbb{R}^{N \times n}, \quad \bar{x} := [\sqrt{1+\sigma^2} a' \quad \sigma b]^\top \in \mathbb{R}^n,$$

so that $u^\top \bar{G} \bar{x} = u^\top (\sqrt{1+\sigma^2} G_1 a' + \sigma G b)$ and $\|\bar{x}\|_2 = \sqrt{(1+\sigma^2)\|a'\|_2^2 + \sigma^2 \|b\|_2^2}$. Then

$$\begin{aligned}\Phi_H \overset{d}{=} \min_{a',b} \max_{u} \frac{2}{N} u^\top \bar{G} \bar{x} - \frac{2\sigma}{\sqrt{1+\sigma^2} N} u^\top x_0 - \frac{1}{N}\|u\|_2^2 \\ + \lambda(\|a'\|_2^2 + \|b\|_2^2) + \lambda \frac{\|k\|_2^2}{(1+\sigma^2)^2} + \frac{2\lambda}{1+\sigma^2} a'^\top k.\end{aligned}$$

The corresponding "easy" problem then is

$$\begin{aligned}\Phi_E = \min_{a',b} \max_{u} \frac{2}{N} \Big(\|\bar{x}\|_2 s^\top u + \|u\|_2 g^\top \bar{x} \Big) - \frac{2\sigma}{\sqrt{1+\sigma^2} N} u^\top x_0 - \frac{1}{N}\|u\|_2^2 \\ + \lambda(\|a'\|_2^2 + \|b\|_2^2) + \lambda \frac{\|k\|_2^2}{(1+\sigma^2)^2} + \frac{2\lambda}{1+\sigma^2} a'^\top k,\end{aligned}$$

where $s \sim N(0, I_N)$ and $g \sim N(0, I_n)$ are independent of each other and of x_0. Writing $g = [g_1 \quad g_2]$ with $g_1 \in \mathbb{R}^d$, $g_2 \in \mathbb{R}^{n-d}$ results in $g^\top \bar{x} = \sqrt{1+\sigma^2} g_1^\top a' + \sigma g_2^\top b$, and hence

$$\begin{aligned}\Phi_E = \min_{a',b} \max_{u} \frac{2}{N} \Big(\sqrt{(1+\sigma^2)\|a'\|_2^2 + \sigma^2 \|b\|_2^2} s^\top u + \|u\|_2 \big(\sqrt{1+\sigma^2} g_1^\top a' + \sigma g_2^\top b \big) \Big) \\ - \frac{2\sigma}{\sqrt{1+\sigma^2} N} u^\top x_0 - \frac{1}{N}\|u\|_2^2 + \lambda(\|a'\|_2^2 + \|b\|_2^2) + \lambda \frac{\|k\|_2^2}{(1+\sigma^2)^2} + \frac{2\lambda}{1+\sigma^2} a'^\top k.\end{aligned}$$

Now we will reduce Φ_E into a problem dependent on scalars rather than vectors. Introduce

$$q := \|a'\|_2 \geq 0, \quad p := \|b\|_2 \geq 0, \quad t := \|u\|_2 \geq 0, \quad r := \sqrt{(1+\sigma^2)q^2 + \sigma^2 p^2}.$$

Write $u = tv$ with $\|v\|_2 = 1$. In the Φ_E, the only terms that depend on v are $s^\top u$ and $x_0^\top u$:

$$\frac{2}{N} r s^\top u - \frac{2\sigma}{\sqrt{1+\sigma^2} N} u^\top x_0 = \frac{2t}{N} v^\top (rs - \frac{\sigma}{\sqrt{1+\sigma^2}} x_0).$$

Therefore,

$$\max_{\|v\|_2=1} \frac{2t}{N} v^\top (rs - \frac{\sigma}{\sqrt{1+\sigma^2}} x_0) = \frac{2t}{N} \|rs - \frac{\sigma}{\sqrt{1+\sigma^2}} x_0\|_2.$$

The only term that depends the direction of b is $\|u\|_2 \sigma g_2^\top b$. So,

$$\min_{\|b\|_2=p} \frac{2}{N} \|u\|_2 \sigma g_2^\top b = -\frac{2t}{N} \sigma \|g_2\|_2 p.$$

Two terms depend on the direction of a': $\frac{2}{N}\|u\|_2\sqrt{1+\sigma^2} g_1^\top a'$ and $\frac{2\lambda}{1+\sigma^2} a'^\top k$. Their sum is

$$\frac{2t}{N}\sqrt{1+\sigma^2} g_1^\top a' + \frac{2\lambda}{1+\sigma^2} k^\top a' = \Big(\frac{2t}{N}\sqrt{1+\sigma^2} g_1 + \frac{2\lambda}{1+\sigma^2} k\Big)^\top a'.$$

Hence

$$\min_{\|a'\|_2=q} \Big(\frac{2t}{N}\sqrt{1+\sigma^2} g_1 + \frac{2\lambda}{1+\sigma^2} k\Big)^\top a' = -q \Big\| \frac{2t}{N}\sqrt{1+\sigma^2} g_1 + \frac{2\lambda}{1+\sigma^2} k \Big\|_2.$$

Then the problem reduces to

$$\Phi_E = \min_{q\geq 0,\ p\geq 0} \max_{t\geq 0} \frac{2t}{N} \|rs - \frac{\sigma}{\sqrt{1+\sigma^2}} x_0\|_2 - \frac{2t}{N}\sigma\|g_2\|_2 p - q \Big\| \frac{2t}{N}\sqrt{1+\sigma^2} g_1 + \frac{2\lambda}{1+\sigma^2} k \Big\|_2 - \frac{t^2}{N} + \lambda(q^2+p^2) + \lambda \frac{\|k\|_2^2}{(1+\sigma^2)^2}.$$

4.2.3 Limit of Φ_E

What is the limit of this expression? Let

$$\frac{N}{n} \to \alpha, \quad \frac{n}{d} \to \beta \geq 1.$$

Note that

$$\|k\|_2^2 = \|U^\top e_i\|_2^2 \approx \frac{d}{n} \to \frac{1}{\beta}.$$

First note that $rs - \frac{\sigma}{\sqrt{1+\sigma^2}} x_0 \sim N(0, \big(r^2 + \frac{\sigma}{\sqrt{1+\sigma^2}}^2 \|k\|_2^2\big) I_N)$. Therefore,

$$\frac{1}{\sqrt{N}} \|rs - \frac{\sigma}{\sqrt{1+\sigma^2}} x_0\|_2 \to \sqrt{(1+\sigma^2)q^2 + \sigma^2 p^2 + \frac{\sigma^2}{1+\sigma^2}\frac{1}{\beta}}.$$

Define the inner part

$$V(p,q) := (1+\sigma^2)q^2 + \sigma^2 p^2 + \frac{\sigma^2}{1+\sigma^2}\frac{1}{\beta}.$$

Then

$$\|rs - \frac{\sigma}{\sqrt{1+\sigma^2}} x_0\|_2 \to \sqrt{N}\sqrt{V(p,q)}.$$

We also have $\|g_2\|_2/\sqrt{n-d} \to 1$. Given that $\frac{n-d}{N} \to \frac{n(1-1/\beta)}{\alpha n} = \frac{1-1/\beta}{\alpha}$, we also have

$$\frac{1}{\sqrt{N}}\|g_2\|_2 \to \sqrt{\frac{1-1/\beta}{\alpha}}.$$

Now expand another term,

$$\left\|\frac{2t}{N}\sqrt{1+\sigma^2}g_1+\frac{2\lambda}{1+\sigma^2}k\right\|_2^2 = \frac{4t^2(1+\sigma^2)}{N^2}\|g_1\|_2^2+\frac{4\lambda^2}{(1+\sigma^2)^2}\|k\|_2^2+\frac{8t\lambda\sqrt{1+\sigma^2}}{N(1+\sigma^2)}g_1^\top k.$$

Resscale term by term using $t = \tau\sqrt{N}$:

$$\frac{4t^2(1+\sigma^2)}{N^2}\|g_1\|_2^2 = \frac{4\tau^2(1+\sigma^2)}{N}\|g_1\|_2^2 \to 4\tau^2(1+\sigma^2)\frac{d}{N} = \frac{4\tau^2(1+\sigma^2)}{\alpha\beta}.$$

Next,

$$\frac{4\lambda^2}{(1+\sigma^2)^2}\|k\|_2^2 \to \frac{4\lambda^2}{\beta(1+\sigma^2)^2}.$$

The cross term vanishes because of N in the denominator and no N, n, d-dependence in the numerator. Therefore,

$$\left\|\frac{2t}{N}\sqrt{1+\sigma^2}g_1 + \frac{2\lambda}{1+\sigma^2}k\right\|_2 \to 2\sqrt{\frac{\tau^2(1+\sigma^2)}{\alpha\beta}+\frac{\lambda^2}{\beta(1+\sigma^2)^2}}.$$

Applying the scaling $t = \tau\sqrt{N}$ to other terms, we get

$$\frac{2t}{N}\|rs - \frac{\sigma}{\sqrt{1+\sigma^2}}x_0\|_2 \to 2\tau\sqrt{V(p,q)},$$

$$\frac{2t}{N}\sigma\|g_2\|_2 p = \frac{2\tau\sqrt{N}}{N}\sigma p\|g_2\|_2 = 2\tau\sigma p\frac{\|g_2\|_2}{\sqrt{N}} \to 2\tau\sigma p\sqrt{\frac{1-1/\beta}{\alpha}},$$

$$-\frac{t^2}{N} = -\tau^2,$$

and

$$\lambda\frac{\|k\|_2^2}{(1+\sigma^2)^2} \to \frac{\lambda}{\beta(1+\sigma^2)^2}.$$

Combining the limits, we have

$$\Phi_E \to \min_{q\geq 0,\ p\geq 0}\ \max_{\tau\geq 0}\ 2\tau\sqrt{V(p,q)} - 2\tau\sigma p\sqrt{\frac{1-1/\beta}{\alpha}} - 2q\sqrt{\frac{\tau^2(1+\sigma^2)}{\alpha\beta}+\frac{\lambda^2}{\beta(1+\sigma^2)^2}} - \tau^2$$
$$+\lambda(q^2+p^2)+\frac{\lambda}{\beta(1+\sigma^2)^2}.$$

Define

$$A := \sqrt{V(p,q)} - \sigma p\sqrt{\frac{1-1/\beta}{\alpha}},$$

and constants

$$B := \frac{1+\sigma^2}{\alpha\beta}, \quad C := \frac{\lambda^2}{\beta(1+\sigma^2)^2}.$$

Then the problem is

$$\Phi_E \to \min_{p\geq 0,\ q\geq 0} \left\{ \max_{\tau\geq 0} \ \phi_{p,q}(\tau) + \lambda(p^2+q^2) + \frac{\lambda}{\beta(1+\sigma^2)^2} \right\},$$

where

$$\phi_{p,q}(\tau) = 2\tau A - 2q\sqrt{B\tau^2 + C} - \tau^2.$$

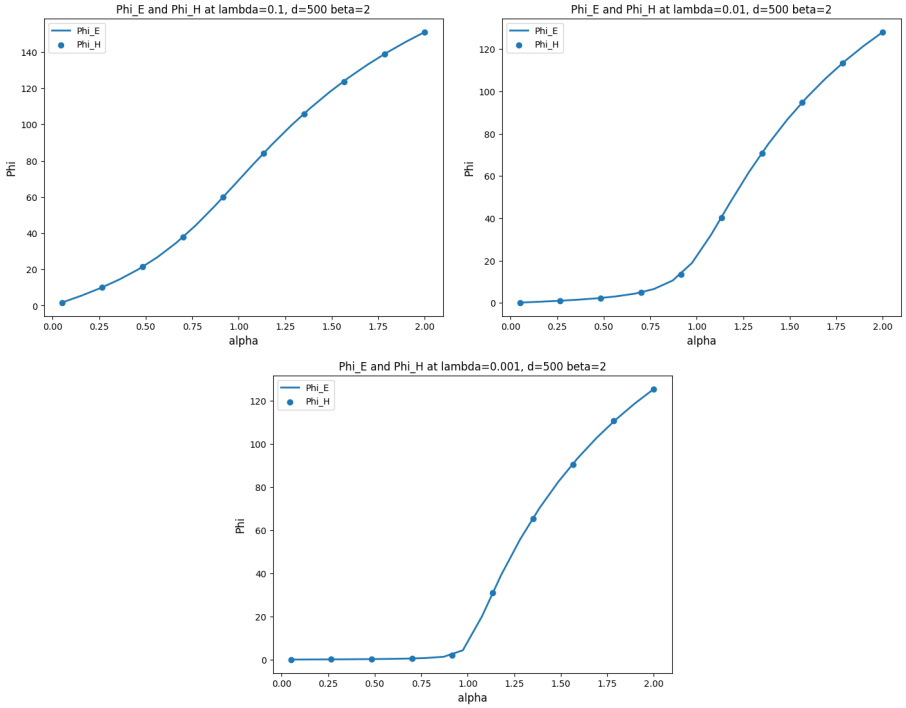
The problem is convex and the optimal τ^* satisfies

$$0 = \phi'_{p,q}(\tau) = 2A - 2\tau - 2q\frac{B\tau}{\sqrt{B\tau^2+C}}.$$

Then the high-dimensional limit is

$$\Phi_E \to \min_{p\geq 0,\ q\geq 0} \left\{ 2\tau^* A - 2q\sqrt{B(\tau^*)^2 + C} - (\tau^*)^2 + \lambda(p^2+q^2) + \frac{\lambda}{\beta(1+\sigma^2)^2} \right\}.$$

The validity of Φ_E is checked empirically against Φ_H, which was computed through the known optimal formula $W^* = XY^\top \left(YY^\top + N\lambda I_n\right)^{-1}$:



4.2.4 Phase transition at $\alpha = 1$

The last image hints that there is a phase transition at $\alpha = 1$. As $\lambda \to 0$,

$$C = \frac{\lambda^2}{\beta(1+\sigma^2)^2} \to 0, \quad \frac{\lambda}{\beta(1+\sigma^2)^2} \to 0, \quad \sqrt{B\tau^2 + C} \to \sqrt{B}\tau.$$

Then the problem becomes a quadratic, allowing for a simple expression of the optimal value of τ, resulting in

$$\min_{p,q\geq 0} \Big[\sqrt{(1+\sigma^2)q^2 + \sigma^2 p^2 + \frac{\sigma^2}{(1+\sigma^2)\beta}} - \sigma p\sqrt{\frac{1-1/\beta}{\alpha}} - q\sqrt{\frac{1+\sigma^2}{\alpha\beta}}\Big]_+^2.$$

Write the linear term as an inner product

$$\sigma p\sqrt{\frac{1-1/\beta}{\alpha}} + q\sqrt{\frac{1+\sigma^2}{\alpha\beta}} = [\sigma p \quad \sqrt{1+\sigma^2}q]^\top [\sqrt{\frac{1-1/\beta}{\alpha}} \quad \frac{1}{\sqrt{\alpha\beta}}].$$

By Cauchy-Schwarz and $\left(\frac{1-1/\beta}{\alpha} + \frac{1}{\alpha\beta}\right)^{1/2} = 1/\sqrt{\alpha}$,

$$\sigma p\sqrt{\frac{1-1/\beta}{\alpha}} + q\sqrt{\frac{1+\sigma^2}{\alpha\beta}} \leq \frac{1}{\sqrt{\alpha}}\sqrt{\sigma^2 p^2 + (1+\sigma^2)q^2}.$$

Equality is achieved with $p, q \geq 0$ when

$$\frac{\sqrt{1+\sigma^2}q}{\sigma p} = \frac{1/\sqrt{\beta}}{\sqrt{1-1/\beta}} = \frac{1}{\sqrt{\beta-1}} \implies q = \frac{\sigma}{\sqrt{1+\sigma^2}}\frac{p}{\sqrt{\beta-1}}.$$

For this choice,

$$\sigma^2 p^2 + (1+\sigma^2)q^2 = \sigma^2 p^2\Big(1 + \frac{1}{\beta-1}\Big) = \sigma^2 p^2 \frac{\beta}{\beta-1},$$

so the 2D minimum equals the 1D minimum

$$\min_{p\geq 0} \Big[\sigma\sqrt{\frac{\beta}{\beta-1}p^2 + \frac{1}{(1+\sigma^2)\beta}} - \frac{\sigma}{\sqrt{\alpha}}\sqrt{\frac{\beta}{\beta-1}}p\Big]_+^2.$$

If $\alpha < 1$, we can always pick large enough p so that the expression inside the bracket is negative, forcing the minimum to be zero.

If $\alpha > 1$, the inside of the bracket is always positive. Let

$$f(p) = \sqrt{\frac{\beta}{\beta-1}p^2 + \frac{1}{(1+\sigma^2)\beta}} - \frac{1}{\sqrt{\alpha}}\sqrt{\frac{\beta}{\beta-1}}p.$$

Setting the derivative to zero gives

$$\begin{aligned}
\frac{\frac{\beta}{\beta-1}p}{\sqrt{\frac{\beta}{\beta-1}p^2+\frac{1}{(1+\sigma^2)\beta}}} &= \frac{1}{\sqrt{\alpha}}\sqrt{\frac{\beta}{\beta-1}} \\
\frac{\frac{\beta^2}{(\beta-1)^2}p^2}{\frac{\beta}{\beta-1}p^2+\frac{1}{(1+\sigma^2)\beta}} &= \frac{\beta}{\alpha(\beta-1)} \\
\frac{\beta^2}{(\beta-1)^2}p^2 &= \frac{\beta^2}{\alpha(\beta-1)^2}p^2+\frac{1}{\alpha(1+\sigma^2)(\beta-1)} \\
\frac{\beta^2}{(\beta-1)^2}p^2\frac{\alpha-1}{\alpha} &= \frac{1}{\alpha(1+\sigma^2)(\beta-1)} \\
(p^*)^2 &= \frac{\beta-1}{(1+\sigma^2)\beta^2(\alpha-1)}.
\end{aligned}$$

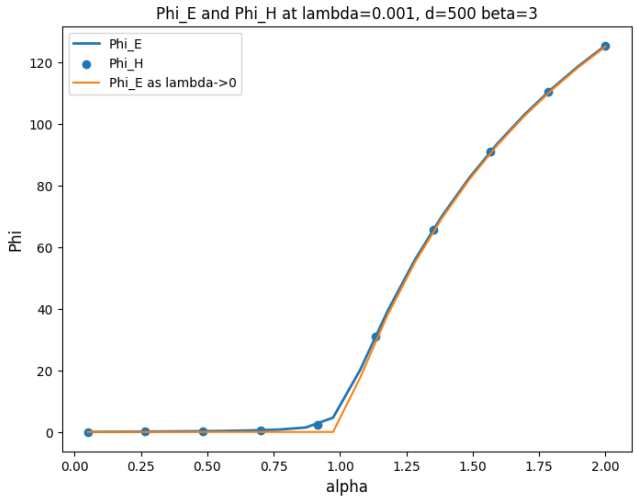
Then

$$\begin{aligned}
f(p) &= \sqrt{\frac{\beta}{\beta-1}\frac{\beta-1}{(1+\sigma^2)\beta^2(\alpha-1)}+\frac{1}{(1+\sigma^2)\beta}}-\frac{1}{\sqrt{\alpha}}\sqrt{\frac{\beta}{\beta-1}}\sqrt{\frac{\beta-1}{(1+\sigma^2)\beta^2(\alpha-1)}} \\
&= \sqrt{\frac{\alpha}{(1+\sigma^2)\beta(\alpha-1)}}-\sqrt{\frac{1}{\alpha(1+\sigma^2)\beta(\alpha-1)}} \\
&= \sqrt{\frac{\alpha-1}{\alpha(1+\sigma^2)\beta}}.
\end{aligned}$$

Therefore for $\alpha > 1$,

$$\min_{p\geq 0}\sigma^2 f(p)^2 = \frac{\sigma^2}{(1+\sigma^2)\beta}\Big(1-\frac{1}{\alpha}\Big).$$

This agrees with empirical findings, where the orange line was constructed using the proposed asymptotic formula and the other parts are constructed like in previous three graphs:



4.3 Per-row test loss

First consider the case $\alpha > 1$.

Define the test loss for any $w \in \mathbb{R}^n$ by

$$R_{\text{test}}(w) := \mathbb{E}\big[(w^\top(Uc + \sigma z) - k^\top c)^2\big].$$

Expanding and using $\mathbb{E}[(v^\top c)^2] = \|v\|^2$ and $\mathbb{E}[(w^\top z)^2] = \|w\|^2$ gives

$$R_{\text{test}}(w) = \|U^\top w - k\|_2^2 + \sigma^2 \|w\|_2^2.$$

Write $C = [c_1, \ldots, c_N]$ and $Z = [z_1, \ldots, z_N]$ for the training sample and recall the training loss

$$R_N(w) := \frac{1}{N}\|C^\top(U^\top w - k) + \sigma Z^\top w\|_2^2.$$

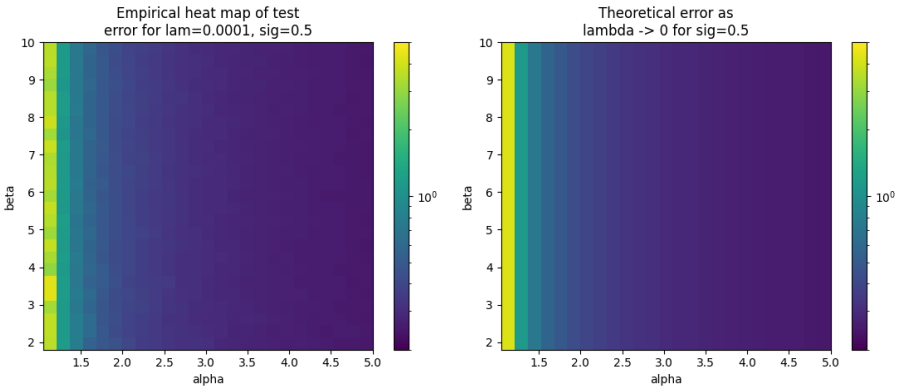
Expandingg gives

$$\begin{aligned} R_N(w) - R_{\text{test}}(w) =& (U^\top w - k)^\top \Big(\frac{1}{N} CC^\top - I_d\Big)(U^\top w - k) \\ &+ \sigma^2 w^\top \Big(\frac{1}{N} ZZ^\top - I_n\Big) w + 2\sigma (U^\top w - k)^\top \Big(\frac{1}{N} CZ^\top\Big) w. \end{aligned}$$

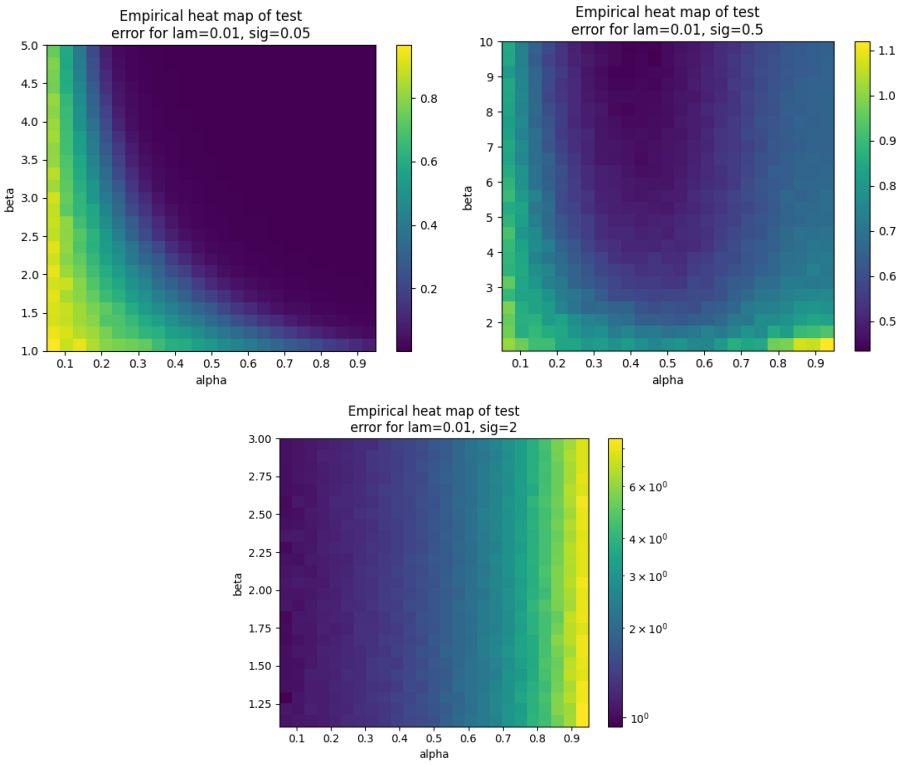
Since $\frac{1}{N} CC^\top \to I_d$, $\frac{1}{N} ZZ^\top \to I_n$, and $\frac{1}{N} CZ^\top \to 0$, the right-hand side vanishes for the trained ridge solution $\hat{w}$. So $R_{\text{test}}(\hat{w}) \approx R_N(\hat{w})$, and the leading term is the same training objective already analyzed by CGMT; hence the Φ_E limit will to the test loss without in the limit.

For fixed $\sigma = 0.5$, there seems to be no β-dependence for the error and the actual test error seems to be well-described by the $\frac{\sigma^2}{(1+\sigma^2)\beta}(1 - \frac{1}{\alpha})$ asymptotic expression derived from CGMT, corresponding to the second descent in the

double descent discussed at the end of the paper. The color scale is fixed at the same values for both heatmaps.



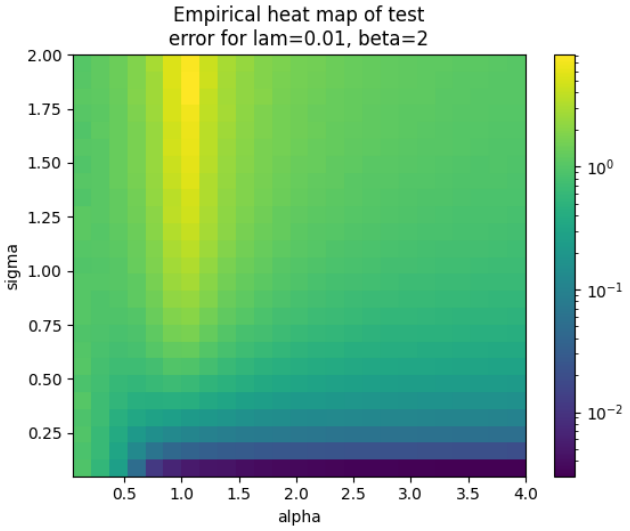
However, for $\alpha < 1$, there are regimes where β-dependence appears. Moreover, different levels of noise lead to different errors:



For the last regime, $\sigma = 2$, noise dominates the data, which might be the

reason as to why it is β-independent. For the first two regimes, $\sigma = 0.05, 0.5$, an extension of this section could be out the application of CGMT to describe the behavior shown, which is more challenging since the value of λ probably plays a role there and therefore we cannot take the limit $\lambda \to 0$ here.

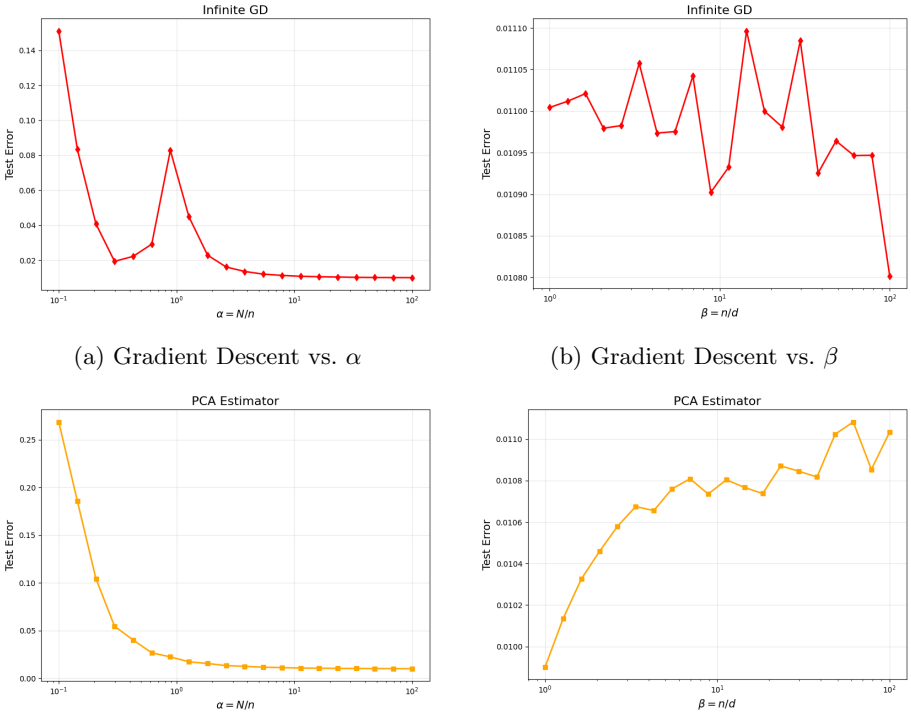
Another interesting empirical finding is the the dependence of the test error on α and σ:



It could be an interesting extensions to analyze the emergence of the bright yellow and dark blue regions in the graph.

5 Further simulations

The figure below displays estimator performance and curves.



(a) Gradient Descent vs. α

(b) Gradient Descent vs. β

(c) PCA Estimator vs. α

(d) PCA Estimator vs. β

Figure 1: Comparison of Gradient Descent and PCA estimators as a function of α and β.

In the gradient descent estimator loss, we notice a double descent and a phase transition at $\alpha = 1$. When varying β we see a more chaotic loss. For the PCA estimator, the graphs follow our intuition.

6 Future Directions

Next, we wish to study some nonlinearity in denoising. Specifically, instead of a linear denoising model, a matrix, we want to both apply a matric and then elementwise nonlinear function, such as ReLU. We believe that such a model can also be studied using known techniques by expanding the nonlinear function using the Hermite Polynomials.

7 Conclusion

In this work, we explored several high-dimensional characterizations of linear denoising, including the phase transition at $\alpha = 1$. In addition, the empirical

findings suggest non-trivial relationship of the test error to parameters α, β, σ, suggesting that denoising performance depends on signal structure and noise, without an obvious scaling law at this point.

References

[1] Alexey Dosovitskiy and et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2021.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising Diffusion Probabilistic Models". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851.

[3] Jordan Hoffmann et al. "Training Compute-Optimal Large Language Models". In: *arXiv preprint arXiv:2203.15556* (2022).

[4] Jared Kaplan et al. "Scaling Laws for Neural Language Models". In: *arXiv preprint arXiv:2001.08361* (2020).

[5] Tobit Klug and Reinhard Heckel. "Scaling Laws for Deep Learning Based Image Reconstruction". In: *arXiv preprint arXiv:2209.13435* (2022). arXiv: `2209.13435 [eess.IV]`.

[6] Marc Mézard and Andrea Montanari. *Information, Physics, and Computation*. Oxford University Press, 2009.

[7] Marc Mézard, Giorgio Parisi, and Miguel Virasoro. *Spin Glass Theory and Beyond*. World Scientific, 1987.

[8] Robin Rombach et al. "High-Resolution Image Synthesis with Latent Diffusion Models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2022).

[9] Mihailo Stojnic. "A Framework to Characterize Performance of Lasso Algorithms". In: *arXiv preprint arXiv:1301.7646* (2013).

[10] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. "Regularized Linear Regression: A Precise Analysis of the Ridge Estimator". In: *Advances in Neural Information Processing Systems* 28 (2015).

[11] Pascal Vincent et al. "Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion". In: *Journal of Machine Learning Research* 11 (2010), pp. 3371–3408.

[12] Kai Zhang et al. "Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising". In: *IEEE Transactions on Image Processing* 26.7 (2017), pp. 3142–3155.