Random matrix methods for high-dimensional machine learning models

A statistical limit theory for ML models

Antoine Bodin

May 20, 2024

EPFL

Table of contents

Introduction & motivational example

Why Random Matrix theory needed?

Random matrices are ubuquitous in machine learning theory. **But** using them is not always straightforward. **So** can we develop simple tools to analyze high-dimensional models involving random matrices?

Example: linear-model

- **Data:** $X \in \mathbb{R}^{n \times d}$ with $X_{ij} \sim \mathcal{N}(0, \frac{1}{d})$ (n samples, d features),
- **Output:** $Y = X\beta^* + \xi$ with $\xi \sim \mathcal{N}(0, \sigma^2 I_n)$ and $\frac{1}{d} \|\beta^*\|^2 = r^2$
- **Model:** $\hat{\beta} = (X^T X + \lambda I_d)^{-1} X^T Y$ (ridge regression)

Generalization error

Averaging over new random data points (x and $y = x\beta^* + \epsilon$)

$$\mathcal{E}_{\text{gen}}(\hat{\beta}) = \mathbb{E}_{\epsilon,x}\left[(x^T\beta^* + \epsilon - x^T\hat{\beta})^2\right] = \sigma^2 + \frac{1}{d}\left\|\beta^* - \hat{\beta}\right\|^2$$

The High-dimensional Limit

Average $\mathbb{E}_{X,Y}\mathcal{E}_{\text{gen}}(\hat{\beta})$ can be decomposed with Bias-variance:

$$\frac{1}{d}\mathbb{E}_{X,\xi}\left\|\beta^* - \hat{\beta}\right\|^2 = \mathcal{B}_X(\hat{\beta}) + \mathcal{V}_X(\hat{\beta}) \tag{1}$$

with:

$$\mathcal{B}_X(\hat{\beta}) = \frac{1}{d}\mathbb{E}\left[\left\|\beta^* - \mathbb{E}[\hat{\beta}|X]\right\|^2\right] \qquad \text{and} \qquad \mathcal{V}_X(\hat{\beta}) = \frac{1}{d}\mathbb{E}\left[\left\|\hat{\beta} - \mathbb{E}[\hat{\beta}|X]\right\|^2\right]$$

... After reducing everything with $\frac{1}{d}\left\|\beta^*\right\|^2 = r^2$*:*

$$\mathcal{B}_X(\hat{\beta}) = \lambda^2 r^2 \mathbb{E}_X\left\{\frac{1}{d}\text{Tr}\left[(X^T X + \lambda I)^{-2}\right]\right\} \tag{2}$$

$$\mathcal{V}_X(\hat{\beta}) = \sigma^2 \mathbb{E}_X\left\{\frac{1}{d}\text{Tr}\left[(X^T X + \lambda I)^{-1}\right] - \lambda\frac{1}{d}\text{Tr}\left[(X^T X + \lambda I)^{-2}\right]\right\} \tag{3}$$

Conclusion

- Both terms depend on the trace $\frac{1}{d}\text{Tr}\left[(X^T X + \lambda I)^{-1}\right] = g_d(-\lambda)$
- Simplifies in the high-dimensional limit $d \to \infty$.

Random matrix theory: Marchenko-Pastur law

High-dimensional limit with $\phi = \frac{n}{d}$:

$$g(z) = \lim_{d\to\infty} \frac{1}{d}\mathbb{E}\mathrm{Tr}\left[(X^TX - zI)^{-1}\right]$$

Then [Marchenko and Pastur, 1967]:

$$zg^2 + (1+z-\phi)g + 1 = 0$$

Then also the derivative w.r.t. z:

$$z(2gg' + g^2) + (1+z-\phi)g' + g = 0$$

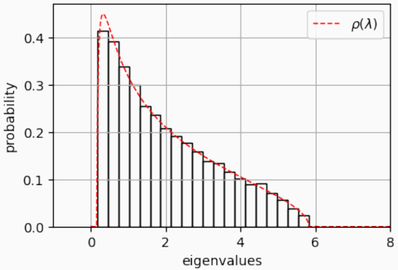


Figure 1: Marchenko-Pastur law with $\phi = 2, n = 2000, d = 1000$ and $\rho(\lambda) = \frac{1}{\pi}\,\mathrm{Im}\,g(\lambda + i0^+)$

Conclusion [Belkin et al., 2020, Hastie et al., 2019]

$$\mathcal{E}_{\text{gen}}(r,\sigma,\phi,\lambda) = \sigma^2 + \lambda^2 r^2 g'(-\lambda) + \sigma^2(g(-\lambda) - \lambda g'(-\lambda))$$

Double-descent phenomenon

Take again the algebraic equations:

$$\begin{cases} -\lambda g^2 + (1-\lambda-\phi)g + 1 = 0 \\ -\lambda(2gg' + g^2) + (1-\lambda-\phi)g' + g = 0 \end{cases}$$

Then in the limit $\lambda \to 0$:

$$\mathcal{B}_X(\hat{\beta}) = \begin{cases} r^2(1-\phi) & \text{if } \phi < 1 \\ 0 & \text{if } \phi > 1 \end{cases}$$

$$\mathcal{V}_X(\hat{\beta}) = \begin{cases} \sigma^2 \frac{\phi}{1-\phi} & \text{if } \phi < 1 \\ \sigma^2 \frac{1}{\phi-1} & \text{if } \phi > 1 \end{cases}$$

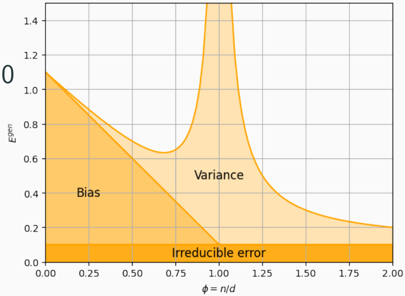


Figure 2: Test error $\sigma^2 = 0.1, r = 1$

Conclusion [Belkin et al., 2020, Hastie et al., 2019]

$$\lim_{\lambda \to 0} \mathcal{E}_{\text{gen}}(r, \sigma, \phi, \lambda) = \begin{cases} \sigma^2 + r^2(1-\phi) + \sigma^2 \frac{\phi}{1-\phi} & \text{if } \phi < 1 \\ \sigma^2 + \sigma^2 \frac{1}{\phi-1} & \text{if } \phi > 1 \end{cases}$$

More general models?

Marchenko-Pastur works well in that case, but what about more complex models?

Non-exhaustive list of interesting models:

1. Teacher-student model $Y = X\beta^*$ and $\hat{Y} = \hat{X}\beta$ with covariation structure $\Sigma = \mathbb{E}[x\hat{x}^T]$
2. Random feature model: $\hat{Y} = \sigma(WX)\beta$ to have a notion of number of parameters
3. Spike Wigner model: $Y = \beta^*{\beta^*}^T + \xi$
4. Matrix denoising: $Y = XX^T + \xi$

Is there a practical tool to analyze them?

Linear-Pencils!

Linear-Pencil Method

Level 1: Linear Pencil, the basics

Let S be a symmetric random matrix ($S_{ij} \sim \mathcal{N}(0, \frac{1}{d})$) and define Stieltjes transform of the spectral density ρ of S:

$$g(z) = \int \frac{\rho(\lambda)\mathrm{d}\lambda}{\lambda - z} = \mathrm{Tr}_d \left[(S - zI)^{-1} \right]$$

Main result [Wigner, 1958]

$$g = (z - g)^{-1} \tag{4}$$

Note 1: above eqn. often represented under the form:

$$g^2 - zg + 1 = 0$$

Note 2: $L = (S - zI)$ *is* a "trivial" linear-pencil of size 1×1! Equation (4) will follow us in the next slides.

Linear-Pencil: 3 examples

Illustration of the method with 3 examples:

1. Marchenko-Pastur:

$$g(z) = \mathrm{Tr}_d\left[(X^T X - zI_d)^{-1}\right]$$

2. Sample Covariance Matrix:

$$g(z) = \mathrm{Tr}_d\left[(C^{\frac{1}{2}} X^T X C^{\frac{1}{2}} - zI_d)^{-1}\right]$$

3. Random Feature Kernel:

$$g(z) = \mathrm{Tr}_d\left[(\sigma(WX)\sigma(WX)^T - zI)^{-1}\right]$$

Each case involves algebraic operations on a matrix level.

The linearization method

Back to Marchenko-Pastur: $L = (X^T X - zI)$ is not a linear-pencil, because it is not "linear" in X.

But we can linearize it with a block-matrix!

Consider $U_1 = (X^T X - zI)^{-1}$ with:

$$\begin{cases} (X^T X - zI)U_1 & = I \\ U_2 & = XU_1 \end{cases}$$

Then we obtain our linear-pencil L:

$$\underbrace{\begin{pmatrix} -zI & X^T \\ X & -I \end{pmatrix}}_{L} \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} = \begin{pmatrix} I \\ 0 \end{pmatrix}$$

And in particular, $\mathrm{Tr}_d\,[U_1]$ is given by the partial trace of L^{-1} because:

$$U_1 = \begin{pmatrix} I & 0 \end{pmatrix} (L^{-1}) \begin{pmatrix} I \\ 0 \end{pmatrix} = (L^{-1})^{(11)}$$

Linear Pencil: main result

Step 1: Encode matrix into a wider block-matrix (variance $\frac{1}{d}$):

$$L = \begin{pmatrix} -zI_d & X^T \\ X & -I_n \end{pmatrix}$$

Step 2: Extract deterministic blocks from random matrices:

$$L_0 = \begin{pmatrix} -zI_d & 0 \\ 0 & -I_n \end{pmatrix}$$

Step 3: Calculate block-wise covariance structure of L:

$$\sigma_{12}^{21} = \sigma_{21}^{12} = 1 \qquad \sigma_{11}^{11} = \sigma_{22}^{22} = 0$$

Step 4: Define the partial-trace of the inverse, and interesting operator:

$$g_{ij} = \mathrm{Tr}_d \left[(L^{-1})^{(ij)} \right] \qquad [\eta_L(g)]_{il} = \sum_{jk \in \mathbb{S}} \sigma_{ij}^{kl} g_{jk}$$

Here, g_{11} is clearly the quantity of interest, and:

$$\eta_L(g) = \begin{pmatrix} \sigma_{12}^{21} g_{22} & 0 \\ 0 & \sigma_{21}^{12} g_{11} \end{pmatrix}$$

Linear Pencil: Main result

Main result [Rashidi Far et al., 2006, Mingo and Speicher, 2017]

$$g = (J \otimes \mathrm{Tr}_d)\left[(L_0 - (\eta_L \otimes I)(g))^{-1}\right] \tag{5}$$

1. $(\eta_L \otimes I)(\cdot)$ is the matrix operator acting on g:

$$((\eta_L \otimes I)(g))^{(ij)} = \begin{cases} [\eta_L(g)]_{ij} I_{N_i} & \text{if } N_i = N_j \\ 0_{N_i, N_j} & \text{otherwise} \end{cases}$$

So in our example:

$$L_0 - (\eta_L \otimes I)(g) = \begin{pmatrix} -zI_d - [\eta_L(g)]_{11} I_d & 0 \\ 0 & -I_n - [\eta_L(g)]_{22} I_n \end{pmatrix}$$

2. $J \odot \mathrm{Tr}_d\,[\cdot]$ is the partial trace operator:

$$[(J \otimes \mathrm{Tr}_d)\,[G]]_{ij} = \begin{cases} \mathrm{Tr}_d\left[G^{(ij)}\right] & \text{if } N_i = N_j \\ 0 & \text{otherwise} \end{cases}$$

Linear Pencil, example 1: Marchenko-Pastur

- From **step 4** we have defined:

$$g_{11} = \mathrm{Tr}_d\left[(X^T X - zI_d)^{-1}\right] \tag{6}$$

- From application of **Main result** we find:

$$\begin{pmatrix} g_{11} & 0 \\ 0 & g_{22} \end{pmatrix} = (J \otimes \mathrm{Tr}_d)\left[\begin{pmatrix} -zI_d - g_{22} I_d & 0 \\ 0 & -I_n - g_{11} I_n \end{pmatrix}^{-1}\right]$$

Conclusion

With $g_C(z) = \mathrm{Tr}_d\left[(C - zI_d)^{-1}\right]$:

$$\begin{cases} g_{11} = \mathrm{Tr}_d\left[(-zI_n - g_{22} I_n)^{-1}\right] = -(z + g_{22})^{-1} \\ g_{22} = \mathrm{Tr}_d\left[(-I_n - g_{11} I_n)^{-1}\right] = -\phi\,(1 + g_{11})^{-1} \end{cases}$$

$$g_{11} = -\frac{1}{z - \phi(1 + g_{11})^{-1}} \implies zg_{11}^2 + (1 + z - \phi)g_{11} + 1 = 0$$

Linear-pencil, example 2: Sample covariance Matrix

Linear-pencil works for more general matrices: for instance, the sample covariance matrix.

Quite often, real-world problems come from samples $u \sim \mathcal{N}(0, C)$. An equivalent model is $u = C^{\frac{1}{2}} x$ with $x \sim \mathcal{N}(0, I_d)$.

Problem

Given n samples $U = (u_1, \ldots, u_n)$ an estimation of C is given by:

$$\tilde{C} = \frac{1}{d} U U^T = C^{\frac{1}{2}} X X^T C^{\frac{1}{2}}$$

with X a random matrix. How to characterize the eigenvalue distribution of $\tilde{C}$ compared to C?

Linear-pencils allow the study of the Stieltjes transform of $\tilde{C}$:

$$g(z) = \mathrm{Tr}_d \left[(C^{\frac{1}{2}} X X^T C^{\frac{1}{2}} - z I_d)^{-1} \right]$$

Linear-pencil, example 2: Sample covariance Matrix

Let's consider instead:

$$g(z) = \mathrm{Tr}_d \left[(X^T C X - z I_n)^{-1} \right]$$

In a similar way with $h(z) = \mathrm{Tr}_d \left[U_1 \right]$ and the linearization:

$$\underbrace{\begin{pmatrix} -zI_n & 0 & X \\ 0 & C & I_d \\ X^T & I_d & 0 \end{pmatrix}}_{L} \begin{pmatrix} U_1 \\ U_2 \\ U_3 \end{pmatrix} = \begin{pmatrix} I_n \\ 0 \\ 0 \end{pmatrix}$$

and with $\sigma_{13}^{31} = \sigma_{31}^{13} = 1$:

$$L_0 = \begin{pmatrix} -zI_n & 0 & 0 \\ 0 & C & I \\ 0 & I & 0 \end{pmatrix} \qquad \text{and} \qquad \eta_L(g) = \begin{pmatrix} g_{33} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & g_{11} \end{pmatrix}$$

Linear-pencil, example 2: Sample covariance Matrix

$$M = (L_0 - (\eta_L \otimes I)(g))^{-1} = \begin{pmatrix} (-z - g_{33})I_n & 0 & 0 \\ 0 & C & I_d \\ 0 & I_d & -g_{11} I_d \end{pmatrix}^{-1}$$

Using a block-matrix inversion formula:

$$M = \begin{pmatrix} -(z + g_{33})^{-1} I_n & 0 & 0 \\ 0 & (C + g_{11}^{-1})^{-1} & (g_{11} C + I_d)^{-1} \\ 0 & (g_{11} C + I_d)^{-1} & -g_{11}^{-1}(I_d - (g_{11} C + I_d)^{-1}) \end{pmatrix}$$

Thus, formula $g = (J \otimes \mathrm{Tr}_d)[M]$ from **Main result** yields:

$$\begin{cases} g_{11} & = g(z) = -\phi(z + g_{33})^{-1} \\ g_{33} & = -\frac{1}{g_{11}}(1 - \mathrm{Tr}_d\left[(g_{11} C + I_d)^{-1}\right]) = \frac{1}{g(z)}(\frac{1}{g(z)} g_C(-\frac{1}{g(z)}) - 1) \end{cases}$$

$$\phi + z\phi g(z) + \frac{1}{g(z)} g_C\left(\frac{-1}{g(z)}\right) = 1$$

Linear-pencil, example 3: Gaussian Equivalence Principle

How to calculate $g(z) = \mathrm{Tr}_d\left[(\sigma(WX)\sigma(WX)^T - zI_d)^{-1}\right]$?

GEP [Pennington and Worah, 2017]

$$\sigma(WX) \equiv \mu WX + \nu\Omega \tag{7}$$

With: $0 = \langle \sigma, H_{e_0} \rangle$ $\quad \mu = \langle \sigma, H_{e_1} \rangle$ $\quad \mu^2 + \nu^2 = \langle \sigma, \sigma \rangle$

Ex.: $U = (\mu W, \nu\Omega)$, $V = (X^T, I)$ so $g(z) = \mathrm{Tr}_d\left[(UV^TVU^T - zI_d)^{-1}\right]$

$$L = \begin{pmatrix} -zI & 0 & 0 & U \\ 0 & 0 & V^T & I \\ 0 & V & I & 0 \\ U^T & I & 0 & 0 \end{pmatrix} = \begin{pmatrix} -zI & 0 & 0 & 0 & \mu W & \nu\Omega \\ 0 & 0 & 0 & X & I & 0 \\ 0 & 0 & 0 & I & 0 & I \\ 0 & X^T & I & I & 0 & 0 \\ \mu W^T & I & 0 & 0 & 0 & 0 \\ \nu\Omega^T & 0 & I & 0 & 0 & 0 \end{pmatrix}$$

Note: no unique encoding! In fact, possible with a 4×4 linear-pencil

GEP: Construction of 4×4 linear pencil

Consider $\mathrm{Tr}_d\left[(\sigma(XW)\sigma(XW)^T - zI)^{-1}\right] = \mathrm{Tr}_d\left[U_0\right]$:

$$\begin{aligned}((\mu WX + \nu\Omega)(\mu X^T W^T + \nu\Omega^T) - zI)U_0 &= I \\ U_1 &= \mu W^T U_0 \\ U_2 &= X^T U_1 + \nu\Omega^T U_0 \\ U_3 &= XU_2\end{aligned}$$

So first equation is: $\mu WU_3 + \nu\Omega U_2 - zU_0 = I$, so:

$$\begin{pmatrix} -zI & 0 & \nu\Omega & \mu W \\ 0 & 0 & X & -I \\ \nu\Omega^T & X^T & -I & 0 \\ \mu W^T & -I & 0 & 0 \end{pmatrix} \begin{pmatrix} U_0 \\ U_1 \\ U_2 \\ U_3 \end{pmatrix} = \begin{pmatrix} I \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

The dynamics of general linear models

Random Feature Model is just a structured linear model

Gaussian covariate model [Loureiro et al., 2021]

A mutual source of randomness Z, and two structures A and B:

$$Y = X\beta^* = (ZB)\beta^* \qquad \hat{Y} = \tilde{X}\beta = (ZA)\beta$$

With the generalization error:

$$\mathcal{E}_{\text{gen}}(\beta) = \mathbb{E}\left[(x^T\beta^* - \tilde{x}^T\beta)^2\right]$$

Ex. Random feature model with $Z = \begin{pmatrix} X_0 & \Omega & \xi \end{pmatrix}$ and:

$$\beta^* = \begin{pmatrix} \beta_0^* \\ 1 \end{pmatrix} \qquad B = \begin{pmatrix} I_d & O \\ O & O \\ 0 & \sigma \end{pmatrix} \qquad A = \begin{pmatrix} \mu W \\ \nu I_N \\ 0 \end{pmatrix}$$

Then:

$$Y = X_0\beta_0^* + \xi \qquad \hat{Y} = (\mu W X_0 + \nu\Omega)\,\beta \equiv \sigma(WX_0)\beta$$

The dynamics of the Gaussian covariate model

Main result [Bodin and Macris, 2022]

With gradient flow on: $\mathcal{E}_{\text{train}}^{\lambda}(\beta) = \frac{1}{n}\|Y - \hat{Y}(\beta)\|^2 + \frac{\lambda}{n}\|\beta\|^2$

$$\bar{\mathcal{E}}_{\text{gen}}(t) = c_0 + r_0^2 \mathcal{B}_0(t) + \mathcal{B}_1(t)$$

With random initialization $\beta(t=0) \sim \mathcal{N}(0, r_0^2 \frac{1}{d} I_d)$:

$$\mathcal{B}_1(t) = \frac{-1}{4\pi^2} \oint_\Gamma \oint_\Gamma \frac{(1-e^{-t(x+\lambda)})(1-e^{-t(y+\lambda)})}{(x+\lambda)(y+\lambda)} f_1(x,y)\mathrm{d}x\mathrm{d}y + \frac{1}{i\pi}\oint_\Gamma \frac{1-e^{-t(z+\lambda)}}{z+\lambda} f_2(z)\mathrm{d}z$$

$$\mathcal{B}_0(t) = \frac{-1}{2i\pi}\oint_\Gamma e^{-2t(z+\lambda)} f_0(z)\mathrm{d}z$$

And with $V = B\beta^*\beta^{*T}B^T$, $U = AA^T$, $c_0 = \mathrm{Tr}_d[V]$ and the self-consistent equations:

$$\begin{aligned}
f_1(x,y) &= f_2(x) + f_2(y) + \tilde{f}_1(x,y) - c_0 \\
\tilde{f}_1(x,y) &= \mathrm{Tr}_d\left[(\phi U + \zeta_x I)^{-1}(\zeta_x\zeta_y V^* + \tilde{f}_1(x,y)\phi U^2)(\phi U + \zeta_y I)^{-1}\right] \\
f_2(z) &= c_0 - \mathrm{Tr}_d\left[\zeta_z V^*(\phi U + \zeta_z I)^{-1}\right] \\
f_0(z) &= -\left(1 + \frac{\zeta_z}{z}\right) \\
\zeta_z &= -z + \mathrm{Tr}_d\left[\zeta_z U(\phi U + \zeta_z I)^{-1}\right]
\end{aligned}$$

The limit $t \to +\infty$ of the Gaussian covariate model

Secondary result [Bodin and Macris, 2022]

$$\bar{\mathcal{E}}_{\text{gen}}(+\infty) = c_0 + f_1(-\lambda, -\lambda) + 2f_2(-\lambda) = \tilde{f}_1$$
$$\bar{\mathcal{E}}^0_{\text{train}}(+\infty) = \lambda^2 \zeta^{-2} \bar{\mathcal{E}}_{\text{gen}}(+\infty)$$

With $U_\star = \phi A^T A$ and $\Xi = \phi A^T B$ and:

$$f_1 = 2f_2 + \tilde{f}_1 - c_0$$
$$f_1 = \mathrm{Tr}_n \left[((\Xi \beta^* \beta^{*T} \Xi^T) + \tilde{f}_1 U_\star) U_\star (U_\star + \zeta I)^{-2} \right]$$
$$f_2 = \mathrm{Tr}_n \left[(\Xi \beta^* \beta^{*T} \Xi^T)(U_\star + \zeta I)^{-1} \right]$$
$$\zeta = \lambda + \mathrm{Tr}_n \left[\zeta U_\star (U_\star + \zeta I)^{-1} \right]$$

See also [Loureiro et al., 2021]

Example 1: Random feature - model/sample double-descents

Ex.: Triple descent [d'Ascoli et al., 2020, Bodin and Macris, 2021]

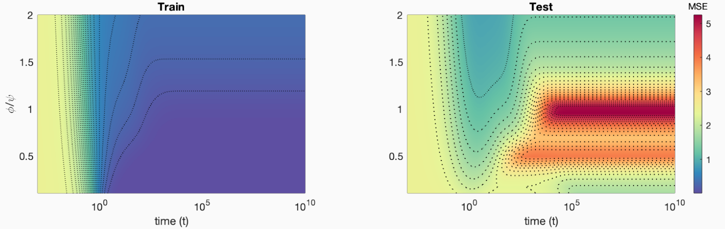


Figure 3: Parameters: $(\mu, \nu, \psi, r, s, \lambda) = (0.9, 0.1, 2, 1, 0.8, 0.0001)$

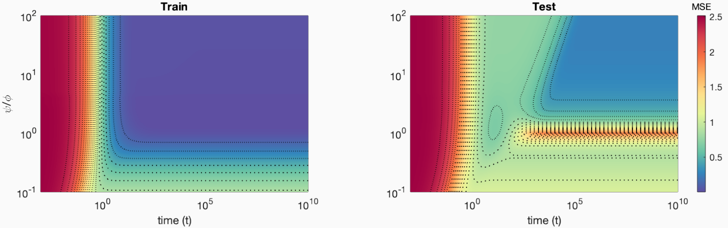
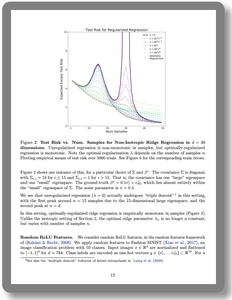


Figure 4: Parameters $(\mu, \nu, \phi, r, s, \lambda) = (0.5, 0.3, 3, 2., 0.4, 0.001)$.

Example 2: Non-Isotropic regression - Descent structures

$$B = I \quad \text{and} \quad A = \mathrm{diag}(\underbrace{\alpha^{-\frac{0}{2}}, \dots, \alpha^{-\frac{0}{2}}}_{\frac{d}{p}\text{times}}, \underbrace{\alpha^{-\frac{1}{2}}, \dots, \alpha^{-\frac{1}{2}}}_{\frac{d}{p}\text{times}}, \alpha^{-\frac{2}{2}}, \dots, \alpha^{-\frac{p-1}{2}})$$



Test Risk for Regularized Regression

Expected Excess Test Risk

Num Samples

Figure 2: **Test Risk vs. Num. Samples for Non-Isotropic Ridge Regression in $d = 30$ dimensions.** Unregularized regression is non-monotonic in samples, but optimally-regularized regression is monotonic. Note the optimal regularization λ depends on the number of samples n. Plotting empirical means of test risk over 5000 trials. See Figure 6 for the corresponding train errors.

Figure 2 shows one instance of this, for a particular choice of Σ and β^*. The covariance Σ is diagonal, with $\Sigma_{i,i} = 10$ for $i \leq 15$ and $\Sigma_{i,i} = 1$ for $i > 15$. That is, the covariance has one "large" eigenspace and one "small" eigenspace. The ground-truth $\beta^* = 0.1\vec{e_1} + \vec{e_{30}}$, which lies almost entirely within the "small" eigenspace of Σ. The noise parameter is $\sigma = 0.5$.

We see that unregularized regression ($\lambda = 0$) actually undergoes "triple descent"[2] in this setting, with the first peak around $n = 15$ samples due to the 15-dimensional large eigenspace, and the second peak at $n = d$.

In this setting, optimally-regularized ridge regression is empirically monotonic in samples (Figure 2). Unlike the isotropic setting of Section 2, the optimal ridge parameter λ_n is no longer a constant, but varies with number of samples n.

Random ReLU Features. We consider random ReLU features, in the random features framework of (Rahimi & Recht, 2008). We apply random features to Fashion-MNIST (Xiao et al., 2017), an image classification problem with 10 classes. Input images $x \in \mathbb{R}^d$ are normalized and flattened to $[-1, 1]^d$ for $d = 784$. Class labels are encoded as one-hot vectors $y \in \{\vec{e_1}, \dots \vec{e_{10}}\} \subset \mathbb{R}^{10}$. For a

[2] See also the "multiple descent" behavior of kernel interpolants in Liang et al. (2020).

12

[Nakkiran et al., 2020]

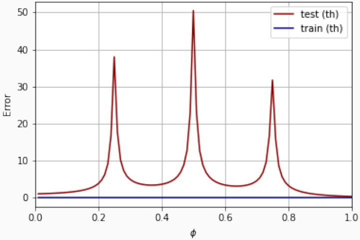


Figure 5: $p = 4, \lambda = 10^{-13}$, $\alpha = 10^4$

Result with $\alpha \to +\infty$ [Bodin and Macris, 2022]

$$\bar{\mathcal{E}}_{\text{gen}}(+\infty) = \frac{1}{p} \sum_{k=0}^{p-1} \frac{\phi(1-\phi)}{\left(\phi - \frac{k}{p}\right)\left(\frac{k+1}{p} - \phi\right)} \mathbb{1}_{]\frac{k}{p}; \frac{k+1}{p}[}(\phi) - \phi + o_\alpha(1)$$

Example 3: Realistic datasets

- **Data:** MNIST dataset $X_0 \in \mathbb{R}^{n_{tot} \times d}$ with normalized entries and $n_{tot} = 60'000$ and $d = 28 \times 28 = 784$
- **Aim:** Learning the parity ($y = \pm 1$ for even/odd numbers)
- **Structure:** cov. matrix $U_\star \simeq \frac{1}{n_{tot}} X^T X$ and $\Xi \beta^* \simeq X^T Y$ (since $X^T X = A^T (Z^T Z) A$ and $X^T Y = A(Z^T Z) B \beta^*$ and $n_{tot} \gg d$)

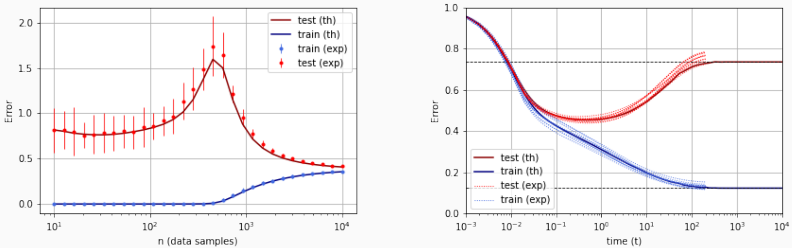


Figure 6: Comparison between the analytical and experimental learning profiles for the minimum least-squares estimator at $\lambda = 10^{-3}$ on the left (20 runs) and the time evolution at $\lambda = 10^{-2}$, $n = 700$ on the right (10 runs).

Questions?

Appendix: Gaussian Equivalence Principle

From [Lu and Yau, 2022, Mei and Montanari, 2019] with $y, x \in \mathbb{S}^{d-1}$, decomposition with Gegenbauer polynomials:

$$\sigma(\sqrt{d}x^T y) = \sum_{k=1}^{p} \alpha_k q_k(\sqrt{d}x^T y)$$

(In high-dimenion, α_k are related to the coefficients of the Hermite polynomials decomposition).
Decoupling with spherical harmonics $Y_{k,a}$:

$$q_k(\sqrt{d}x^T y) = \frac{1}{\sqrt{N_k}} \sum_{a=0}^{N_k} Y_{k,a}(x) Y_{k,a}(y)$$

Appendix: Bias-variance decomposition

Classical bias-variance decomposition over new data point x:

$$\mathbb{E}[(y-\hat{y})^2] = \mathbb{E}[y-\hat{y}]^2 + \mathrm{Var}(\hat{y}) + \mathrm{Var}(y) \tag{8}$$

Is true when:

$$\mathrm{Var}(y-\hat{y}) = \mathrm{Var}(\hat{y}) + \mathrm{Var}(y) \implies \mathrm{Cov}(y,\hat{y}) = 0 \tag{9}$$

This is true for the R.V. $\beta_0, \xi, X, \epsilon$ (because each of them only affect either y and $\hat{y}$) but not for x (and β^* if considered random). So in fact we implicitly always require the conditional expectation on x:

$$\mathbb{E}[(y-\hat{y})^2] = \mathbb{E}_x\left[\mathbb{E}[y-\hat{y}|x]\right]^2 + \mathbb{E}_x\mathrm{Var}(\hat{y}|x) + \mathbb{E}_x\mathrm{Var}(y|x) \tag{10}$$

Belkin, M., Hsu, D., and Xu, J. (2020).
Two models of double descent for weak features.
SIAM Journal on Mathematics of Data Science, 2(4):1167–1180.

Bodin, A. and Macris, N. (2021).
Model, sample, and epoch-wise descents: exact solution of gradient flow in the random feature model.
Advances in Neural Information Processing Systems, 34.

Bodin, A. and Macris, N. (2022).
Gradient flow in the gaussian covariate model: exact solution of learning curves and multiple descent structures.

d'Ascoli, S., Sagun, L., and Biroli, G. (2020).
Triple descent and the two kinds of overfitting: where and why do they appear?
In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3058–3069. Curran Associates, Inc.

Hastie, T., Montanari, A., Rosset, S., and Tibshirani, R. J. (2019).
Surprises in High-Dimensional Ridgeless Least Squares Interpolation.
arXiv e-prints, page arXiv:1903.08560.

Loureiro, B., Gerbelot, C., Cui, H., Goldt, S., Krzakala, F., Mezard, M., and Zdeborová, L. (2021).
Learning curves of generic features maps for realistic datasets with a teacher-student model.
Advances in Neural Information Processing Systems, 34:18137–18151.

Lu, Y. M. and Yau, H.-T. (2022).
An equivalence principle for the spectrum of random inner-product kernel matrices.
arXiv preprint arXiv:2205.06308.

Marchenko, V. A. and Pastur, L. A. (1967).
Distribution of eigenvalues for some sets of random matrices.
Matematicheskii Sbornik, 114(4):507–536.

Mei, S. and Montanari, A. (2019).
The generalization error of random features regression: Precise asymptotics and double descent curve.
arXiv e-prints, page arXiv:1908.05355.

Mingo, J. A. and Speicher, R. (2017).
Free probability and random matrices, volume 35.
Springer.

Nakkiran, P., Venkat, P., Kakade, S. M., and Ma, T. (2020).
Optimal regularization can mitigate double descent.
In *International Conference on Learning Representations*.

Pennington, J. and Worah, P. (2017).
Nonlinear random matrix theory for deep learning.
In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 2637–2646.

Rashidi Far, R., Oraby, T., Bryc, W., and Speicher, R. (2006).
Spectra of large block matrices.
arXiv e-prints, page cs/0610045.

Wigner, E. P. (1958).
On the distribution of the roots of certain symmetric matrices.
Annals of Mathematics, 67(2):325–327.