DEEP LEARNING AND FREE PROBABILITY: TRAINING AND GENERALIZATION DYNAMICS IN HIGH DIMENSIONS

JEFFREY PENNINGTON

GOOGLE BRAIN

COLUMBIA

11-1-19

OUTLINE

1. Motivation / Introduction
2. Case Study: Linear Regression
3. Linearization pt 1: High-Dimensional Kernels
4. Linearization pt 2: The Linear Pencil
5. Linearization pt 3: Neural Tangent Kernel

OUTLINE

1. Motivation / Introduction
2. Case Study: Linear Regression
3. Linearization pt 1: High-Dimensional Kernels
4. Linearization pt 2: The Linear Pencil
5. Linearization pt 3: Neural Tangent Kernel

DATASETS ARE OFTEN HIGH-DIMENSIONAL

Many common datasets have both a large number of samples **and** a large number of features

- CIFAR-10 (10^5 samples, 10^4 features)
- Imagenet (10^7 samples, 10^5 features)

DATASETS ARE OFTEN HIGH-DIMENSIONAL

Many common datasets have both a large number of samples **and** a large number of features

- CIFAR-10 (10^5 samples, 10^4 features)
- Imagenet (10^7 samples, 10^5 features)

Many modalities are intrinsically high-dimensional:

- Speech (high frequency, large dynamic range)
- Video (high frame rates, high resolution)
- DNA sequences (large number of base pairs)

DEEP LEARNING MODELS ARE HIGH-DIMENSIONAL

Deep learning models employ large numbers of parameters.
At least two practically-relevant high-dimensional regimes:

1. Linearly overparameterized ($p \sim m$)
2. Quadratically overparameterized ($n_l \sim m$)

Examples:

	Width n_l	# Samples m	# Parameters p
FC/ CIFAR-10	10^3	10^4	10^6
ResNet/ ImageNet	10^3	10^7	10^8

HIGH-DIMENSIONAL SCALING LIMITS

We will focus on the following high-dimensional asymptotics of zero and one hidden-layer networks:

1. Dataset size $m \to \infty$
2. Input dimensionality $n_0 \to \infty$
3. Hidden-layer size $n_1 \to \infty$

with the ratios $\phi = \dfrac{n_0}{m}$ and $\psi = \dfrac{n_0}{n_1}$ held constant

MARCHENKO–PASTUR DISTRIBUTION

In the low-dimensional (standard) regime, certain statistics may be simple:

- Dataset $X \in \mathbb{R}^{n_0 \times m}$, $X_{ij} \sim \mathcal{N}(0,1)$
- For n_0 finite, infinite samples ($m \to \infty$), $\frac{1}{m} XX^T \to I_{n_0}$

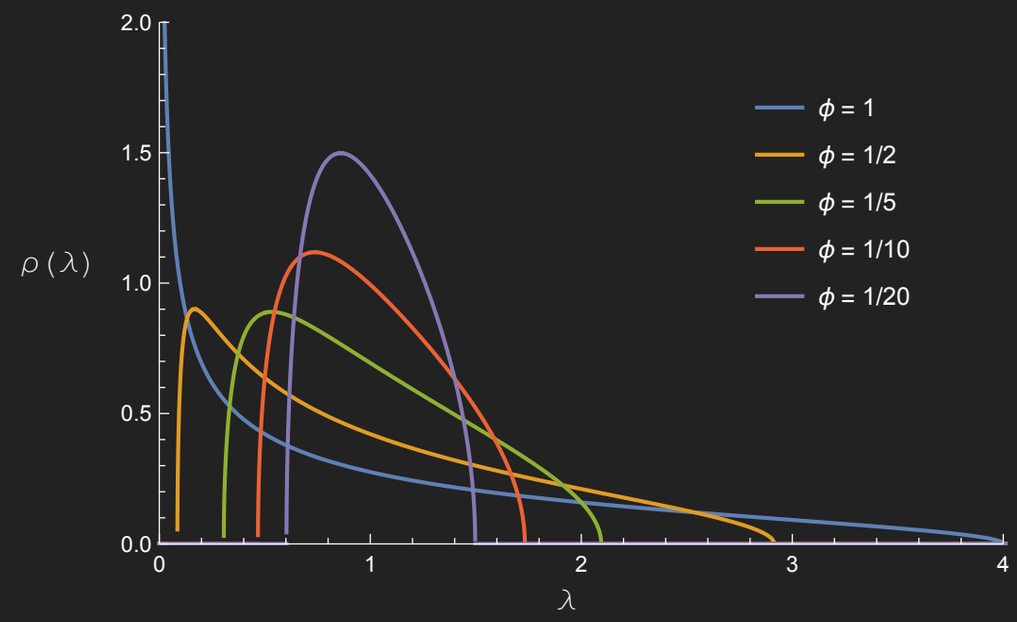
MARCHENKO-PASTUR DISTRIBUTION

In the low-dimensional (standard) regime, certain statistics may be simple:

- Dataset $X \in \mathbb{R}^{n_0 \times m}$, $X_{ij} \sim \mathcal{N}(0,1)$
- For n_0 finite, infinite samples ($m \to \infty$), $\frac{1}{m} XX^T \to I_{n_0}$

In the high-dimensional regime, spectrum can be non-trivial

- $n_0/m \to \phi$ as $n_0, m \to \infty$
- $\rho(\frac{1}{m} XX^T) \to MP(\phi)$



OUTLINE

1. Motivation / Introduction
2. Case Study: Linear Regression
3. Linearization pt 1: High-Dimensional Kernels
4. Linearization pt 2: The Linear Pencil
5. Linearization pt 3: Neural Tangent Kernel

LINEAR REGRESSION

Consider one of the simplest possible learning problems, linear ridge regression with iid Gaussian inputs and targets.

$$L = \|WX - Y\|_F^2 + \gamma \|W\|_F^2 , \qquad X_{ij} \sim \mathcal{N}(0,1) , \; Y_{ij} \sim \mathcal{N}(0,1)$$

LINEAR REGRESSION

Consider one of the simplest possible learning problems, linear ridge regression with iid Gaussian inputs and targets.

$$L = \|WX - Y\|_F^2 + \gamma \|W\|_F^2 , \qquad X_{ij} \sim \mathcal{N}(0,1) , \; Y_{ij} \sim \mathcal{N}(0,1)$$

$$W^* = YQX^T , \qquad Q = (X^T X + \gamma I)^{-1}$$

$$E_{train} = \|W^*X - Y\|_F^2 = tr[(YQX^T X - Y)^T (YQX^T X - Y)]$$

$$= tr[X^T X Q Y^T Y Q X^T X] - 2tr[X^T X Q Y^T Y] + tr[Y^T Y]$$

LINEAR REGRESSION

Consider one of the simplest possible learning problems, linear ridge regression with iid Gaussian inputs and targets.

$$L = \|WX - Y\|_F^2 + \gamma \|W\|_F^2 , \qquad X_{ij} \sim \mathcal{N}(0,1) , \; Y_{ij} \sim \mathcal{N}(0,1)$$

$$W^* = YQX^T , \qquad Q = (X^T X + \gamma I)^{-1}$$

$$\begin{aligned} E_{train} &= \|W^*X - Y\|_F^2 = tr[(YQX^TX - Y)^T(YQX^TX - Y)] \\ &= tr[X^TXQY^TYQX^TX] - 2tr[X^TXQY^TY] + tr[Y^TY] \\ &= tr[X^TXQ^2X^TX] - 2tr[X^TXQ] + 1 \\ &= \gamma^2 tr[Q^2] \\ &= -\gamma^2 \partial_\gamma tr[Q] \end{aligned}$$

$$X^TX = Q^{-1} - \gamma I$$

RESOLVENT AND STIELTJES TRANSFORM

The training error depends on the trace of the resolvent Q

$$E_{train} = -\gamma^2 \partial_\gamma tr[Q] \qquad Q = (X^T X + \gamma I)^{-1}$$

This trace $tr[Q]$ is known as the Cauchy transform $G : \mathbb{C}^+ \to \mathbb{C}^+$,

$$G(z) = -tr[(X^T X - zI)^{-1}] = \int \frac{1}{z-\lambda} \rho_{X^T X}(\lambda) d\lambda$$

RESOLVENT AND STIELTJES TRANSFORM

The training error depends on the trace of the resolvent Q

$$E_{train} = -\gamma^2 \partial_\gamma tr[Q] \qquad Q = (X^T X + \gamma I)^{-1}$$

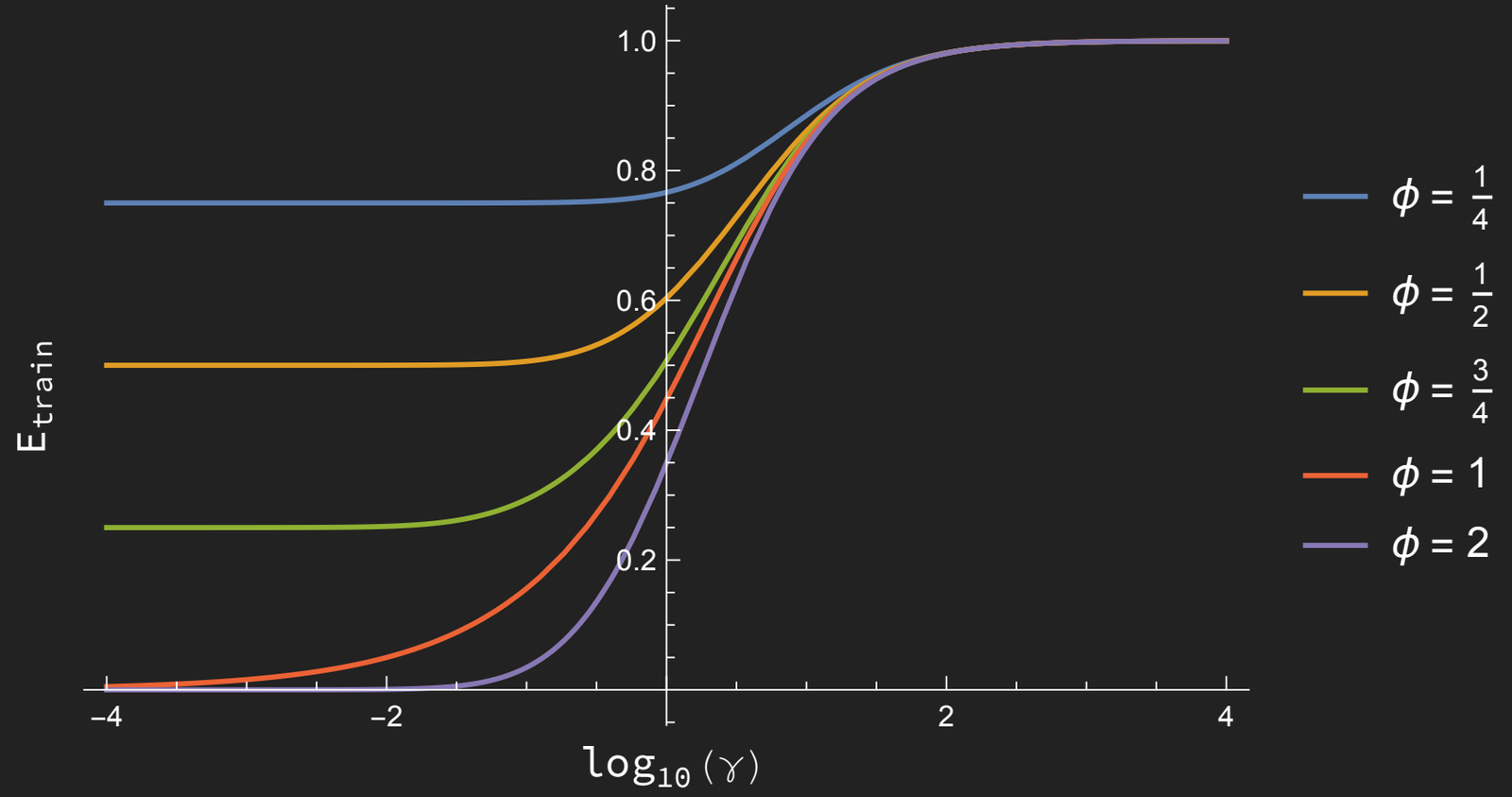
This trace $tr[Q]$ is known as the Cauchy transform $G : \mathbb{C}^+ \to \mathbb{C}^+$,

$$G(z) = -tr[(X^T X - zI)^{-1}] = \int \frac{1}{z-\lambda} \rho_{X^T X}(\lambda) d\lambda$$

$$= \frac{1 - (1-z)\phi + \sqrt{(1-(1-z)\phi)^2 - 4z\phi}}{2z} \qquad \phi = \frac{n_0}{m}$$

HIGH-DIMENSIONAL TRAINING ERROR

$$E_{train} = \frac{\sqrt{(\gamma\phi + \phi - 1)^2 + 4\gamma\phi(\phi(\gamma\phi + \gamma + \phi - 2) + 1)}}{2\phi(\gamma((\gamma + 2)\phi + 2) + \phi - 2) + 2} + \frac{1 - \phi}{2} \qquad \phi = \frac{n_0}{m}$$



GRADIENT DESCENT

Let's optimize the regression weights using gradient descent.

$$L = \|WX - Y\|_F^2 + \gamma \|W\|_F^2, \qquad X_{ij} \sim \mathcal{N}(0,1), \; Y_{ij} \sim \mathcal{N}(0,1)$$

$$W(t) = YQ(t)X^T$$

$$Q(t) = K^{-1}(I - (I - 2\eta K)^t)$$

$$K = X^T X + \gamma I$$

GRADIENT DESCENT

Let's optimize the regression weights using gradient descent.

$$L = \|WX - Y\|_F^2 + \gamma \|W\|_F^2, \qquad X_{ij} \sim \mathcal{N}(0,1), \; Y_{ij} \sim \mathcal{N}(0,1)$$

$$W(t) = YQ(t)X^T \qquad Q(t) = K^{-1}(I - (I - 2\eta K)^t)$$

$$K = X^T X + \gamma I$$

Now the training error has a simple time-dependent expression:

$$\begin{aligned} E_{train}(t) &= tr[X^T X Q(t)^2 X^T X] - 2tr[X^T X Q(t)] + 1 \\ &= tr[(K - \gamma I)^2 Q(t)^2] - 2tr[(K - \gamma I) Q(t)] + 1 \\ &= tr[K^{-2}((K - \gamma I)(I - 2\eta K)^t + \gamma I)] \end{aligned}$$

TIME-DEPENDENCE THROUGH CAUCHY'S FORMULA

$$E_{train}(t) = tr[f(K)] \qquad f(K) = K^{-2}((K - \gamma I)(I - 2\eta K)^t + \gamma I)^2$$

Recalling Cauchy's integral formula for matrix functions,

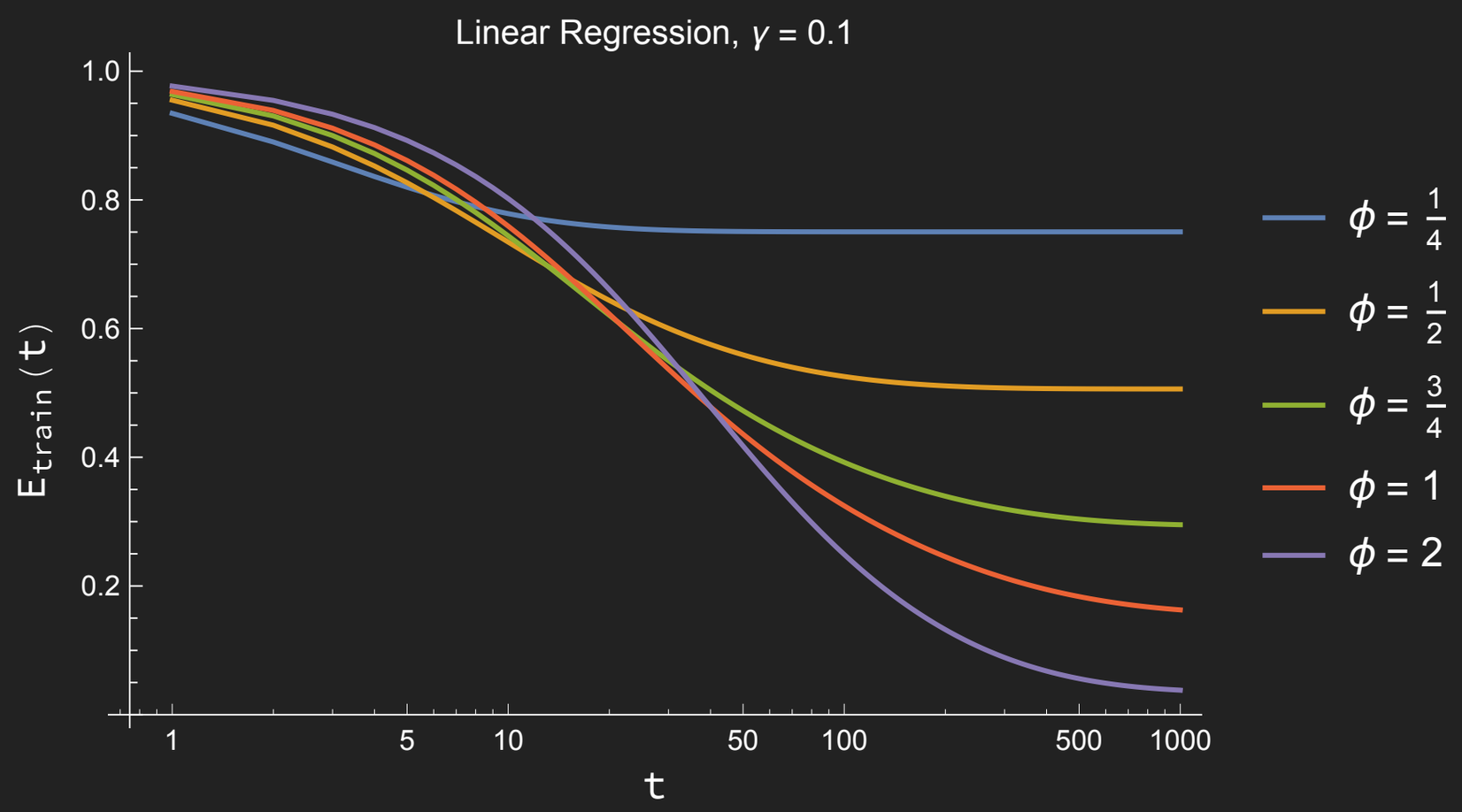
$$f(A) = \frac{1}{2\pi i}\int_C f(z)(A - zI)^{-1}dz$$

Taking the trace of this equation gives,

$$E_{train}(t) = \frac{1}{2\pi i}\int_C \frac{((z - \gamma)(1 - 2\eta z)^t + \gamma)^2}{z^2} G(z - \gamma)dz$$

TIME-DEPENDENCE THROUGH CAUCHY'S FORMULA

$$E_{train}(t) = \frac{1}{2\pi i}\int_C \frac{((z-\gamma)(1-2\eta z)^t+\gamma)^2}{z^2} G(z-\gamma)dz$$



OUTLINE

1. Motivation / Introduction
2. Case Study: Linear Regression
3. Linearization pt 1: High-Dimensional Kernels
4. Linearization pt 2: The Linear Pencil
5. Linearization pt 3: Neural Tangent Kernel

KERNEL RIDGE REGRESSION

Consider random nonlinear features $F = f(W_1 X)$

$$L = \|WF - Y\|_F^2 + \gamma \|W\|_F^2, \qquad X_{ij}, Y_{ij}, [W_1]_{ij} \sim \mathcal{N}(0,1)$$

$$W(t) = YQ(t)F^T$$

$$Q(t) = K^{-1}(I - (I - 2\eta K)^t)$$

$$K = F^T F + \gamma I$$

KERNEL RIDGE REGRESSION

Consider random nonlinear features $F = f(W_1 X)$

$$L = \|WF - Y\|_F^2 + \gamma \|W\|_F^2, \qquad X_{ij}, Y_{ij}, [W_1]_{ij} \sim \mathcal{N}(0,1)$$

$$W(t) = YQ(t)F^T \qquad Q(t) = K^{-1}(I - (I - 2\eta K)^t)$$

$$K = F^T F + \gamma I$$

Identical to linear regression, but $X^T X \to F^T F$

$$E_{train}(t) = \frac{1}{2\pi i} \int_C \frac{((z-\gamma)(1-2\eta z)^t + \gamma)^2}{z^2} G(z-\gamma) dz$$

$$G(z) = -\, tr[(F^T F - zI)^{-1}] = \int \frac{1}{z-\lambda} \rho_{F^T F}(\lambda) d\lambda$$

CAUCHY TRANSFORM OF $F = f(WX)$

1. Naive option: method of moments

$$G(z) = tr[(zI - F^T F)^{-1}] = \frac{1}{n_1} \sum_k \frac{1}{z^{k+1}} \mathbb{E}\, tr[(F^T F)^k]$$

CAUCHY TRANSFORM OF $F = f(WX)$

1. Naive option: method of moments

$$G(z) = tr[(zI - F^T F)^{-1}] = \frac{1}{n_1} \sum_k \frac{1}{z^{k+1}} \mathbb{E}\, tr[(F^T F)^k]$$

$$\mathbb{E}\, \frac{1}{n_1} tr[(F^T F)^k] = \frac{1}{n_1} \frac{1}{m^k}\, \mathbb{E} \Bigg[\sum_{\substack{i_1, \ldots, i_k \in [n_1] \\ \mu_1, \ldots, \mu_k \in [m]}} F_{i_1 \mu_1} F_{i_2 \mu_1} F_{i_2 \mu_2} F_{i_3 \mu_2} \cdots F_{i_k \mu_k} F_{i_1 \mu_k} \Bigg]$$

Can be evaluated to leading order

CAUCHY TRANSFORM OF $F = f(WX)$

2. Better option: "strong universality" + free probability

    i) "Strong universality" – can replace $F = f(WX)$ by another matrix that has the same second moments

CAUCHY TRANSFORM OF $F = f(WX)$

2. Better option: "strong universality" + free probability

    i) "Strong universality" – can replace $F = f(WX)$ by another matrix that has the same second moments

$$F \simeq F^{lin} \equiv \sqrt{\zeta} WX + \sqrt{\eta - \zeta}\, A$$

$$\eta = \int dz\, \frac{e^{-z^2/2}}{\sqrt{2\pi}} f(\sigma_w \sigma_x z)^2 \qquad \zeta = \left[\sigma_w \sigma_x \int dz\, \frac{e^{-z^2/2}}{\sqrt{2\pi}} f'(\sigma_w \sigma_x z) \right]^2 \qquad A_{ij} \sim \mathcal{N}(0,1)$$

CAUCHY TRANSFORM OF $F = f(WX)$

2. Better option: “strong universality” + free probability

    ii) Free probability – algebraic formalism that allows adding and multiplying “freely independent” noncommutative random variables

If A, W, X are free then the Cauchy transform of F can be obtained from the Cauchy transforms of A, W, X.

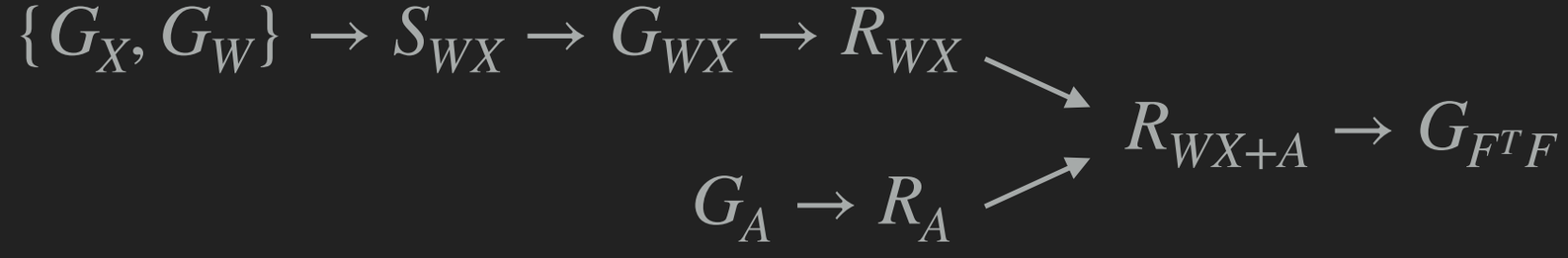
$$F \simeq F^{lin} \equiv \sqrt{\zeta} WX + \sqrt{\eta - \zeta}\, A$$

KERNEL RIDGE REGRESSION

The linearized feature matrix $F^{lin} \equiv \sqrt{\zeta} WX + \sqrt{\eta - \zeta}\, A$ consists of freely independent matrices W, X, A. Can therefore compute Cauchy transform G using R-transform and S-transform.

KERNEL RIDGE REGRESSION

The linearized feature matrix $F^{lin} \equiv \sqrt{\zeta} WX + \sqrt{\eta - \zeta}\, A$ consists of freely independent matrices W, X, A. Can therefore compute Cauchy transform G using R-transform and S-transform.



KERNEL RIDGE REGRESSION

The linearized feature matrix $F^{lin} \equiv \sqrt{\zeta}WX + \sqrt{\eta - \zeta}\,A$ consists of freely independent matrices W, X, A. Can therefore compute Cauchy transform G using R-transform and S-transform.

$$\{G_X, G_W\} \to S_{WX} \to G_{WX} \to R_{WX} \searrow$$

$$G_A \to R_A \nearrow \quad R_{WX+A} \to G_{F^TF}$$

$$E_{train}(t) = \frac{1}{2\pi i}\int_C \frac{((z-\gamma)(1-2\eta z)^t + \gamma)^2}{z^2} G_{F^TF}(z-\gamma)dz$$

KERNEL RIDGE REGRESSION

The linearized feature matrix $F^{lin} \equiv \sqrt{\zeta}WX + \sqrt{\eta - \zeta}\,A$ consists of freely independent matrices W, X, A. Can therefore compute Cauchy transform G using R-transform and S-transform.

$$\{G_X, G_W\} \rightarrow S_{WX} \rightarrow G_{WX} \rightarrow R_{WX} \searrow$$

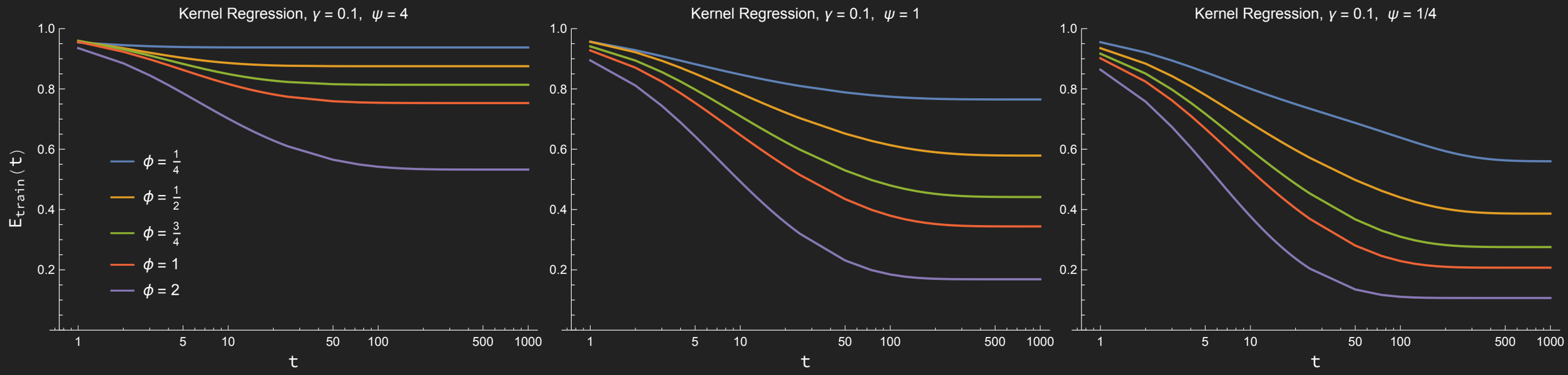
$$G_A \rightarrow R_A \nearrow \quad R_{WX+A} \rightarrow G_{F^T F}$$

Does not require Gaussian X!
Only need spectrum of X.

$$E_{train}(t) = \frac{1}{2\pi i}\int_C \frac{((z-\gamma)(1-2\eta z)^t + \gamma)^2}{z^2} G_{F^T F}(z-\gamma)dz$$

KERNEL RIDGE REGRESSION

$$E_{train}(t) = \frac{1}{2\pi i}\int_C \frac{((z-\gamma)(1-2\eta z)^t+\gamma)^2}{z^2} G_{F^T F}(z-\gamma)dz$$



$$\phi = \frac{n_0}{m} \qquad \psi = \frac{n_0}{n_1}$$

OUTLINE

1. Motivation / Introduction
2. Case Study: Linear Regression
3. Linearization pt 1: High-Dimensional Kernels
4. Linearization pt 2: The Linear Pencil
5. Linearization pt 3: Neural Tangent Kernel

GENERALIZATION ERROR

To discuss generalization, need a non-trivial model for the joint (X, Y) distribution.

For concreteness, consider the student-teacher setup, where $Y = V_2 g(V_1 X)$ for fixed, random weights.

GENERALIZATION ERROR

To discuss generalization, need a non-trivial model for the joint (X, Y) distribution.

For concreteness, consider the student-teacher setup, where $Y = V_2 g(V_1 X)$ for fixed, random weights.

As we saw for F, in high dimensions Y can also be replaced with a linearized version having the correct second moments,

$$Y \simeq Y^{lin} \equiv \sqrt{\zeta_g} V_2 V_1 X + \sqrt{\eta_g - \zeta_g}\, V_2 B \qquad B_{ij} \sim \mathcal{N}(0,1)$$

$$\eta_g = \int dz\, \frac{e^{-z^2/2}}{\sqrt{2\pi}} g(\sigma_w \sigma_x z)^2 \qquad \zeta_g = \left[\sigma_w \sigma_x \int dz\, \frac{e^{-z^2/2}}{\sqrt{2\pi}} g'(\sigma_w \sigma_x z) \right]^2$$

STUDENT-TEACHER KERNEL REGRESSION IN HIGH DIMENSIONS

$$L = \|WF - Y\|_F^2 + \gamma \|W\|_F^2 , \qquad Y = V_2 g(V_1 X) \qquad F = f(W_1 X)$$

$$W^* = YQF^T , \qquad Q = (F^T F + \gamma I)^{-1}$$

STUDENT-TEACHER KERNEL REGRESSION IN HIGH DIMENSIONS

$$L = \|WF - Y\|_F^2 + \gamma\|W\|_F^2, \qquad Y = V_2 g(V_1 X) \qquad F = f(W_1 X)$$

$$W^* = YQF^T, \qquad Q = (F^T F + \gamma I)^{-1}$$

Consider an unseen test point $\tilde{x}$, with random features $\tilde{f} = f(W_1\tilde{x})$ and targets $\tilde{y} = V_2 g(V_1\tilde{x})$.

$$E_{test} = \mathbb{E}_{\tilde{x}}\|W^*\tilde{f} - \tilde{y}\|_F^2 = \mathbb{E}_{\tilde{x}}\, tr[(YQF^T\tilde{f} - \tilde{y})^T(YQF^T\tilde{f} - \tilde{y})]$$

$$= \mathbb{E}_{\tilde{x}}\, tr[\tilde{f}^T FQY^T YQF^T\tilde{f}] - 2\mathbb{E}_{\tilde{x}}\, tr[\tilde{f}^T FQY^T\tilde{y}] + \mathbb{E}_{\tilde{x}}\, tr[\tilde{y}^T\tilde{y}]$$

STUDENT-TEACHER KERNEL REGRESSION IN HIGH DIMENSIONS

$$L = \|WF - Y\|_F^2 + \gamma \|W\|_F^2 , \qquad Y = V_2 g(V_1 X) \qquad F = f(W_1 X)$$

$$W^* = YQF^T , \qquad Q = (F^T F + \gamma I)^{-1}$$

Consider an unseen test point $\tilde{x}$, with random features $\tilde{f} = f(W_1 \tilde{x})$ and targets $\tilde{y} = V_2 g(V_1 \tilde{x})$.

$$E_{test} = \mathbb{E}_{\tilde{x}} \|W^* \tilde{f} - \tilde{y}\|_F^2 = \mathbb{E}_{\tilde{x}}\, tr[(YQF^T \tilde{f} - \tilde{y})^T (YQF^T \tilde{f} - \tilde{y})]$$

$$= \mathbb{E}_{\tilde{x}}\, tr[\tilde{f}^T FQY^T YQF^T \tilde{f}] - 2\mathbb{E}_{\tilde{x}}\, tr[\tilde{f}^T FQY^T \tilde{y}] + \mathbb{E}_{\tilde{x}}\, tr[\tilde{y}^T \tilde{y}]$$

Now, utilize “strong universality" to apply the linearization,

$$Y \to Y^{lin} \equiv \sqrt{\zeta_g} V_2 V_1 X + \sqrt{\eta_g - \zeta_g}\, V_2 B \qquad F \to F^{lin} \equiv \sqrt{\zeta} W_1 X + \sqrt{\eta - \zeta}\, A$$

$$\tilde{y} \to \tilde{y}^{lin} \equiv \sqrt{\zeta_g} V_2 V_1 \tilde{x} + \sqrt{\eta_g - \zeta_g}\, V_2 \tilde{b} \qquad \tilde{f} \to \tilde{f}^{lin} \equiv \sqrt{\zeta} W_1 \tilde{x} + \sqrt{\eta - \zeta}\, \tilde{a}$$

STUDENT-TEACHER KERNEL REGRESSION IN HIGH DIMENSIONS

$$E_{test} = \mathbb{E}_{\tilde{x}}\, tr[\tilde{f}^T F Q Y^T Y Q F^T \tilde{f}] - 2\mathbb{E}_{\tilde{x}}\, tr[\tilde{f}^T F Q Y^T \tilde{y}] + \mathbb{E}_{\tilde{x}}\, tr[\tilde{y}^T \tilde{y}]$$

After applying the linearization,

$$Y \to Y^{lin} \equiv \sqrt{\zeta_g} V_2 V_1 X + \sqrt{\eta_g - \zeta_g}\, V_2 B \qquad F \to F^{lin} \equiv \sqrt{\zeta} W_1 X + \sqrt{\eta - \zeta}\, A$$

$$\tilde{y} \to \tilde{y}^{lin} \equiv \sqrt{\zeta_g} V_2 V_1 \tilde{x} + \sqrt{\eta_g - \zeta_g}\, V_2 \tilde{b} \qquad \tilde{f} \to \tilde{f}^{lin} \equiv \sqrt{\zeta} W_1 \tilde{x} + \sqrt{\eta - \zeta}\, \tilde{a}$$

The expectations over $V_1, V_2, B, \tilde{b}, \tilde{a}$ are trivial because

$$Q \to ((F^{lin})^T F^{lin} + \gamma I)^{-1} = \left((\sqrt{\zeta} W_1 X + \sqrt{\eta - \zeta}\, A)^T (\sqrt{\zeta} W_1 X + \sqrt{\eta - \zeta}\, A)\right)^{-1}$$

depends only on W_1, X, A.

STUDENT–TEACHER KERNEL REGRESSION IN HIGH DIMENSIONS

After applying linearization and performing the trivial expectations, the result can be written as

$$E_{test} = \sum_i tr[R_i Q S_i Q] + \sum_i tr[T_i Q]$$

where R_i, S_i, T_i are low-order polynomials in W_1, X, A.

STUDENT–TEACHER KERNEL REGRESSION IN HIGH DIMENSIONS

After applying linearization and performing the trivial expectations, the result can be written as

$$E_{test} = \sum_i tr[R_i Q S_i Q] + \sum_i tr[T_i Q]$$

where R_i, S_i, T_i are low-order polynomials in W_1, X, A.

Q: How to evaluate the trace of a *rational function* of random matrices?

STUDENT–TEACHER KERNEL REGRESSION IN HIGH DIMENSIONS

After applying linearization and performing the trivial expectations, the result can be written as

$$E_{test} = \sum_i tr[R_i Q S_i Q] + \sum_i tr[T_i Q]$$

where R_i, S_i, T_i are low-order polynomials in W_1, X, A.

Q: How to evaluate the trace of a *rational function* of random matrices?

A: Linearization + operator-valued free probability

RATIONAL FUNCTIONS AS BLOCK MATRIX OPERATIONS

Any rational function of non-commutative variables can be represented in terms of the inverse of a matrix whose entries are linear in the variables.

$$R(x_1, \ldots, x_k) = u^T M^{-1} v\,, \qquad M = M_0 + \sum_i M_i x_i$$

This representation is called the linear pencil.

Constructive proof by induction: manifestly true for $k = 1$, and higher k follow if the representation is closed under addition, multiplication, and inversion. These follow from Schur complement formula.

EXAMPLE OF LINEAR PENCIL

Consider the resolvent as a function in W, X, A,

$$Q = ((F^{lin})^T F^{lin} - zI)^{-1} = ((WX + A)^T (WX + A) - zI)^{-1}$$

$$= u^T M^{-1} v = (I \quad 0 \quad 0 \quad 0) \begin{pmatrix} -zI & A^T & X^T & 0 \\ -A & I & 0 & -W \\ 0 & -W^T & I & 0 \\ -X & 0 & 0 & I \end{pmatrix}^{-1} \begin{pmatrix} I \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

EXAMPLE OF LINEAR PENCIL

Consider the resolvent as a function in W, X, A,

$$Q = ((F^{lin})^T F^{lin} - zI)^{-1} = ((WX + A)^T (WX + A) - zI)^{-1}$$

$$= u^T M^{-1} v = (I \quad 0 \quad 0 \quad 0) \begin{pmatrix} -zI & A^T & X^T & 0 \\ -A & I & 0 & -W \\ 0 & -W^T & I & 0 \\ -X & 0 & 0 & I \end{pmatrix}^{-1} \begin{pmatrix} I \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

M is linear in the W, X, A:

$$M = \begin{pmatrix} -zI & 0 & 0 & 0 \\ 0 & I & 0 & 0 \\ 0 & 0 & I & 0 \\ 0 & 0 & 0 & I \end{pmatrix} + \begin{pmatrix} 0 & 0 & X^T & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -X & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -W \\ 0 & -W^T & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & A^T & 0 & 0 \\ -A & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

but the additive terms are not free, owing to the block structure.

EXAMPLE OF LINEAR PENCIL

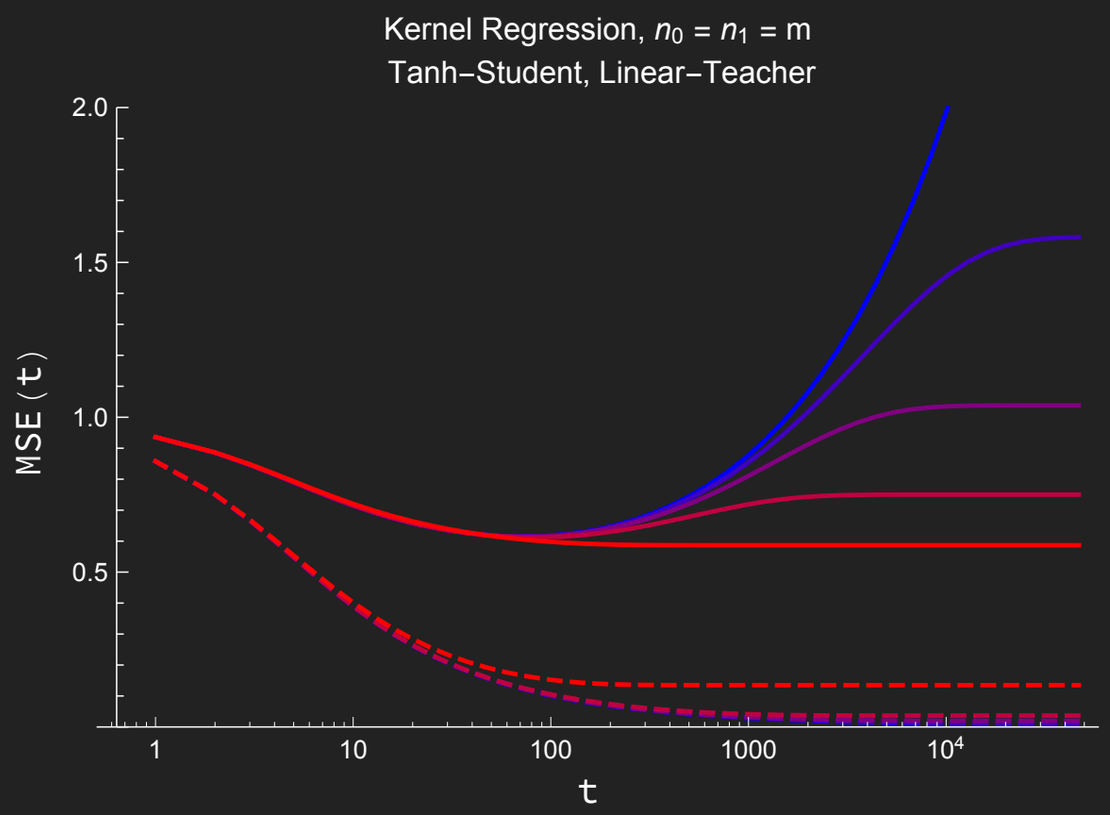
However, we can view M as a linear function of the W, X, A with coefficients in $M_4(\mathbb{C})$

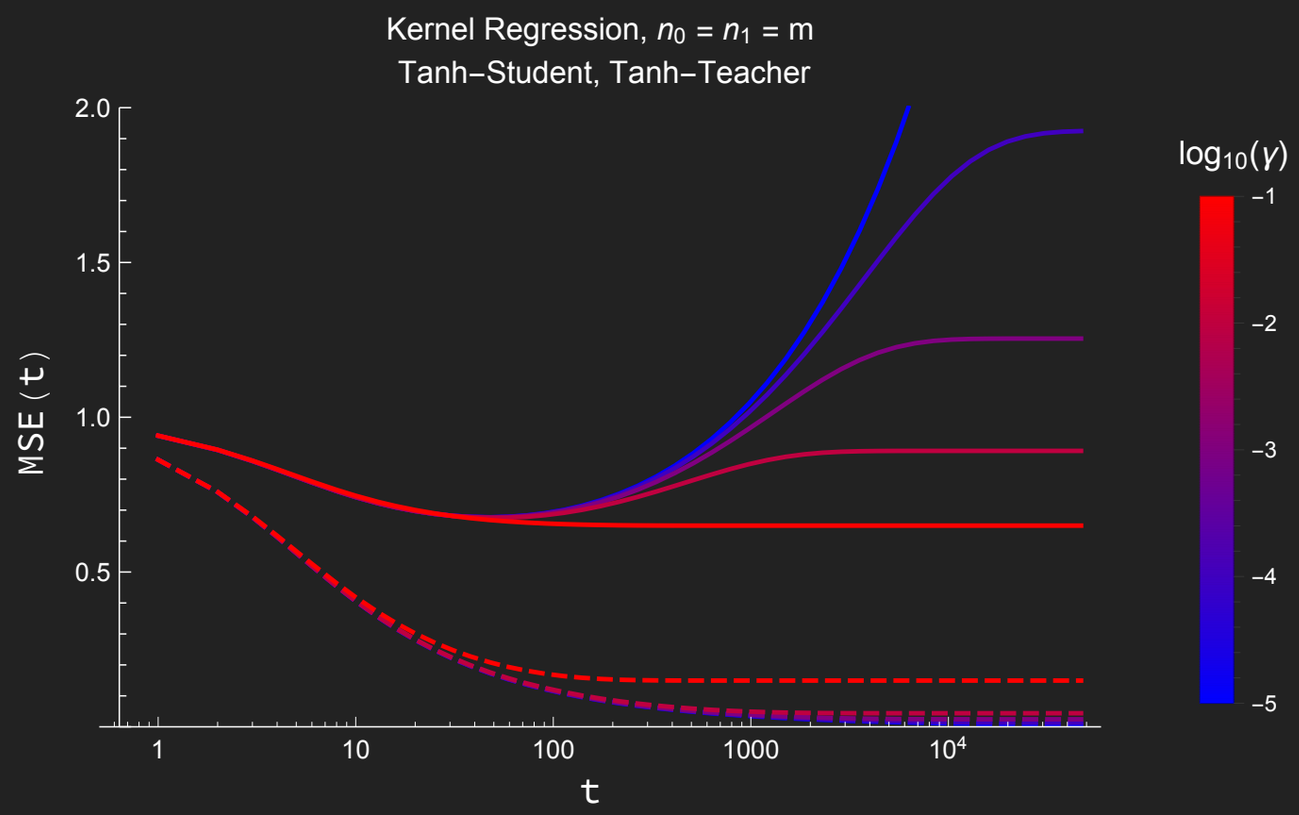
$$M = \begin{pmatrix} -z & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \otimes I + \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{pmatrix} \otimes X + \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \otimes X^T$$

$$+ \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \otimes W + \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \otimes W^T + \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \otimes A + \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} \otimes A^T$$

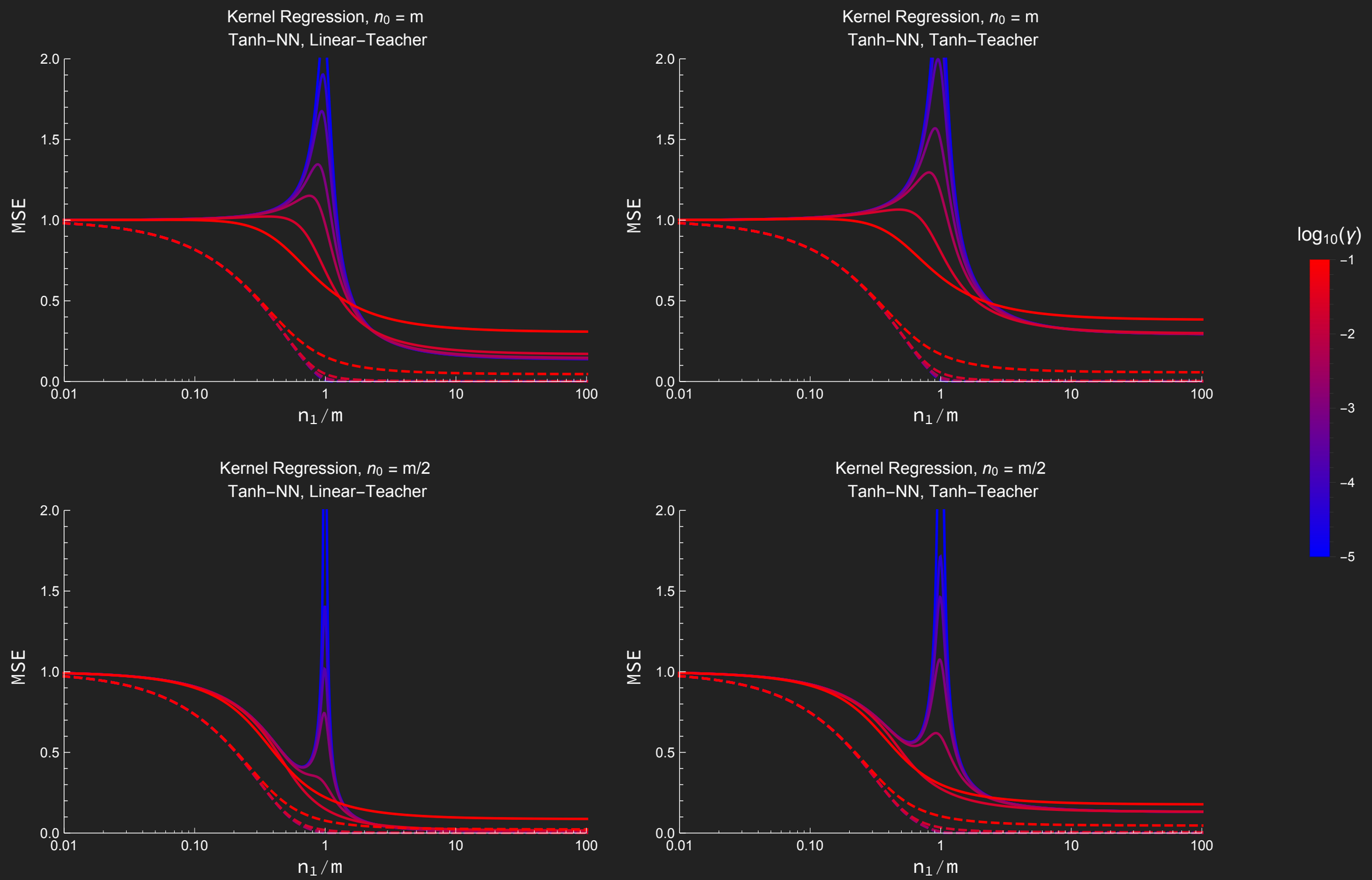
and then freeness can be salvaged, but one must account for the non-commutativity of the coefficients in $M_4(\mathbb{C})$

GENERALIZATION ERROR





GENERALIZATION ERROR



OUTLINE

1. Motivation / Introduction
2. Case Study: Linear Regression
3. Linearization pt 1: High-Dimensional Kernels
4. Linearization pt 2: The Linear Pencil
5. Linearization pt 3: Neural Tangent Kernel

FROM KERNELS TO NEURAL NETWORKS AND BACK AGAIN

Now consider a single-layer neural network in which all the parameters are trained, $N(x; \theta = \{W_1, W_2\}) = W_2 f(W_1 x)$

FROM KERNELS TO NEURAL NETWORKS AND BACK AGAIN

Now consider a single-layer neural network in which all the parameters are trained, $N(x; \theta = \{W_1, W_2\}) = W_2 f(W_1 x)$

As the width grows, parameters move less during the course of gradient descent, i.e. $\theta(t) \approx \theta(0)$

FROM KERNELS TO NEURAL NETWORKS AND BACK AGAIN

Now consider a single-layer neural network in which all the parameters are trained, $N(x; \theta = \{W_1, W_2\}) = W_2 f(W_1 x)$

As the width grows, parameters move less during the course of gradient descent, i.e. $\theta(t) \approx \theta(0)$

This motivates a linear approximation

$$N(x; \theta(t)) \approx N(x; \theta(0)) + \frac{\partial N}{\partial \theta}\Big|_{\theta=\theta(0)} (\theta(t) - \theta(0)) + \mathcal{O}(\theta(t) - \theta(0))^2$$

FROM KERNELS TO NEURAL NETWORKS AND BACK AGAIN

Now consider a single-layer neural network in which all the parameters are trained, $N(x; \theta = \{W_1, W_2\}) = W_2 f(W_1 x)$

As the width grows, parameters move less during the course of gradient descent, i.e. $\theta(t) \approx \theta(0)$

This motivates a linear approximation

$$N(x; \theta(t)) \approx N_0 + J_0(\theta(t) - \theta(0))$$

FROM KERNELS TO NEURAL NETWORKS AND BACK AGAIN

Now consider a single-layer neural network in which all the parameters are trained, $N(x; \theta = \{W_1, W_2\}) = W_2 f(W_1 x)$

As the width grows, parameters move less during the course of gradient descent, i.e. $\theta(t) \approx \theta(0)$

This motivates a linear approximation

$$N(x; \theta(t)) \approx N_0 + J_0(\theta(t) - \theta(0))$$

The dynamics are determined by the Neural Tangent Kernel

$$\Theta = J_0^T J_0 = \Theta_1 + \Theta_2 = (F')^T D_{W_2} F' \odot X^T X + F^T F \qquad F' = f'(W_1 X)$$

NEURAL TANGENT KERNEL

The offset N_0 contributes unnecessary variance. Can set $N_0 = 0$ by subtracting two copies of the model with same initialization

$$N^{VR}(x; \{\theta_1, \theta_2\}) = \frac{1}{\sqrt{2}}\big(N(x;\theta_1) - N(x;\theta_2)\big) \qquad \theta_1(0) = \theta_2(0)$$

$$N_0^{VR} = 0, \quad \Theta^{VR} = \Theta$$

NEURAL TANGENT KERNEL: SECOND-LAYER KERNEL

The component of the kernel from the second layer is the same random features kernel studied before, $\Theta_2 = K = F^T F$

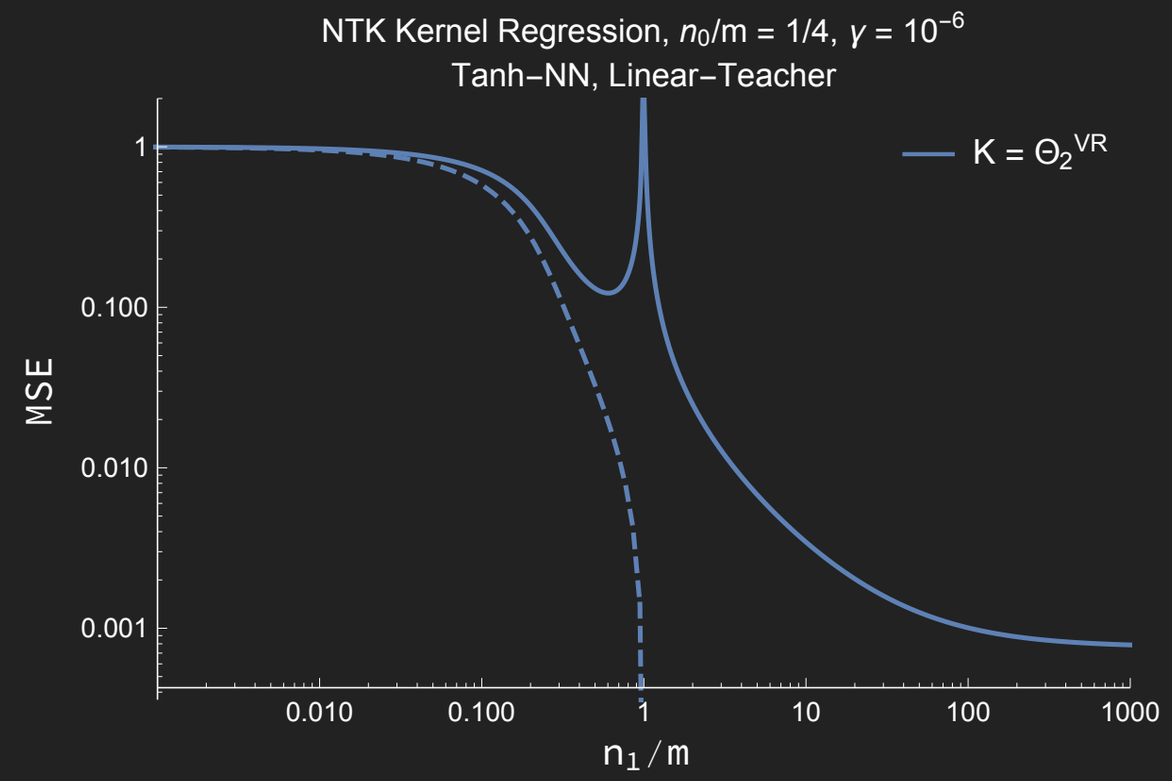
It has non-trivial random matrix behavior in the high-dimensional limit when $n_0 \sim n_1 \sim m$

NEURAL TANGENT KERNEL: FIRST-LAYER KERNEL

The first layer kernel has a Hadamard product structure, $\Theta_1 = (F')^T D_{W_2} F' \odot X^T X$. It has two non-trivial scaling regimes:

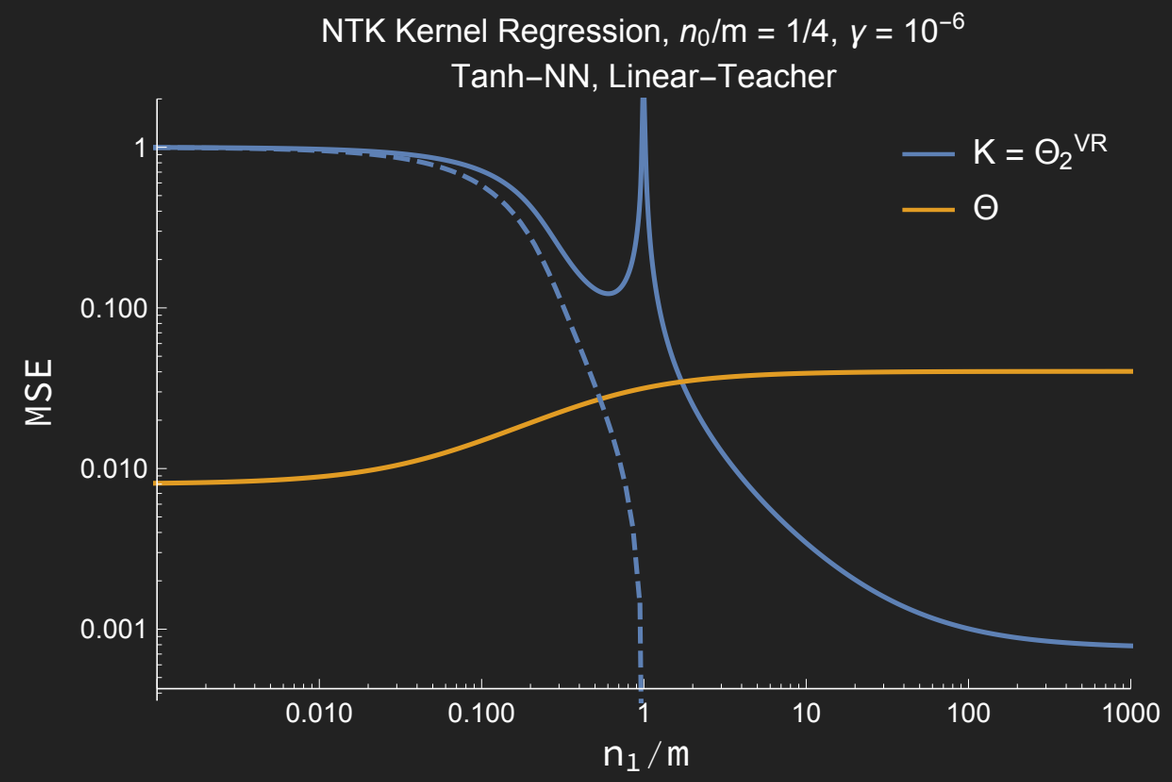
1. Linearly overparameterized ($n_0 n_1 \sim m$)
   - Fluctations of $(F')^T D_{W_2} F'$ are important
   - n eigenvalues of $\mathcal{O}(n)$ and n^2 of $\mathcal{O}(1)$
2. Quadratically overparameterized ($n_l \sim m$)
   - Only the mean of $(F')^T D_{W_2} F'$ is important
   - $\Theta_1 \simeq \Theta_1^{lin} = c_1 I + c_2 X^T X$

QUADRATIC OVERPARAMETERIZATION



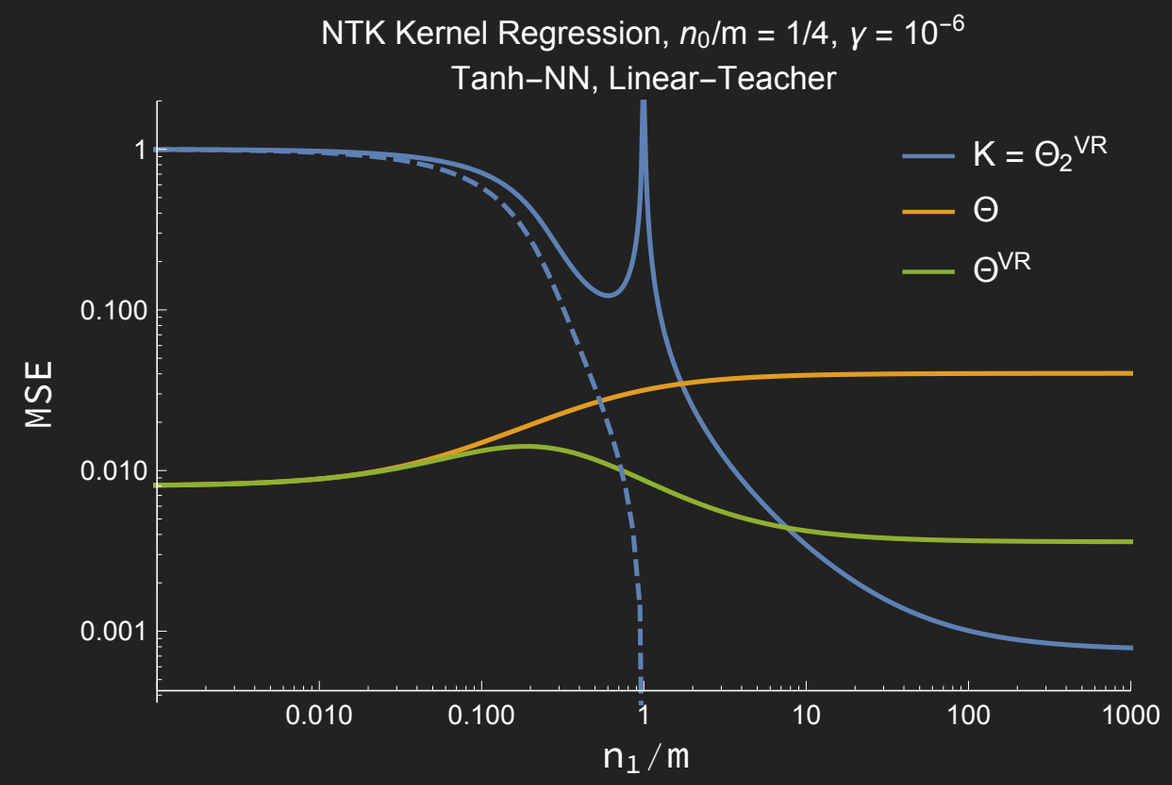
QUADRATIC OVERPARAMETERIZATION

The network can be too overparametrized

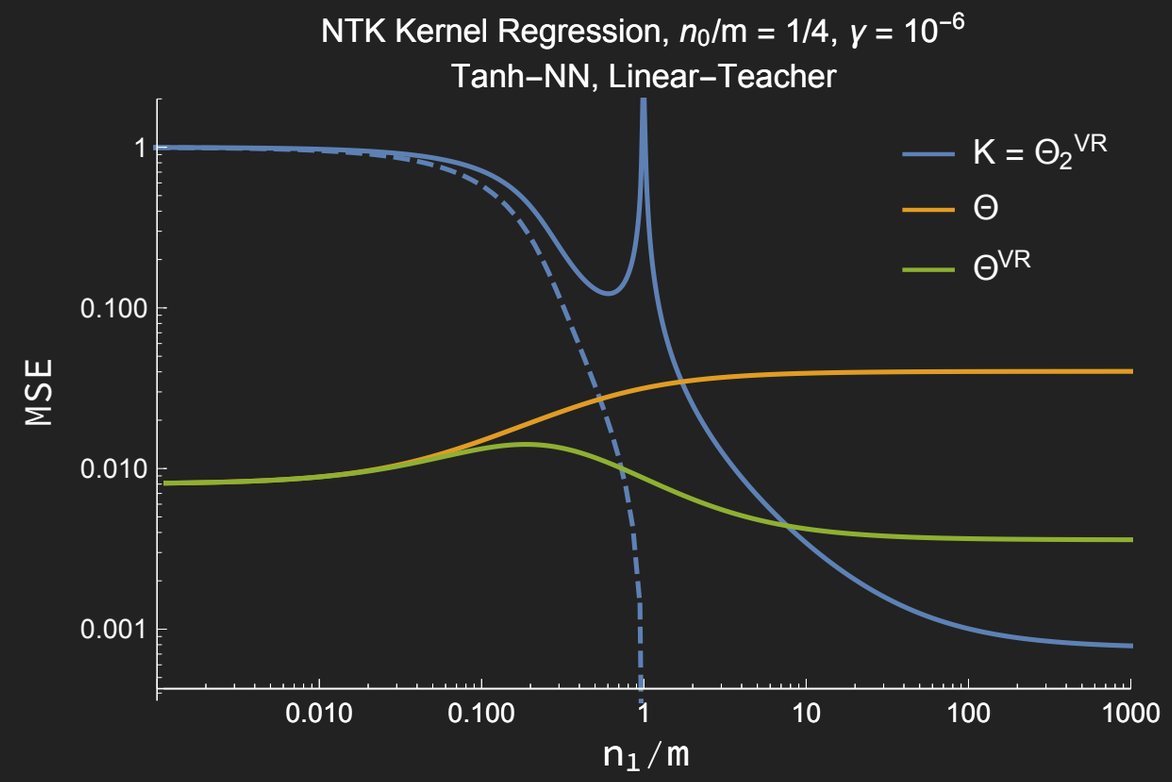


QUADRATIC OVERPARAMETERIZATION

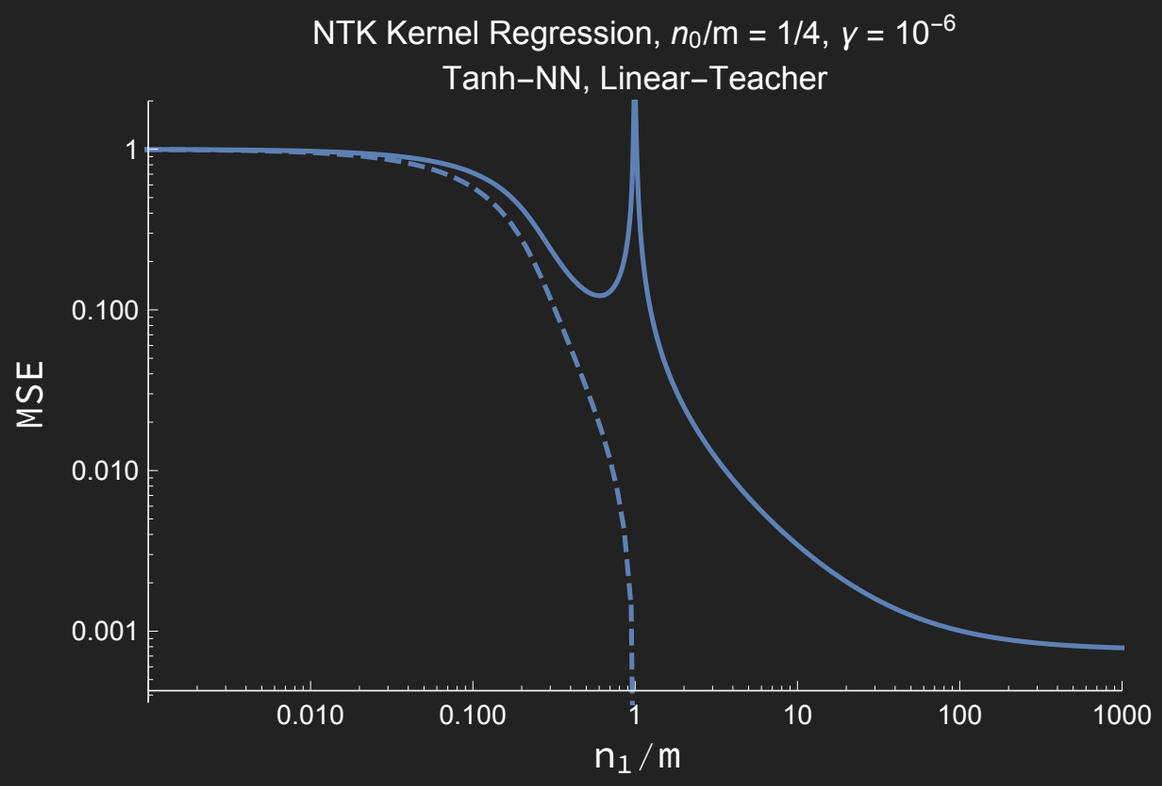
Reducing the variance helps, but a peak emerges

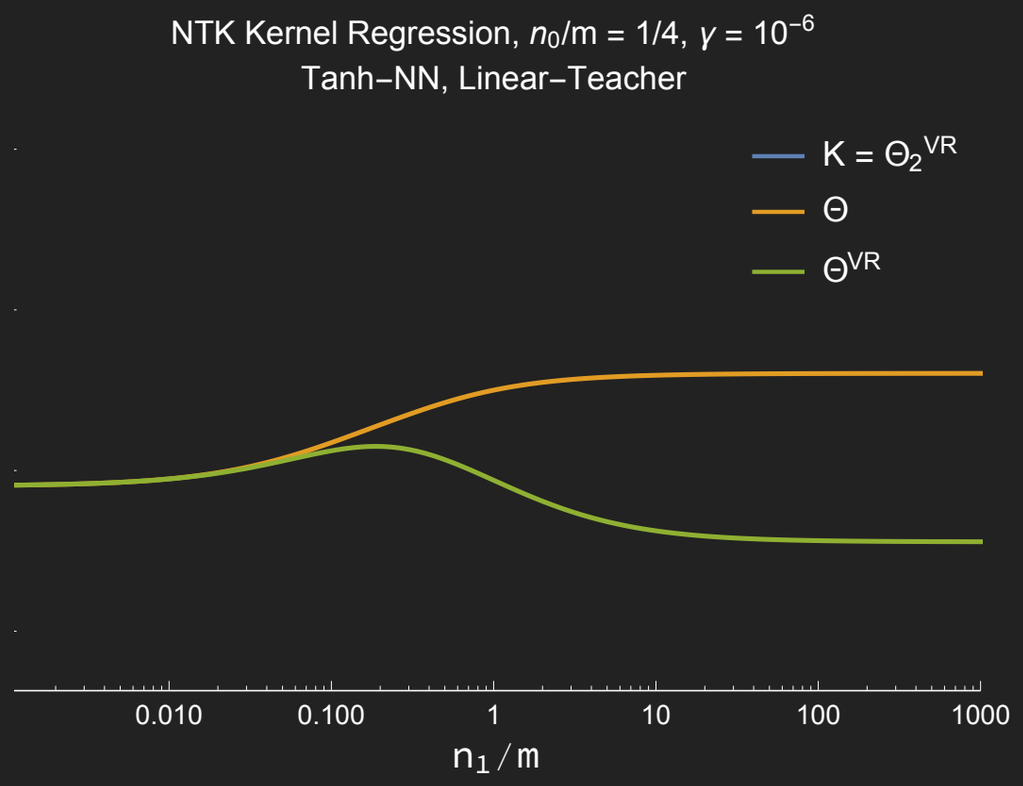


TWO OVERPARAMETERIZATION SCALES



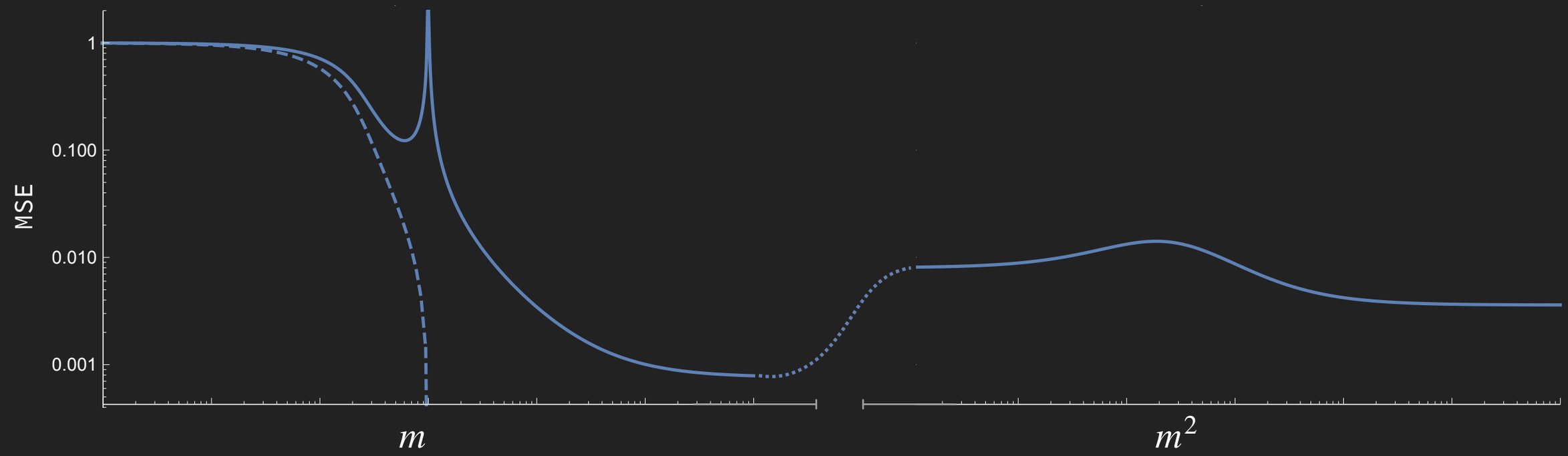
TWO OVERPARAMETERIZATION SCALES



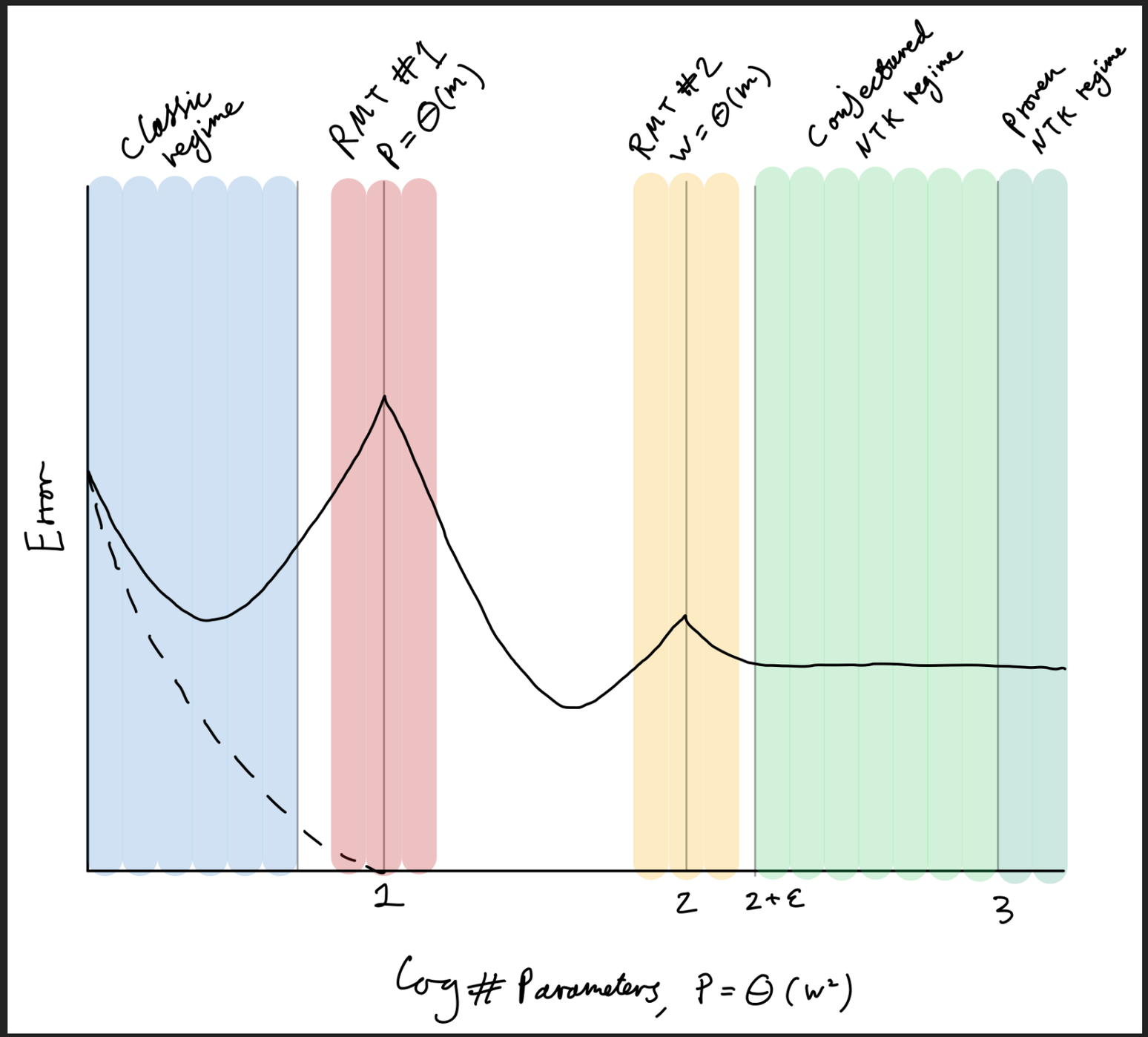


TWO OVERPARAMETERIZATION SCALES

NTK Kernel Regression, $n_0/m = 1/4$, $\gamma = 10^{-6}$
Tanh–NN, Linear–Teacher



TRIPLE DESCENT?



EXTRA SLIDES

CUMULANTS AND CLASSICAL INDEPENDENCE

The cumulant generating function K generates connected correlation functions via the relation

$$K(t_1, \ldots, t_n) = \log \mathbb{E}\, e^{\sum_{i=1}^{n} t_i X_i}$$

The cumulants κ are defined by the moments via a sum over partitions π:

$$\mathbb{E}[X_1 \cdots X_n] = \sum_{\pi} \kappa_\pi[X_1, \ldots, X_n] \qquad \kappa_\pi[X_1, \ldots, X_n] = \prod_{B \in \pi} \kappa[X_i : i \in B]$$

CUMULANTS AND CLASSICAL INDEPENDENCE

The cumulant generating function K generates connected correlation functions via the relation

$$K(t_1, \ldots, t_n) = \log \mathbb{E}\, e^{\sum_{i=1}^{n} t_i X_i}$$

The cumulants κ are defined by the moments via a sum over partitions π:

$$\mathbb{E}[X_1 \cdots X_n] = \sum_{\pi} \kappa_\pi[X_1, \ldots, X_n] \qquad \kappa_\pi[X_1, \ldots, X_n] = \prod_{B \in \pi} \kappa[X_i : i \in B]$$

For example,

n=1: $\mathbb{E}[X_1] = \kappa[X_1]$

n=2: $\mathbb{E}[X_1 X_2] = \kappa[X_1 X_2] + \kappa[X_1]\kappa[X_2]$

CUMULANTS AND CLASSICAL INDEPENDENCE

n=3: $$\mathbb{E}[X_1X_2X_3] = \kappa[X_1X_2X_3] + \kappa[X_1X_2]\kappa[X_3] + \kappa[X_1X_3]\kappa[X_2] + \kappa[X_2X_3]\kappa[X_1] + \kappa[X_1]\kappa[X_2]\kappa[X_3]$$

n=4: $$\begin{aligned}\mathbb{E}[X_1X_2X_3X_4] = {} & \kappa[X_1X_2X_3X_4] + \kappa[X_1X_2X_3]\kappa[X_4] + \kappa[X_1X_2X_4]\kappa[X_3] \\ & + \kappa[X_1X_3X_4]\kappa[X_2] + \kappa[X_2X_3X_4]\kappa[X_1] + \kappa[X_1X_2]\kappa[X_3X_4] \\ & + \kappa[X_1X_3]\kappa[X_2X_4] + \kappa[X_1X_4]\kappa[X_2X_3] + \kappa[X_3X_4]\kappa[X_1]\kappa[X_2] \\ & + \kappa[X_2X_4]\kappa[X_1]\kappa[X_3] + \kappa[X_2X_3]\kappa[X_1]\kappa[X_4] + \kappa[X_1X_4]\kappa[X_2]\kappa[X_3] \\ & + \kappa[X_1X_3]\kappa[X_2]\kappa[X_4] + \kappa[X_1X_2]\kappa[X_3]\kappa[X_4] + \kappa[X_1]\kappa[X_2]\kappa[X_2]\kappa[X_3]\end{aligned}$$

The mixed cumulants vanish for independent random variables

FREE CUMULANTS AND FREE INDEPENDENCE

Free cumulants: sum over non-crossing partitions $\pi \in NC(n)$:

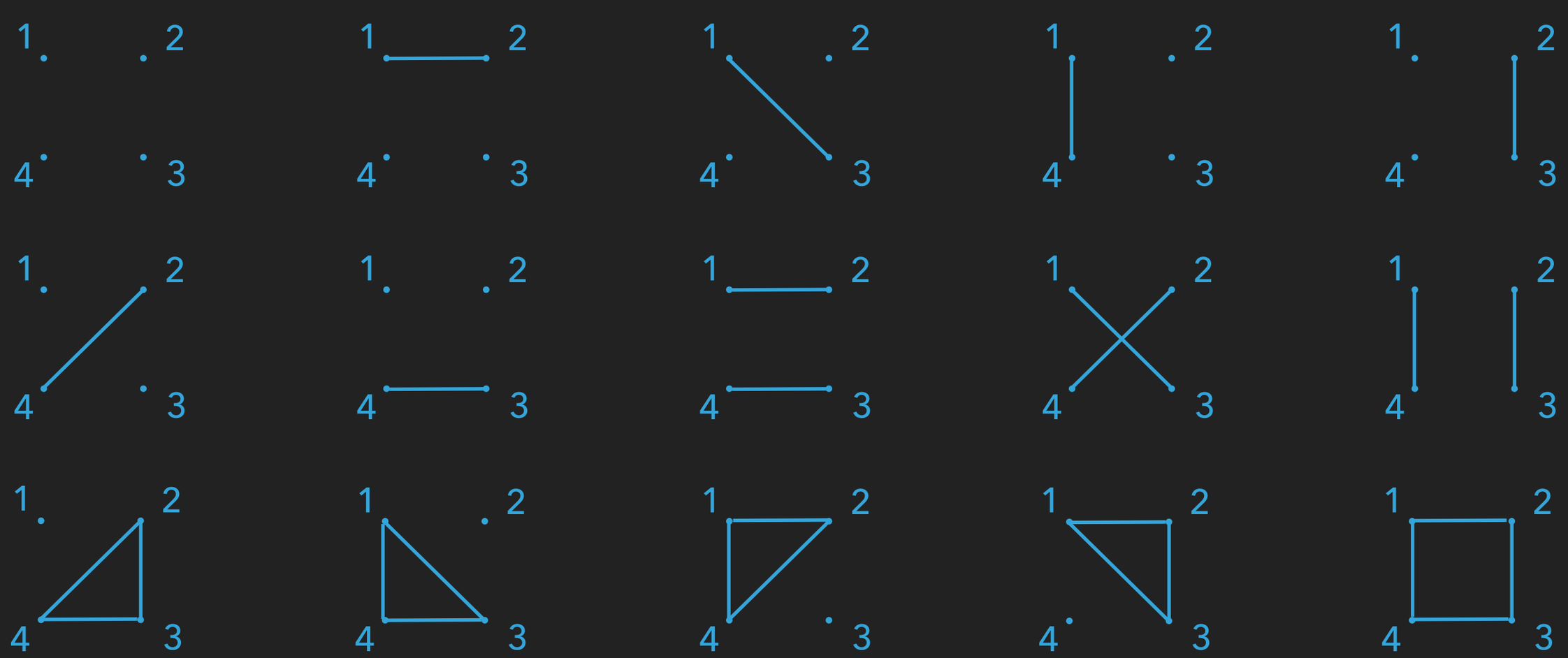
$$\mathbb{E}[X_1 \cdots X_n] = \sum_{\pi \in NC(n)} \kappa_\pi[X_1, \ldots, X_n] \qquad \kappa_\pi[X_1, \ldots, X_n] = \prod_{B \in \pi} \kappa[X_i : i \in B]$$

FREE CUMULANTS AND FREE INDEPENDENCE

Free cumulants: sum over non-crossing partitions $\pi \in NC(n)$:

$$\mathbb{E}[X_1 \cdots X_n] = \sum_{\pi \in NC(n)} \kappa_\pi[X_1, \ldots, X_n] \qquad \kappa_\pi[X_1, \ldots, X_n] = \prod_{B \in \pi} \kappa[X_i : i \in B]$$

For example, at $n = 4$, the partitions are

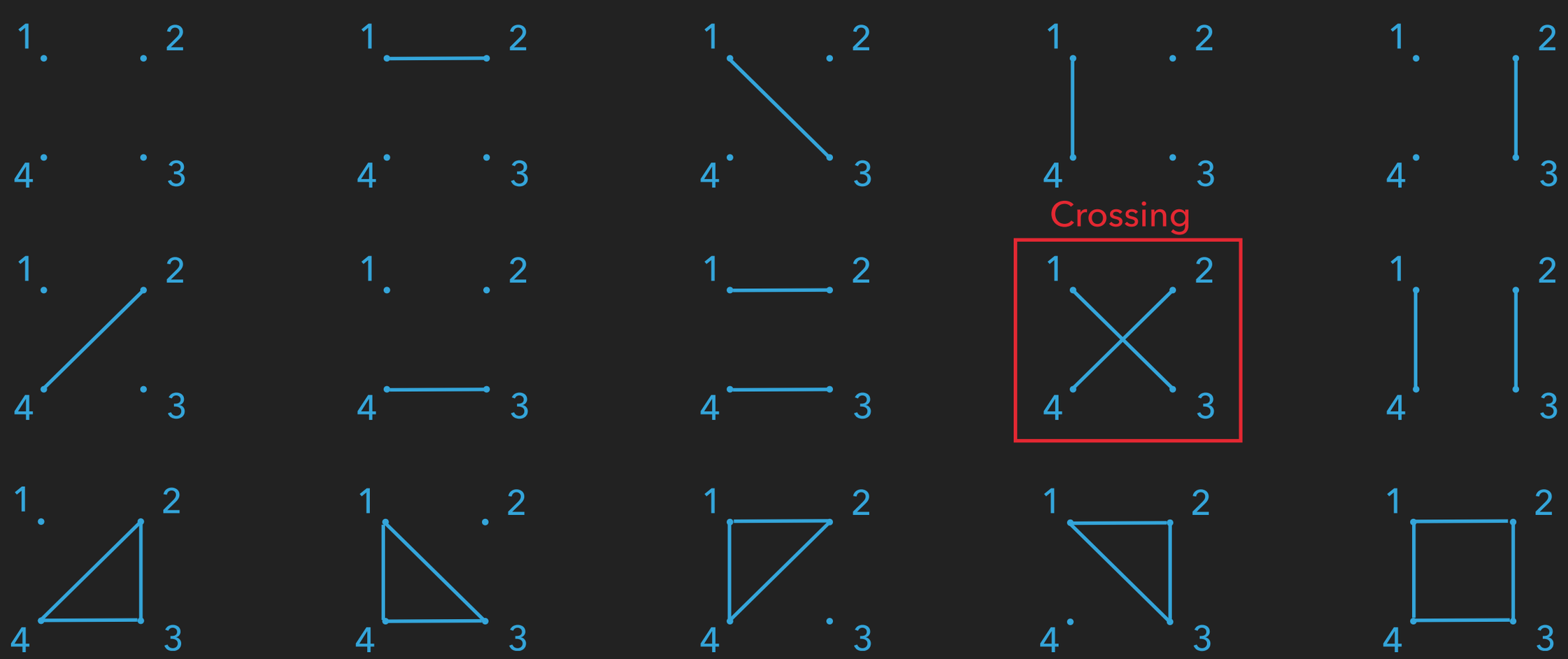


FREE CUMULANTS AND FREE INDEPENDENCE

Free cumulants: sum over non-crossing partitions $\pi \in NC(n)$:

$$\mathbb{E}[X_1 \cdots X_n] = \sum_{\pi \in NC(n)} \kappa_\pi[X_1, \ldots, X_n] \qquad \kappa_\pi[X_1, \ldots, X_n] = \prod_{B \in \pi} \kappa[X_i : i \in B]$$

For example, at $n = 4$, the partitions are

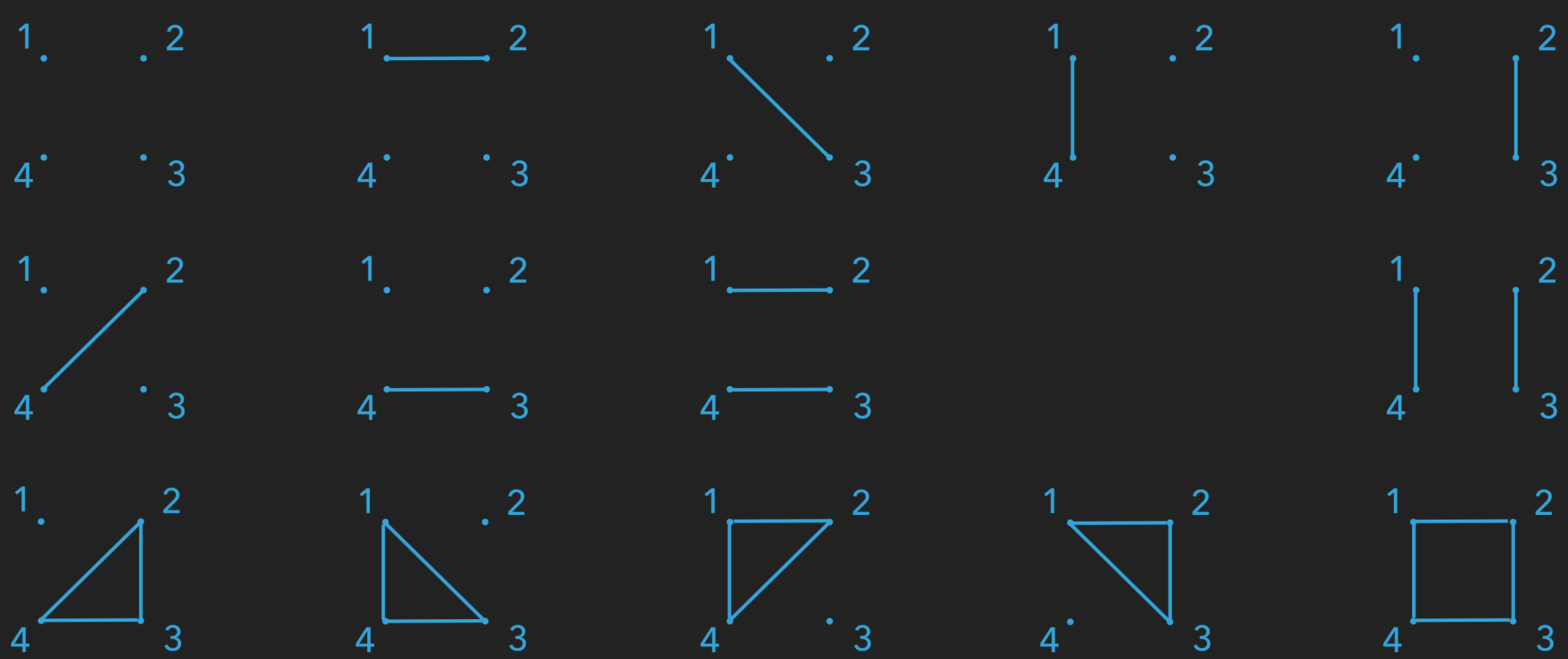


FREE CUMULANTS AND FREE INDEPENDENCE

Free cumulants: sum over non-crossing partitions $\pi \in NC(n)$:

$$\mathbb{E}[X_1 \cdots X_n] = \sum_{\pi \in NC(n)} \kappa_\pi[X_1, \ldots, X_n] \qquad \kappa_\pi[X_1, \ldots, X_n] = \prod_{B \in \pi} \kappa[X_i : i \in B]$$

For example, at $n = 4$, the non-crossing partitions are



FREE CUMULANTS AND FREE INDEPENDENCE

Free cumulants: sum over non-crossing partitions $\pi \in NC(n)$:

$$\mathbb{E}[X_1 \cdots X_n] = \sum_{\pi \in NC(n)} \kappa_\pi[X_1, \ldots, X_n] \qquad \kappa_\pi[X_1, \ldots, X_n] = \prod_{B \in \pi} \kappa[X_i : i \in B]$$

For example, at $n = 4$, the free cumulants obey

n=4: $$\begin{aligned}\mathbb{E}[X_1X_2X_3X_4] = {} & \kappa[X_1X_2X_3X_4] + \kappa[X_1X_2X_3]\kappa[X_4] + \kappa[X_1X_2X_4]\kappa[X_3] \\ & + \kappa[X_1X_3X_4]\kappa[X_2] + \kappa[X_2X_3X_4]\kappa[X_1] + \kappa[X_1X_2]\kappa[X_3X_4] \\ & + \cancel{\kappa[X_1X_3]\kappa[X_2X_4]} + \kappa[X_1X_4]\kappa[X_2X_3] + \kappa[X_3X_4]\kappa[X_1]\kappa[X_2] \\ & + \kappa[X_2X_4]\kappa[X_1]\kappa[X_3] + \kappa[X_2X_3]\kappa[X_1]\kappa[X_4] + \kappa[X_1X_4]\kappa[X_2]\kappa[X_3] \\ & + \kappa[X_1X_3]\kappa[X_2]\kappa[X_4] + \kappa[X_1X_2]\kappa[X_3]\kappa[X_4] + \kappa[X_1]\kappa[X_2]\kappa[X_2]\kappa[X_3]\end{aligned}$$

The mixed free cumulants vanish for freely independent random variables.

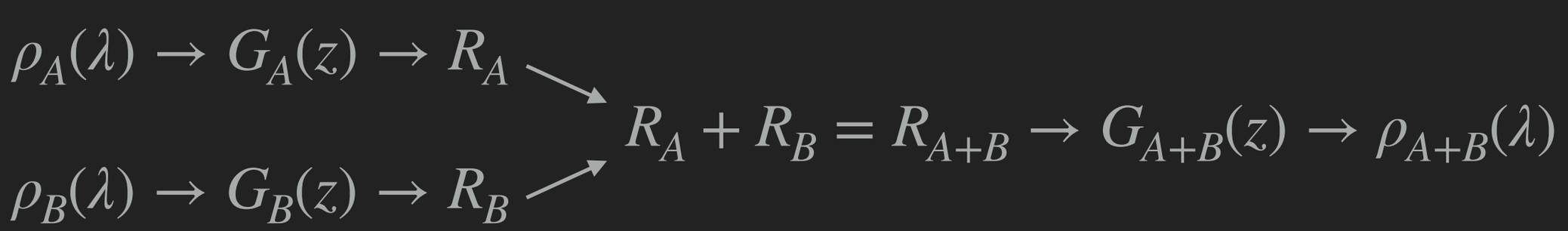
R-TRANSFORM AND S-TRANSFORM

Given free random matrices A and B, can add and multiply using auxiliary objects: the R-transform and the S-transform

R-TRANSFORM AND S-TRANSFORM

Given free random matrices A and B, can add and multiply using auxiliary objects: the R-transform and the S-transform

R-transform: $zG(z) = 1 + R(G(z))G(z)$



R-TRANSFORM AND S-TRANSFORM

Given free random matrices A and B, can add and multiply using auxiliary objects: the R-transform and the S-transform

R-transform: $zG(z) = 1 + R(G(z))G(z)$

$$\left.\begin{array}{l}\rho_A(\lambda) \to G_A(z) \to R_A \\ \rho_B(\lambda) \to G_B(z) \to R_B\end{array}\right\rangle \; R_A + R_B = R_{A+B} \to G_{A+B}(z) \to \rho_{A+B}(\lambda)$$

S-transform: $G(z) = S(zG(z) - 1)(z(G(z) - 1))$

$$\left.\begin{array}{l}\rho_A(\lambda) \to G_A(z) \to S_A \\ \rho_B(\lambda) \to G_B(z) \to S_B\end{array}\right\rangle \; S_A S_B = S_{AB} \to G_{AB}(z) \to \rho_{AB}(\lambda)$$

OPERATOR-VALUED FREE PROBABILITY

Basic idea: perform as much of the calculation as possible in $M_d(\mathbb{C})$ before projecting down to $\mathbb{C}$.

OPERATOR-VALUED FREE PROBABILITY

Basic idea: perform as much of the calculation as possible in $M_d(\mathbb{C})$ before projecting down to $\mathbb{C}$.

The operator-valued Cauchy transform $G : M_d(\mathbb{C})^+ \to M_d(\mathbb{C})^+$

Scalar-valued		Operator-valued
$G(z) = tr[(zI - M)^{-1}]$	$\Rightarrow$	$G(b) = (id \otimes tr)\left[(b \otimes I - M_x \otimes X)^{-1}\right]$

OPERATOR-VALUED FREE PROBABILITY

Basic idea: perform as much of the calculation as possible in $M_d(\mathbb{C})$ before projecting down to $\mathbb{C}$.

The operator-valued Cauchy transform $G : M_d(\mathbb{C})^+ \to M_d(\mathbb{C})^+$

Scalar-valued | Operator-valued

$$G(z) = tr[(zI - M)^{-1}] \quad \Rightarrow \quad G(b) = (id \otimes tr)\left[(b \otimes I - M_x \otimes X)^{-1}\right]$$

Operator-valued R-transform obeys same relation as scalar case:

$$bG(b) = I + R(G(b))G(b)$$

If A and B are free over $M_d(\mathbb{C})$, their operator-valued R-transforms add

OPERATOR-VALUED FREENESS

Operator-valued freeness is analogous to standard freeness, but the cumulants are operator-valued so the ordering matters.

Standard: $\mathbb{E}[X_1X_2X_3] = \kappa[X_1X_2X_3] + \kappa[X_1X_2]\kappa[X_3] + \kappa[X_1X_3]\kappa[X_2]$
$+\kappa[X_2X_3]\kappa[X_1] + \kappa[X_1]\kappa[X_2]\kappa[X_3]$

Operator: $\mathbb{E}[X_1X_2X_3] = \kappa[X_1X_2X_3] + \kappa[X_1X_2\kappa[X_3]] + \kappa[X_1\kappa[X_2]X_3]$
$+\kappa[\kappa[X_1]X_2X_3] + \kappa[X_1]\kappa[X_2]\kappa[X_3]$

OPERATOR-VALUED FREENESS

Operator-valued freeness is analogous to standard freeness, but the cumulants are operator-valued so the ordering matters.

Standard: $\mathbb{E}[X_1X_2X_3] = \kappa[X_1X_2X_3] + \kappa[X_1X_2]\kappa[X_3] + \kappa[X_1X_3]\kappa[X_2]$
$+\kappa[X_2X_3]\kappa[X_1] + \kappa[X_1]\kappa[X_2]\kappa[X_3]$

Operator: $\mathbb{E}[X_1X_2X_3] = \kappa[X_1X_2X_3] + \kappa[X_1X_2\kappa[X_3]] + \kappa[X_1\kappa[X_2]X_3]$
$+\kappa[\kappa[X_1]X_2X_3] + \kappa[X_1]\kappa[X_2]\kappa[X_3]$

A and B are free over $M_d(\mathbb{C})$ if their mixed operator-valued cumulants vanish

OPERATOR-VALUED FREENESS

Operator-valued freeness is analogous to standard freeness, but the cumulants are operator-valued so the ordering matters.

Standard: $\mathbb{E}[X_1X_2X_3] = \kappa[X_1X_2X_3] + \kappa[X_1X_2]\kappa[X_3] + \kappa[X_1X_3]\kappa[X_2]$
$+\kappa[X_2X_3]\kappa[X_1] + \kappa[X_1]\kappa[X_2]\kappa[X_3]$

Operator: $\mathbb{E}[X_1X_2X_3] = \kappa[X_1X_2X_3] + \kappa[X_1X_2\kappa[X_3]] + \kappa[X_1\kappa[X_2]X_3]$
$+\kappa[\kappa[X_1]X_2X_3] + \kappa[X_1]\kappa[X_2]\kappa[X_3]$

A and B are free over $M_d(\mathbb{C})$ if their mixed operator-valued cumulants vanish

$\Rightarrow$ True for the linear pencils needed for the test error

OPERATOR-VALUED FREE PROBABILITY

Basic idea: perform as much of the calculation as possible in $M_d(\mathbb{C})$ before projecting down to $\mathbb{C}$.

OPERATOR-VALUED FREE PROBABILITY

Basic idea: perform as much of the calculation as possible in $M_d(\mathbb{C})$ before projecting down to $\mathbb{C}$.

The operator-valued Cauchy transform $G : M_d(\mathbb{C})^+ \to M_d(\mathbb{C})^+$

Scalar-valued		Operator-valued
$G(z) = tr[(zI - M)^{-1}]$	$\Rightarrow$	$G(b) = (id \otimes tr)\left[(b \otimes I - M_x \otimes X)^{-1}\right]$

OPERATOR-VALUED FREE PROBABILITY

Basic idea: perform as much of the calculation as possible in $M_d(\mathbb{C})$ before projecting down to $\mathbb{C}$.

The operator-valued Cauchy transform $G : M_d(\mathbb{C})^+ \to M_d(\mathbb{C})^+$

Scalar-valued | Operator-valued

$$G(z) = tr[(zI - M)^{-1}] \quad \Rightarrow \quad G(b) = (id \otimes tr)\left[(b \otimes I - M_x \otimes X)^{-1}\right]$$

Operator-valued R-transform obeys same relation as scalar case:

$$bG(b) = I + R(G(b))G(b)$$

If A and B are free over $M_d(\mathbb{C})$, their operator-valued R-transforms add

OPERATOR-VALUED FREENESS

Operator-valued freeness is analogous to standard freeness, but the cumulants are operator-valued so the ordering matters.

Standard: $\mathbb{E}[X_1X_2X_3] = \kappa[X_1X_2X_3] + \kappa[X_1X_2]\kappa[X_3] + \kappa[X_1X_3]\kappa[X_2]$
$+\kappa[X_2X_3]\kappa[X_1] + \kappa[X_1]\kappa[X_2]\kappa[X_3]$

Operator: $\mathbb{E}[X_1X_2X_3] = \kappa[X_1X_2X_3] + \kappa[X_1X_2\kappa[X_3]] + \kappa[X_1\kappa[X_2]X_3]$
$+\kappa[\kappa[X_1]X_2X_3] + \kappa[X_1]\kappa[X_2]\kappa[X_3]$

OPERATOR-VALUED FREENESS

Operator-valued freeness is analogous to standard freeness, but the cumulants are operator-valued so the ordering matters.

Standard: $\mathbb{E}[X_1X_2X_3] = \kappa[X_1X_2X_3] + \kappa[X_1X_2]\kappa[X_3] + \kappa[X_1X_3]\kappa[X_2]$
$+\kappa[X_2X_3]\kappa[X_1] + \kappa[X_1]\kappa[X_2]\kappa[X_3]$

Operator: $\mathbb{E}[X_1X_2X_3] = \kappa[X_1X_2X_3] + \kappa[X_1X_2\kappa[X_3]] + \kappa[X_1\kappa[X_2]X_3]$
$+\kappa[\kappa[X_1]X_2X_3] + \kappa[X_1]\kappa[X_2]\kappa[X_3]$

A and B are free over $M_d(\mathbb{C})$ if their mixed operator-valued cumulants vanish

OPERATOR-VALUED FREENESS

Operator-valued freeness is analogous to standard freeness, but the cumulants are operator-valued so the ordering matters.

Standard: $\mathbb{E}[X_1X_2X_3] = \kappa[X_1X_2X_3] + \kappa[X_1X_2]\kappa[X_3] + \kappa[X_1X_3]\kappa[X_2]$
$+\kappa[X_2X_3]\kappa[X_1] + \kappa[X_1]\kappa[X_2]\kappa[X_3]$

Operator: $\mathbb{E}[X_1X_2X_3] = \kappa[X_1X_2X_3] + \kappa[X_1X_2\kappa[X_3]] + \kappa[X_1\kappa[X_2]X_3]$
$+\kappa[\kappa[X_1]X_2X_3] + \kappa[X_1]\kappa[X_2]\kappa[X_3]$

A and B are free over $M_d(\mathbb{C})$ if their mixed operator-valued cumulants vanish

$\Rightarrow$ True for the linear pencils needed for the test error