Outline

1. Practice
2. Theory
3. More practice
4. Theorems
   - Introduction
   - Sample covariance matrix for large dimensional data: from concentration to RMT
   - A random matrix perspective of the “curse of dimensionality”
   - Kernel spectral clustering for large dimensional data
   - A random matrix approach to large neural networks and random features
Introduction
Sample covariance matrix for large dimensional data: from concentration to RMT
A random matrix perspective of the “curse of dimensionality”
Kernel spectral clustering for large dimensional data
A random matrix approach to large neural networks and random features
Lots of DNNs analyzed: Look at nearly every publicly-available SOTA model in CV and NLP

- Don’t evaluate your method on one/two/three NNs, evaluate it on:
  - dozens (2017)
  - hundreds (2019)
  - thousands (2021)

- Don’t use bad/toy models, use SOTA models.
  - If you do, don’t be surprised if low-quality/toy models are different than high-quality/SOTA models.

- Don’t train models, instead validate pre-trained models.
  - Validating models is harder than training models.
Results: LeNet5 (an old/small NN example)

Figure: Full and zoomed-in ESD for LeNet5, Layer FC1.

Older and/or smaller and/or less well-trained models look like bulk+spike.
Results: AlexNet (a typical modern/large DNN example)

Figure: Zoomed-in ESD for Layer FC1 and FC3 of AlexNet.

Newer SOTA models have heavy-tail structure in their weight matrix correlations (i.e., not elements but eigenvalues).
Random Matrix Theory 101: Wigner and Tracy-Widom

- Wigner: *global bulk statistics* approach universal semi-circular form
- Tracy-Widom: *local edge statistics* fluctuate in universal way

Problems with Wigner and Tracy-Widom:
- Weight matrices usually not square
- Typically do only a single training run
Random Matrix Theory 102': Marchenko-Pastur

(c) Vary aspect ratios  (d) Vary variance parameters

Figure: Marchenko-Pastur (MP) distributions.

Important points:

- **Global bulk stats**: The overall shape is deterministic, fixed by $Q$ and $\sigma$.

- **Local edge stats**: The edge $\lambda^+$ is very crisp, i.e.,
  \[
  \Delta \lambda_M = |\lambda_{\text{max}} - \lambda^+| \sim O(M^{-2/3}),
  \]
  plus Tracy-Widom fluctuations.

We use both global bulk statistics as well as local edge statistics in our theory.
Random Matrix Theory 103: Heavy-tailed RMT

Go beyond the (relatively easy) Gaussian Universality class:

- *model* strongly-correlated systems ("signal") with heavy-tailed random matrices.

<table>
<thead>
<tr>
<th>Generative Model w/ elements from Universality class</th>
<th>Finite-N Global shape $\rho_N(\lambda)$</th>
<th>Limiting Global shape $\rho(\lambda), \ N \to \infty$</th>
<th>Bulk edge Local stats $\lambda \approx \lambda^+$</th>
<th>(far) Tail Local stats $\lambda \approx \lambda_{\max}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic MP</td>
<td>Gaussian</td>
<td>MP distribution</td>
<td>MP</td>
<td>No tail.</td>
</tr>
<tr>
<td>Spiked-Covariance</td>
<td>Gaussian, + low-rank perturbations</td>
<td>MP + Gaussian spikes</td>
<td>MP</td>
<td>TW</td>
</tr>
<tr>
<td>Heavy tail, $4 &lt; \mu$</td>
<td>(Weakly) Heavy-Tailed</td>
<td>MP + PL tail</td>
<td>MP</td>
<td>Heavy-Tailed*</td>
</tr>
<tr>
<td>Heavy tail, $2 &lt; \mu &lt; 4$</td>
<td>(Moderately) Heavy-Tailed (or &quot;fat tailed&quot;)</td>
<td>PL** $\sim \lambda^{-(a\mu+b)}$</td>
<td>PL $\sim \lambda^{-(\frac{1}{2}\mu+1)}$</td>
<td>No edge. Frechet</td>
</tr>
<tr>
<td>Heavy tail, $0 &lt; \mu &lt; 2$</td>
<td>(Very) Heavy-Tailed</td>
<td>PL$^{**}$ $\sim \lambda^{-(\frac{1}{2}\mu+1)}$</td>
<td>PL $\sim \lambda^{-(\frac{1}{2}\mu+1)}$</td>
<td>No edge. Frechet</td>
</tr>
</tbody>
</table>

Basic MP theory, and the spiked and Heavy-Tailed extensions we use, including known, empirically-observed, and conjectured relations between them. Boxes marked "*" are best described as following “TW with large finite size corrections” that are likely Heavy-Tailed, leading to bulk edge statistics and far tail statistics that are indistinguishable. Boxes marked "**" are phenomenological fits, describing large ($2 < \mu < 4$) or small ($0 < \mu < 2$) finite-size corrections on $N \to \infty$ behavior.
RMT-based 5+1 Phases of Training (in pictures)

(a) **Random-like.**

(b) **Bleeding-out.**

(c) **Bulk+Spikes.**

(d) **Bulk-decay.**

(e) **Heavy-Tailed.**

(f) **Rank-collapse.**

**Figure:** The 5+1 phases of learning we identified in DNN training.
Bulk + Spikes: Small Models $\sim$ Tikhonov regularization

Low-rank perturbation

$$W_l \approx W_l^{\text{rand}} + \Delta^{\text{large}}$$

Perturbative correction

$$\lambda_{\text{max}} = \sigma^2 \left( \frac{1}{Q} + \frac{|\Delta|^2}{N} \right) \left( 1 + \frac{N}{|\Delta|^2} \right)$$

$$|\Delta| > (Q)^{-\frac{1}{4}}$$

Simple scale threshold

$$x = \left( \hat{X} + \alpha I \right)^{-1} \hat{W}^T y$$

eigenvalues $> \alpha$ (Spikes) carry most of the signal/information

Smaller, older models like LeNet5 exhibit traditional regularization and can be described perturbatively with Gaussian RMT.
Heavy-tailed Self-regularization

\( W \) is strongly-correlated and highly non-random

- We model strongly-correlated systems by heavy-tailed random matrices
- We model signal (not noise) by heavy-tailed random matrices

Then RMT/MP ESD will also have heavy tails.

- The eigenvalues are heavy-tailed; the weights are NOT.

“All” larger, modern DNNs exhibit novel Heavy-tailed self-regularization
Introduction
Sample covariance matrix for large dimensional data: from concentration to RMT
A random matrix perspective of the “curse of dimensionality”
Kernel spectral clustering for large dimensional data
A random matrix approach to large neural networks and random features
Watching weights with WeightWatcher

https://github.com/CalculatedContent/WeightWatcher

Analyzing DNN Weight matrices with **WeightWatcher**

1. Take a model
2. Take a weight matrix
3. Do Spectral analysis
4. Histogram of eigenvalues

- Analyze one layer of pre-trained model
- Compare multiple layers of pre-trained model
- Monitor NN properties as you train your own model

“pip install weightwatcher”
Using the theory

Different ways one could *use* a theory.

- Perform diagnostics for model validation, to develop hypotheses, etc.*
- Make predictions about model quality, generalization, transferability, etc.*
- Did post-training modifications damage my model?*
- Will buying more data help?*
- Will training longer help?*
- Will quantizing or distilling help?*
- Construct a regularizer to do model training.**

*Ideally, by peeking at very little or no data.

**If you have lots of data, lots of GPUs, etc.
Predicting test accuracies ... lots of metrics ...

- **Average log norm** (a VC-like data-dependent capacity metric):
  \[
  \langle \log \|\mathbf{W}\| \rangle = \frac{1}{N} \sum_{l,i} \log \|\mathbf{W}_{l,i}\| = \frac{1}{N} \sum_{l,i} \log(\lambda_{l,i}^{\text{max}})
  \]

- **Average alpha** (also data-dependent, from HT-SR theory):
  \[
  \alpha = \frac{1}{N} \sum_{l,i} \alpha_{l,i}
  \]

- **Combine the two** into a weighted average (weighted to compensate for different size and scale of feature maps):
  \[
  \hat{\alpha} = \frac{1}{N} \sum_{l,i} \log(\lambda_{l,i}^{\text{max}})\alpha_{l,i}
  \]

- In a special case (\(\alpha \approx 2\)), for each layer:

  **PL–Norm Relation:** \(\alpha \log \lambda^{\text{max}} \approx \log \|\mathbf{W}\|^2_F\).

“pip install weightwatcher”
(The first) large-scale study (meta-analysis) of hundreds of 
SOTA pretrained models ‡

Summary statistics: VGG; ResNet; DenseNet.

Table 1: Quality metrics (for RMSE, smaller is better; for $R^2$, larger is better; and for Kendall-τ rank correlation, larger magnitude is better) for reported top 1 test error for pretrained models in each architecture series. Column # refers to number of models. VGG, ResNet, and DenseNet were pretrained on ImageNet. ResNet-1K was pretrained on ImageNet-1K.

Table 3: Comparison of linear regression fits for different average Log Norm and Weighted Alpha metrics across 5 CV datasets, 17 architectures, covering 108 (out of over 400) different pretrained models.

Different metrics on pre-trained VGG.

Summary statistics: hundreds of models.

Lots more plots to prove we can “predict trends . . . without access . . . ”

Using a theory: on SOTA models

Analyzing pre-trained models: properties of VGG vs ResNet vs DenseNet leads to the idea of correlation flow.

![Graphs showing depth vs VGG, ResNet, DenseNet, and ResNet (overlaid)]

Figure 4: PL exponent (α) versus layer id, for the least and the most accurate models in VGG (a), ResNet (b), and DenseNet (c) series. (VGG is without BN; and note that the Y axes on Alpha versus depth: VGG, ResNet, DenseNet.
Using a theory: on SOTA models

Analyzing pre-trained models: properties of GPTx series leads to the idea of *scale collapse*.

Figure 6: Histogram of PL exponents and Log Spectral Norms for weight matrices from the OpenAI GPT and GPT2-small pretrained models.

Figure 7: Log Spectral Norms (in (a)) and PL exponents (in (b)) for weight matrices from the OpenAI GPT and GPT2-small pretrained models. (Note that the quantities shown on each Y axis are different.) In the text, this is interpreted in terms of Scale Collapse and Correlation Flow.

Histogram and depth plots of $\alpha_{l,i}$ and $\lambda_{l,i}^{\text{max}}$. 
Using a theory: easy to break popular SLT metrics

Easy to “break” popular SLT metrics:
- they are *not* validated counterfactually
- (but they drive the development of models)

Intel’s distillation “broke” their models.

GPTx series: how does a model trained to “bad” data differ from one trained to “good” data?
Using a theory: leads to predictions

Based on analyzing hundreds of pre-trained SOTA models:

- **“Correlation flow”**:  
  ▶ “Shape” of ESD of adjacent layers, as well as overlap between eigenvectors of adjacent layers, should be well-aligned.

- **“Scale collapse”**:  
  ▶ “Size” of ESD of one or more layers changes dramatically, while the size of other layers changes very little, as a function of some perturbation of a model, during training (or post-training modification).

- **“Correlation traps”**:  
  ▶ Spuriously large eigenvalues§ may appear, and they may even be important for model convergence.

We can measure these quantities with Weightwatcher—so can you!

---

Table of Contents

1 Practice

2 Theory

3 More practice

4 Theorems
  - Introduction
  - Sample covariance matrix for large dimensional data: from concentration to RMT
  - A random matrix perspective of the “curse of dimensionality”
  - Kernel spectral clustering for large dimensional data
  - A random matrix approach to large neural networks and random features
Understanding the mechanism of large dimensional machine learning

- Big Data era: exploit large $n, p, N$
- counterintuitive phenomena when $n \gg p$, e.g., the "curse of dimensionality"
- complete change of understanding of many ML algorithms
- **RMT** provides the tools!

Large dimensional data

$x_1, \ldots, x_n \in \mathbb{R}^p$

Large learning systems of size $N$
From low to high dimensional machine learning

- **low dimension**: data vectors $x_i \in \mathbb{R}^p, p = 2, 3$, gathered in different “groups” can be classified using distance-based approach

- **high dimension**:
  1. easy or trivial scenario where low dimensional intuition holds and a pairwise distance-based classification approach via, e.g., Johnson–Lindenstrauss lemma, is efficient;
  2. hard or non-trivial scenario where such intuition collapses: data vectors at approximately the same Euclidean distance, regardless their arising from same or different classes.

Figure: Visual representation of classification in (left) small and (right) large dimensions.
Non-trivial high dimensional classification beyond the JL regime

In the high dimensional regime where data dimension $p$ and sample size $n$ both large, a dual phenomenon:

(i) data points not pairwise classifiable: Euclidean distance between any two data points $x_i \in C_a$ and $x_j \in C_b$ approximately constant $\approx \tau = O(1)$ independent of their classes $C_a, C_b$: $\|x_i - x_j\|^2 / p = \tau + o(1)$ as $n, p \to \infty$ and data pairs neither close nor far from each other;

(ii) classification remains possible by exploiting the spectral information of large Euclidean distance matrix $E = \{\|x_i - x_j\|^2 / p\}_{i,j=1}^n$, thanks to a collective behavior of all data belonging to same (and large) classes.
Figure: Euclidean distance matrices $E$, the histogram of the entries of $E$, and the second top eigenvectors $v_2$, for small (left, $p = 5$) and large (right, $p = 250$) dimensional data $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p \times n}$ with $x_1, \ldots, x_{n/2} \in \mathcal{C}_1$ and $x_{n/2+1}, \ldots, x_n \in \mathcal{C}_2$ for $n = 5000$ and different values of $p$. 
Sample covariance matrix in the large $n, p$ regime

- For $x_i \sim \mathcal{N}(0, C)$, estimate **population covariance** $C \in \mathbb{R}^{p \times p}$ from $n$ data samples $X = [x_1, \ldots, x_n] \in \mathbb{R}^{p \times n}$.

- Maximum likelihood sample covariance matrix with **entry-wise** convergence

$$
\hat{C} = \frac{1}{n} \sum_{i=1}^{n} x_i x_i^T = \frac{1}{n} XX^T \in \mathbb{R}^{p \times p}, \quad [\hat{C}]_{ij} \to [C]_{ij}
$$

almost surely as $n \to \infty$: optimal for $n \gg p$ (or, for $p$ “small”).

- In the regime $n \sim p$, conventional wisdom breaks down:
  for $C = I_p$ with $n < p$, $\hat{C}$ has at least $p - n$ zero eigenvalues.

  $$
  \|\hat{C} - C\| \nrightarrow 0, \quad n, p \to \infty
  $$

  ⇒ eigenvalue mismatch and not consistent! ⇒ matrix norms not equivalent in large dimensions!

- due to $\|A\|_{\infty} \leq \|A\| \leq p \|A\|_{\infty}$ for $A \in \mathbb{R}^{p \times p}$ and $\|A\|_{\infty} \equiv \max_{ij} |A_{ij}|$. 
Quantitative spectral characterization of sample covariance

**Theorem (Concentration of sample covariance, [Ver18, Theorem 4.6.1])**

Let $X \in \mathbb{R}^{p \times n}$ be a random matrix with i.i.d. sub-gaussian columns $x_i \in \mathbb{R}^p$ such that $E[x_i] = 0$ and $E[x_i x_i^T] = I_p$, one has, with probability at least $1 - 2 \exp(-t^2)$ for any $t \geq 0$ that

$$\|\hat{C} - I_p\| \leq C_1 \max(\delta, \delta^2), \quad \delta = C_2(\sqrt{p/n} + t/\sqrt{n})$$

for some constants $C_1, C_2 > 0$ independent of $n, p$.

- Non-asymptotic and high probability characterization
- However, not precise in the $p \sim n$ regime, since $\delta = O(\sqrt{p/n}) = O(1)$

**Theorem (Marčenko-Pastur law, [MP67])**

Under the same setting of Theorem 1, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, with probability one, the empirical spectral measure $\mu_{\hat{C}} = \frac{1}{p} \sum_{i=1}^{p} \delta_{\lambda_i(\hat{C})}$ of $\hat{C} = \frac{1}{n}XX^T$ converges weakly to a probability measure $\mu$ given explicitly by

$$\mu(dx) = (1 - c^{-1})^+ \delta_0(x) + \frac{1}{2\pi cx} \sqrt{(x - E_-)^+(E_+ - x)^+} \, dx$$

where $E_{\pm} = (1 \pm \sqrt{c})^2$ and $(x)^+ = \max(0, x)$, and is known as the Marčenko-Pastur law.
Two ways of spectral characterization of sample covariance

- matrix concentration-type characterization
  \[ \| \hat{C} - I_p \| \leq C_1 \max(\delta, \delta^2), \quad \delta = C_2(\sqrt{p/n} + t/\sqrt{n}) \]
  ⇒ non-asymptotic characterization of small dimensional intuition: how \( \hat{C} \) concentrates around \( I_p \);
- random matrix-type characterization of precise eigenvalue distribution
  \[ \mu(dx) = (1 - c^{-1}) + \delta(x) + \frac{1}{2\pi cx} \sqrt{(x - E_-) + (E_+ - x)} + dx \]
  ⇒ asymptotic characterization (as \( n, p \to \infty \)) of large dimensional intuition: how \( \hat{C} \) differs from \( I_p \! \! \! \)!

**Figure:** Histogram of the eigenvalues of \( \hat{C} \) (blue) versus the Marčenko-Pastur law (red), for \( X \) having standard Gaussian entries in different settings: (left: small versus large dimensional intuition) \( p = 20, n = 1000p \) versus \( p = 20, n = 100p \); and (right: non-asymptotic versus asymptotic MP law) \( p = 20, n = 100p \) versus \( p = 500, n = 100p \).
When is one in the random matrix regime? Almost always!

What about \( n = 100p \)? For \( \mathbf{C} = \mathbf{I}_p \), as \( n, p \to \infty \) with \( p/n \to c \in (0, \infty) \): the Marčenko–Pastur law

\[
\mu(dx) = (1-c^{-1})^+ \delta(x) + \frac{1}{2\pi cx} \sqrt{(x-E_-)^+(E_+-x)^+} dx
\]

where \( E_- = (1-\sqrt{c})^2 \), \( E_+ = (1+\sqrt{c})^2 \) and \( (x)^+ \equiv \max(x, 0) \). Close match!

Figure: Eigenvalue distribution of \( \hat{\mathbf{C}} \) versus Marčenko-Pastur law, \( p = 500, n = 50\,000 \).

- eigenvalues span on \([E_- = (1-\sqrt{c})^2, E_+ = (1+\sqrt{c})^2]\).
- for \( n = 100p \), on a range of \( \pm 2\sqrt{c} = \pm 0.2 \) around the population eigenvalue 1.
Beyond eigenvalue distribution: a modern RMT approach via the resolvent

This **change-of-intuition** leads to very different behavior for small- versus large-dimensional ML:

- **linear** models: low-rank approximation, spectral classification/clustering, and linear least squares regression in high dimensions different from their small dimensional counterparts
- as well as more involved **nonlinear** models: kernel spectral clustering, nonlinear neural nets, etc.

**Technical challenges:**

- classical RMT focuses on eigenvalue distribution
- ML applications need **eigenvectors** and more complex matrix **functionals**!

**Figure:** Different objects of interest and their corresponding technical tools for “old” and “new school” RMT.
"Curse of dimensionality": loss of relevance of Euclidean distance

- Binary Gaussian mixture classification $x \in \mathbb{R}^p$:
  
  $C_1 : x \sim \mathcal{N}(\mu_1, C_1)$, versus $C_2 : x \sim \mathcal{N}(\mu_2, C_2)$;

- Neyman-Pearson test: classification is possible only when [CLM18]
  
  $\|\mu_1 - \mu_2\| \geq C_\mu$, or $\|C_1 - C_2\| \geq C_C \cdot p^{-1/2}$

  for some constants $C_\mu, C_C > 0$.

- In this non-trivial setting, for $x_i \in C_a, x_j \in C_b$:
  
  $$\max_{1 \leq i \neq j \leq n} \left\{ \frac{1}{p} \|x_i - x_j\|^2 - \frac{2}{p} \text{tr} C^\circ \right\} \xrightarrow{a.s.} 0$$

  as $n, p \to \infty$ (i.e., $n \sim p$), for $C^\circ \equiv \frac{1}{2}(C_1 + C_2)$, \underline{regardless} of the classes $C_a, C_b$! (In fact even for $n = p^m$.)

  $\Rightarrow$ Direct consequence to various distance-based machine learning methods (e.g., kernel spectral clustering)!

---

Reminder on kernel spectral clustering

Two-step classification of $n$ data points based on distance kernel matrix $K \equiv \{f(\|x_i - x_j\|^2/p)\}_{i,j=1}^n$:

\[ K \equiv \{f(\|x_i - x_j\|^2/p)\}_{i,j=1}^n \]

isolated eigenvalues

Top eigenvectors

$\downarrow$ Top eigenvectors $\downarrow$

Eigenv. 1

Eigenv. 2
Reminder on kernel spectral clustering

\[ \downarrow K\text{-dimensional representation} \downarrow \]

EM or k-means clustering.
(Three classes/clusters in this example.)
**Objective:** “cluster” Gaussian data $x_1, \ldots, x_n \in \mathbb{R}^p$ into $C_1$ or $C_2$.

Consider Gaussian kernel matrix $K_{ij} = \exp(-\|x_i - x_j\|^2 / 2p)$ and the second top eigenvectors $v_2$ for small (left) and large (right) dimensional data.

(a) $p = 5, n = 500$

$$K = \begin{bmatrix}
C_1 & C_2 \\
C_1 & C_2
\end{bmatrix}$$

$$v_2 = \begin{bmatrix}
\end{bmatrix}$$

(b) $p = 250, n = 500$

$$K = \begin{bmatrix}
\end{bmatrix}$$

$$v_2 = \begin{bmatrix}
\end{bmatrix}$$

**Figure:** Kernel matrices $K$ and the second top eigenvectors $v_2$ for small (left, $p = 5, n = 500$) and large (right, $p = 250, n = 500$) dimensional data.
Kernel matrices for large dimensional real-world data

(a) MNIST

\[
K = \begin{bmatrix}
5 & 0 & 4 & 1 \\
0 & 0 & 0 & 0 \\
4 & 0 & 0 & 0 \\
1 & 0 & 0 & 0
\end{bmatrix}
\]

\[
v_2 = \begin{bmatrix}
5 & 0 & 4 & 1 \\
0 & 0 & 0 & 0 \\
4 & 0 & 0 & 0 \\
1 & 0 & 0 & 0
\end{bmatrix}
\]

(b) Fashion-MNIST

\[
K = \begin{bmatrix}
5 & 0 & 4 & 1 \\
0 & 0 & 0 & 0 \\
4 & 0 & 0 & 0 \\
1 & 0 & 0 & 0
\end{bmatrix}
\]

\[
v_2 = \begin{bmatrix}
5 & 0 & 4 & 1 \\
0 & 0 & 0 & 0 \\
4 & 0 & 0 & 0 \\
1 & 0 & 0 & 0
\end{bmatrix}
\]
A spectral viewpoint of large kernel matrices in large dimensions

▶ “local” linearization of nonlinear kernel matrices in large dimensions, e.g., Gaussian kernel matrix $K_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2} / 2p\right)$ with $C_1 = C_2 = I_p$ (e.g., $C_1 : x_i = \mu_1 + z_i$ versus $C_2 : x_j = \mu_2 + z_j$) so that

$$\|x_i - x_j\|^2 / p \overset{a.s.}{\to} 2, \text{ and } K = \exp\left(-\frac{2}{2}\right) \left(1_n 1_n^T + \frac{1}{p} Z^T Z\right) + g(\|\mu_1 - \mu_2\|) \frac{1}{p} jj^T + * + o_{\|\|}(1)$$

with Gaussian matrix $Z = [z_1, \ldots, z_n] \in \mathbb{R}^{p \times n}$ and $j = \left[1_{n/2}; -1_{n/2}\right]$, the class-information vector

▶ accumulated effect of small “hidden” statistical information ($\|\mu_1 - \mu_2\|$ in this case)

Therefore

▶ entry-wise:

$$K_{ij} = \exp\left(-1\right)\left(1 + \frac{1}{p} z_i^T z_j\right) \pm \frac{1}{p} g(\|\mu_1 - \mu_2\|) + *, \text{ so that } \frac{1}{p} g(\|\mu_1 - \mu_2\|) \ll \frac{1}{p} z_i^T z_j,$$

$O(p^{-1/2})$

$O(p^{-1})$

▶ spectrum-wise: (i) $\|K - \exp(-1)1_n 1_n^T\| \not\to 0$; (ii) $\|\frac{1}{p} Z^T Z\| = O(1)$ and $\|g(\|\mu_1 - \mu_2\|) \frac{1}{p} jj^T\| = O(1)$!

▶ Same phenomenon as the sample covariance example: $[\hat{C} - C]_{ij} \to 0 \not\Rightarrow \|\hat{C} - C\| \to 0$!

$\Rightarrow$ With modern RMT, we understand kernel spectral clustering (eigenvectors!) for large dimensional data!
kernel matrices \( K \in \mathbb{R}^{n \times n} \) from pairwise comparison of \( n \) data points: expansive for \( n \) large

**idea:** find easy-to-compute \( \hat{K} \) to approximate \( K \), e.g., \( \| \hat{K} - K \| \) is small

**example:** random Fourier feature [RR08] \( \Sigma^T = [\cos(WX)^T, \sin(WX)^T] \in \mathbb{R}^{2N \times n} \) of data \( X = [x_1, \ldots, x_n] \in \mathbb{R}^{p \times n} \) with standard Gaussian \( W \in \mathbb{R}^{N \times p} \), i.e., \( W_{ij} \sim \mathcal{N}(0,1) \)

approximates Gaussian kernel \( \exp(\frac{1}{2} \| x_i - x_j \|^2) \): entry-wise convergence of RFF Gram \( \frac{1}{N} [\Sigma^T \Sigma]_{ij} \to [K_{\text{Gauss}}]_{ij} \) Gaussian kernel matrix as number of features \( N \to \infty \)

**proof:** (strong) law of large numbers:

\[
\frac{1}{N} [\Sigma^T \Sigma]_{ij} = \frac{1}{N} \sum_{k=1}^{N} \cos(x_i^T w_k) \cos(w_k^T x_j) + \sin(x_i^T w_k) \sin(w_k^T x_j) \\
\to \mathbb{E}_{w \sim \mathcal{N}(0,I_p)} [\cos(x_i^T w) \cos(w^T x_j) + \sin(x_i^T w) \sin(w^T x_j)] = [K_{\text{cos}} + K_{\text{sin}}]_{ij} = [K_{\text{Gauss}}]_{ij}
\]

for \( K_{\text{cos}} = e^{-\frac{1}{2}(\|x_i\|^2 + \|x_j\|^2)} \cosh(x_i^T x_j) \) and \( K_{\text{sin}} = e^{-\frac{1}{2}(\|x_i\|^2 + \|x_j\|^2)} \sinh(x_i^T x_j) \).

---

Random features-based ridge regression and neural networks

\[
X = \mathbb{R}^{p \times n}
\]
\[
\hat{X} = \mathbb{R}^{p \times \hat{n}}
\]
\[
\Sigma_X^T = \Sigma^T = [\cos(WX)^T, \sin(WX)^T]
\]
\[
\Sigma_{\hat{X}}^T = [\cos(W\hat{X})^T, \sin(W\hat{X})^T]
\]

Figure: Illustration of random Fourier features regression model.

- **RFF ridge regressor** \( \beta \in \mathbb{R}^{2N} \) given by, for regularization penalty \( \gamma \geq 0 \),
  \[
  \beta \equiv \frac{1}{n} \Sigma \left( \frac{1}{n} \Sigma^T \Sigma + \gamma \mathbf{I}_n \right)^{-1} \mathbf{1}_{2N> n} + \left( \frac{1}{n} \Sigma \Sigma^T + \gamma \mathbf{I}_{2N} \right)^{-1} \frac{1}{n} \Sigma \mathbf{y} \cdot \mathbf{1}_{2N<n}. \quad (3)
  \]

- **Performance**: training and test Mean Squared Error (MSE):
  \[
  E_{\text{train}} = \frac{1}{n} \| \mathbf{y} - \Sigma_X^T \beta \|^2 \quad \text{and} \quad E_{\text{test}} = \frac{1}{\hat{n}} \| \hat{\mathbf{y}} - \Sigma_{\hat{X}}^T \beta \|^2,
  \]
  with \( \Sigma_{\hat{X}}^T \in \mathbb{R}^{\hat{n} \times 2N} \) RFFs of a test set \((\hat{X}, \hat{y})\) of size \( \hat{n} \).

- **single-hidden-layer** neural network with cos + sin activations, connected to neural tangent kernel (NTK)

---

Random Fourier features approximate Gaussian kernel, but in which sense?

- [RR08]: entry-wise convergence of RFF Gram $\frac{1}{N}[\Sigma^T \Sigma]_{ij} \rightarrow [K_{Gauss}]_{ij}$ Gaussian kernel matrix as $N \rightarrow \infty$

- again, not true in spectral norm sense, i.e., $\|\Sigma^T \Sigma / N - K_{Gauss}\| \not\rightarrow 0$ unless $N \gg n$
  - e.g., $\Sigma^T \Sigma \in \mathbb{R}^{n \times n}$ of rank at most $N$ if $N \leq n$, while $K_{Gauss}$ of rank $n$ (for distinct $x_i$)
  - significant impact on various RFF-based algorithms

Figure: Training MSEs of RFF ridge regression on MNIST data (class 3 versus 7) as a function of regression penalty $\lambda$.

- effective kernel can be derived with RMT in the large $n, p, N$ regime
- provides precise training and test performances of RFF for any ratio $N/n$, more practical and more flexible, recover Gaussian kernel result with $N/n \rightarrow \infty$
- data-dependent theory with no strong assumption on data
Sharp analysis of RFF ridge regression performance via RMT

Figure: MSEs of RFF ridge regression on Fashion- (left two) and Kannada-MNIST (right two).

Figure: Test MSEs of RFF regression as a function of the ratio $N/n$, on MNIST data set.
“Recap” for double descent phenomenon for over-parameterized models

Figure: Comparison between training risk (blue) and true/test risk (red).

- empirically observed for various large-scale machine learning models, e.g., RF-based methods, decision trees, ensemble methods, and deep NNs
- proved here for RFF on real-world data!
- phase transition from under- to over-param of resolvent \((\Sigma^T \Sigma + \lambda I_n)^{-1}\) in the ridgeless \(\lambda \to 0\) limit


Take-away messages:

▶ RF methods: classical statistical learning theory provides performance guarantee for $N \gg n, p$
▶ here we derive (limiting) kernel in the more practical large $n, p, N$ regime
▶ fast tuning of regularization parameter $\lambda$
▶ double descent theory for novel understanding of over-parameterized neural networks

References:

Random matrix theory (RMT) for machine learning:

- **change of intuition** from small to large dimensional learning paradigm!
- **better understanding** of existing methods: why they work if they do, and what the issue is if they do not
- **improved novel methods** with performance guarantee!

Thank you! Q & A?