A Random Matrix Theory Analysis of Linear Generative Models

Mohamed El Amine Seddik
Technology Innovation Institute
`melaseddik.github.io`

Khalifa University Mathematics Seminar

Abu Dhabi November 23rd 2023

Technology Innovation Institute

Outline

Large Language Models

- LLMs became popular due to assistant chatbots (e.g., chatGPT).
- Rely on foundational models through **self-supervised pre-training**.

Given a corpus of vocabulary size k:

$$\operatorname*{arg\,min}_{\mathbf{W}\in\mathbb{R}^{d\times k},\,\phi} -\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i^\top \log\sigma\left(\mathbf{W}^\top\phi(\mathbf{X}_i)\right) \qquad \sigma(v) = \frac{\exp(\boldsymbol{v})}{\sum_{j=1}^{k}\exp(v_j)}$$

- $\mathbf{X}_i \in \mathbb{R}^{d\times \ell_i}$ is a context sequence (of *embeddings*).
- $\boldsymbol{y}_i \in \mathbb{R}^k$ is a canonical vector encoding the next *token*.
- $\phi : \mathbb{R}^{d\times \ell_i} \to \mathbb{R}^d$ is a sequence encoder (*transformer* architecture).

NLP terminology:

- *Tokenization*: breaking down text into smaller units "tokens".
- *Embedding*: convert tokens into high-dimensional vectors.

`“Mathematics is the giving of the same name to different things.” HP.`

[2118, 8991, 34805, 374, 279, 7231, 315, 279, 1890, 836, 311, 2204, 2574, 2029, 12478, 13][1]

[1] https://platform.openai.com/tokenizer

Transformers: The Core Mechanism

Let $\mathbf{X}_1 = \mathbf{X} \in \mathbb{R}^{d\times\ell}$ be an input sequence of ℓ embeddings of dimension d.

The unmasked self-attention[2] layer $g_l : \mathbf{X}_l \in \mathbb{R}^{d\times\ell} \mapsto \mathbf{X}_{l+1} \in \mathbb{R}^{d\times\ell}$ is:

$$\mathbf{Y}_l = \underbrace{\mathbf{W}_v\mathbf{X}_l}_{\text{value}}\ \underbrace{\mathbf{A}_l}_{\text{attention}} + \mathbf{X}_l \quad \text{with} \quad \mathbf{A}_l = \sigma\left(d^{-\frac{1}{2}}\, \underbrace{(\mathbf{W}_k\mathbf{X}_l)^\top}_{\text{key}}\, \underbrace{\mathbf{W}_q\mathbf{X}_l}_{\text{query}}\right) \in \mathbb{R}^{\ell\times\ell}$$

$$\mathbf{X}_{l+1} = \mathbf{W}_2^\top f\left(\mathbf{W}_1^\top \mathbf{Y}_l\right) + \mathbf{Y}_l$$

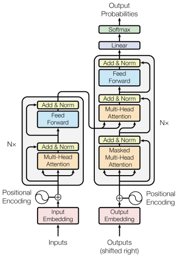
A transformer is a composition of D layers and $\phi : \mathbb{R}^{d\times\ell} \to \mathbb{R}^d$ is:

$$\phi(\mathbf{X}) = [g_D \circ \cdots \circ g_1(\mathbf{X})]_{:,\ell}$$

where $[\mathbf{M}]_{:,i}$ is the i-th column of $\mathbf{M}$.

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com



[2] Ashish Vaswani, et al. **"Attention is all you need"**, Neurips 2017.

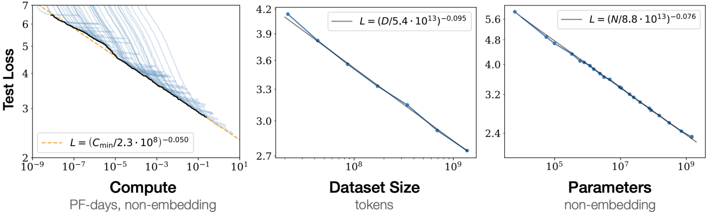
Motivation for a Theoretical Framework

Ultimately:

- Understand generalization: Express **test loss** in terms of hyperparameters.
- Uncertainty estimation: Control output model **bias** and **variance**.
- Predict **scaling laws**[3] theoretically:

$$\mathcal{L}(d,n) = \left[\left(\frac{d_c}{d} \right)^{\frac{\alpha_d}{\alpha_n}} + \frac{n_c}{n} \right]^{\alpha_n}$$

where d is number of parameters and n is dataset size.



[3] Jared Kaplan, et al. **"Scaling laws for neural language models"**, arXiv:2001.08361 (2020).

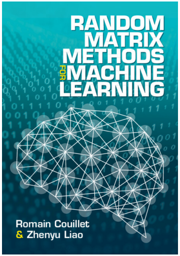
Outline

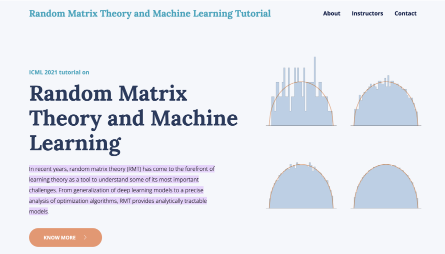
Why RMT?

- The problem involves high-dimensions: both d and n are large!
- Estimating test loss: a scalar quantity function of a random matrix (e.g. data matrix).

RMT has been applied to analyze a wide range of ML problems[4][5]:

- Kernel Methods.
- Large Neural Networks & NTKs.
- Implicit Convex Optimization Problems.
- Unsupervised, Semi-supervised, Transfer and Multi-task Learning.





[4] Romain Couillet and Zhenyu Liao, "**Random matrix methods for machine learning**", Cambridge University Press, 2022.

[5] `https://random-matrix-learning.github.io/`

Example: Large Sample Covariance Matrices

Let $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d\times n}$ with $\boldsymbol{x}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

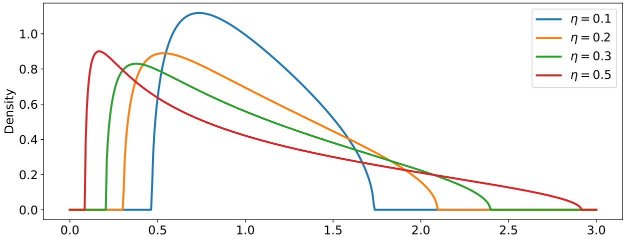
- Classical statistics: when $n \to \infty$ and d is **fixed**, with maximum likelihood:

$$\hat{\mathbf{C}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top = \frac{1}{n}\mathbf{X}\mathbf{X}^\top \xrightarrow{\text{a.s.}} \boldsymbol{\Sigma}$$

- RMT regime: both $d, n \to \infty$, curse of dimension occurs:

$$\|\hat{\mathbf{C}} - \boldsymbol{\Sigma}\| \not\to 0 \quad \text{as} \quad \frac{d}{n} \to \eta \in (0, \infty)$$

- Marchenko-Pastur Law (1967): when $\boldsymbol{\Sigma} = \mathbf{I}_d$:



$$f(x) = \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{2\pi\eta x}\mathbf{1}_{x\in[\lambda_-, \lambda_+]} \quad \text{with} \quad \lambda_\pm = (1 \pm \sqrt{\eta})^2$$

RMT Tools: Spectral Measure & Stieltjes Transform

- Let $\mathbf{S} \in \mathbb{R}^{d \times d}$ some **symmetric** random matrix and λ_i its eigenvalues.
- Originally, RMT is about characterizing the **spectrum** of $\mathbf{S}$ when $d \to \infty$.
- Under control of the moments of the entries of $\mathbf{S}$:

$$\nu_d = \frac{1}{d} \sum_{i=1}^{d} \delta_{\lambda_i} \xrightarrow[d\to\infty]{w} \nu \quad \text{(in the weak sense)}$$

  where ν is a **deterministic** probability measure.
- **Stieltjes Transform** of a probability measure ν is:

$$g_\nu(z) = \int \frac{d\nu(\lambda)}{\lambda - z} \quad z \in \mathbb{C} \setminus \mathrm{Supp}(\nu)$$

- **Equivalence**: Let $(\nu_d)_{d\in\mathbb{N}}$ be a sequence of probability measures. Then:

$$\nu_d \xrightarrow[d\to\infty]{w} \nu \quad \Leftrightarrow \quad g_{\nu_d}(z) \xrightarrow[d\to\infty]{\text{a.s.}} g_\nu(z) \quad \text{for all} \quad z \in \mathbb{C} \setminus \mathrm{Supp}(\nu)$$

- **Resolvent**: Let $\mathbf{Q}(z) = (\mathbf{S} + z\mathbf{I}_d)^{-1}$, we have:

$$g_{\nu_d}(z) = \frac{1}{d} \sum_{i=1}^{d} \frac{1}{\lambda_i - z} = \frac{1}{d} \operatorname{Tr} \mathbf{Q}(-z)$$

- $g_{\nu_d}(z)$ is a linear form of $\mathbf{Q}(-z)$.

RMT Tools: Concentration & Deterministic Equivalent

Concentration (Trace Lemma):

- $\boldsymbol{x} \in \mathbb{R}^d$ random with finite $2m$ order moment and let $\boldsymbol{\Sigma} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$.
- $\mathbf{A} \in \mathbb{R}^{d\times d}$ independent of $\boldsymbol{x} \sim \mathcal{L}(\mathbf{0}, \boldsymbol{\Sigma})$ and $\|\mathbf{A}\|, \|\boldsymbol{\Sigma}\| < \infty$.

Then:

$$\mathbb{E}_{\boldsymbol{x}}\left[\left|\frac{1}{d}\boldsymbol{x}^\top \mathbf{A}\boldsymbol{x} - \frac{1}{d}\operatorname{Tr}(\boldsymbol{\Sigma}\mathbf{A})\right|^m\right] \leq C\, d^{-\frac{m}{2}} \quad \Rightarrow \quad \boxed{\frac{1}{d}\boldsymbol{x}^\top \mathbf{A}\boldsymbol{x} \xrightarrow{\text{a.s.}} \frac{1}{d}\operatorname{Tr}(\boldsymbol{\Sigma}\mathbf{A})}$$

Deterministic Equivalent:

- **Definition**: $\mathbf{Q} \leftrightarrow \bar{\mathbf{Q}}$ if $u(\mathbf{Q} - \bar{\mathbf{Q}}) \xrightarrow{\text{a.s.}} 0$ for any bounded linear form $u : \mathbb{R}^{d\times d} \to \mathbb{R}$.

Let $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_d] \in \mathbb{R}^{d\times n}$ with $\boldsymbol{x}_i \sim \mathcal{L}(\mathbf{0}, \boldsymbol{\Sigma})$ and independent, then[6]:

$$\boxed{\mathbf{Q}(z) = \left(\frac{1}{n}\mathbf{X}\mathbf{X}^\top + z\mathbf{I}_d\right)^{-1} \leftrightarrow \bar{\mathbf{Q}}(z) = \left(\frac{\boldsymbol{\Sigma}}{1+\delta(z)} + z\mathbf{I}_d\right)^{-1}}$$

where $\delta(z) = \frac{1}{n}\operatorname{Tr}\left(\boldsymbol{\Sigma}\bar{\mathbf{Q}}(z)\right)$.

- **Limiting Stieltjes transform** is given by $g_\nu(z) = \frac{1}{d}\operatorname{Tr}\bar{\mathbf{Q}}(-z)$.

[6] Cosme Louart and Romain Couillet, **"Concentration of measure and large random matrices with an application to sample covariance matrices"**, arXiv:1805.08295 (2018).

RMT Tools: Sketch of Proof

Let $\mathbf{Q}_{-i} = \left(\frac{1}{n}\mathbf{X}\mathbf{X}^\top - \frac{1}{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top + z\mathbf{I}_d\right)^{-1}$ and $\bar{\mathbf{Q}} = (\mathbf{F} + z\mathbf{I}_d)^{-1}$, with:

$$\mathbf{Q} = \mathbf{Q}_{-i} - \frac{\mathbf{Q}_{-i}\frac{1}{n}\boldsymbol{x}_i\boldsymbol{x}_i^\top\mathbf{Q}_{-i}}{1 + \frac{1}{n}\boldsymbol{x}_i^\top\mathbf{Q}_{-i}\boldsymbol{x}_i} \quad \mathbf{Q}\boldsymbol{x}_i = \frac{\mathbf{Q}_{-i}\boldsymbol{x}_i}{1 + \frac{1}{n}\boldsymbol{x}_i^\top\mathbf{Q}_{-i}\boldsymbol{x}_i}$$

and $\mathbf{A}^{-1} - \mathbf{B}^{-1} = \mathbf{A}^{-1}(\mathbf{B} - \mathbf{A})\mathbf{B}^{-1}$.

$\mathbf{Q}$ concentrates around $\mathbb{E}[\mathbf{Q}]$ is the sense of deterministic equivalents[7], and:

$$\mathbb{E}\left[\mathbf{Q} - \bar{\mathbf{Q}}\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left[\mathbf{Q}_{-i}\left(\frac{\boldsymbol{x}_i\boldsymbol{x}_i^\top}{1 + \frac{1}{n}\boldsymbol{x}_i^\top\mathbf{Q}_{-i}\boldsymbol{x}_i} - \mathbf{F}\right)\bar{\mathbf{Q}}\right] + \mathcal{O}(n^{-1})$$

By trace lemma:

$$\frac{1}{n}\boldsymbol{x}_i^\top\mathbf{Q}_{-i}\boldsymbol{x}_i \xrightarrow{\text{a.s.}} \frac{1}{n}\operatorname{Tr}\left(\boldsymbol{\Sigma}\mathbb{E}\left[\mathbf{Q}_{-i}\right]\right) = \frac{1}{n}\operatorname{Tr}\left(\boldsymbol{\Sigma}\bar{\mathbf{Q}}\right) + \mathcal{O}(n^{-1})$$

$$\Rightarrow \quad \bar{\mathbf{Q}} = \left(\frac{\boldsymbol{\Sigma}}{1+\delta} + z\mathbf{I}_d\right)^{-1} \quad \text{with} \quad \delta = \frac{1}{n}\operatorname{Tr}\left(\boldsymbol{\Sigma}\bar{\mathbf{Q}}\right)$$

- **Remark**: $\delta \to 0$ if $n \to \infty$ with d fixed.

[7] Walid Hachem, Philippe Loubaton and Jamal Najim, **"Deterministic equivalents for certain functionals of large random matrices"**, (2007): 875-930.

Outline

Statistical Model & Linear Generative Model

Statistical Data Model:

- Denote k vocabulary size and ℓ context length (possible contexts $c = k^\ell$).
- n context representations $\boldsymbol{x}_i = \phi(\mathbf{X}_i) \in \mathbb{R}^d$ and next tokens $\boldsymbol{y}_i \in \mathbb{R}^k$:

$$\boldsymbol{x}_i = \boldsymbol{z}_a \sim \mathcal{L}(\mathbf{0}, \mathbf{I}_d) \quad \text{with} \quad \mathbb{P}\{\boldsymbol{x}_i = \boldsymbol{z}_a\} = \alpha_a / c \quad a \in [c]$$

$$\boldsymbol{y}_i \sim \mathbb{P}\{\cdot \mid \boldsymbol{x}_i = \boldsymbol{z}_a\} \quad \text{s.t.} \quad p_{aj} = \mathbb{P}\{y_{ij} = 1 \mid \boldsymbol{x}_i = \boldsymbol{z}_a\}$$

- We want a generative model to learn:

$$\boldsymbol{p}_a = (p_{aj})_{j \in [k]} \in \mathbb{R}^k \quad \text{and} \quad \boxed{\mathbf{P} = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_c] \in \mathbb{R}^{k \times c}}$$

- From data matrix and labels:

$$\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n} \quad \text{and} \quad \mathbf{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n] \in \mathbb{R}^{k \times n}$$

Linear Generative Model:

- Consider a *linear Ridge generative model*:

$$\mathcal{L}(\mathbf{W}) = \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{y}_i - \mathbf{W}^\top \boldsymbol{x}_i\|^2 + \gamma \|\mathbf{W}\|_\mathsf{F}^2$$

- For a given context $a \in [c]$, forward pass is:

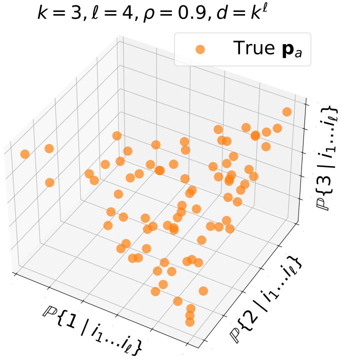
$$\boxed{\hat{\boldsymbol{p}}_a = \mathbf{W}^\top \boldsymbol{z}_a \in \mathbb{R}^k} \quad \mathbf{W} = \frac{1}{n} \mathbf{Q}(\gamma) \mathbf{X} \mathbf{Y}^\top, \quad \mathbf{Q}(z) = \left(\frac{1}{n} \mathbf{X} \mathbf{X}^\top + z \mathbf{I}_d \right)^{-1}$$

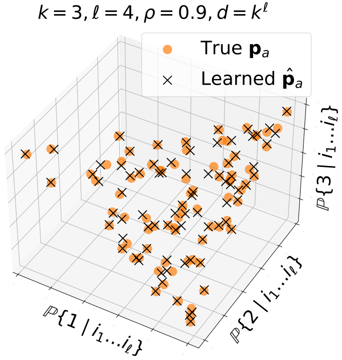
Example: $k = 3$ & $\ell = 4$

A toy model for p_{aj} is:

$$p_{aj} = \frac{\exp\left(G_{aj}/\rho\right)}{\sum_{b=1}^{k} \exp\left(G_{ab}/\rho\right)}$$

where $\mathbf{G} = (G_{ab}) \in \mathbb{R}^{c \times k}$ is random with $\mathcal{N}(0, 1)$ i.i.d. entries and $\rho > 0$.





Model Output: Expectation & Covariance

Proposition: As $d, c \to \infty$ with $\frac{d}{c} \to \eta \in (0, \infty)$, for all $a \in [c]$:

$$\left\| \mathbb{E}\left[\hat{\boldsymbol{p}}_a\right] - \boldsymbol{m}_a \right\| \leq \mathcal{O}(d^{-\frac{1}{2}}) \quad \text{and} \quad \left\| \mathbb{E}\left[\hat{\boldsymbol{p}}_a \hat{\boldsymbol{p}}_a^\top\right] - \mathbf{C}_a \right\| \leq \mathcal{O}(d^{-\frac{1}{2}})$$

where:

$$\boldsymbol{m}_a = \frac{\alpha_a \delta}{1 + \alpha_a \delta} \boldsymbol{p}_a, \quad \mathbf{C}_a = \left(\frac{\alpha_a \delta}{1 + \alpha_a \delta} \right)^2 \boldsymbol{\Sigma}_a + \frac{\frac{\kappa}{c} \sum_{b \neq a}^{c} \frac{\alpha_b^2}{(1+\alpha_b \delta)^2} \boldsymbol{\Sigma}_b}{(1 + \alpha_a \delta)^2 (1 - \beta \kappa)}$$

where $\boldsymbol{\Sigma}_a = \frac{c}{\alpha_a n} \operatorname{Diag}(\boldsymbol{p}_a) + \left(1 - \frac{c}{\alpha_a n}\right) \boldsymbol{p}_a \boldsymbol{p}_a^\top$ and

$$\delta = \frac{\eta}{\alpha + \gamma}, \; \alpha = \frac{1}{c} \sum_{a=1}^{c} \frac{\alpha_a}{1 + \alpha_a \delta}, \; \kappa = \frac{\eta}{(\alpha + \gamma)^2}, \; \beta = \frac{1}{c} \sum_{a=1}^{c} \left(\frac{\alpha_a}{1 + \alpha_a \delta} \right)^2$$

Some remarks:

- If $\alpha_a = 1$ then $\delta = \frac{\eta - \gamma - 1 + \sqrt{(\eta - \gamma - 1)^2 + 4\gamma}}{2\gamma}$ (Marchenko-Pastur result).
- The model is unbiased if δ is large.
- Variance reduces if n is large ($n \gg c$).
- The model has a larger variance on unrepresented contexts (small α_a).

Model Output: Simulations

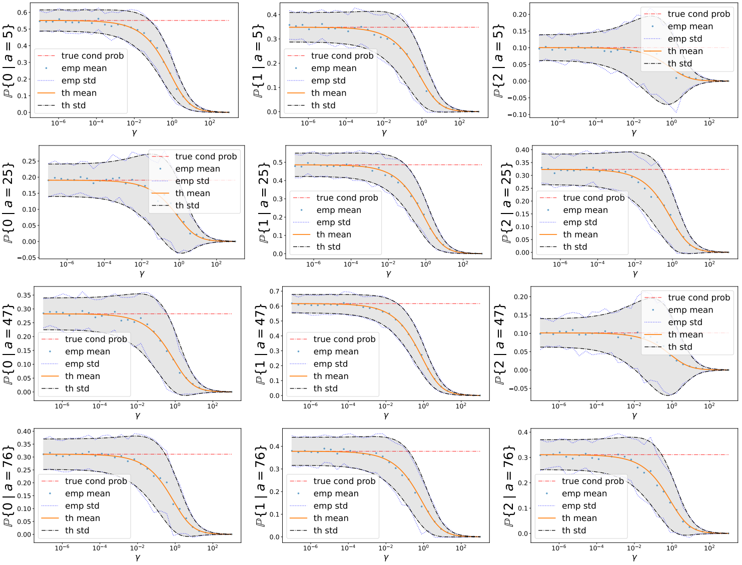
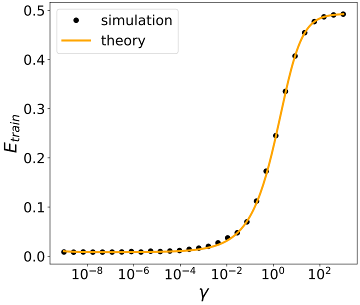


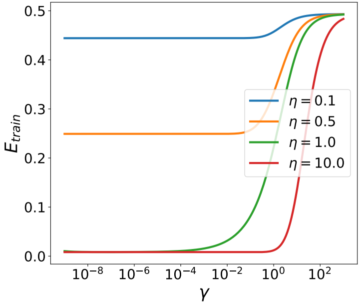
Figure: $k = 3$, $\ell = 4$, $d = k^\ell = 81$, $\rho = .9$, $\alpha_a = 1$ and $n = 5000$.

Training Error (Coincides with Test Error)

Let $\mathbf{Z} = [\boldsymbol{z}_1, \ldots, \boldsymbol{z}_c] \in \mathbb{R}^{d \times c}$ and $\mathbf{D} = \operatorname{Diag}\left(\frac{\alpha_a \delta}{1+\alpha_a \delta} \mid a \in [c]\right)$, then:

$$E_{\text{train}} = \frac{1}{c}\|\mathbf{P} - \mathbf{W}^\top \mathbf{Z}\|_{\mathsf{F}}^2 \xrightarrow{\text{a.s.}} \frac{1}{c}\operatorname{Tr}\left[\mathbf{P}\left(\mathbf{I}_c - 2\mathbf{D}\right)\mathbf{P}^\top\right] + \frac{1}{c}\sum_{a=1}^{c}\operatorname{Tr}\left(\mathbf{C}_a\right)$$





- The model learns with high-dimensional embeddings ($\eta = \frac{d}{c} \geq 1$).
- The statistical model does not allow us to understand generalization!

Understanding Generalization

Statistical Data Model:

- Denote k vocabulary size and c possible contexts.
- n context representations $\boldsymbol{x}_i = \phi(\mathbf{X}_i) \in \mathbb{R}^d$ and next tokens $\boldsymbol{y}_i \in \mathbb{R}^k$:

$$\boldsymbol{x}_i = \boldsymbol{\mu}_a + \boldsymbol{z}_i \quad \text{with} \quad \boldsymbol{z}_i \sim \mathcal{L}(\mathbf{0}, \mathbf{I}_d) \quad \text{and} \quad \mathbb{P}\{\boldsymbol{x}_i \in \mathcal{C}_a\} = \pi_a,\, a \in [c]$$

$$\boldsymbol{y}_i \sim \mathbb{P}\{\cdot \mid \boldsymbol{x}_i \in \mathcal{C}_a\} \quad \text{s.t.} \quad p_{aj} = \mathbb{P}\{y_{ij} = 1 \mid \boldsymbol{x}_i \in \mathcal{C}_a\}$$

- We want a generative model to learn:

$$\boldsymbol{p}_a = (p_{aj})_{j \in [k]} \in \mathbb{R}^k \quad \text{and} \quad \boxed{\mathbf{P} = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_c] \in \mathbb{R}^{k \times c}}$$

- From data matrix and labels:

$$\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \mathbb{R}^{d \times n} \quad \text{and} \quad \mathbf{Y} = [\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n] \in \mathbb{R}^{k \times n}$$

Linear Generative Model:

- Consider a *linear Ridge generative model*:

$$\mathcal{L}(\mathbf{W}) = \frac{1}{n} \sum_{i=1}^{n} \|\boldsymbol{y}_i - \mathbf{W}^\top \boldsymbol{x}_i\|^2 + \gamma \|\mathbf{W}\|_{\mathsf{F}}^2$$

- For $a \in [c]$, forward pass for $\tilde{\boldsymbol{x}}_a = \boldsymbol{\mu}_a + \tilde{\boldsymbol{z}}_a$ with $\tilde{\boldsymbol{z}}_a$ independent of $\mathbf{X}$:

$$\boxed{\hat{\boldsymbol{p}}_a = \mathbf{W}^\top \tilde{\boldsymbol{x}}_a \in \mathbb{R}^k} \quad \mathbf{W} = \frac{1}{n}\mathbf{Q}(\gamma)\mathbf{X}\mathbf{Y}^\top, \quad \mathbf{Q}(z) = \left(\frac{1}{n}\mathbf{X}\mathbf{X}^\top + z\mathbf{I}_d\right)^{-1}$$

Generalization Error

Let $\tilde{\mathbf{X}} = [\tilde{\boldsymbol{x}}_1, \dots, \tilde{\boldsymbol{x}}_c] \in \mathbb{R}^{d\times c}$ and denote $E_{\text{test}} = \frac{1}{c}\|\mathbf{P} - \mathbf{W}^\top \tilde{\mathbf{X}}\|_{\mathsf{F}}^2$.

Proposition: As $d, n \to \infty$ with $\frac{d}{n} \to \eta \in (0, \infty)$ and $c, \|\boldsymbol{\mu}_a\| = \mathcal{O}(1)$:

$$\boxed{\forall \varepsilon > 0, \quad n^{\frac{1}{2}-\varepsilon}\left(E_{\text{test}} - \bar{E}_{\text{test}}\right) \xrightarrow{\text{a.s.}} 0}$$

where, for $\pi_a = \frac{1}{c}$ and $\mathbf{M} = [\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_c] \in \mathbb{R}^{d\times c}$, $\bar{E}_{\text{test}}$ is:

$$\bar{E}_{\text{test}} = \frac{1}{c}\operatorname{Tr}\left[\mathbf{P}\mathbf{P}^\top\left(\mathbf{I}_c - \frac{2\mathbf{M}^\top\bar{\mathbf{Q}}\mathbf{M}}{(1+\delta)c}\right)\right] + \frac{1}{c}\sum_{a=1}^{c}\operatorname{Tr}\left(\mathbf{C}_a\right)$$

$$\mathbf{C}_a = \frac{\tau\sum_{b=1}^{c}\operatorname{Diag}(\boldsymbol{p}_b)}{(1+\delta)^2 c} + \mathbf{P}\mathbf{M}^\top\left(\frac{\bar{\mathbf{Q}}\boldsymbol{\mu}_a\boldsymbol{\mu}_a^\top\bar{\mathbf{Q}} + \mathbf{R}}{(1+\delta)^2c^2} - \frac{2\tau\bar{\mathbf{Q}}}{(1+\delta)^3c^2}\right)\mathbf{M}\mathbf{P}^\top$$

with $\delta = \frac{\eta-\gamma-1+\sqrt{(\eta-\gamma-1)^2+4\gamma}}{2\gamma}$, $\zeta = \gamma + \frac{1}{1+\delta}$, $\tau = \frac{\eta(1+\delta)^2}{\zeta^2(1+\delta)^2-\eta}$, $\kappa = \frac{\eta}{\zeta^2}$

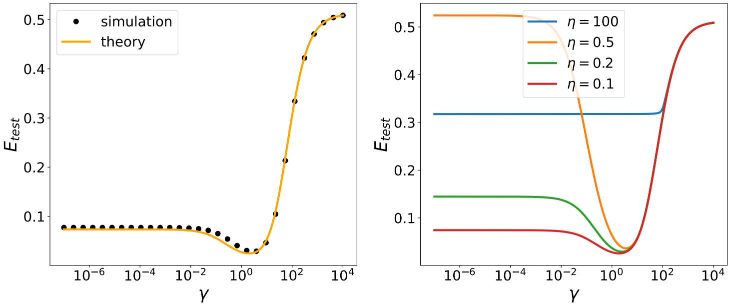
$$\bar{\mathbf{Q}} = \frac{1}{\zeta}\mathbf{I}_p - \frac{1}{\zeta^2}\mathbf{M}\left((1+\delta)c\mathbf{I}_c + \frac{1}{\zeta}\mathbf{M}^\top\mathbf{M}\right)^{-1}\mathbf{M}^\top,\ \mathbf{R} = \frac{\bar{\mathbf{Q}}^2 + \frac{\kappa\bar{\mathbf{Q}}\mathbf{M}\mathbf{M}^\top\bar{\mathbf{Q}}}{(1+\delta)^2c}}{1 - \frac{\kappa}{(1+\delta)^2}}$$

Generalization Error: Simulations

Recall

$$\mathbf{W} = \frac{1}{n}\left(\frac{1}{n}\mathbf{X}\mathbf{X}^\top + \gamma \mathbf{I}_d\right)^{-1}\mathbf{X}\mathbf{Y}^\top$$

- Large γ yields **simple** model: $\mathbf{W} \approx \frac{1}{n\gamma}\mathbf{X}\mathbf{Y}^\top$.
- Small γ yields **complex** model: $\mathbf{W} \approx \left(\mathbf{X}\mathbf{X}^\top\right)^{-1}\mathbf{X}\mathbf{Y}^\top$.



- Generalization depends on **optimal** γ and for small $\eta = \frac{d}{n}$.

Outline

Take Away Messages

- RMT provides tools to assess ML performance when both **sample size and data dimension are large**.
- In this talk, we used these tools for a simple **linear generative model**.
- Provided exact characterization of **train and test errors**.

Limitations:

- Considered square loss, but an **extension**[8] is possible with:

$$\arg\min_{\mathbf{W}\in\mathbb{R}^{d\times k}} -\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i^\top \log\sigma\left(\mathbf{W}^\top \boldsymbol{x}_i\right) + \gamma\|\mathbf{W}\|_\mathsf{F}^2$$

- Extension beyond **convex** problems is required.
- Include **attention** mechanism to understand **feature learning**.

Thank you for your attention!
melaseddik.github.io

[8] Mohamed El Amine Seddik, et al. **"The unexpected deterministic and universal behavior of large softmax classifiers"** AISTATS, PMLR, 2021.