Random Matrix Theory

Analysis of LoRA vs. Full Fine-Tuning

B. Khodabandeh, S. Heidari

Sharif University of Technology
Electrical Engineering Department

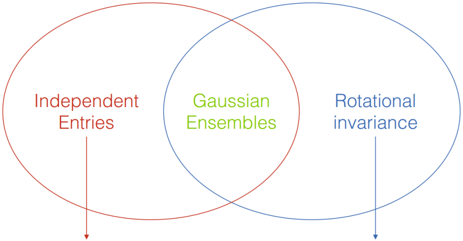
Fall 2024

Outline

Introduction

- Random matrix theory explores the statistical properties of matrices with random entries.
- Applications span Quantum Information Theory, Machine Learning, Statistical Physics, Finance, Trust Fund, 6'5", Blue eyes. . . .
- Key areas of focus include:
  - Joint Distribution of Eigenvalues
  - Joint Distribution of Eigenvectors
  - Expected Empirical Distribution of Eigenvalues
  - The Distribution of Spacings of Eigenvalues
  - Spiked Matrix Models
  - Perturbation Analysis
  - . . .

Layman Classification and Gaussian Ensembles



$\rho[H] \propto \prod_{i=1}^{N} f_i(H_{ii}) \prod_{i<j} f_{ij}(H_{ij})$ $\qquad$ $\rho[H] = \rho[UHU^{-1}]$

$$\rho\left[H_s\right] \propto \exp\left(-\frac{1}{2}\mathrm{Tr}\left(H_s^2\right)\right)$$

- GOE (Gaussian Orthogonal Ensemble): Symmetric matrices.
- GUE (Gaussian Unitary Ensemble): Hermitian matrices.
- GSE (Gaussian Symplectic Ensemble): Quaternionic matrices.

Wishart Ensemble

- Used in statistics for covariance matrices.
- Matrix $W = \frac{1}{n}XX^T$, where $X \in \mathbb{R}^{p \times n}$ is a rectangular matrix with independent entries.
- The entries of X are assumed to have zero mean and variance σ^2.
- **When it applies:** The Marchenko-Pastur law applies to such Wishart (or sample covariance) ensembles.

Marchenko-Pastur Distribution:
For entries with variance σ^2, the distribution is given by:

$$\mu_{\mathrm{MP}} = \frac{\sqrt{(\lambda_+ - x)(x - \lambda_-)}}{2\pi\sigma^2 cx} \mathbf{1}_{[\lambda_-, \lambda_+]}(x)$$

- Here, $\lambda_\pm = \sigma^2(1 \pm \sqrt{c})^2$ and $c = p/n$ represents the limiting aspect ratio.

Empirical Spectral Measure:

$$\mu_{\mathbf{M}} = \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i(\mathbf{M})}$$

- Represents the average of Dirac masses placed at each eigenvalue λ_i of the matrix $\mathbf{M} \in \mathbb{C}^{n \times n}$.
- For large matrices, $\mu_{\mathbf{M}}$ converges to a deterministic limit, thereby capturing the asymptotic spectral distribution.

Key Features:

- Encodes the complete spectral information of the matrix.
- Serves as a fundamental building block for asymptotic spectral analysis.

Stieltjes Transform

Definition:

$$m_\mu(z) = \int \frac{1}{t - z}\mu(dt), \quad z \in \mathbb{C} \setminus \mathbb{R}$$

- This transform is analytic on $\mathbb{C} \setminus \mathbb{R}$ and uniquely characterizes the measure μ.
- For matrices, it is written as:

$$m_{\mu_{\mathbf{M}}}(z) = \frac{1}{n}\mathrm{Tr}\left((\mathbf{M} - z\mathbf{I})^{-1}\right)$$

Key Properties:

- There exists an invertible relationship between the Stieltjes transform and the measure μ.
- The transform is stable under convergence, making it a robust tool for spectral analysis.

Inverse Stieltjes Transform

Density Recovery:

$$f(x) = \frac{1}{\pi} \lim_{y \to 0^+} \operatorname{Im}\left\{ m_\mu(x + iy) \right\}$$

- This relation recovers the spectral density from the boundary behavior of the Stieltjes transform.
- Similarly, the measure of an interval is given by:

$$\mu([a, b]) = \frac{1}{\pi} \lim_{y \to 0^+} \int_a^b \operatorname{Im}\left\{ m_\mu(x + iy) \right\} dx$$

Complex Integration:

$$\mathbb{E}[g(\lambda)] = -\frac{1}{2\pi i} \oint_\Gamma g(z) m_\mu(z) dz$$

- This contour integration formula is particularly useful for computing expectations of functions of eigenvalues.

Cauchy's Integral Formula in Spectral Analysis

Cauchy's Integral Formula:

$$\frac{1}{2\pi i}\oint_\Gamma \frac{f(z)}{z-z_0}dz = \begin{cases} f(z_0), & z_0 \text{ is enclosed by } \Gamma \\ 0, & \text{otherwise} \end{cases}$$

For Matrices:

$$f(\mathbf{M}) = \frac{1}{2\pi i}\oint_\Gamma \frac{f(z)}{zI-\mathbf{M}}dz = -\frac{1}{2\pi i}\oint_\Gamma f(z)\mathbf{Q}_\mathbf{M}(z)dz$$

$$\mathbb{E}_{\Lambda\sim\mu_\mathbf{M}}\left[f(\Lambda)\right] = -\frac{1}{2\pi i n}\oint_\Gamma f(z)\mathrm{Tr}(\mathbf{Q}_\mathbf{M}(z))dz = -\frac{1}{2\pi i}\oint_\Gamma f(z)m_{\mu_\mathbf{M}}(z)dz$$

- Here, Γ is a contour enclosing all the eigenvalues of $\mathbf{M}$.

Wigner Matrices & Semicircle Law

Definition:

- Consider a symmetric matrix $\mathbf{M} \in \mathbb{R}^{n\times n}$ with random entries.
- The entries satisfy: $\mathbb{E}[M_{ij}] = 0$, with variances

$$\mathbb{E}[M_{ij}^2] = \frac{\sigma^2}{n} \quad \text{for } i \neq j, \quad \mathbb{E}[M_{ii}^2] = \frac{2\sigma^2}{n}.$$

- **Ensembles:** This law applies to Wigner ensembles (such as the GOE, GUE for real and complex cases, respectively) where the entries have general variance σ^2.

Semicircle Density:

$$f_{\text{sc}}(x) = \frac{1}{2\pi\sigma^2}\sqrt{4\sigma^2 - x^2}\,\mathbf{1}_{[-2\sigma,2\sigma]}(x)$$

Universality:

- The semicircular distribution persists even for non-Gaussian entries (with the same variance σ^2), provided they have finite moments.

Marchenko-Pastur Law: Complete Form

General Case:

$$\mu_{\mathrm{MP}} = \max\left(1 - \tfrac{1}{c}, 0\right)\delta_0 + \frac{\sqrt{(b-x)(x-a)}}{2\pi c x}\,\mathbf{1}_{[a,b]}(x)$$

- Here, $a = \sigma^2(1-\sqrt{c})^2$ and $b = \sigma^2(1+\sqrt{c})^2$, and $c = p/n$ represents the limiting aspect ratio.
- **Ensembles:** The Marchenko-Pastur law applies to sample covariance (Wishart) ensembles, where the data matrix has independent entries with variance σ^2.

Phase Transitions:

- For $c < 1$: The spectrum is purely continuous on $[a, b]$.
- For $c > 1$: A point mass $\frac{c-1}{c}\delta_0$ appears alongside the continuous part.
- For $c = 1$: A square-root singularity is observed at 0.

Setup and Perturbation Expansion

- Consider a Hermitian matrix $A_0 \in \mathbb{C}^{n\times n}$ with real eigenvalues λ_k^0 and an orthonormal set of eigenvectors $\{v_k^0\}$.
- Introduce a small perturbation:

$$A(\epsilon) = A_0 + \epsilon\, B,$$

  where B is Hermitian and $\epsilon \ll 1$.
- The eigenproblem becomes:

$$A(\epsilon)v_k(\epsilon) = \lambda_k(\epsilon)v_k(\epsilon).$$

- Series expansions:

$$\lambda_k(\epsilon) = \lambda_k^0 + \epsilon\, \lambda_k^{(1)} + \epsilon^2\, \lambda_k^{(2)} + O(\epsilon^3)$$
$$v_k(\epsilon) = v_k^0 + \epsilon\, v_k^{(1)} + \epsilon^2\, v_k^{(2)} + O(\epsilon^3).$$

- Note: The nondegenerate case assumes all λ_k^0 are distinct, while in the degenerate case some eigenvalues have multiplicity greater than one.

First-Order Corrections (Nondegenerate)

- **Eigenvalue Correction:**

$$\lambda_k^{(1)} = {v_k^0}^T B \, v_k^0 = B_{kk}.$$

- **Eigenvector Correction:**

$$v_k^{(1)} = \sum_{j \neq k} \frac{B_{jk}}{\lambda_k^0 - \lambda_j^0} \, v_j^0.$$

- Remarks:
  - $v_k^{(1)}$ is orthogonal to v_k^0.
  - Its magnitude is controlled by the spectral gaps $\lambda_k^0 - \lambda_j^0$.

Second-Order Corrections (Nondegenerate)

- **Eigenvalue Correction:**

$$\lambda_k^{(2)} = \sum_{j \neq k} \frac{|B_{jk}|^2}{\lambda_k^0 - \lambda_j^0}.$$

- **Eigenvector Correction:**

$$v_k^{(2)} = \sum_{j \neq k} \sum_{m \neq k} \frac{B_{jm} B_{mk}}{(\lambda_k^0 - \lambda_j^0)(\lambda_k^0 - \lambda_m^0)} v_j^0 - \sum_{j \neq k} \frac{B_{jk} B_{kk}}{(\lambda_k^0 - \lambda_j^0)^2} v_j^0$$

- **LoRA (Low-Rank Adaptation):**
  - Fine-tunes a pre-trained model using a low-rank update.
  - Update is of the form

$$\mathbf{W}_{\mathrm{ft}} = \mathbf{W} + \Delta\mathbf{W}_{\mathrm{LoRA}}, \qquad \Delta\mathbf{W}_{\mathrm{LoRA}} = \mathbf{B}\mathbf{A}^\top, \quad \mathbf{B}, \mathbf{A} \in \mathbb{R}^{N \times r}, \quad r \ll N.$$

- **Full Fine-Tuning:**
  - Updates every entry of the weight matrix with a dense, small perturbation.
- **Goal:**
  - Use Random Matrix Theory (RMT) to analyze how these two methods affect the spectral structure of the weight matrix.

Intruder Dimensions and the Issue

- **Full Fine-Tuning:** Dense, small perturbations preserve the bulk MP spectrum.
- **LoRA:** The low-rank update $\Delta\mathbf{W}_{\mathrm{LoRA}}$ can introduce new singular values (*intruder dimensions*) outside the MP bulk.
- **Issue:** These intruder dimensions represent new directions that are **not** aligned with the pre-trained features, potentially leading to overfitting on the fine-tuning task and reduced generalization.

Intruder Dimensions and the Issue

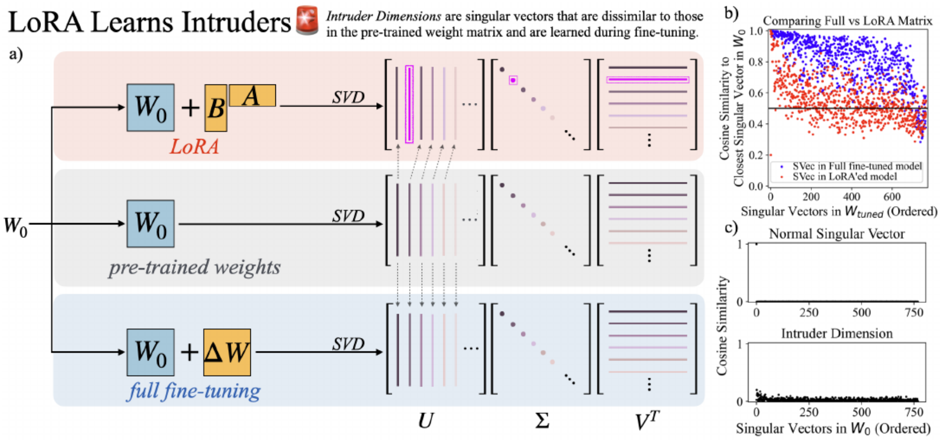


Figure: LoRA learns intruder dimentions

Pre-Trained Weight Matrix

- Model the pre-trained weight matrix as:

$$\mathbf{W}_{\text{pre}} \in \mathbb{R}^{N \times N} \quad \text{with } W_{ij} \sim \mathcal{N}\left(0, \frac{\sigma^2}{N}\right).$$

- In the large N limit, the singular values follow the **Marchenko-Pastur (MP) law**:

$$\rho_{\text{MP}}(\lambda) = \frac{1}{2\pi\sigma^2} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda}, \quad \lambda_\pm = \sigma^2(1 \pm \sqrt{c})^2,$$

  where c is the aspect ratio.

- **Intuition:** The MP distribution represents the typical spectral structure of the pre-trained model.

2. Intruder Dimensions: The Issue

- **Full Fine-Tuning:**
  - Dense perturbations preserve the MP bulk.
  - Singular vectors shift slightly but remain aligned with the original structure.

$$W_{\text{ft}} = W + \Delta W_{\text{ft}}, \qquad \tilde{u}_i = u_i + \epsilon \sum_{j \neq i} \frac{u_j^\top \Delta W_{\text{ft}} v_i}{\lambda_i - \lambda_j} u_j$$

- **LoRA:**
  - The low-rank update

$$\Delta \mathbf{W}_{\text{LoRA}} = \mathbf{B}\mathbf{A}^\top = \sum_{k=1}^{r} \gamma_k \, \mathbf{p}_k \, \mathbf{q}_k^\top$$

    concentrates energy in a few directions.
  - This can introduce **outlier singular values** outside the MP bulk.
- **Issue:**
  - Intruder dimensions are not aligned with the pre-trained model, potentially leading to overfitting and poor generalization.

Intruder Dimensions and the Issue

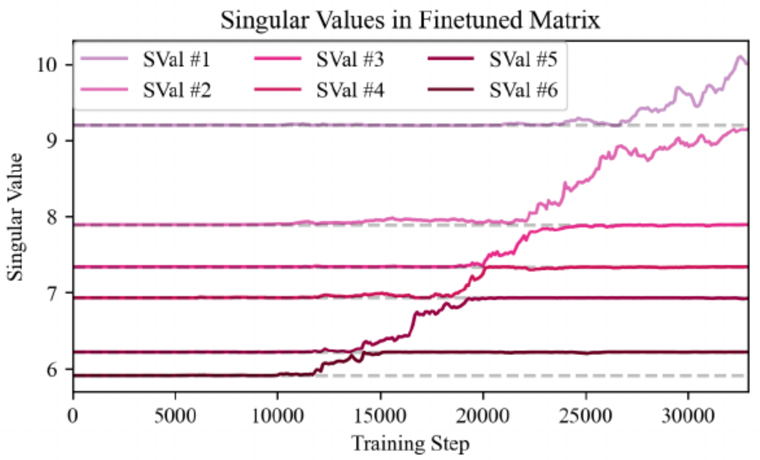


Figure: LoRA learns intruder dimentions

3. Low-Rank Updates and Resolvent Analysis

- Consider the perturbed matrix:

$$\mathbf{M} = \mathbf{W}_{\text{pre}} + \Delta\mathbf{W}_{\text{LoRA}},$$

where, for simplicity, we begin with a rank-1 update:

$$\Delta\mathbf{W} = \theta\, uv^\top, \quad \|u\| = \|v\| = 1.$$

- The resolvent (Green's function) is defined as:

$$G(z) = (z\mathbf{I} - \mathbf{W}_{\text{pre}})^{-1}.$$

- Using the **Sherman–Morrison formula**:

$$(z\mathbf{I} - \mathbf{W}_{\text{pre}} - \theta\, uv^\top)^{-1} = G(z) - \frac{\theta\, G(z)\, uv^\top\, G(z)}{1 + \theta\, v^\top G(z)u}.$$

- The pole of $G(z)$ corresponds to an eigenvalue of $\mathbf{M}$, determined by:

$$1 + \theta\, v^\top G(z)u = 0.$$

BBP Transition for a Rank-1 Update

- For large N, the quadratic form $v^\top G(z)u$ concentrates around the Stieltjes transform, when $v^\top u \approx 1$:

$$m(z) = \frac{1}{N}\mathrm{tr}\, G(z) = \int \frac{\rho_{\mathrm{MP}}(\lambda)\, d\lambda}{z-\lambda}.$$

- The condition for an outlier eigenvalue is:

$$1 + \theta\, m(\lambda_{\mathrm{out}}) = 0 \quad \Longrightarrow \quad \theta\, m(\lambda_{\mathrm{out}}) = -1.$$

- For λ_{out} outside the MP bulk (say, $\lambda_{\mathrm{out}} > \lambda_+$), the Stieltjes transform takes the form:

$$m(z) = \frac{z - \sqrt{z^2 - 4\sigma^2}}{2\sigma^2}.$$

- After some algebra, one obtains:

$$\lambda_{\mathrm{out}} \approx \theta + \frac{\sigma^2}{\theta} \Rightarrow \mu_\theta(dx) = \frac{\sqrt{4-x^2}}{2\pi(\theta^2+1-\theta x)}\mathbf{1}_{|x|<2}dx + 1_{|\theta|\geq\sqrt{\lambda_+}}(1-\frac{1}{\theta^2})\delta_{\theta+\frac{\sigma^2}{\theta}}(dx)$$

provided that $\theta > \sqrt{\lambda_+}$.

BBP Transition for Rank-r Updates

- For a rank-r update:

$$\Delta \mathbf{W}_{\text{LoRA}} = \sum_{k=1}^{r} \gamma_k \, \mathbf{p}_k \mathbf{q}_k^\top ,$$

  each spike γ_k leads to an outlier approximately if:

$$\gamma_k > \sqrt{\lambda_+}.$$

- The outlier singular values are approximately given by:

$$\lambda_{\text{out},k} \approx \gamma_k + \frac{\sigma^2}{\gamma_k}.$$

- **Corollary:** If the singular vectors $\mathbf{p}_k, \mathbf{q}_k$ are nearly orthogonal to the pre-trained singular vectors, these outliers represent *intruder dimensions*. which is especially important in high dimensions.

4. Mitigation: Increasing Rank

- **Idea:** Increase the rank r of the LoRA update.
- **Effect:**
  - Spreads the update energy over more directions.
  - Reduces the dominance of any single spike.
- **Result:** The overall spectral distortion is more distributed, and the outlier effects become less severe.
- **Shortcomings:** Full fine-tuning updates have a higher effective rank than LoRA updates, even when LoRA is performed with a full-rank matrix. For example, with the high rank of $r = 768$ for RoBERTa, LoRA updates have an average effective rank of 300. This suggests that LoRA is under utilizing its full capacity.

5. Mitigation: Rank Stabilization

- **Orthogonality Constraints:**

$$\mathbf{B}^\top \mathbf{B} = \mathbf{I}_r \quad \text{and} \quad \mathbf{A}^\top \mathbf{A} = \mathbf{I}_r.$$

  Finally represent the orthogonal LoRA updates as:

$$W_{\text{ft}} = W + \frac{\alpha}{r} BA^\top$$

  The normalization constant is added since $\|BA^\top\| \leq \|B\|\|A\| \leq C\sqrt{r} \cdot C\sqrt{r} = C'r$
- **Effect:** Suppresses the formation of dominant, misaligned intruder dimensions.
- **Outcome:**
  - The singular vectors of the LoRA update now exhibit higher cosine similarity with $\mathbf{W}_{\text{pre}}$.
  - The overall spectrum more closely resembles that of full fine-tuning.

6. Mitigation: Spectral Fine-Tuning

- **Idea:** Leverage the SVD of the pretrained weight matrix $\mathbf{W} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$ to guide fine-tuning.
- **Mechanisms:**
  - **Additive:**
$$\mathrm{Adapter}_A(\mathbf{W}) = [\mathbf{U}_1 + \mathbf{A}_U\,\mathbf{U}_2]\,\mathbf{S}\,[\mathbf{V}_1 + \mathbf{A}_V\,\mathbf{V}_2]^\top.$$
  - **Rotational:**
$$\mathrm{Adapter}_R(\mathbf{W}) = [\mathbf{U}_1\,\mathbf{R}_U\,\mathbf{U}_2]\,\mathbf{S}\,[\mathbf{V}_1\,\mathbf{R}_V\,\mathbf{V}_2]^\top.$$
- **Benefit:** Aligns fine-tuning with the pretrained spectrum, suppressing outlier (intruder) dimensions and preserving generalization.

Summary

- **LoRA** employs a low-rank update:

$$\Delta \mathbf{W}_{\text{LoRA}} = \mathbf{B}\mathbf{A}^\top,$$

  which is efficient but can introduce spectral outliers (intruder dimensions) via the BBP transition.
- **Full Fine-Tuning** uses dense, small perturbations that preserve the pre-trained MP bulk.
- **BBP Transition:**
  - A rank-r update with singular values γ_k creates outlier singular values at:

$$\lambda_{\text{out},k} \approx \gamma_k + \frac{\sigma^2}{\gamma_k},$$

    if $\gamma_k > \sqrt{\lambda_+}$.
- **Mitigation:**
  - Increasing the update rank or applying rank stabilization (orthogonality constraints) can reduce the adverse impact of intruder dimensions.

Conclusion

Key Takeaways

- **RMT & Perturbation Theory provide a rigorous framework** to understand how low-rank updates (LoRA) affect the spectral properties of pre-trained models.
- The emergence of **intruder dimensions** via the BBP transition explains differences between LoRA and full fine-tuning.
- By **increasing the rank** or enforcing **rank stabilization**, one can mitigate these effects, aligning LoRA's behavior closer to that of full fine-tuning..

Future Work

- Using this theoretical understanding, **novel frameworks can be introduced** to mitigate this increasingly important issue
- Utilize higher-order perturbations for the Non-Asymptotic case.

Questions?

Any Questions?

References

- Baik, J., Ben Arous, G., & Péché, S. (2005). Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices.
- Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis.
- Nathan Noiry (2020) Spectral Measures of Spiked Random Matrices
- Reece Shuttleworth and Jacob Andreas and Antonio Torralba and Pratyusha Sharma, (2024), LoRA vs Full Fine-tuning: An Illusion of Equivalence
- 1. Couillet R, Liao Z. Random Matrix Methods for Machine Learning. Cambridge University Press; 2022.
- Livan, G., Novaes, M., Vivo, P. (2017). Introduction to Random Matrices - Theory and Practice. ArXiv. https://doi.org/10.1007/978-3-319-70885-0
- Recent work on LoRA and parameter-efficient fine-tuning.