Understanding Incremental Learning in Overparamerterized Matrix Factorization

Hancheng Min[1], René Vidal[2]

[1] INS, Shanghai Jiao Tong University
[2] IDEAS, University of Pennsylvania

Dec. 10, 2025

Overparametrization is critical in DL: Double Descent

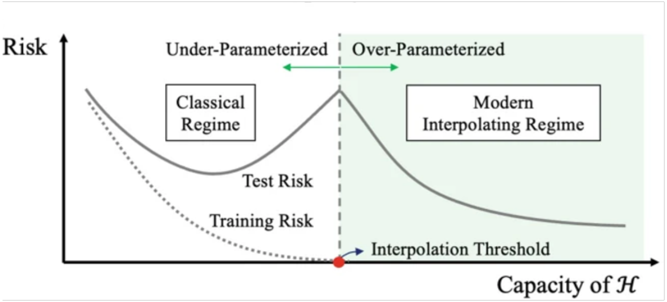


Figure 1: The "Double Descent" phenomenon: Generalization performance of ML models (Loss/risk/cost during test time) increases as the model capacity/complexity increases once beyond the interpolating threshold (Overparametrized regime)

Classic vs. DL views on overparametrization

Classic view:

- A problem is overparametrized if underdetermined
- Explicit regularization for finding simple solutions (Occam's razor)

Classic vs. DL views on overparametrization

Classic view:

- A problem is overparametrized if underdetermined
- Explicit regularization for finding simple solutions (Occam's razor)

DL view:

- A problem is overparametrized if underdetermined, and the model class can be parametrized by many more parameters than needed
- Implicit regularization induced by model parametrization when training with gradient-based algorithms under proper initialization

Example: Matrix Sensing under linear operator $\mathcal{A}$

Classic vs. DL views on overparametrization

Classic view:

- A problem is overparametrized if underdetermined
- Explicit regularization for finding simple solutions (Occam's razor)

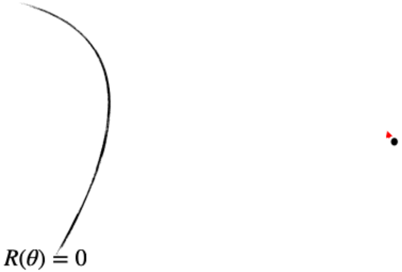
DL view:

- A problem is overparametrized if underdetermined, and the model class can be parametrized by many more parameters than needed
- Implicit regularization induced by model parametrization when training with gradient-based algorithms under proper initialization

Example: Matrix Sensing under linear operator $\mathcal{A}$

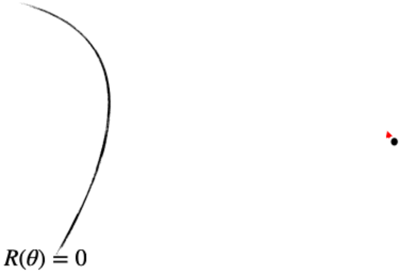
(Explicit reg.) $\min_{W \in \mathbb{R}^{d \times d}} \|y - \mathcal{A}(W)\|^2 + \gamma \|W\|_*$ (Implicit Reg.) $\min_{\substack{W_i \in \mathbb{R}^{d_i \times d_{i+1}} \\ d_1 = d, d_{L+1} = d}} \|y - \mathcal{A}(\underbrace{W_1 W_2 \cdots W_L}_{:= W})\|^2$

Implicit biases of training dynamics



- Network parameters (weights) θ updated through some optimization algorithm to minimize some loss/risk/cost function $\mathcal{L}(\theta)$

Implicit biases of training dynamics



- Network parameters (weights) θ updated through some optimization algorithm to minimize some loss/risk/cost function $\mathcal{L}(\theta)$
- **Implicit Bias**: depending on the choice of initialization scale, step size, gradient stochasticity, etc., one obtains different θ^*

Incremental learning phenomenon

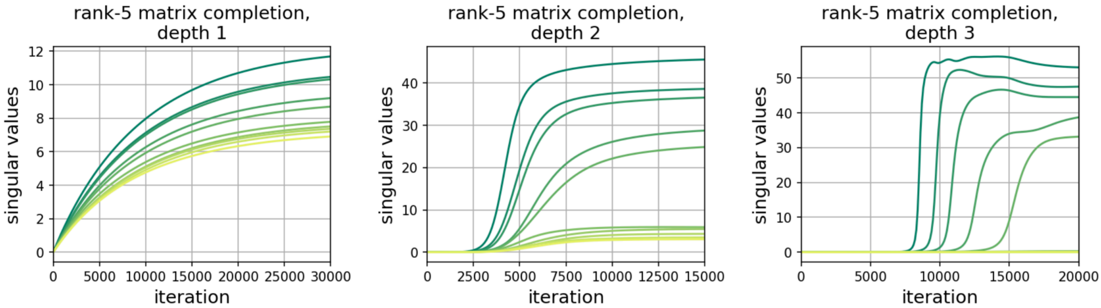


Figure 2: Deep matrix factorization exhibits the **incremental learning** phenomenon.

GD on $\|y - \mathcal{A}(W_1 W_2 \cdots W_L)\|^2$ with large depth L, starting from a small initialization:

Incremental learning phenomenon

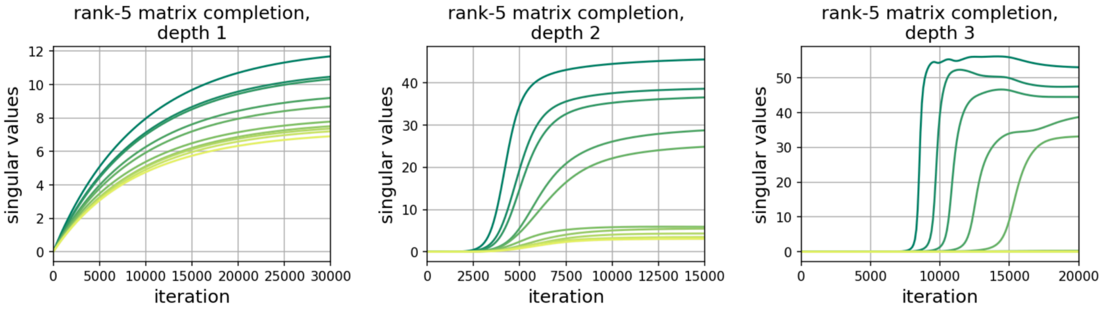


Figure 2: Deep matrix factorization exhibits the **incremental learning** phenomenon.

GD on $\|y - \mathcal{A}(W_1 W_2 \cdots W_L)\|^2$ with large depth L, starting from a small initialization:

1. The singular values of the target/ground-truth matrix are learned sequentially;

Incremental learning phenomenon

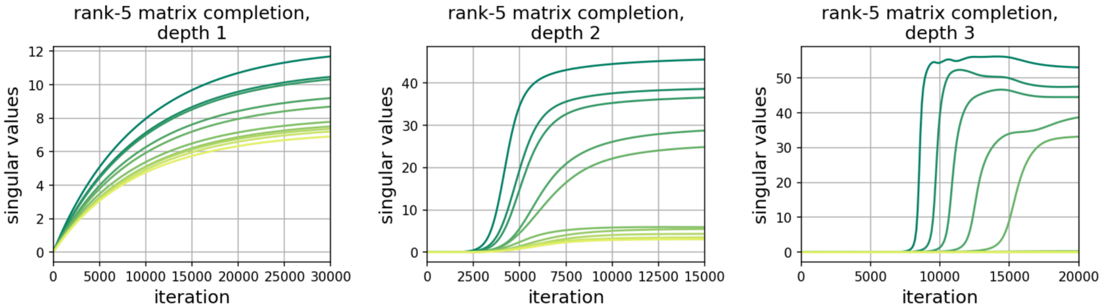


Figure 2: Deep matrix factorization exhibits the **incremental learning** phenomenon.

GD on $\|y - \mathcal{A}(W_1 W_2 \cdots W_L)\|^2$ with large depth L, starting from a small initialization:

1. The singular values of the target/ground-truth matrix are learned sequentially;
2. Large singular values are learned first.

Related works

- Incremental learning in matrix factorization: Initially studied by Saxe et al. [2014]; More indepth analyses by Arora et al. [2019], Gunasekar et al. [2017].

Related works

- Incremental learning in matrix factorization: Initially studied by Saxe et al. [2014]; More indepth analyses by Arora et al. [2019], Gunasekar et al. [2017].
- Manifested in other learning problems:
  - Spectral bias/frequency priciple in deep learning [Rahaman et al., 2019, Xu et al., 2019]: Low-frequency components of the target function are learned first;

Related works

- Incremental learning in matrix factorization: Initially studied by Saxe et al. [2014]; More indepth analyses by Arora et al. [2019], Gunasekar et al. [2017].
- Manifested in other learning problems:
  - Spectral bias/frequency priciple in deep learning [Rahaman et al., 2019, Xu et al., 2019]: Low-frequency components of the target function are learned first;
  - Incremental learning when learning a state-space model, or linear Recurrent Neural Networks(RNNs) [Proca et al., 2025]: Sequential learning singular values of input-output correlation matrix;

Related works

- Incremental learning in matrix factorization: Initially studied by Saxe et al. [2014]; More indepth analyses by Arora et al. [2019], Gunasekar et al. [2017].
- Manifested in other learning problems:
  - Spectral bias/frequency priciple in deep learning [Rahaman et al., 2019, Xu et al., 2019]: Low-frequency components of the target function are learned first;
  - Incremental learning when learning a state-space model, or linear Recurrent Neural Networks(RNNs) [Proca et al., 2025]: Sequential learning singular values of input-output correlation matrix;
  - Incremental learning when training a transformer [Abbe et al., 2023]: Sequential learning of tasks from low to high complexity.

Related works

- Incremental learning in matrix factorization: Initially studied by Saxe et al. [2014]; More indepth analyses by Arora et al. [2019], Gunasekar et al. [2017].
- Precise characterization of incremental learning in matrix factorization problems is limited to the two-layer problems (Symmetric, Asymmertic):
  - Gradient flow (analyzing closed-form solutions):
    Spectral initialization [Gidel et al., 2019, Tarmoun et al., 2021]; General initialization (**Our work**)
  - Gradient descent (analyzing iterates):
    Spectral initialization [Gidel et al., 2019]; Random initialization [Jiang et al., 2023, Jin et al., 2023]

Problem Settings

- Loss $\mathcal{L}(U) = \frac{1}{4}\|Y - UU^\top\|_F^2$, $Y = Y^\top \succeq 0$, $U \in \mathbb{R}^{n\times r}$ (This talk: $r \geq n$)
- Gradient Flow (GF) on U:

$$\dot{U} = -\nabla_U \mathcal{L}(U) = (Y - UU^\top)U, \; U(0) = U_0 \tag{1}$$

Problem Settings

- Loss $\mathcal{L}(U) = \frac{1}{4}\|Y - UU^\top\|_F^2$, $Y = Y^\top \succeq 0$, $U \in \mathbb{R}^{n\times r}$ (This talk: $r \geq n$)
- Gradient Flow (GF) on U:

$$\dot{U} = -\nabla_U \mathcal{L}(U) = (Y - UU^\top)U, \; U(0) = U_0 \tag{1}$$

- Induced dynamics on $W = UU^\top$:

$$\dot{W} = \dot{U}U^\top + U^\top \dot{U} = (Y - W)W + W(Y - W), \; W(0) = U_0 U_0^\top := W_0 \tag{2}$$

Problem Settings

- Loss $\mathcal{L}(U) = \frac{1}{4}\|Y - UU^\top\|_F^2$, $Y = Y^\top \succeq 0$, $U \in \mathbb{R}^{n\times r}$ (This talk: $r \geq n$)
- Gradient Flow (GF) on U:

$$\dot{U} = -\nabla_U \mathcal{L}(U) = (Y - UU^\top)U,\ U(0) = \alpha^{1/2}U_0 \tag{3}$$

- Induced dynamics on $W = UU^\top$:

$$\dot{W} = \dot{U}U^\top + U^\top\dot{U} = (Y - W)W + W(Y - W),\ W(0) = U_0U_0^\top := \alpha W_0 \tag{4}$$

- Split the initial condition into **Initialization scale** α and **Initialization shape** $U_0(W_0)$: We are interested in *how incremental learning phenomenon emerges as the initialization scale decreases*.

Close-form solution

For the induced dynamics on W (A matrix Riccati differential equation):

$$\dot{W} = (Y - W)W + W(Y - W), \; W(0) = \alpha W_0 \,. \tag{5}$$

Proposition

If $\operatorname{rank}(Y) = k$ *and the full SVD of* Y *is* $\Phi \begin{bmatrix} \Sigma_Y & 0 \\ 0 & 0 \end{bmatrix} \Phi^\top$*, then* (5) *has a unique solution:*

$$W(t) = \Phi S(t) \alpha \tilde{W}_0 \left(I_n + \alpha G(t) \tilde{W}_0 \right)^{-1} S^\top(t) \Phi^\top, \tag{6}$$

where $\tilde{W}_0 = \Phi^\top W_0 \Phi$ *and* $G(t) = \begin{bmatrix} \Sigma_Y^{-1}(e^{2\Sigma_Y t} - I_K) & 0 \\ 0 & 2I_{n-K}t \end{bmatrix}, S(t) = \begin{bmatrix} e^{\Sigma_Y t} & 0 \\ 0 & I_{n-K} \end{bmatrix}.$

Solution under spectral initialization

Definition

W_0 is a *spectral initialization* if W_0 and Y are codiagonalizable, i.e., $\tilde{W}_0 = \Phi^\top W_0 \Phi$ is diagonal

Corollary

Let $\Sigma_Y = \mathrm{diag}\{\sigma_{i,Y}\}_{i=1}^K$. *If* $U_0 = \Phi \Sigma_{U_0} V_{U_0}^\top$ *renders* $W_0 = U_0 U_0^\top$ *a spectral initialization, then the solution to* (5) *has the form* $W(t) = \Phi \mathrm{diag}\{\sigma_{i,W}(t)\}_{i=1}^n \Phi^\top$ *with*

$$\sigma_{i,W}(t) = \frac{\alpha \sigma_{i,Y} \sigma_{i,0} e^{2\sigma_{i,Y} t}}{\sigma_{i,Y} + \alpha \sigma_{i,0} (e^{2\sigma_{i,Y} t} - 1)}, \text{ if } i \leq K; \qquad \sigma_{i,W}(t) = \frac{\alpha \sigma_{i,0}}{1 + 2\alpha \sigma_{i,0} t}, \text{ if } i > K, \tag{7}$$

where $\sigma_{i,0} = [\tilde{W}_0]_{ii} = [\Sigma_{U_0}^2]_{ii} \geq 0, \forall i$.

Dynamic modes $\sigma_{i,W}(t)$ are decoupled, each learns one singular value of Y.

Learning singular value of Y under small initialization scale

- $\forall \varepsilon > 0, \exists C_\varepsilon > c_\varepsilon > 0$, such that for sufficiently small α (details later)

$$\sigma_{i,W}(t) \leq \varepsilon, \quad \forall t \leq \frac{1}{2\sigma_{i,Y}} \log \frac{c_\varepsilon}{\alpha}$$
$$\sigma_{i,W}(t) \geq \sigma_{i,Y} - \varepsilon, \quad \forall t \geq \frac{1}{2\sigma_{i,Y}} \log \frac{C_\varepsilon}{\alpha}$$

- $\sigma_{i,W}(t)$ remains small until $\Theta(\frac{1}{\sigma_{i,Y}} \log \frac{1}{\alpha})$ time, followed by an sharp transition phase of learning $\sigma_{i,Y}$

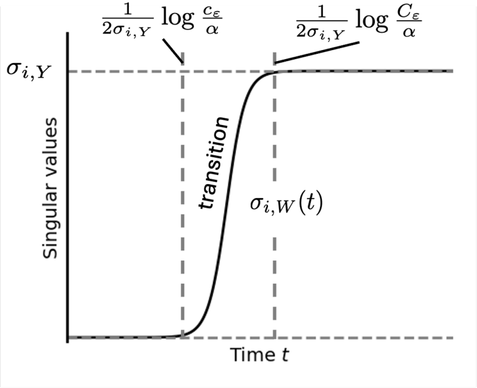


Figure 3: Learning curve $\sigma_{i,W}(t)$ for singular value $\sigma_{i,Y}$

Incremental learning under small initialization scale

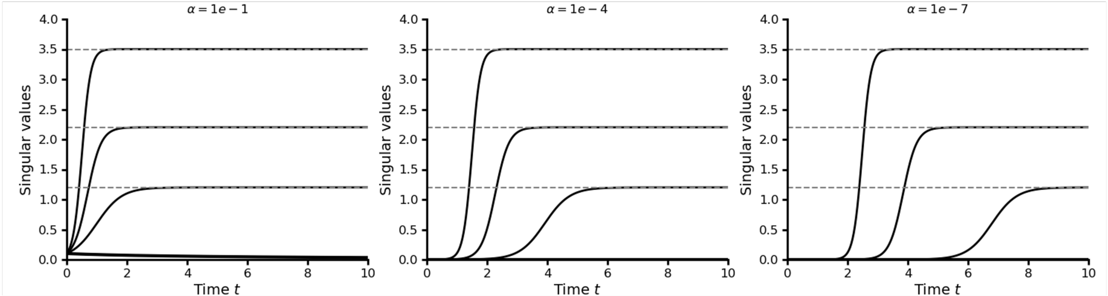


Figure 4: Incremental learning emerges as initialization scale decreases

- When init. scale decreases $\alpha \to e^{-M}\alpha$, transition phase for $\sigma_{i,Y}$ is delayed by $\frac{M}{\sigma_{i,Y}}$.
- (**Incremental learning**) For sufficiently small α:

  1) (Sequential learning) Transition phases for different $\sigma_{i,Y}$ become non-overlapping;

  2) (Low-rank approximations) Those for larger singular values happen earlier.

Main result under general small initialization

Theorem (Incremental learning under general small initialization)

Suppose the target Y has K distinct non-zero singular values and:

- *The initialization* $\tilde{W}_0 = \Phi^\top \bar{U}_0 \bar{U}_0^\top \Phi$ *has an inverse V; let* $M := \max\{\|V\|, \|V^{-1}\|\}$*;*

Main result under general small initialization

Theorem (Incremental learning under general small initialization)

Suppose the target Y has K distinct non-zero singular values and:

- *The initialization $\tilde{W}_0 = \Phi^\top \bar{U}_0 \bar{U}_0^\top \Phi$ has an inverse V; let $M := \max\{\|V\|, \|V^{-1}\|\}$;*
- *Given some $0 < \varepsilon \leq \min\{\sigma_{K,Y}, 1\}$, let $c_\varepsilon = \frac{\varepsilon}{16M^2}$, $C_\varepsilon = \frac{16\sigma_{1,Y}^2 M^2}{\varepsilon}$;*

Main result under general small initialization

Theorem (Incremental learning under general small initialization)

Suppose the target Y has K distinct non-zero singular values and:

- *The initialization $\tilde{W}_0 = \Phi^\top \bar{U}_0 \bar{U}_0^\top \Phi$ has an inverse V; let $M := \max\{\|V\|, \|V^{-1}\|\}$;*
- *Given some $0 < \varepsilon \leq \min\{\sigma_{K,Y}, 1\}$, let $c_\varepsilon = \frac{\varepsilon}{16M^2}$, $C_\varepsilon = \frac{16\sigma_{1,Y}^2 M^2}{\varepsilon}$;*
- *The init. scale α is sufficiently small so that $\alpha \leq \frac{c_\varepsilon}{M}$ and $\mathcal{I}_k := \left[\frac{1}{2\sigma_{k,Y}} \log \frac{C_\varepsilon}{\alpha}, \frac{1}{2\sigma_{k+1,Y}} \log \frac{c_\varepsilon}{\alpha}\right]$ are non-empty;*

Main result under general small initialization

Theorem (Incremental learning under general small initialization)

Suppose the target Y has K distinct non-zero singular values and:

- *The initialization $\tilde{W}_0 = \Phi^\top \bar{U}_0 \bar{U}_0^\top \Phi$ has an inverse V; let $M := \max\{\|V\|, \|V^{-1}\|\}$;*
- *Given some $0 < \varepsilon \leq \min\{\sigma_{K,Y}, 1\}$, let $c_\varepsilon = \frac{\varepsilon}{16M^2}$, $C_\varepsilon = \frac{16\sigma_{1,Y}^2 M^2}{\varepsilon}$;*
- *The init. scale α is sufficiently small so that $\alpha \leq \frac{c_\varepsilon}{M}$ and $\mathcal{I}_k := \left[\frac{1}{2\sigma_{k,Y}} \log \frac{C_\varepsilon}{\alpha}, \frac{1}{2\sigma_{k+1,Y}} \log \frac{c_\varepsilon}{\alpha}\right]$ are non-empty;*

then the solution $W(t)$ to (5) *satisfies that $\forall 1 \leq k \leq K$,*

$$\|W(t) - \hat{Y}_k\| \leq \varepsilon, \quad \forall t \in \mathcal{I}_k, \tag{8}$$

where $\hat{Y}_k := \arg\min_{\mathrm{rank}(Z)=k} \|Y - Z\|_F$ is the best rank-k approximation of Y.

Conclusion and future work

To summarize, we studied incremental learning in matrix factorization with closed-form solutions. Future work:

1. Removing the assumption that $\tilde{W}_0$ is invertible.
2. Extension to asymmetric factorization with the symmetrization trick [Burer and Monteiro, 2005].
3. From identity operator $\mathcal{A} = \mathrm{Id}$ to those with Ristricted Isometry Property.

Conclusion and future work

To summarize, we studied incremental learning in matrix factorization with closed-form solutions. Future work:

1. Removing the assumption that $\tilde{W}_0$ is invertible.
2. Extension to asymmetric factorization with the symmetrization trick [Burer and Monteiro, 2005].
3. From identity operator $\mathcal{A} = \mathrm{Id}$ to those with Ristricted Isometry Property.

Thank you!

References

E. Abbe, S. Bengio, E. Boix-Adsera, E. Littwin, and J. Susskind. Transformers learn through gradual rank increase. NeurIPS, 36, 2023.

S. Arora, N. Cohen, W. Hu, and Y. Luo. Implicit regularization in deep matrix factorization. NeurIPS, 2019.

S. Burer and R. D. C. Monteiro. Local minima and convergence in low-rank semidefinite programming. Math. Program., 103(3):427–444, July 2005.

G. Gidel, F. Bach, and S. Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In NeurIPS, 2019.

S. Gunasekar, B. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. In NeurIPS, 2017.

L. Jiang, Y. Chen, and L. Ding. Algorithmic regularization in model-free overparametrized asymmetric matrix factorization. SIAM Journal on Mathematics of Data Science, 5(3): 723–744, 2023.

J. Jin, Z. Li, K. Lyu, S. S. Du, and J. D. Lee. Understanding incremental learning of gradient descent: A fine-grained analysis of matrix sensing. In ICML, 2023.