Implicit bias of SGD in L_2-regularized linear DNNs: One-way jumps from high to low rank

Zihan Wang, Arthur Jacot

Courant Institute of Mathematical Sciences
New York University

Problem setting

- We study a matrix completion problem with true matrix A^* using deep linear networks (DLNs): $A_\theta = W_L \cdots W_1$.

Problem setting

- We study a matrix completion problem with true matrix A^* using deep linear networks (DLNs): $A_\theta = W_L \cdots W_1$.
- The optimization objective is the matrix completion (MC) loss with L_2 regularization: $\frac{1}{2N}\sum_{(i,j)\in I}(A^*_{ij} - A_{ij})^2 + \lambda\|\theta\|^2$, where I is the set of all observed entries of A^* and $N = |I|$.

Problem setting

- We study a matrix completion problem with true matrix A^* using deep linear networks (DLNs): $A_\theta = W_L \cdots W_1$.
- The optimization objective is the matrix completion (MC) loss with L_2 regularization: $\frac{1}{2N}\sum_{(i,j)\in I}(A^*_{ij} - A_{ij})^2 + \lambda\|\theta\|^2$, where I is the set of all observed entries of A^* and $N = |I|$.
- We examine the implicit bias of SGD with finite learning rate:

$$\theta_{t+1} = (1 - 2\eta\lambda)\theta_t - \frac{\eta}{2}\nabla_\theta \left(A^*_{i_t j_t} - A_{\theta_t, i_t j_t}\right)^2 .$$

Low rank region

1. Approximately balanced: for all ℓ, $\|W_\ell^T W_\ell - W_{\ell-1} W_{\ell-1}^T\|_F^2 \leq \epsilon_1$.

Low rank region

1. Approximately balanced: for all ℓ, $\|W_\ell^T W_\ell - W_{\ell-1} W_{\ell-1}^T\|_F^2 \leq \epsilon_1$.
2. Approximately rank r: for all ℓ, $\sum_{i=1}^{\mathrm{Rank} W_\ell} f_\alpha(s_i(W_\ell^\top W_\ell)) \leq r + \epsilon_2$ where $s_i(A)$ is the i-th singular value of A and $f_\alpha(x)$ is a concave and increasing function such that $f_\alpha(0) = 0$ and $f_\alpha(x) = 1$ for $x > \alpha$ with some mild assumptions.

Low rank region

1. Approximately balanced: for all ℓ, $\|W_\ell^T W_\ell - W_{\ell-1} W_{\ell-1}^T\|_F^2 \leq \epsilon_1$.
2. Approximately rank r: for all ℓ, $\sum_{i=1}^{\mathrm{Rank} W_\ell} f_\alpha(s_i(W_\ell^\top W_\ell)) \leq r + \epsilon_2$ where $s_i(A)$ is the i-th singular value of A and $f_\alpha(x)$ is a concave and increasing function such that $f_\alpha(0) = 0$ and $f_\alpha(x) = 1$ for $x > \alpha$ with some mild assumptions.
3. Bounded: for all ℓ, $\|W_\ell\|_F^2 \leq C$.

Low rank region

1. Approximately balanced: for all ℓ, $\|W_\ell^T W_\ell - W_{\ell-1} W_{\ell-1}^T\|_F^2 \leq \epsilon_1$.
2. Approximately rank r: for all ℓ, $\sum_{i=1}^{\mathrm{Rank} W_\ell} f_\alpha(s_i(W_\ell^\top W_\ell)) \leq r + \epsilon_2$ where $s_i(A)$ is the i-th singular value of A and $f_\alpha(x)$ is a concave and increasing function such that $f_\alpha(0) = 0$ and $f_\alpha(x) = 1$ for $x > \alpha$ with some mild assumptions.
3. Bounded: for all ℓ, $\|W_\ell\|_F^2 \leq C$.

- We denote the region in parameter space satisfying the conditions above by B_r.

Main results

Theorem (Informal)

For any initialization θ_0 and any $r \geq 0$, there exists T such that

$$\mathbb{P}(\theta_t \in B_r, \forall t > T | \theta_0) = 1.$$

Main results

Theorem (Informal)

For any initialization θ_0 *and any* $r \geq 0$, *there exists* T *such that*

$$\mathbb{P}(\theta_t \in B_r, \forall t > T | \theta_0) = 1.$$

Lemma

For any critical point $\hat{\theta}$ *in* B_r, *we have* $\sum_{i=1}^{\mathrm{Rank} A_{\hat{\theta}}} f_\alpha(s_i(A_{\hat{\theta}})^{2/L}) \leq r + \epsilon_2$.

Proof sketch

1. For SGD, the set B_r is closed:

$$\theta_t \in B_r \Rightarrow \theta_{t+1} \in B_r.$$

Proof sketch

1. For SGD, the set B_r is closed:

$$\theta_t \in B_r \Rightarrow \theta_{t+1} \in B_r.$$

2. For any parameter θ_t, there exists a time T such that

$$\mathbb{P}\left(\theta_{t+T} \in B_r | \theta_t\right) \geq O(r^T).$$

Experiments

- We observe 3 out of 4 entries of a 2×2 matrix: $\begin{pmatrix} 1 & * \\ \epsilon & 1 \end{pmatrix}$. The ground truth is the rank-1 matrix where the missing entry * is ϵ^{-1}.

- We observe 3 out of 4 entries of a 2×2 matrix: $\begin{pmatrix} 1 & * \\ \epsilon & 1 \end{pmatrix}$. The ground truth is the rank-1 matrix where the missing entry $*$ is ϵ^{-1}.

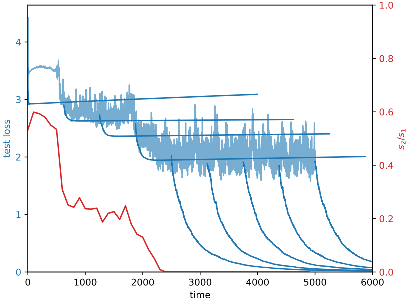

Figure: **Red:** ratio of second to first singular value of A_θ. **Light blue:** test error. **Dark blue:** test error after offshoots at different time with smaller η and λ.

Conclusion

- We have shown that SGD in DLNs has a non-zero probability of jumping from any higher rank region to a lower rank one, but the inverse is impossible.

Conclusion

- We have shown that SGD in DLNs has a non-zero probability of jumping from any higher rank region to a lower rank one, but the inverse is impossible.
- Our analysis does not rely on continuous approximation of SGD and the absorbing phenomenon cannot be recovered with a continuous approximation.

Conclusion

- We have shown that SGD in DLNs has a non-zero probability of jumping from any higher rank region to a lower rank one, but the inverse is impossible.
- Our analysis does not rely on continuous approximation of SGD and the absorbing phenomenon cannot be recovered with a continuous approximation.

Thank you!