Implicit Regularization in Deep Matrix Factorization

SeokHoon Park

June 29, 2021

Seoul National University

Table of Contents

Table of Contents

Deep matrix factorization

- Def of Deep matrix factorization

  Since standard matrix factorization can be viewed as a two-layer neural network, a natural extension is to consider deeper models. A deep matrix factorization of W $\in \mathrm{R}^{d \times d'}$, with hidden dimensions $d_1, ..., d_{N-1} \in \mathbb{N}$ is the parameterization:

$$W = W_N W_{N-1} \cdots W_1$$

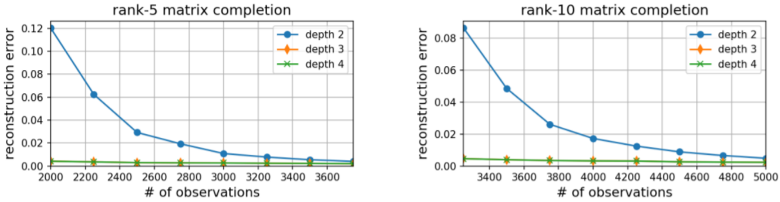
  where $W_j \in \mathbb{R}^{d_j \times d_{j-1}} j = 1, ..., N$ with $d_N = d, d_0 = d'$.

Conjecture 1 from former papar

With small enough learning rate and initialization close enough to the origin, gradient descent on a full-dimensional matrix factorization converges to the minimum nuclear norm solution.

Table of Contents

Can the implicit regularization be captured by norms?



Can the implicit regularization be captured by norms?

- Hypothesis:
  gradient descent on a depth-N matrix factorization implicitly minimizes some norm that approximates rank, with the approximation being more accurate the larger N is.

- Schatten-p quasi-norm $||W||_{S_p}^p = \sum_{r=1}^{min(d,d')} \sigma_r^p(W)$
  where $\sigma_i(W)$: singular value of W.

Current theory does not distinguish depth-N from depth-2

- Implicit regularization and matrix sensing
  former paper studied implicit regularization in shallow matrix factorization by considering recovery of a positive semidefinite matrix from sensing via symmetric measurements.

  $\min_{W \in S_+^d} l(W) = \min_{W \in S_+^d} \frac{1}{2} \sum_{i=1}^{m} (y_i - < A_i, W >)^2 ....(2)$

  where $A_1, ..., A_m \in \mathbb{R}^{d,d}$ are symmtric and linearly independent

Thm 1

- Thm 1:
  Assume the measurement matrices $A_1, ..., A_m$ commute. Then, if $\bar{W}_{\text{sha}} := \lim_{\alpha \to 0} W_{\text{sha},\infty}(\alpha)$ exists and is a global optimum for (2) with $l(\bar{W}_{\text{sha}}) = 0$, it holds that $\bar{W}_{\text{sha}} \in \text{argmin}_{W \in S_+^d, l(W)=0} ||W||_*$ i.e $\bar{W}_{\text{sha}}$ is a global optimum with minimal nuclear norm.

- $W_{sha,}(\alpha)$: The final solution $W = ZZ^T$ obtained from running gradient flow on $l(ZZ^T)$ with initialization αI

Extend to depth-N

$\min_{W \in S_+^d} l(W)$
$= \min_{W \in S_+^d} \frac{1}{2} \sum_{i=1}^{m} (y_i - < A_i, W_N W_{N-1} ... W_1 >)^2 ....(3)$

- Thm 2
  Suppose $N \geq 3$, and that the matrices $A_1, ..., A_m$ commute. Then if $\bar{W}_{\text{deep}} := \lim_{\alpha \to 0} W_{\text{deep},\infty}(\alpha)$ exists and is a global optimum for (3) with $l(\bar{W}_{\text{deep}}) = 0$, it holds that $\bar{W}_{\text{deep}} \in \text{argmin}_{W \in S_+^d, l(W)=0} ||W||_*$ i.e $\bar{W}_{\text{deep}}$ is a global optimum with minimal nuclear norm.

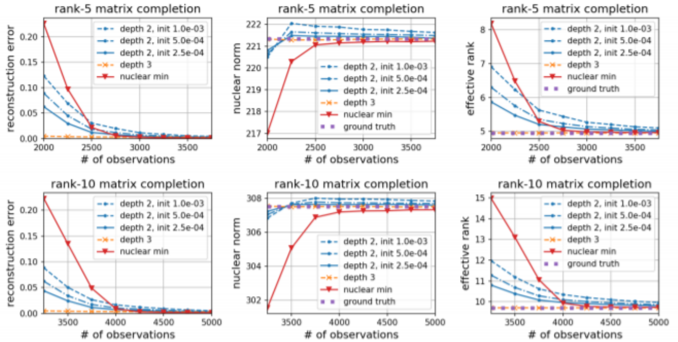
Cannot explain implicit regularization with Schatten quasi-norm

- Proposition 1

  For any dimension $d \geq 3$, there exist linearly independent symmetric and commutable measurement matrices $A_1, ..., A_m \in \mathbb{R}^{d,d}$, and corresponding labels $y_1, .., y_m \in \mathbb{R}$, such that the limit solution defined in Thm2 which has been shown to satisfy $\bar{W}_{\text{deep}} \in \text{argmin}_{W \in S_+^d, l(W)=0} ||W||_*$ is not a local minimum of the following program for any $0 < p < 1$

  $\min_{W \in S_+^d, l(W)=0} ||W||_{S_p}$

Experiment



rank-5 matrix completion
reconstruction error
nuclear norm
effective rank
rank-10 matrix completion
# of observations
depth 2, init 1.0e-03
depth 2, init 5.0e-04
depth 2, init 2.5e-04
depth 3
nuclear min
ground truth

Experiment

- Compare minimum nuclear norm solution to those brought forth by running gradient descent on matrix factorization of different depths.

- When there are less entries observed, neither shallow nor deep factorization minimize nuclear norm.

- $erank(A) = exp(H(p_1, ..., p_Q))$

  where $\sigma_1, ..., \sigma_Q$: A의 singular values , $p_k = \frac{\sigma_k}{||\sigma||_1}$
  $H(p_1, ..., p_Q) = -\sum_{k=1}^{Q} p_k \log p_k$

Hypothesis

- Capturing implicit regularization in matrix factorization through a single mathematical norm may not be posiible.

Table of Contents

Dynamical analysis

- We derive differential equations governing the dynamics of singular values and singular vectors for the product matrix W.
- Evolution rates of singular values turn out to be proportional to their size exponentiated by 2-2/N , where N is the depth of the factorization.
- We explain how our findings imply a tendency towards low-rank solutions, which intensifies with depth.

Dynamical analysis

- $\phi(W_1, ..., X_N) = l(W_N...W_1)$, where l : general analytic loss

- gradient flow over factorization:

  $\dot{W}_j(t) := \frac{d}{dt}W_j(t) = -\frac{d}{dW_j}\phi(W_1(t), ..., W_N(t)),$

  $j = 1, ..., N, t \geq 0$

Dynamical analysis

- Lemma:

  The product matrix $W(t)$ can be expressed as:
  $W(t) = U(t)S(t)V^T(t)$

  where $U(t) \in \mathbb{R}^{d,min(d,d')}$, $S(t) \in \mathbb{R}^{min(d,d'),min(d,d')}$, $V(t) \in \mathbb{R}^{d',min(d,d')}$ are analytic functions of t

Dynamical analysis

- The diagonal elements of S(t) , which we denote by $\sigma_1(t), ..., \sigma_{min(d,d')}(t)$ are signed singular values of W(t)
- The columns of U(t) and V(t), denoted $u_1(t), ..., u_{min(d,d')}(t)$ and $v_1, ..., v_{min(d,d')}(t)$ are the corresponding left and right singular vectors

Dynamical analysis

- Thm3

  The signed singular values of the product matrix W(t) evolve by:

  $\dot{\sigma_r}(t) = -N(\sigma_r^2(t))^{1-\frac{1}{N}} < \nabla \mathrm{l}(W(t)), u_r(t)v_r^T(t) >$,
  $r = 1, ..., min(d, d')$

  If the matric factorization is non-degenerate , i.e has depth $N \geq 2$, the singular values need not be signed(we may assume $\sigma_r(t) \geq 0$ for all t

Dynamical analysis

- Lemma

  Assume that at initialization, the singular values of the product matirx W(t) are distinct and different from zero. Then, its singular vectors evolve by:

$$
\begin{aligned}
\dot{U}(t) &= -U(t)\left(F(t)\odot\left[U^\top(t)\nabla\ell(W(t))V(t)S(t)+S(t)V^\top(t)\nabla\ell^\top(W(t))U(t)\right]\right) \\
&\quad -\left(I_d-U(t)U^\top(t)\right)\nabla\ell(W(t))V(t)(S^2(t))^{\frac{1}{2}-\frac{1}{N}} \\
\dot{V}(t) &= -V(t)\left(F(t)\odot\left[S(t)U^\top(t)\nabla\ell(W(t))V(t)+V^\top(t)\nabla\ell^\top(W(t))U(t)S(t)\right]\right) \\
&\quad -\left(I_{d'}-V(t)V^\top(t)\right)\nabla\ell^\top(W(t))U^\top(t)(S^2(t))^{\frac{1}{2}-\frac{1}{N}},
\end{aligned}
$$

  where I_d and $I_{d'}$ are the identity matrices of sizes $d \times d$ and $d' \times d'$ respectively, $\odot$ stands for the *Hadamard* product , and the matrix $F(t) \in \mathbb{R}^{min(d,d'}$ is skew-symmetric with $((\sigma^2_{r'})^{\frac{1}{n}} - (\sigma^2_r(t))^{\frac{1}{n}})^{-1}$ in its (r,r')'th entry, $r \neq r'$

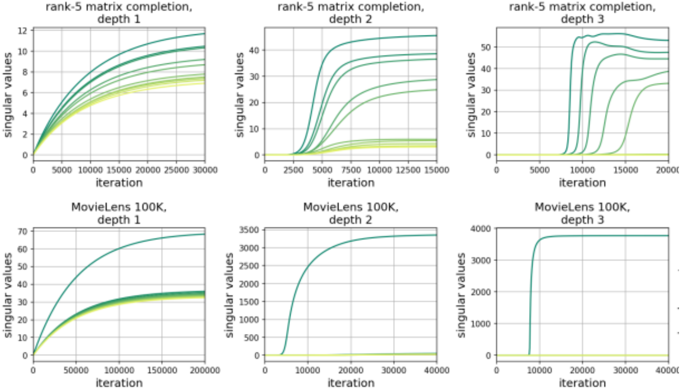
- Corollary 1:

  Assume the conditions of Lemma , and the matrix factorization is non-degenerative i.e has depth $N \geq 2$. Then, for any time t such that the singular vectors of the product matrix W(t) are stationary, i.e $\dot{W}(t) = 0$ and $\dot{V}(t) = 0$, it holds that $U^T(t)\triangledown l(W(t))V(t)$ is diagonal, meaning they align with the singular vectors of $\triangledown l(W(t))$

Dynamical analysis

- Lemma and Corollary suggests that a "goal" of gradient flow on a deep matrix factorization is to align singular vectors of the product matrix with those of the gradient.

Empirical demonstration



Interpretation

- It shows that for a non-degenerate deep matrix factorization, i.e one with depth $N \geq 2$, under gradient descent with small learning rate and near-zero initialization, singular values of the product matrix are subject to an enhancement/attenuation effect as described above.
- Singular value is an implicit regularization towards low rank, which intensifies with depth.

Table of Contents

Conclusion

- Through theory and experiments, we questioned prevalent norm-based explanations for implicit regularization in matrix factorization , and offered an alternative, dynamical approach.