I) Implicit Regularization in Deep Learning

DNNs generalize w/o explicit regularization when # of weights >> training set size

Conventional wisdom: gradient descent (GD) induces an implicit regularization

Can we mathematically understand this effect in concrete settings?

II) Setting: Matrix Completion

Matrix completion: recover low rank matrix (ground truth) given subset of entries

Convex Programming Approach

Minimize nuclear norm s.t. fitting observations: \( \min_{W} \|W\|_F \), s.t. \( W_{ij} = b_{ij} \forall (i, j) \in \Omega \)

Perfectly recovers if observations are sufficiently many \(^2\)

Depth leads to larger gaps between singular vals (lower rank)!

Langage of standard regularizers (e.g. norms) may not suffice

Current theory is oblivious to depth (always points to nuclear norm)

III) Deep Matrix Factorization

Deep Learning Approach ("deep matrix factorization")

Parameterize by depth \( N \) linear neural network and minimize \( \|W\|_F \) loss with GD:

\[ \min_{W_N, W_{N-1}, \ldots, W_1} \sum_{i,j \in \Omega} (W_NW_{N-1}\cdots W_1 - b_{ij})^2 \]  

No explicit regularization!

Post Work (Gunasekar et al. 2017)

For depth \( 2 \) only — \( \min_{W_1, W_2} \sum_{i,j \in \Omega} (W_2W_1 - b_{ij})^2 \):

- Experiments: low rank ground truth matrix oftentimes recovered accurately
- Conjecture: GD implicitly finds min nuclear norm solution (cf. convex approach)
- Theorem: conjecture holds for a certain restricted setting

\[ \sum_{t} \sigma^2_r(W) \text{ where } \langle r(W) \rangle, \text{ are singular vals of } W \]

Capturing implicit regularization via single norm may not be possible

V) Can the Stronger Implicit Regularization Be Explained via Norms?

Gunasekar et al. 2017: depth 2 implicitly minimizes nuclear norm (surrogate for rank)

Perhaps higher depths implicitly minimize other norms closer to rank?

Example

Schatten-\( p \) quasi-norm to the power of \( p \):

- \( p = 1 \): nuclear norm, corresponds to depth 2 according to Gunasekar et al. 2017
- \( p \in (0, 1) \): closer to rank, may correspond to higher depths

Theorem

In restricted setting where Gunasekar et al. 2017 proved depth 2 minimizes nuclear norm, any depth \( \geq 2 \) does so as well

Proposition

\( \exists \) instances of this setting where nuclear norm minimization contradicts minimization (even locally) of Schatten-\( p \) quasi-norm \( \forall p \in (0, 1) \)

IV) The Effect of Depth

Experiment

Matrix completion via GD over deep factorizations:

\[ \text{RMSE versus # of observations for rank-5 and rank-30 100x100 matrix completion} \]

Depth enhances implicit regularization towards low rank!

VI) Trajectory Analysis

Definition

Product matrix for depth \( N \) factorization: \( W_{N:N} = W_NW_{N-1}\cdots W_1 \)

Denote:

- \( \langle r(W) \rangle \): singular vals of \( W_{N:N} \)
- \( \{u_i\}_i / \{v_i\}_i \): corresponding left/right singular vecs

Theorem

GD over depth \( N \) matrix factorization on loss \( f(W) \) leads \( \{r(W)\} \) to evolve by:

\[ \sigma(t+1) - \sigma(t) - \frac{\sigma^2_r(W(t))}{\|\nabla f(W(t))\|} \]

Interpretation

- Given \( W_{N:N}(t) \), depth \( N \) affects evolution only via factors \( \sigma_r^2(W(t)) \)
- \( N = 1 \) (classic linear model): factors reduce to 1
- \( N \geq 2 \): factors slow down (speed up) small (large) singular vals, more significantly for larger \( N \) (higher depth)

Experiment

Singular vals during GD over deep factorizations for matrix completion:

\[ \text{RMSE versus # of observations for rank-5 100x100 matrix completion} \]

Theoretical Example

In matrix completion with one observation, GD over deep factorization leads singular vals to have relationship:

- \( \text{depth 1: linear} \)
- \( \text{depth 2: polynomial} \)
- \( \text{depth 3: asymptotic} \)

Depth leads to larger gaps between singular vals (lower rank)!

VII) Takeaway

To understand implicit regularization in deep learning:

- Language of standard regularizers (e.g. norms) may not suffice
- Carefully analyzing trajectories of GD might be necessary