Implicit Regularization in Matrix Factorization

Xiaojie Mao

Department of Statistical Science
Cornell University

xm77@cornell.edu

February 14, 2019

*Based on Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, Nathan Srebro (2017).

Quiz

1. The gradient descent on the factored form matrix factorization may minimize:

1. Frobenius Norm;
2. Spectral Norm;
3. Nuclear Norm;
4. Max Norm.

2. What kind of problem is considered in this paper?

1. Over-determined problem;
2. Under-determined problem;
3. Exact-determined problem.

Overview

1. Introduction
    - Motivating example: underdetermined least squares
    - Matrix Factorization
2. Analysis
3. Conclusion
4. Discussion

Introduction

Motivating example: underdetermined least squares

Consider least squares problem

$$\min_{x\in\mathbb{R}^p} \|Ax - y\|_2^2$$

where $A \in \mathbb{R}^{n\times p}$, $y \in \mathbb{R}^n$.

- Underdetermined: $p > n$;
- Many global minimizers with zero objective value;
- Gradient descent leads to minimum Euclidean norm solution if the initialization $x_0 \in \mathcal{R}(A)$, while other algorithms may not.

Motivating example: underdetermined least squares

Proof: Denote converging point of gradient descent as $x^* = \lim_{t\to\infty} x_t$.

- Convex optimization $\implies x^*$ is a global minimizer, i.e., $Ax^* = y$;
- Gradient $A^\top(Ax - y) \in \mathcal{R}(A)$ and $x_0 \in \mathcal{R}(A) \implies x^* \in \mathcal{R}(A)$;
- $x^* = A^\top z$ such that $AA^\top z = y \implies x^* = A^\top(AA^\top)^{-1}y = A^+y$ is the solution to the minumum Euclidean norm problem:

$$\min_{x\in\mathbb{R}^p} \|x\|_2 \quad \text{s.t.} \quad Ax = y$$

Phenomenon: in underdetermined problems, optimization algorithms may introduce implicit regularization.

Matrix Factorization

Convex formulation:

$$\min_{X \succeq 0} F(X) = \|\mathcal{A}(X) - y\|_2^2$$

where

- $X \in \mathbb{R}^{n \times n}$ is p.s.d, $y \in \mathbb{R}^m$;
- $\mathcal{A} : \mathbb{R}^{n \times n} \to \mathbb{R}^m$ is a linear operator specified $\mathcal{A}(X)_i = \langle A_i, X \rangle$ with symmetric linearly independent A_i;
- Underdetermined: $m \ll n^2$.

Non-convex formulation:

$$\min_{U \in \mathbb{R}^{n \times d}} f(U) = \|\mathcal{A}(UU^\top) - y\|_2^2$$

where we are mostly interested in the case $d = n$.

Gradient descent

Denote $r_t = \mathcal{A}(X_t) - y$ or $r_t = \mathcal{A}(U_t U_t^\top) - y$.

Projected gradient descent on X:

$$X_{t+1} = \mathcal{P}(X_t - \alpha_t \nabla F(X_t)) = \mathcal{P}(X_t - \alpha_t \mathcal{A}^*(r_t))$$

Gradient descent on U:

$$U_{t+1} = U_t - \alpha_t \nabla f(U_t) = U_t - \mathcal{A}^*(r_t) U_t$$

where $\mathcal{A}^* : \mathbb{R}^m \to \mathbb{R}^{n \times n}$ is the adjoint of $\mathcal{A}$ given by $\mathcal{A}^*(r) = \sum_{i=1}^m r_i A_i$.

Low Rank Matrix Completion

Suppose we can only **partially observe** entries of matrix X^* with **low rank** r, and we hope to **reconstruct** the whole matrix. The most straightforward optimization problem is:

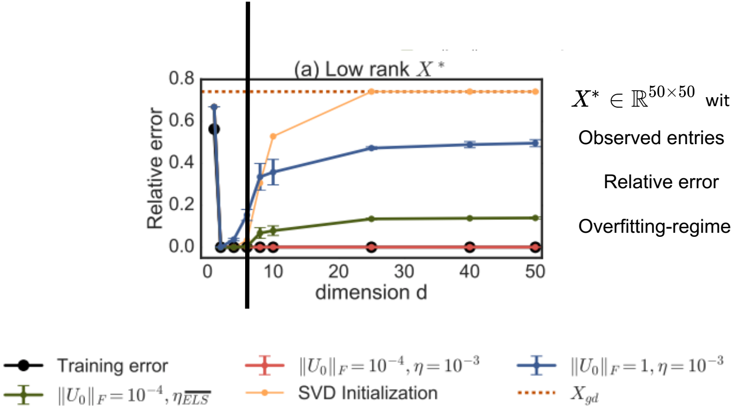
$$\min_{X \in \mathbb{R}^{n \times n}} \sum_{(i,j) \in \Omega} \|X_{ij} - X_{ij}^*\|_2^2 \quad \text{s.t.} \quad \mathrm{Rank}(X) \leq r.$$

where $\Omega \subset \{1, \ldots, n\} \times \{1, \ldots, n\}$ is the index set for observed entries. The solution to the convex relaxation problem

$$\min_{X \in \mathbb{R}^{n \times n}} \sum_{(i,j) \in \Omega} \|X_{ij} - X_{ij}^*\|_2^2 + \lambda \|X\|_*.$$

guaranteed to recover X^* when the number of observations $m = |\Omega|$ is large enough. Reconstruction is generally not guaranteed when $m < nr$ or $r > \frac{m}{n}$.

Empirical Observations: generalization



$X^* \in \mathbb{R}^{50\times 50}$ with $r = 2$

Observed entries $m = 3nr$

Relative error $\frac{\|X - X^*\|_F}{\|X^*\|_F}$

Overfitting-regime $d \gg \frac{m}{n}$

The solution given by gradient descent on the factors (with **initialization ≈ 0 and small step size**) achieves great **generalization** even without rank constraint.

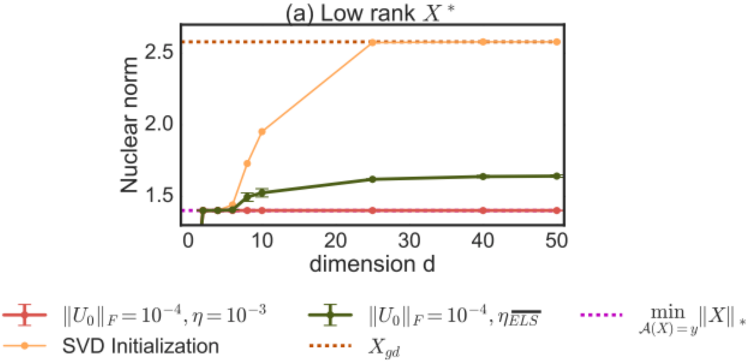
Question

Why gradient descent on factors has good generalization?

- Not because of model complexity: $d \gg \frac{m}{n}$;
- Not because of explicit regularization: $d = n$;
- **The optimization algorithm may introduce implicit regularization!**

But what's the regularization here?

Empirical Observations: nuclear norm



The solution given by gradient descent on the factors (with **initialization** ≈ 0 **and small step size**) achieves minimum nuclear norm.

Conjecture (Informal)

With **small step size** and **intialization** ≈ 0, if the gradient descent on the factored form matrix factorization

$$\min_{U \in \mathbb{R}^{n \times d}} f(U) = \|\mathcal{A}(UU^\top) - y\|_2^2$$

converges to global minimum with $\mathcal{A}(X) = y$, then it also converges to a minimum nuclear norm solution

$$\operatorname{argmin}_{X \succeq 0} \|X\|_* \quad \text{s.t.} \quad \mathcal{A}(X) = y$$

Analysis

Analysis Roadmap

- **Nuclear Norm Minimization** (NNM): characterize the solution by KKT condition;
- **Factored Matrix Factorization** (MF): characterize the trajectory of gradient descent (GD) by the solution **manifold** of a differential equation;
- Prove that

$$\underbrace{\{\text{GD Solution Manifold}\} \cap \{\text{Global Optimizer}\}}_{MF} \subset \underbrace{\{\text{KKT}\}}_{NNM}$$

Nuclear norm minimization

Consider

$$\operatorname{argmin}_{X \succeq 0} \|X\|_* \quad \text{s.t.} \quad \mathcal{A}(X) = y$$

whose KKT optimality conditions are

$$\begin{aligned} &\mathcal{A}(X) = y, \qquad\qquad X \succeq 0 \\ &\exists v \in \mathbb{R}^m \quad \text{s.t. } I - \mathcal{A}^*(v) \succeq 0, \quad \mathcal{A}^*(v)X = X \end{aligned}$$

Factored matrix factorization

Gradient descent with **infinitesimally small** step size:

$$\dot{U}_t = -\nabla f(U_t) = -\mathcal{A}^*(\mathcal{A}(U_t U_t^\top) - y)U_t$$

which can be equivalently characterized by

$$\dot{X}_t = \dot{U}_t U_t^\top + U_t \dot{U}_t^\top = -\mathcal{A}^*(r_t)X_t - X_t\mathcal{A}^*(r_t)$$

where $r_t = \mathcal{A}(X_t^\top) - y$.

Define the limit point $X_\infty(X_{\text{init}}) := \lim_{t\to\infty} X_t$ with intialization $X_0 = X_{\text{init}}$.

Restate the conjecture (formal)

Conjecture: for any full rank X_{init} (**implicitly requires factor dimension** $d = n$), if

- $\hat{X} = \lim_{\alpha \to 0} X_\infty(\alpha X_{\text{init}})$ exists,
- and $\hat{X}$ is a global optima for the matrix factorization with $\mathcal{A}(\hat{X}) = y$,

then $\hat{X}$ satisfies the KKT conditions for the nuclear norm minimization:

- $\mathcal{A}(X) = y$, $X \succeq 0$;
- There exists $v \in \mathbb{R}^m$ such that $I - \mathcal{A}^*(v) \succeq 0$, and $\mathcal{A}^*(v)\hat{X} = \hat{X}$.

Problem: how can we characterize $\hat{X}$ and find v?

Gradient flow: $m = 1$

Consider the number of measurements $m = 1$: $A_1 = A$, $y_1 = y$, and the gradient flow is characterized by

$$\dot{X}_t = -r_t(AX_t + X_tA)$$

Step 1: characterize $\hat{X}$. The solution to the differential equation is

$$X_t = \exp(s_tA)(\alpha X_{\text{init}})\exp(s_tA)$$

with $s_T = -\int_0^T r_t dt$.

As long as $y \neq 0$, $\mathcal{A}(\hat{X}) = y$ implies that $\hat{X} \neq 0$. Therefore, the existence of $\hat{X} = \lim_{\alpha \to 0} X_\infty(\alpha X_{\text{init}})$ implies that $s_\infty = \infty$.

Gradient flow: measurements $m = 1$

Denote the eigen-decomposition $A = QDQ^\top$, then

$$X_\infty = Q \exp(D^{s_\infty})(\alpha X_{\text{init}}) \exp(D^{s_\infty}) Q^\top$$

So $\hat{X} = \lim_{\alpha \to 0} X_\infty(\alpha X_{\text{init}})$ is effectively spanned by the top eigenvector(s) of A. For example, consider

$$\hat{X} = Q \begin{bmatrix} 100 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} Q^\top$$

where Q_1 is the top eigenvector.

Step 2: find the valid v to verify KKT condition. Set $v \in \mathbb{R}$ such that $v\lambda_{\max}(A) = 1$, then $I - vA \succeq 0$, and $vA = Q(vD)Q^\top$ has eigenvalue 1 corresponding to top eigenvector(s) of A. For example, consider

$$vA = Q \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0 \end{bmatrix} Q^\top.$$

Gradient flow: measurements $m = 1$

In summary, $\hat{X}$ is spanned by the top eigenvector(s) of A, and vA has eigenvalue 1 corresponding to top eigenvector(s) of A. As a result, $vA\hat{X} = \hat{X}$. For example,

$$
\begin{aligned}
vA\hat{X} &= Q \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 100 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} Q^\top \\
&= Q \begin{bmatrix} 100 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} Q^\top = \hat{X}
\end{aligned}
$$

Key steps:

- A and $\hat{X}$ can both be characterized by the same eigen-basis;
- vA recovers the nonzero diagonal block in the decomposition of $\hat{X}$.

Gradient flow: measurements $m > 1$

Theorem

When symmetric matrices $\{A_i\}_{i=1}^m$ **commute** *(i.e.,* $A_iA_j = A_jA_i$*), if* $\hat{X} = \lim_{\alpha \to 0} X_\infty(\alpha I)$ *exists, and is a global optimum for matrix factorization with* $\mathcal{A}(\hat{X}) = y$*, then* $\hat{X} \in \operatorname{argmin}_{X \succeq 0} \|X\|_*$ *such that* $\mathcal{A}(X) = y$*.*

Why commutative $\{A_i\}_{i=1}^m$?

- The gradient flow has nice form:

$$X_t = \alpha \exp(\mathcal{A}^*(s_t)) \exp(\mathcal{A}^*(s_t)).$$

- $\{A_i\}_{i=1}^m$ are simultaneously diagonalizable by the same set of eigen-basis, and so are $\mathcal{A}^*(s)$ for any $s \in \mathbb{R}^m$, X_t, and $\hat{X}$.

Proof for commutative case

Goal: we need to find a $v \in \mathbb{R}^m$ such that $I - \mathcal{A}^*(v) \succeq 0$, and $\mathcal{A}^*(v)\hat{X} = \hat{X}$.

Observations:

- $\mathcal{A}^*(v)$ and $\hat{X}$ have the same eigen-basis, and so is $\mathcal{A}^*(v)\hat{X}$;
- we need to find v such that

$$\lambda_k(\mathcal{A}^*(v)\hat{X}) = \lambda_k(\mathcal{A}^*(v))\lambda_k(\hat{X}) = \lambda_k(\hat{X});$$

- we need to find v such that $\lambda_k(\mathcal{A}^*(v)) = 1$ if $\lambda_k(\hat{X}) \neq 0$ and $\lambda_{\max}(\mathcal{A}^*(v)) \leq 1$.

Proof for commutative case

Recall $\hat{X} = \lim_{\alpha \to 0} X_\infty(\alpha I)$ with

$$X_\infty(\alpha I) = \alpha \exp(\mathcal{A}^*(s_\infty)) \exp(\mathcal{A}^*(s_\infty))$$

So **analyzing the eigenvalues of $\hat{X}$ requires analyzing eigenvalues of $X_\infty(\alpha I)$**: with $\beta = -\frac{1}{2}\log\alpha$,

$$\begin{aligned}\lambda_k(X_\infty(\alpha I)) &= \alpha\left[\exp\left(\lambda_k\left(\mathcal{A}^*(s_\infty)\right)\right)\right]^2 \\ &= \exp\left(2\lambda_k\left(\mathcal{A}^*(s_\infty)\right) - 2\beta\right)\end{aligned}$$

$$\lambda_k(\mathcal{A}^*(\frac{s_\infty}{\beta})) - \frac{\log \lambda_k(X_\infty(\exp(-2\beta)I))}{2\beta} = 1.$$

which implies that as $\beta \to \infty$ ($\alpha \to 0$),

$$\lambda_k(\mathcal{A}^*(\frac{s_\infty}{\beta})) - \frac{\log \lambda_k(\hat{X})}{2\beta} \to 1.$$

Proof for commutative case

So far, we know that as $\beta \to \infty$ ($\alpha \to 0$), with $v(\beta) = \frac{s_\infty}{\beta}$

$$\lambda_k(\mathcal{A}^*(v(\beta))) - \frac{\log \lambda_k(\hat{X})}{2\beta} \to 1.$$

- For $\lambda_k(\hat{X}) \neq 0$, apparently

$$\lim_{\beta \to \infty} \lambda_k(\mathcal{A}^*(v(\beta))) = 1.$$

- For $\lambda_k(\hat{X}) = 0$,

$$\left[\exp(\lambda_k(\mathcal{A}^*(v(\beta))) - 1)\right]^{2\beta} = \lambda_k(X_\infty(\exp(-2\beta)I)) \to \lambda_k(\hat{X}) = 0,$$

so for sufficiently large β and $\epsilon \in (0, 1]$,

$$\exp(\lambda_k(\mathcal{A}^*(v(\beta))) - 1) < \epsilon^{\frac{1}{2\beta}} < 1 \Longrightarrow \lim_{\beta \to \infty} \lambda_k(\mathcal{A}^*(v(\beta))) < 1$$

So $v = \lim_{\beta\to\infty} v(\beta)$ satisfies the condition. [QED]

Big picture

Recall the plan:

$$\underbrace{\{\text{GD Solution Manifold}\} \cap \{\text{Global Optimizer}\}}_{MF} \subset \underbrace{\{\text{KKT}\}}_{NNM}$$

In the commutative case, the gradient descent with infinitesimal step size traverses on the solution manifold

$$\{X \in \mathbb{R}^{n\times n} : X = \alpha \exp(\mathcal{A}^*(s)) \exp(\mathcal{A}^*(s)), s \in \mathbb{R}^m\},$$

and with the extra restriction as a global optimum with $\mathcal{A}(X) = y$, the final solution satisfies the KKT condition.

Conclusion

Conclusion

- Optimization algorithms for **underdetermined problems** can introduce **implicit regularization** that bias the solution towards a specific global minimum (not necessarily useful);
  - The algorithms may actually solve a nicer problem!
- Gradient descent on **factored form** of matrix factorization converging lead to **minimmum nuclear norm** solution, if it converges to a global minimum;
- The **solution manifold** characterizing the dynamics of gradient descent and the **global minimum** jointly determines the regularization.

Discussion

Question I

How practically useful is the result (e.g., considering the assumptions of small step size, intialization close to 0, convergence to global minimum)?

How practically useful is this result?

- This paper mainly points out that optimization algorithms introduce regularization, but does not answer if it's **useful** regularization;
  - **Usefulness** of the regularization depends on **problem structure**. For example the Euclidean norm regularization in underdetermined least squares may not be useful.
- Small step size $\Longrightarrow$ the trajectory of gradient descent doesn't fall off of the solution manifold.
- Initialization ≈ 0 will result in **slow convergence**:

$$U_{t+1} = U_t - \alpha_t \nabla f(U_t) = U_t - \mathcal{A}^*(r_t)U_t$$

- It assumes that gradient descent converges to global minima;
  - Gradient descent converges to local minima under **random initialization** (Lee at. al., 2016);
  - All local minima in matrix factorization are global minima (Rong Ge, Jason D. Lee, and Tengyu Ma, 2016).

Question II

What's your view on the analysis in this paper? Can it be extended to more general problems? What are the challenges of the extension?

Challenges for extensions

- Characterize the solution manifold: it is very challenging even for non-commutative $\{A_i\}_{i=1}^m$ in matrix factorization.

$$X_t = \lim_{\varepsilon \to 0} \left(\prod_{\tau=\frac{t}{\varepsilon}}^{1} \exp(-\varepsilon \mathcal{A}^*(r_{\tau\varepsilon})) \right) X_0 \left(\prod_{1}^{\tau=\frac{t}{\varepsilon}} \exp(-\varepsilon \mathcal{A}^*(r_{\tau\varepsilon})) \right)$$

  The $\exp(\mathcal{A}^*(s))$ term is very intractable because

$$\exp(A_1 + A_2) \neq \exp(A_1) \exp(A_2)$$

  unless A_1 and A_2 commute. Also, we lose the nice characterization with the same eigen-basis.

- Characterize the global optimum: the zero error set is hard to summarize for very complicated models, e.g., deep neural net.

Thanks!