Implicit Regularization in Matrix Factorization

Jeong Hwichang

Seoul National Universitiy

2021.05.13.

Outline

- Introduction
- Matrix regression
- Experiment
- Theorem
- Proof

Introduction

- Deep models often generalize well when trained purely by minimizing the training error, and when optimization problem is underdetermined.
- Even though there are many zero training error solutions, optimization algorithm seems to prefer solutions that do generalize well.
- This bias is not explicitly specified in the objective or problem formulation.(Implicit bias)

Introduction

- It seems that the optimization algorithm minimizes some implicit regularization measure.
- This paper analyze implicit regularization in matrix factorization models.
- Identify the implicit regularizer as the nuclear norm.

Matrix Regression

- Consider least squares objectives over matrices $X \in \mathbb{R}^{n \times n}$ of the form:

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

  where $\mathcal{A} : \mathbb{R}^{n \times n} \to \underline{\mathbf{R}}^m$ is a linear operator specified by $\mathcal{A}(X)_i = \langle A_i, X \rangle, A_i \in \mathbb{R}^{n \times n}$, and $y \in \mathbb{R}^m$.
- Consider only symmetric positive semidefinite X and symmetric linearly independent A_i.

Matrix Regression

- Instead of working on X directly, use a factorization of $X = UU^T$.

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

- If $m \ll n^2$, then above problem is underdetermined and can be optimized in many ways.
- Estimating a global optima cannot ensure generalization.

Matrix Regression

- To simulate matrix reconstruction problem, generate $m \ll n^2$ random measurement matrices and set $y = \mathcal{A}(X^*)$ for some planted $X^* \succeq 0$.
- By performing gradient descent on U to convergence and then measure the relative reconstruction error $\|X - X^*\|_F$.
- Here η is learning rate and U_0 is initial value.

Experiment

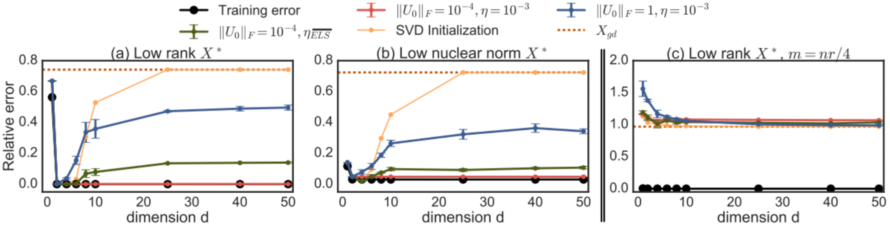

Figure 1: Reconstruction error of the solutions for the planted 50×50 matrix reconstruction problem. In (a) X^* is of rank $r = 2$ and $m = 3nr$, in (b) X^* has a spectrum decaying as $O(1/k^{1.5})$ normalized to have $\|X^*\|_* = \sqrt{r}\|X^*\|_F$ for $r = 2$ and $m = 3nr$, and in (c) we look at a non-reconstructable setting where the number of measurements $m = nr/4$ is much smaller than the requirement to reconstruct a rank $r = 2$ matrix. The plots compare the reconstruction error of gradient descent on U for different choices initialization U_0 and step size η, including fixed step-size and exact line search clipped for stability ($\eta_{\overline{ELS}}$). Additonally, the orange dashed reference line represents the performance of X_{gd} — a rank unconstrained global optima obtained by projected gradient descent on X space for (1), and 'SVD-Initialization' is an example of an alternate rank d global optima, where initialization U_0 is picked based on SVD of X_{gd} and gradient descent with small stepsize is run on factor space. The results are averaged across 3 random initialization and (nearly zero) errorbars indicate the standard deviation.

Experiment

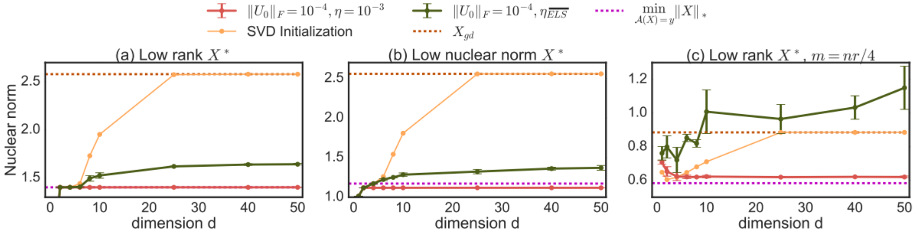

Figure 2: Nuclear norm of the solutions from Figure 1. In addition to the reference of X_{gd} from Figure 1, the magenta dashed line (almost overlapped by the plot of $\|U\|_F = 10^{-4}, \eta = 10^{-3}$) is added as a reference for the (rank unconstrained) minimum nuclear norm global optima. The error bars indicate the standard deviation across 3 random initializations. We have dropped the plot for $\|U\|_F = 1, \eta = 10^{-3}$ to reduce clutter.

Matrix Regression

Theorem

In the case where matrices $\{A_i\}_{i=1}^m$ *commute, if* $\widehat{X} = \lim_{\alpha \to 0} X_\infty(\alpha I)$ *exists and is a global optimum for* $\min_{X \succeq 0} \|\mathcal{A}(X) - y\|_2^2$ *with* $\mathcal{A}(\widehat{X}) = y$*,then* $\widehat{X} \in \operatorname{argmin}_{X \succeq 0} \|X\|_*$ *s.t.* $\mathcal{A}(X) = y$.

- Here limit point $X_\infty(X_{init}) := \lim_{t \to \infty} X_t$ for the factorized gradient flow initialized at $X_0 = X_{init}$.

Proof

Using the chain rule

$$\dot{X}_t = \dot{U}_t U_t^\top + U_t \dot{U}_t^\top = -\mathcal{A}^*(r_t) X_t - X_t \mathcal{A}^*(r_t) \cdots (1)$$

where $\mathcal{A}^* : \underline{\mathbf{R}}^m \to \mathbb{R}^{n \times n}$ is the adjoint of $\mathcal{A}$ and is given by $\mathcal{A}^*(r) = \sum_i r_i A_i$ and $r_t = \mathcal{A}(X_t) - y$.
When A_i commute, Defining $s_T = -\int_0^T r_t dt-$ a vector integral, we can verify by differentiating that solution of (1) is

$$X_t = \exp(\mathcal{A}^*(s_t)) X_0 \exp(\mathcal{A}^*(s_t)) \cdots (2)$$

Proof

Our problem is

$$\min_{X \succeq 0} \|X\|_* \text{ s.t. } \mathcal{A}(X) = y \cdots (3)$$

The KKT optimality conditions for (1) are:

$$\exists \nu \in \mathbb{R}^m \text{ s.t.} \quad \mathcal{A}(X) = y \quad X \succeq 0 \quad \mathcal{A}^*(\nu) \preceq I \quad (I - \mathcal{A}^*(\nu))\, X = 0 \cdots (4)$$

It suffices to show that such a $\widehat{X}$ satisfies the complementary slackness and dual feasibility KKT conditions in (4). Since the matrices A_i commute and are symmetric, they are simultaneously diagonalizable by a basis $v_1, .., v_n$, and so is $\mathcal{A}^*(s)$ for any $s \in \mathbb{R}^m$. This implies that for any $\alpha, X_\infty(\alpha I)$ given by (2) and its limit $\widehat{X}$ also have the same eigenbasis. Furthermore, since $X_\infty(\alpha I)$ converges to $\widehat{X}$, the scalars $v_k^\top X_\infty(\alpha I) v_k \to v_k^\top \widehat{X} v_k$ for each $k \in [n]$. Therefore, $\lambda_k\left(X_\infty(\alpha I)\right) \to \lambda_k(\widehat{X})$, where $\lambda_k(\cdot)$ is defined as the eigenvalue corresponding to eigenvector v_k and not necessarily the k^{th} largest eigenvalue.

Proof

Let $\beta = -\log \alpha$, then $\lambda_k\left(X_\infty(\alpha I)\right) = \exp\left(2\lambda_k\left(\mathcal{A}^*\left(s_\infty(\beta)\right)\right) - 2\beta\right)$. For all k such that $\lambda_k(\widehat{X}) > 0$, by the continuity of log, we have

$$2\lambda_k\left(\mathcal{A}^*\left(s_\infty(\beta)\right)\right) - 2\beta - \log \lambda_k(\widehat{X}) \to 0 \Longrightarrow \lambda_k\left(\mathcal{A}^*\left(\frac{s_\infty(\beta)}{\beta}\right)\right) - 1 - \frac{\log \lambda_k(\widehat{X})}{2\beta} \to 0$$

Defining $\nu(\beta) = s_\infty(\beta)/\beta$, we conclude that for all k such that $\lambda_k(\widehat{X}) \neq 0, \lim_{\beta\to\infty} \lambda_k\left(\mathcal{A}^*(\nu(\beta))\right) = 1$ Similarly, for each k such that $\lambda_k(\widehat{X}) = 0$

$$\exp\left(2\lambda_k\left(\mathcal{A}^*\left(s_\infty(\beta)\right)\right) - 2\beta\right) \to 0 \Longrightarrow \exp\left(\lambda_k\left(\mathcal{A}^*(\nu(\beta))\right) - 1\right)^{2\beta} \to 0$$

Thus, for every $\epsilon \in (0, 1]$, for sufficiently large β

$$\exp\left(\lambda_k\left(\mathcal{A}^*(\nu(\beta))\right) - 1\right) < \epsilon^{\frac{1}{2\beta}} < 1 \Longrightarrow \lambda_k\left(\mathcal{A}^*(\nu(\beta))\right) < 1$$

Therefore, we have shown that $\lim_{\beta\to\infty} \mathcal{A}^*(\nu(\beta)) \preceq I$ and $\lim_{\beta\to\infty} \mathcal{A}^*(\nu(\beta))\hat{X} = \widehat{X}$ establishing the optimality of $\widehat{X}$ for (3).