# Implicit Regularization in Matrix Factorization

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, Nathan Srebro

TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO

## Matrix Estimation from Linear Measurementnts

$$\min_{X\in\mathbb{R}^{n\times n}} F(X)=\|\mathcal{A}(X)-b\|_2^2=\sum_{i=1}^m(\langle A_i,X\rangle-b_i)^2$$

e.g matrix completion (), linear neural networks,…

$\mathbf{m \ll n^2}$ underdetermined $\Rightarrow$ **many global minima**

$\longrightarrow$ Gradient descent on $X$ converges to the global optimum with minimum Frobenius norm $X^*_F=\operatorname{argmin}_{\mathcal{A}(X)=b}\|X\|_F^2$

For matrix completion, $X^*_F$ is a trivial imputation with zeros

$$\min_{U,V\in\mathbb{R}^{n\times n}} f(U,V)=F(UV^\top)=\|\mathcal{A}(UV^\top)-b\|_2^2$$

No explicit regularization, equivalent problem with many global minima

$\longrightarrow$ Gradient descent on $f(U,V)$

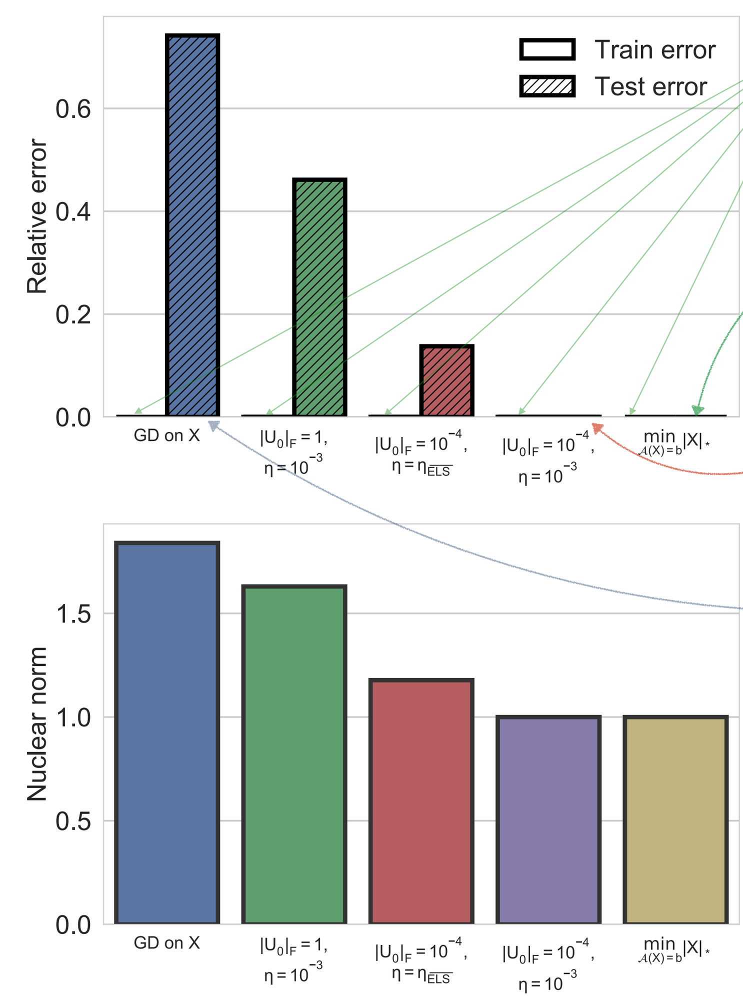
$$U_{k+1}=U_k-\eta\nabla_U f(U_k,V_k)$$
$$V_{k+1}=V_k-\eta\nabla_V f(U_k,V_k)$$

Empirically, does not reach trivial global minima and generalizes (see Example —>)

**Question:** Which global minimum does gradient descent on $f(U,V)$ get to?

## Example

$n=50, m=300,$ $A_i$ random Gaussian
$y=\mathcal{A}(X^*)$ generated from ground truth $X^*\in\mathbb{R}^{n\times n}$ of rank-2




Train error $\approx 0$ for all estimates, i.e. global optima reached by all algorithms

$X^*$ is low-rank and $A_i$'s are random
$\Rightarrow \min_{\mathcal{A}(X)=b}\|X\|_*$ recovers $X^*$

Under some conditions on initialization and step size, solution from gradient descent on $f(U,V)$ generalizes without any explicit regularization

Some other global optima do not generalize

## Our Conjecture

Gradient descent on $f(U,V)$ converges to the minimum nuclear norm solution when:

- The initialization is very close to 0
- The step size is very small

### Symmetric psd factorization

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

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

An asymmetric factorization $\|\tilde{\mathcal{A}}(\tilde{U}\tilde{V}^\top)-b\|_2^2$ is a special case of symmetric factorization $\|\mathcal{A}(UU^\top)-b\|_2^2$ with $A_i=\begin{bmatrix}0 & \tilde{A}_i\\ \tilde{A}_i^\top & 0\end{bmatrix}$

- Henceforth, we only consider the symmetric psd factorization which is **more general** and subsumes the asymmetric factorization

### Gradient flow

Gradient flow is the continuous-time limit of gradient descent as step size goes to zero, dynamics given by:

$$\frac{dU_t}{dt}=\dot{U}_t=-\nabla_U f(U_t)=-\mathcal{A}^*(r_t)U_t$$
$$\boxed{r_t=\mathcal{A}(U_tU_t^\top)-b}$$

**Conjecture:** As $\|U_0\|\to 0$ if gradient flow converges to a global minimum $U_\infty=\lim_{t\to\infty}U_t$, i.e. $f(U_\infty)=0$, then

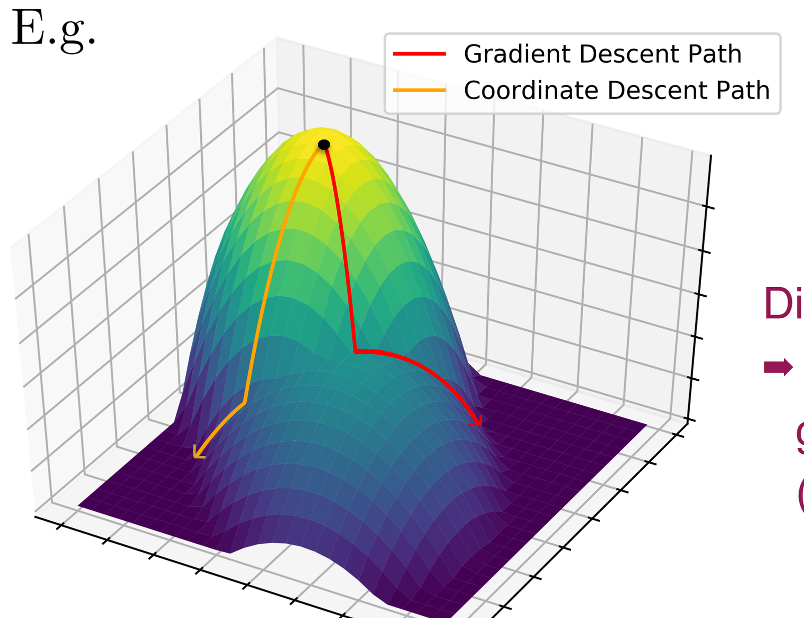
$$X_\infty=U_\infty U_\infty^\top=X^*_{\mathrm{NN}}=\operatorname{argmin}_{\mathcal{A}(X)=b}\|X\|_*$$

## Motivation

**Choice of optimization algorithm** $\Longleftrightarrow$ **Implicit bias towards certain global optima**

Understand the bias associated with different algorithms

E.g.




Different global optima $\Rightarrow$ different generalization (test error)

“implicit bias” plays a crucial role in neural network learning

- neural networks are highly overparametrized — most global minima are bad for generalization
- local search methods (like SGD) find non trivial solutions that generalize well (even without explicit regularization) — implicit regularization key in explaining generalization
- matrix factorization is a simple 2 layer linear network — segue for understanding the impact bias rigorously

## What We Can Prove: Commutative Measurements

**Theorem:** Let $U_\infty(\alpha)$ be the solution of gradient flow initialized at $U_0=\alpha I$. If measurements $A_i$ commute, i.e. $A_iA_j=A_jA_i$, and if $\bar{X}_\infty=\lim_{\alpha\to 0}U_\infty(\alpha)U_\infty(\alpha)^\top$ exists and satisfies $\mathcal{A}(\bar{X}_\infty)=b$, then $\bar{\mathbf{X}}_\infty=\mathbf{X}^*_{\mathrm{NN}}$

**Corollary:** Consider a non-negative *vector* least squares problem $\min_{x\in\mathbb{R}^n_+}\bar{F}(x)=\|Ax-b\|_2^2$.

Let $x=u^2$ (element wise square) for $u\in\mathbb{R}^n$, then gradient flow for $\min_{u\in\mathbb{R}^n}\bar{f}(u)=\|Au^2-b\|_2^2$ initialized at $u_0=\alpha\vec{1}$, as $\alpha\to 0$ will converge to a global optimum $u_\infty$ such that $\mathbf{u}^2_\infty=\operatorname{argmin}_{\mathbf{Ax=b}}\|\mathbf{x}\|_1$

### Proof strategy

1. Characterize the gradient flow path

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

$$\xrightarrow{(A_i\text{'s commute})} X_t=\exp\left(-\mathcal{A}^*(s_t)\right)U_0U_0^\top\exp\left(-\mathcal{A}^*(s_t)\right)$$
$$\xrightarrow{(U_0=\alpha I)} \quad =\exp\left(-2\mathcal{A}^*(s_t)+2\log(\alpha)I\right)\quad \boxed{s_t=\int_0^t r_s ds}$$

2. KKT conditions of $\min_{X\succeq 0}\|X\|_*$ s.t. $\mathcal{A}(X)=y$

$\mathcal{A}(X)=b \longrightarrow$ (satisfied by global optimality)
$X\succeq 0 \longrightarrow$ (guaranteed by psd factorization)

$\mathcal{A}^*(\nu)\preceq I$
$\mathcal{A}^*(\nu)X=X$ $\longrightarrow$ Construct dual certificate $\bar{\nu}$ for $\bar{X}_\infty$

- $\nu_\alpha\stackrel{\text{def}}{=}\lim_{t\to\infty}\frac{s_t}{\log(\alpha)}$, and $\bar{\nu}=\lim_{\alpha\to 0}\nu_\alpha$
- $\lim_{\alpha\to 0}\lambda_{\max}\left(\mathcal{A}^*(\nu_\alpha)\right)=1\Rightarrow\mathcal{A}^*(\bar{\nu})\preceq I$
  as $\|\bar{X}_\infty\|=0$ (if $\lambda_{\max}(\mathcal{A}^*(\bar{\nu}))<1$) or $\|\bar{X}_\infty\|=\infty$ (if $\lambda_{\max}(\mathcal{A}^*(\bar{\nu})>1$)
- $\bar{X}_\infty=\lim_{\alpha\to 0}e^{-2(\mathcal{A}^*(\nu_\alpha)-I)\log\alpha}$
  spanned by top eigenvectors of $\mathcal{A}^*(\bar{\nu})$

Proof **does not** depend on $r_t$ being residuals in the gradient flow path, but just on the form of $X_t$ (the $m$-dimensional manifold parametrized by $s_t\in\mathbb{R}^m$)

- $\eta\to 0$ necessary to remain in the (non-linear) manifold

## General Case

If measurements do not commute:

- no simple expression for gradient flow path
- path described a “time ordered exponential”

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

- unlike commutative case, using arbitrary steering $r_t$ — not generated from residuals along the gradient flow path — can lead to **any** p.s.d matrix
- need to exploit the specific structure of *residuals* — $r_t=\mathcal{A}(U_tU_t^\top)-b$ from gradient flow to construct dual certificate