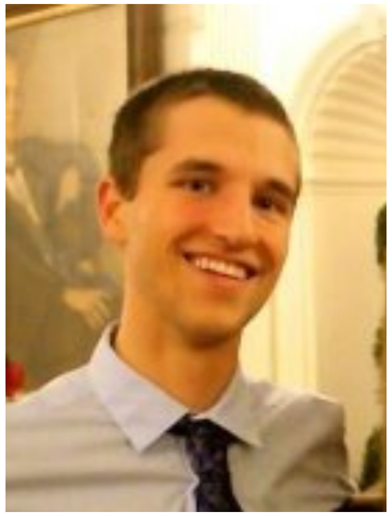
Implicit Regularization in Matrix Factorization

Suriya Gunasekar

TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO

http://www.ttic.edu/

Joint work with…


Blake Woodworth


Srinadh Bhojanapalli


Behnam Neyshabur


Nati Srebro

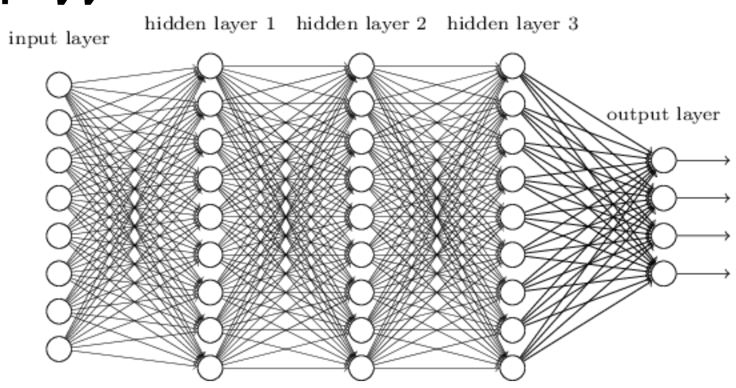
Modern machine learning algorithms

h_W parameterized by W $\in \mathbb{R}^d$

$$\min_W f_S(W) = \sum_{i=1}^{m} \ell(h_W(x_i), y_i) + \mathcal{R}(W)$$

With regularization:
$\mathcal{H}_R = \{h_W : W \in \mathbb{R}^d, \mathcal{R}(W) \le t\}$

- $h_w(x)$ is highly non-convex function of w
  e.g. $h_w(x) = W_k * \sigma\Big(\dots\sigma\big(W_2 * \sigma(W_1 * x)\big)\Big)$
  where $W = vec([W_k, \dots, W_2, W_1])$, $\sigma(z) = \max(0, z)$



- w is very high dimensional $d \gg m$
  Datasets CIFAR ~ 60K images, ImageNet ~14M images, ~1M annotations
  Deep learning architectures
  AlexNet (2012): 8 layers, 60M parameters
  VGG-16 (2014): 16 layers, 138M parameters
  ResNet (2015): 152 layers, …

Mysteries of deep learning I

$$\min_W f_S(W) = \sum_{i=1}^m \ell(h_W(x_i), y_i) + \mathcal{R}(W)$$

- $h_w(x)$ is highly non-convex function of w

  e.g. $h_w(x) = W_k * \sigma\left(\ldots \sigma(W_2 * \sigma(W_1 * x))\right)$

  where $W = vec([W_k, \ldots, W_2, W_1]), \quad \sigma(z) = \max(0, z)$

→ Finding the global optimum is hard

- "local search" methods like (stochastic) gradient descent can be guaranteed to converge only to a local optimum

→ Easy to get 0-training error solutions using local search

- (informally) over-parameterization makes optimization easy

Mysteries of deep learning II

$$\min_{W} f_S(W) = \sum_{i=1}^{m} \ell(h_W(x_i), y_i) + \mathcal{R}(W)$$

- w is very high dimensional $d \gg m$

  Datasets CIFAR ~ 60K images, ImageNet ~14M images, ~1M annotations

  Deep learning architectures AlexNet (2012): 8 layers, 60M parameters; VGG-16 (2014): 16 layers, 138M parameters

→ High variance and bad generalization on test dataset

- Many global minima. e.g. $\min_{\mathrm{w} \in \mathbb{R}^{100}} \sum_{i=1}^{10} (w^\top x_i - y_i)^2$
- MOST global optima are bad for generalization

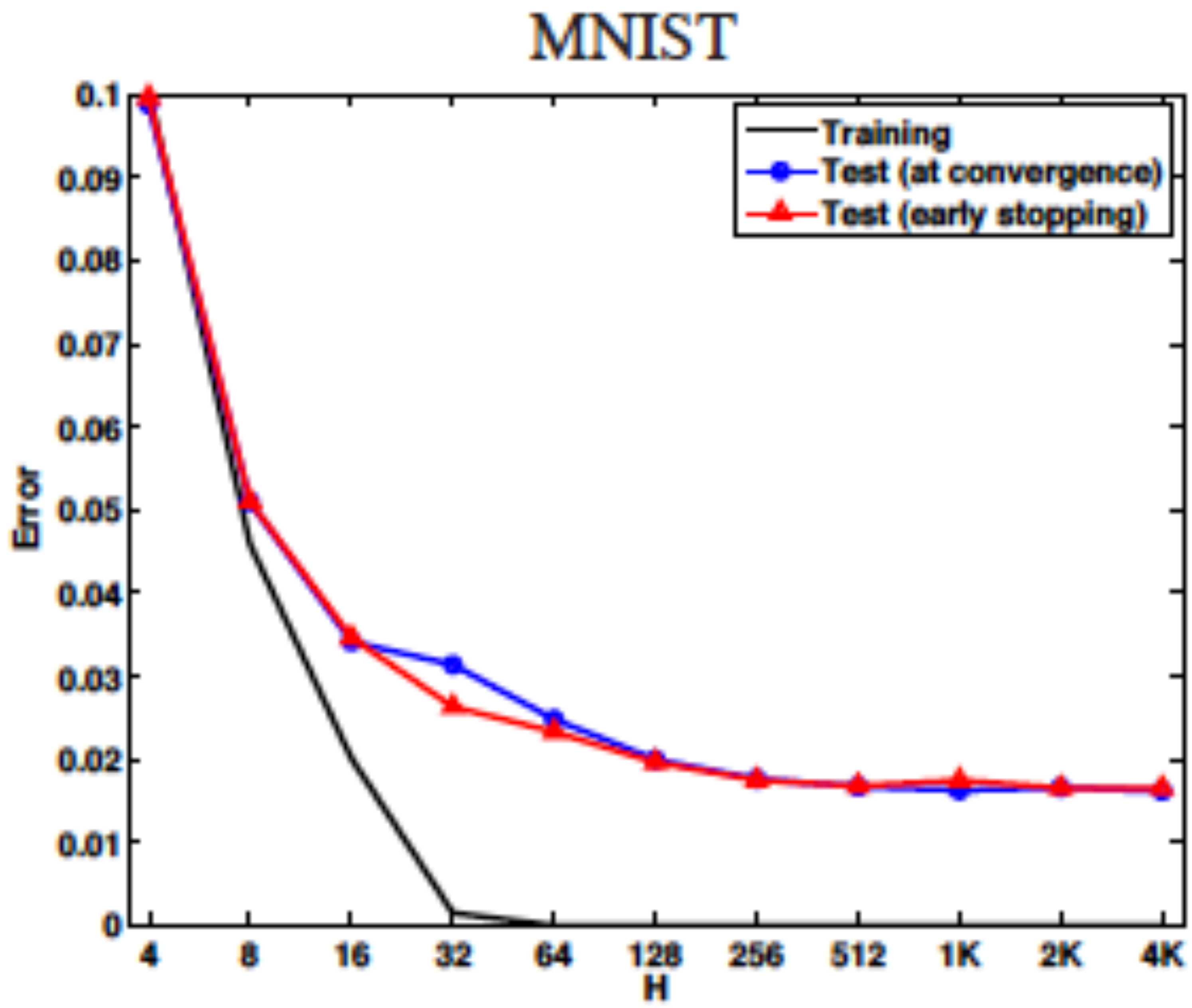
→ Pre-DNN models had strong regularization for high dimensional estimation

→ Solutions from (S)GD-like algorithms do not overfit

- small test error even without any regularization

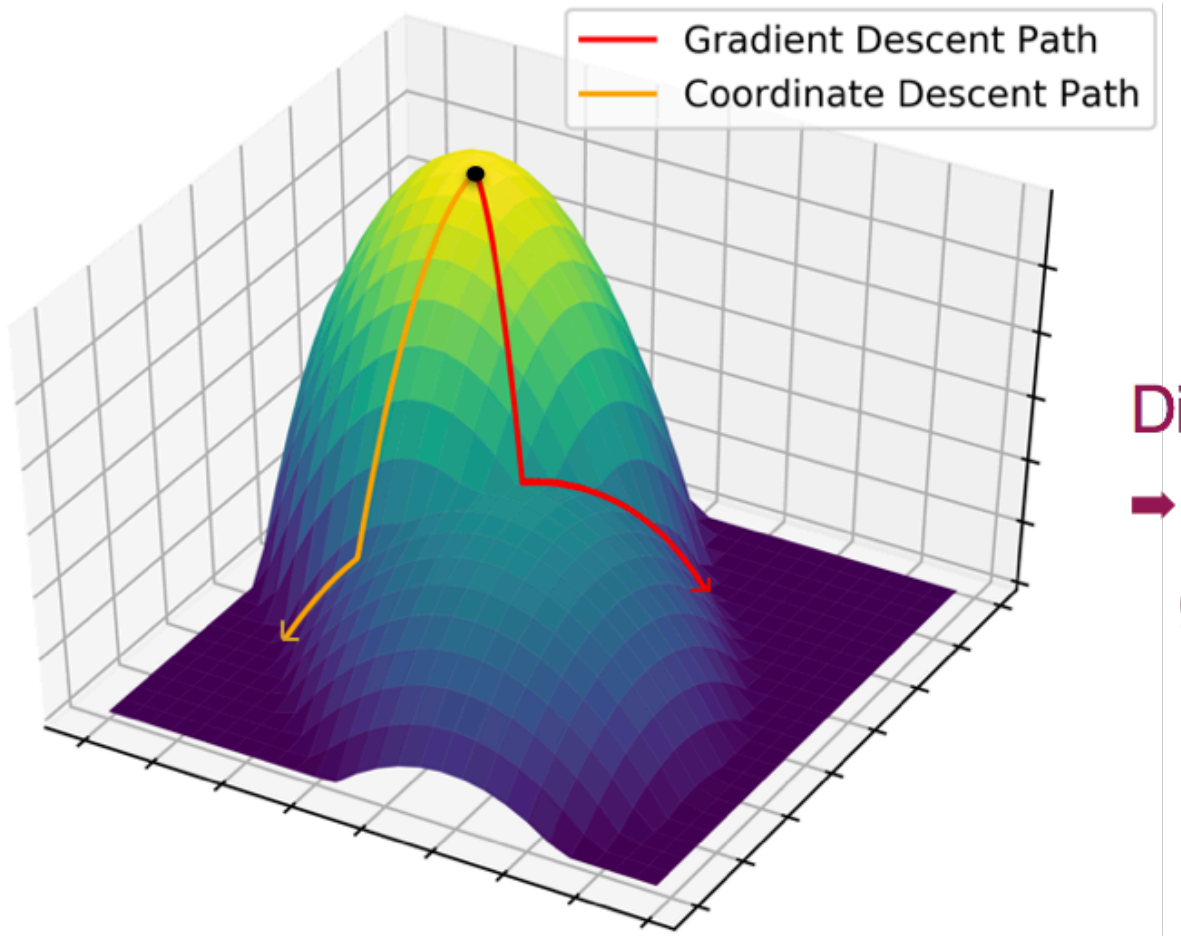
Example

Neyshabur, Tomioka, Srebro ICLR 2015



Implicit Regularization

Inductive bias induced by choice of optimization algorithm

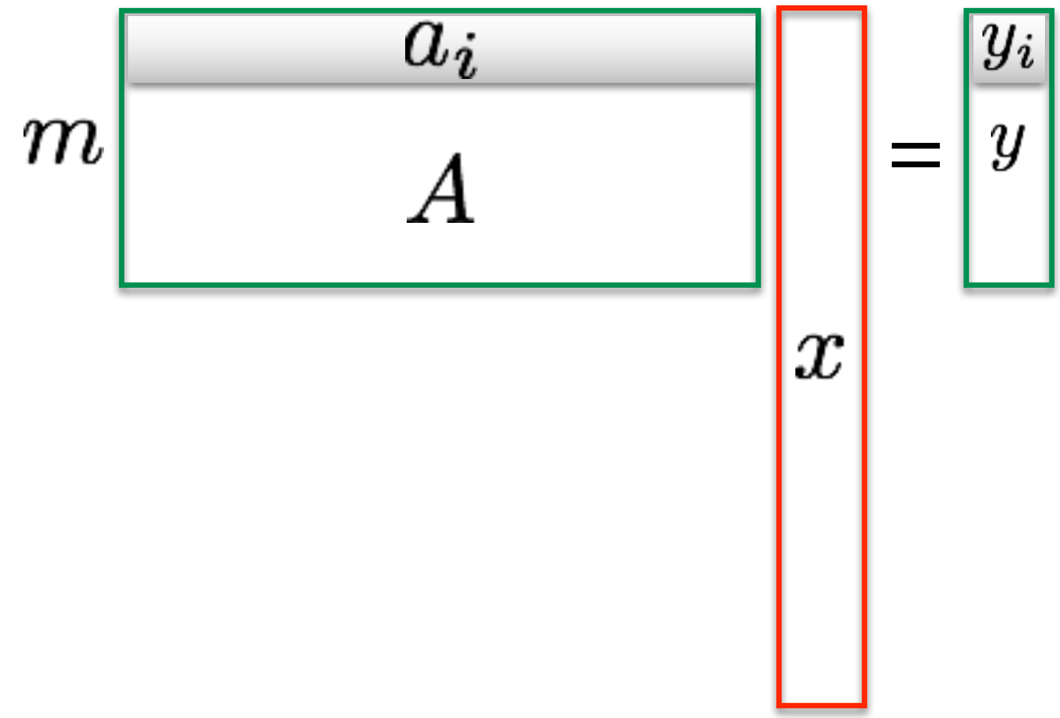


Different global optima
➡ different generalization (test error)

Choice of optimization algorithm ⟺ **Implicit bias towards "good" global optima**

Least squares

$$\min_x f(x) = \|Ax - y\|_2^2$$

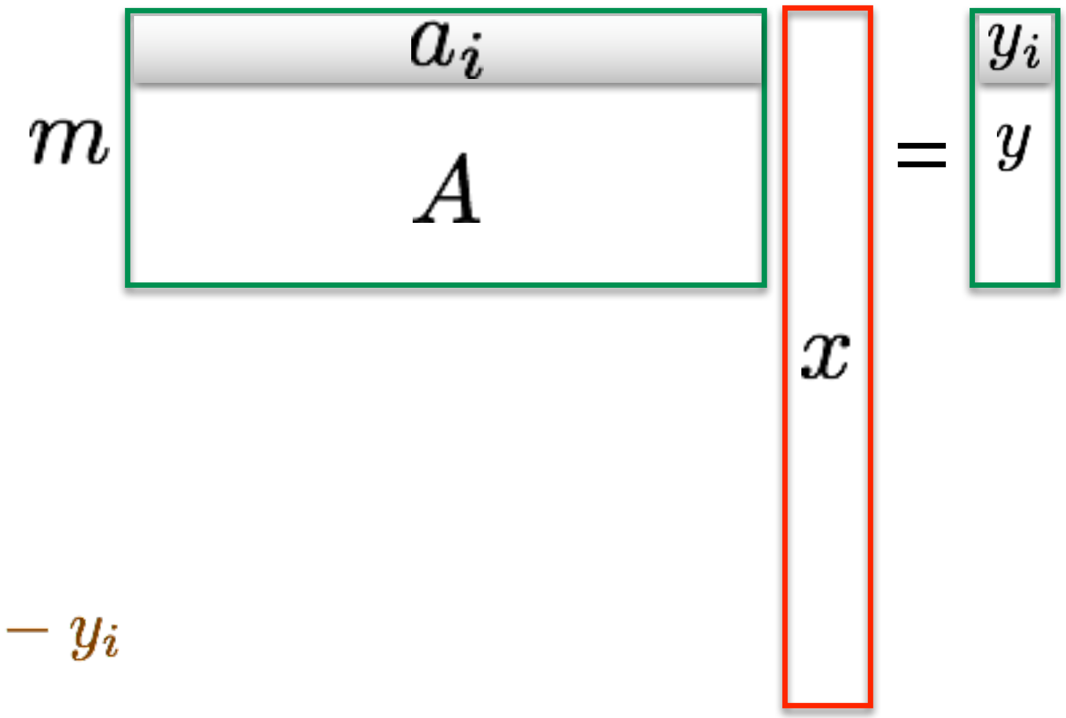


Gradient descent initialized at x_0

$$x_{t+1} = x_t - \eta \nabla_x f(x_t)$$

Least squares

$$\min_x f(x) = \|Ax - y\|_2^2$$



Gradient descent initialized at x_0

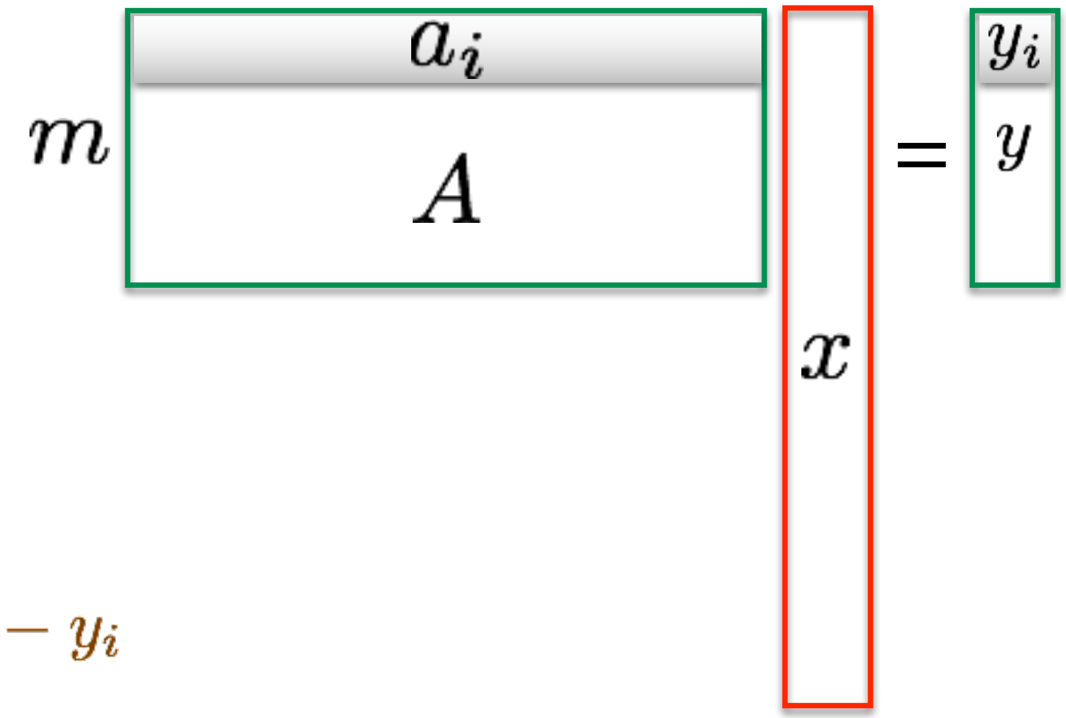
$$x_{t+1} = x_t - \eta \nabla_x f(x_t)$$

$$\nabla_x f(x) = A^\top (Ax - y) = \sum_{i=1}^{m} r_i a_i, \text{ where } r_i = a_i^\top x - y_i$$

Gradient based updates will only change component of x_0 along the span of a_i's

Least squares

$$\min_x f(x) = \|Ax - y\|_2^2$$



Gradient descent initialized at x_0

$$x_{t+1} = x_t - \eta \nabla_x f(x_t)$$

$$\nabla_x f(x) = A^\top (Ax - y) = \sum_{i=1}^{m} r_i a_i, \ \text{where } r_i = a_i^\top x - y_i$$

Gradient based updates will only change component of x_0 along the span of a_i's

$$\text{If} \quad x_0 = 0, \quad \text{then} \quad \widehat{x}_{(s)gd} = \underset{Ax=y}{\operatorname{argmin}} \|x\|_2$$

Goal

For DNNs, we want to characterize the “complexity” implicitly minimized by common optimization algorithms?

Previous work

- Generalization in terms of bounded norm DNNs. (Neyshabur et al. 2015, 2017)
- DNNs can fit random data. (Zhang et al. 2016)
- Comparing performance of adaptive optimization algorithms (Wilson et al. 2017)
- SGD biases towards "flat" global optima. (Hochreiter and Schmidhuber 1997; Keskar et al. 2017)

  → "flatness" not necessary for generalization (Dihn et al. 2017)

Why do we care about?

Guide choice of optimization

- Modified (S)GD and acceleration techniques for faster convergence and/or better generalization

New regularization techniques

- Deep learning still needs1000s of examples per class
- Explicit regularization for low "complexity"

Efficiency

- Can potentially train smaller networks more efficiently
  - reduce test time computation and memory requirements
  - Model compression

Low rank matrix estimation

$$\min_X F(X) = \|\mathcal{A}(X) - y\|_2^2, \quad \text{s.t. } \operatorname{rank}(X) \leq d$$

$\mathcal{A} : \mathbb{R}^{n \times n} \to \mathbb{R}^m$ $\qquad$ $y \in \mathbb{R}^m$

$\mathcal{A}(X)_i = \langle A_i, X \rangle$ for $i = 1, 2, \ldots m$

Matrix completion

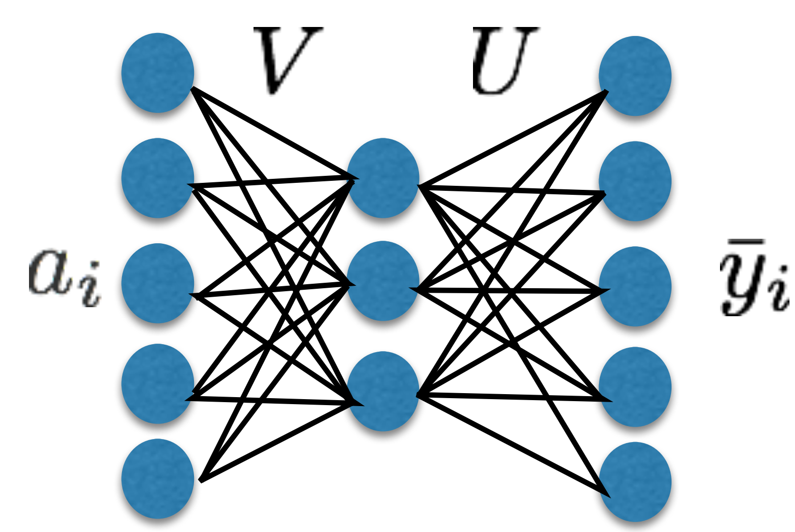
A_i has 1 on i^{th} observation, 0 otherwise

Linear networks

$X = UV$

$\bar{y}_i = X a_i$

$\mathcal{A} \equiv \{A_{ik} = a_i e_k^\top : i \in [m], k \in \dim(\bar{y})\}$



Matrix factorization

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

$d \sim m/n$ low rank matrix regression
often *unique global minima*

Jain, Netrapalli, Sanghavi 2012; … Bhojanapalli, Neyshabur, Srebro 2016 & Ge, Lee Ma 2016;

$d = n$ unconstrained problem,
equivalent to $\min_X F(X)$
easy to get global minima

What happens when f is optimized using gradient descent?

Gradient descent for matrix regression

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

$$U_{t+1} = U_t - \eta \mathcal{A}^*(r_t) U_t,$$

where $r_t = \mathcal{A}(U_t U_t^\top) - y, \mathcal{A}^*(r) = \sum_{i=1}^m r_i A_i$

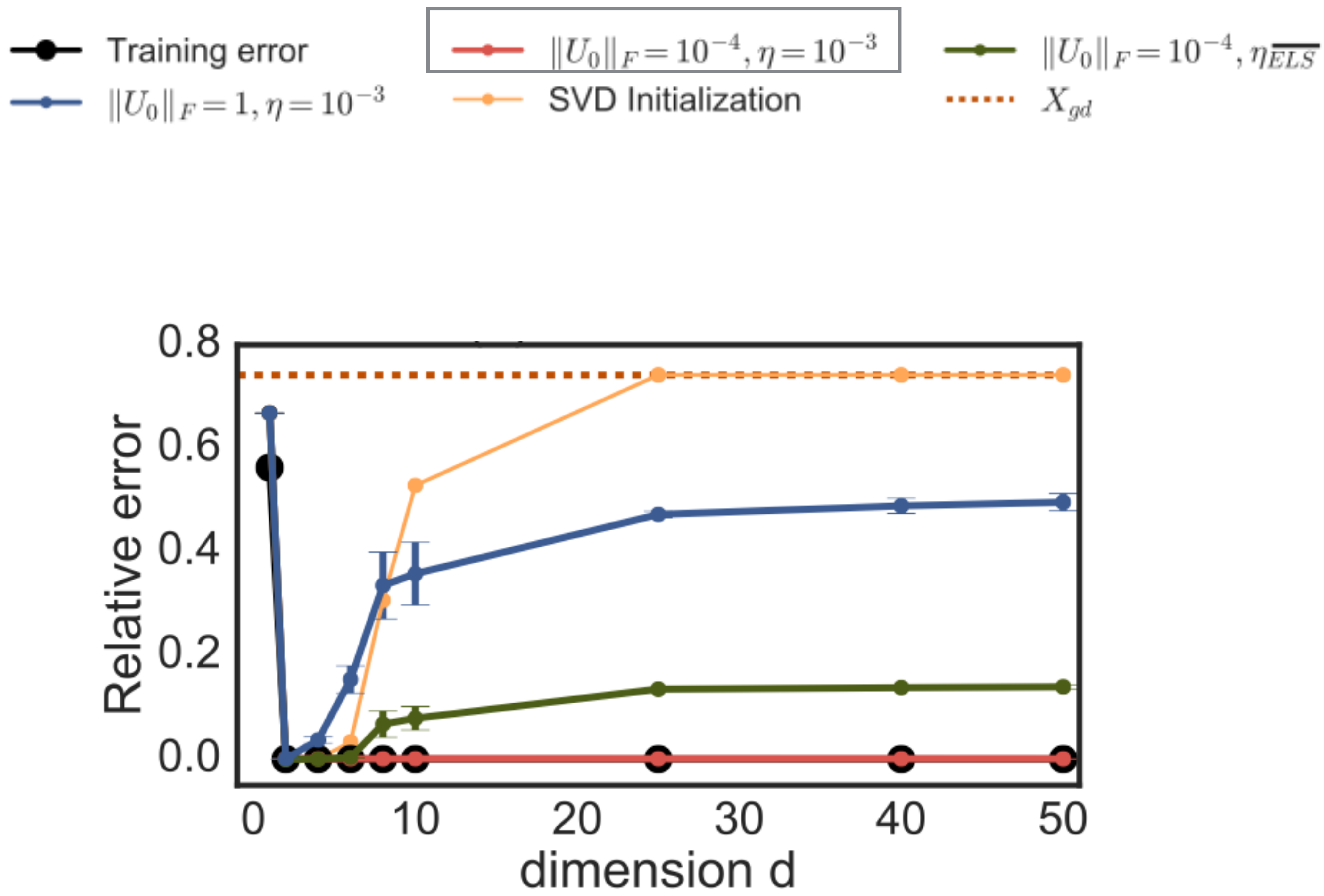
$n = 50, m = 300$

A_i symmetric random

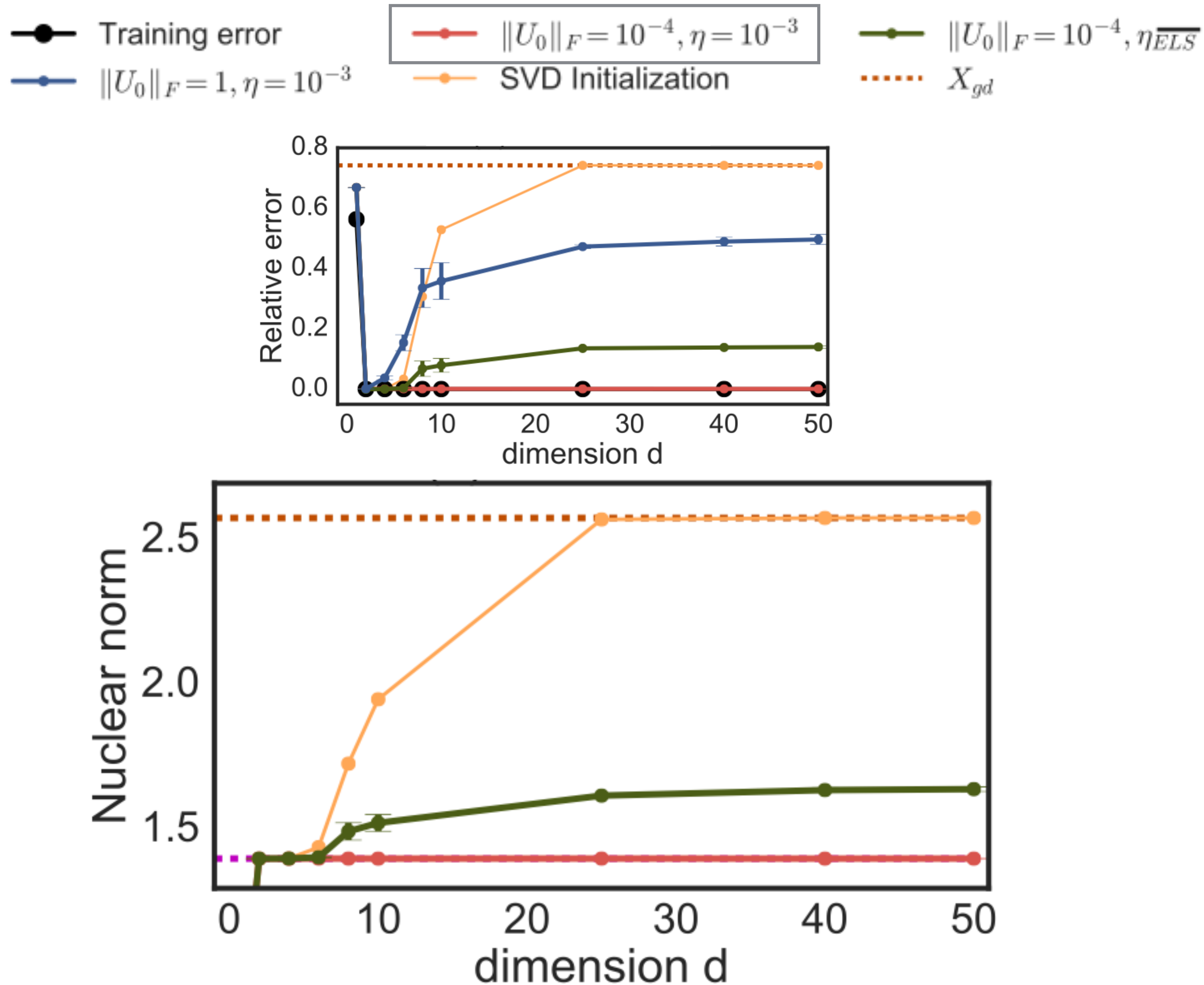
$y = \mathcal{A}(X^*)$ generated from ground truth
$X^* \succcurlyeq 0$ of rank-2

Test data: $\mathcal{A}_{\text{test}}$ and $y_{\text{test}} = \mathcal{A}_{\text{test}}(X^*)$
$\rightarrow$ low rank or low nuclear norm solutions generalize well

Gradient descent for matrix regression



Gradient descent for matrix regression



Gradient Flow

Gradient descent with infinitesimal step size

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

$$\boxed{\begin{array}{l} r_t = \mathcal{A}(U_t U_t^\top) - y \\ \mathcal{A}^*(r) = \sum_{i=1}^m r_i A_i \end{array}}$$

Gradient Flow

Gradient descent with infinitesimal step size

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

$$\boxed{\begin{array}{l} r_t = \mathcal{A}(U_t U_t^\top) - y \\ \mathcal{A}^*(r) = \sum_{i=1}^m r_i A_i \end{array}}$$

Induced dynamics on $X_t = U_t U_t^\top$

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

Independent of U for gradient flow
— not true for gradient descent

For $X_0 = X_{\text{init}},\ X_\infty(X_{\text{init}}) := \lim_{t\to\infty} X_t$

Gradient Flow

Gradient descent with infinitesimal step size

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

Induced dynamics on $X_t = U_t U_t^\top$

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

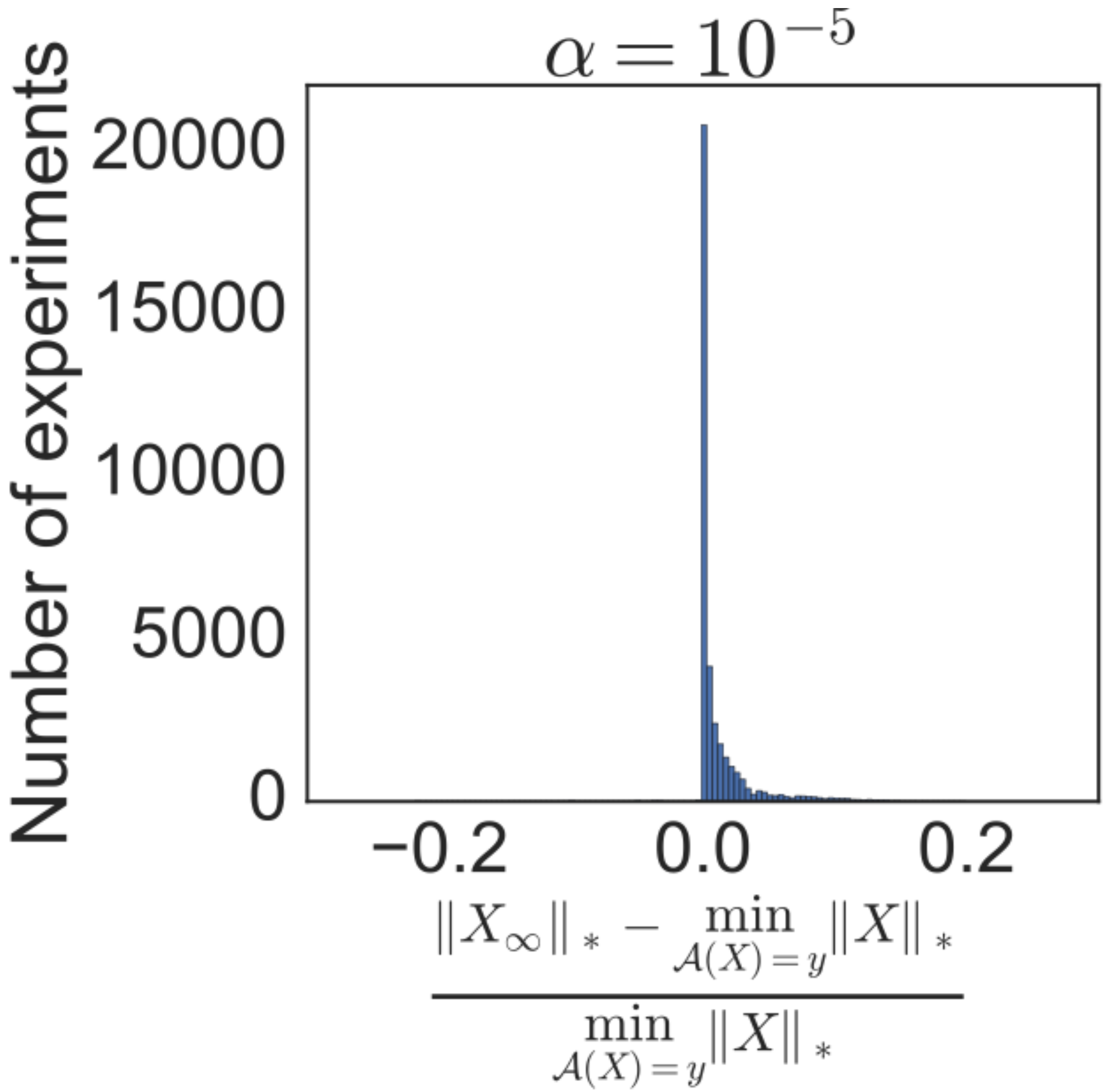
For $X_0 = X_{\text{init}}$, $X_\infty(X_{\text{init}}) := \lim_{t\to\infty} X_t$

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

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

Exhaustive Search

All PSD matrix completion problems
in 3x3 matrices with m=4



Commutative A_i

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

If $A_iA_j = A_jA_i$ for all $i, j \in [m]$, then
$X_t = e^{\mathcal{A}^*(s_t)}X_0e^{\mathcal{A}^*(s_t)}$ where $s_t = -\int_0^t r_t$

KKT conditions: $\mathcal{A}^*(\nu)X = X$ and $\lambda_{\max}(\mathcal{A}^*(\nu)) \leq 1$

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 $\bar{\mathbf{X}}_\infty = \mathbf{X}^*_{\mathrm{NN}}$

Commutative A_i

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

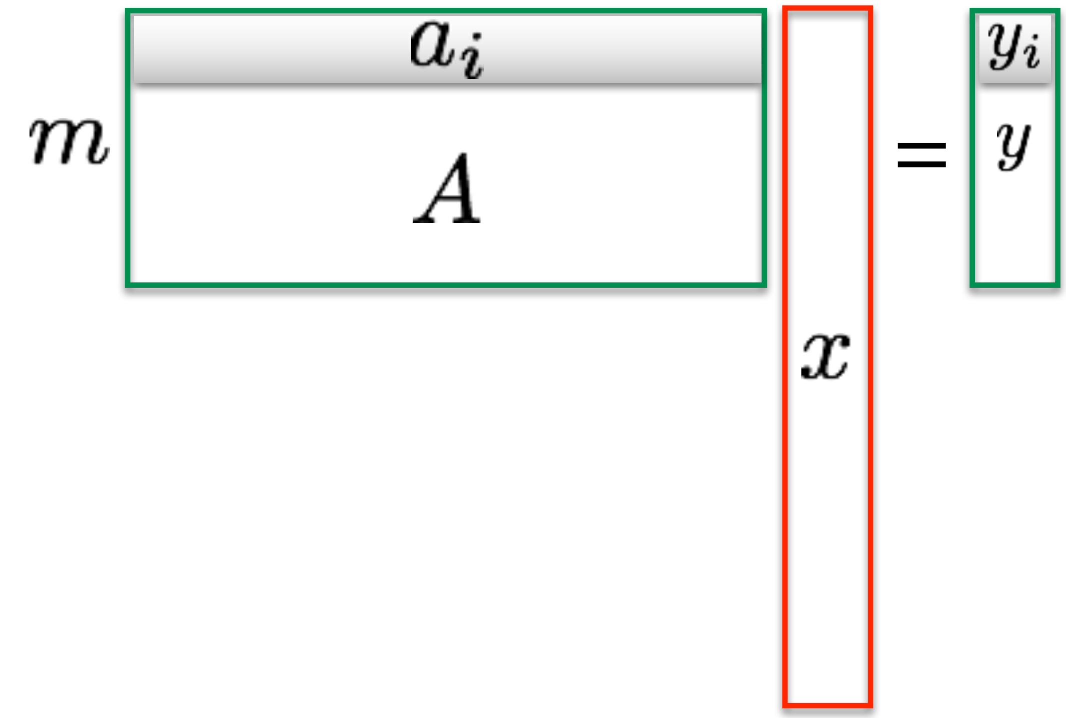
If $A_iA_j = A_jA_i$ for all $i, j \in [m]$, then
$X_t = e^{\mathcal{A}^*(s_t)}X_0\mathcal{A}^*(s_t)$ where $s_t = -\int_0^t r_t$

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_∞ such that $\mathbf{u_\infty^2} = \underset{\mathbf{Ax=b}}{\operatorname{argmin}} \|\mathbf{x}\|_\mathbf{1}$

Proof strategy: revisit Least squares

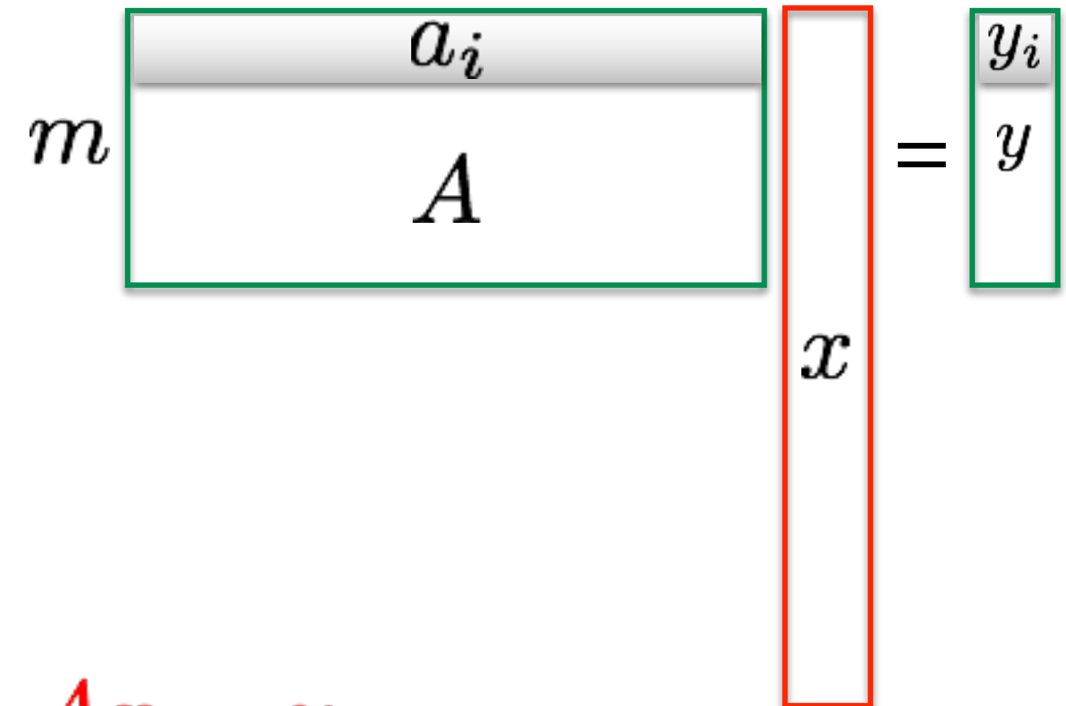
$$\min_x f(x) = \|Ax - y\|_2^2$$



$$\nabla_x f(x) = \sum_{i=1}^{m} r_i a_i \in \mathrm{span}\{a_i\}$$

Proof strategy: revisit Least squares

$$\min_x f(x) = \|Ax - y\|_2^2$$



$$\nabla_x f(x) = \sum_{i=1}^{m} r_i a_i \in \text{span}\{a_i\}$$

KKT conditions for $\min_x \|x\|_2^2$ s.t. $Ax = y$

$$Ax = y \qquad \exists \nu \in \mathbb{R}^m, x = A^\top \nu = \sum_i \nu_i a_i$$

✓ True for any global optima

✓ If $x_0 = \mathbf{0}$, true for any point on gradient descent path

Proof strategy

KKT conditions for $\min_{X \succcurlyeq 0} \|X\| * \text{ s.t. } \mathcal{A}(X) = y$

$\mathcal{A}(X) = y$ ✓ True for any global optima on gradient flow path as $X_t = U_t U_t^\top \succcurlyeq 0$

$X \succcurlyeq 0$

$$\boxed{\begin{gathered}\mathcal{A}^*(\nu) X = X \\ \mathcal{A}^*(\nu) \preccurlyeq I\end{gathered}}$$

Eigenvalues of $\mathcal{A}^*(\nu)$ are ≤ 1

✓ Columns of X spanned by eigenvectors of $\mathcal{A}^*(\nu)$ corresponding to eigenvalue 1

1. Characterize the 'set' of reachable points in gradient descent path
2. Characterize the asymptotic limit of the 'set'
3. Show KKT conditions for asymptotic reachable points

Gradient flow for single observation m=1

$$\dot{X}_t = -r_t(AX + XA) \propto -AX - XA$$

$$\Downarrow$$

$$X_t = e^{s_t A} X_0 e^{s_t A} \text{ where } s_t = \int_0^t r_t dt$$

KKT conditions: $\nu AX = X,\ \nu \leq 1/\lambda_{\max}(A)$

- As $X_0 \to 0$, if $y > 0$ and $\langle A, X_\infty(X_0) \rangle = y$, then $s_\infty(X_0) \to \infty$

Gradient flow for single observation m=1

$$\dot{X}_t = -r_t(AX + XA) \propto -AX - XA$$

$$\Downarrow$$

$$X_t = e^{s_t A} X_0 e^{s_t A} \text{ where } s_t = \int_0^t r_t dt$$

KKT conditions: $\nu AX = X,\ \nu \leq 1/\lambda_{\max}(A)$

- As $X_0 \to 0$, if $y > 0$ and $\langle A, X_\infty(X_0)\rangle = y$, then $s_\infty(X_0) \to \infty$
- $e^{s_\infty A} = \sum_{r=1}^n e^{s_\infty \lambda_r} \bar{a}_i \bar{a}_i = e^{s_\infty \lambda_{\max}} \sum_{r=1}^n e^{s_\infty(\lambda_r - \lambda_{\max})} \bar{a}_i \bar{a}_i$

Gradient flow for single observation m=1

$$\dot{X}_t = -r_t(AX + XA) \propto -AX - XA$$

$$\Downarrow$$

$$X_t = e^{s_t A} X_0 e^{s_t A} \text{ where } s_t = \int_0^t r_t dt$$

KKT conditions: $\nu AX = X,\ \nu \leq 1/\lambda_{\max}(A)$

- As $X_0 \to 0$, if $y > 0$ and $\langle A, X_\infty(X_0)\rangle = y$, then $s_\infty(X_0) \to \infty$
- $e^{s_\infty A} = \sum_{r=1}^n e^{s_\infty \lambda_r} \bar{a}_i \bar{a}_i = e^{s_\infty \lambda_{\max}} \sum_{r=1}^n e^{s_\infty(\lambda_r - \lambda_{\max})} \bar{a}_i \bar{a}_i$
- $X_\infty(X_0)$ spanned only by eigenvectors of $\lambda_{\max}(A)$
- $X_\infty(X_0) = \nu A X_\infty$ with $\nu = 1/\lambda_{\max}(A)$

Commutative A_i

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

If $A_i A_j = A_j A_i$ for all $i, j \in [m]$, then
$X_t = e^{\mathcal{A}^*(s_t)} X_0 e^{\mathcal{A}^*(s_t)}$ where $s_t = -\int_0^t r_t$

- Every point on gradient flow lies in $\mathcal{M} = \{\alpha e^{\mathcal{A}^*(s)} : s \in \mathbb{R}^m\}$
- Complementary slackness and dual feasibility hold for any point on $\mathcal{M}_\infty = \{\alpha e^{\mathcal{A}^*(s)} : \alpha \to 0, \|s\|_2 \to \infty\}$
- Proof does not depend on particular choice of $r_t = \mathcal{A}(X_t) - y$
- $\mathcal{M}$ suggests infinitesimal stepwise necessary
- Result holds for infinitesimal SGD, but not accelerated gradient versions.

Non-commutative A_i

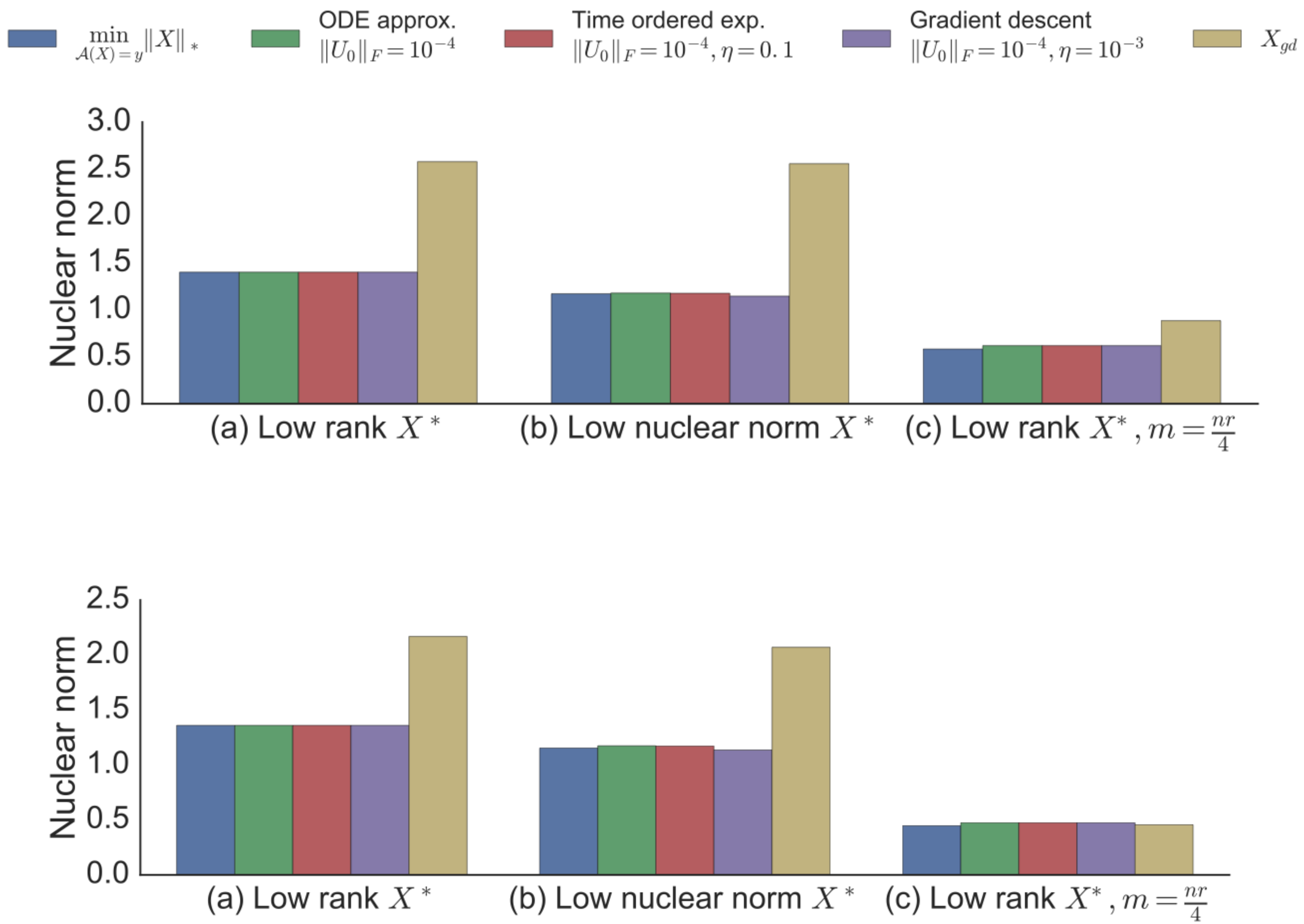
$$\dot{X}_t = -\mathcal{A}^*(r_t) X_t - X_t \mathcal{A}^*(r_t)$$

bilinear dynamical system with r_t as 'control' variables

$$X_t = \left(\lim_{\epsilon \to 0} \prod_{k=t/\epsilon}^{0} e^{-\epsilon \mathcal{A}^*(r_t)} \right) X_0 \left(\lim_{\epsilon \to 0} \prod_{0}^{k=t/\epsilon} e^{-\epsilon \mathcal{A}^*(r_t)} \right)$$

- with arbitrary r_t (not necessarily residuals from gradient flow), can reach any psd matrix even for m=2
- exploit structure of r_t to show: asymptotically, only components of leading eigenvectors of $\mathcal{A}^*(\nu)$ remain
  - empirically seems to hold for r_t from gradient flow
  - empirically also holds for random r_t

Summary of matrix reconstruction results



Conclusion

- Optimization algorithms traditionally studied as a tools to minimize training objective
- It is becoming increasingly evident that choice of optimization algorithm also influences generalization performance
- Such implicit regularization is potentially a key component for understanding learning in complex models
- Matrix factorization is the simplest non-convex problem where such properties can be analyzed