Inference via Alternating Minimization

Sujay Sanghavi

Electrical and Computer Engg.
University of Texas, Austin

Mathematics used to be about finding the best arrow to hit your target …

… nowadays, a lot of it is about painting the best target around your arrow.

- H. Narayanan (Prof., IIT Bombay)

This talk ..

Three problems of object recovery with missing information:

- Matrix completion (STOC 2013)
- Phase Recovery (NIPS 2013)
- Mixed Linear Regression (preprint)

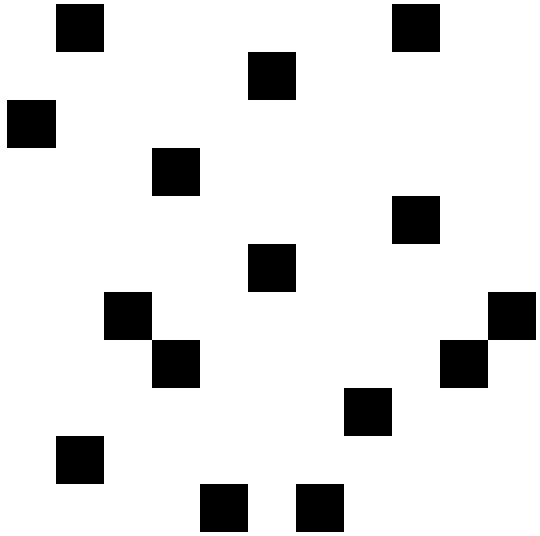
"One" algorithmic approach: Alternating Minimization

- problems naturally formulated as non-convex optimization
- AltMin: hold one set of variables, optimize over the other, alternate
- super fast, widely applied

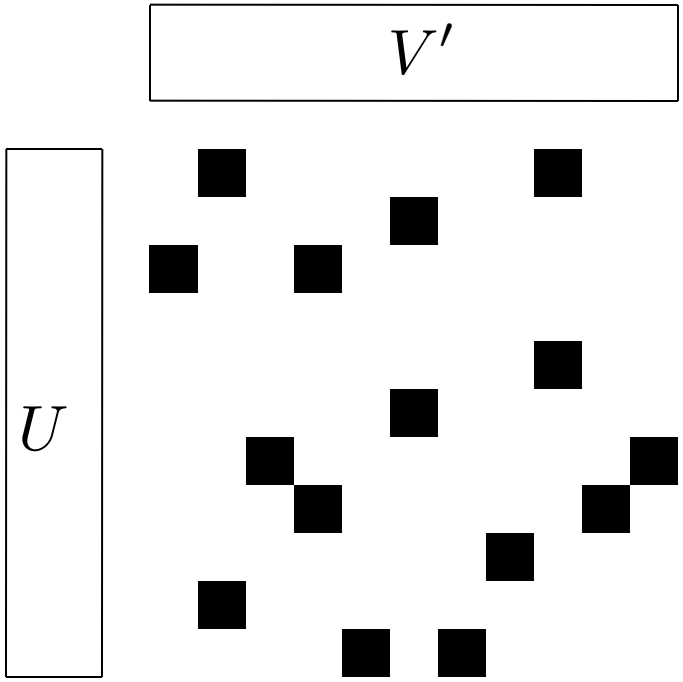
Our work: the first guarantees of statistical performance

Problem 1: Matrix Completion

Find a low-rank matrix from a few (randomly sampled) elements



Problem 1: Matrix Completion



Empirically popular approach:

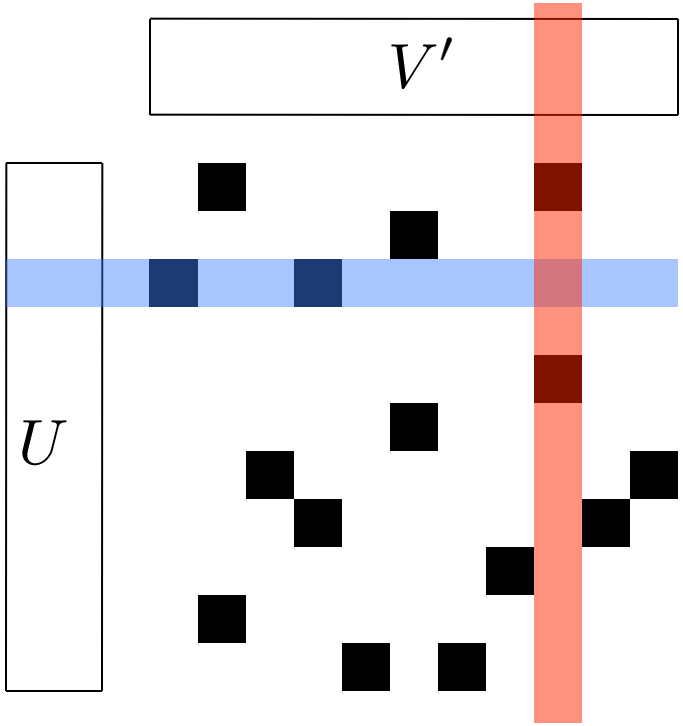
(1) Write as non-convex problem

$$\min_{U,V} \|\mathcal{P}_\Omega(M - UV')\|_F$$

(2) Alternately optimize U and V
(from random initialization)

Part of the BellKor winning entry of the Netflix prize.

AltMin for Matrix Completion



Naturally decouples into small least-squares problems

(a) For all i

$$u_i \leftarrow \min_u \sum_{j:(i,j)\in\Omega} (m_{ij} - \langle u, v_j \rangle)^2$$

(b) For all j

$$v_j \leftarrow \min_v \sum_{i:(i,j)\in\Omega} (m_{ij} - \langle u_i, v \rangle)^2$$

No theoretical guarantees on exact/approximate recovery

Matrix Completion

[Candes, Recht '08] : First method with any rigorous guarantees on recovery

- based on convex optimization over $n \times n$ matrices

$$\min_X \quad \|X\|_*$$

$$s.t. \quad P_\Omega(X) = P_\Omega(M)$$

Input and output : $\widetilde{O}(nr)$

But this needs $\widetilde{O}(n^2)$ memory (and computation) !

Theorem [CR,08] (and several others since):

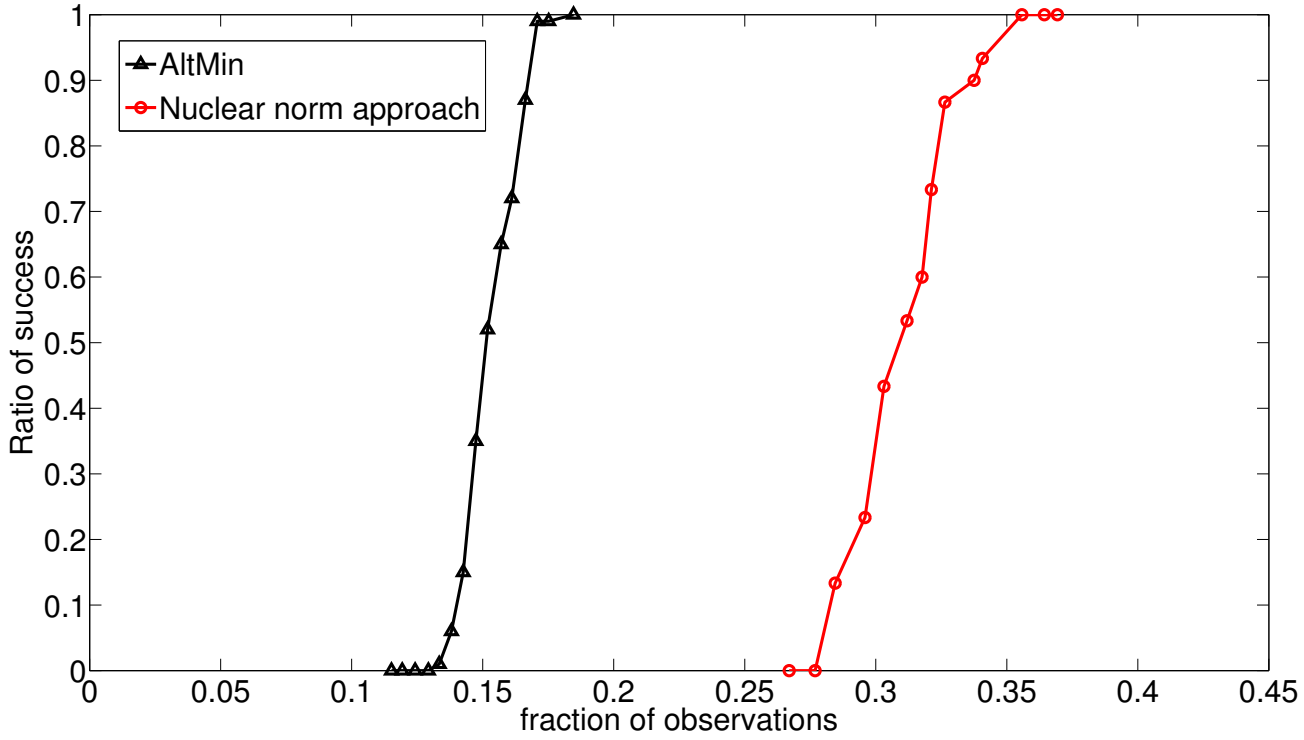
Random samples Ω + incoherent matrix M $\Rightarrow$ exact recovery

See also: [Keshavan, Montanari, Oh] – SVD + gradient descent on grassman manifold

Matrix Completion

Surprisingly: AltMin seems to need **fewer samples** than trace-norm minimization

(empirically)



Problem 2: Phase recovery

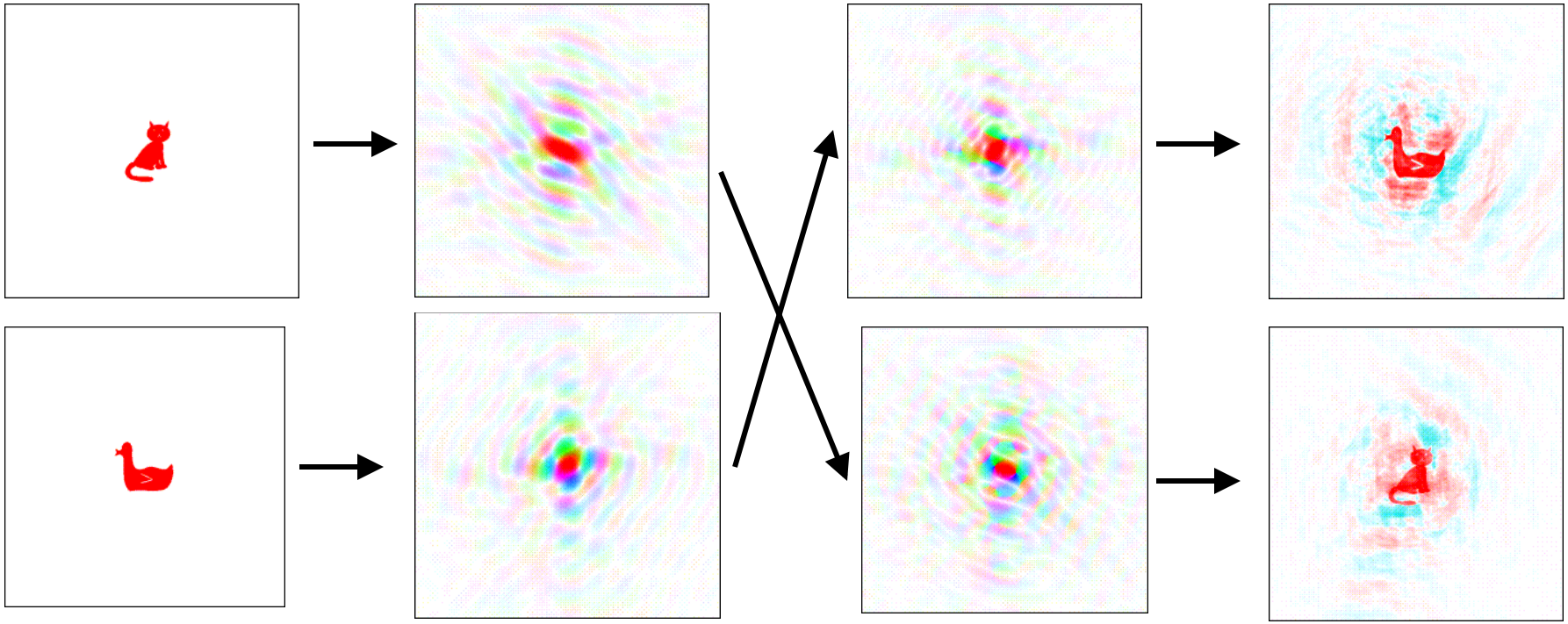
Recover complex vector given only **magnitudes** of linear eq.s

$$y = |Ax^*|$$

recover $x^* \in \mathbb{C}^n$ from A and y

<an abstraction of>
Application: diffraction imaging
(e.g. in crystallography)

Problem 2: Phase recovery



FFT

Switch phases
Keep magnitudes

IFFT

Phases contain crucial information …

Problem 2: Phase recovery

Recover complex vector given only **magnitudes** of linear eq.s

$$y = |Ax^*|$$

recover $x^* \in \mathbb{C}^n$ from A and y

<an abstraction of>
Application: diffraction imaging
(e.g. in crystallography)

Empirically popular approach [Gerchberg-Saxton '72], [Fineup '80] etc.

(1) Write as non-convex problem

$$\min_{C,x} \|Cy - Ax\|_2$$

Diagonal matrix of phases

(2) Alternately optimize over x and C starting from random initialization.

AltMin for Phase Recovery

(a) Solve a least-squares problem $x \leftarrow \arg\min_x \|Cy - Ax\|_2$

(b) Record the resulting phases $c_{ii} \leftarrow Ph(\langle a_i, x \rangle)$

*This is nothing but **Expectation-Maximization (EM)** for the noiseless case*

Phase Recovery

[Candes, Strohmer, Voroninski '12] etc. : lifting + SDP relaxation of rank

$$\min_X \quad tr(X)$$

$$s.t. \quad a_i' X a_i = y_i^2$$

$$X \quad psd$$

Makes an $O(n)$ problem into

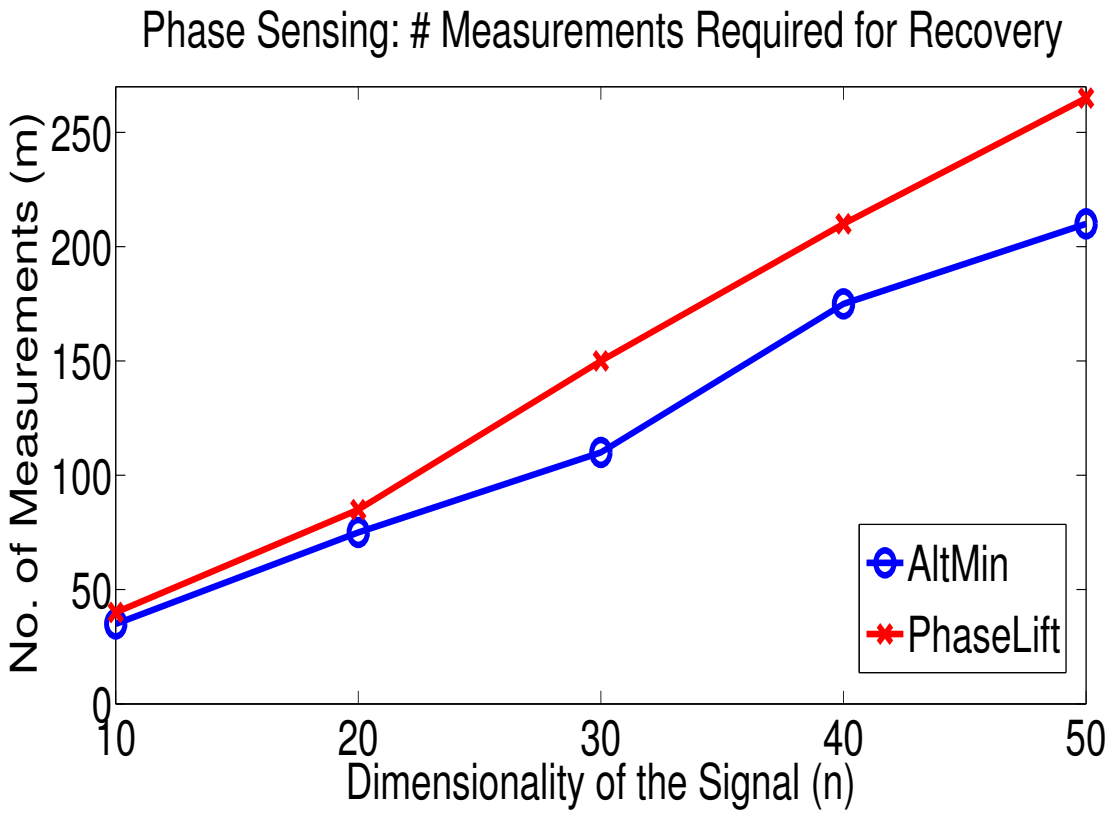
an $O(n^2)$ problem

Theorem [CSV'12], [CL'13]:

If $a_i \sim \mathcal{CN}(0, I)$ then $\widehat{X} = x^*(x^*)'$ whp, from $O(n)$ samples

See also: [Waldspurger, d'Aspermont, Mallat] for alternate convex formulation

AltMin for Phase Recovery



Again, lower number of samples than convex methods

(empirically)

Problem 3: Mixed linear regression

Solve linear equations, except that each is either

$$y_i = \langle x_i, \beta_0^* \rangle \qquad \text{or} \qquad y_i = \langle x_i, \beta_1^* \rangle$$

Find β_1^*, β_0^* given $\{y_i, x_i\}$

Natural for settings where linear prediction / modeling with latent classes

- Evolutionary biology: separating out mutant behavior / expression
- Quantitative Finance: detecting regime change
- Healthcare: separating patient classes for differential treatment

….

Several specialized R packages (see [Grun,Leisch] for overview)

- all implement variants / optimizations of EM

Mixed Linear Regression

… my netflix problem …

Problem 3: Mixed linear regression

Solve linear equations, except that each is either

$$y_i = \langle x_i, \beta_0^* \rangle \qquad \text{or} \qquad y_i = \langle x_i, \beta_1^* \rangle$$

Find β_1^*, β_0^* given $\{y_i, x_i\}$

Only existing algorithm: Expectation Maximization (EM)

= AltMin on the non-convex problem

$$\min_{\beta_1, \beta_0} \sum_i \min_{z_i \in \{0,1\}} \left(y_i - z_i \langle x_i, \beta_1 \rangle + (1 - z_i) \langle x_i, \beta_0 \rangle\right)^2$$

… starting from random initialization.

No theoretical guarantees for any method, in any setting.

Mixed Linear Regression

(a) Assign labels to the samples, based on current estimates $\widehat{\beta}_1, \widehat{\beta}_0$

$$\widehat{z}_i = 1 \quad \Leftrightarrow \quad (y_i - \langle x_i, \widehat{\beta}_1 \rangle)^2 < (y_i - \langle x_i, \widehat{\beta}_0 \rangle)^2$$

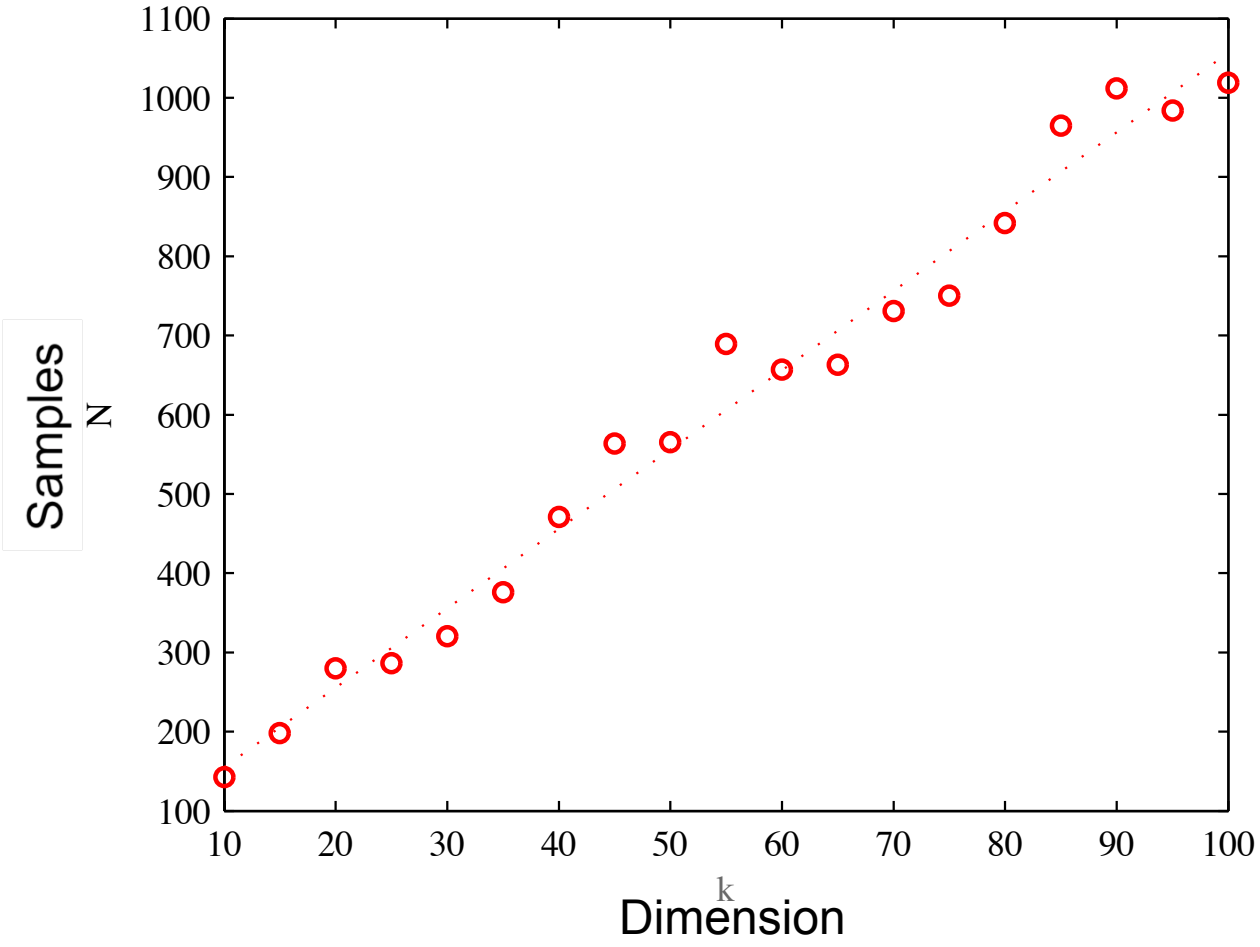
$$\widehat{z}_i = 0 \qquad \text{else}$$

(b) Update estimates using new labels

$$\widehat{\beta}_1 \leftarrow \arg\min_{\beta} \sum_{i:z_i=1} (y_i - \langle x_i, \beta \rangle)^2$$

Mixed Linear Regresion

Synthetic experiment with isotropic gaussian samples



(empirically)
Number of samples scale linearly with dimension

(for EM with our initialization)

The story so far

Three problems: matrix completion, phase retrieval, mixed linear regression

Empirically: Best methods involve AltMin on natural non-convex formulation

No statistical guarantees on consistent recovery, in any setting

Methods with statistical guarantees: convex optimization

(under statistical assumptions) establish consistent recovery

Slower, involve optimization in higher dimensions than warranted by the data or output

(often) need more samples than non-convex methods,

Our motivation

Is it possible to obtain statistical guarantees for AltMin algorithms that work in the dimension specified by the input and output ?

Equivalently:

Does the fact that convex methods have statistical consistency represent a genuine algorithmic advance ?

Or is it just that the statistical setting is “easy” enough for faster methods as well ?

Our Results

Statistical guarantees for exact recovery:

Global convergence + statistical consistency for AltMin

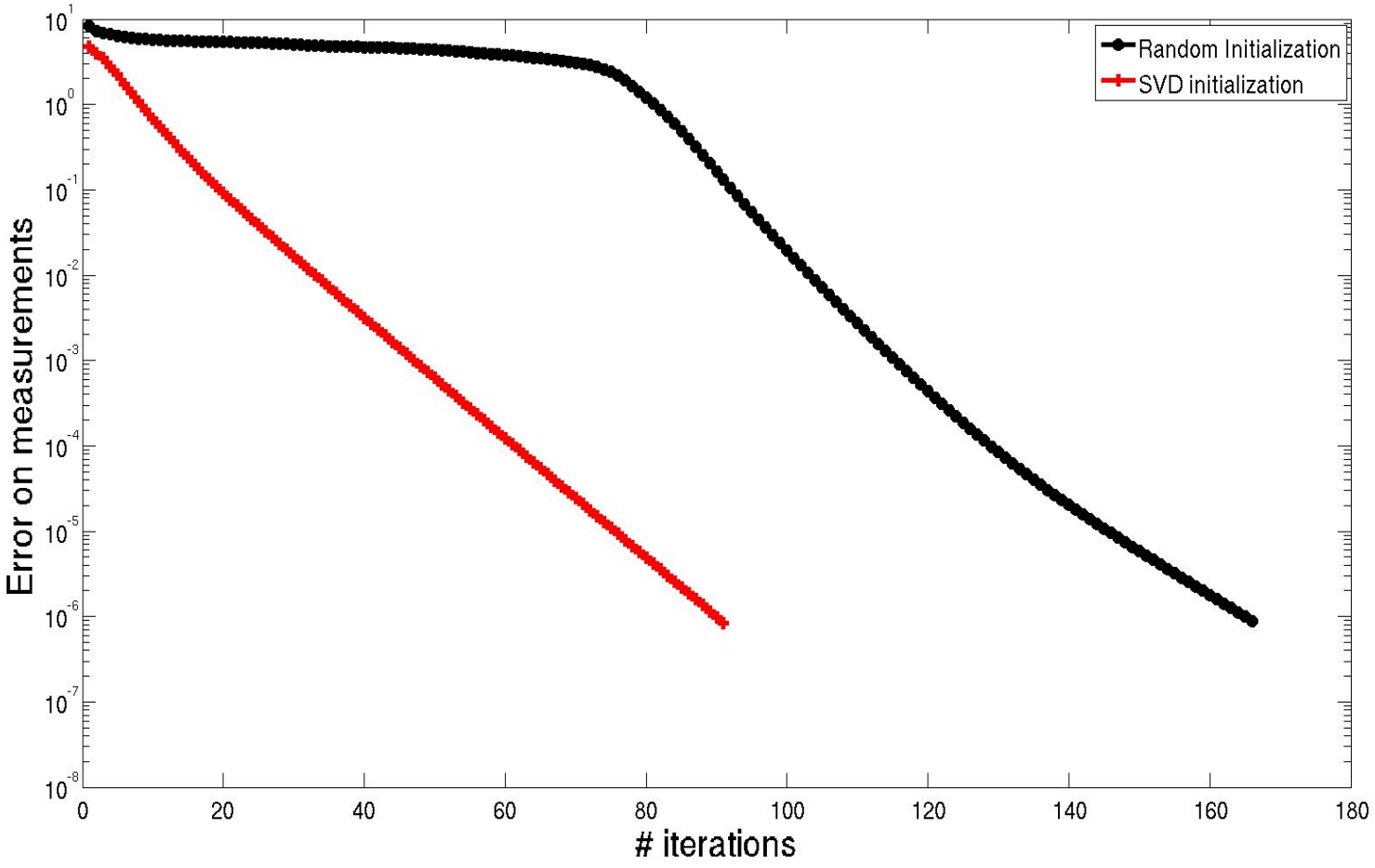
… in the standard settings

Two key components of our analysis:

Initialization:
via leading eigenvector(s) of appropriate matrix

Re-sampling: (analytical trick)
as a work-around to vexing dependency issues

Initialization



Phase Recovery: Initialization

Problem: solve $y = |Ax|$ i.e. equations $y_i = |\langle a_i, x \rangle| \quad , \; i = 1, \ldots, N$

Make a matrix $M = \frac{1}{N} \sum_i y_i^2 \, a_i a_i'$ $\qquad a_i \sim \mathcal{CN}(0, I)$

Key observation: as $N \to \infty$ the top eigenvector of $M \to x^*$

$$M = \frac{1}{N} \sum_i |\langle a_i, x^* \rangle| \, a_i a_i'$$

$$\to \; I + 2x^*(x^*)'$$

Phase Recovery: Initialization

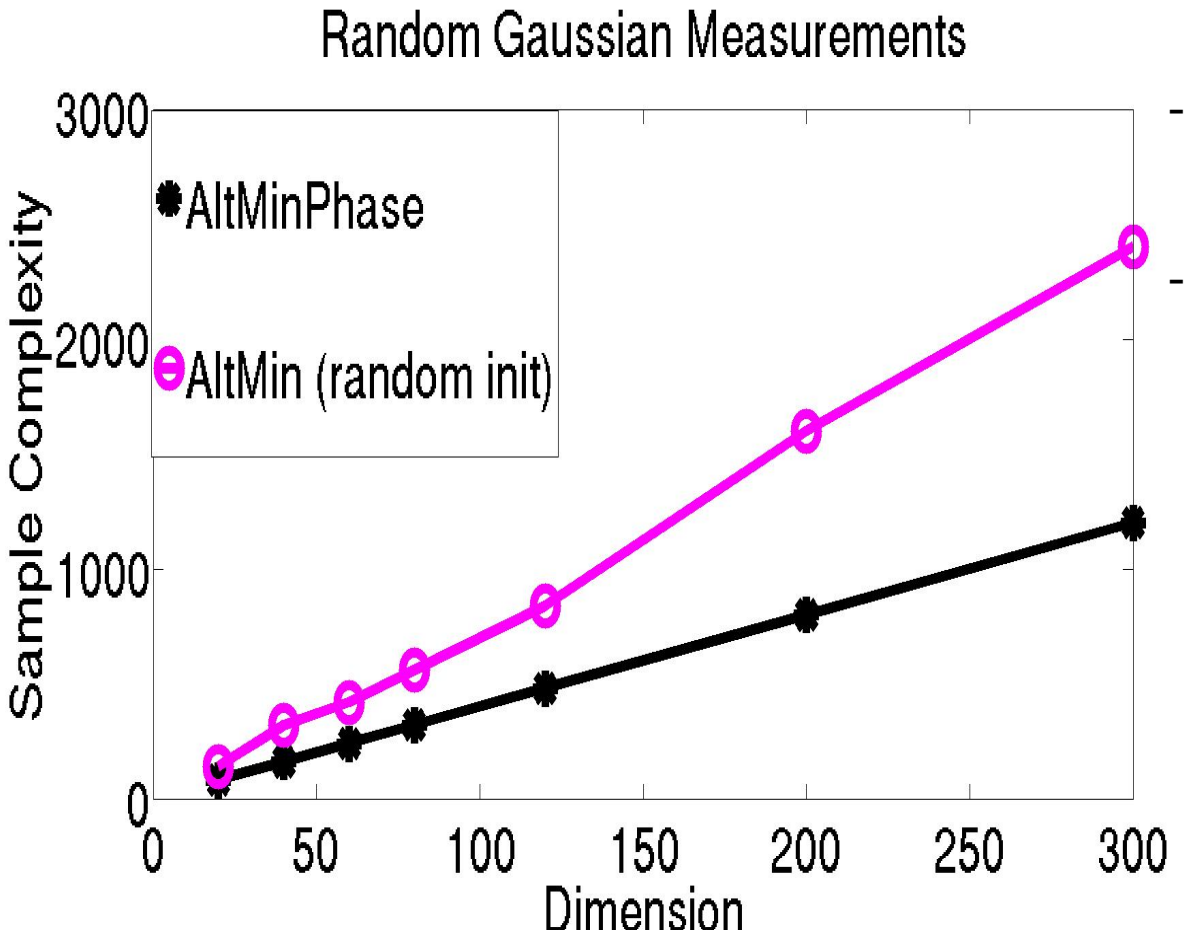
Given N samples, $x^{(0)} \leftarrow$ top eigenvector of $M = \frac{1}{N} \sum_i y_i^2 \, a_i a_i'$

Lemma: with $N = \frac{C}{\epsilon^2} n \log^2 n$ samples, can get $\|x^{(0)} - x^*\| < \epsilon$

Asymptotically consistent, but slow $O(1/\epsilon^2)$ convergence

Not satisfactory by itself, but useful for initialization

Effects of Initialization



- allows for geometric convergence
- **and** reduces the number of samples required

(and is crucial for analysis)

Initializations …

Matrix Completion:

$U_0 \leftarrow$ top left singular vectors of 0-filled matrix M_Ω

Theorem: $N = cr^{2.5} n \log n$ samples for constant distance

Mixed linear equations: $y_i = \langle x_i, \beta_1 \rangle$ or $y_i = \langle x_i, \beta_2 \rangle$

$\beta_1^{(0)}, \beta_2^{(0)} \leftarrow$ Top two eigenvectors of $M = \frac{1}{N} \sum_i y_i^2 \, x_i x_i'$

Theorem: $N = cn \log^2 n$ samples for constant distance

All three: **convergence to truth requires too many samples**. So use only for init.

Re-sampling

Empirically: use all samples in every iteration

- after initialization, geometric decay of error observed

Analysis of this is hard

- because concentration results require independence between samples and current iterate.

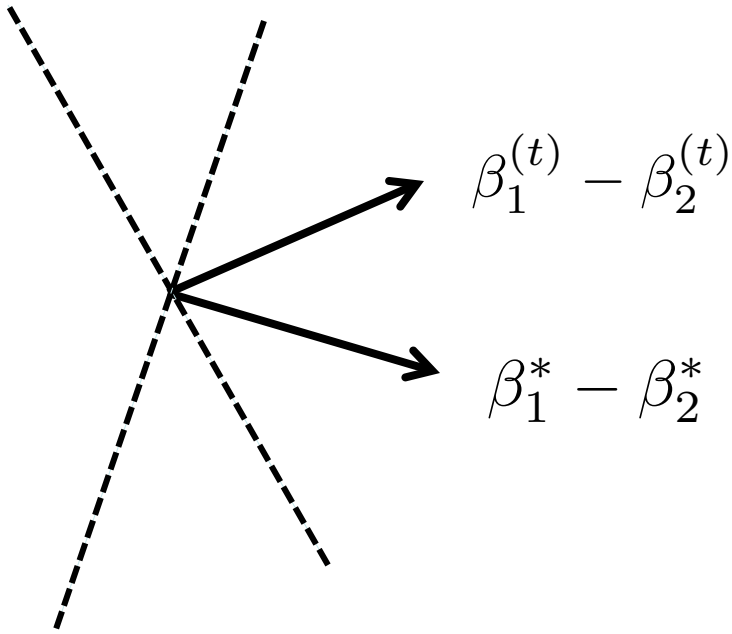
"Solution" : use fresh samples in every iteration

- making them independent of current iterate
- by pre-partitioning the given samples

Example: Mixed Linear Equations

Intuition: current iterate $\beta_1^{(t)}, \beta_2^{(t)}$ truth β_1^*, β_2^*

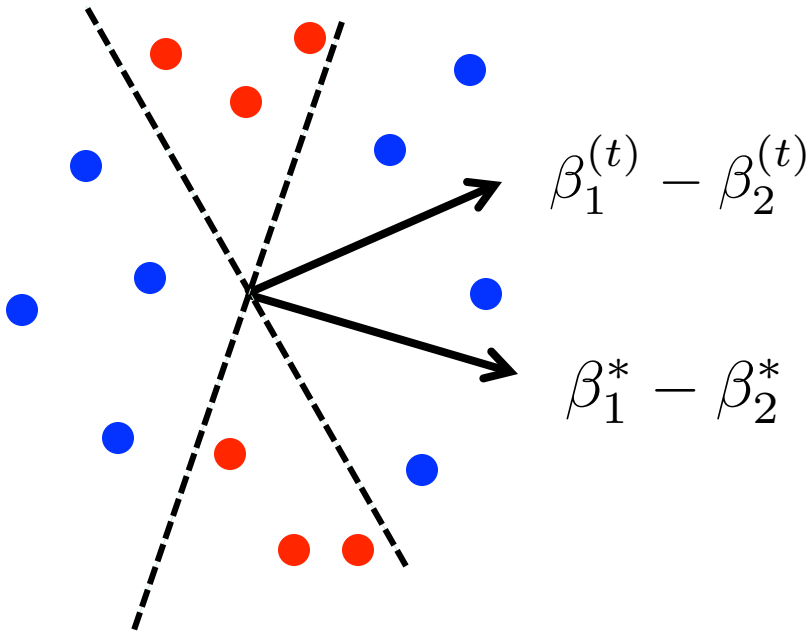
samples $\{y_i, x_i\}$ $x_i \sim \mathcal{N}(0, I)$



Example: Mixed Linear Equations

Intuition: current iterate β_1, β_2 truth β_1^*, β_2^*

samples $\{y_i, x_i\}$ $x_i \sim \mathcal{N}(0, I)$



If β_1, β_2 not too far from β_1^*, β_2^*

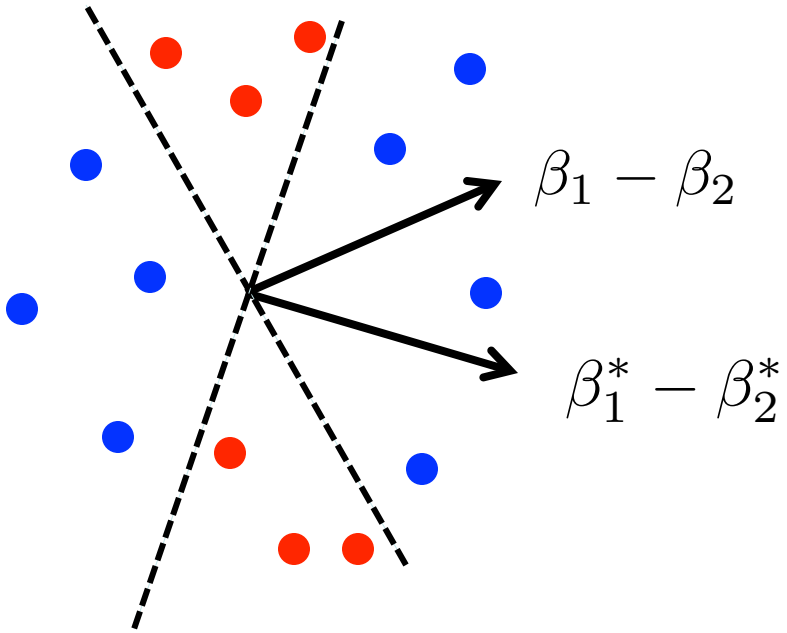
Then **majority** points will be correctly assigned.

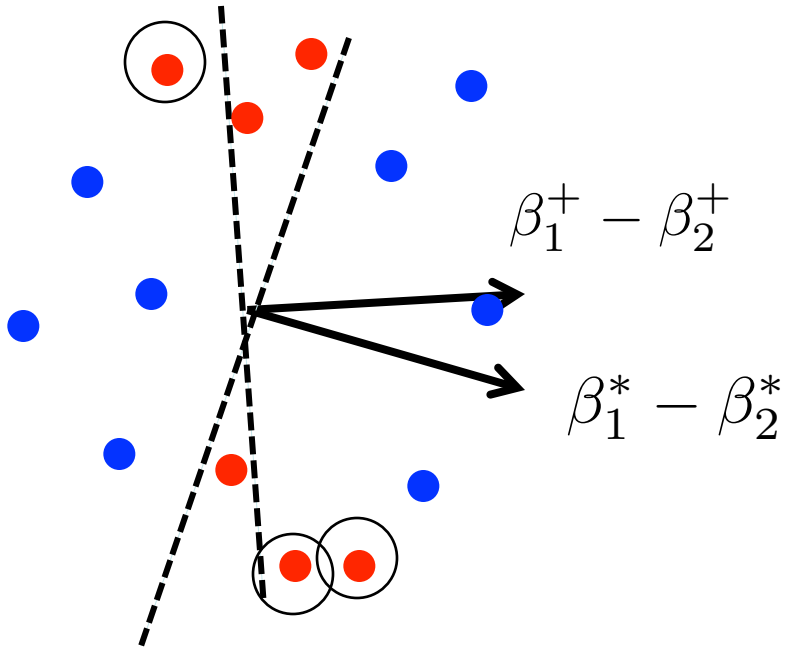
So, running least-squares on these will yield better next iterate.

Example: Mixed Linear Equations

Intuition: current iterate β_1, β_2 truth β_1^*, β_2^*

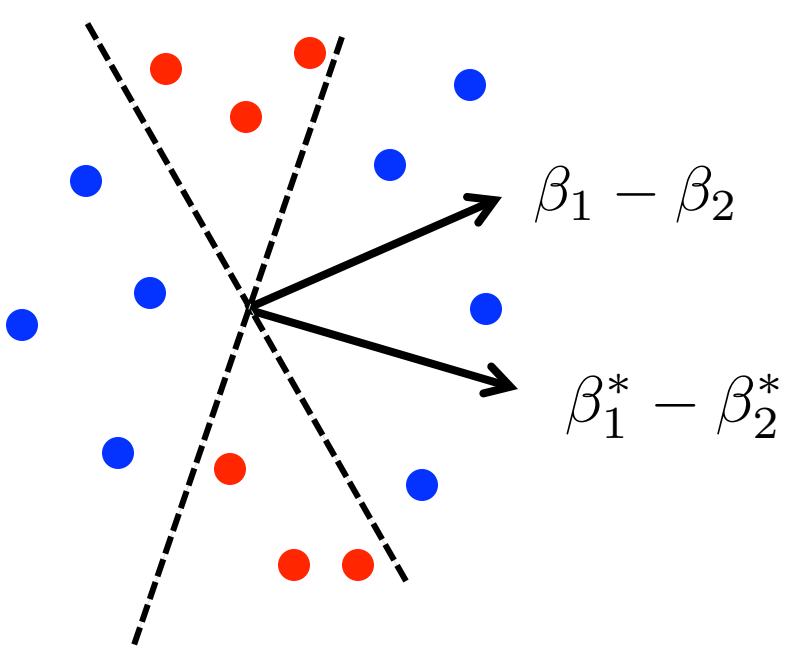
samples $\{y_i, x_i\}$ $x_i \sim \mathcal{N}(0, I)$

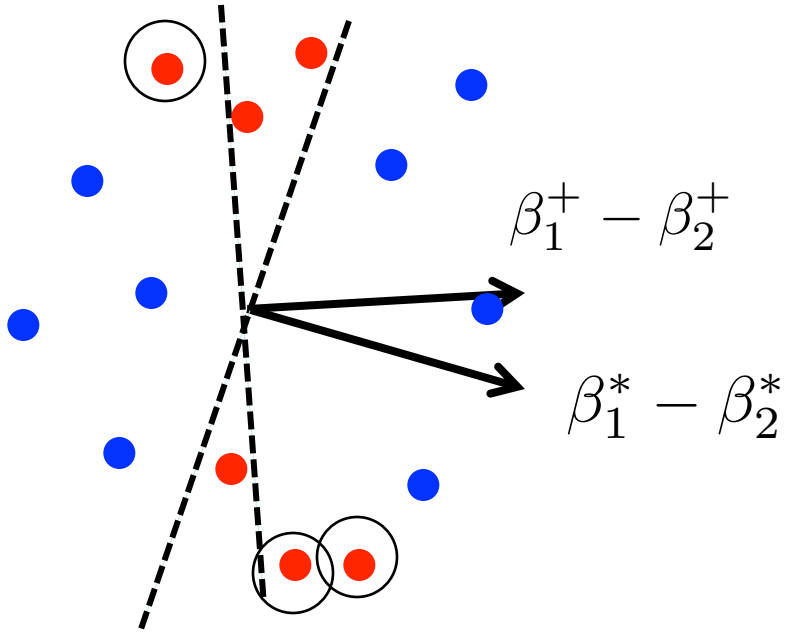




… which will give fewer error samples …

Example: Mixed Linear Equations





Analysis of this is hard because, after first step, **samples are dependent** the β s

Idea: make them independent, by re-sampling at every iteration

Resampling

Resampling == forcing independence between samples and estimate

by modifying the algorithm to use fresh samples in every iteration.

Matrix completion: new elements in every iteration

Phase Recovery: new measurements

Mixed linear equations: new samples

Note: seems to be NOT needed empirically.

Proving this is the case would be very interesting.

Mixed Linear Regression

Theorem: [Yi, Caramanis, Sanghavi '13]

If the current iterate satisfies $\|\beta_i - \beta_i^*\| < c\|\beta_1^* - \beta_2^*\|$

and we use new, independent samples, then the new error satisfies

$$\|\beta_i^+ - \beta_i^*\| < \frac{1}{2}\|\beta_i - \beta_i^*\|$$

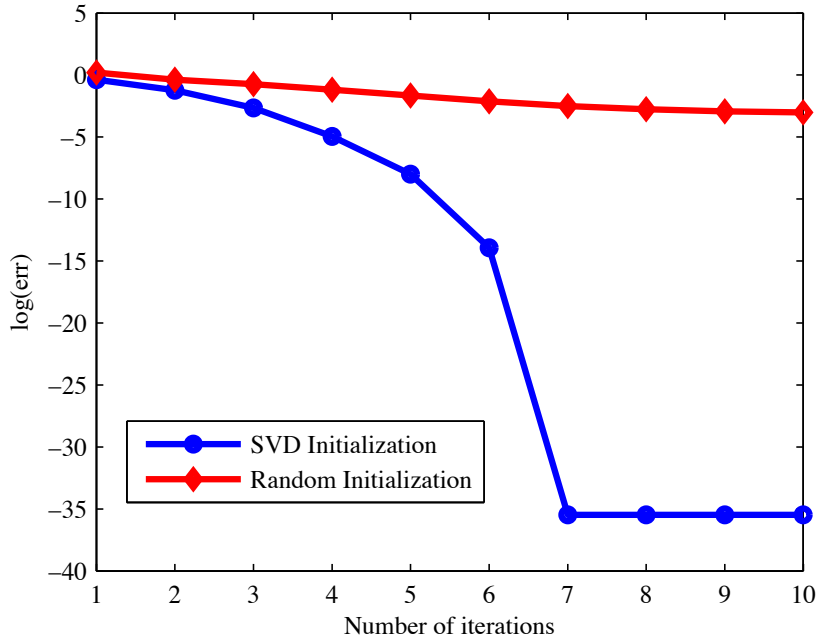
provided the number of samples is greater than $\dfrac{n}{\min\{p_1, p_2\}}$

Similar results (i.e. halving of error in each step) for matrix completion and phase retrieval.

Iterations with Re-sampling

So: **geometric decay in the error** – halving in every step.
- better than rate of convergence for (the non-smooth) convex methods

Good initialization is crucial to showing this decay in error



But: need $\log(1/\epsilon)$ extra samples for accuracy of $\|\beta_i - \beta_i^*\| < \epsilon$

Summary

Practice: **+** fast **+** low number of samples - no guarantees

Theory: **+** statistical guarantees under assumptions
- slower **-** (often) more samples

Our work:

An (imperfect, but first) attempt to bridge this divide, via two key ideas
- **Initialization** – important both empirically and for proof
- **Resampling** – only for proof, not important empirically

We show: AltMin with these two works under **similar statistical assumptions**

Future:

Vast number of applications where EM etc. are the most popular methods

Removal of the re-sampling requirement (?)

Conclusion

Papers: (also on my website)

Matrix completion: STOC 2013
w/ Praneeth Netrapalli and Prateek Jain

Phase Retrieval: NIPS 2013
w/ Praneeth Netrapalli and Prateek Jain

Mixed Linear Regression: preprint
w/ Xinyang Yi and Constantine Caramanis

Thanks !

Mixed Linear Regression: Initialization

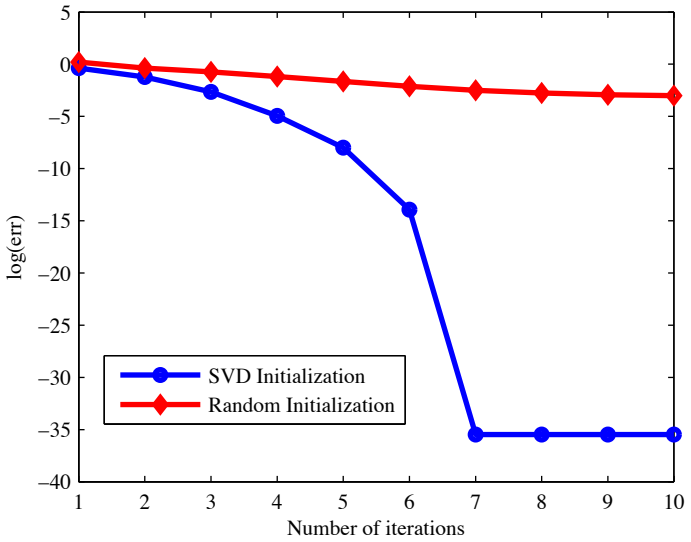
Recall: $y_i = \langle x_i, \beta_0^* \rangle$ or $y_i = \langle x_i, \beta_1^* \rangle$

Make matrix $M := \frac{1}{N} \sum_i y_i^2 x_i x_i'$

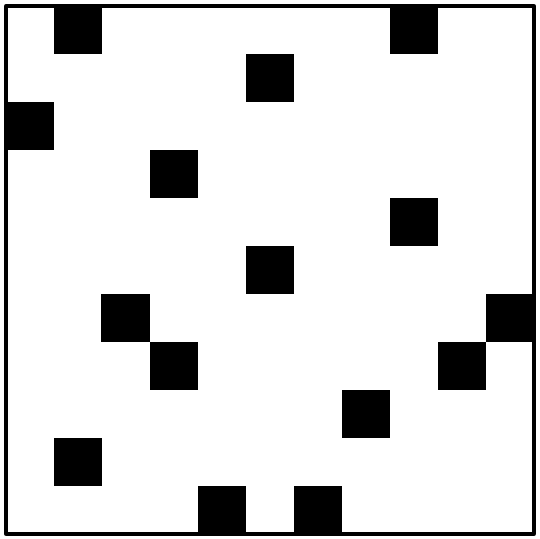
Then, $M \rightarrow I + 2\beta_1^*(\beta_1^*)' + 2\beta_2^*(\beta_2^*)'$

So, even with finite samples, **top-two eigenspace of M** a good approximation to $span(\beta_1^* \beta_2^*)$

Our initialization: a 1-d grid search for the βs in this space

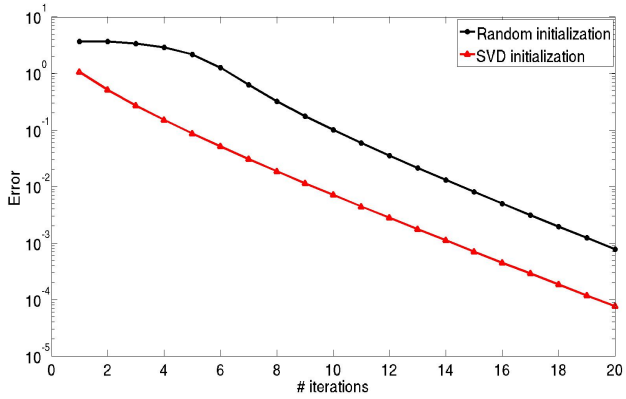


Matrix Completion: Initialization



Consider the 0-filled matrix $P_\Omega(M)$

$U^{(0)} \leftarrow$ Top r-dimensional column space of $P_\Omega(M)$



Lemma [Jain, Netrapalli, Sanghavi]:

We have $dist(U^{(0)}, U^*) \leq \frac{1}{2}$ if the matrix is μ-incoherent and number of samples is $\Omega(\kappa^4 \mu^2 r^{2.5} n \log n)$

Condition number