1-Bit Matrix Completion

Mark A. Davenport

School of Electrical and Computer Engineering
Georgia Institute of Technology

Yaniv Plan  Mary Wootters  Ewout van den Berg
Matrix Completion

- When is it possible to recover the original matrix?
- How can we do this efficiently?
- How many samples will we need?
Low-Rank Matrices

Singular value decomposition:

\[ M = U \Sigma V^* \]

\( \approx dr \ll d^2 \)

degrees of freedom
Collaborative Filtering

The “Netflix Problem”

\[ M_{i,j} = \text{how much user } i \text{ likes movie } j \]

Rank 1 model: \( u_i = \text{how much user } i \text{ likes romantic movies} \)

\[ v_j = \text{amount of romance in movie } j \]

\[ M_{i,j} = u_i v_j \]

Rank 2 model: \( w_i = \text{how much user } i \text{ likes zombie movies} \)

\[ x_j = \text{amount of zombies in movie } j \]

\[ M_{i,j} = u_i v_j + w_i x_j \]
Beyond Netflix

- Recovery of incomplete survey data
- Analysis of voting data
- Sensor localization
- Quantum state tomography
- ...

Low-Rank Matrix Recovery

Given:
- a $d \times d$ matrix $M$ of rank $r$
- samples of $M$ on the set $X$ : $Y = M$

How can we recover $M$?

$$\hat{M} = \arg \inf_{X: X = Y} \text{rank}(X)$$

Can we replace this with something computationally feasible?
Convex relaxation!

Replace $\text{rank}(X)$ with $\|X\|_* = \sum_{j=1}^{d} |\sigma_j|$

$$\hat{M} = \arg\inf_{X: X = Y} \|X\|_*$$

If $|\cdot| = O(r d \log d)$, under certain assumptions, this procedure can recover $\hat{M}$!
Matrix Completion in Practice

- **Noise**
  \[ Y = (M + Z) \]

- **Quantization**
  - Netflix: Ratings are integers between 1 and 5
  - Survey responses: True/False, Yes/No, Agree/Disagree
  - Voting data: Yea/Nay
  - Quantum state tomography: Binary outcomes

Extreme quantization *destroys low-rank structure*
1-Bit Matrix Completion

Extreme case

\[ Y = \text{sign}(M) \]

Claim: Recovering \( M \) from \( Y \) is impossible!

\[ M = \begin{bmatrix} \lambda & \lambda & \lambda & \lambda \\ \lambda & \lambda & \lambda & \lambda \\ \lambda & \lambda & \lambda & \lambda \\ \lambda & \lambda & \lambda & \lambda \end{bmatrix} \]

No matter how many samples we obtain, all we can learn is whether \( \lambda > 0 \) or \( \lambda < 0 \)
1-Bit Matrix Completion

Extreme case

\[ Y = \text{sign}(M) \]

Claim: Recovering \( M \) from \( Y \) is impossible!

\[ M = uv^* \]

\[ \tilde{u} = \text{sign}(u) \quad \tilde{v} = \text{sign}(v) \]

\[ \tilde{M} = \tilde{u}\tilde{v}^* \]

\[ \text{sign}(\tilde{M}) = \text{sign}(M) \]
If we consider a noisy version of the problem, recovery becomes feasible!

\[ Y = \text{sign}(M + Z) \]

Fraction of positive/negative observations tells us something about \( \lambda \)

Example of the power of \textit{dithering}
Observation Model

For \((i, j) \in \mathbb{Z}^2\) we observe

\[
Y_{i,j} = \begin{cases} 
+1 & \text{with probability } f(M_{i,j}) \\
-1 & \text{with probability } 1 - f(M_{i,j})
\end{cases}
\]

If \(f\) behaves like a CDF, then this is equivalent to

\[
Y_{i,j} = \text{sign}(M_{i,j} + Z_{i,j})
\]

where \(Z_{i,j}\) is drawn according to a suitable distribution.

We will assume that \(Z_{i,j}\) is drawn uniformly at random.
Examples

- Logistic regression / Logistic noise
  
  \[ f(x) = \frac{e^x}{1 + e^x} \]

  \[ Z_{i,j} \sim \text{logistic distribution} \]

- Probit regression / Gaussian noise
  
  \[ f(x) = \Phi(x/\sigma) \]

  \[ Z_{i,j} \sim \mathcal{N}(0, \sigma^2) \]
Assumptions

\[ M = \begin{bmatrix} d & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \]

- If the upper-left corner of \( M \) is not sampled, we have no information

- Solution: Assume that \( M \) is “spread”

\[
\frac{1}{d \alpha} \|M\|_* \leq \sqrt{r}
\]

\[
\|M\|_\infty = \max_{i,j} |M_{i,j}| \leq \alpha \approx O(1)
\]
Maximum Likelihood Estimation

Log-likelihood function:

\[ F(X) = \sum_{(i,j) \in +} \log(f(X_{i,j})) + \sum_{(i,j) \in -} \log(1 - f(X_{i,j})) \]

\[ \hat{M} = \arg\max_X F(X) \]

s.t. \( \text{rank}(X) \leq r \)
Maximum Likelihood Estimation

Log-likelihood function:

\[ F(X) = \sum_{(i,j)\in +} \log(f(X_{i,j})) + \sum_{(i,j)\in -} \log(1 - f(X_{i,j})) \]

\[ \hat{M} = \arg \max_x F(X) \]

s.t. \[ \frac{1}{d\alpha} \|X\|_* \leq \sqrt{r} \]

\[ \|X\|_\infty \leq \alpha \]
Recovery of the Matrix

**Theorem (Upper bound achieved by convex ML estimator)**

Assume that $\frac{1}{d^\alpha} \| M \|_* \leq \sqrt{r}$ and $\| M \|_\infty \leq \alpha$. If $\| E \|_1 = m > d \log d$, then with high probability

$$\frac{1}{d^2} \| \widehat{M} - M \|_F^2 \leq C \alpha L_\alpha \beta_\alpha \sqrt{\frac{r d}{m}}$$

where

$$L_\alpha := \sup_{|x| \leq \alpha} \frac{|f'(x)|}{f(x)(1 - f(x))} \quad \beta_\alpha := \sup_{|x| \leq \alpha} \frac{f(x)(1 - f(x))}{(f'(x))^2}$$

Is this bound tight?
Recovery of the Matrix

**Theorem (Upper bound achieved by convex ML estimator)**

Assume that $\frac{1}{d^\alpha} \|M\|_* \leq \sqrt{r}$ and $\|M\|_\infty \leq \alpha$. If $M$ is chosen at random with $\mathbb{E}|M| = m > d \log d$, then with high probability

$$\frac{1}{d^2} \|\hat{M} - M\|_F^2 \leq C\alpha L_\alpha \beta_\alpha \sqrt{\frac{rd}{m}}$$

**Theorem (Lower bound on any estimator)**

There exist $M$ satisfying the assumptions above such that for any set $\mathcal{E}$ with $|\mathcal{E}| = m$, we have (under mild technical assumptions) that

$$\inf_{\hat{M}} \mathbb{E} \left[ \frac{1}{d^2} \|\hat{M} - M\|_F^2 \right] \geq c\alpha \sqrt{\beta \frac{3}{4} \alpha} \sqrt{\frac{rd}{m}}$$
Logistic Model

\[ L_\alpha = 1 \quad \beta_\alpha \approx e^\alpha \]

**Theorem (Upper bound achieved by convex ML estimator)**

\[
\frac{1}{d^2} \| \hat{M} - M \|_F^2 \leq C \alpha e^\alpha \sqrt{\frac{rd}{m}}
\]

**Theorem (Lower bound on any estimator)**

\[
\inf_{\hat{M}} \mathbb{E} \left[ \frac{1}{d^2} \| \hat{M} - M \|_F^2 \right] \geq c \alpha e^{\frac{3}{8} \alpha} \sqrt{\frac{rd}{m}}
\]
Probit Model

\[ L_\alpha \approx \frac{\alpha}{\sigma} + 1 \quad \beta_\alpha \approx \sigma^2 e^{\alpha^2/2\sigma^2} \]

**Theorem (Upper bound achieved by convex ML estimator)**

\[
\frac{1}{d^2} \| \hat{M} - M \|_F^2 \leq C \left( \frac{\alpha}{\sigma} + 1 \right) e^{\alpha^2/2\sigma^2} \sigma \alpha \sqrt{\frac{rd}{m}}
\]

Two regimes

- High signal-to-noise ratio: \( \sigma \leq \alpha \)
- Low signal-to-noise ratio: \( \sigma \geq \alpha \)

Compare to how well we can estimate \( M \) from unquantized, noisy measurements
Probit Model (High SNR)

**Theorem** *(Upper bound achieved by convex ML estimator)*

\[
\frac{1}{d^2} \| \hat{M} - M \|_F^2 \leq C \alpha^2 e^{\alpha^2/2\sigma^2} \sqrt{\frac{rd}{m}}
\]

**Theorem** *(Lower bound on any estimator with unquantized measurements)*

\[
\inf_{\hat{M}} \mathbb{E} \left[ \frac{1}{d^2} \| \hat{M} - M \|_F^2 \right] \geq c\alpha\sigma \sqrt{\frac{rd}{m}}
\]
Probit Model (Low SNR)

**Theorem** (*Upper bound achieved by convex ML estimator*)

\[
\frac{1}{d^2} \| \hat{M} - M \|_F^2 \leq C\alpha\sigma \sqrt{\frac{rd}{m}}
\]

**Theorem** (*Lower bound on any estimator with unquantized measurements*)

\[
\inf_{\hat{M}} \mathbb{E} \left[ \frac{1}{d^2} \| \hat{M} - M \|_F^2 \right] \geq c\alpha\sigma \sqrt{\frac{rd}{m}}
\]

More noise can lead to *improved* performance!
Recovery of the Distribution

• It is also possible to establish bounds concerning the recovery of the distribution $f(M)$, i.e., the matrix where each entry gives us the probability of observing +1 when we sample that entry.

• We obtain matching upper and lower bounds on the average Hellinger distance between $f(M)$ and $f(\hat{M})$.

• When $\lim_{\alpha \to \infty} L_\alpha < \infty$, we can recover the distribution $f(M)$ without any assumptions on $\|M\|_\infty$:
  - logistic model
  - not probit model
  - any model where the noise has heavy tails.
Proof Methods

• Upper bounds
  - Probability in Banach spaces
  - Random matrix theory

• Lower bounds
  - Information theoretic arguments
  - Fano’s inequality
  - Packing sets of low-rank matrices
Tiny Sketch of Proof of Upper Bound

Recall that we maximize the log-likelihood $F(X)$

- For a fixed matrix $X$, $\mathbb{E}[F(M) - F(X)] = c \cdot D(f(X)\|f(M))$

- Lemma: Let $K = \{X : \frac{1}{\|X\|_*} \leq \sqrt{r}\}$. With high probability, $\sup_{X \in K} |F(X) - \mathbb{E}F(X)| \leq \delta$

- By definition, $F(\hat{M}) \geq F(M)$

$$0 \geq F(M) - F(\hat{M}) \geq \mathbb{E} \left[ F(M) - F(\hat{M}) \right] - 2\delta = c \cdot D(f(\hat{M})\|f(M)) - 2\delta$$

- Thus, $D(f(\hat{M})\|f(M)) \leq \frac{2}{c}\delta$
Synthetic Simulations

\[ d = 500 \quad m = .15 d^2 \]

\[ \frac{\| \hat{M} - M \|_F}{\| M \|_F} \]

\[ \log_{10} \sigma \]

\[ r = 5 \]
\[ r = 3 \]
\[ r = 2 \]
Voting Simulation

Binary (incomplete) data:
Voting history of 105 US senators on 299 bills from 2008-2010

First singular vector of $\hat{M}$

Senator party affiliations

First singular vector of $Y$
Voting Simulation

Randomly delete 90% of the entries

First singular vector of $\hat{M}$

Senator party affiliations

First singular vector of $Y$
Voting Simulation

Randomly delete 95% of the entries

First singular vector of $\hat{M}$

Senator party affiliations

First singular vector of $Y$

86% of missing votes correctly predicted
MovieLens Data Set

- 100,000 movie ratings (1000 users, 1700 movies) on a scale from 1 to 5

- Convert to binary outcomes by comparing each rating to the average rating in the data set

- Evaluate by checking if we predict the correct sign

- Training on 95,000 ratings and testing on remainder
  - “standard” matrix completion: 68% accuracy
  - 1-bit matrix completion: 74% accuracy
Thank You!