Phase Transitions in Coding, Communications, and Inference

Andrea Montanari and Rüdiger Urbanke

Stanford University and EPFL, Lausanne

January 5, 2009

From coding to probabilistic inference

Binary symmetric channel

$$\underline{y} \;=\; \underline{x} \oplus \underline{z}\,, \qquad\qquad \mathbb{H}\underline{x} = 0\,.$$

$$\underline{z} \;=\; (z_1, z_2, \ldots, z_n)\,, \qquad z_i\text{'s iid Bernoulli}(p).$$

Binary symmetric channel

$$\underline{y} = \underline{x} \oplus \underline{z}\,, \qquad \mathbb{H}\underline{x} = 0\,.$$

$$\underline{z} = (z_1, z_2, \dots, z_n)\,, \qquad z_i\text{'s iid Bernoulli}(p).$$

Syndrome decoding

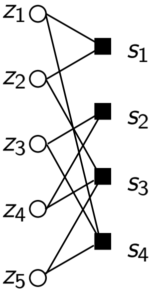
$$\underline{s} = \mathbb{H}\underline{y} = \mathbb{H}\underline{z}$$

Syndrome decoding

$$\underline{s} = \mathbb{H}\underline{y} = \mathbb{H}\underline{z}$$

Syndrome decoding

$(z_1, \ldots, z_n) \in \{0, 1\}$ is $n\,p$ sparse

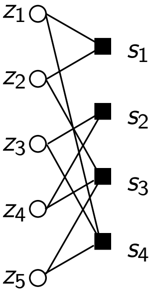


$\underline{s} = \mathbb{H}\underline{z} \quad \text{mod } 2$

$s_1, \ldots, s_m$: linear observations of the noise vector.

Syndrome decoding

$(z_1, \ldots, z_n) \in \{0, 1\}$ is $n\,p$ sparse



$\underline{s} = \mathbb{H}\underline{z} \quad \mod 2$

$s_1, \ldots, s_m$: linear observations of the noise vector.

Outline

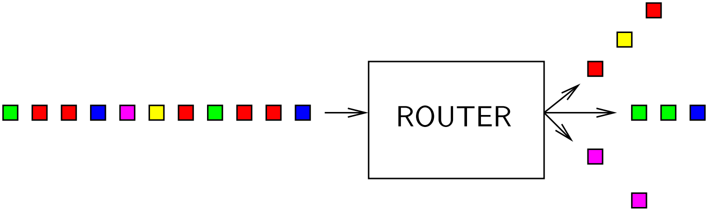
SPARSE VECTORS

Why sparse vectors? Network measurements

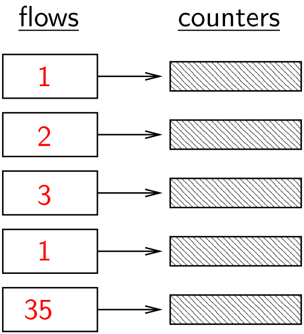
Network measurements

Internet core link :

- $10^{10} \div 10^{11}$ bits/sec
- Packet size 10^3 bits
- $10^6 \div 10^7$ flows per hour (mice. . . elephants)



A naive approach (1 flow $\leftrightarrow$ 1 counter)



Problem 1: Memory

Processing time: 12 nanosec per packet

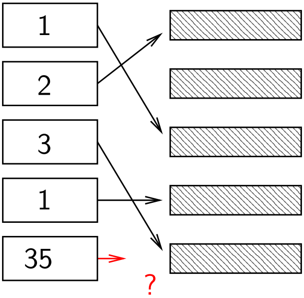
Space: 10^6 flows $\times$ 64 bits counters $=$ 8 MBytes

Problem 1: Memory

Processing time: 12 nanosec per packet

Space: 10^6 flows $\times$ 64 bits counters $=$ 8 MBytes

Problem 2: Flow-to-Counter association



Previous approaches and Related work

Hybrid architectures

[Shah, Iyer, Prabhakar, McKeown 2002]

Sampling

[Estan, Varghese 2001; CISCO's NetFlow]

Compressed sensing

[Candes, Donoho, Romberg, Tao, Indyk, Gilbert, Tanner 2006-. . .]

Previous approaches and Related work

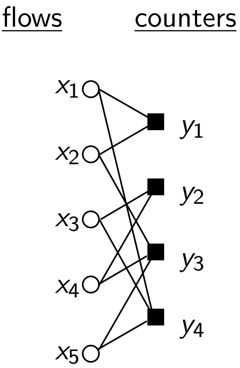
Hybrid architectures
[Shah, Iyer, Prabhakar, McKeown 2002]

Sampling
[Estan, Varghese 2001; CISCO's NetFlow]

Compressed sensing
[Candes, Donoho, Romberg, Tao, Indyk, Gilbert, Tanner 2006-. . .]

Counter braids: Vanilla version

Counter braid: one layer



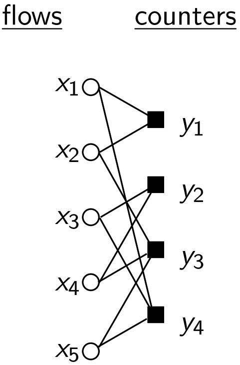
$$y = \mathbb{H}x$$

$$x = (x_1, \ldots, x_n) \in \mathbb{N}^n$$

$$y = (y_1, \ldots, y_m) \in \mathbb{N}^m$$

$\mathbb{H}$ adjacency matrix

Decoding: probabilistic inference

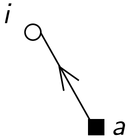


$$y = \mathbb{H}x$$

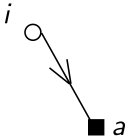
$$\mu(\underline{x}) = \frac{1}{Z} \prod_{a=1}^{m} \mathbb{I}(y_a = h_a^T x) \prod_{i=1}^{n} p_0(x_i)$$

Message passing algorithm

$$\text{BP messages} \quad \longrightarrow \quad \nu^{(t)}_{i\to a}(x_i),\ \widehat{\nu}^{(t)}_{a\to i}(x_i).$$



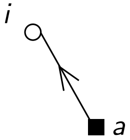
$\widehat{\nu}^{(t)}_{a\to i}$: counter a to flow i



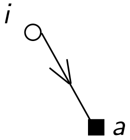
$\nu^{(t)}_{i\to a}$: flow i to counter a

$\nu^{(t)}_{i\to a}, \widehat{\nu}^{(t)}_{a\to i} \in \mathbb{N}$, $\nu^{(0)}_{i\to a} = 0$

Low complexity message passing algorithm



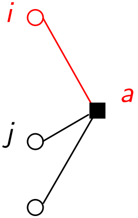
$\widehat{\nu}^{(t)}_{a\to i}$: counter a to flow i



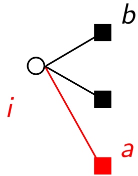
$\nu^{(t)}_{i\to a}$: flow i to counter a

$\nu^{(t)}_{i\to a}, \widehat{\nu}^{(t)}_{a\to i} \in \mathbb{N}$, $\qquad \nu^{(0)}_{i\to a} = 0$

Decoding a counter braid



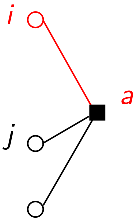
$$\widehat{\nu}^{(t)}_{a\to i} = \left[y_a - \textstyle\sum_{j\in\partial a\setminus i} \nu^{(t)}_{j\to a} \right]_+$$



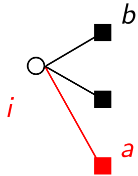
$$\nu^{(t+1)}_{i\to a} = \min\left\{ \widehat{\nu}^{(t)}_{b\to i} \;:\; b\in\partial i\setminus a \right\} \quad \text{for } t \text{ even}$$

$$\nu^{(t+1)}_{i\to a} = \max\left\{ \widehat{\nu}^{(t)}_{b\to i} \;:\; b\in\partial i\setminus a \right\} \quad \text{for } t \text{ odd}$$

Decoding a counter braid



$$\widehat{\nu}^{(t)}_{a\to i} = \left[y_a - \textstyle\sum_{j\in\partial a\setminus i} \nu^{(t)}_{j\to a}\right]_+$$



$$\nu^{(t+1)}_{i\to a} = \min\left\{\widehat{\nu}^{(t)}_{b\to i} \;:\; b\in\partial i\setminus a\right\} \quad \text{for } t \text{ even}$$

$$\nu^{(t+1)}_{i\to a} = \max\left\{\widehat{\nu}^{(t)}_{b\to i} \;:\; b\in\partial i\setminus a\right\} \quad \text{for } t \text{ odd}$$

Analysis and performances

The sandwich property

Proposition

$$\nu_i^{(0)} \leq \nu_i^{(2)} \leq \nu_i^{(4)} \leq \nu_i^{(6)} \leq \cdots \leq x_i$$

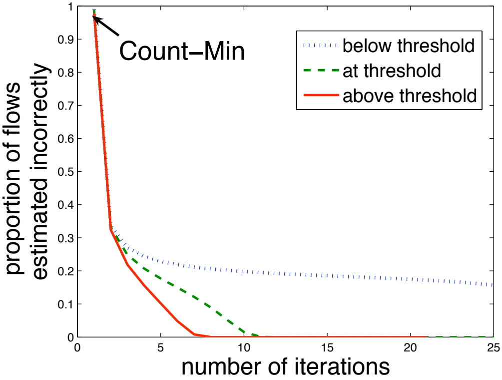
$$\nu_i^{(1)} \geq \nu_i^{(3)} \geq \nu_i^{(5)} \geq \nu_i^{(7)} \geq \cdots \geq x_i$$

A statistical model?

$X_1, X_2, \dots X_n$ iid, $X_i \geq X_{\min}$

$$\mathbb{P}\{X_i > X_{\min}\} = \epsilon\,.$$

Typical runs



Density evolution

l : flows degree

r : counters degree

$z_t \equiv \mathbb{P}\{\nu_{i\to a}^{(t)} \neq x_i\}$

$$z_{t+1} = \begin{cases} (1-(1-z_t)^{r-1})^{l-1} & \text{for } t \text{ even,} \\ \epsilon(1-(1-z_t)^{r-1})^{l-1} & \text{for } t \text{ odd,} \end{cases}$$

Density evolution

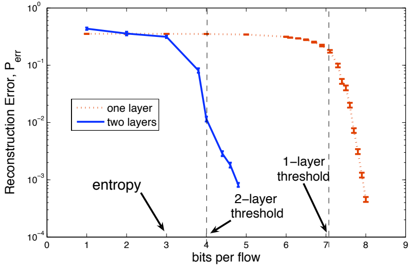
l : flows degree

r : counters degree

$z_t \equiv \mathbb{P}\{\nu_{i\to a}^{(t)} \neq x_i\}$

$$z_{t+1} = \begin{cases} (1-(1-z_t)^{r-1})^{l-1} & \text{for } t \text{ even,} \\ \epsilon(1-(1-z_t)^{r-1})^{l-1} & \text{for } t \text{ odd,} \end{cases}$$

Threshold in memory space



$$n = 1000,\ \mathbb{P}\{X_1 \geq x\} = x^{-3/2}$$

Dimensionality reduction

$$\gamma = \frac{\#\text{counters}}{\#\text{flows}} \, .$$

Optimal dimensionality reduction

Theorem (Donoho, Tanner, 2006)

Let $\gamma_{dens}(\epsilon)$ be the dimensionality reduction rate for ϵ-sparse sources, with Gaussian random matrices and LP decoding. Then

$$\gamma_{dens}(\epsilon) = 2 \cdot \epsilon \log(1/\epsilon) + O(\epsilon).$$

Theorem (Lu, M, Prabhakar, 2008)

Let $\gamma_{sparse}(\epsilon)$ be the dimensionality reduction rate for ϵ-sparse sources, sparse matrices and message passing decoding. Then

$$\gamma_{sparse}(\epsilon) \leq 2.09 \cdot \epsilon \log(1/\epsilon) + O(\epsilon).$$

Optimal dimensionality reduction

Theorem (Donoho, Tanner, 2006)

Let $\gamma_{dens}(\epsilon)$ be the dimensionality reduction rate for ϵ-sparse sources, with Gaussian random matrices and LP decoding. Then

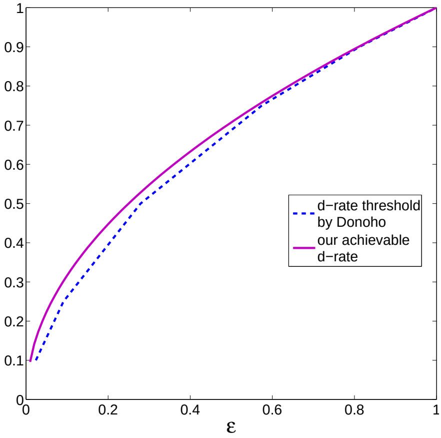
$$\gamma_{dens}(\epsilon) = 2 \cdot \epsilon \log(1/\epsilon) + O(\epsilon).$$

Theorem (Lu, M, Prabhakar, 2008)

Let $\gamma_{sparse}(\epsilon)$ be the dimensionality reduction rate for ϵ-sparse sources, sparse matrices and message passing decoding. Then

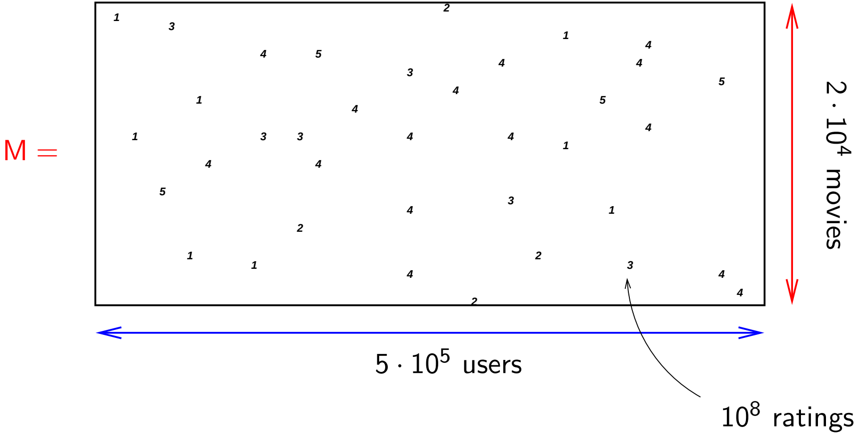
$$\gamma_{sparse}(\epsilon) \leq 2.09 \cdot \epsilon \log(1/\epsilon) + O(\epsilon).$$

Optimal dimensionality reduction

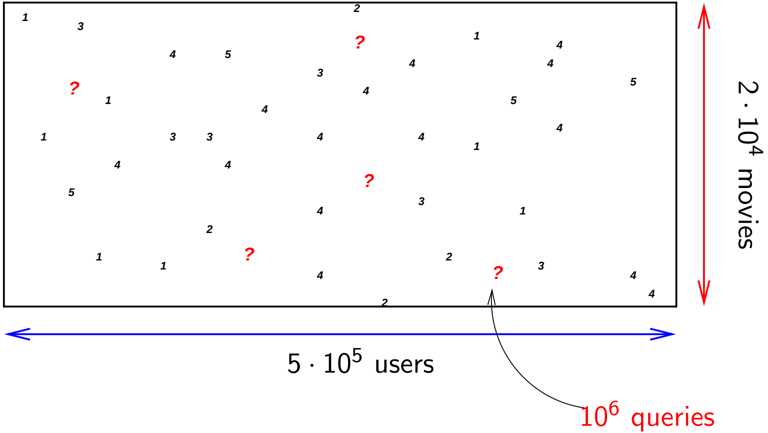


COLLABORATIVE FILTERING

Netflix dataset: A big (!) matrix



A big (!) matrix

$M =$ 

You get a prize if...

$$\mathrm{RMSE} < 0.8563 \qquad\qquad ;-)$$

Is this possible?

You get a prize if. . .

$$\mathrm{RMSE} < 0.8563 \qquad\qquad ;-)$$

Is this possible?

You get a prize if. . .

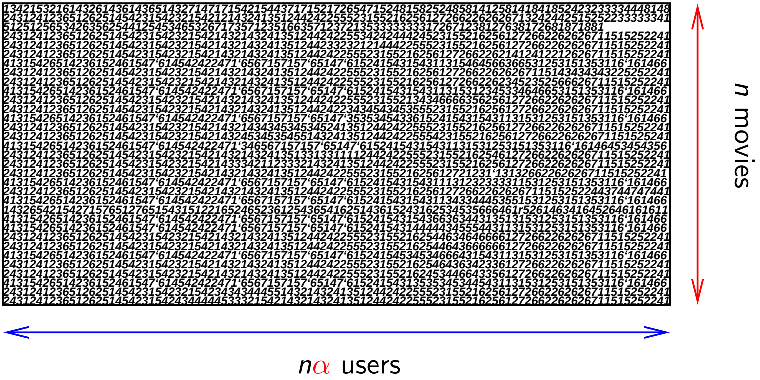
$$\mathrm{RMSE} < 0.8563 \qquad\qquad ;-)$$

Is this possible?

A model: Random low-rank matrices

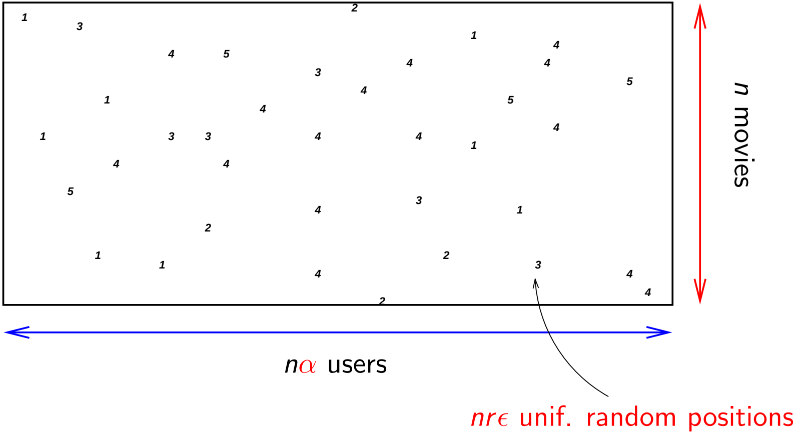
The observations

$M =$

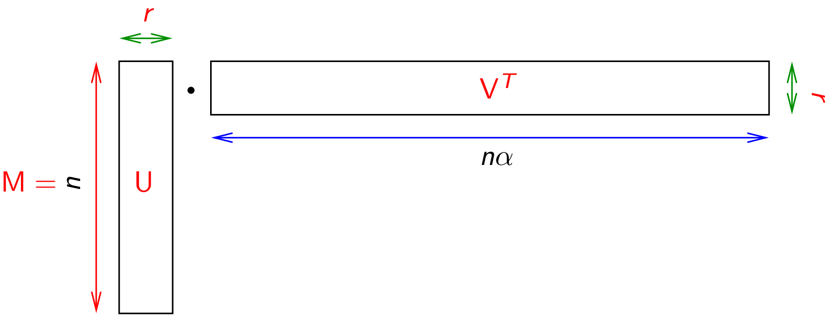


The observations

$\mathsf{M}^E =$

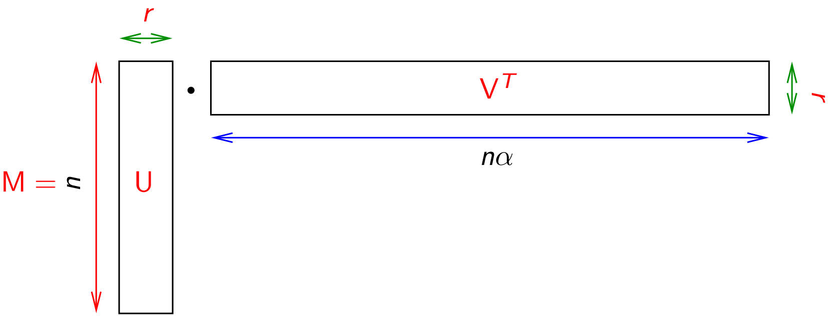


You need some structure!



$r \leq n^{1/2}$

You need some structure!



$r \leq n^{1/2}$

Factors

- U_{ik}, V_{ak} i.i.d.
- U, V random orthogonal.
- U, V 'incoherent'.

Factors

- U_{ik}, V_{ak} i.i.d.
- U, V random orthogonal.
- U, V ‘incoherent’.

Factors

- U_{ik}, V_{ak} i.i.d.
- U, V random orthogonal.
- U, V ‘incoherent’.

Metric (RMSE)

$$D(\mathsf{M}, \hat{\mathsf{M}}) \equiv \left\{ \frac{1}{n^2\alpha} \sum_{i,a} |\mathsf{M}_{ia} - \hat{\mathsf{M}}_{ia}|^2 \right\}^{1/2}$$

Previous work

Theorem (Candés, Recht, 2008)

If

$$\epsilon \geq C\, n^{1/5} \log n$$

then whp

1. M *is unique given the observed entries.*
2. M *is the unique minimum of a SDP.*

cf. also [Recht, Fazel, Parrilo 2007]

Previous work

Theorem (Candés, Recht, 2008)

If

$$\epsilon \geq C\, n^{1/5} \log n$$

then whp

1. M *is unique given the observed entries.*
2. M *is the unique minimum of a SDP.*

cf. also [Recht, Fazel, Parrilo 2007]

Previous work

Theorem (Candés, Recht, 2008)

If

$$\epsilon \geq C\, n^{1/5} \log n$$

then whp

1. M *is unique given the observed entries.*
2. M *is the unique minimum of a SDP.*

cf. also [Recht, Fazel, Parrilo 2007]

Previous work

Theorem (Candés, Recht, 2008)

If

$$\epsilon \geq C\, n^{1/5} \log n$$

then whp

1. M *is unique given the observed entries.*
2. M *is the unique minimum of a SDP.*

cf. also [Recht, Fazel, Parrilo 2007]

Previous work

Theorem (Candés, Recht, 2008)

If

$$\epsilon \geq C\, n^{1/5} \log n$$

then whp

1. M *is unique given the observed entries.*
2. M *is the unique minimum of a SDP.*

cf. also [Recht, Fazel, Parrilo 2007]

Great, but. . .

1. $n^{1/5}$ observations for 1 bit of information?
2. $\mathrm{RMSE} = 0$?
3. SDP $=$ $O(n^{4\dots6})$. Substitute $n = 10^5$. . .

Great, but...

1. $n^{1/5}$ observations for 1 bit of information?

2. $\mathrm{RMSE} = 0$?

3. SDP $=$ $O(n^{4\cdots6})$. Substitute $n = 10^5\ldots$

Great, but...

1. $n^{1/5}$ observations for 1 bit of information?

2. $\mathrm{RMSE} = 0$?

3. SDP = $O(n^{4\dots6})$. Substitute $n = 10^5$...

Great, but...

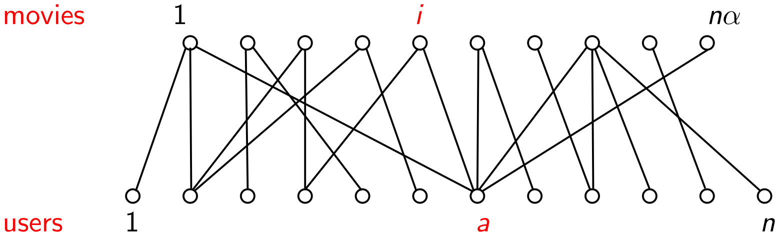
1. $n^{1/5}$ observations for 1 bit of information?

2. $\mathrm{RMSE} = 0$?

3. SDP = $O(n^{4\ldots6})$. Substitute $n = 10^5$...

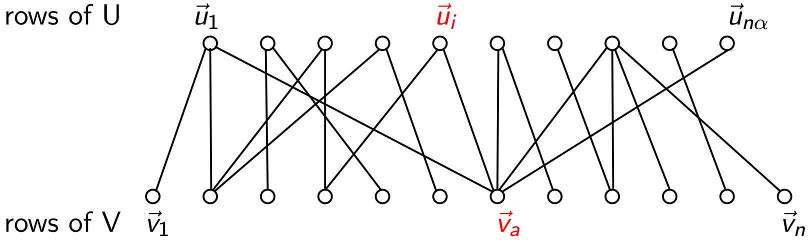
A graphical model

The graph



$(i, a) \in E \quad \Leftrightarrow$ User a rated movie i.

The graphical model



$$\mu(\{\vec{u}_i\},\{\vec{v}_a\}) = \frac{1}{Z}\prod_{(i,a)\in E}\mathbb{I}(\vec{u}_i\cdot\vec{v}_a = \mathrm{M}_{ia})\prod_{i=1}^{n\alpha}p_0(\vec{u}_i)\prod_{a=1}^{n}p_0(\vec{v}_i)\,.$$

Belief propagation

Messages $\nu_{i\to a}(\vec{u}_i)$, $\nu_{a\to i}(\vec{v}_a)$.

A small simulation

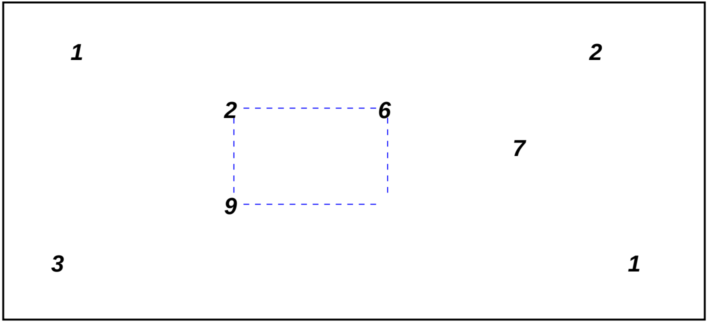
Messages $\nu_{i\to a}(\vec{u}_i)$, $\nu_{a\to i}(\vec{v}_a)$.

A small simulation

$O(n)$ entries are enough (practice)

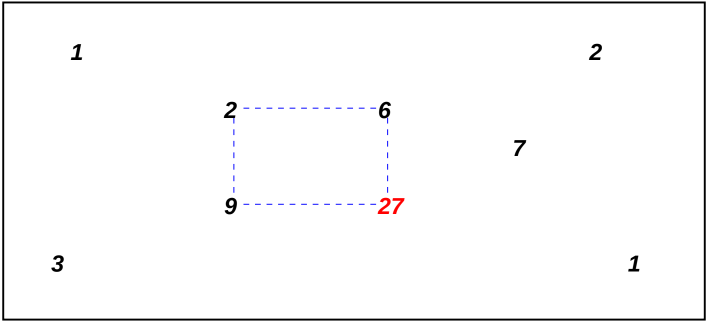
Rank = 1: an easy trick

$M =$

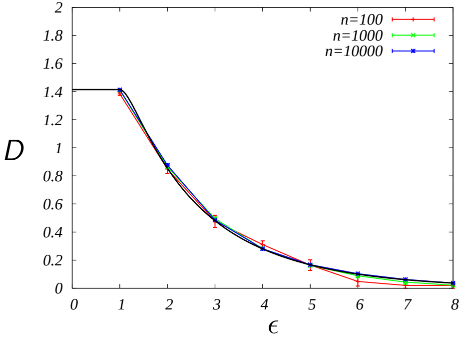


Rank = 1: an easy trick

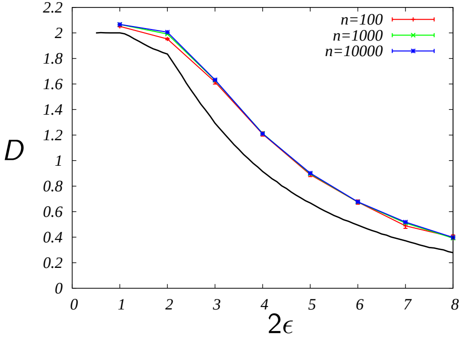
$M =$



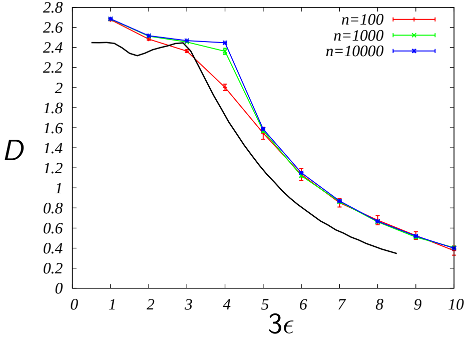
Rank = 1: Trick vs. Belief Propagation



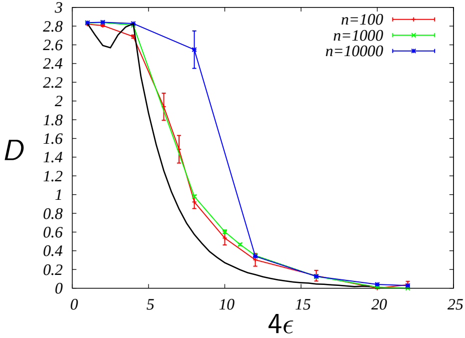
Rank = 2: Belief Propagation



Rank = 3: Belief Propagation



Rank = 4: Belief Propagation



$O(n)$ entries are enough (theory)

Can you prove anything about BP?

Assume $p_0(\cdot)$ uniform over $\{-1,+1\}^r$

Uniform fixed point $\nu^*_{i\to a}(\cdot) = \nu^*_{a\to i}(\cdot) = p_0(\cdot)$

- Becomes unstable for $r\epsilon > 1$.
- Two new $+/-$ symmetric fixed points.
- Become unstable for $\epsilon >$ const. :-(

Can you prove anything about BP?

Assume $p_0(\,\cdot\,)$ uniform over $\{-1, +1\}^r$

Uniform fixed point $\nu^*_{i\to a}(\,\cdot\,) = \nu^*_{a\to i}(\,\cdot\,) = p_0(\,\cdot\,)$

- Becomes unstable for $r\epsilon > 1$.
- Two new $+/-$ symmetric fixed points.
- Become unstable for $\epsilon >$ const. :-(

Can you prove anything about BP?

Assume $p_0(\cdot)$ uniform over $\{-1,+1\}^r$

Uniform fixed point $\nu^*_{i\to a}(\cdot) = \nu^*_{a\to i}(\cdot) = p_0(\cdot)$

- Becomes unstable for $r\epsilon > 1$.
- Two new $+/-$ symmetric fixed points.
- Become unstable for $\epsilon >$ const. :-(

Can you prove anything about BP?

Assume $p_0(\,\cdot\,)$ uniform over $\{-1,+1\}^r$

Uniform fixed point $\nu^*_{i\to a}(\,\cdot\,) = \nu^*_{a\to i}(\,\cdot\,) = p_0(\,\cdot\,)$

- Becomes unstable for $r\epsilon > 1$.
- Two new $+/-$ symmetric fixed points.
- Become unstable for $\epsilon >$ const. :-(

$$\mathrm{M}^E_{ia} = \begin{cases} \mathrm{M}_{ia} & \text{if } (i,a) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Projection

$$\mathrm{M}^E = \sum_{i=1}^{n} \sigma_i x_i y_i^T, \qquad \sigma_1 \geq \sigma_2 \geq \dots$$

$$\mathrm{T}_r(\mathrm{M}_E) = \frac{n\sqrt{\alpha}}{\epsilon} \sum_{i=1}^{r} \sigma_i x_i y_i^T.$$

Naive spectral algorithm

$$\mathrm{M}^E_{ia} = \begin{cases} \mathrm{M}_{ia} & \text{if } (i,a) \in E, \\ 0 & \text{otherwise.} \end{cases}$$

Projection

$$\begin{aligned} \mathrm{M}^E &= \sum_{i=1}^{n} \sigma_i x_i y_i^T\,, \qquad \sigma_1 \ge \sigma_2 \ge \dots \\ \mathrm{T}_r(\mathrm{M}_E) &= \frac{n\sqrt{\alpha}}{\epsilon} \sum_{i=1}^{r} \sigma_i x_i y_i^T\,. \end{aligned}$$

Troubles and solution

If $\epsilon = O(1)$, 'spurious' singular values $\Omega(\sqrt{\log n/(\log\log n)})$.

Trimming

$$\widetilde{M}^E_{ia} = \begin{cases} M^E_{ia} & \text{if } \deg(i) \le 2\,\mathbb{E}\deg(i), \ \ \deg(a) \le 2\,\mathbb{E}\deg(a)\,, \\ 0 & \text{otherwise.} \end{cases}$$

Troubles and solution

If $\epsilon = O(1)$, 'spurious' singular values $\Omega(\sqrt{\log n/(\log\log n)})$.

Trimming

$$\widetilde{\mathsf{M}}^E_{ia} = \begin{cases} \mathsf{M}^E_{ia} & \text{if } \deg(i) \le 2\,\mathbb{E}\deg(i), \;\; \deg(a) \le 2\,\mathbb{E}\deg(a)\,, \\ 0 & \text{otherwise.} \end{cases}$$

Not-as-naive spectral algorithm

SPECTRAL MATRIX COMPLETION(matrix M^E)
1: Trim M^E, and let $\widetilde{\mathrm{M}}^E$ be the output;
2: Project $\widetilde{\mathrm{M}}^E$ to $\mathrm{T}_r(\widetilde{\mathrm{M}}^E)$;
3: Clean residual errors by coordinate descent in the factors.

Complexity

SVD of $\widetilde{\mathrm{M}}^{E}$

- Standard algorithms $\longrightarrow O(n^3)$
- Iterative $\longrightarrow O(nr\epsilon \log n)$

Theorem (Keshavan, Oh, M, 2009)

For each $\delta > 0$, if $\epsilon \geq C(\alpha, \delta)$, then with high probability

$$||\mathsf{M} - \mathrm{T}_r(\widetilde{\mathsf{M}}^E)||_{\mathrm{F}}^2 \leq n^2 r\, \delta.$$

Theorem (Keshavan, Oh, M, 2009)

If $\epsilon \geq C'(\alpha) \log n$, then SPECTRAL MATRIX COMPLETION *returns, with high probability, the matrix* M.

Theorem (Keshavan, Oh, M, 2009)

For each $\delta > 0$, if $\epsilon \geq C(\alpha, \delta)$, then with high probability

$$||\mathsf{M} - \mathrm{T}_r(\widetilde{\mathsf{M}}^E)||_{\mathrm{F}}^2 \leq n^2 r\,\delta.$$

Theorem (Keshavan, Oh, M, 2009)

If $\epsilon \geq C'(\alpha) \log n$, then Spectral Matrix Completion *returns, with high probability, the matrix* M.

Key technical result

$$M = \sum_{i=1}^{r} \Sigma_i \underline{u}_i \underline{v}_i^T \, ,$$

$$||\underline{u}_i|| = \sqrt{n}\, , \;\; \underline{u}_i^T \underline{u}_j = 0\, , \qquad ||\underline{v}_i|| = \sqrt{n\alpha}\, , \;\; \underline{v}_i^T \underline{v}_j = 0\, .$$

Theorem

If $\{\underline{u}_i\}$, $\{\underline{v}_i\}$ are incoherent, then, w.h.p.

$$\begin{aligned} |\sigma_q - \epsilon r \Sigma_q| &\leq C\, r\sqrt{\epsilon}\log\epsilon \qquad &\text{for } q \leq r\, ,\\ \sigma_q &\leq C\, r\sqrt{\epsilon} \qquad &\text{for } q > r\, . \end{aligned}$$

Key technical result

$$\mathrm{M} = \sum_{i=1}^{r} \Sigma_i \underline{u}_i \underline{v}_i^T \,,$$

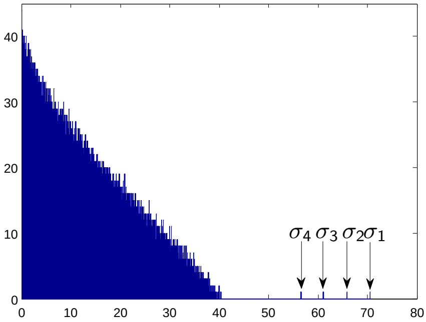
$$||\underline{u}_i|| = \sqrt{n}\,, \quad \underline{u}_i^T \underline{u}_j = 0\,, \qquad ||\underline{v}_i|| = \sqrt{n\alpha}\,, \quad \underline{v}_i^T \underline{v}_j = 0\,.$$

Theorem

If $\{\underline{u}_i\}$, $\{\underline{v}_i\}$ are incoherent, then, w.h.p.

$$\begin{aligned} |\sigma_q - \epsilon r \Sigma_q| &\le C\, r \sqrt{\epsilon} \log \epsilon && \text{for } q \le r\,, \\ \sigma_q &\le C\, r \sqrt{\epsilon} && \text{for } q > r\,. \end{aligned}$$

$n = 10000,\ r = 4,\ \epsilon = 12.5$



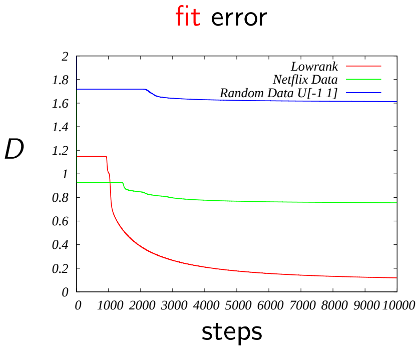
Back to the data

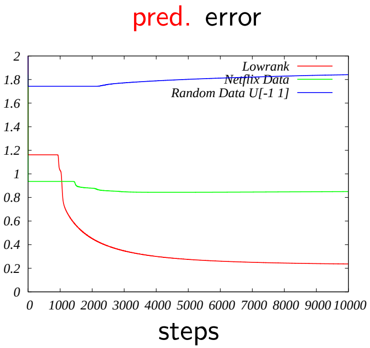
Is Neflix a random low-rank matrix?

Compare for coordinate descent (SimonFunk).

$$(n = 5 \cdot 10^3,\ \alpha = 1)$$

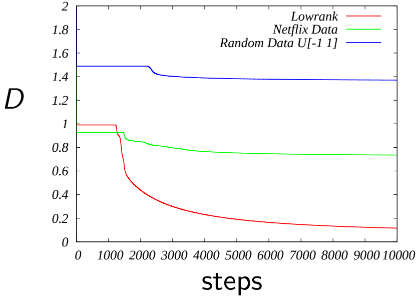
Rank = 3


fit error
D
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Lowrank
Netflix Data
Random Data U[-1 1]
0 1000 2000 3000 4000 5000 6000 7000 8000 9000 10000
steps

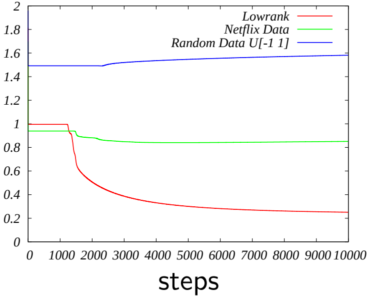

pred. error
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
Lowrank
Netflix Data
Random Data U[-1 1]
0 1000 2000 3000 4000 5000 6000 7000 8000 9000 10000
steps

Rank $= 4$

fit error

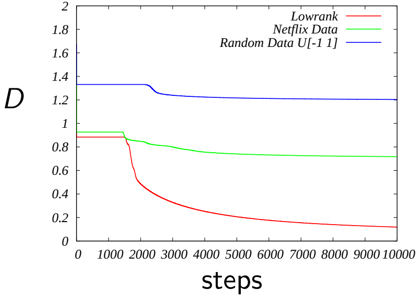


pred. error

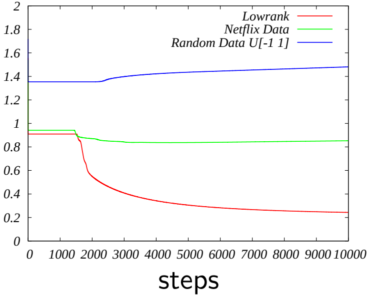


Rank = 5

fit error



pred. error



CONCLUSION

Conclusion

Enough information, measurements,... $\Rightarrow$ Threshold

Engineering phase tranditions.

Conclusion

Enough information, measurements,. . . $\Rightarrow$ Threshold

Engineering phase tranditions.