Compression via Sparse Linear Regression

Ramji Venkataramanan

University of Cambridge

(Acknowledgements: S. Tatikonda, T. Sarkar, A. Joseph, A. Barron)

May 9, 2014

Information Theory deals with

- Communication
- Compression (Lossless and Lossy)
- Multi-terminal communication and compression:
  Multiple-access channels, Broadcast channels, Distributed compression, ...
- Sharp characterization of achievable rates for many of these problems

Information Theory deals with

- Communication
- Compression (Lossless and Lossy)
- Multi-terminal communication and compression:

  Multiple-access channels, Broadcast channels, Distributed compression, ...
- Sharp characterization of achievable rates for many of these problems

Textbook code constructions are based on:

- *Random coding* for point-to-point communication and compression
- Superposition and binning for multi-terminal problems
- High complexity of storage and coding: *exponential* in "n"

GOAL:

- Codes with compact representation + fast encoding/decoding
  'Fast' $\Rightarrow$ *polynomial* in n

- In the last 20 years, many advances:
  LDPC/LDGM codes, Polar codes for finite-alphabet sources & channels

- We will focus on Gaussian sources and channels here

In this talk . . .

- Ensemble of codes based on sparse linear regression
- *Provably* achieve rates close to info-theoretic limits with fast encoding + decoding
- Based on construction of Barron & Joseph for AWGN channel
  - Achieve capacity with fast decoding [IT Trans. '12, '14]

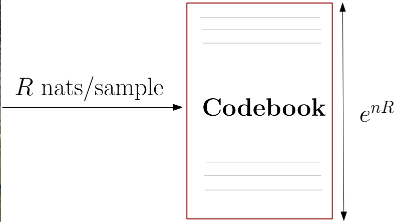
In this talk ...

- Ensemble of codes based on sparse linear regression
- *Provably* achieve rates close to info-theoretic limits with fast encoding + decoding
- Based on construction of Barron & Joseph for AWGN channel
  - Achieve capacity with fast decoding [IT Trans. '12, '14]

Outline

- We'll focus on the compression problem:
  - Fundamental limits of the code (with optimal encoding)
  - Computationally efficient compression algorithm & analysis
- Extension to multi-terminal communication and compression

Lossy Compression

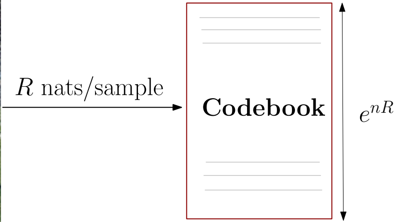


$\mathbf{S} = S_1, \ldots, S_n$ $\qquad\qquad$ $\hat{\mathbf{S}} = \hat{S}_1, \ldots, \hat{S}_n$

- Distortion criterion: $\frac{1}{n}\|\mathbf{S} - \hat{\mathbf{S}}\|^2 = \frac{1}{n}\sum_k (S_k - \hat{S}_k)^2$
- To achieve $\frac{1}{n}\|\mathbf{S} - \hat{\mathbf{S}}\|^2 \leq D$, need
$$R > R^*(D) = \min_{P_{\hat{S}|S}} I(S; \hat{S})$$
- For i.i.d $\mathcal{N}(0, \sigma^2)$ source, $R^*(D) = \frac{1}{2}\log\frac{\sigma^2}{D}$, $\quad D < \sigma^2$

$\Rightarrow$ Minimum possible distortion $D^*(R) = \sigma^2 e^{-2R}$

Lossy Compression



$\mathbf{S} = S_1, \ldots, S_n$ $\qquad\qquad$ $\hat{\mathbf{S}} = \hat{S}_1, \ldots, \hat{S}_n$

- Distortion criterion: $\frac{1}{n}\|\mathbf{S} - \hat{\mathbf{S}}\|^2 = \frac{1}{n}\sum_k (S_k - \hat{S}_k)^2$
- To achieve $\frac{1}{n}\|\mathbf{S} - \hat{\mathbf{S}}\|^2 \leq D$, need
$$R > R^*(D) = \min_{P_{\hat{S}|S}} \; I(S; \hat{S})$$
- For i.i.d $\mathcal{N}(0, \sigma^2)$ source, $R^*(D) = \frac{1}{2}\log\frac{\sigma^2}{D}$, $\quad D < \sigma^2$

$$\Rightarrow \quad \text{Minimum possible distortion } D^*(R) = \sigma^2 e^{-2R}$$

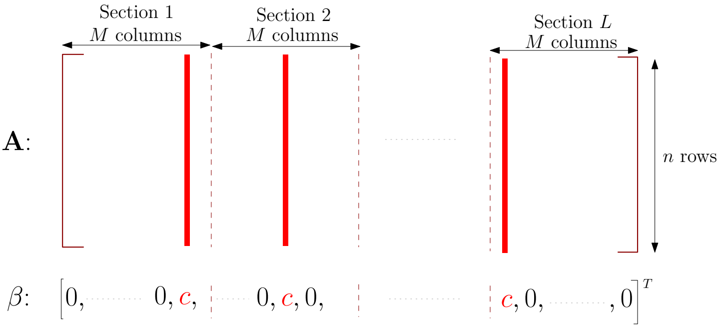
Can we achieve this with *low-complexity* algorithms?

Sparse Regression Codes (SPARC)

$$\beta: \quad [0, \cdots\cdots 0, {\color{red}c}, \cdots 0, {\color{red}c}, 0, \cdots\cdots {\color{red}c}, 0, \cdots\cdots, 0]^T$$

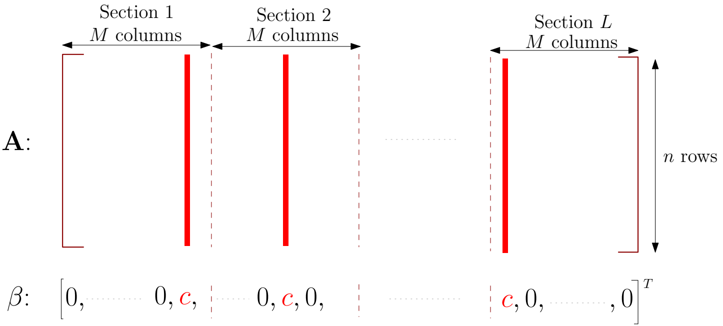
- **A**: design matrix or 'dictionary' with ind. $\mathcal{N}(0,1)$ entries
- Codewords $\mathbf{A}\beta$ - *sparse* linear combinations of columns of **A**

SPARC Construction



n rows, ML columns

SPARC Construction

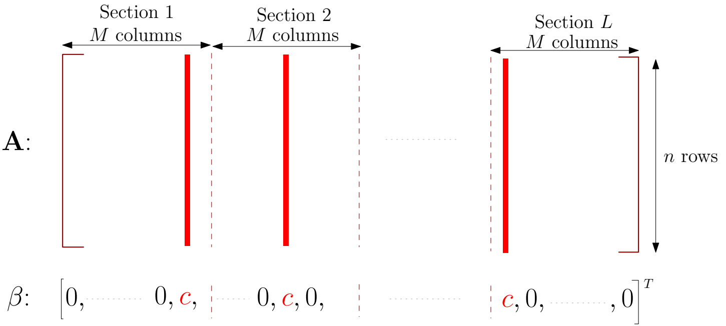


n rows, ML columns

Choosing M and L:

- For rate R codebook, need $M^L = e^{nR}$
- Choose $M = L^b$ for $b > 1 \Rightarrow bL \log L = nR$

SPARC Construction

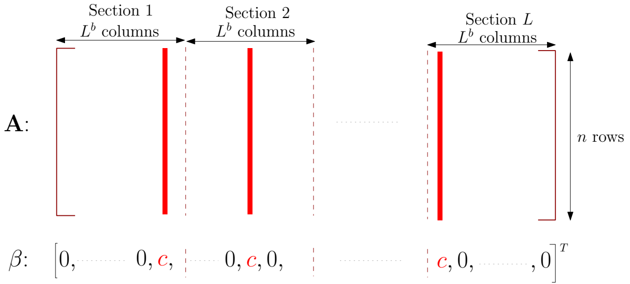


n rows, ML columns

Choosing M and L:

- For rate R codebook, need $M^L = e^{nR}$
- Choose $M = L^b$ for $b > 1 \Rightarrow bL \log L = nR$
- $L \sim n/\log n$ and $M \sim$ polynomial in n
- Storage Complexity $\leftrightarrow$ Size of **A**: polynomial in n

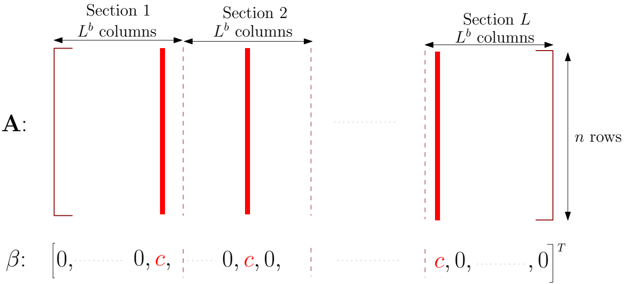
Minimum Distance Encoding



Given source sequence $\mathbf{S}$ with variance σ^2:

- *Encoder*: Find $\hat{\beta} = \underset{\beta}{\text{argmin}} \, \|\mathbf{S} - \mathbf{A}\beta\|^2$
- *Decoder*: Reconstruct $\hat{\mathbf{S}} = \mathbf{A}\hat{\beta}$

Minimum Distance Encoding



Given source sequence $\mathbf{S}$ with variance σ^2:

- *Encoder*: Find $\hat{\beta} = \underset{\beta}{\text{argmin}} \; \|\mathbf{S} - \mathbf{A}\beta\|^2$
- *Decoder*: Reconstruct $\hat{\mathbf{S}} = \mathbf{A}\hat{\beta}$

$$P_n = P\left(\tfrac{1}{n} \|\mathbf{S} - \mathbf{A}\hat{\beta}\|^2 > D \right)$$

1. Want to show that $P_n \to 0$ if $R > \frac{1}{2} \log \frac{\sigma^2}{D}$
2. Also want asymptotic rate of decay (*error exponent*)

SPARC Rate-Distortion Function

Theorem (RV-Joseph-Tatikonda '12, RV-Tatikonda '14)

For a source with variance σ^2, SPARCs with minimum-distance encoding achieve distortion D for all rates

$$R > \frac{1}{2}\log\frac{\sigma^2}{D}$$

when $b > b_{min}$ where

$$b_{min} = \begin{cases} \frac{2.5R}{R-1+D/\sigma^2} & \text{if } R > (1-\frac{D}{\sigma^2}) \\ \frac{40R}{\left(\frac{2R}{(1-D/\sigma^2)}-1\right)^2\left((1-\frac{D}{\sigma^2})(2+\frac{D}{\sigma^2})-2R\right)} & \text{if } R \le (1-\frac{D}{\sigma^2}) \end{cases}$$

SPARC Rate-Distortion Function

Theorem (RV-Joseph-Tatikonda '12, RV-Tatikonda '14)

For a source with variance σ^2, SPARCs with minimum-distance encoding achieve distortion D for all rates

$$R > \frac{1}{2}\log\frac{\sigma^2}{D}$$

when $b > b_{min}$ where

$$b_{min} = \begin{cases} \frac{2.5R}{R-1+D/\sigma^2} & \text{if } R > (1-\frac{D}{\sigma^2}) \\[2ex] \frac{40R}{\left(\frac{2R}{(1-D/\sigma^2)}-1\right)^2\left((1-\frac{D}{\sigma^2})(2+\frac{D}{\sigma^2})-2R\right)} & \text{if } R \le (1-\frac{D}{\sigma^2}) \end{cases}$$

Note:

$$\frac{D}{\sigma^2} \in (0.203, 1) \quad \Leftrightarrow \quad \left(1-\frac{D}{\sigma^2}\right) > \frac{1}{2}\log\frac{\sigma^2}{D}$$

Setting up the analysis

Call β a *solution* if $|\mathbf{S} - \mathbf{A}\beta|^2 \leq D$

For $i = 1, \ldots, e^{nR}$, define

$$U_i = \begin{cases} 1 & \text{if } \beta(i) \text{ is a solution ,} \\ 0 & \text{otherwise.} \end{cases}$$

The number of solutions X is

$$X = U_1 + \ldots + U_{e^{nR}}$$

Want to show $P(X > 0) \to 1$ as $n \to \infty$

Setting up the analysis

Call β a *solution* if $|\mathbf{S} - \mathbf{A}\beta|^2 \leq D$

For $i = 1, \ldots, e^{nR}$, define

$$U_i = \begin{cases} 1 & \text{if } \beta(i) \text{ is a solution }, \\ 0 & \text{otherwise.} \end{cases}$$

The number of solutions X is

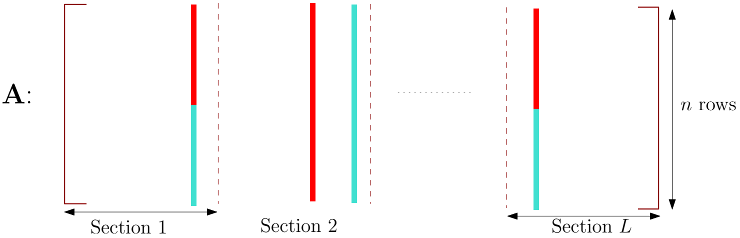
$$X = U_1 + \ldots + U_{e^{nR}}$$

Want to show $P(X > 0) \to 1$ as $n \to \infty$

Notice that the U_i's are *dependent*!

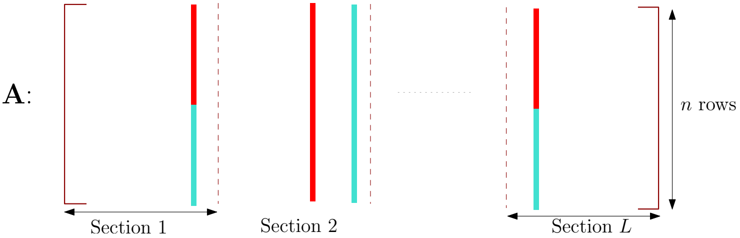
Dependent Codewords

- Each codeword sum of L columns
- Codewords $\beta(i), \beta(j)$ *dependent* if they have common columns



Dependent Codewords

- Each codeword sum of L columns
- Codewords $\beta(i), \beta(j)$ *dependent* if they have common columns



The number of codewords sharing r common terms with any $\beta(i)$ is

$$\binom{L}{r}(M-1)^{L-r}, \quad r = 0, 1, \ldots, L$$

$\#$ codewords dependent with $\beta(i) = M^L - 1 - (M-1)^L$

The Second Moment Method (2nd MoM)

$$X = U_1 + \ldots + U_{e^{nR}}$$

To show $P(X > 0)$ w.h.p., we use the 2nd MoM:

$$P(X > 0) \;\geq\; \frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]}$$

Proof : $(\mathbb{E}[XY])^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$ with $Y = \mathbf{1}_{\{X>0\}}$.

The Second Moment Method (2nd MoM)

$$X = U_1 + \ldots + U_{e^{nR}}$$

To show $P(X > 0)$ w.h.p., we use the 2nd MoM:

$$P(X > 0) \;\geq\; \frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]}$$

Proof : $(\mathbb{E}[XY])^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$ with $Y = \mathbf{1}_{\{X>0\}}$.

The *expected* number of solutions is

$$\mathbb{E}X \;=\; e^{nR} P(U_1 = 1) \;\doteq\; e^{n(R \,-\, \frac{1}{2}\log\frac{\sigma^2}{D})}$$

$\mathbb{E}X \to \infty$ if $R > \frac{1}{2}\log\frac{\sigma^2}{D}$, but is $X > 0$ w.h.p. ?

The Second Moment

$$
\begin{aligned}
\mathbb{E}[X^2] &= \mathbb{E}[\ (U_1 + \ldots + U_{e^{nR}})^2\ ] \\
&= e^{nR} \sum_{r=0}^{L} \binom{L}{r} (M-1)^{L-r}\, \mathbb{E}[U_1 U_2 \mid \beta_1, \beta_2 \text{ share } r \text{ terms }]
\end{aligned}
$$

The Second Moment

$$\mathbb{E}[X^2] = \mathbb{E}[\ (U_1 + \ldots + U_{e^{nR}})^2\ ]$$
$$= e^{nR} \sum_{r=0}^{L} \binom{L}{r}(M-1)^{L-r}\,\mathbb{E}[U_1 U_2 \mid \beta_1, \beta_2 \text{ share } r \text{ terms }]$$

The key ratio is

$$\frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]} \doteq \left(\frac{D}{\sigma^2}\right)^n \left[\sum_{r=0}^{L}\binom{L}{r}(M-1)^{-r}\ \underbrace{\mathbb{E}[U_1 U_2 \mid \beta_1, \beta_2 \text{ share } r \text{ terms}]}_{\text{can compute Chernoff bound}}\right]^{-1}$$

$$\vdots$$

The Second Moment

$$\mathbb{E}[X^2] = \mathbb{E}[\,(U_1 + \ldots + U_{e^{nR}})^2\,]$$

$$= e^{nR} \sum_{r=0}^{L} \binom{L}{r} (M-1)^{L-r}\, \mathbb{E}[U_1 U_2 \mid \beta_1, \beta_2 \text{ share } r \text{ terms }]$$

The key ratio is

$$\frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]} \doteq \left(\frac{D}{\sigma^2}\right)^n \left[\sum_{r=0}^{L} \binom{L}{r} (M-1)^{-r} \underbrace{\mathbb{E}[U_1 U_2 \mid \beta_1, \beta_2 \text{ share } r \text{ terms}]}_{\text{can compute Chernoff bound}} \right]^{-1}$$

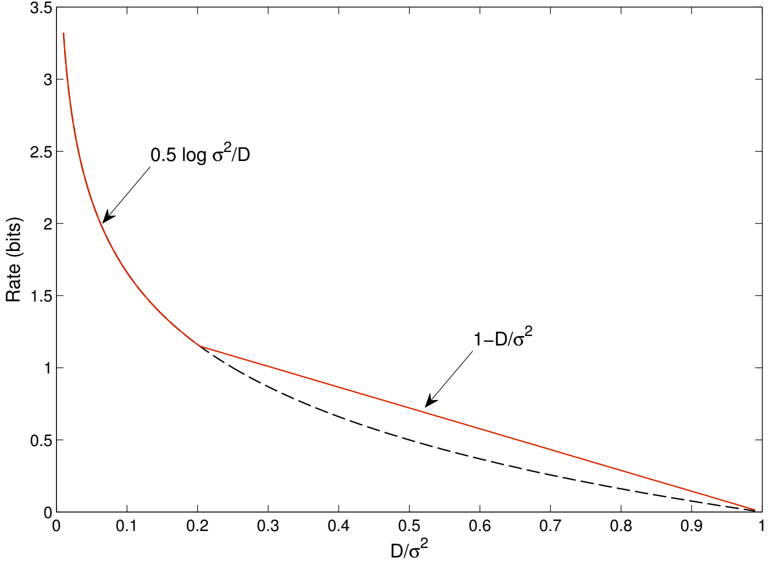
$$\vdots$$

$$P(X = 0) < L^{-\frac{1}{R}(b - b_{min})(R - (1 - \frac{D}{\sigma^2}))}$$

We've shown that rates $R > \max\left\{(1 - \frac{D}{\sigma^2}),\ \frac{1}{2}\log\frac{\sigma^2}{D}\right\}$ are achievable

What we have shown . . .

$$\text{Plot of max } \left\{ \left(1 - \frac{D}{\sigma^2}\right),\ \frac{1}{2}\log\frac{\sigma^2}{D} \right\}$$



Key Question: For

$$(0.203) \;<\; \frac{D}{\sigma^2} \;<\; 1$$

- Is the SPARC inherently a suboptimal code?
- Or, is it a shortcoming of the proof technique?

Why does the 2nd MoM fail ?

$$\mathbb{E}[X^2] = \mathbb{E}[\ (U_1 + \ldots + U_{2^{nR}})^2\ ] = \mathbb{E}[X]\,\mathbb{E}[X | U_1 = 1]$$

Hence

$$P(X > 0) \ \geq \ \frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]} \ = \ \frac{\mathbb{E}[X]}{\mathbb{E}[X | U_1 = 1]}$$

Why does the 2nd MoM fail ?

$$\mathbb{E}[X^2] = \mathbb{E}[\ (U_1 + \ldots + U_{2^{nR}})^2\ ] = \mathbb{E}[X]\,\mathbb{E}[X|U_1 = 1]$$

Hence

$$P(X > 0) \ \geq\ \frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]} \ =\ \frac{\mathbb{E}[X]}{\mathbb{E}[X|U_1 = 1]}$$

We want $\mathbb{E}[X|\,\beta(1)$ is a solution$] \ \sim\ \mathbb{E}[X]$

Why does the 2nd MoM fail ?

$$\mathbb{E}[X^2] = \mathbb{E}[\ (U_1 + \ldots + U_{2^{nR}})^2\ ] = \mathbb{E}[X]\,\mathbb{E}[X | U_1 = 1]$$

Hence

$$P(X > 0)\ \geq\ \frac{(\mathbb{E}X)^2}{\mathbb{E}[X^2]}\ =\ \frac{\mathbb{E}[X]}{\mathbb{E}[X|U_1 = 1]}$$

We want $\mathbb{E}[X|\,\beta(1) \text{ is a solution}]\ \sim\ \mathbb{E}[X]$

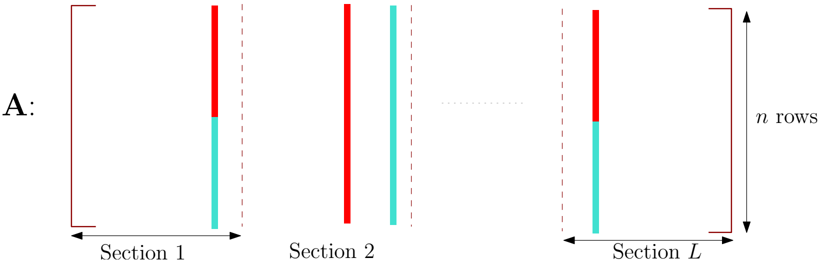
But when $\frac{1}{2}\log\frac{\sigma^2}{D}\ <\ R\ <\ \left(1 - \frac{D}{\sigma^2}\right)$

$$\frac{\mathbb{E}[X|\,\beta(1) \text{ is a solution}]}{\mathbb{E}[X]}\ \to\ \infty$$

- The expected number of solutions *given* that we have one solution blows up!
- Similar phenomenon in random hypergraph 2-colouring [Coja-Oghlan, Zdeborova '12]

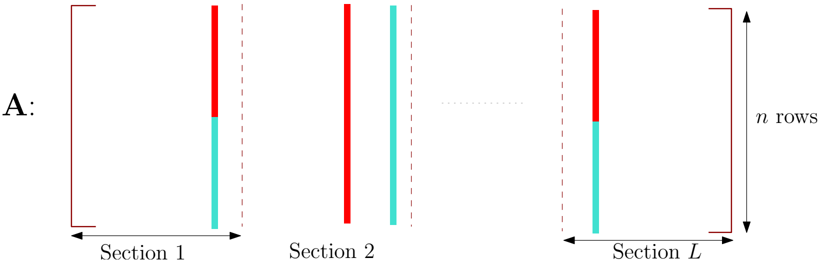
Q: Why is $\mathbb{E}[X \mid \beta(1) \text{ is a solution}] \gg \mathbb{E}[X]$?

- There are many codewords $\beta(i)$ that are *dependent* with $\beta(1)$
- If $\beta(1), \beta(i)$ are dependent: *given that* $|\mathbf{S} - \mathbf{A}\beta(1)|^2 \leq D$, *the probability of* $|\mathbf{S} - \mathbf{A}\beta(i)|^2 \leq D$ *increases*



Q: Why is $\mathbb{E}[X \mid \beta(1) \text{ is a solution}] \gg \mathbb{E}[X]$?

- There are many codewords $\beta(i)$ that are *dependent* with $\beta(1)$
- If $\beta(1), \beta(i)$ are dependent: *given that* $|\mathbf{S} - \mathbf{A}\beta(1)|^2 \leq D$, *the probability of* $|\mathbf{S} - \mathbf{A}\beta(i)|^2 \leq D$ *increases*



- Even a *small increase* in the probability may be enough to blow up $\mathbb{E}[X \mid \beta(1) \text{ is a solution}]$

A Stylized Example

Assume that the number of solutions X can only take one of two values

$$X = \begin{cases} 2^n & \text{with probability } 1 - 2^{-np} \\ 2^{1.1n} & \text{with probability } 2^{-np} \end{cases}$$

Note:

- There are always at least 2^n solutions $\Rightarrow$ $P(X > 0) = 1$

A Stylized Example

Assume that the number of solutions X can only take one of two values

$$X = \begin{cases} 2^n & \text{with probability } 1 - 2^{-np} \\ 2^{1.1n} & \text{with probability } 2^{-np} \end{cases}$$

Note:

- There are always at least 2^n solutions $\Rightarrow$ $P(X > 0) = 1$
- The expected number of solutions is

$$\begin{aligned} \mathbb{E}X &= 2^{1.1n}2^{-np} \; + \; 2^n(1 - 2^{-np}) \\ &\approx 2^n \quad \text{if } p > 0.1 \end{aligned}$$

A Stylized Example

Assume that the number of solutions X can only take one of two values

$$X = \begin{cases} 2^n & \text{with probability } 1 - 2^{-np} \\ 2^{1.1n} & \text{with probability } 2^{-np} \end{cases}$$

Note:

- There are always at least 2^n solutions $\Rightarrow$ $P(X > 0) = 1$
- The expected number of solutions is

$$\begin{aligned} \mathbb{E}X &= 2^{1.1n} 2^{-np} + 2^n (1 - 2^{-np}) \\ &\approx 2^n \quad \text{if } p > 0.1 \end{aligned}$$

- For 2nd MoM to predict existence of solutions, we need

$$\frac{\mathbb{E}[X]}{\mathbb{E}[X | \beta \text{ is a solution}]} \approx 1$$

Example ctd.

$$\mathbb{E}[X|\ \beta \text{ is a solution}]$$
$$= P(X = 2^{1.1n}|\beta \text{ is a soln.})\, 2^{1.1n} \ + \ P(X = 2^{n}|\beta \text{ is a soln.})\, 2^{n}$$

Example ctd.

$$
\begin{aligned}
&\mathbb{E}[X|\ \beta \text{ is a solution}] \\
&= P(X = 2^{1.1n} | \beta \text{ is a soln.})\, 2^{1.1n} \;+\; P(X = 2^{n} | \beta \text{ is a soln.})\, 2^{n} \\
&\approx \underbrace{\frac{2^{1.1n}\, 2^{-np}}{2^{n} + 2^{1.1n}\, 2^{-np}}}_{\approx\ 2^{-n(p-.1)}}\, 2^{1.1n} \;+\; \underbrace{\frac{2^{n}}{2^{n} + 2^{1.1n}\, 2^{-np}}}_{\approx\ 1-2^{-n(p-.1)}}\, 2^{n}
\end{aligned}
$$

Example ctd.

$$\mathbb{E}[X|\ \beta \text{ is a solution}]$$

$$= P(X = 2^{1.1n}|\beta \text{ is a soln.})\, 2^{1.1n} \ + \ P(X = 2^{n}|\beta \text{ is a soln.})\, 2^{n}$$

$$\approx \underbrace{\frac{2^{1.1n}\, 2^{-np}}{2^{n} + 2^{1.1n}\, 2^{-np}}}_{\approx\ 2^{-n(p-.1)}} 2^{1.1n} \ + \ \underbrace{\frac{2^{n}}{2^{n} + 2^{1.1n}\, 2^{-np}}}_{\approx\ 1-2^{-n(p-.1)}} 2^{n}$$

$$\mathbb{E}[X|\ \beta \text{ is a solution}] \approx \ 2^{1.2n}\, 2^{-np} \ + \ 2^{n}$$

$$\approx \begin{cases} 2^{n} & \text{if } p > 0.2 \\ 2^{1.2n-p} & \text{if } 0.1 < p < 0.2 \end{cases}$$

Example ctd.

$$
\begin{aligned}
&\mathbb{E}[X|\ \beta \text{ is a solution}] \\
&= P(X = 2^{1.1n}|\beta \text{ is a soln.})\, 2^{1.1n} \;+\; P(X = 2^{n}|\beta \text{ is a soln.})\, 2^{n} \\
&\approx \underbrace{\frac{2^{1.1n}\,2^{-np}}{2^{n} \,+\, 2^{1.1n}\,2^{-np}}}_{\approx\, 2^{-n(p-.1)}}\, 2^{1.1n} \;+\; \underbrace{\frac{2^{n}}{2^{n} \,+\, 2^{1.1n}\,2^{-np}}}_{\approx\, 1-2^{-n(p-.1)}}\, 2^{n}
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}[X|\ \beta \text{ is a solution}] &\approx \quad 2^{1.2n}\,2^{-np} \;+\; 2^{n} \\
&\approx \begin{cases} 2^{n} & \text{if } p > 0.2 \\ 2^{1.2n-p} & \text{if } 0.1 < p < 0.2 \end{cases}
\end{aligned}
$$

When $0.1 < p < 0.2$, the 2nd MoM fails because:

- Conditioned on β being a soln., probability of $X = 2^{1.1n}\ \uparrow$
- $\mathbb{E}[X|\ \beta$ is a solution$] \gg \mathbb{E}X$ *although* $X|\ \beta$ is a solution $\approx \mathbb{E}X$ w.h.p

Back to SPARCs

For a *low-probability set* of design matrices:

- Columns of β are unusually well-aligned with $\mathbf{S}$
- $\Rightarrow$ lots of *neighbours* of a solution are also solutions.
- Due to these atypical matrices,

$$\mathbb{E}[X \mid \beta \text{ is a solution}] \gg \mathbb{E}[X]$$

Back to SPARCs

For a *low-probability set* of design matrices:

- Columns of β are unusually well-aligned with $\mathbf{S}$
- $\Rightarrow$ lots of *neighbours* of a solution are also solutions.
- Due to these atypical matrices,

$$\mathbb{E}[X \,|\, \beta \text{ is a solution}] \gg \mathbb{E}[X]$$

Lemma

Given that β is a solution, the number of neighbours of β that are also solutions is less than $L^{-1/2}\,\mathbb{E}[X]$ with prob. at least $1 - L^{-2}$, when $b > b^*$

The lemma implies

$$X \,|\, \beta \text{ is a solution} \;\sim\; \mathbb{E}[X] \qquad \text{with prob. at least } 1 - L^{-2}$$

Fixing the 2nd MoM

Call a solution β *good* if fewer than $L^{-1/2}\,\mathbb{E}[X]$ of its neighbours are also solutions

- Lemma says w.h.p any solution β is good.

$$X_{good} = V_1 + V_2 + \ldots + V_{e^{nR}}$$

where

$$V_i = \begin{cases} 1 & \text{if } \beta(i) \text{ is a good solution} \\ 0 & \text{otherwise.} \end{cases},$$

Fixing the 2nd MoM

Call a solution β *good* if fewer than $L^{-1/2}\,\mathbb{E}[X]$ of its neighbours are also solutions

- Lemma says w.h.p any solution β is good.

$$X_{good} = V_1 + V_2 + \ldots + V_{e^{nR}}$$

where

$$V_i = \begin{cases} 1 & \text{if } \beta(i) \text{ is a good solution} \\ 0 & \text{otherwise.} \end{cases},$$

- Apply 2nd MoM to show that $X_g > 0$ *w.h.p.*

This works because $\mathbb{E}[X_{good} | \beta \text{ is a solution}] \approx \mathbb{E}X_{good} \approx \mathbb{E}X$

Summary

- To show $X > 0$, 2nd MoM method requires $\mathbb{E}[X|\beta] \approx \mathbb{E}X$
- This may not hold *although* $X|\beta \approx \mathbb{E}[X]$ w.h.p

Summary

- To show $X > 0$, 2nd MoM method requires $\mathbb{E}[X|\beta] \approx \mathbb{E}X$
- This may not hold *although* $X|\beta \approx \mathbb{E}[X]$ w.h.p

Two-step fix

1. Show that most solutions are *good*, i.e., not many neighbours are solutions
2. Apply 2nd MoM to count the good solutions

Summary

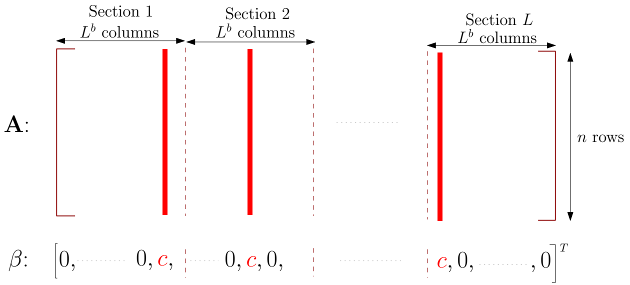
- To show $X > 0$, 2nd MoM method requires $\mathbb{E}[X|\beta] \approx \mathbb{E}X$
- This may not hold *although* $X|\beta \approx \mathbb{E}[X]$ w.h.p

Two-step fix

1. Show that most solutions are *good*, i.e., not many neighbours are solutions
2. Apply 2nd MoM to count the good solutions

- Similar situation in random hypergraph 2-colouring [Coja-Oghlan, Zdeberova SODA '12]
- Step 1 is key, and is problem-specific; Step 2 generic
- This two-step recipe potentially useful in many problems

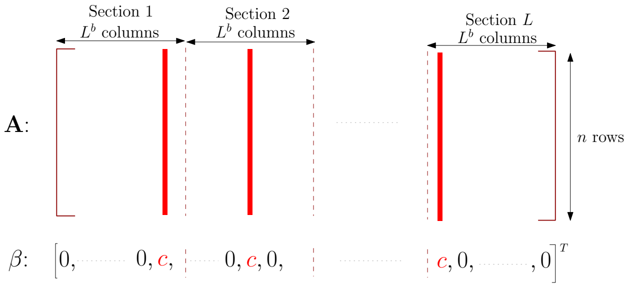
So far . . .



$$P_n = P\left(\tfrac{1}{n} \|\mathbf{S} - \mathbf{A}\hat{\beta}\|^2 > D \right)$$

For any ergodic source with variance σ^2 and distortion $D < \sigma^2$, $P_n \to 0$ for all rates $R > \frac{1}{2} \log \frac{\sigma^2}{D}$, when $b > b_{min}$

So far . . .



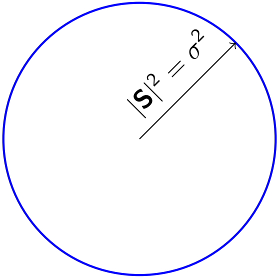
$$P_n = P\left(\tfrac{1}{n}\|\mathbf{S} - \mathbf{A}\hat{\beta}\|^2 > D \right)$$

For any ergodic source with variance σ^2 and distortion $D < \sigma^2$, $P_n \to 0$ for all rates $R > \frac{1}{2}\log\frac{\sigma^2}{D}$, when $b > b_{min}$

- We would also like to know the *error exponent*:
  $T = -\limsup_n \frac{1}{n}\log P_n \;\Rightarrow\; P_n \lesssim e^{-nT}$
- The 2nd MoM only gives a polynomial decay of P_n in n

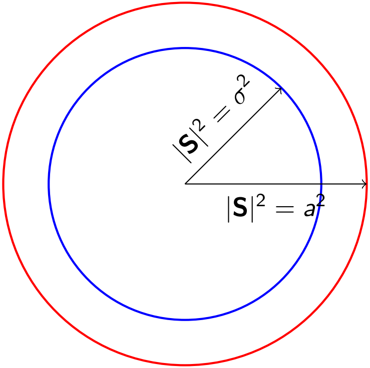
Refined Error Analysis for SPARC

$$\mathbf{S} \sim \text{ i.i.d. } \mathcal{N}(0, \sigma^2)$$



Refined Error Analysis for SPARC

$\mathbf{S} \sim$ i.i.d. $\mathcal{N}(0, \sigma^2)$

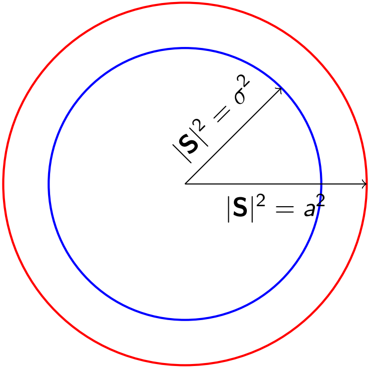


$R = \frac{1}{2} \log \frac{a^2}{D}$

$$P_n < P(|\mathbf{S}|^2 \geq a^2) \; + \; P(\text{ error } \mid |\mathbf{S}|^2 < a^2)$$

Refined Error Analysis for SPARC

$\mathbf{S} \sim$ i.i.d. $\mathcal{N}(0, \sigma^2)$



$R = \frac{1}{2} \log \frac{a^2}{D}$

$$P_n < P(|\mathbf{S}|^2 \geq a^2) \; + \; P(\text{ error } \mid |\mathbf{S}|^2 < a^2)$$

- The first term $\leq \exp(-n\mathcal{D}(a^2 \parallel \sigma^2))$
- $\mathcal{D}(a^2 \parallel \sigma^2)$: KL divergence between $\mathcal{N}(0, a^2)$ and $\mathcal{N}(0, \sigma^2)$

Error Analysis

$$P_n \; < \underbrace{P(|\mathbf{S}|^2 \geq a^2)}_{\text{KL divergence}} + \underbrace{P(\text{ error } \; | \; |\mathbf{S}|^2 < a^2)}_{?}.$$

$$P(\text{ error } \; | \; |\mathbf{S}|^2 < a^2) = P\left(X = 0 \mid |\mathbf{S}|^2 < a^2\right)$$

where $X = \sum_{i=1}^{e^{nR}} U_i$ and

$$U_i = \begin{cases} 1 & \text{if } \beta(i) \text{ is a solution }, \\ 0 & \text{otherwise.} \end{cases}$$

Error Analysis

$$P_n < \underbrace{P(|\mathbf{S}|^2 \geq a^2)}_{\text{KL divergence}} + \underbrace{P(\text{ error } \mid |\mathbf{S}|^2 < a^2)}_{?}.$$

$$P(\text{ error } \mid |\mathbf{S}|^2 < a^2) = P\left(X = 0 \mid |\mathbf{S}|^2 < a^2\right)$$

where $X = \sum_{i=1}^{e^{nR}} U_i$ and

$$U_i = \begin{cases} 1 & \text{if } \beta(i) \text{ is a solution }, \\ 0 & \text{otherwise.} \end{cases}$$

Error Analysis

$$P_n \; < \underbrace{P(|\mathbf{S}|^2 \geq a^2)}_{\text{KL divergence}} + \underbrace{P(\text{ error } \mid |\mathbf{S}|^2 < a^2)}_{?}.$$

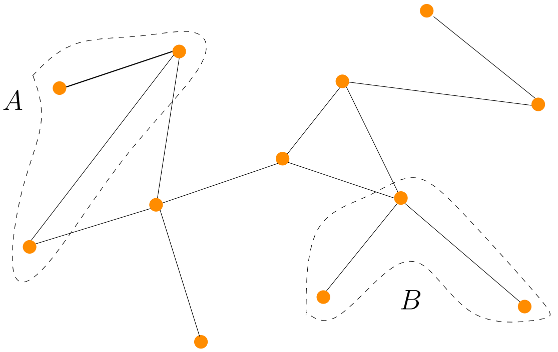
$$P(\text{ error } \mid |\mathbf{S}|^2 < a^2) = P\left(X = 0 \mid |\mathbf{S}|^2 < a^2\right)$$

where $X = \sum_{i=1}^{e^{nR}} U_i$ and

$$U_i = \begin{cases} 1 & \text{if } \beta(i) \text{ is a solution }, \\ 0 & \text{otherwise.} \end{cases}$$

We get a sharp bound on $P\left(X = 0 \mid |\mathbf{S}|^2 < a^2\right)$ using *Suen's inequality*

Dependency Graph



For random variables $\{U_i\}_{i \in \mathcal{I}}$, any graph with vertex set $\mathcal{I}$ s.t:

> *If A and B are two disjoint subsets of $\mathcal{I}$ such that there are no edges with one vertex in A and the other in B, then the families $\{U_i\}_{i \in A}$ and $\{U_i\}_{i \in B}$ are independent.*

For our problem ...

$$U_i = \begin{cases} 1 & \text{if } \beta(i) \text{ is a solution ,} \\ 0 & \text{otherwise.} \end{cases}, \quad i = 1, \dots, e^{nR}$$

For the family $\{U_i\}$,

$$\{i \sim j : i \neq j \text{ and } \beta(i), \beta(j) \text{ share at least one common term}\}$$

is a dependency graph.

Suen's correlation inequality

Let $\{U_i\}_{i\in\mathcal{I}}$, be Bernoulli rvs with dependency graph Γ. Then

$$P\left(\sum_{i\in\mathcal{I}} U_i = 0\right) \le \exp\left(-\min\left\{\frac{\lambda}{2}, \frac{\lambda^2}{8\Delta}, \frac{\lambda}{6\delta}\right\}\right)$$

where

$$\lambda = \sum_{i\in\mathcal{I}} \mathbb{E} U_i,$$

$$\Delta = \frac{1}{2}\sum_{i\in\mathcal{I}}\sum_{j\sim i} \mathbb{E}(U_i U_j),$$

$$\delta = \max_{i\in\mathcal{I}} \sum_{k\sim i} \mathbb{E} U_k.$$

Bounding the error

$$P_n \leq P(|\mathbf{S}|^2 \geq a^2) \; + \; P\left(\sum_{i=1}^{e^{nR}} U_i = 0 \mid |\mathbf{S}|^2 < a^2 \right)$$

$$\leq \exp\left(-n\mathcal{D}(a^2 \parallel \sigma^2)\right) \; + \; \exp\left(- \min\left\{ \frac{\lambda}{2}, \frac{\lambda}{6\delta}, \frac{\lambda^2}{8\Delta} \right\} \right)$$

Bounding the error

$$P_n \leq P(|\mathbf{S}|^2 \geq a^2) \; + \; P\left(\sum_{i=1}^{e^{nR}} U_i = 0 \mid |\mathbf{S}|^2 < a^2\right)$$
$$\leq \exp\left(-n\mathcal{D}(a^2 \parallel \sigma^2)\right) \; + \; \exp\left(-\min\left\{\frac{\lambda}{2}, \frac{\lambda}{6\delta}, \frac{\lambda^2}{8\Delta}\right\}\right)$$

where for sufficiently large n

$$\lambda > e^{n\left(R - \frac{1}{2}\log\frac{a^2}{D} - \epsilon_n\right)}, \qquad \frac{\lambda}{\delta} > L^{b-1}, \qquad \frac{\lambda^2}{\Delta} > L^{(b-b_{min})(1-(1-D/a^2)/R)}$$

- For large n, the first KL divergence term dominates P_n
- $\lambda, \frac{\lambda}{\delta}, \frac{\lambda^2}{\Delta}$ all grow polynomially in n for $b > b^*$

  $\Rightarrow$ second term decays *super-exponentially*
- Need to use refinement technique when $R < (1 - D/a^2)$

Error Exponent of SPARC with Min-Distance Encoding

$$P_n = P\left(\tfrac{1}{n}\|\mathbf{S} - \mathbf{A}\hat{\beta}\|^2 > D \right)$$

Theorem (RV,Joseph,Tatikonda '12, '14)

1. *For $R > \frac{1}{2}\log\frac{\sigma^2}{D}$, the probability of error P_n decays exponentially in n for $b > b^*$*
2. *The error-exponent $\mathcal{D}(a^2 \,\|\, \sigma^2)$, with $a^2 = De^{2R}$, is optimal for Gaussian sources with squared-error distortion.*

Error Exponent of SPARC with Min-Distance Encoding

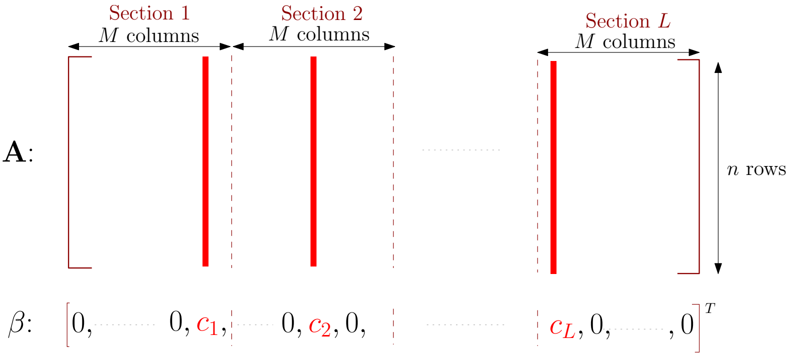
$$P_n = P\left(\tfrac{1}{n}\|\mathbf{S} - \mathbf{A}\hat{\beta}\|^2 > D \right)$$

Theorem (RV,Joseph,Tatikonda '12, '14)

1. *For $R > \frac{1}{2}\log\frac{\sigma^2}{D}$, the probability of error P_n decays exponentially in n for $b > b^*$*
2. *The error-exponent $\mathcal{D}(a^2 \parallel \sigma^2)$, with $a^2 = De^{2R}$, is optimal for Gaussian sources with squared-error distortion.*

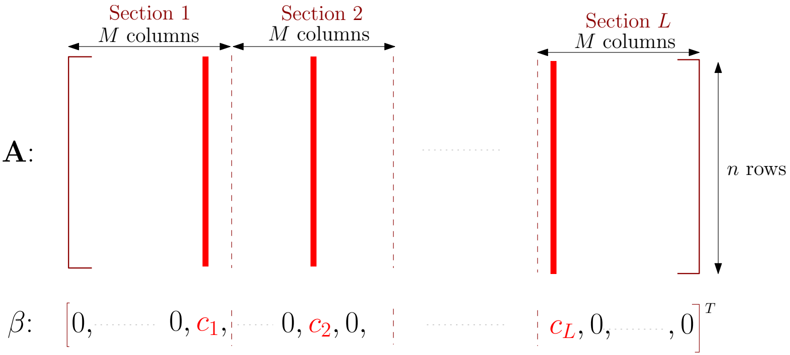
- This result shows that SPARCs are structurally good codes
- But minimum-distance encoding is infeasible — what about practical algorithms?

SPARC Construction



Main Idea: Vary the coefficients across sections

SPARC Construction

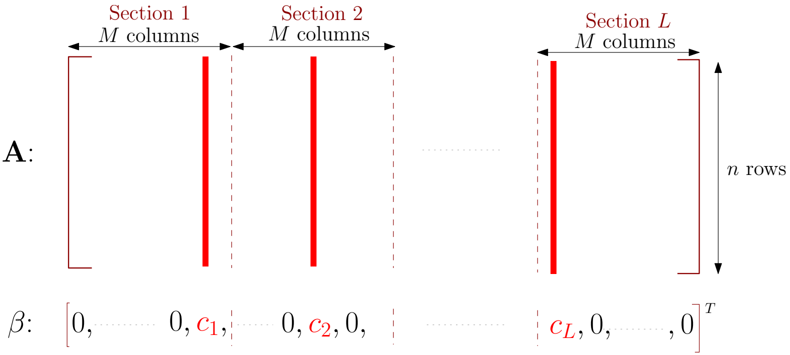


Main Idea: Vary the coefficients across sections

As before:

- For rate R codebook, need $M^L = e^{nR}$

SPARC Construction

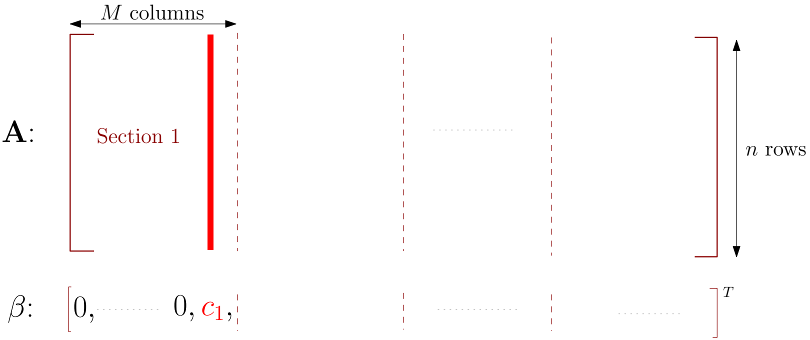


Main Idea: Vary the coefficients across sections

As before:

- For rate R codebook, need $M^L = e^{nR}$
- Choose M polynomial of $n \Rightarrow L \sim n/\log n$

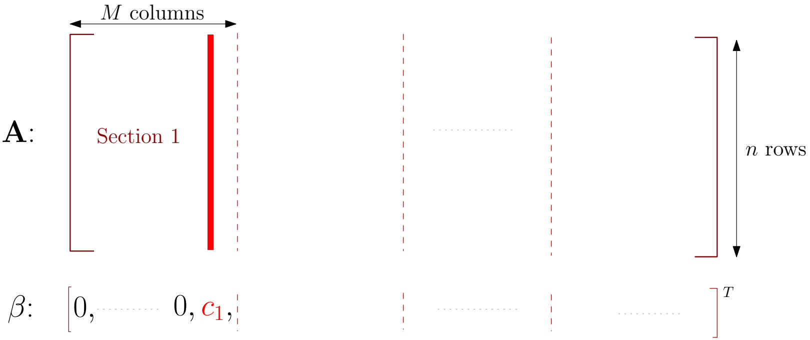
An Encoding Algorithm



Step 1: Choose column in Sec.1 that minimizes $\|\mathbf{S} - c_1 \mathbf{A}_j\|^2$

- $c_1 = \sqrt{2R\sigma^2/L}$

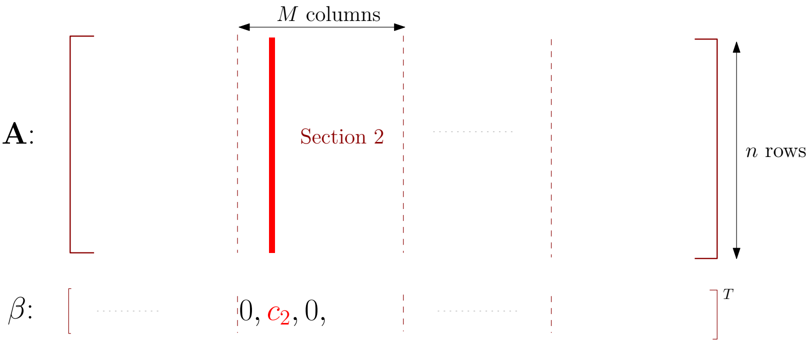
An Encoding Algorithm



Step 1: Choose column in Sec.1 that minimizes $\|\mathbf{S} - c_1 \mathbf{A}_j\|^2$

- $c_1 = \sqrt{2R\sigma^2/L}$
- Max among inner products $\langle \mathbf{S}, \mathbf{A}_j \rangle$
- Residue $\mathbf{R}_1 = \mathbf{S} - c_1 \hat{\mathbf{A}}_1$

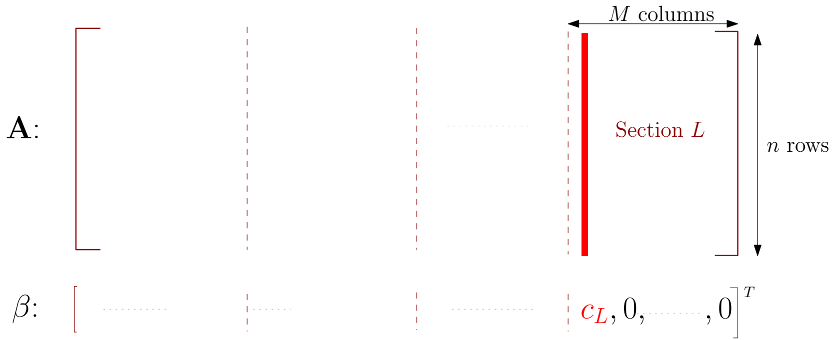
An Encoding Algorithm



Step 2: Choose column in Sec.2 that minimizes $\|\mathbf{R}_1 - c_2\mathbf{A}_j\|^2$

- $c_2 = \sqrt{\frac{2R\sigma^2}{L}\left(1 - \frac{2R}{L}\right)}$
- Max among inner products $\langle \mathbf{R}_1, \mathbf{A}_j \rangle$
- Residue $\mathbf{R}_2 = \mathbf{R}_1 - c_2\hat{\mathbf{A}}_2$

An Encoding Algorithm



Step L: Choose column in Sec. L that minimizes $\|\mathbf{R}_{L-1} - c_L \mathbf{A}_j\|^2$

- $c_L = \sqrt{\frac{2R\sigma^2}{L}\left(1 - \frac{2R}{L}\right)^L}$
- Max among inner products $\langle \mathbf{R}_{L-1}, \mathbf{A}_j \rangle$
- Residue $\mathbf{R}_L = \mathbf{R}_{L-1} - c_L \hat{\mathbf{A}}_L$

Performance

Theorem (RV, Sarkar, Tatikonda '13)

For an ergodic source $\mathbf{S}$ *with mean* 0 *and variance* σ^2*, the encoding algorithm produces a codeword* $\mathbf{A}\hat{\beta}$ *that satisfies the following for sufficiently large* M, L.

$$P\left(\, |\mathbf{S} - \mathbf{A}\hat{\beta}|^2 \; > \; \sigma^2 e^{-2R} + \Delta \, \right) < \exp\left(-\kappa n \left(\Delta - \frac{c \log\log M}{\log M}\right)\right)$$

Deviation Δ is $O(\frac{\log\log n}{\log n})$

Performance

Theorem (RV, Sarkar, Tatikonda '13)

For an ergodic source $\mathbf{S}$ *with mean* 0 *and variance* σ^2*, the encoding algorithm produces a codeword* $\mathbf{A}\hat{\beta}$ *that satisfies the following for sufficiently large* M, L*.*

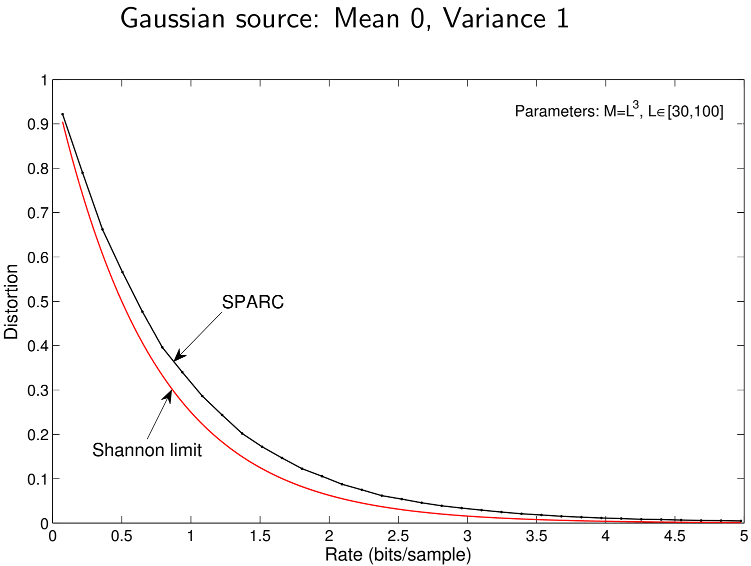
$$P\left(|\mathbf{S} - \mathbf{A}\hat{\beta}|^2 \; > \; \sigma^2 e^{-2R} + \Delta \right) < \exp\left(-\kappa n \left(\Delta - \frac{c \log \log M}{\log M} \right) \right)$$

Deviation Δ is $O(\frac{\log \log n}{\log n})$

Encoding Complexity

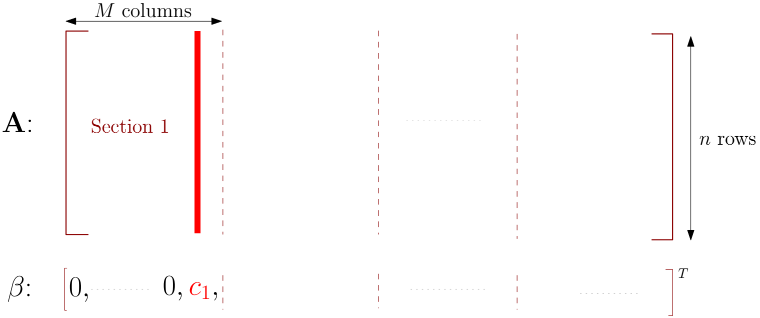
ML inner products and comparisons $\Rightarrow$ *polynomial* in n

Simulation



Gaussian source: Mean 0, Variance 1
Parameters: M=L^3, L∈[30,100]
SPARC
Shannon limit
Distortion
Rate (bits/sample)
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5

Why does the algorithm work?

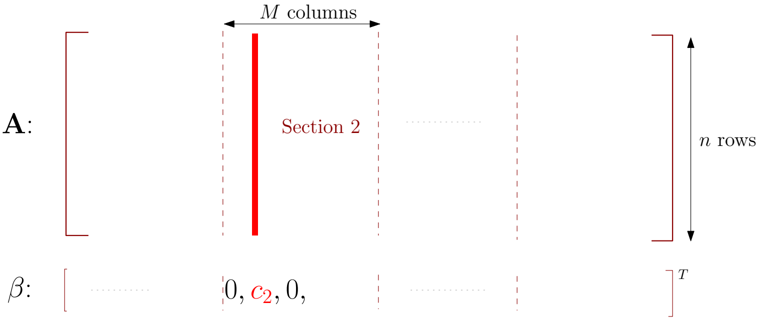


Each section is a code of rate $R/L \quad (L \sim \frac{n}{\log n})$

- Step 1: $\mathbf{S} \quad \longrightarrow \quad \mathbf{R}_1 = \mathbf{S} - c_1 \hat{\mathbf{A}}_1$

$$|\mathbf{R}_1|^2 \approx \sigma^2 e^{-2R/L} \approx \sigma^2 \left(1 - \frac{2R}{L}\right) \qquad \text{for } c_1 = \sqrt{2R\sigma^2/L}$$

Why does the algorithm work?



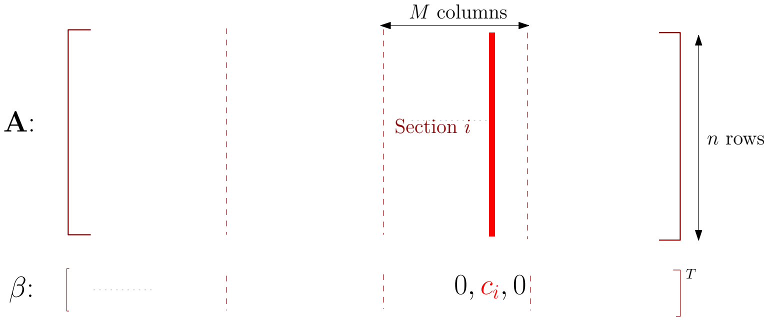
Each section is a code of rate R/L $\quad (L \sim \frac{n}{\log n})$

- Step 1: $\mathbf{S} \quad \longrightarrow \quad \mathbf{R}_1 = \mathbf{S} - c_1 \hat{\mathbf{A}}_1$

$$|\mathbf{R}_1|^2 \approx \sigma^2 e^{-2R/L} \approx \sigma^2 \left(1 - \frac{2R}{L}\right) \qquad \text{for } c_1 = \sqrt{2R\sigma^2/L}$$

- Step 2: 'Source' $\mathbf{R}_1 \quad \longrightarrow \quad \mathbf{R}_2 = \mathbf{R}_1 - c_2 \hat{\mathbf{A}}_2$

Why does the algorithm work?



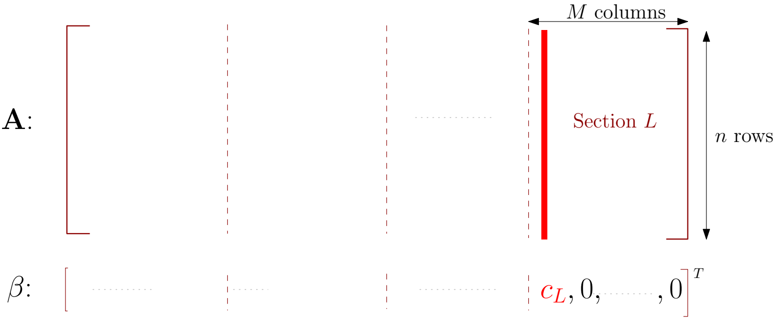
Each section is a code of rate $R/L \quad (L \sim \frac{n}{\log n})$

- Step i: 'Source' $\mathbf{R}_{i-1} \quad \longrightarrow \quad \mathbf{R}_i = \mathbf{R}_{i-1} - c_i \hat{\mathbf{A}}_2$

  With $c_i^2 = \frac{2R\sigma^2}{L}\left(1 - \frac{2R}{L}\right)^{i-1}$,

$$|\mathbf{R}_i|^2 \approx |\mathbf{R}_{i-1}|^2 \left(1 - \frac{2R}{L}\right) \approx \sigma^2 \left(1 - \frac{2R}{L}\right)^i$$

Why does the algorithm work?

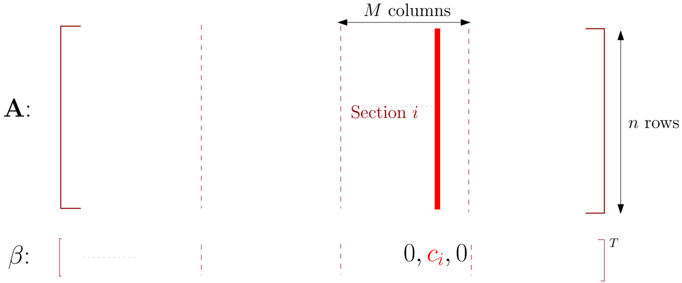


Each section is a code of rate R/L $\quad (L \sim \frac{n}{\log n})$

$$\text{Final Distortion:} \quad |\mathbf{R}_L|^2 \approx \sigma^2 \left(1 - \frac{2R}{L}\right)^L \leq \sigma^2 e^{-2R}$$

L-stage successive refinement $\quad L \sim n / \log n$

Successive Refinement Interpretation



- The encoder successively refines the source over $\sim \frac{n}{\log n}$ stages
- The deviations in each stage can be significant!

$$|\mathbf{R}_i|^2 = \underbrace{\sigma^2 \left(1 - \frac{2R}{L}\right)^i}_{\text{'Typical Value'}} (1 + \Delta_i)^2, \quad i = 0, \ldots, L$$

- KEY to result: Controlling the final deviation Δ_L

Proof involves controlling deviations due to:

Proof involves controlling deviations due to:

- *Source*: $|\mathbf{S}|^2 = \sigma^2(1 + \Delta_0)^2$

Proof involves controlling deviations due to:

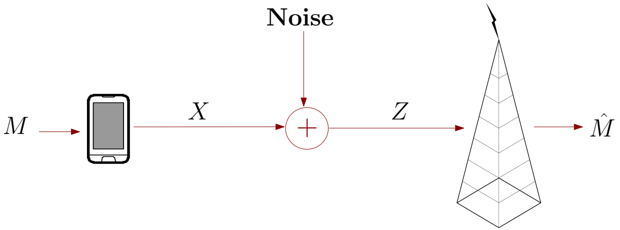
- *Source*: $|\mathbf{S}|^2 = \sigma^2(1 + \Delta_0)^2$
- *Dictionary columns*: $|\mathbf{A}_j|^2 = 1 + \gamma_j, \quad 1 \le j \le ML$

Proof involves controlling deviations due to:

- *Source*: $|\mathbf{S}|^2 = \sigma^2(1+\Delta_0)^2$
- *Dictionary columns*: $|\mathbf{A}_j|^2 = 1+\gamma_j, \quad 1 \le j \le ML$
- *Computed value*:

$$\max_j \left\langle \frac{\mathbf{R}_{i-1}}{\|\mathbf{R}_{i-1}\|}, \mathbf{A}_j \right\rangle = \sqrt{2\log M}\,(1+\epsilon_i), \quad 1 \le i \le L$$

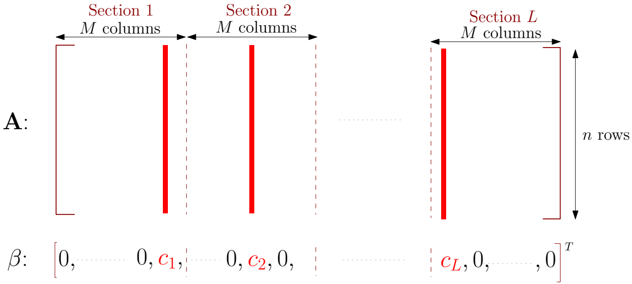
SPARCs for Communicating over Gaussian Channels



$Z = X +$ Noise $\qquad \frac{\|\mathbf{X}\|^2}{n} \leq P, \qquad \text{Noise} \sim \mathcal{N}(0, N)$

GOAL: Achieve rates close to capacity $\mathcal{C} = \frac{1}{2}\log\left(1 + \frac{P}{N}\right)$

Efficient Decoder



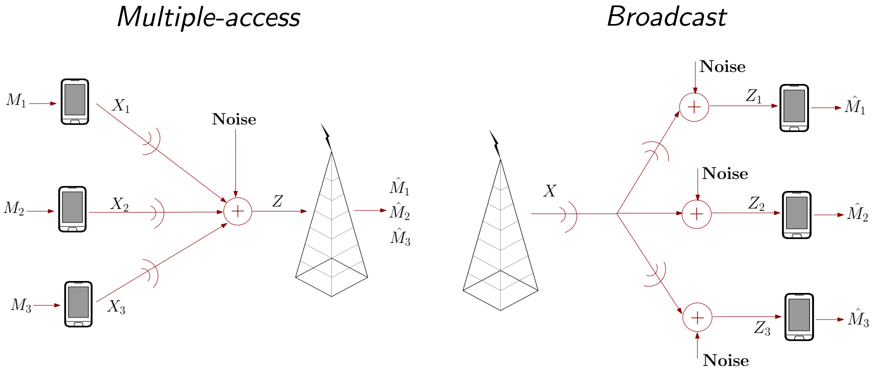
$$\mathbf{Z} = \mathbf{A}\beta + \text{Noise}$$

- Each β corresponds to a message $\Rightarrow$ M^L messages
- Efficient decoders proposed by [Barron-Joseph '12], [Barron-Cho '13]:

Achieve rates $R < \mathcal{C} - O\left(\frac{\log\log M}{\log M}\right)$ with $P_e < e^{-cL(\mathcal{C}-R)^2}$

Multi-terminal networks

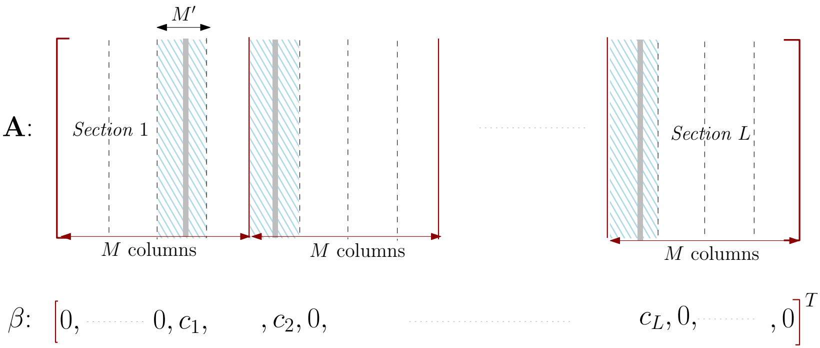
Examples:



Key ingredients

- Superposition (Multiple-access, Broadcast)
- Random binning (e.g., distributed compression, source/channel coding with side-information)

Binning with SPARCs



[RV-Tatikonda, Allerton '12]

Any random coding scheme that consists of point-to-point source and channel codes combined via binning/superposition can be implemented with SPARCs.

Summary

Sparse Regression Codes

- Rate-optimal for Gaussian compression and communication
- Low-complexity coding algorithms that provably attain Shannon limits

Future Directions

- Better channel decoders and source encoders:

  Approximate message passing, ℓ_1 minimization etc.?
- Simplified design matrices

  Can we prove that the results hold for ± 1 design matrices
- Network information theory: Multiple descriptions, Interference channels ...
- Finite-field analogues: binary SPARCs?

Papers at `http://www2.eng.cam.ac.uk/~rv285/pub.html`