Communication by Regression,
Reliable at Rates up to Channel Capacity

Andrew Barron
Department of Statistics
Yale University

Coauthors: Antony Joseph, David Smalling

May 17, 2009
Innovation and Inventiveness in Statistical Methodology

Workshop honoring John Hartigan

Shannon Formulation

- Input bit string: $u = (u_1, u_2, \ldots\ldots, u_K)$

$\downarrow$

- Encoded string: $x = (x_1, x_2, \ldots, x_n)$

$\downarrow$

- Channel: $p(y|x)$

$\downarrow$

- Received string: $y = (y_1, y_2, \ldots, y_n)$

$\downarrow$

- Decoded string: $\hat{u} = (\hat{u}_1, \hat{u}_2, \ldots\ldots, \hat{u}_K)$

- Rate: $R = \frac{K}{n}$ input bits per uses of channel
- Reliability: Want small probability of error $\{\hat{u} \neq u\}$

Gaussian Noise Channel

- Input bit string: $u = (u_1, u_2, \ldots\ldots, u_K)$
- Encoded string: $x = (x_1, x_2, \ldots, x_n)$
- Power: $|x|^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2$

  Power constraint: $|x|^2 \leq P$
- Gaussian noise: $\varepsilon = (\epsilon_1, \epsilon_2, \ldots, \epsilon_n)$

  Noise variance: σ^2

  Signal to noise: $v = \frac{P}{\sigma^2}$
- Received string: $y = (y_1, y_2, \ldots, y_n)$

  $y = x + \varepsilon$
- Decoded string: $\hat{u} = (\hat{u}_1, \hat{u}_2, \ldots\ldots, \hat{u}_K)$

Shannon Theory

- Channel Capacity:
  Supremum of rates R such that reliable communication is possible, with arbitrarily small error probability
- Information Capacity: $C = \max_{P_X} I(X;Y)$

  Where $I(X;Y)$ is the mutual information, aka Kullback divergence between $P_{X,Y}$ and $P_X \times P_Y$
- Shannon Channel Capacity Theorem:
  The supremum of achievable communication rates R equals the information capacity C
- Books:

  Shannon (49), Gallager (68), Cover & Thomas (06)

Gaussian Channel Capacity

- Gaussian Channel Capacity:

$$C = \max_{E|X|^2 \le P} I(X;Y) = \frac{1}{2}\log\left(1 + \frac{P}{\sigma^2}\right)$$

  Normal$(0, P)$ is the maximizing distribution on X

- Foundational model for
  wireless communication
  (radio waves, cell phones, television, satellite, space)
  wired communication
  (internet, telephone, cable)

- Relation to sphere packing: Conway and Sloane (88)

- No fast encoding and decoding algorithm has been mathematically proven to achieve rates up to capacity

Binary Channel Capacity

- Binary Symmetric Channel:
  Bits are received in error a fraction α of the time
- Capacity: $1 - h(\alpha)$, where
  $h(\alpha) = \alpha \log \frac{1}{\alpha} + (1-\alpha) \log \frac{1}{1-\alpha}$
- No fast encoding and decoding algorithm has been mathematically proven to achieve rates up to capacity
- Bayesian belief propagation: empirically shown to achieve near capacity performance in a particular Gallager inspired design by Luby, Mitzenmacher,Shokrollahi,Spielman (01)
- Reed-Solomon codes: Algebraic code on finite fields GF(2^m) corrects a fraction of ϵ errors with rate $1 - 2\epsilon$. Convertable to code for binary strings of same rate, near 1 for small ϵ, with guaranteed correction of a fraction $\alpha = \epsilon / \log n$ of errors. MacWilliams & Sloane (77)

Regression Formulation for the Gaussian Channel

- Random design matrix X is n by N $\quad (p = N)$
- Input bit string: u of length K

  $\downarrow$
- Coefficient vector: β of length N, with constraints on form

  $\downarrow$
- Codeword sent: $X\beta$ of length n

  $\downarrow$
- Received string: $Y = X\beta + \varepsilon$

  $\downarrow$
- Least squares: $\hat{\beta} = \text{argmin} \| Y - X\beta \|^2$

  This decoder maximizes likelihood; maximizes posterior probability given the data X, Y; minimizes probability of error with uniform distribution on input strings

Sparse Superposition Code Formulation

- Design matrix: X is n by N, each entry indep. N$(0, P/L)$
- Constraint on form of β of length N:
  Only L nonzero coordinates, of absolute value 1
  Count $2^K \leq \binom{N}{L}$ near $(Ne/L)^L$; more available if use signs
- Codewords: $X\beta$ of length n; each entry indep. N$(0, P)$
  Codewords are sums of L columns of X
- $\hat{\beta}$ decoded by least squares is reliable if correctly determine most of the L terms sent.
  Yields small bit error rate $\frac{1}{K}\sum_{i=1}^{K} 1_{\{\hat{u}_i \neq u_i\}}$
  Also small probability of error $\{\hat{u} \neq u\}$?

Partitioned Superposition Code

- Split u into L sections, each of length $\log B$ bits

  Total length $K = L \log B$.

- Split β into L sections, each of length B,
  with one non-zero value in each section.

  Total length $N = LB$.

- Input mapping:
  In each section, the bit string of length $\log B$ specifies
  in binary the location of the non-zero coefficient value.
  Optional: an extra bit per section used to specify the sign.

- Split columns of X into L sections, B choices in each.

- Codeword: a sum of L columns, one from each section

Communication Rate

- Codewords formed from L terms, one from each section
- Number of codewords: $2^K = B^L$
- Number of input bits communicated: $K = L \log B$
- Section size $B = L^a$ and dictionary size $N = L^{1+a}$
- Number of input bits communicated for a rate R code

$$nR = K = L \log B = aL \log L$$

- Choice of sample size to achieve a rate R code

$$n = (a/R)\, L \log L$$

Sufficient section size rate

- Polynomial section size:

$$B = L^a$$

- An expression a_v is determined:

  a_v is decreasing with signal to noise ratio v

  a_v is near 1 for large v

- Significance:

  If $a > a_v$, then the error probability of least squares is shown to be exponentially small for any communication rate $R < C$.

Precursors

- Forney (66) Concatenated codes
- Cover (72) Superposition codes for broadcast channels. Section sizes $B_1 = 2^{nR_1}, \ldots, B_L = 2^{nR_L}$ exponentially large. Codeword sent is sum of codewords for respective users. Similar setting for multiple access channels Rimoldi and Urbanke (01), Cao and Yeh (07).
- Wainwright (09) Information theory bounds on size of sets of sparse coefficients correctly decodeable by least squares. Correspond to positive rates (associated with $n = const\, L \log N/L$), but not all the way to capacity. Related work Candes & Tao (06), Candes & Plan (08).
- Interpretation of our conclusion: Compressed sensing capacity. Minimal number of measurements of average power P needed to determine the L out of N terms with small average prob error is $n = (1/C)L \log(N/L)$ where C is Shannon's channel capacity.

Error Probability Bound

- Codeword sent: $X\beta^*$
- Least squares or approximate least squares estimate $\hat{\beta}$ satisfies $|Y - X\hat{\beta}|^2 \leq |Y - X\beta^*|^2$
- Error event of a fraction of $\alpha = \ell/L$ section mistakes, contained in the event

  $E_\alpha = \{|Y - X\beta| \leq |Y - X\beta^*|^2 \text{ for some } \beta \in Wrong_\alpha\}$

  where $Wrong_\alpha$ is the set of β differing from β^* in ℓ sections.

- Error probability: Bound on $\mathbb{P}[E_\alpha]$ using

$$\binom{L}{\alpha L} exp\{-nD_\alpha\}$$

  where the exponent D_α is sufficiently large to cancel the combinatorial coefficient and produce an exponentially small error, provided the section size $a > a_v$ and $R < C$.

Some ingredients of the error exponent

Ingredients in $D_\alpha = D(\Delta_\alpha, \rho_\alpha^2)$

$$\Delta_\alpha = \alpha(C - R) + (C_\alpha - \alpha C)$$

$$C_\alpha = (1/2)\log(1 + \alpha v)$$

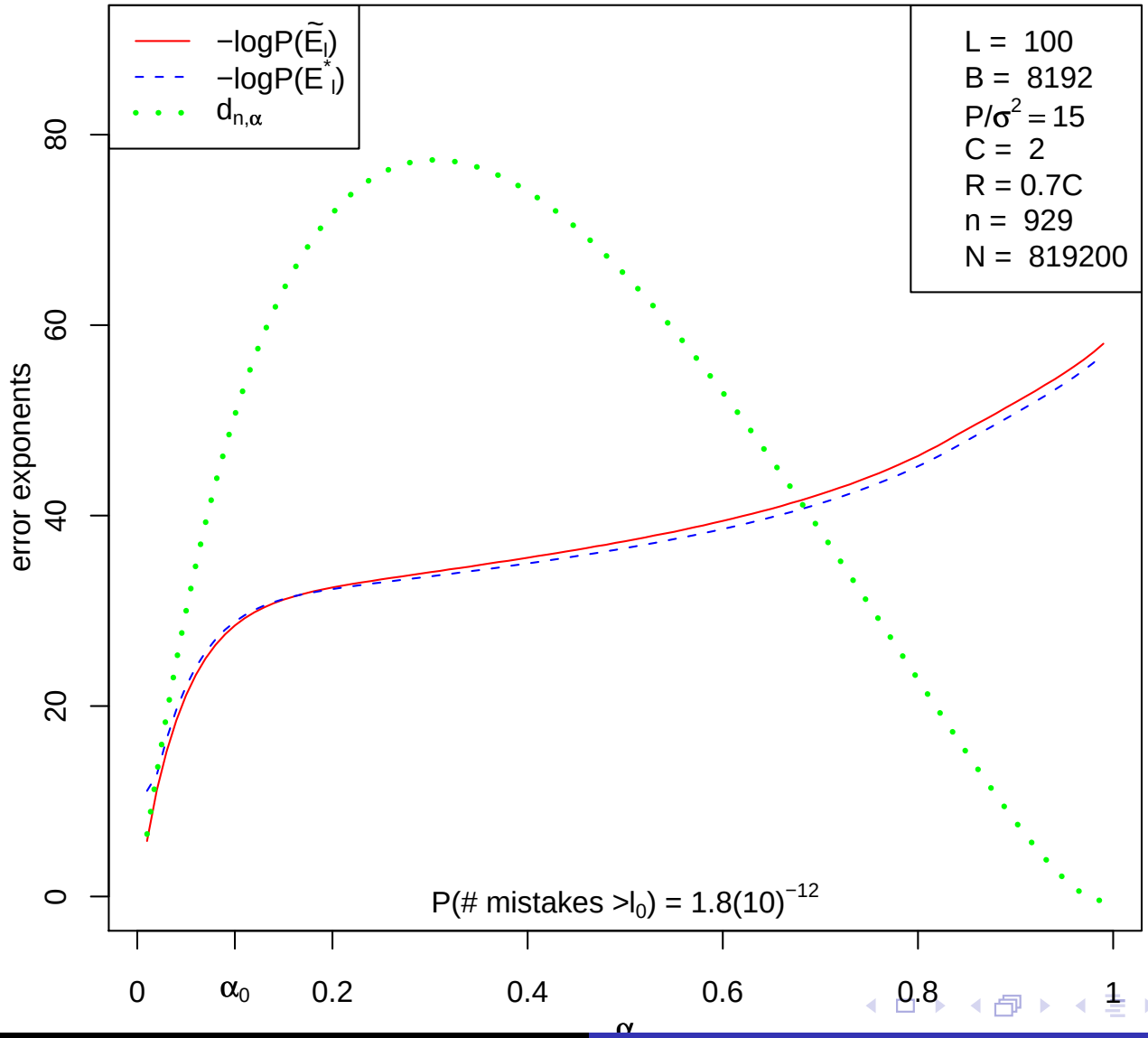
$$1 - \rho_\alpha^2 = \alpha(1-\alpha)v/(1 + \alpha^2 v)$$

where capacity $C = \frac{1}{2}\log(1+v)$ and $v = P/\sigma^2$ is signal/noise.

Here $D(\Delta, \rho^2)$ is the large deviation exponent associated with the cumulant generating function for $(1/2)(Z_1^2 - Z_2^2)$ with Z_1, Z_2 bivariate normal, mean zero, unit variance and correlation ρ. Near $(1/2)\Delta^2/(1-\rho^2)$ for small Δ.

Complete story includes tradeoff with another term.

Contributions to Error Exponent



Iterative decoding for approximate least squares

- Convex hull algorithm
  Let $\mathcal{A}$ be the convex hull of the allowed β.
  Initialize with β maximizing inner product $Y \cdot (X\beta)$.
  Relaxed greedy update (as in Jones (92)):

$$\beta(k+1) = (1-w)\beta(k) + w\beta^{update}$$

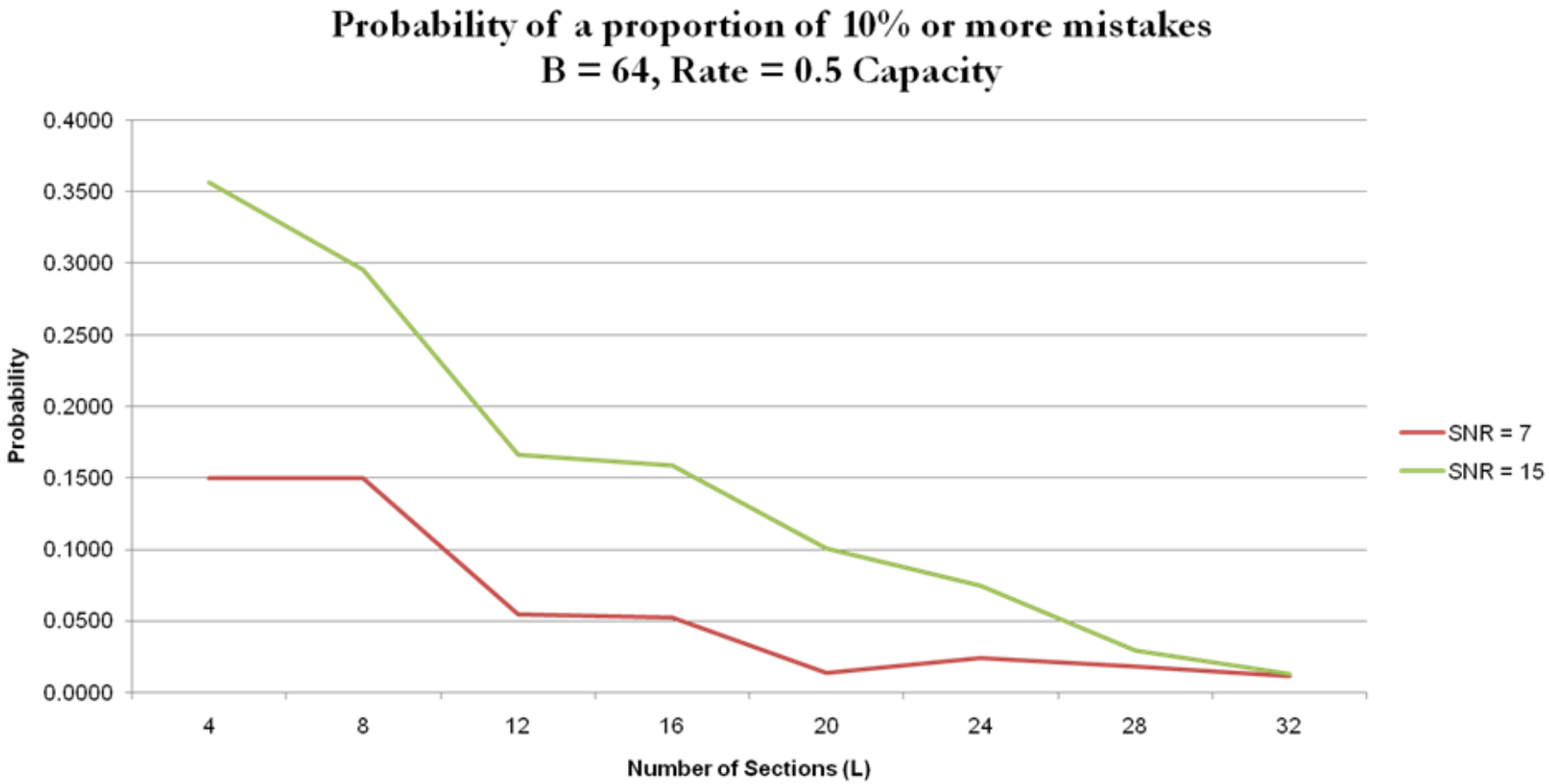
  Update chosen to maximize the inner product $Res_k \cdot (X\beta)$ with the residuals $Res_k = Y - (1-w_k)X\beta(k)$.
  Here w in $[0,1]$ is optimized by least squares.
- Computation and accuracy tradeoff

$$|Y - X\beta(k)| \leq |Y - X\beta^{proj}|^2 + \frac{4|X|^2}{k}$$

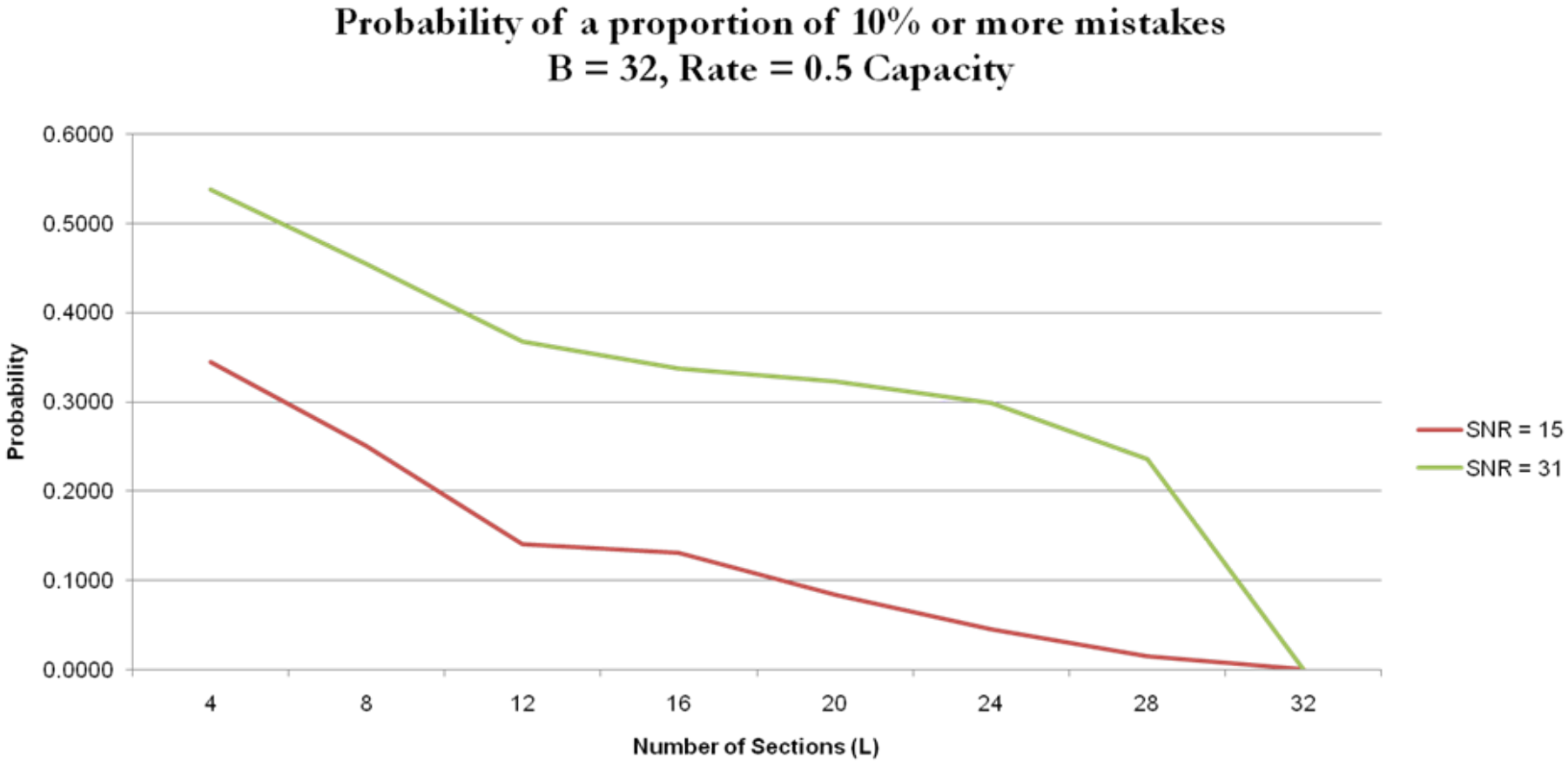
- Vertex move algorithm Similar analysis for randomized vertex move algorithm in the manuscript.

Simulations using approximate least squares

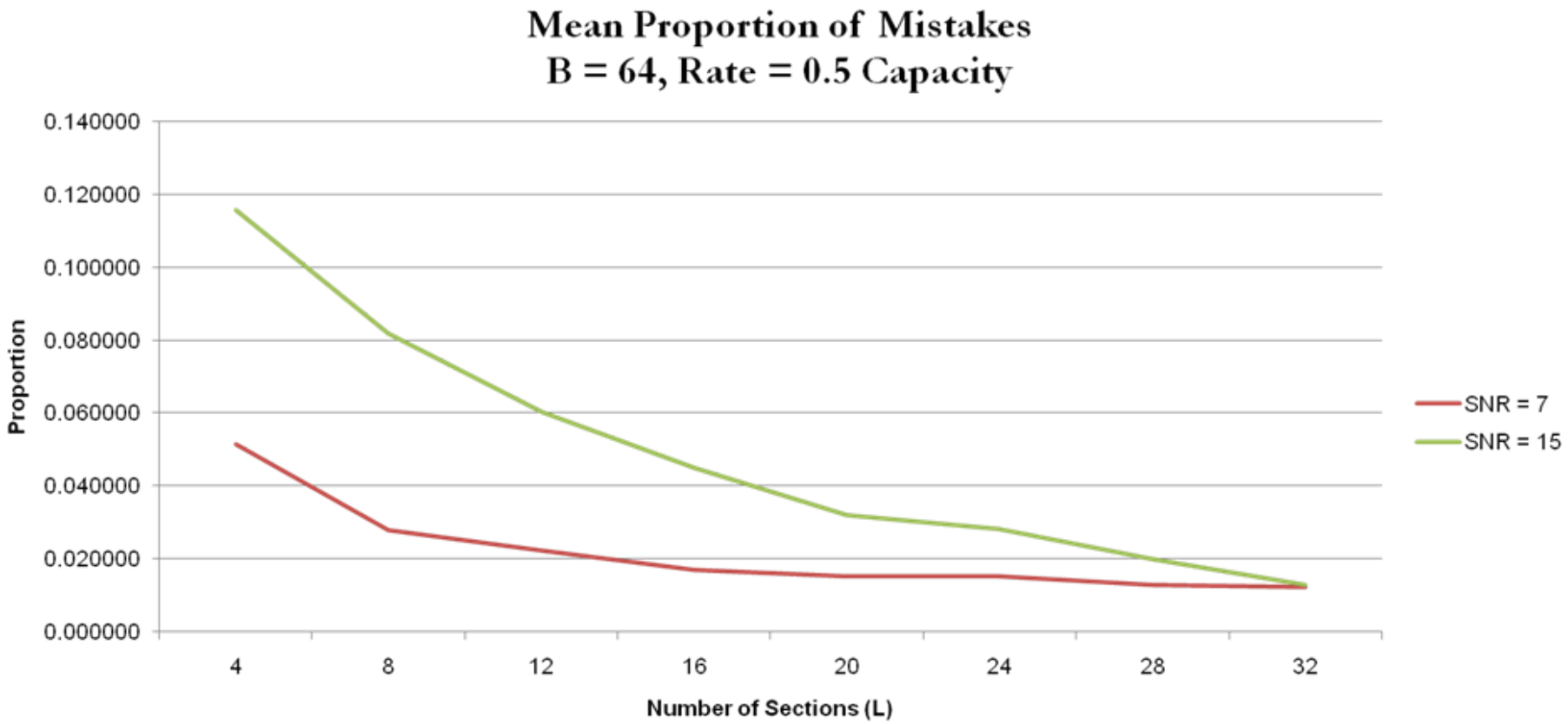
- Estimator $\hat{\beta}$ found by first doing the convex hull optimization getting $\beta(k)$ and then taking the closest vertex.
- Simulations performed by Yale senior David Smalling for his senior project in Applied Mathematics



L	SNR = 7	SNR = 15
4	0.1500	0.3565
8	0.1500	0.2955
12	0.0540	0.1665
16	0.0520	0.1590
20	0.0130	0.1010
24	0.0240	0.0750
28	0.0180	0.0295
32	0.0110	0.0130



L	SNR = 15	SNR = 31
4	0.3450	0.5380
8	0.2505	0.4540
12	0.1405	0.3680
16	0.1305	0.3375
20	0.0840	0.3230
24	0.0450	0.2990
28	0.0150	0.2355
32	0.0000	0.0000



L	SNR = 7	SNR = 15
4	0.051500	0.115625
8	0.027750	0.081875
12	0.022167	0.060167
16	0.016875	0.044875
20	0.015300	0.032083
24	0.015158	0.028297
28	0.012821	0.019950
32	0.012250	0.012935

Summary

- Sparse superposition coding is reliable at rates up to channel capacity
- Analysis blends modern statistical regression and information theory