Recovery of Sparse Active Inputs in general systems

M. Malyutov

Northeastern University, Boston, MA

Tampere, January 8, 2010

Simulations are made by Dr. H. Sadaka assisted by Dr. D. Malioutov

INTRODUCTION

Around 600 papers were published last 5 years on COMPRESSIVE SENSING after
Donoho (*genius of advertising*) et al applied practical L_1 optimization (see e.g. www.dsp.ece.rice.edu/cs).

Multi-Access Information Theory based applications obtained since 1979
can greatly enhance the current understanding of sparse recovery.

Some of the alternative competitive *practical* algorithms *known since 1959*
outperform L_1 optimization in a wide range of models
are applicable to general *nonparametric noisy models* including *'noise with memory'*.

Typical Expected Applications

i. Detecting the *change-point* in users' profiles in a large computer network possibly caused by unauthorized intrusion into the system.

ii. Monitoring natural language texts, or the phone call traffic in 'hot' areas for their profiles matching those of special interest.

iii. Mine detection using routine road profile monitoring with relevant sensors.

Outline

Set up: Input a level in i-th trial, $x_i(a), i = 1, \ldots, N$, $a = 1, \ldots, t$, fixed *before* trials.
Design matrix: $X := x_i(a), i = 1, \ldots, N$, $a = 1, \ldots, t$.

Outputs z_i: noisy measurements of $g(x_i(A))$

A – *unknown* subset of *active* inputs,
to be *recovered*, $||A||_0 = s =$ sparsity.

Illustration: Active Senders $a \in A, 1 \leq a \leq t$ announce existence of packages to transmit over Multi-Access Channel with sending the column $X(a)$. The set of Active Senders is to be recovered.

Outline2

ONLY MARGINAL REMARKS ON:
Sparse Estimation under SOME conditions on $N \times t$ DESIGN matrix X:

WHAT I DO DISCUSS:

Lower and Upper Bounds and Simulated performance of three Analysis methods under
Asymptotically OPTIMAL = RANDOM X
in terms of CAPACITY $C(s)$**.**

MEP and Capacity

MEP is the Mean Error recovery Probability, if s-subsets of *Active* inputs (AIs) are equally likely among $[t] := \{1, \ldots, t\}$.
Threshold phenomenon under $N \to \infty$, $t \to \infty$,
fixed s:
If $\liminf N/N^*(s,t) > 1$, recovery has MEP **exponentially small in** N under optimal X.
If $\limsup N/N^*(s,t) < 1$, small recovery MEP impossible FOR ANY DESIGN X.

Capacity (C. Shannon) is $C(s) = \log t/N^*(s,t)$.
$N^*(s,t) = \log t/C(s)$.
$C(s)$ is found for our two analysis methods and IID unknown noise. This result extended
for stationary ergodic noise.

Elementary noiseless classical models

Find $A, ||A||_0 = s$, analyzing outputs $y_i = g(x_i(A))$, i=1,. . . , N,
1. $g(x(A)) = \cup_{a \in A} x(a)$ is a boolean sum $\cup$: $\cup$=model.
2. $g(x(A)) = \sum_{a \in A} x(a)$ is the ordinary summation (False Coin (FC)-model)
3. : $g(\cdot)$ is a non-symmetric general linear model:

$$g(x(A)) = \sum_{\substack{1 \le \lambda \le t \\ \lambda \in A}} b_\lambda x(\lambda). \tag{1}$$

Non-null $b_\lambda, \lambda \in A$ are *Significant* coefficients with *unknown* assignment. The capacity depends solely on cardinality of their linear combinations with coefficients ± 1.

Analysis methods

N_γ^T is the minimal sample size for a (T, γ)-separating design to exist, i.e. such that the MEP of T to misidentify s-tuple A is less than γ , $0 < \gamma < 1$.
In models 1,2 A are *unordered* subsets, assignment of different coefficients is *ordered* in model 3.

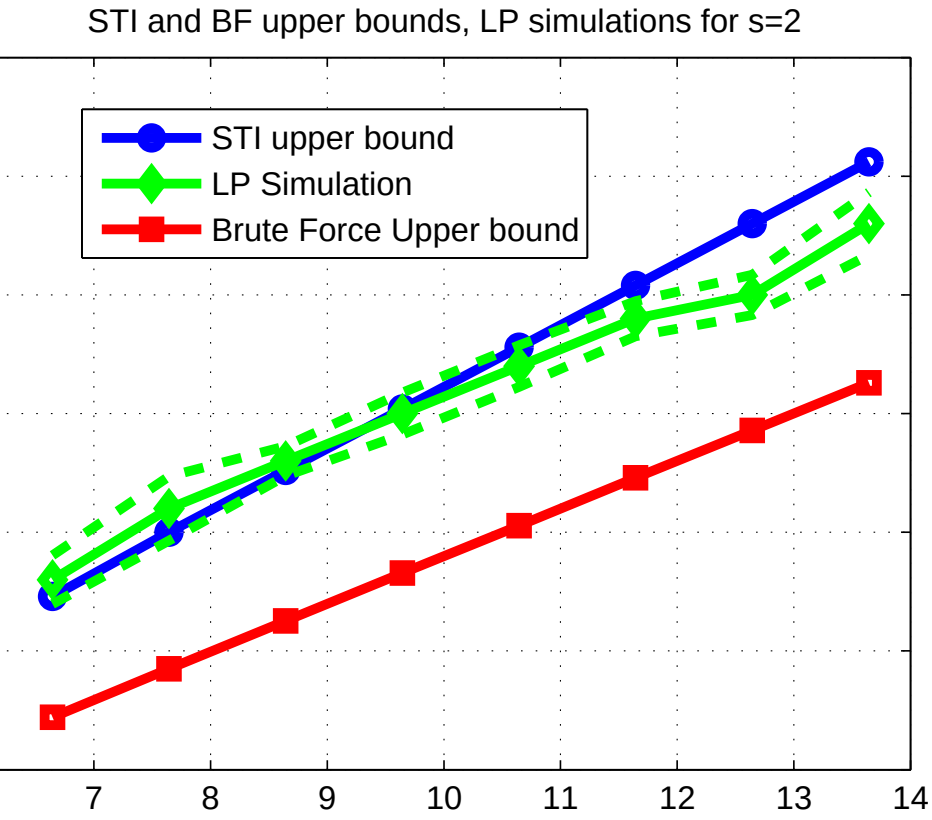
Analysis methods: 1. Brute Force (BF) or L_0-minimization,
2. Separate Testing of Inputs (STI),
3. Lasso, Linear Programming (LP), or L_1-minimization

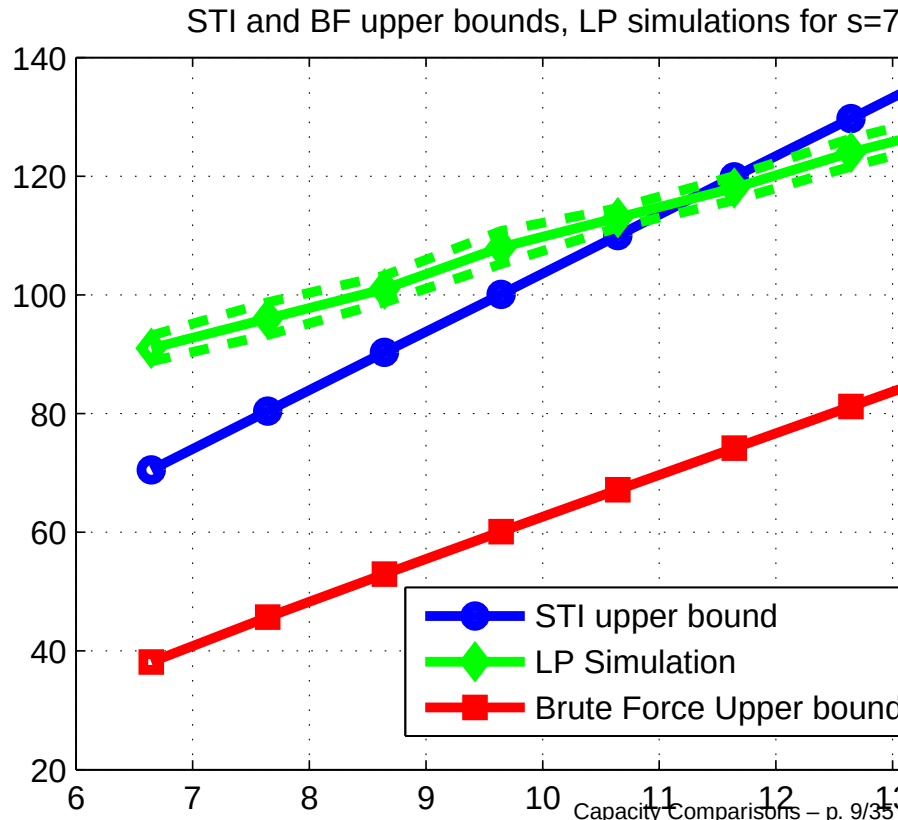
The **straightforward LP** in the $\cup$-model is compared with a more powerful LP-BF hybrid in simulations: Outputs 0 admit simpler analysis. We simulated both N^{LP} and N^{LP-BF}.

Simulation 1

Figures 1-2 summarize capacities under alternative methods of analysis for $\cup$-and FC-models with only respectively s=2, 7.

The simulated **straightforward** N^{LP} for $\cup$-model are **too big and above the frame** of figure 1.





Noisy measurements

We assign arbitrary t-tuple of binary inputs $\mathbf{x} := (x(j), j \in [t])$ and measure the **noisy output** ζ in a measurable space $\mathcal{Z}$ such that $P(\zeta|y)$ is its conditional distribution given $y = g(x(A))$ as in (1), and $P(\zeta|x(A))$ is their superposition (*Multi Access Channe*l (MAC)), where $\mathbf{x}(\mathbf{A})$ is an s-tuple of $\mathbf{x}$ of AIs.
Successive measurements $\zeta_i, i = 1, \ldots, N,$ are independent given an N-sequence of input t-tuples ($(N \times t)$-design).

To avoid algebraic technicalities, we review first *symmetric* $g(x(A))$ and finite output alphabet,
Malyutov and Mateev (1980), Malyutov and Tsitovich (2000) studied a general case.

Capacity

For a sequence $\mathcal{D}$ of N-designs, $N = 1, \ldots$, the asymptotic rate is

$$C^T(\mathcal{D}) = \lim \sup_{t \to \infty} \frac{\log t}{N^T(\gamma)} \tag{2}$$

of a test T. Here $N^T(\gamma)$ is the minimal sample size such that the MEP of T over the product of noise and prior A distribution is less than γ , $0 < \gamma < 1$.

For a general class of tests and designs, $C^T(s)$ does not depend on γ, $0 < \gamma < 1$, and remains the same for *slowly decreasing* $\gamma(t) : |\log \gamma(t)| = o(\log t)$ *as* $t \to \infty$.

Universal BF

MAC: superposition $(x(A) \to y \to \zeta)$ is the adequate model in our set up.
For **known MAC** $P(\zeta|x(A))$, the Maximum Likelihood (ML)-decision minimizes the MEP for any design. Thus C^{ML} is maximal among all tests, *if applicable*.

The **universal nonparametric** $\mathcal{BF}$-test inspired by a similar development in Csiszar and Koerner (1981) provides the maximal $C_{\mathcal{T}}$ under arbitrary finite inputs and assumptions

i. strict positivity of the cross-entropy between the output distributions corresponding to different $g(\cdot)$ and A;
ii. compactifiability of continuous output distributions (Malyutov and Tsitovich (2000)).

ESI

Test $\mathcal{BF}$ chooses as AIs the s-tuple A maximizing the Empirical Shannon Information (ESI)

$$\mathcal{I}(\tau_N^N(A)) = \tag{3}$$

$$\sum_{x(A)\in\mathcal{B}^{|A|}} \sum_{z\in\mathcal{Z}_\mathcal{N}} \tau(x(A),z)\log(\tau(z|x(A))/\tau(z)), \tag{4}$$

$\tau = \tau_N(\cdot)$ is the marginal (quantized) empirical distribution of the output.

$C^{BF}(s)$ is the maximum over randomization of similar expression with ESI replaced with the **Shannon Information**.

Nonparametric ESI

Test $\mathcal{STI}$ is defined similarly by replacing $\tau(x(A), z)$ with $\tau(x_\lambda, z)$.

Intuitive ideas in choosing these statistics
1. for s- tuple of AIs (and its subsets) these tests are strictly positive while for s-tuples of non-significant inputs $\mathcal{T}$ asymptotically vanish for large samples.

2. The large deviation probabilities for this test are unexpectedly easy to bound using Sanov's Large Deviations theorem and its conditional version.

$C^{\mathcal{STI}}$ and $C^{\mathcal{BF}}$ for symmetric models

$P_\beta^{\mathbf{m}}$ is the joint distribution of Ξ, ζ under random design with $P_\beta(x_i(\alpha) = 1) = \beta$ for all i, α, MAC $\mathbf{m}$ and $A = \{1, \ldots, s\}$.

$$I_\beta^{\mathbf{m}}(\xi(A) \wedge \zeta) = \mathbf{E} \log \frac{P_\beta^{\mathbf{m}}(\zeta|\xi([s]))}{P_\beta(\zeta)}, \tag{5}$$

where $P_\beta(\cdot)$ is the marginal distribution of ζ, expected value is over $P_\beta^{\mathbf{m}}$, and

$$C_{\mathbf{m}}^{\mathcal{STI}}(s) = \max_{\beta \in B} I_\beta^{\mathbf{m}}(X(\lambda) \wedge \zeta). \tag{6}$$

$$C_{\mathbf{m}}^{\mathcal{BF}}(s) = \max_{\beta \in B} I_\beta^{\mathbf{m}}(X(A) \wedge \zeta)/s. \tag{7}$$

STI for a MAC with Memory: in progress

Given arbitrary 'preliminary' vector–output $\mathbf{y}$, the sequence $\mathbf{z}$ distribution $\mathcal{P}_{\zeta-\mathbf{y}}$ is that of a stationary ergodic random process taking values from a finite alphabet $\mathcal{Z}$.

Our lower bounds hold with the 'entropy rate' $\lim_{N\to\infty}(I_{\beta}^{\mathbf{m}}(X_1^N(\lambda)\wedge\zeta_1^N)/N)$ instead of constant $I_{\beta}^{\mathbf{m}}(X(\lambda)\wedge\zeta)$. The $\lim$ exists due to stationarity of the pair (IID X_i,ζ_i).

We use generated product of marginal sequences with distributions ζ and ξ_1 respectively (if x_1 is inactive).

Universal Compressors as Tests

Choose one of weakly universal compressors (see e.g. Cover and Thomas (1992) or Ziv (1988) and denote

$$\mathcal{U} = \mathcal{B} \times \mathcal{Z},$$

and consider for a given $j = 1, \ldots, N$ two N-sequences with letters from $\mathcal{U}$:

$$u_j{}^N := (x_j(i), z(i)), i = 1 \ldots, N; v_j^N :$$

$$v_j^N := (x_j(i)(\times)z(i)), i = 1 \ldots, N,$$

Cont-d

taken respectively from the original and generated product-distributions (H_0) and evaluate the Ziv's (1988) goodness of fit statistic of the product distribution vs. original distribution in both sequences of the two introduced N-sequences.

The Ziv's test is based on the length comparison between compressed sequences u_j^N and v_j^N. It is shown to have the same exponentially decreasing tail of the error probability as ML under H_0 with growing N.

Thus the capacity is the same as for the I.I.D. noise with the same entropy rate of the Shannon Information between input and output per observation.

STI

The STI consists of testing significance of each input separately regarding influence of all other SI's as noise which is especially relevant for random designs:
for each input k we test the null 'randomness' hypothesis: product distribution
$Q_0(x,z) = P(\xi(k) = x)P(\zeta = z)$ versus the 'significance' alternative

$$Q_k(x,z) = P(x)P(\zeta = z|\xi(k) = x). \tag{8}$$

Due to the independence of trials, these distributions generate the product marginal distributions

$$Q_j(\mathbf{x}(k,\mathbf{z})), j = 0, k. \tag{9}$$

Cont-d

We denote the mean and variance over Q_j as $\mathbf{E}_j$ and σ_j^2 respectively, the likelihood ratios $l_k(\mathbf{x}(k,\mathbf{z}))$ and critical regions

$$\Delta_k = \{\mathbf{x},\mathbf{z} : l_k(\mathbf{x},\mathbf{z}) > \max\{\kappa_u(k), \tag{10}$$

$$\max_{m\neq k} l_m(\mathbf{x},\mathbf{z}\}\}, k = 1,\ldots,s,) \tag{11}$$

and their complement Δ_0.
The critical value is

$$\kappa_u(k) = \mathbf{E}l_k - u\sigma(l_k). \tag{12}$$

STI Decision

Let us define the STI decision as $(f(1), \ldots, f(t))$, where $f(k) = \sum k\mathbf{1}_k$, and $\mathbf{1}_k$ is the indicator of set Δ_k.
The decision is correct, if $f(\lambda_i) = i$, $i = 1, \ldots, s$, where λ_i is the i-th SI, and $f(j) = 0$ for all other j.

Definition. SI k is **hidden in noise**(HiN), if $Q_k = Q_0$ for all values p of the randomization parameter.

Malyutov and Mateev (1980) give elementary examples of HiN SI, a necessary and sufficient condition for SI to be HiN and sufficient conditions for non-existence of HiN AIs, prove that **there are no HiN AIs, if the model is symmetric.**

Theorem 1'. For symmetric model $N^{STI}(s, t, \gamma)/\log t \to 1/C^{STI}(s)$ as $t \to \infty$, where $C^{STI}(s) = \mathbf{E}l_k$ does not depend on k.

STI capacity examples

For the $\cup$ - model

$$C^{STI}(s) = \mathbf{H}(\zeta) - \mathbf{H}(\zeta|\xi) \tag{13}$$

$$= s+1)/(2s) + (2^{-1/s} - 1/2)\log(1 - 2^{-(s-1)/s}). \tag{14}$$

$C^{STI}(2) = .38$ is around $.76$ of the corresponding value for the BF-analysis. $C^{STI}(s)s \to \ln 2$. Thus $C^{STI}(s)$ is around $.7$ of the corresponding value for the BF-analysis.
2. For the FC-model

$$C^{STI}(s) = \mathbf{H}(\zeta) - \mathbf{H}(\zeta|\xi)$$

$$= H(B_s^{(1/2)}) - H(B_{(s-1)}^{1/2}). \tag{15}$$

General Linear Model

A1. cardinality of the set of all possible combinations $\{\sum_{\lambda \in A_s} b_\lambda x_i(\lambda) | x_i(\lambda) = \pm 1, \}$ is 2^s.

This maximal cardinality of the range of the outputs takes place almost sure, if coefficients b_λ are chosen randomly with whatever non-degenerate continuous distribution in R^s.

Theorem (Meshalkin (1970)) Under condition A1, the system of equations (1) determines the set A_s unambiguously with probability not less than $1 - \gamma$, if

$$N \geq \bar{N}^{BF}(s, t, \gamma) = s + \log([t - s + 1]/\gamma). \qquad (16)$$

STI-detection of SI's under A1

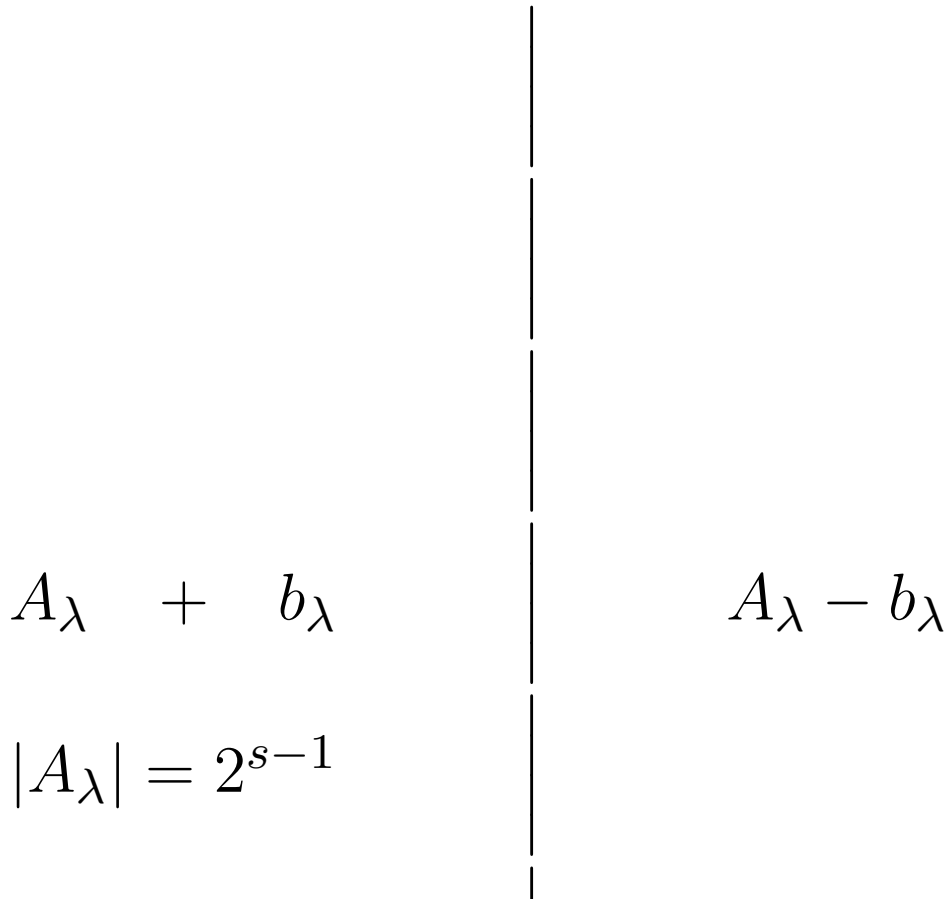
We examine all pairs $(x^N(\lambda), y^N)$,
$x^N(\lambda)$ is the binary input column and y^N is the output column with components taking 2^s values.

The *Leaders* in the ordered sequence of these test statistics are chosen as AIs. Certain sequential improvements were studied in Malyutov (1977).

Fixing the value of one of them we have 2^{s-1} equally likely combinations of the rest.

STI Illustration

In the left (right hand) side of the scatter diagrams for $x(\lambda) = \pm 1$ for each SI we have non-overlapping sets of outputs $y - b_\lambda$ $(y + b_\lambda)$, respectively: a separate partition of the outputs $\mathcal{Z}$ into two subsets $\{\pm b_\lambda + A_\lambda\}$ of size 2^{s-1}.



C^{STI}

$\mathbf{I}(x_1 \wedge Y) = \mathbf{H}(y) - \mathbf{H}(Y|x_1) = 1$. Thus $C^{STI} = C^{BF}$.
I doubt that this holds also for $N^{LP}(s,t,\gamma)$, taking into account the results of our simulation (figure 3).

For the detection of the set A in a somewhat more general model the "Random Balance Method"(RBM) (F. E. Satterthwaite, T. A. Budne, Technometrics, **1**, No 2, (1959)), was proposed

and applied successfully to numerous cases of posterior finding disorders of industrial production.

The visual inspection of scatter diagrams of data N-sequences $(X^N(\lambda), Z^N)$ for each $\lambda = 1, \ldots, t$,

Falsity of $s/t = \text{const}$ asymptotics

Under A1 any fraction of t can be restored with $N(s,t,\gamma)$ less than t for sufficiently large t.

If A1 is not valid, then the information per measurement I is strictly less than s, implying asymptotically:
$N^{BF}(s,t,\gamma) \geq s \log t/I = (s/I) \log t$.
Putting $s = kt$ for whatever $k < 1$ implies
$N^{BF}(s,t,\gamma) \geq (kt/I) \log t > t$ for sufficiently large t showing that even under the brute force analysis without noise the sparse designs are more economic, than the trivial ones **only if A1 holds**.

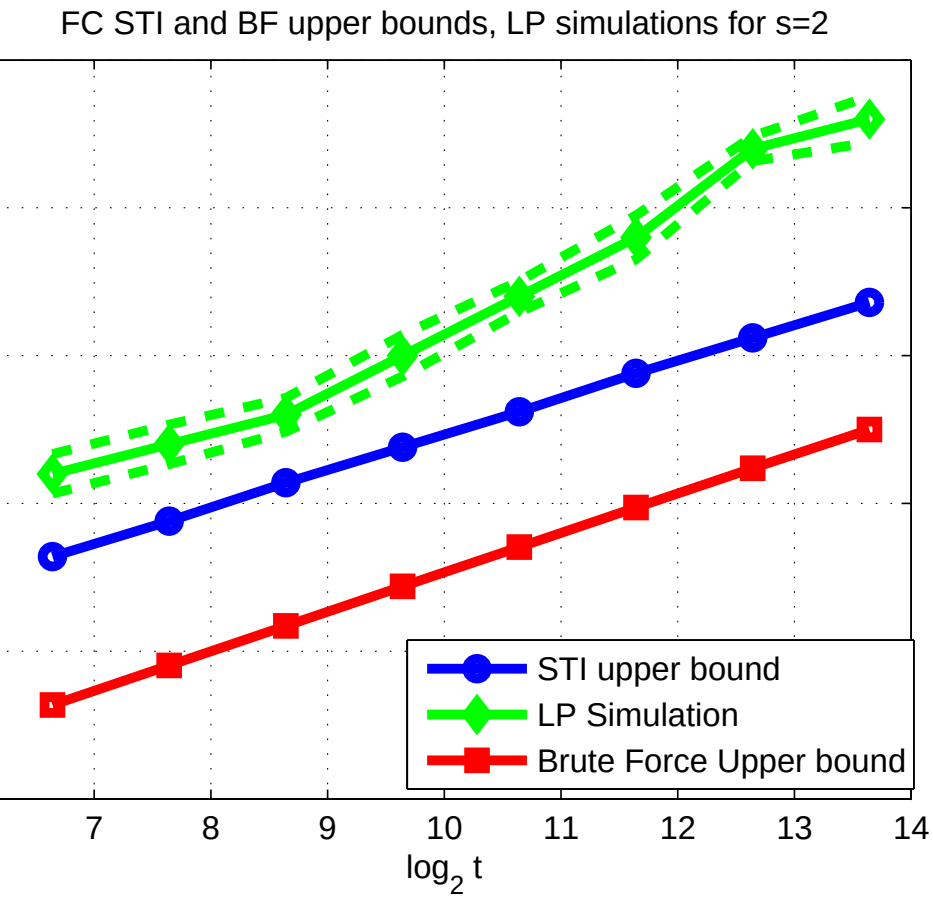
Although A1 is valid with Probability 1 under non-degenerate prior distribution of significant coefficients, it is doubtful to be taken for granted in practice.

Capacity comparisons cont-d

The left hand part of plot 3 shows simulated $N^{STI}(s,t)$ and $N^{LP}(s,t)$ performance for the linear model, $s = 2, 3, 4$ with significant coefficients 1, e, π, $\sqrt{2}$.

In the right hand side of figure 3 the Meshalkin (1970) upper bound is shown
$N^{BF}(s,t) \leq s + \log_2([t - s + 1]/.05)$
for the brute force analysis of the same model.

Figure 2


FC STI and BF upper bounds, LP simulations for s=2
STI upper bound
LP Simulation
Brute Force Upper bound
7
8
9
10
11
12
13
14
log2 t

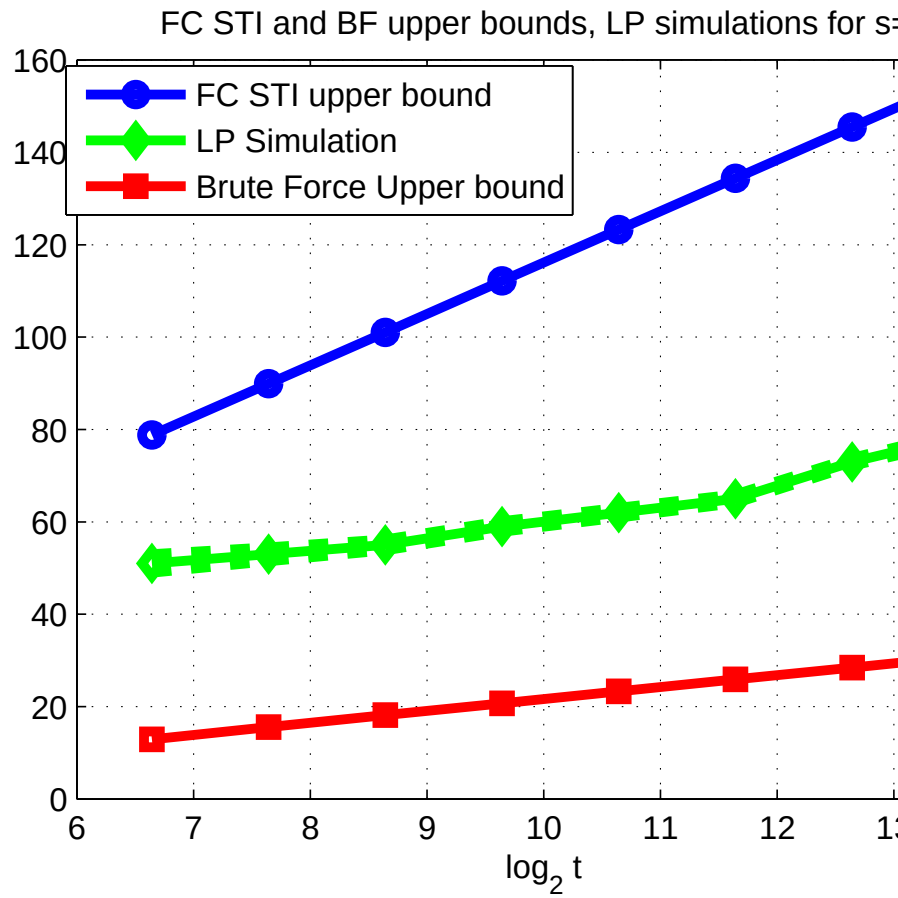
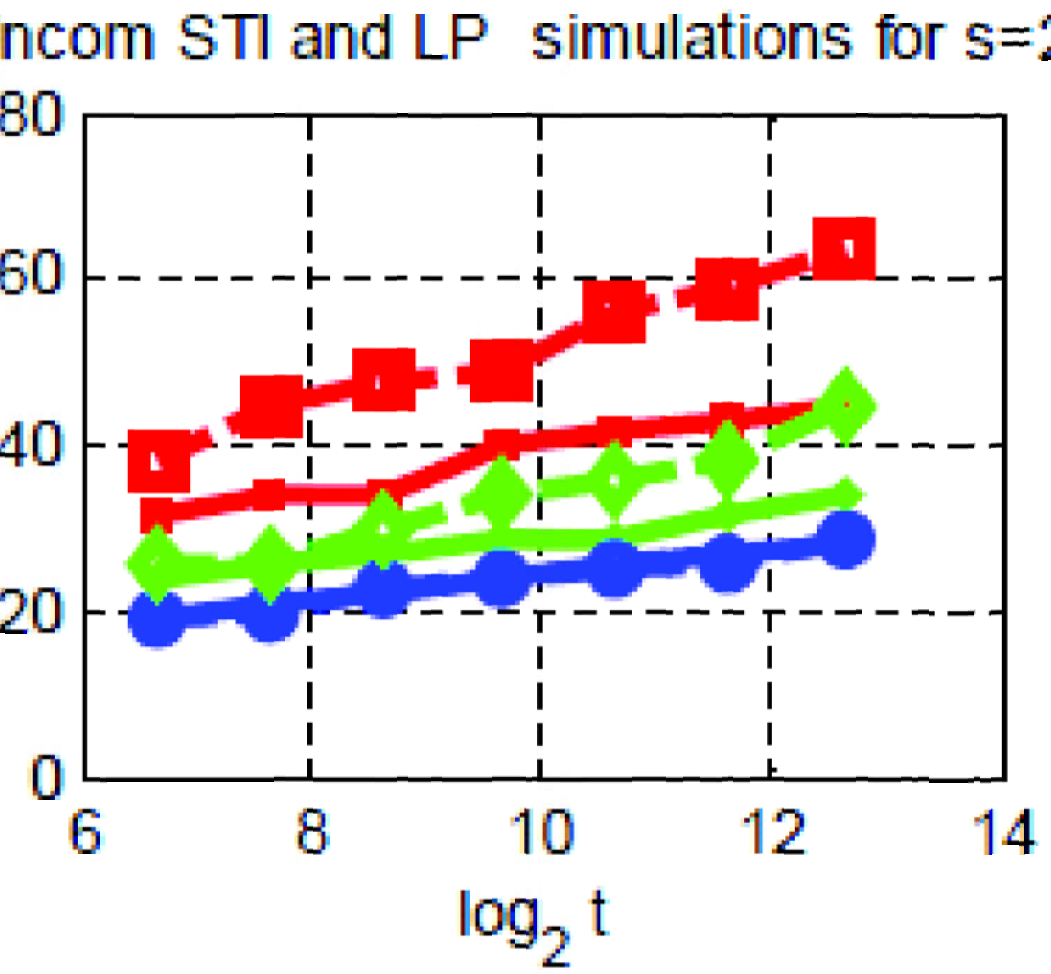
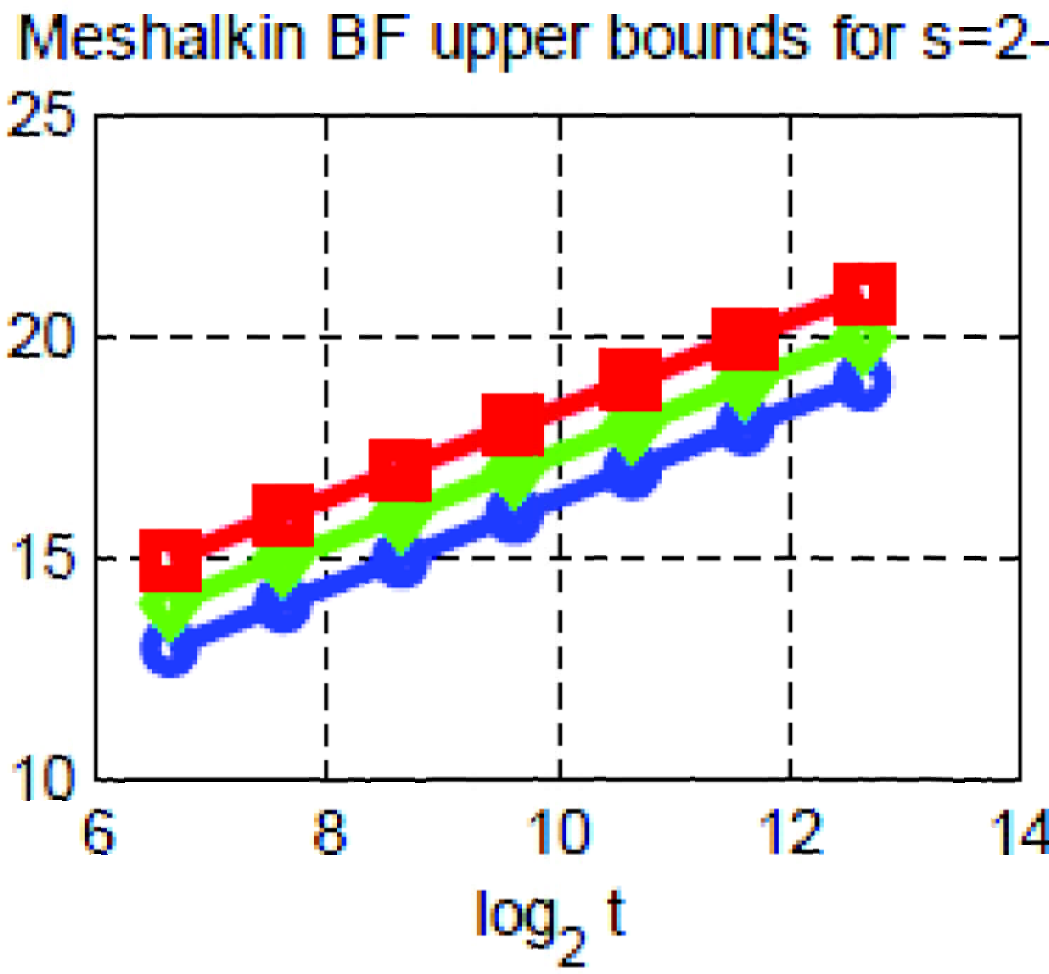

FC STI and BF upper bounds, LP simulations for s=
FC STI upper bound
LP Simulation
Brute Force Upper bound
160
140
120
100
80
60
40
20
0
6
7
8
9
10
11
12
log2 t

Figure 3


ncom STI and LP simulations for s=2-4
80
60
40
20
0
6
8
10
12
14
log2 t


Meshalkin BF upper bounds for s=2-
25
20
15
10
6
8
10
12
14
log2 t


Capacity Comparisons – p. 30/35

References I

- Ahlswede, R. (1973). Multi-way communication channels.
Proceedings of 2nd International Symposium on Information Theory, Tsahkadzor, 1971,
Akademiai Kiado, Budapest, 23-52.
- Chen, S. S., Donoho, D. L. and Saunders, M. A. (1998) Atomic de-composition by basis pursuit,
SIAM J. Scientific Computing, **20**, 33-61.
- Csiszar, I. and Körner, J. (1981).
Information Theory: Coding Theorems for Discrete Memoryless Systems, Academic Press and
Akadémiai Kiadó, Budapest.

References II

- Donoho, D.L. and Elad, M. (2003). Maximal Sparsity Representation via L_1 Minimization", *Proc. Nat. Acad. Sci.*, Vol. 100, pp. 2197-2202, March 2003.
- Donoho, D.L. and Tanner, J. (2005). Sparse nonnegative solution of underdetermined linear equations by linear programming, *Proc. National Academy of Sciences*, **102**, no. 27
- Erdös, P. and Renyi, A. (1963). On two Problems of Information Theory, *Publ. Math. Inst. of Hung. Acad.of Sc.,* **8**, 229-243.

References III

- Malyutov, M.B. and Tsitovich, I.I. (2000) Non-parametric Search for Significant Inputs of Unknown System, In *N. Callaos editor, Proceedings of SCI'2000/ISAS 2000 World Multiconference on Systemics, Cybernetics... Orlando, FL,* vol. XI, 75-83.
- Malyutov, M. B. and Pinsker, M. S. (1972). Note on the Simplest Model of the Random Balance Method. *Probabilistic Methods of Research.* Moscow University Press (ed. A. N. Kolmogorov) (In Russian).
- Ziv, J. (1988): On classification and universal data compression. *IEEE Trans. on Inform. Th.*, **34:2**, 278-286.

References IY

- Malyutov, M.B. and Sadaka H. (1998). Jaynes Principle in Testing Significant Variables of Linear Model, *Random Operators and Stochastic Equations*, **6**, 311-330.
- Malyutov, M.B. and Mateev, P.S. (1980). Screening Designs for Non-Symmetric Response Function. *Mat. Zametki,* **27**, l09-127.
- Malyutov, M.B. (1979). On the maximal rate of screening designs. *Theory Probab. and Appl.*, XXIV, 655-657.

References Y

- Malyutov, M.B. (1977). Mathematical Models and Results in Theory of Screening Experiments. In *Theoretical Problems of Experimental Design,* ed. by Malyutov M.B.„ 5-69, Soviet Radio, Moscow (In Russian).
- Malyutov, M.B. (1976). On Planning of Screening Experiments. In *Proceedings of 1975 IEEE-USSR Workshop on Inform. Theory*, N.Y. ,IEEE Inc., 1976, 144-147.
- Tibshirani, R. (1996). Regression shrinkage and selection via the LASSO. *J. Royal. Statist. Soc. B*, **58**, 267-288.