Info-Greedy Sequential Adaptive Compressed Sensing

Yao Xie
Joint work with Gabor Braun and Sebastian Pokutta

Georgia Institute of Technology

Presented at Allerton Conference 2014
Information sensing for big data

- How to extract useful information from big data for statistical inference?
- many measurements are made **sequentially**
- example: real-time estimation of state of a large network, detect changes/anomaly quickly
- we need to make minimum number of measurements
Sequential adaptive compressed sensing

linear measurement:

\[ y_i = a_i^\top x + w_i, \quad i = 1, \ldots, m. \]

\[ \beta_i = \|a_i\|_2^2 : \text{power allocated} \]

total resource constraint:

\[ \sum_{i=1}^{m} \|a_i\|_2^2 \leq P \]

- adaptiveness: measurements made sequentially, next measurement can be designed based on previous results
- compressive measurement
- how to acquire as much new information as possible?
Info-Greedy Sensing

- use conditional mutual information as metric
- incorporate distributional knowledge about $x$

Algorithm 1 Info-Greedy Sensing

Require: distributions of signal $x$ and noise $w$, error tolerance $\varepsilon$ or maximum number of iterations $M$

1: $i \leftarrow 1$
2: repeat
3: $a_i \leftarrow \text{argmax}_{a} \mathbb{I}[x; a_i^\top x + w_i | y_j, a_j, j < i]$
4: measure $y_i = a_i^\top x + w_i$
5: $i \leftarrow i + 1$
6: until $\mathbb{I}[x; a_i^\top x + w_i | y_j, a_j, j \leq i] \leq \varepsilon$ or $i > M$.

- Info-Greedy optimal
- problem $\max_{a_i} \mathbb{I}[x; a_i^\top x + w | y_j, a_j, j < i]$ non-concave in general [PalomarVerdu2006]
Related work

- mutual information one-shot compressed sensing: [CarsonChenCalderbankCarin2012]

\[
\begin{align*}
\text{maximize}_A & \quad \mathbb{I}(x; Ax + w) \\
\text{subject to} & \quad \frac{1}{m} \text{tr}(AA^\top) \leq 1
\end{align*}
\]

- direct adaptive sensing:
  distilled sensing [HauptCastroNowak2011],
  multistage support estimate [WeiHero2013]

- adaptive compressive sensing for \(k\)-sparse signal:
  compressive binary search [DavenportCastro2012],
  CASS [MalloyNowak2012]
What’s new in Info-Greedy sensing

- systematical look at sequential adaptive compressive sensing from information theory perspective

- new insights (Info-Greedy optimality, when adaptiveness help) into existing algorithms: bisection, CASS, Gaussian

- more general signal and noise model

- new algorithms: sparse measurement
Information theoretical bounds

- to obtain certain small recovery error $\|x - \hat{x}\|_2^2 \leq \epsilon$
- sequentially acquire measurements to reconstruct a signal
- equivalent to learning the $\epsilon$-ball that $x$ is constrained in

$$\|x - \hat{x}\|_2 < \epsilon$$
let $\mathcal{F}$ be a family of signals of interest

$F \in \mathcal{F}$ a random variable with uniform distribution

**Lemma: general lower bound on the number of measurements**

Suppose that for some constant $C > 0$,

$$\mathbb{H}[y_i|a_i, a_j, y_j, j \leq i, M \geq i] \leq C$$

for every round $i$. Then $\mathbb{E}[M] \geq \frac{\log |\mathcal{F}|}{C}$.

Moreover, for all $t$ we have

$$\mathbb{P}[M < t] \leq \frac{(Ct)/\mathbb{H}(F)}{\mathbb{H}(F)}$$

and

$$\mathbb{P}[M = O(\mathbb{H}(F))] = 1 - o(1).$$
\(k\)-sparse signal: windfarm example

- \(n\) wind turbines on a windfarm
- a sensor equipped on each turbine and senses \(x_i \in \{0, 1\}, i = 1, \ldots, n\)
- a central hub measures total sum of a subset of sensor readings at each time
- suppose there are \(k\) turbines broken down: \(k \ll n\)
- how do hub make measurements to quickly \textit{localize} defective turbines
Bisection algorithm

\begin{enumerate}
\item \textbf{Require:} dimension $n$ and sparsity $k$ of $x$
\item $L \leftarrow \{([n], k)\}$.
\item \textbf{while} $\exists (S, \ell) \in L : |S| > 1$ \textbf{do}
\item $L' \leftarrow \emptyset$.
\item \textbf{for all} $(S, \ell) \in L$ \textbf{do}
\item Partition $S = S_1 \cup S_2$ with $||S_1| - |S_2|| \leq 2$
\item Measure: $y = a^T x$ with $[a]_i = 1$ iff $i \in S$
\item if $0 < y < |S_1|$ then
\item $L' \leftarrow L' \cup \{(S_1, y)\}$
\item end if
\item if $\ell - |S_2| < y < \ell$ then
\item $L' \leftarrow L' \cup \{(S_2, \ell - y)\}$
\item end if
\item end for
\item $L \leftarrow L'$
\item \textbf{end while}
\item \textbf{return} $L$ as the estimated signal support
\end{enumerate}
$k$-sparse signal

... Use results from information theoretic lower bounds:

- uniform amplitude signal

**Lemma: upper bound for bisection algorithm**

Let $x \in \mathbb{R}^n$, $[x]_i \in \{0, 1\}$ be a $k$-sparse signal. Then in the absence of noise, the bisection algorithm recovers $x$ exactly with at most $k \lceil \log n \rceil$ measurements.

**Lemma: Info-Greedy optimality**

For $k = 1$ in the absence of noise the bisection algorithm is Info-Greedy optimal with respect to random support. When $k > 1$, the bisection algorithm is Info-Greedy optimal up to a factor of $\log k$.

- can also be extended to non-uniform non-negative amplitude signals: CASS-type algorithms is Info-Greedy optimal
Low-rank Gaussian signals

assume \( x \sim \mathcal{N}(0, \Sigma_x) \)

\[
\mathbb{I}[x; y_1] = \frac{1}{2} \ln \left( a_1^\top \Sigma_x a_1 \sigma^2 + 1 \right).
\]

\( a_1 = \arg \max_a \mathbb{I}[x; y_1] \) = leading eigenvector of \( \Sigma_x \)

conditioned on \( y_1 \), random vector \([x, y_2]\) is again Gaussian with updated mean and covariance matrix
Info-Greedy Sensing for low-rank Gaussian

Recursive form

Given: $\Sigma_x, \sigma^2, p$: probability of error less than $\varepsilon$
Repeat

$$\beta \leftarrow (\chi_n^2(p) / \varepsilon^2 - 1/\|\Sigma_x\|)\sigma^2$$
$$a \leftarrow \sqrt{\beta} \cdot \text{leading eigenvector of } \Sigma_x$$
$$\mu \leftarrow \mu + \Sigma_x a (a^\top \Sigma_x a + \sigma^2)^{-1} (y - a^\top \mu)$$
$$\Sigma_x \leftarrow \Sigma_x - \Sigma_x a (a^\top \Sigma_x a + \sigma^2)^{-1} a^\top \Sigma_x$$

Until $\|\Sigma_x\| \leq \varepsilon^2 / \chi_n^2(p)$
Uncertainty reduction

After measuring in the direction of a unit norm eigenvector with eigenvalue $\lambda$, using power $\beta$, conditional covariance matrix

$$
\Sigma_x - \sqrt{\beta} a \left( \sqrt{\beta} a^\top \Sigma_x \sqrt{\beta} a + \sigma^2 \right)^{-1} \sqrt{\beta} a^\top \Sigma_x
$$

$$
= \frac{\lambda \sigma^2}{\beta \lambda + \sigma^2} a^\top a + \Sigma_{x}^{\perp a},
$$

- $\Sigma_{x}^{\perp a}$: component of $\Sigma_x$ in the orthogonal complement of $a$
- after measurement: eigenvalue of $a$:

$$
\lambda \rightarrow \frac{\lambda \sigma^2}{(\beta \lambda + \sigma^2)}
$$

uncertainty in direction $a$ is reduced
Insight

$$\lambda \rightarrow \frac{\lambda \sigma^2}{(\beta \lambda + \sigma^2)}$$

Making repeated measurements is equivalent to allocating more power.

- after one measurement with $\beta = 1$:
  $$\lambda \rightarrow \frac{\lambda \sigma^2}{\lambda + \sigma^2} = \frac{\sigma^2}{1 + \sigma^2/\lambda}$$

- Suppose after $k$ measurements: $$\frac{\lambda \sigma^2}{k\lambda + \sigma^2}$$
- with $k + 1$ measurement in that direction and $\beta = 1$

$$\frac{\sigma^2}{1 + \sigma^2 \frac{k\lambda + \sigma^2}{\lambda \sigma^2}} = \frac{\lambda (\sigma^2)^2}{\lambda \sigma^2 + k\lambda \sigma^2 + (\sigma^2)^2} = \frac{\lambda \sigma^2}{(k + 1)\lambda + \sigma^2}$$
Info-Greedy Sensing for low-rank Gaussian

... based on above insights

one-shot form: power allocation

▶ choose $a_1, a_2, \ldots$ to be eigenvectors of $\Sigma_x$ in a decreasing order of eigenvalues

▶ amplitude $\beta_1, \beta_2, \ldots$ to be $(\chi^2_n(p)/\varepsilon^2 - 1/\lambda_i)\sigma^2$

one-shot form: repeated measurements

▶ choose $a_1, a_2, \ldots$ to be eigenvectors of $\Sigma_x$ in a decreasing order of eigenvalues

▶ $\beta_i = 1$

▶ $i$th measurement repeats $\lceil (\chi^2_n(p)/\varepsilon^2 - 1/\lambda_i)\sigma^2 \rceil$ times
“Sequential” and “One-shot solutions” are same for Gaussian?

- **Sequential:** \( \max_{a_i} \mathbb{I}[x; a_i^\top x + w| y_j, a_j, j < i] \)
- **One-shot:** \( \max_A \mathbb{I}[x; Ax] \)
  (solution: dominated eigenvector [CarsonChenCalderbankCarin2012])

Still, “sequential” may be beneficial

- Sparse \( \Sigma_x \in \mathbb{R}^{n\times n} \) with \( v \) non-zero entries \( \Rightarrow \) sparse power’s method, \( O(t(n + v)) \)
- when knowledge of \( \Sigma_x \) is imprecise, sequential measurement allows updating \( \Sigma_x \) from data
Theoretical performance guarantee

**Low bound on \( m \): white noise**

Info-Greedy Sensing computes a reconstruction \( \hat{x} \) with 
\[ \|x - \hat{x}\|_2^2 < \varepsilon \]
with probability at least \( p \) with at most the following number of measurements with \( \beta_i = 1 \),

\[
\sum_{\substack{i=1 \atop \lambda_i \neq 0}}^{k} \max \left\{ 0, \left[ \left( \frac{\chi_n^2(p)}{\varepsilon^2} - \frac{1}{\lambda_i} \right) \sigma^2 \right] \right\}.
\]

Or with \( k \) measurements, and total power at most

\[
\sum_{\substack{i=1 \atop \lambda_i \neq 0}}^{k} \max \left\{ 0, \left[ \left( \frac{\chi_n^2(p)}{\varepsilon^2} - \frac{1}{\lambda_i} \right) \sigma^2 \right] \right\}.
\]
Recovery of power consumption vector

- Recovery power consumption data of 58 counties in California
- Gaussian is reasonable fit: data year 2006 to year 2011
- Even knowledge learned from 5 years of limited data can make a difference!
Low-rank Gaussian mixture model signals

\[ p(x) = \sum_{c=1}^{C} \pi_c N(\mu_c, \Sigma_c) \]

- no analytic expression for conditional mutual information
- two approaches: gradient descent based approach, greedy approach
- adaptiveness makes a difference

![Graphs showing the comparison between greedy and gradient approaches for different values of \( M = 20 \) and \( \sigma = 0.01 \).]
Handwritten digit recognition

**Figure**: Comparison of true and recovered handwritten digit 2 by the greedy heuristic and the gradient descent algorithm, respectively.

<table>
<thead>
<tr>
<th>Method</th>
<th>Random</th>
<th>Greedy</th>
<th>Gradient</th>
</tr>
</thead>
<tbody>
<tr>
<td>prob. false classification</td>
<td>0.192</td>
<td>0.152</td>
<td>0.144</td>
</tr>
</tbody>
</table>
Sparse measurement vector

- add additional constraint $\|a\|_0 \leq k_0$ in the formulation
- can be solved using **outer approximation**
- mixed-integer linear program

$$\begin{align*}
\text{maximize} & \quad z \\
\text{subject to} & \quad \sum_{i=1}^{n} r_i \leq k_0 \\
& \quad a_i \leq r_i, \quad -a_i \leq r_i \\
& \quad 0 \leq z \leq c, \quad r_i \in \{0, 1\}, \ i = 1, \ldots, n \\
& \quad a \in \mathbb{R}^n, \quad z \in \mathbb{R},
\end{align*}$$
Summary

- Info-Greedy sequential adaptive compressed sensing

$$\max_{a_i} \mathbb{I}[x; a_i^\top x + w | y_j, a_j, j < i], i = 1, 2, \ldots$$

- Info-Greedy optimality: bisection and CASS-type algorithm

- Gaussian signals and insights

- Sparse measurements

- Future: towards online compressed sensing
  interplay of learning signal distribution and designing sequential measurements