Sparse Cost Function Optimization

Petros Boufounos
petrosp@merl.com

with Sohail Bahmani and Bhiksha Raj, CMU

BIRS
March 10, 2011
CS AT A GLANCE
Compressed Sensing Measurement Model

\[ M \text{ measurements} \quad y = A x \quad N \times 1 \]

- \( x \) is \( K \)-sparse or \( K \)-compressible
- \( A \) random, satisfies a restricted isometry property (RIP)
  - \( A \) has RIP of order \( 2K \) with constant \( \delta \)
  - If there exists \( \delta \) s.t. for all \( 2K \)-sparse \( x \):
    \[ (1 - \delta) \| x \|_2^2 \leq \| A x \| \leq (1 + \delta) \| x \|_2^2 \]
- \( M = O(K \log N/K) \)
- \( A \) also has small coherence \( \mu \triangleq \max_{i \neq j} | \langle a_i, a_j \rangle | \)
Compressed Sensing Measurement Model

- $x$ is $K$-sparse or $K$-compressible
- $A$ random, satisfies a restricted isometry property (RIP)
  - $A$ has RIP of order $2K$ with constant $\delta$
  - If there exists $\delta$ s.t. for all $2K$-sparse $x$:
    \[
    (1 - \delta)\|x\|_2^2 \leq \|Ax\| \leq (1 + \delta)\|x\|_2^2
    \]
- $M=O(K\log N/K)$
- $A$ also has small coherence $\mu \triangleq \max_{i \neq j} |\langle a_i, a_j \rangle|$
CS RECONSTRUCTION
CS Reconstruction

• Reconstruction using **sparse approximation**:  
  – Find sparsest \( x \) such that \( y \approx Ax \)

• **Convex optimization** approach:  
  – Minimize \( \ell_1 \) norm: e.g.,  
    \[
    \hat{x} = \arg \min_x \|x\|_1 \text{ s.t. } y \approx Ax
    \]

• **Greedy algorithms** approach:  
  – MP, OMP, ROMP, StOMP, CoSaMP, …

• If coherence \( \mu \) or RIP \( \delta \) is **small**: Exact reconstruction

**Semi-ignored question:**
How do we measure “\( \approx \)”?
Approximation Cost

- **Convex optimization** formulations

\[ \hat{x} = \arg \min_{x} \| x \|_1 + \frac{\mu}{2} \| y - Ax \|_2^2 \]

\[ \hat{x} = \arg \min_{x} \| x \|_1 \text{ s.t. } \| y - Ax \|_2^2 \leq \epsilon \]

- **Greedy pursuits** (implicit) goal

\[ \hat{x} = \arg \min_{x} \| y - Ax \|_2^2 \text{ s.t. } \| x \|_0 \leq K \]

All approaches attempt to minimize \( f(x) = \| y - Ax \|_2^2 \) such that the argument \( x \) is sparse.

Can we do it for general \( f(x) \)?
SPARSITY-CONSTRAINED FUNCTION MINIMIZATION
Problem Formulation

\[ \mathbf{x}^* = \arg \min_{\mathbf{x}} f(\mathbf{x}) \quad \text{s.t.} \quad \|\mathbf{x}\|_0 \leq K \]

- **Objective:** minimize an arbitrary cost function

- **Applications:**
  - Sparse logistic regression
  - Quantized and saturation-consistent Compressed Sensing
  - De-noising and Compressed Sensing with non-gaussian noise models

- **Questions:**
  - What algorithms can we use?
  - What functions can we minimize?
  - What are the conditions on \( f(\mathbf{x}) \)?
  - What guarantees can we provide?
Commonalities in Sparse Recovery Algorithms

• Most greedy and $l_1$ algorithms have several common steps:
  – **Maintain** a current *estimate*
  – **Compute** a *residual*
  – **Compute** a gradient, *proxy*, correlation, or some other name
  – **Update estimate** based on proxy
  – **Prune** (soft or hard threshold)
  – **Iterate**

• Key step: proxy/correlation $A^T(y-Ax)$
  – This is the *gradient* of $f(x)=||y-Ax||_2^2$
  – Can we substitute it with the general gradient $\nabla f(x)$?

**YES**

What **guarantees** can we prove?
What becomes of the **RIP**?
GraSP (Gradient Subspace Pursuit)

State Variables: Signal estimate, \( \hat{x} \) support estimate: \( T \)

- Initialize estimate and support:
  \[
  \hat{x} = 0, \quad T = \text{supp}(\hat{x})
  \]

- Compute Gradient at Current Estimate:
  \[
  \nabla f (\hat{X})
  \]

- Select location of largest \( 2K \) gradient directions:
  \[
  \text{supp}(g|_{2K})
  \]

- Add to support set:
  \[
  \Omega = \text{supp}(g|_{2K}) \cup T
  \]

- Minimize over support:
  \[
  b = \arg \min_x f(x) \quad \text{s.t. } x_{\Omega^c} = 0
  \]

- Truncate result:
  \[
  \hat{x} = b|_K
  \]
  \[
  T = \text{supp} (b|_K)
  \]

Iterate using residual.
\[ f(x) = \|y - Ax\|_2^2 \Rightarrow \textbf{CoSaMP (Compressive Sampling MP)} \] [Needell and Tropp]

**State Variables:** Signal estimate, \( \hat{x} \) support estimate: \( T \)

- Initialize estimate, residual and support
  \[ \hat{x} = 0, \ T = \text{supp}(\hat{x}), \ r = y \]

- Correlate residual with dictionary → signal proxy
  \[ \langle a_k, r \rangle = p_k \]

- Select location of largest \( 2K \) correlations
  \[ \Omega = \text{supp}(p|_{2K}) \cup T \]

- Add to support set
  \[ \Omega = \text{supp}(p|_{2K}) \cup T \]

- Invert over support
  \[ b = A_{\Omega}^{\dagger} y \]

- Truncate and compute residual
  \[ T = \text{supp}(b|_{K}) \]
  \[ \hat{x} = b|_{K} \]
  \[ r \leftarrow y - A\hat{x} \]

Iterate using residual
CONDITIONS AND GUARANTEES
Stable Hessian Property

- Guarantees based on the Hessian of the function $H_f(x)$
- Some definitions:
  
  for all $\|u\|_0 \leq K$
  
  $A_K (u) = \sup \left\{ \frac{v^T H_f(u)v}{\|v\|^2_2} \mid \text{supp}(v) = \text{supp}(u), \text{and } v \neq 0 \right\}$
  
  $B_K (u) = \inf \left\{ \frac{v^T H_f(u)v}{\|v\|^2_2} \mid \text{supp}(v) = \text{supp}(u), \text{and } v \neq 0 \right\}$

- Stable Hessian Property (SHP) of order $K$, with constant $\mu_K$:

  $\frac{A_K (u)}{B_K (u)} \leq \mu_K$, for all $\|u\|_0 \leq K$

- Bounds the local curvature of $f(x)$
Recovery Guarantees

- Denote the **global optimum** using $x^*$:
  \[ x^* = \arg \min_x f(x) \text{ s.t. } \|x\|_0 \leq K \]

- Assume $f(x)$ satisfies and order $4K$ SHP with:
  \[ \text{for all } \|u\|_0 \leq 4K, \quad \frac{A_{4K}(u)}{B_{4K}(u)} \leq \mu_{4K} \leq \sqrt{2} \]

- And its restriction is **convex**:
  \[ \text{for all } \|u\|_0 \leq 4K, \quad B_{4K} > \epsilon \]

- Then the **estimate** after the $p$th iteration, $\hat{x}^{(p)}$, satisfies:
  \[ \left\| \hat{x}^{(p)} - x^* \right\|_2 \leq 2^{-p} \|x^*\|_2 + \frac{4(2 + \sqrt{2})}{\epsilon} \|\nabla f(x^*)|_I\|_2 \]

  where $I$ is the set of the largest $3K$ components of $\nabla f(x^*)$ in magnitude.
Connections to CS

- CS uses $f(x) = \|y - Ax\|_2$

- **SHP** bounds $A_K(u)$, $B_K(u)$, reduce to **RIP** bounds $(1 \pm \delta_K)$

- $\mu_K$ reduces to $(1+\delta_K)/(1-\delta_K)$

- **GraSP** reduces to **CoSaMP**

- Reconstruction guarantees reduce to classical CS guarantees
APPLICATIONS
Given: **Bit budget** $B$ bits/sample, **Signal norm** $\|x\|_2$

- **Set quantization threshold** $G$
  - Implicitly sets quantization interval $\Delta = 2^{B+1}G$
  - Implicitly sets saturation rate at $2Q(G/\|x\|_2)$

- **Classical heuristic**: Set $G$ large (avoid saturation)

  - **Wrong! Will revisit!**

- **Note**:
  - Equivalent to fixing $G$ and varying signal amplification
  - $Q(\cdot)$ denotes the tail of the Gaussian distribution
Exploit Saturation Information

\[
\hat{x} = \arg\min_x \|y - \tilde{A}x\|_2 + \|G - A^+ x\|_2^+ + \|G + A^- x\|_2^+ + \|x\|_0 \leq K
\]

Saturation provides information:
The measurement magnitude is larger than \(G\). But how to handle it?

**Option 1:** Just use the measurement as if unsaturated

**Option 2:** Discard saturated measurements

**Option 3:** Treat measurement as a constraint! (consistent reconstruction)
Experimental Results

Note: optimal performance requires 10% saturation
Reconstruction Results: Real Data [Wei, Boufounos]

Synthetic Aperture Radar (SAR) acquisition

(a) CSA unsaturated
(b) CSA 30% sat.
(c) Robust 30% sat.

Loss of fine features
Significant intensity loss due to saturation
Intensity loss restored Crisper image
Reconstruction Results: Real Data, log scale

Synthetic Aperture Radar (SAR) acquisition

(a) CSA unsaturated  

(b) CSA 30% sat.  

(c) Robust 30% sat.

Significant Reconstruction Noise  

Image model (wavelet sparsity) performs denoising
Sparse Logistic Regression

- Examples in data points $d_i$, each has a label $l_i$ (±1)

- Need to find coefficients $x_i$ that predict labels from data
  - Prediction through the logistic function
  - Feature selection: find a sparse set $x$

- Resulting problem is a sparse minimization:
  $$f(x) = \sum_{i=1}^{N} \log \left(1 + \exp \left( -l_i x^T d^i \right) \right)$$

- We can use GraSP!

- **Alternative:** $\ell_1$ regularization (e.g., IRLS-LARS, [Lee et al, 2006]):
  $$\hat{x} = \arg \min_x f(x) + \lambda \|x\|_1$$
Simulation Results Classification Accuracy

- Data: UCI Adult Data Set
  - Goal: Predict household income $\leq 50K$ from 14 variables, 123 features

- Note: Prediction accuracy ≠ optimization performance
  - We actually also achieve a smaller sparse minimum.
Open Problems

• Several questions:
  – What is the appropriate $\ell_1$ formulation?
  – What about other greedy algorithms? (e.g., OMP, IHT)
  – Can the **Stable Hessian Property** help with those?
  – What does the **SHP** really mean for $f(x)$? What about its convexity?
  – How to interpret the guarantees?
  – What other conditions can we use instead?
    • Related work, different context, by Blumensath, SCP
  – Can we derive equivalents of coherence or NSP?
  – Can we accommodate functions that are not twice differentiable?

**Questions/Comments?**

More info: petrosb@merl.com