MITSUBISHI ELECTRIC RESEARCH LABORATORIES
Cambridge, Massachusetts

Sparse Cost Function Optimization

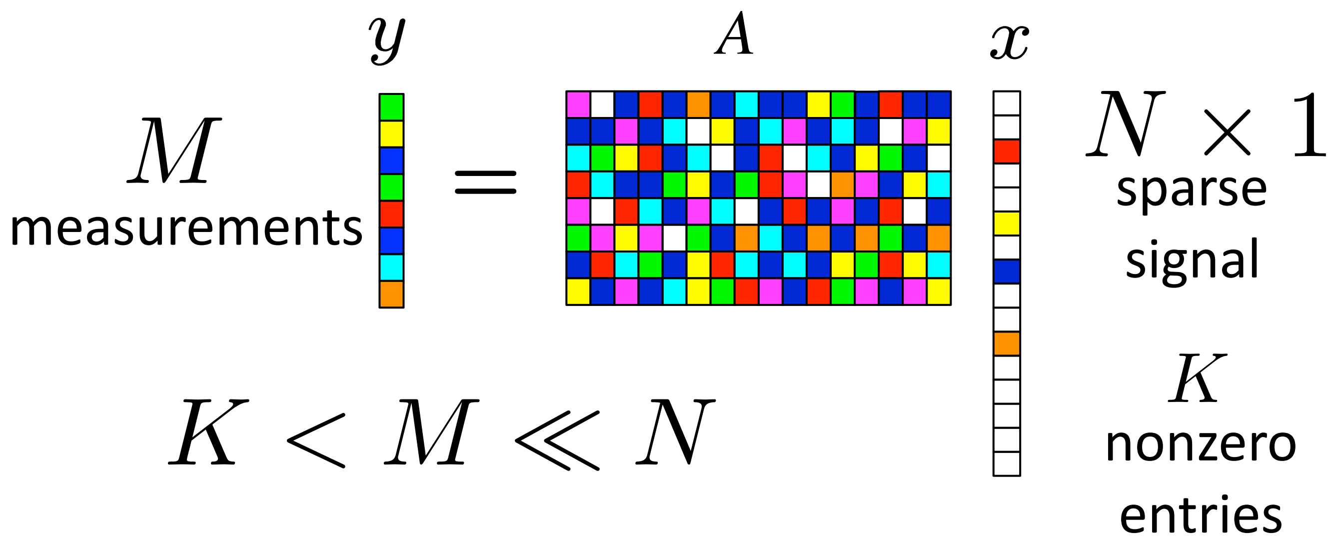
Petros Boufounos
petrosb@merl.com

with Sohail Bahmani and Bhiksha Raj, CMU

BIRS
March 10, 2011

CS AT A GLANCE

Compressed Sensing Measurement Model



$$K < M \ll N$$

- x is K-sparse or K-compressible
- A **random**, satisfies a *restricted isometry property (RIP)*

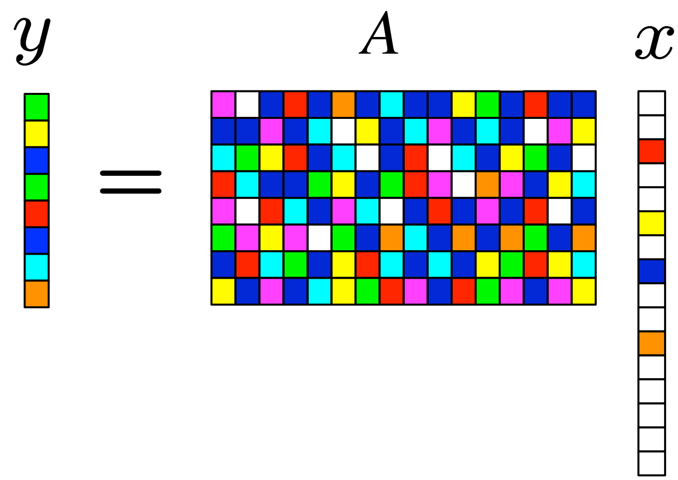
  A has RIP of order $2K$ with constant δ

  If there exists δ s.t. for all $2K$-sparse x:

  $$(1-\delta)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\| \leq (1+\delta)\|\mathbf{x}\|_2^2$$

- $M = O(K \log N/K)$
- A also has small *coherence* $\mu \triangleq \max_{i \neq j} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle|$

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 δ

  If there exists δ s.t. for all $2K$-sparse x:

  $$(1-\delta)\|\mathbf{x}\|_2^2 \leq \|\mathbf{A}\mathbf{x}\| \leq (1+\delta)\|\mathbf{x}\|_2^2$$

- $M = O(K \log N/K)$
- A also has small *coherence* $\mu \triangleq \max_{i \neq j} |\langle \mathbf{a}_i, \mathbf{a}_j \rangle|$

CS RECONSTRUCTION

CS Reconstruction

- Reconstruction using **sparse approximation**:
  - Find sparsest $\mathbf{x}$ such that $\mathbf{y} \approx \mathbf{A}\mathbf{x}$
- **Convex optimization** approach:
  - Minimize ℓ_1 norm: e.g.,

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \text{ s.t. } \mathbf{y} \approx \mathbf{A}\mathbf{x}$$

- **Greedy algorithms** approach:
  - MP, OMP, ROMP, StOMP, CoSaMP, …
- If coherence μ or RIP δ is ***small***: Exact reconstruction

Semi-ignored question:
How do we measure "≈"?

Approximation Cost

- **Convex optimization** formulations

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 + \frac{\mu}{2}\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$$

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \text{ s.t. } \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \leq \epsilon$$

- **Greedy pursuits** (implicit) goal

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \text{ s.t. } \|\mathbf{x}\|_0 \leq K$$

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

Can we do it for general $f(\mathbf{x})$?

MITSUBISHI ELECTRIC
Changes for the Better

SPARSITY-CONSTRAINED FUNCTION MINIMIZATION

Problem Formulation

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

- **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 $\mathbf{A}^T(\mathbf{y}-\mathbf{A}\mathbf{x})$
  - This is the **gradient** of $f(\mathbf{x})=\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2^2$
  - Can we substitute it with the general gradient $\nabla f(\mathbf{x})$?

YES

What **guarantees** can we prove?

What becomes of the **RIP**?

GraSP (Gradient Subspace Pursuit)

State Variables: Signal estimate, $\hat{\mathbf{x}}$ support estimate: T

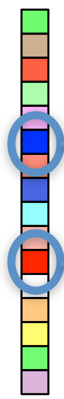
Initialize estimate and support

$$\hat{\mathbf{x}}=0,\; T=\mathrm{supp}(\hat{\mathbf{x}})$$

Compute Gradient at Current Estimate

$$\nabla f(\widehat{\mathbf{x}}) = $$

$\mathbf{g}$



Select location of largest *2K* gradient directions

$$\mathrm{supp}(\mathbf{g}|_{2K})$$

Add to support set

$$\Omega = \mathrm{supp}(\mathbf{g}|_{2K}) \cup T$$

Minimize over support

$$\mathbf{b} = \arg\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{s.t. } \mathbf{x}_{\Omega^c} = 0$$

Truncate result

$$\widehat{\mathbf{x}} = \mathbf{b}|_K$$

$$T = \mathrm{supp}\,(\mathbf{b}|_K)$$

Iterate using residual

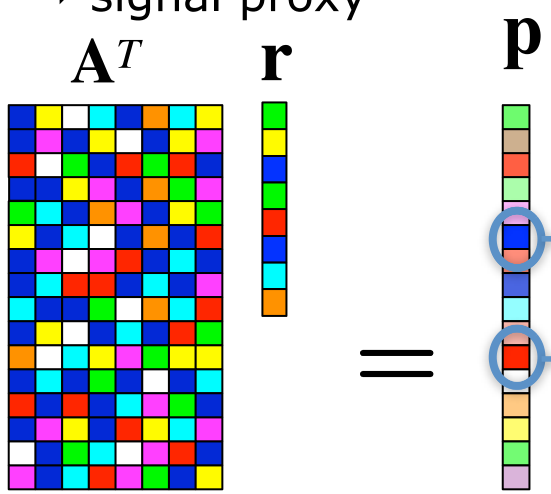
$f(x)=\|y-Ax\|_2^2 \Rightarrow$ CoSaMP (Compressive Sampling MP) [Needell and Tropp]

State Variables: Signal estimate, $\hat{\mathbf{x}}$ support estimate: T

Initialize estimate, residual and support

$$\hat{\mathbf{x}}=0,\ T=\mathrm{supp}(\hat{\mathbf{x}}),\ \mathbf{r}=\mathbf{y}$$

Correlate residual
with dictionary
→ signal proxy



$$\langle a_k, r\rangle = p_k$$

Select location
of largest
2K correlations

$$\mathrm{supp}(\mathbf{p}|_{2K})$$

Add to
support set

$$\Omega = \mathrm{supp}(p|_{2K}) \cup T$$

Invert over
support

$$b = A_\Omega^\dagger y$$

Truncate and
compute residual

$$T = \mathrm{supp}(b|_K)$$

$$\hat{x} = b|K$$

$$r \leftarrow y - A\hat{x}$$

Iterate using residual

CONDITIONS AND GUARANTEES

Stable Hessian Property

- Guarantees based on the Hessian of the function $\mathbf{H}_f(\mathbf{x})$
- Some definitions:

for all $\|\mathbf{u}\|_0 \leq K$

$$A_K(\mathbf{u}) = \sup\left\{\frac{\mathbf{v}^{\mathrm{T}}\mathbf{H}_f(\mathbf{u})\mathbf{v}}{\|\mathbf{v}\|_2^2}\,\middle|\,\mathrm{supp}(\mathbf{v}) = \mathrm{supp}(\mathbf{u}),\ \text{and } \mathbf{v} \neq 0\right\}$$

$$B_K(\mathbf{u}) = \inf\left\{\frac{\mathbf{v}^{\mathrm{T}}\mathbf{H}_f(\mathbf{u})\mathbf{v}}{\|\mathbf{v}\|_2^2}\,\middle|\,\mathrm{supp}(\mathbf{v}) = \mathrm{supp}(\mathbf{u}),\ \text{and } \mathbf{v} \neq 0\right\}$$

- Stable Hessian Property (SHP) of order K, with constant μ_K:

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

- Bounds the local curvature of $f(\mathbf{x})$

Recovery Guarantees

- Denote the **global optimum** using $\mathbf{x}^*$:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} f(\mathbf{x}) \ \text{ s.t. } \ \|\mathbf{x}\|_0 \le K$$

- Assume $f(\mathbf{x})$ satisfies and order $4K$ **SHP** with:

$$\text{for all } \|\mathbf{u}\|_0 \le 4K, \ \frac{A_{4K}(\mathbf{u})}{B_{4K}(\mathbf{u})} \le \mu_{4K} \le \sqrt{2}$$

- And its **restriction** is **convex**:

$$\text{for all } \|\mathbf{u}\|_0 \le 4K, \ B_{4K} > \epsilon$$

- Then the **estimate** after the p^{th} iteration, $\widehat{\mathbf{x}}^{(p)}$, **satisfies**:

$$\left\|\widehat{\mathbf{x}}^{(p)} - \mathbf{x}^\star\right\|_2 \le 2^{-p} \|\mathbf{x}^\star\|_2 + \frac{4\left(2+\sqrt{2}\right)}{\epsilon} \|\nabla f(\mathbf{x}^\star)|_{\mathcal{I}}\|_2$$

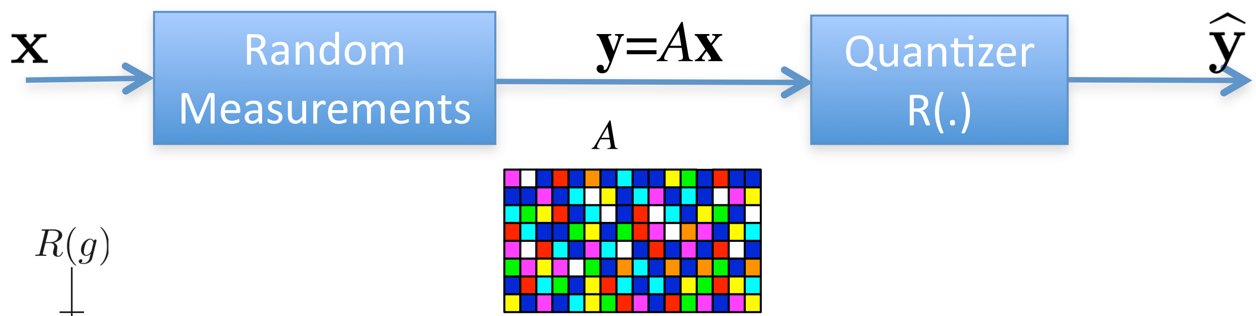
where $\mathcal{I}$ is the set of the largest $3K$ components of $\nabla f(\mathbf{x}^*)$ in magnitude

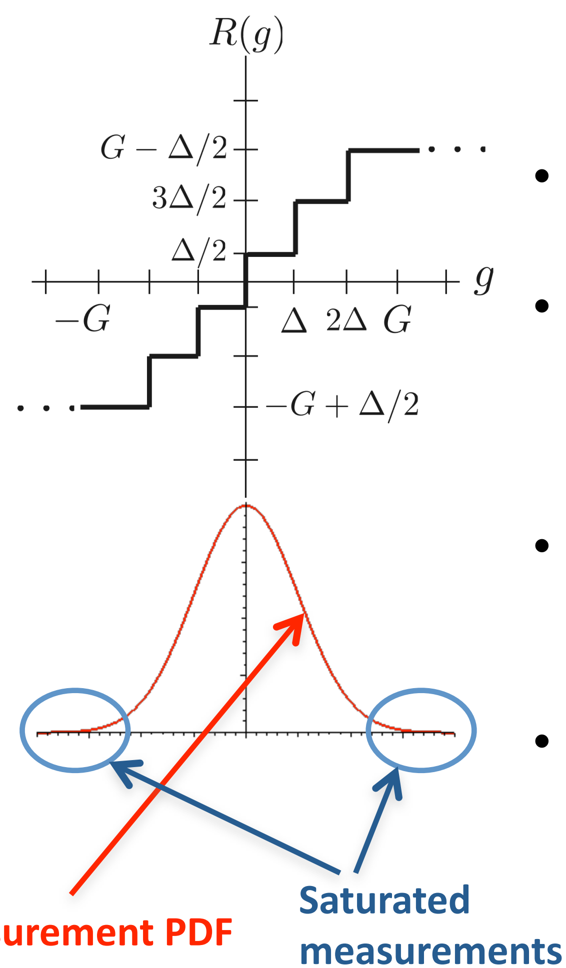
Connections to CS

- CS uses $f(\mathbf{x})=\|\mathbf{y}-\mathbf{A}\mathbf{x}\|_2^2$
- **SHP** bounds $A_K(\mathbf{u})$, $B_K(\mathbf{u})$, reduce to **RIP** bounds $(1\pm\delta_K)$
- μ_K reduces to $(1+\delta_K)/(1-\delta_K)$
- **GraSP** reduces to **CoSaMP**
- Reconstruction guarantees reduce to classical CS guarantees

APPLICATIONS

CS and Saturation [Laska, Boufounos, Davenport, Baraniuk]



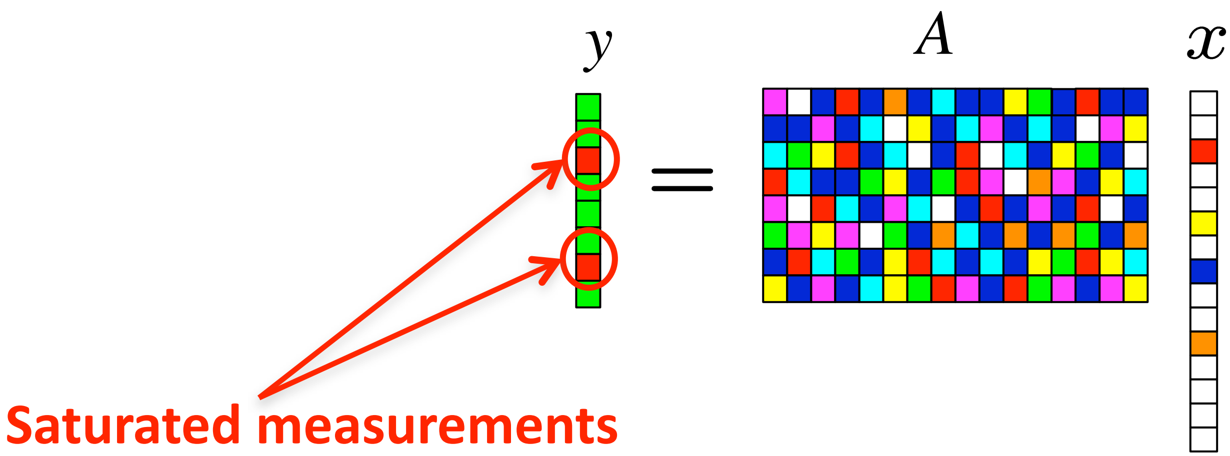


- Given: **Bit budget** B bits/sample, **Signal norm** $\|\mathbf{x}\|_2$
- Set quantization **threshold** G
  - Implicitly sets quantization interval $\Delta = 2^{-B+1}G$
  - Implicitly sets saturation rate at $2Q(G/\|x\|_2)$
- ~~Classical heuristic: Set G large (avoid saturation)~~

Wrong! Will revisit!

- Note:
  - equivalent to fixing G and varying signal amplification
  - $Q(\cdot)$ denotes the tail of the Gaussian distribution

Exploit Saturation Information



Saturation provides information:

The measurement magnitude is **larger** than G. But **how to handle it**?

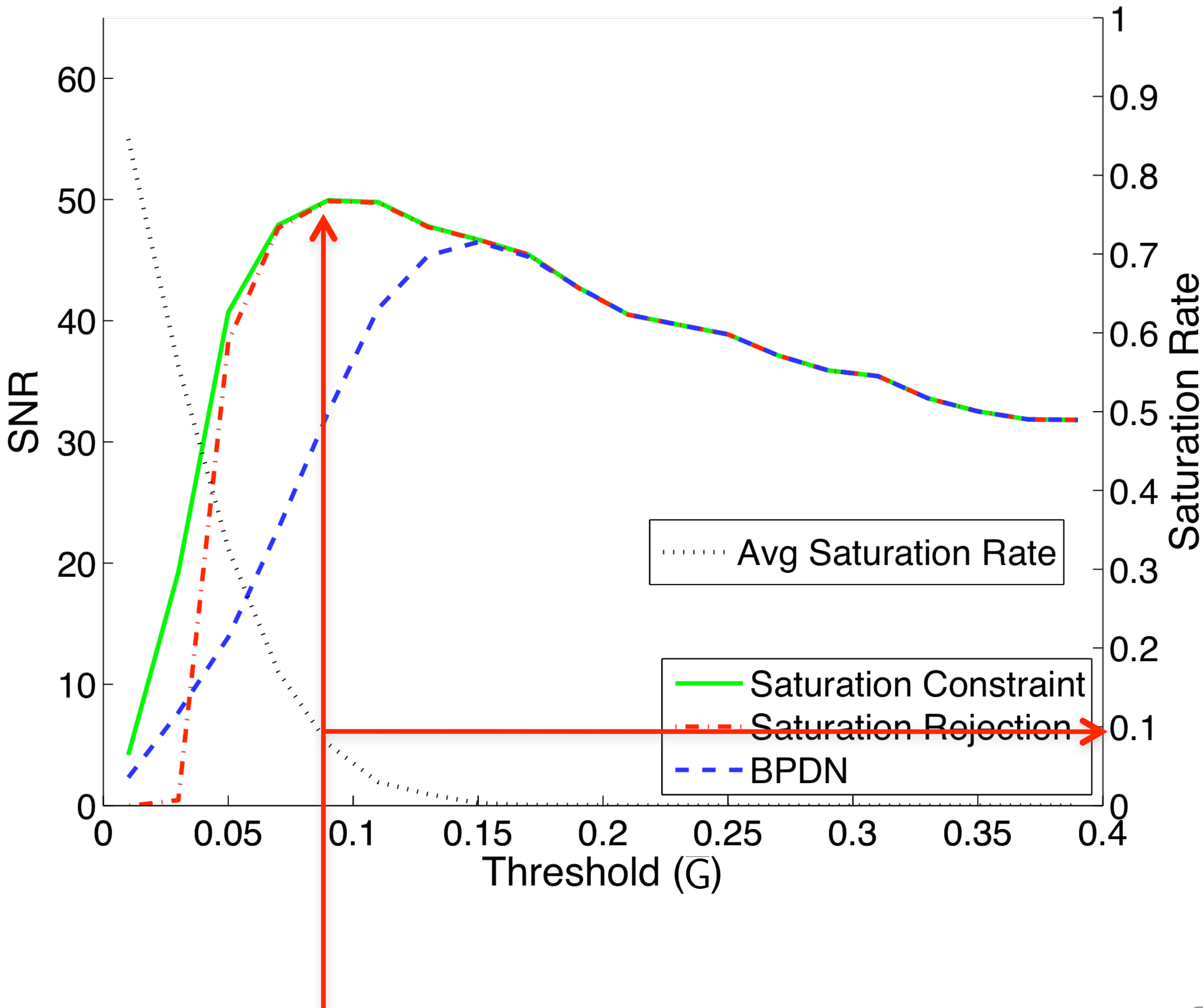
Option 1: Just **use** the measurement **as if unsaturated**

Option 2: **Discard** saturated measurements

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

$$
\begin{aligned}
\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} \quad & \|\mathbf{y} - \widetilde{\mathbf{A}}\mathbf{x}\|_2 + && \text{Unsaturated} \\
& \left\| \left(G - \mathbf{A}^+\mathbf{x}\right)_+ \right\|_2 + && \text{Positive Saturation} \\
& \left\| \left(G + \mathbf{A}^-\mathbf{x}\right)_+ \right\|_2 && \text{Negative Saturation} \\
\text{s.t.} \quad & \|\mathbf{x}\|_0 \le K
\end{aligned}
$$

Experimental Results



Note: optimal performance **requires** 10% saturation

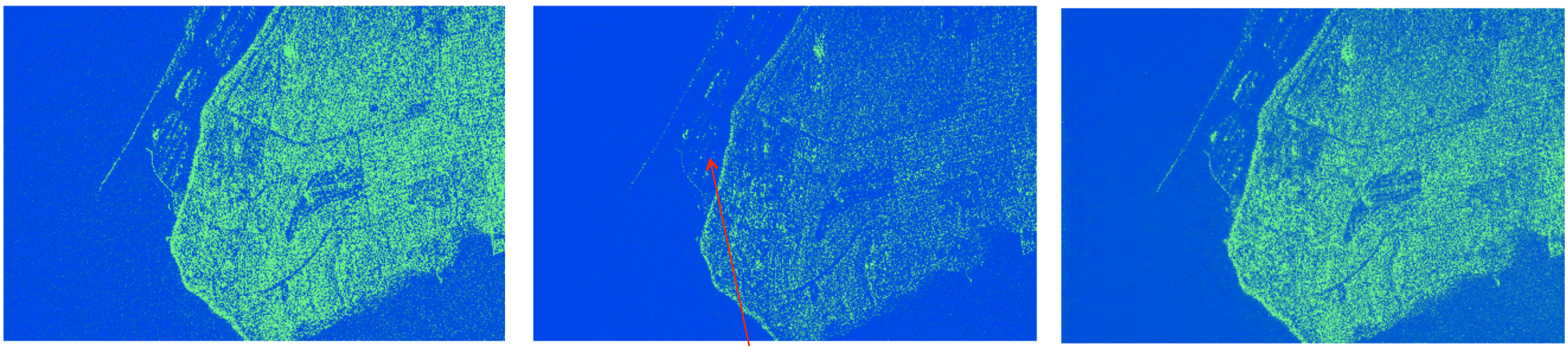
Reconstruction Results: Real Data [Wei, Boufounos]

Synthetic Aperture Radar (SAR) acquisition

(a) CSA unsaturated

(b) CSA 30% sat.

(c) Robust 30% sat.



Loss of fine features

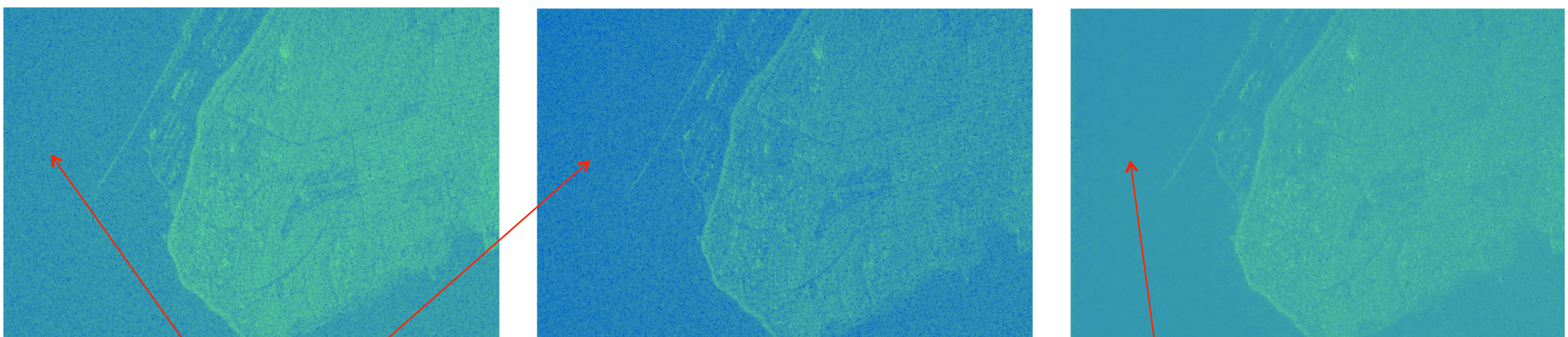
Significant intensity loss
due to saturation

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 though the logistic function
  - Feature selection: find a sparse set $\mathbf{x}$
- Resulting problem is a sparse minimization:

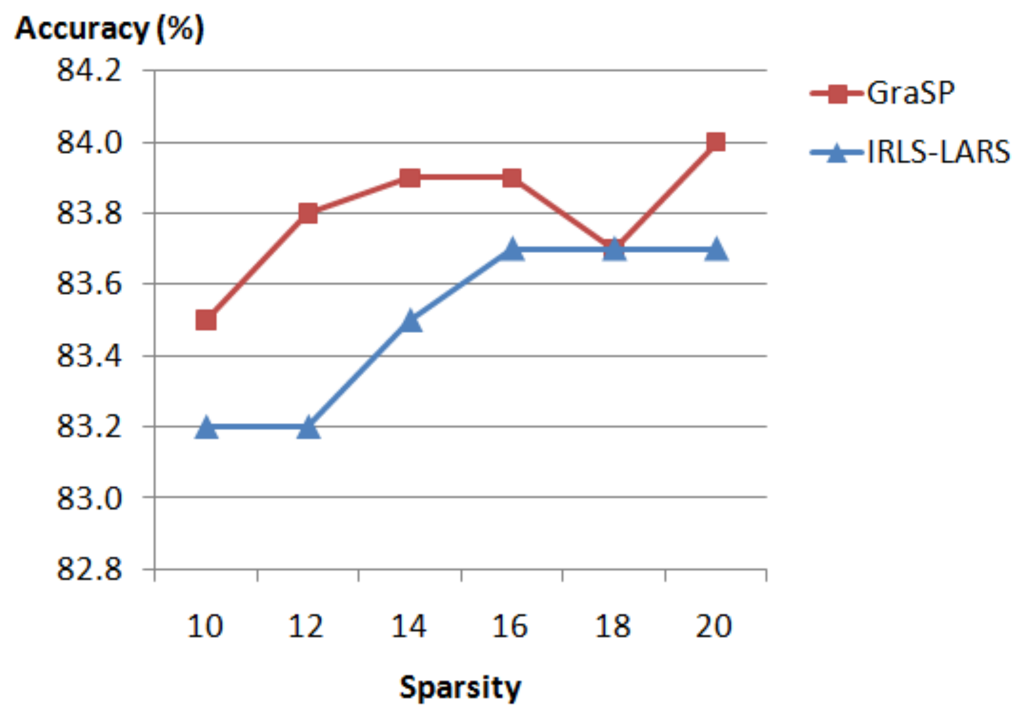
$$f(\mathbf{x}) = \sum_{i=1}^{N} \log\left(1 + \exp\left(-l^i \mathbf{x}^{\mathrm{T}} \mathbf{d}^i\right)\right)$$

- We can use **GraSP**!
- **Alternative:** ℓ_1 regularization (e.g., IRLS-LARS, [Lee et al, 2006]):

$$\widehat{\mathbf{x}} = \arg\min_{\mathbf{x}} f(\mathbf{x}) + \lambda \|\mathbf{x}\|_1$$

Simulation Results Classification Accuracy

- Data: UCI Adult Data Set
  - Goal: Predict household income ≶\$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 ℓ_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 $f(\mathbf{x})$? 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*