# Greedy Sparsity-Constrained Optimization

Sohail Bahmani

with Petros Boufounos and Bhiksha Raj

# Outline

- Background
  - Compressed Sensing
- Problem Formulation
  - Generalizing Compressed Sensing
  - Example
  - Prior Work
- GraSP Algorithm
  - Main Result
  - Required Conditions
  - Example: $\ell_2$-regularized Logistic Regression

# Compressed Sensing (1)

| Linear Inverse Problem | |
|---|---|
| **Sparse** signal | $\mathbf{x}^\star \in \mathbb{R}^p$ |
| Measurement matrix | $\mathbf{A} \in \mathbb{R}^{n \times p}$ |
| Measurement | $\mathbf{y} = \mathbf{A}\mathbf{x}^\star + \mathbf{e}$ |
| Noise | $\mathbf{e} \in \mathbb{R}^n$ |
| Given $\mathbf{y}$ and $\mathbf{A}$ with $n \ll p$, estimate $\mathbf{x}^\star$ | |

**Applications:**

- Biomedical Imaging, Image Denoising, Image Segmentation, Filter Design, System Identification, etc

# Compressed Sensing (2)

$$\|\mathbf{x}\|_0 = |\mathrm{supp}(\mathbf{x})| = \sum_{i=1}^{p} \mathbb{I}(x_i \neq 0)$$

$$\|\mathbf{x}\|_1 = \sum_{i=1}^{p} |x_i|$$

**$\ell_0$-minimization**

$$\arg\min_{\mathbf{x}} \quad \|\mathbf{x}\|_0 \qquad \text{(L0)}$$

$$\text{subject to} \quad \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \leq \epsilon$$

**$\ell_1$-minimization**

$$\arg\min_{\mathbf{x}} \quad \|\mathbf{x}\|_1 \qquad \text{(L1)}$$

$$\text{subject to} \quad \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2 \leq \epsilon$$

NP-hard

(1) Convexify

**$\ell_0$-constrained LS**

$$\arg\min_{\mathbf{x}} \quad \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 \qquad \text{(C0)}$$

$$\text{subject to} \quad \|\mathbf{x}\|_0 \leq s$$

**$\ell_1$-constrained LS**

$$\arg\min_{\mathbf{x}} \quad \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 \qquad \text{(C1)}$$

$$\text{subject to} \quad \|\mathbf{x}\|_1 \leq R$$

Use $\ell_1$-norm as a proxy for $\ell_0$-pseudonorm

(2)

(Greedy) Approximate Solvers

# Generalizing Compressed Sensing

## Common assumptions in CS

consider nonlinear relations

- The relation between the input and response has a **linear** form: $\boldsymbol{y} = \mathbf{A}\mathbf{x} + \mathbf{e}$
- The error is usually measured in **squared error**: $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$

other measures of fidelity

### General Formulation

Let $f: \mathbb{R}^p \to \mathbb{R}$ be a cost function. Approximate the solution to

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} f(\mathbf{x})$$
$$\text{subject to } \|\mathbf{x}\|_0 \leq s.$$

- For $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$ we get the $\ell_0$-constrained least squares formulation in CS
- We will see **$\ell_2$-regularized logistic loss** as another example for $f(\mathbf{x})$
- More generally, $f(\mathbf{x})$ can be the **empirical loss** associated with some observations in statistical estimation problems

# Example

## Gene selection problem

- **Data points $\mathbf{a} \in \mathbb{R}^p$:** *Gene expression coefficients* obtained from tissue samples
- **Labels $y \in \{0,1\}$:** Determines healthy ($y = 0$) vs. cancer ($y = 1$) samples
- **Observation:** $n$ copies of $(\mathbf{a}, y)$ namely iid instances $\{(\mathbf{a}_i, y_i)\}_{i=1}^n$
- **Restriction:** Fewer samples than dimensions, i.e., $n < p$
- **Goal:** Find $s \ll p$ entries (i.e., variables) of data points $\mathbf{a}$ using which label $y$ can be predicted with least "error"

## MLE

Nonlinearity

- $y|\mathbf{a}$ has a **likelihood function** that depends on a $s$-sparse parameter vector $\mathbf{x}$

Empirical loss: $f(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n -\log l(\mathbf{x}; \mathbf{a}_i, y_i)$

- Min. the loss (equivalent to max. joint likelihood) to estimate true parameter $\mathbf{x}^\star$



# Prior Work

In statistical estimation framework: convex $f$ + $\ell_1$-regularization

- *Kakade et al.* [AISTAT'09] : Loss functions from exponential family
- *Negahban et al.* [NIPS'09] : M-estimators and "decomposable" norms
- *Agarwal et al.* [NIPS'10] : Projected Gradient Descent with $\ell_1$-constraint

**Issue:** Sparsity cannot be guaranteed to be optimal, because

- Nonlinearity causes solution-dependent error bounds that can become very large
- $\ell_1$-regularization is merely a proxy to induce sparsity

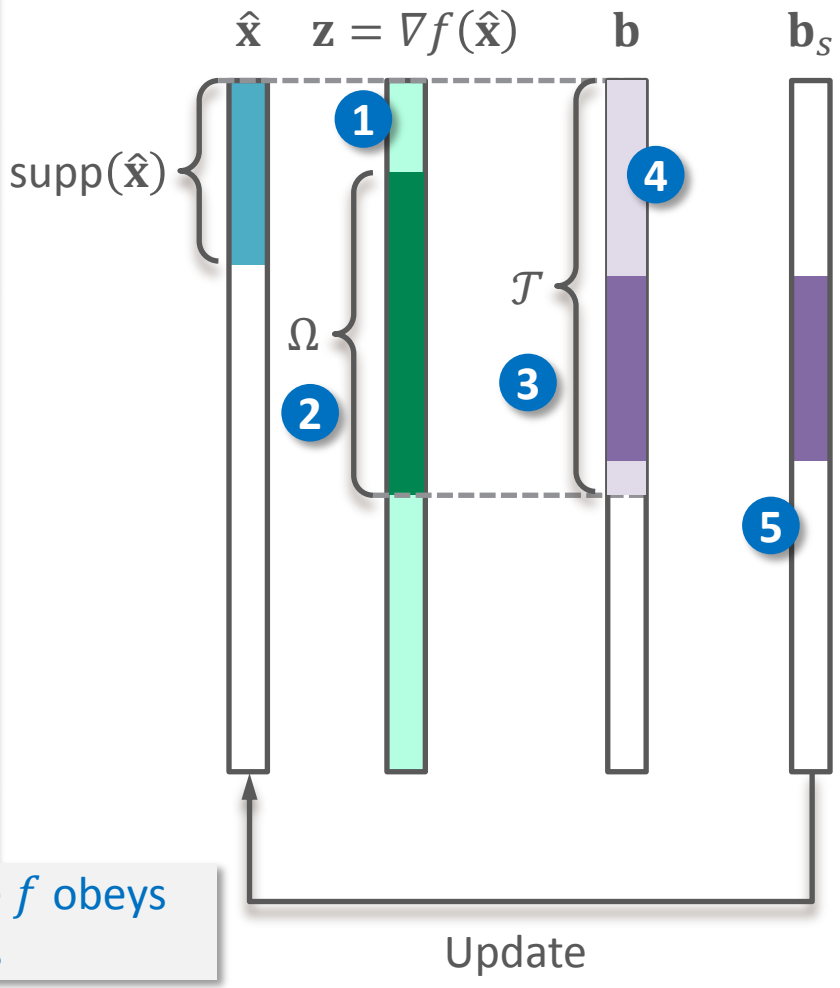
We consider a **greedy algorithm** for the problem

- Algorithm enforces sparsity directly
- Generally has lower computational complexity

# Algorithm

| Gradient Support Pursuit | |
|---|---|
| Input | $f(\cdot)$ and $s$ |
| Output | $\hat{\mathbf{x}}$ |
| **0. Initialize** | $\hat{\mathbf{x}} = \mathbf{0}$ |
| Repeat | |
| **1. Compute Gradient** | $\mathbf{z} = \nabla f(\hat{\mathbf{x}})$ |
| **2. Identify Coordinates** | $\Omega = \text{supp}(\mathbf{z}_{2s})$ |
| **3. Merge Supports** | $\mathcal{T} = \text{supp}(\hat{\mathbf{x}}) \cup \Omega$ |
| **4. Find Crude Estimate** | $\mathbf{b} = \arg\min_{\mathbf{x}} f(\mathbf{x})$ s.t. $\mathbf{x}\|_{\mathcal{T}^c} = \mathbf{0}$ |
| **5. Prune** | $\hat{\mathbf{x}} = \mathbf{b}_s$ |
| Until Halting Condition Holds | |

Inspired by the CoSaMP algorithm [Needell & Tropp '09]




Tractable because $f$ obeys certain conditions

# Main Result

## Theroem

If $f$ satisfies certain properties then the estimate obtained at the $i$-th iteration of GraSP obeys

$$\left\|\hat{\mathbf{x}}^{(i)} - \mathbf{x}^\star\right\|_2 \leq \kappa^i \|\mathbf{x}^\star\|_2 + C \left\| \left| \nabla f(\mathbf{x}^\star) \right|_{\mathcal{I}} \right\|_2,$$

where $\mathcal{I}$ contains the indices of the $3s$ largest coordinates of $\nabla f(\mathbf{x}^\star)$ in magnitude.

- For $\kappa < 1$ (ie., **contraction factor**) we get *linear rate* of convergence up to an **approximation error**

- In statistical estimation problems $\|\nabla f(\mathbf{x}^\star)|_{\mathcal{I}}\|$ can be related to the **statistical precision** of the estimator

# Required Conditions

## Definition (Stable Hessian Property)

For $f: \mathbb{R}^p \longrightarrow \mathbb{R}$ with **Hessian** $\mathbf{H}_f(\cdot)$ let

$$A_k(\mathbf{x}) := \sup_{\substack{|\mathrm{supp}(\mathbf{x}) \cup \mathrm{supp}(\boldsymbol{\Delta})| \leq k \\ \|\boldsymbol{\Delta}\|_2 = 1}} \boldsymbol{\Delta}^{\mathrm{T}} \mathbf{H}_f(\mathbf{x}) \boldsymbol{\Delta}$$

$$B_k(\mathbf{x}) := \inf_{\substack{|\mathrm{supp}(\mathbf{x}) \cup \mathrm{supp}(\boldsymbol{\Delta})| \leq k \\ \|\boldsymbol{\Delta}\|_2 = 1}} \boldsymbol{\Delta}^{\mathrm{T}} \mathbf{H}_f(\mathbf{x}) \boldsymbol{\Delta}.$$

Then we say $f$ satisfies SHP of order $k$ with constant $\mu_k$ if we have

$$\frac{A_k(\mathbf{x})}{B_k(\mathbf{x})} \leq \mu_k$$

for all $k$-sparse vectors $\mathbf{x}$.

- SHP basically says that **symmetric restrictions** of the Hessian are **well-conditioned**
- For $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2$ as in CS, SHP implies the *Restricted Isometry Property*

$$\frac{1+\delta_k}{1-\delta_k} \leq \mu_k \Rightarrow \delta_k \leq \frac{\mu_k - 1}{\mu_k + 1}$$



# Example

**Logistic model:**

$$y \mid \mathbf{a}; \mathbf{x} \sim \text{Bernoulli}\left(\frac{1}{1+e^{-\langle \mathbf{a}, \mathbf{x} \rangle}}\right)$$

For iid observation pairs $\{(\mathbf{a}_i, y_i)\}_{i=1}^n$ write the logistic loss as

$$\mathcal{L}(\mathbf{x}) := \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{\langle \mathbf{a}_i, \mathbf{x} \rangle}\right) - y_i \langle \mathbf{a}_i, \mathbf{x} \rangle .$$

**$\ell_2$-regularized logistic regression with sparsity constraint:**

$$\arg\min_{\mathbf{x}} f(\mathbf{x}) = \mathcal{L}(\mathbf{x}) + \frac{\eta}{2} \|\mathbf{x}\|_2^2$$

$$\text{subject to } \|\mathbf{x}\|_0 \leq s.$$

We can show $\mu_k \leq 1 + \frac{\alpha_k}{4\eta}$, where

$$\alpha_k = \max_{\mathcal{K}} \lambda_{\max}\left(\mathbf{A}_{\mathcal{K}}\right) \quad \text{subject to } |\mathcal{K}| \leq k.$$



# Main Result Revisited

If $f$ satisfies certain properties then the estimate obtained at the $i$-th iteration of GraSP obeys

$$\left\|\hat{\mathbf{x}}^{(i)} - \mathbf{x}^\star\right\|_2 \leq \kappa^i \|\mathbf{x}^\star\|_2 + C \left\|\nabla f(\mathbf{x}^\star)\Big|_{\mathcal{I}}\right\|_2,$$

where $\mathcal{I}$ contains the indices of the $3s$ largest coordinates of $\nabla f(\mathbf{x}^\star)$ in magnitude.

## Theorem

If $f$ satisfies **SHP of order $4s$ with constant $\mu_{4s} < \sqrt{2}$ and $B_{4s}(\mathbf{x}) > \epsilon$,** then the estimate obtained at the $i$-th iteration of GraSP obeys

$$\left\|\hat{\mathbf{x}}^{(i)} - \mathbf{x}^\star\right\|_2 \leq (\mu_{4s}^2 - 1)^i \|\mathbf{x}^\star\|_2 + \frac{2(\mu_{4s}+2)}{\epsilon(2-\mu_{4s}^2)} \left\|\nabla f(\mathbf{x}^\star)\Big|_{\mathcal{I}}\right\|_2,$$

where $\mathcal{I}$ contains the indices of the $3s$ largest coordinates of $\nabla f(\mathbf{x}^\star)$ in magnitude.

# Summary

## Extend CS results to *Nonlinear Models* and *Different Error Measures*

- $\ell_1$-regularization may not yield sufficiently sparse solutions because of the type of cost functions introduced by nonlinearities in the model

## GraSP Algorithm

- Greedy method that always gives a sparse solution
- Accuracy is guaranteed for the class of functions that satisfy SHP
- Linear rate of convergence up to the approximation error

## Some interesting problems to study

- Deterministic results, e.g., using equivalent of incoherence
- Relax SHP to an entirely local condition