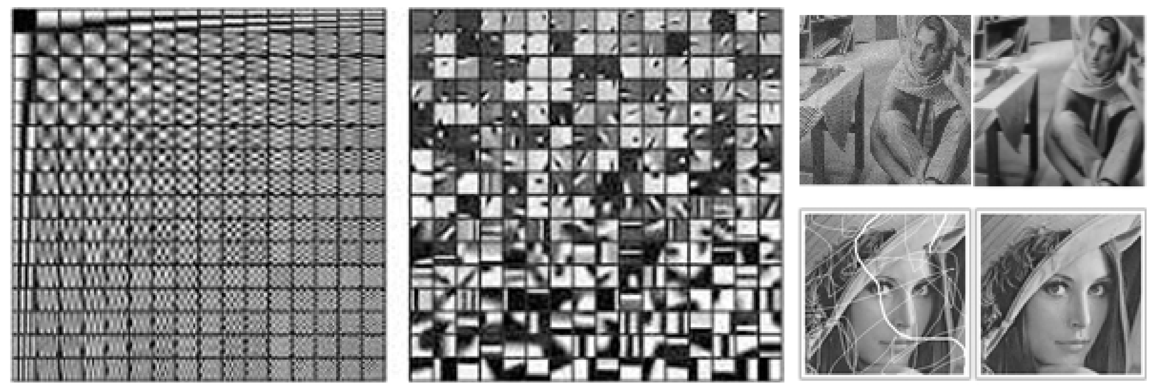
Information Processing & Algorithms Laboratory
Everything can be optimized

ADAPTIVE MATCHING PURSUIT FOR SPARSE SIGNAL RECOVERY

Tiep Vu, Hojjat Mousavi and Vishal Monga
School of Electrical Engineering and Computer Science, The Pennsylvania State University, USA
PennState College of Engineering

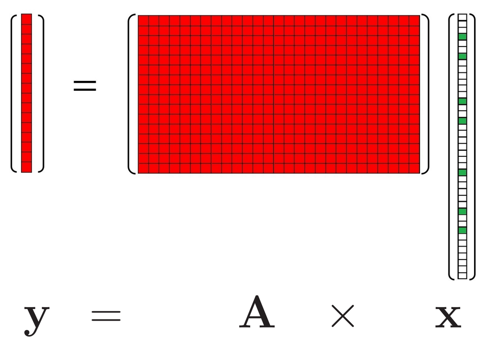
1. Sparsity in Signal/Image Processing

- Parsimony in signal representation has value in many applications.
- A large class of signals can be represented in a compact manner with respect to some basis (dictionary).
- Recent advances using spare representation for a variety of applications:



16x16 DCT bases — Adaptive bases — Denoising/Inpainting

- A measurement (dictionary) matrix $\mathbf{A} \in \mathbb{R}^{q \times p}$
- Compressive Sensing framework ($\mathbf{x} \in \mathbb{R}^p$):
  $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_0$ subject to $\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 \le \epsilon$
- The ℓ_1 relaxation of regularized problem:
  $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 + \lambda\|\mathbf{x}\|_1$



2. Motivation and Idea

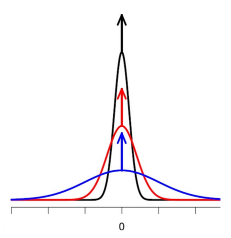
- Model-based Compressive Sensing:

$$\max_{\mathbf{x}} f(\mathbf{x}) \text{ subject to } \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2 < \epsilon$$

Laplacian MAP estimation

$$\arg\max_{\mathbf{x}} \exp\left\{-\lambda \sum_i |x_i|\right\} \Leftrightarrow \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1$$

- Introduce a prior $f_{\mathbf{x}}$: which can capture the sparse structure
- Spike-and-Slab Priors: The gold standard to induce sparsity [1]
- Goal: Find the sparse signal $\mathbf{x} \in \mathbb{R}^p$ from a set of fewer measurements $\mathbf{y} \in \mathbb{R}^q$. $(q \ll p)$



3. Bayesian Interpretation

Bayesian formulation

$$\mathbf{y}|\mathbf{A}, \mathbf{x}, \gamma, \sigma^2 \sim \mathcal{N}(\mathbf{A}\mathbf{x}, \sigma^2\mathbf{I}),$$
$$\mathbf{x}|\gamma, \lambda, \sigma^2 \sim \prod_{i=1}^{p} \gamma_i \mathcal{N}(0, \sigma^2\lambda^{-1}) + (1-\gamma_i)\mathbb{I}(x_i = 0),$$
$$\gamma \sim \prod_{i=1}^{p} \text{Bernoulli}(\kappa_i),$$

- The optimal $\mathbf{x}^*, \gamma^*$ are obtained by MAP estimation on joint posterior density:

$$(\mathbf{x}^*, \gamma^*) = \arg\max_{\mathbf{x},\gamma} \left\{p(\mathbf{x}, \gamma | \mathbf{A}, \mathbf{y}, \boldsymbol{\kappa}, \lambda, \sigma^2)\right\}. \quad (1)$$

Optimization problem [2]

$$(1) \Leftrightarrow (\mathbf{x}^*, \gamma^*) = \arg\min_{\mathbf{x},\gamma} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_2^2 + \sum_{i=1}^{p} \rho_i\gamma_i \quad (2)$$

where $\rho_i \triangleq \sigma^2 \log\left(\frac{2\pi\sigma^2(1-\kappa_i)^2}{\lambda\kappa_i^2}\right)$.

- This is a hard nonconvex mixed-integer programming to find both the support and the values of the solution.

5. Existing Solutions/Special Cases

1. Markov Chain Monte Carlo (MCMC algorithm) [3]
2. Majorization minimization (MM algorithm) [4] by assuming only one κ

$$\arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_2^2 + \rho\|\mathbf{x}\|_0$$

3. Adaptive Elastic Net [5] by relaxing ℓ_0 norm to ℓ_1 norm

$$\arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_2^2 + \rho\|\mathbf{x}\|_1$$

4. Sparse Reconstruction by Separable Approximation(SpaRSA) [6]

6. AMP: Adaptive Matching Pursuit - First observations

- First, let $\mathbf{D} = \begin{bmatrix}\mathbf{A} \\ \sqrt{\lambda}\mathbf{I}\end{bmatrix}$ and $\mathbf{z} = \begin{bmatrix}\mathbf{y} \\ \mathbf{0}\end{bmatrix}$ with $\mathbf{I} \in \mathbb{R}^{p \times p}$ and $\mathbf{0} \in \mathbb{R}^{p \times 1}$ being the identity matrix and zero vector, we can rewrite (2) as:

$$(\mathbf{x}^*, \gamma^*) = \arg\min_{\mathbf{x},\gamma} \|\mathbf{z} - \mathbf{D}\mathbf{x}\|_2^2 + \sum_{i=1}^{p} \rho_i\gamma_i. \quad (3)$$

where $\|\mathbf{d}_i\|_2^2 = \|\mathbf{a}_i\|_2^2 + \lambda = 1 + \lambda,\ \forall i = 1, \dots, p$

- **If we know the true support of the signal**, i.e. $\mathcal{S} = \{i : \gamma_i \neq 0\}$, **we can infer the solution** of (3) by:

$$\mathbf{x}_{\mathcal{S}} = \arg\min_{\mathbf{x}_{\mathcal{S}}} \|\mathbf{z} - \mathbf{D}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}\|_2^2 \Rightarrow \boxed{\mathbf{D}_{\mathcal{S}}^T\mathbf{D}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}} = \mathbf{D}_{\mathcal{S}}^T\mathbf{z} \triangleq \mathbf{w}_{\mathcal{S}}} \text{ and } \mathbf{r}_{\mathcal{S}} = \mathbf{z} - \mathbf{D}_{\mathcal{S}}\mathbf{x}_{\mathcal{S}}$$

7. AMP: The central idea

Define: $g(\mathcal{S}) = \min_{\mathbf{x}^{\mathcal{S}}} \|\mathbf{z} - \mathbf{D}_{\mathcal{S}}\mathbf{x}^{\mathcal{S}}\|_2^2 + \sum_{i \in \mathcal{S}} \rho_i.$

best improvement if we **insert**: $U_{\mathcal{S}} = \min_{i \notin \mathcal{S}} g(\mathcal{S} \cup \{i\}) - g(\mathcal{S}),$

best improvement if we **remove**: $V_{\mathcal{S}} = \min_{j \in \mathcal{S}} g(\mathcal{S} \backslash \{j\}) - g(\mathcal{S}).$

- If $U_{\mathcal{S}}, V_{\mathcal{S}} > 0 \Rightarrow$ stop. If $U_{\mathcal{S}} < V_{\mathcal{S}} \Rightarrow$ insert, else remove.
- **Challenge:** calculating $U_{\mathcal{S}}, V_{\mathcal{S}}$ is expensive.
- **Solution:** find their *tight upper bounds.*

8. AMP: Supporting lemmas and Cholesky decomposition

Lemma 1: Initialization of the support set $\mathcal{S}$

If $\rho_i < 0$ ($\kappa_i >$ a threshold), then $i \in \hat{\mathcal{S}}$ – the optimal active set $\Rightarrow$ initialize $\mathcal{S}_0 = \{i : \rho_i < 0\}$

Lemma 2: Approximation of $U_{\mathcal{S}}$

$$U_{\mathcal{S}} \le \min_{i \notin \mathcal{S}} \left\{\rho_i - \frac{(\mathbf{r}_{\mathcal{S}}^T\mathbf{d}_i)^2}{1+\lambda}\right\} \triangleq \bar{U}_{\mathcal{S}}$$

Lemma 3: Approximation of $V_{\mathcal{S}}$

$$V_{\mathcal{S}} \le \min_{j \in \mathcal{S}} \left\{(1+\lambda)(x_j^{\mathcal{S}})^2 + 2\mathbf{d}_j^T\mathbf{r}_{\mathcal{S}}x_j^{\mathcal{S}} - \rho_j\right\} \triangleq \bar{V}_{\mathcal{S}}.$$

- Challenge: Once $\mathcal{S}$ is found, we need to solve: $\mathbf{D}_{\mathcal{S}}^T\mathbf{D}_{\mathcal{S}}\mathbf{x}^{\mathcal{S}} = \mathbf{w}^{\mathcal{S}}$ which may be expensive.
- Solution: Use the Cholesky decomposition: $\mathbf{D}_{\mathcal{S}}^T\mathbf{D}_{\mathcal{S}} = \mathbf{L}_{\mathcal{S}}^T\mathbf{L}_{\mathcal{S}}$ where $\mathbf{L}_{\mathcal{S}}$ is a low triangular matrix.

Forward Cholesky decomposition

If $\mathbf{v}$ is the solution of $\mathbf{L}_{\mathcal{S}}\mathbf{v} = \mathbf{D}_{\mathcal{S}}^T\mathbf{d}_i$ via forward substitution, then:

$$\mathbf{L}_{\mathcal{S}\cup\{i\}} = \begin{bmatrix}\mathbf{L}_{\mathcal{S}} & \mathbf{0} \\ \mathbf{v}^T & \sqrt{1+\lambda-\mathbf{v}^T\mathbf{v}}\end{bmatrix}.$$

Backward Cholesky decomposition

$$\text{If } \mathbf{L}_{\mathcal{S}} = \begin{bmatrix}\mathbf{L}_{11} & 0 & 0 \\ \mathbf{l}_{21}^T & l_{22} & 0 \\ \mathbf{L}_{31} & \mathbf{l}_{32} & \mathbf{L}_{33}\end{bmatrix} \leftarrow j^{\text{th}} \text{ row}$$

$$\Rightarrow \mathbf{L}_{\mathcal{S}\backslash\{j\}} = \begin{bmatrix}\mathbf{L}_{11} & 0 \\ \mathbf{L}_{31} & \bar{\mathbf{L}}_{33}\end{bmatrix} \text{ where } \bar{\mathbf{L}}_{33}\bar{\mathbf{L}}_{33}^T = \mathbf{L}_{33}\mathbf{L}_{33}^T + \mathbf{l}_{32}\mathbf{l}_{32}^T$$

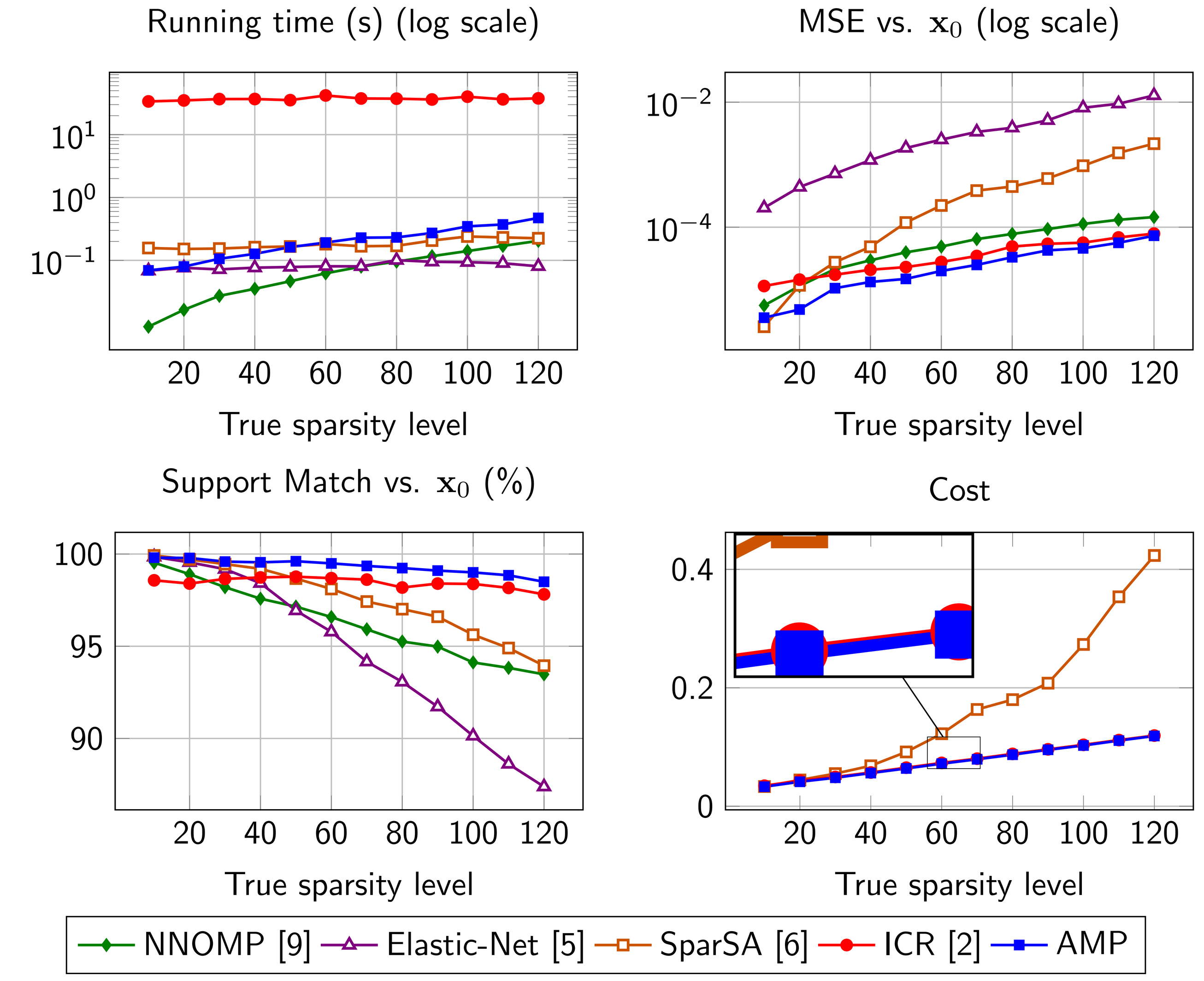
9. Experiments with simulated data

- 1000 realizations of $\mathbf{A}$, sparse $\mathbf{x}_0$ and Gaussian noise ε: $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \varepsilon$
- Coefficient vector $\mathbf{x}_0 \in \mathbb{R}^{512}$. Test vector $\mathbf{y} \in \mathbb{R}^{256}$. Sparsity level $= 100$.

Table: Comparison of methods for $p = 512$, $q = 256$.

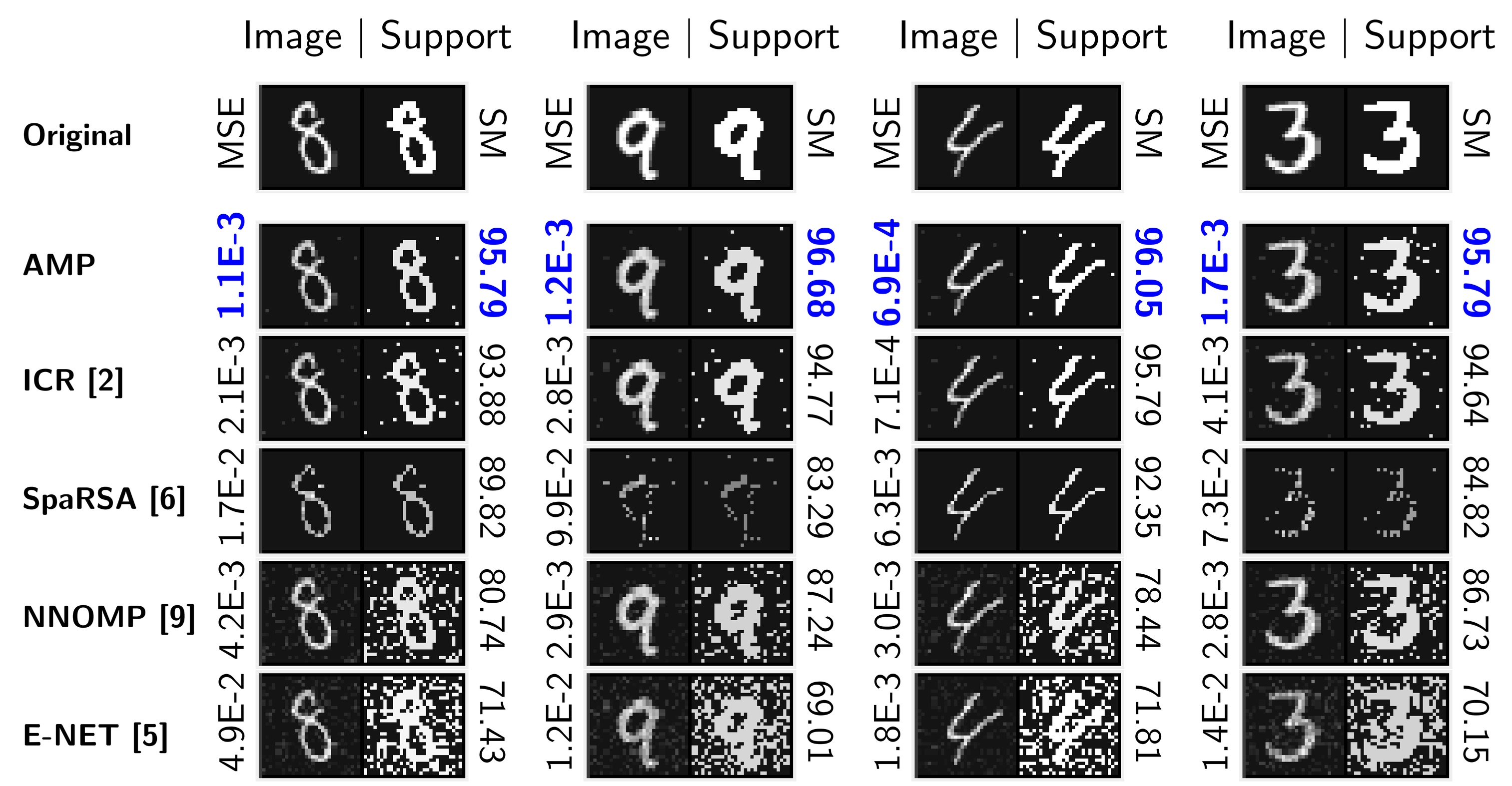
Average	CoSaMP [7]	E-NET [5]	RW-L1 [8]	SpaRSA [6]	ICR [2]	**AMP**
Time (s)	1.1E0	**7.9E-2**	4.7E-1	2.2E-1	2.9E1	***1.6E-1***
MSE	3.0E-4	4.5E-3	2E-4	5.4E-4	3.1E-3	**6.1E-5**
Cost	-	-	-	1.6E-1	2.1E-1	**9.5E-2**
SM (%)	93.52	85.55	92.71	93.87	83.79	**97.32**

- With Nonnegativity constraints:



10. Experiments with real data

- Reconstruction of handwritten digit images from the MNIST dataset.
- MNIST contains 60000 digit images (0 to 9) of size 28×28 pixels.
- Sparse signal $\mathbf{x}_0$ (vectorized image) is to be reconstructed from a smaller set of random measurements $\mathbf{y}$: $\mathbf{y} = \mathbf{A}\mathbf{x}_0 + \varepsilon$
- Coefficient vector $\mathbf{x}_0 \in \mathbb{R}^{784}$. measured vector $\mathbf{y} \in \mathbb{R}^{350}$.



[1] M. Andersen et. al, NIPS 2014
[2] H. Mousavi et. al, SPL 2015
[3] H. Mousavi et. al, ICIP 2014
[4] T.J. Yen et. al., Ann. Statist., 2011
[5] H. Zou et. al., Ann. Statist., 2009.
[6] S. J. Wright, et. al. TSP 2009
[7] J. Tropp et. al, App. Com. Har. Ana. 2009
[8] E. Candes, et. al., J. Four. Anal. App., 2008.
[9] A. Bruckstein et. al, TIT 2008

Research was supported by NSF CAREER award and Office of Naval Research (ONR) grant.
IEEE ICASSP 2017 (March 5-9, New Orleans, USA)
Email: tiepvu@psu.edu,
Web: http://signal.ee.psu.edu/AMP.html