ℓ_1-Minimization, Group Sparsity, and Algorithm Parallelization

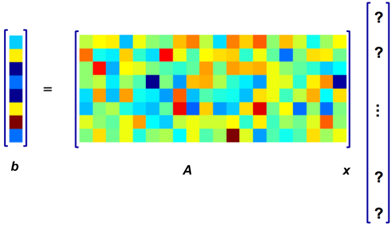
Allen Y. Yang
yang@eecs.berkeley.edu

IJCB 2011 Tutorial –
Sparse Representation and Low-Rank Representation for Biometrics

Compressive Sensing Theory: An Introduction

- Compressive Sensing (CS) deals with an estimation problem in **underdetermined systems** of linear equations, A in general is full rank:

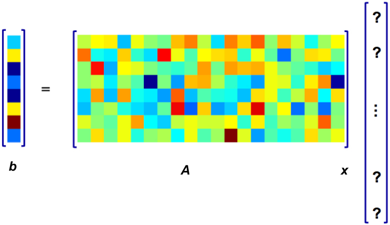
$$\mathbf{b} = A\mathbf{x} \quad \text{where } A \in \mathbb{R}^{d\times n}, (d < n)$$



Compressive Sensing Theory: An Introduction

- Compressive Sensing (CS) deals with an estimation problem in **underdetermined systems** of linear equations, A in general is full rank:

$$\mathbf{b} = A\mathbf{x} \quad \text{where } A \in \mathbb{R}^{d \times n}, (d < n)$$

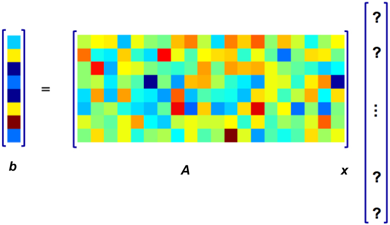


- Two interpretations:
  1. Compression: A as a sensing matrix.
  2. Representation: A as a prior dictionary.

Compressive Sensing Theory: An Introduction

- Compressive Sensing (CS) deals with an estimation problem in **underdetermined systems** of linear equations, A in general is full rank:

$$\mathbf{b} = A\mathbf{x} \quad \text{where } A \in \mathbb{R}^{d \times n}, (d < n)$$



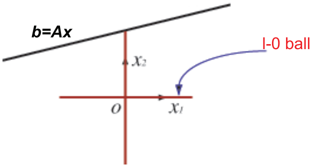
- Two interpretations:
  1. Compression: A as a sensing matrix.
  2. Representation: A as a prior dictionary.
- Infinitely many solutions for $\mathbf{x}$, without extra constraints

ℓ_0/ℓ_1 Equivalence Relationship for Sparsest Solutions

- ℓ_0-Minimization (**NP-Hard**)

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{subj. to } \mathbf{b} = A\mathbf{x}.$$

$\|\cdot\|_0$ simply counts the number of nonzero terms.

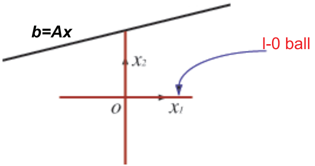


ℓ_0/ℓ_1 Equivalence Relationship for Sparsest Solutions

- ℓ_0-Minimization (**NP-Hard**)

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{subj. to } \mathbf{b} = A\mathbf{x}.$$

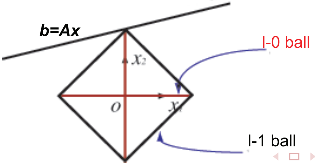
$\|\cdot\|_0$ simply counts the number of nonzero terms.



- ℓ_1-Minimization (**Linear Program**)

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{subj. to } \mathbf{b} = A\mathbf{x}.$$

$\|\mathbf{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n|$.



Feasibility and Uniqueness: ℓ_0-Minimization

Spark Condition

- Spark(A): smallest number of columns that are linearly dependent
  1. Example I: Identity matrix $I \in \mathbb{R}^{d \times d}$, Spark$(A)$ = d+1;
  2. Example II: $\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$, Spark$(A) = 2$;
  3. Example III: Random matrix $[\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n] \in \mathbb{R}^{d \times n}$, Spark$(A)$ = d+1 (with high probability);

Feasibility and Uniqueness: ℓ_0-Minimization

Spark Condition

- Spark(A): smallest number of columns that are linearly dependent
  1. Example I: Identity matrix $I \in \mathbb{R}^{d\times d}$, Spark($A$) = d+1;
  2. Example II: $\begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$, Spark($A$) = 2;
  3. Example III: Random matrix $[\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n] \in \mathbb{R}^{d\times n}$, Spark($A$) = d+1 (with high probability);
- Sparse signal $\mathbf{x}$ can be **uniquely** recovered by ℓ_0-min if

$$\|\mathbf{x}\|_0 < \frac{\text{Spark}(A)}{2}$$

Proof.

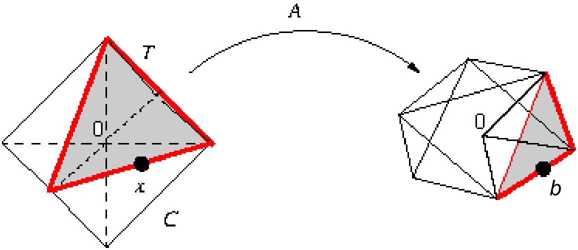
1. Suppose $\mathbf{x}_1 \neq \mathbf{x}_2$ both satisfy the spark condition, and $\mathbf{b} = A\mathbf{x}_1$, $\mathbf{b} = A\mathbf{x}_2$.
2. $A(\mathbf{x}_1 - \mathbf{x}_2) \doteq A\mathbf{y} = \mathbf{b} - \mathbf{b} = 0$.
3. But $\|\mathbf{y}\|_0 < \frac{\text{Spark}(A)}{2} + \frac{\text{Spark}(A)}{2} = \text{Spark}(A)$. **Contradiction.**

$\square$

Estimating Spark(A) is as expensive as ℓ_0-min itself!

Feasibility and Uniqueness: ℓ_1-Minimization

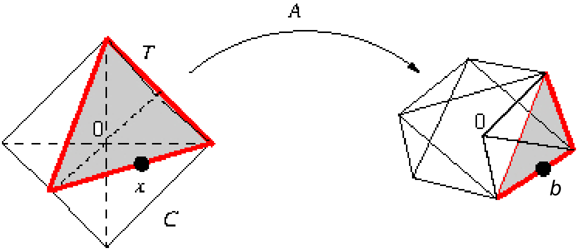
k-Neighborliness Condition



- Define **cross polytope** C and **quotient polytope** P such that $P = AC$.
- $\mathbf{x}$ is k-sparse $\Leftrightarrow$ $\mathbf{x}$ lies on a unique $(k-1)$-face of C.

Feasibility and Uniqueness: ℓ_1-Minimization

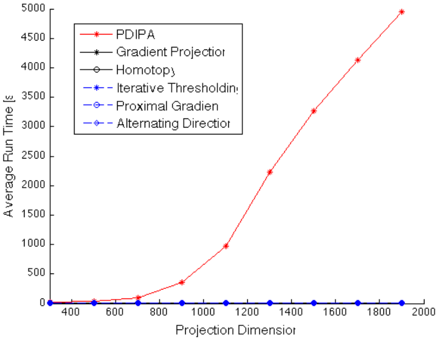
k-Neighborliness Condition



- Define **cross polytope** C and **quotient polytope** P such that $P = AC$.
- $\mathbf{x}$ is k-sparse $\Leftrightarrow$ $\mathbf{x}$ lies on a unique $(k-1)$-face of C.
- **Necessary and Sufficient:**
  1. If the $(k-1)$-face where $\mathbf{x}$ lies maps to a face of P, then ℓ^1/ℓ^0 holds for this specific $\mathbf{x}$.
  2. If all $(k-1)$-faces of C map to the faces of P on the boundary, ℓ^1/ℓ^0 holds for all k-sparse signals.

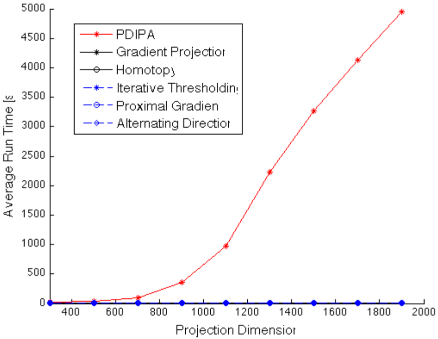
Why ℓ_1-Minimization is still a difficult problem?

- General linear-programming toolboxes do exist: **cvx**, **SparseLab**. However, interior-point methods are **very** expensive in HD space.



Why ℓ_1-Minimization is still a difficult problem?

- General linear-programming toolboxes do exist: **cvx**, **SparseLab**. However, interior-point methods are **very** expensive in HD space.



- Improve speed via new numerical algorithms.
- Improve accuracy by exploiting finer (group) data structure of the problems.

$$\mathbf{b} = [A_1, A_2, \cdots, A_K]\mathbf{x} + \mathbf{e}.$$

- Implement ℓ_1-min and SRC on multi-core CPUs/GPUs.

ℓ_1-Min Literature

1. **Primal-Dual Interior-Point**
   - Log-Barrier [Frisch '55, Karmarkar '84, Megiddo '89, Monteiro-Adler '89, Kojima-Megiddo-Mizuno '93]
2. **Homotopy**
   - Homotopy [Osborne-Presnell-Turlach '00, Malioutov-Cetin-Willsky '05, Donoho-Tsaig '06]
   - Polytope Faces Pursuit (PFP) [Plumbley '06]
   - Least Angle Regression (LARS) [Efron-Hastie-Johnstone-Tibshirani '04]
3. **Gradient Projection**
   - Gradient Projection Sparse Representation (GPSR) [Figueiredo-Nowak-Wright '07]
   - Truncated Newton Interior-Point Method (TNIPM) [Kim-Koh-Lustig-Boyd-Gorinevsky '07]
4. **Iterative Thresholding**
   - Soft Thresholding [Donoho '95]
   - Sparse Reconstruction by Separable Approximation (SpaRSA) [Wright-Nowak-Figueiredo '08]
5. **Proximal Gradient** [Nesterov '83, Nesterov '07]
   - FISTA [Beck-Teboulle '09]
   - Nesterov's Method (NESTA) [Becker-Bobin-Candés '09]
6. **Augmented Lagrangian Methods** [Yang-Zhang '09, AY et al '10]
   - Bergman [Yin et al. '08]
   - YALL1 [Yang-Zhang '09]
   - SALSA [Figueiredo et al. '09]
   - Primal ALM, Dual ALM [AY et al '10]

Reference:

AY, *et al.*, *A review of fast ℓ_1-minimization algorithms for robust face recognition*. ICIP, 2010.

Homotopy Methods

- The existence of measurement errors (assume Gaussian)

$$\mathbf{x}^* = \arg\min \|\mathbf{x}\|_1 \quad \text{subj. to } \|\mathbf{e}\|_2 = \|\mathbf{b} - A\mathbf{x}\|_2 < \epsilon$$

Homotopy Methods

- The existence of measurement errors (assume Gaussian)

$$\mathbf{x}^* = \arg\min \|\mathbf{x}\|_1 \quad \text{subj. to } \|\mathbf{e}\|_2 = \|\mathbf{b} - A\mathbf{x}\|_2 < \epsilon$$

- Lagrangian method

$$\begin{aligned} \mathbf{x}^* = \arg\min F(\mathbf{x}) &= \arg\min \frac{1}{2}\|\mathbf{b} - A\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1 \\ &\doteq \arg\min f(\mathbf{x}) + \lambda g(\mathbf{x}) \end{aligned}$$

Homotopy Methods

- The existence of measurement errors (assume Gaussian)

$$\mathbf{x}^* = \arg\min \|\mathbf{x}\|_1 \quad \text{subj. to } \|\mathbf{e}\|_2 = \|\mathbf{b} - A\mathbf{x}\|_2 < \epsilon$$

- Lagrangian method

$$\begin{aligned} \mathbf{x}^* = \arg\min F(\mathbf{x}) &= \arg\min \frac{1}{2}\|\mathbf{b} - A\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1 \\ &\doteq \arg\min f(\mathbf{x}) + \lambda g(\mathbf{x}) \end{aligned}$$

- **Homotopy** refers to the fact

$$\begin{aligned} \text{When } \lambda \to +\infty \quad & \mathbf{x}^* \to 0; \\ \text{When } \lambda \to 0 \quad & \mathbf{x}^* \to \mathbf{b} = A\mathbf{x}. \end{aligned}$$

Homotopy Methods

- The existence of measurement errors (assume Gaussian)

$$\mathbf{x}^* = \arg\min \|\mathbf{x}\|_1 \quad \text{subj. to } \|\mathbf{e}\|_2 = \|\mathbf{b} - A\mathbf{x}\|_2 < \epsilon$$

- Lagrangian method

$$\begin{aligned} \mathbf{x}^* = \arg\min F(\mathbf{x}) &= \arg\min \frac{1}{2}\|\mathbf{b} - A\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1 \\ &\doteq \arg\min f(\mathbf{x}) + \lambda g(\mathbf{x}) \end{aligned}$$

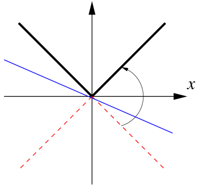
- **Homotopy** refers to the fact

$$\begin{aligned} \text{When } \lambda \to +\infty \qquad & \mathbf{x}^* \to 0; \\ \text{When } \lambda \to 0 \qquad & \mathbf{x}^* \to \mathbf{b} = A\mathbf{x}. \end{aligned}$$

- $F(\mathbf{x}) = f(\mathbf{x}) + \lambda g(\mathbf{x})$ is called a **composite objective function**
  1. $f(\mathbf{x}) = \frac{1}{2}\|\mathbf{b} - A\mathbf{x}\|_2^2$ is convex and smooth.
  2. $g(\mathbf{x}) = \|\mathbf{x}\|_1$ is convex but **not smooth**.
  3. As a result, $\nabla F(\mathbf{x})$ does not exist!

Subgradient Method

- The anomaly of $\nabla \|\mathbf{x}\|_1$ occurs exactly at those coefficients where $x_i = 0$.

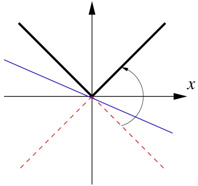


- **Subdifferential**

$$\partial |x_i| \doteq u_i = \begin{cases} +1 & \text{when } x_i > 0 \\ -1 & \text{when } x_i < 0 \\ [-1, 1] & \text{when } x_i = 0 \end{cases}$$

Subgradient Method

- The anomaly of $\nabla\|\mathbf{x}\|_1$ occurs exactly at those coefficients where $x_i = 0$.



- **Subdifferential**

$$\partial|x_i| \doteq u_i = \begin{cases} +1 & \text{when } x_i > 0 \\ -1 & \text{when } x_i < 0 \\ [-1, 1] & \text{when } x_i = 0 \end{cases}$$

Homotopy Algorithm

1. Initialization: $\mathbf{x} = 0$; set a large value for λ.
2. In kth iteration: Set $\partial F = 0 \Leftrightarrow \partial g(\mathbf{x}) = -(1/\lambda)\partial f(\mathbf{x})$.
3. Update $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + \gamma\Delta\mathbf{x}$ based on $\partial g(\mathbf{x})$.
4. Reduce $\lambda \to 0$ and jump to **(2)**.

Augmented Lagrangian Method (ALM)

- ℓ_1-Min:

$$\mathbf{x}^* = \arg\min \|\mathbf{x}\|_1 \quad \text{subj. to} \quad \mathbf{b} = A\mathbf{x}$$

(adding a penalty term for the equality constraint)

$$L_\mu(\mathbf{x}) = \|\mathbf{x}\|_1 + \frac{\mu}{2}\|\mathbf{b} - A\mathbf{x}\|_2^2 \quad \text{subj. to} \quad \mathbf{b} = A\mathbf{x}.$$

Augmented Lagrangian Method (ALM)

- ℓ_1-Min:

$$\mathbf{x}^* = \arg\min \|\mathbf{x}\|_1 \quad \text{subj. to} \quad \mathbf{b} = A\mathbf{x}$$

(adding a penalty term for the equality constraint)

$$L_\mu(\mathbf{x}) = \|\mathbf{x}\|_1 + \frac{\mu}{2}\|\mathbf{b} - A\mathbf{x}\|_2^2 \quad \text{subj. to} \quad \mathbf{b} = A\mathbf{x}.$$

- Augmented Lagrange Function [Bertsekas '03]:

$$L_\mu(\mathbf{x}, \mathbf{y}) = \|\mathbf{x}\|_1 + \langle \mathbf{y}, \mathbf{b} - A\mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{b} - A\mathbf{x}\|_2^2,$$

where $\mathbf{y}$ is the Lagrange multipliers for the constraint $\mathbf{b} = A\mathbf{x}$.

Convergence of ALM [Hestenes '69, Powell '69, Bertsekas '03]

$$L_\mu(\mathbf{x}, \mathbf{y}) = \|\mathbf{x}\|_1 + \langle \mathbf{y}, \mathbf{b} - A\mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{b} - A\mathbf{x}\|_2^2,$$

1. When $\mathbf{y}$ close to $\mathbf{y}^*$, by Lagrange Multiplier Theory,

$$\arg\min_{\mathbf{x}} L_\mu(\mathbf{x}, \mathbf{y}^*) = \arg\min_{\mathbf{x}} L_\mu(\mathbf{x}).$$

Convergence of ALM [Hestenes '69, Powell '69, Bertsekas '03]

$$L_\mu(\mathbf{x},\mathbf{y}) = \|\mathbf{x}\|_1 + \langle \mathbf{y}, \mathbf{b} - A\mathbf{x}\rangle + \frac{\mu}{2}\|\mathbf{b} - A\mathbf{x}\|_2^2,$$

1. When $\mathbf{y}$ close to $\mathbf{y}^*$, by Lagrange Multiplier Theory,

$$\arg\min_{\mathbf{x}} L_\mu(\mathbf{x},\mathbf{y}^*) = \arg\min_{\mathbf{x}} L_\mu(\mathbf{x}).$$

2. When μ is very large, high cost of infeasibility implies

$$L_\mu(\mathbf{x},\mathbf{y}) \approx (P_1).$$

Convergence of ALM [Hestenes '69, Powell '69, Bertsekas '03]

$$L_\mu(\mathbf{x},\mathbf{y}) = \|\mathbf{x}\|_1 + \langle \mathbf{y}, \mathbf{b} - A\mathbf{x}\rangle + \frac{\mu}{2}\|\mathbf{b} - A\mathbf{x}\|_2^2,$$

1. When $\mathbf{y}$ close to $\mathbf{y}^*$, by Lagrange Multiplier Theory,

$$\arg\min_{\mathbf{x}} L_\mu(\mathbf{x},\mathbf{y}^*) = \arg\min_{\mathbf{x}} L_\mu(\mathbf{x}).$$

2. When μ is very large, high cost of infeasibility implies

$$L_\mu(\mathbf{x},\mathbf{y}) \approx (P_1).$$

Theorem: Convergence of ALM [Bertsekas '03]

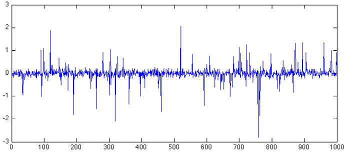
When optimize $L_\mu(\mathbf{x},\mathbf{y})$ w.r.t. a sequence $\mu^k \to \infty$, and $\{\mathbf{y}^k\}$ is bounded, then the limit point of $\{\mathbf{x}^k\}$ is the global minimum of the original problem, namely, ℓ_1-min.

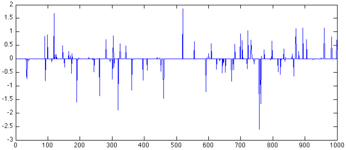
Minimize Augmented Lagrangian

- **Update $\mathbf{y}^{k+1}$: The Method of Multipliers** [Rockafellar '73]
  Assume $(\mathbf{x}^k, \mu^k)$ fixed,
  $$\mathbf{y}^{k+1} = \mathbf{y}^k + \mu^k \nabla_{\mathbf{y}} L_{\mu^k}(\mathbf{x}^k, \mathbf{y}^k)$$
  with complexity $O(dn)$.

Minimize Augmented Lagrangian

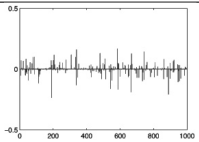
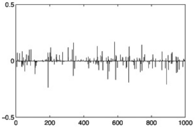
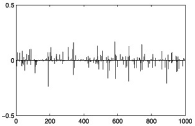
- **Update $\mathbf{y}^{k+1}$: The Method of Multipliers** [Rockafellar '73]
  Assume $(\mathbf{x}^k, \mu^k)$ fixed,
  $$\mathbf{y}^{k+1} = \mathbf{y}^k + \mu^k \nabla_{\mathbf{y}} L_{\mu^k}(\mathbf{x}^k, \mathbf{y}^k)$$
  with complexity $O(dn)$.
- **Update $\mathbf{x}^{k+1}$: Nesterov's Method** [Nesterov '07, Becker et al. '09]
  - Let $f(\mathbf{x}) = \frac{\mu}{2}\|\mathbf{b} - A\mathbf{x}\|_2^2 + \langle \mathbf{y}^k, \mathbf{b} - A\mathbf{x}\rangle$ and $g(\mathbf{x}) = \|\mathbf{x}\|_1$:
    $$L_{\mu^k}(\mathbf{x}, \mathbf{y}^k) = f(\mathbf{x}) + g(\mathbf{x})$$
  - Form a second-order upper bound of $L_{\mu^k}(\mathbf{x}, \mathbf{y}^k)$ based on two step history $(\mathbf{x}^k, \mathbf{x}^{k-1})$:
    $$\begin{aligned} \mathbf{z}^k &= \alpha_1 \mathbf{x}^k + \alpha_2 \mathbf{x}^{k-1} \\ Q(\mathbf{x}, \mathbf{z}) &\doteq f(\mathbf{z}) + \langle \nabla f(\mathbf{z}), \mathbf{x} - \mathbf{z}\rangle + \frac{L}{2}\|\mathbf{x} - \mathbf{z}\|^2 + g(\mathbf{x}). \end{aligned} \tag{1}$$
  - Minimize $Q(\mathbf{x}, \mathbf{z})$ via **soft-thresholding**: $\text{soft}(x, a) = \text{sgn}(x)\max(|x| - a, 0)$





Simulation: Speed of ℓ_1-Min Solvers

Table: Source signal in 1000-D: sparsity = 200; random projection = 600-D.

Algorithm	Estimate	Runtime
PDIPA		63 s
Homotopy		1.7 s
ALM		0.16 s

Group Sparsity Minimization

- Convexification of **entry-wise sparsity** [Donoho & Elad '03, Donoho '05, Candès & Tao '06]

$$
\begin{aligned}
(P_0): \quad & \mathbf{x}_0^* = \arg\min \|\mathbf{x}\|_0 \quad \text{subj. to} \quad A\mathbf{x} = \mathbf{b} \\
(P_1): \quad & \mathbf{x}_1^* = \arg\min \|\mathbf{x}\|_1 \quad \text{subj. to} \quad A\mathbf{x} = \mathbf{b}
\end{aligned}
$$

Group Sparsity Minimization

- Convexification of **entry-wise sparsity** [Donoho & Elad '03, Donoho '05, Candès & Tao '06]

$$(P_0): \quad \mathbf{x}_0^* = \arg\min \|\mathbf{x}\|_0 \quad \text{subj. to} \quad A\mathbf{x} = \mathbf{b}$$
$$(P_1): \quad \mathbf{x}_1^* = \arg\min \|\mathbf{x}\|_1 \quad \text{subj. to} \quad A\mathbf{x} = \mathbf{b}$$

- Convexification of **group sparsity** $\ell_{0,p}$-min: Let $A = [A_1, \cdots, A_K]$

$$(P_{0,p}): \quad \mathbf{x}_{0,p}^* = \underset{\mathbf{x}}{\operatorname{argmin}} \sum_{i=1}^{K} \mathcal{I}(\|\mathbf{x}_i\|_p > 0), \quad \text{subj. to} \quad A\mathbf{x} \doteq [A_1 \quad \cdots \quad A_K] \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_K \end{bmatrix} = \mathbf{b}$$

1. Uniqueness and stability [Lu & Do '08, Blumensath & Davies '09]

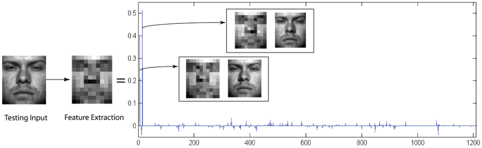
   group Spark condition, group RIP condition, etc.

2. Efficient convex surrogates [Eldar & Mishali '09, Stojnic et al. '09, Sprechmann et al. '10, Elhamifar & Vidal '11]

$$(P_{1,2}): \quad x_{1,2}^* = \arg\min \sum_{i=1}^{K} \|\mathbf{x}_i\|_2 \quad \text{subj. to} \quad A\mathbf{x} = \mathbf{b}$$

Application: Robust Face Recognition

1. Face subspace model [Belhumeur et al. '97, Basri & Jacobs '03]
   Assume $\mathbf{b}$ belongs to Class i:
   $$\mathbf{b} = A_i \mathbf{x}_i$$
2. **Sparse representation** encodes membership [Wright et al. '09, '10]
   $$\begin{aligned} & \mathbf{b} = [A_1, A_2, \cdots, A_K][\mathbf{x}_1; \mathbf{x}_2; \cdots; \mathbf{x}_K] = A\mathbf{x} \\ \Rightarrow \quad & \mathbf{x}^* = [0; \cdots; 0; \mathbf{x}_i; 0; \cdots; 0] \end{aligned}$$



3. In the presence of gross image corruption: **Cross-and-Bouquet model**
   $$\min \|\mathbf{x}\|_1 + \|\mathbf{e}\|_1 \quad \text{subj. to} \quad \mathbf{b} = A\mathbf{x} + \mathbf{e}$$

Robust Face Recognition as a Group Sparsity Recovery Problem

Can we frame robust face recognition using group sparsity?

1. **Negative:**

$$\mathbf{b} = [A_1, A_2, \cdots, A_K, I][\mathbf{x}_1; \mathbf{x}_2; \cdots; \mathbf{x}_K; \mathbf{e}]$$

Standard group sparsity formulation, whereby $\mathbf{e}$ is treated as the $(K+1)$th group, has a trivial solution of 1-group-sparsity:

$$\mathbf{e} = \mathbf{b}; \mathbf{x} = \mathbf{0}.$$

Robust Face Recognition as a Group Sparsity Recovery Problem

Can we frame robust face recognition using group sparsity?

1. **Negative:**

$$\mathbf{b} = [A_1, A_2, \cdots, A_K, I][\mathbf{x}_1; \mathbf{x}_2; \cdots; \mathbf{x}_K; \mathbf{e}]$$

Standard group sparsity formulation, whereby $\mathbf{e}$ is treated as the $(K+1)$th group, has a trivial solution of 1-group-sparsity:

$$\mathbf{e} = \mathbf{b}; \mathbf{x} = \mathbf{0}.$$

2. **Proper formulation:** Mixed sparsity minimization (MSM) problem

$$(MP_{0,p}): \quad \{\mathbf{x}^*_{0,p}, \mathbf{e}^*_0\} = \operatorname*{argmin}_{(\mathbf{x},\mathbf{e})} \ell_{0,p}(\mathbf{x}) + \gamma\|\mathbf{e}\|_0, \quad \text{subj. to} \quad \begin{bmatrix} A_1 & \cdots & A_K \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_K \end{bmatrix} = \mathbf{b} + \mathbf{e}$$

Robust Face Recognition as a Group Sparsity Recovery Problem

Can we frame robust face recognition using group sparsity?

1. **Negative:**

$$\mathbf{b} = [A_1, A_2, \cdots, A_K, I][\mathbf{x}_1; \mathbf{x}_2; \cdots; \mathbf{x}_K; \mathbf{e}]$$

Standard group sparsity formulation, whereby $\mathbf{e}$ is treated as the $(K+1)$th group, has a trivial solution of 1-group-sparsity:

$$\mathbf{e} = \mathbf{b}; \mathbf{x} = \mathbf{0}.$$

2. **Proper formulation:** Mixed sparsity minimization (MSM) problem

$$(MP_{0,p}): \quad \{\mathbf{x}^*_{0,p}, \mathbf{e}^*_0\} = \operatorname*{argmin}_{(\mathbf{x},\mathbf{e})} \ell_{0,p}(\mathbf{x}) + \gamma\|\mathbf{e}\|_0, \quad \text{subj. to} \quad \begin{bmatrix} A_1 & \cdots & A_K \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_K \end{bmatrix} = \mathbf{b} + \mathbf{e}$$

3. **How to convexify the NP-Hard problem?**
   - Out of all $p \geq 1$, which value should we choose to convexify (mixed) group sparsity problems?
   - Does this choice make any difference in accuracy and speed?

Convexify Mixed Sparsity Minimization via Lagrange Biduality

- Consider a generalization of entry-wise sparsity and group sparsity:

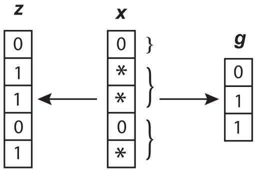
$$\mathbf{x}^* = \operatorname*{argmin}_{\mathbf{x}} \sum_{k=1}^{K} \left[\alpha_k \mathcal{I}(\|\mathbf{x}_k\|_p > 0) + \beta_k \|\mathbf{x}_k\|_0\right], \quad \text{subj. to} \quad \begin{bmatrix} A_1 & \cdots & A_K \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_K \end{bmatrix} = \mathbf{b}$$

- Regularize $\|\mathbf{x}^*\|_\infty \leq M$

$$\mathbf{x}^*_{\text{primal}} = \operatorname*{argmin}_{\mathbf{x}} \sum_{k=1}^{K} \left[\alpha_k \mathcal{I}(\|\mathbf{x}_k\|_p > 0) + \beta_k \|\mathbf{x}_k\|_0\right], \quad \text{subj. to} \quad A\mathbf{x} = \mathbf{b} \text{ and } \|\mathbf{x}\|_\infty \leq M$$

A Mixed Integer Program

- Introduce two sparsity indicator variables:
  $\mathbf{z} \in \{0,1\}^n$ for entry-wise sparisty; $\mathbf{g} \in \{0,1\}^K$ for group sparsity



- The primal problem becomes a **mixed integer program:**

$$\{\mathbf{x}_+^*, \mathbf{x}_-^*, \mathbf{z}^*, \mathbf{g}^*\} = \underset{\{\mathbf{x}_+ \geq 0, \mathbf{x}_- \geq 0, \mathbf{z}, \mathbf{g}\}}{\operatorname{argmin}} \left(\boldsymbol{\alpha}^T \mathbf{g} + \boldsymbol{\beta}^T \mathbf{z}\right) \quad \text{subj. to}$$

$$A(\mathbf{x}_+ - \mathbf{x}_-) = \mathbf{b}, \Pi\mathbf{g} \geq \frac{1}{M}(\mathbf{x}_+ + \mathbf{x}_-), \mathbf{z} \geq \frac{1}{M}(\mathbf{x}_+ + \mathbf{x}_-)$$

where $\Pi \in \{0,1\}^{n \times K}$ is a membership matrix for the group sparsity:

$$\Pi = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix} \Rightarrow \Pi g = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \tag{2}$$

Sketch of the Biduality Approach

Primal (NP-Hard)

$$\underset{\{\mathbf{x}_+\geq 0, \mathbf{x}_-\geq 0, \mathbf{z}, \mathbf{g}\}}{\operatorname{argmin}} (\boldsymbol{\alpha}^T \mathbf{g} + \boldsymbol{\beta}^T \mathbf{z})$$
$$\mathbf{g} \in \{0,1\}^K, \mathbf{z} \in \{0,1\}^n,$$
$$A(\mathbf{x}_+ - \mathbf{x}_-) = \mathbf{b}$$
$$\Pi \mathbf{g} \geq \frac{1}{M}(\mathbf{x}_+ + \mathbf{x}_-)$$
$$\mathbf{z} \geq \frac{1}{M}(\mathbf{x}_+ + \mathbf{x}_-)$$

$\Rightarrow$

Lagrangian Dual

Concave and LP

$\Leftrightarrow$

Bidual (Convex and LP)

$$\underset{\{\mathbf{x}_+\geq 0, \mathbf{x}_-\geq 0, \mathbf{z}, \mathbf{g}\}}{\operatorname{argmin}} (\boldsymbol{\alpha}^T \mathbf{g} + \boldsymbol{\beta}^T \mathbf{z})$$
$$\mathbf{g} \in [0,1]^K, \mathbf{z} \in [0,1]^n,$$
$$A(\mathbf{x}_+ - \mathbf{x}_-) = \mathbf{b}$$
$$\Pi \mathbf{g} \geq \frac{1}{M}(\mathbf{x}_+ + \mathbf{x}_-)$$
$$\mathbf{z} \geq \frac{1}{M}(\mathbf{x}_+ + \mathbf{x}_-)$$

Sketch of the Biduality Approach

Primal (NP-Hard)

$$\begin{gathered}\underset{\{\mathbf{x}_+\geq 0, \mathbf{x}_-\geq 0, \mathbf{z}, \mathbf{g}\}}{\operatorname{argmin}} (\boldsymbol{\alpha}^T\mathbf{g} + \boldsymbol{\beta}^T\mathbf{z})\\ \mathbf{g}\in\{0,1\}^K, \mathbf{z}\in\{0,1\}^n,\\ A(\mathbf{x}_+ - \mathbf{x}_-) = \mathbf{b}\\ \Pi\mathbf{g} \geq \tfrac{1}{M}(\mathbf{x}_+ + \mathbf{x}_-)\\ \mathbf{z} \geq \tfrac{1}{M}(\mathbf{x}_+ + \mathbf{x}_-)\end{gathered}$$

$\Rightarrow$

Lagrangian Dual

Concave and LP

$\Leftrightarrow$

Bidual (Convex and LP)

$$\begin{gathered}\underset{\{\mathbf{x}_+\geq 0, \mathbf{x}_-\geq 0, \mathbf{z}, \mathbf{g}\}}{\operatorname{argmin}} (\boldsymbol{\alpha}^T\mathbf{g} + \boldsymbol{\beta}^T\mathbf{z})\\ \mathbf{g}\in[0,1]^K, \mathbf{z}\in[0,1]^n,\\ A(\mathbf{x}_+ - \mathbf{x}_-) = \mathbf{b}\\ \Pi\mathbf{g} \geq \tfrac{1}{M}(\mathbf{x}_+ + \mathbf{x}_-)\\ \mathbf{z} \geq \tfrac{1}{M}(\mathbf{x}_+ + \mathbf{x}_-)\end{gathered}$$

Lagrangian Bidual of Mixed Sparsity Minimization Problem

$$\mathbf{x}^*_{\text{bidual}} = \underset{\mathbf{x}}{\operatorname{argmin}} \frac{1}{M}\sum_{k=1}^{K}(\alpha_k\|\mathbf{x}_k\|_\infty + \beta_k\|\mathbf{x}_k\|_1) \quad \text{subj. to (a) } A\mathbf{x} = \mathbf{b} \text{ and (b) } \|\mathbf{x}\|_\infty \leq M.$$

- ℓ_∞-norm promotes dense signal within the groups; while ℓ_1-norm promotes sparsity.
- Biduality approach provides a rigorous and operable method to convexify an NP-hard problem.

Corollary I: Bidual of Group Sparsity

Lagrangian Bidual

$$\mathbf{x}^*_{\text{bidual}} = \operatorname*{argmin}_{\mathbf{x}} \frac{1}{M} \sum_{k=1}^{K} [\alpha_k \|\mathbf{x}_k\|_\infty + \beta_k \|\mathbf{x}_k\|_1] \quad \text{subj. to (a) } A\mathbf{x} = \mathbf{b} \text{ and (b) } \|\mathbf{x}\|_\infty \leq M.$$

- Let $\boldsymbol{\alpha} = \mathbf{1}$ and $\boldsymbol{\beta} = \mathbf{0}$, then with a conservative M, the bidual of $(P_{0,p})$ is

$$(P_{1,\infty}): \quad \mathbf{x}^*_{1,\infty} = \operatorname*{argmin}_{\mathbf{x}} \sum_{k=1}^{K} \|\mathbf{x}_k\|_\infty \quad \text{subj. to} \quad A\mathbf{x} = \mathbf{b} \tag{3}$$

Corollary I: Bidual of Group Sparsity

Lagrangian Bidual

$$\mathbf{x}^*_{\text{bidual}} = \operatorname*{argmin}_{\mathbf{x}} \frac{1}{M} \sum_{k=1}^{K} \left[\alpha_k \|\mathbf{x}_k\|_\infty + \beta_k \|\mathbf{x}_k\|_1\right] \quad \text{subj. to (a) } A\mathbf{x} = \mathbf{b} \text{ and (b) } \|\mathbf{x}\|_\infty \leq M.$$

- Let $\boldsymbol{\alpha} = \mathbf{1}$ and $\boldsymbol{\beta} = \mathbf{0}$, then with a conservative M, the bidual of $(P_{0,p})$ is

$$(P_{1,\infty}): \quad \mathbf{x}^*_{1,\infty} = \operatorname*{argmin}_{\mathbf{x}} \sum_{k=1}^{K} \|\mathbf{x}_k\|_\infty \quad \text{subj. to} \quad A\mathbf{x} = \mathbf{b} \tag{3}$$

- Multiple Measurement Vector (MMV) problem [Eldar & Mishali '09]

$$Y = AX \Leftrightarrow \text{vec}(Y^T) = (A \otimes I)\text{vec}(X^T)$$

Bidual of group sparsity leads to the same **MMV convex relaxation** [Tropp '06]:

$$(MMV_0): \quad \min \|X\|_{\text{row}-0} \quad \text{subj. to} \quad Y = AX$$
$$(MMV_{1,\infty}): \quad \min \textstyle\sum_i \max_j |x_{i,j}| \quad \text{subj. to} \quad Y = AX$$

Corollary II: Bidual of Sparsity-based Classification

1. For robust face recognition, we consider MSM

$$(MP_{0,p}): \quad \{\mathbf{x}^*_{0,p}, \mathbf{e}^*_0\} = \underset{(\mathbf{x},\mathbf{e})}{\text{argmin}}\ \ell_{0,p}(\mathbf{x}) + \gamma\|\mathbf{e}\|_0, \quad \text{subj. to} \quad [A_1 \quad \cdots \quad A_K] \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_K \end{bmatrix} + \mathbf{e} = \mathbf{b}$$

Its bidual is

$$\{\mathbf{x}^*_{1,\infty}, \mathbf{e}^*_1\} = \underset{\{\mathbf{x},\mathbf{e}\}}{\text{argmin}} \sum_{k=1}^{K} \|\mathbf{x}_k\|_\infty + \gamma\|\mathbf{e}\|_1 \quad \text{subj. to} \quad A\mathbf{x} + \mathbf{e} = \mathbf{b}.$$

2. Numerical implementation

- As an LP, available standard packages include CVX and MOSEK.
- Specialized toolboxes exist: TFOCS and iCAP.
- Other accelerated linear programming algorithms: ALM.

Corollary II: Bidual of Sparsity-based Classification

1. For robust face recognition, we consider MSM

$$(MP_{0,p}): \quad \{\mathbf{x}^*_{0,p}, \mathbf{e}^*_0\} = \underset{(\mathbf{x},\mathbf{e})}{\operatorname{argmin}}\ \ell_{0,p}(\mathbf{x}) + \gamma\|\mathbf{e}\|_0, \quad \text{subj. to} \quad \begin{bmatrix} A_1 & \cdots & A_K \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_K \end{bmatrix} + \mathbf{e} = \mathbf{b}$$

   Its bidual is

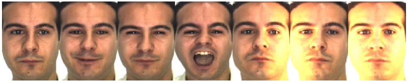
$$\{\mathbf{x}^*_{1,\infty}, \mathbf{e}^*_1\} = \underset{\{\mathbf{x},\mathbf{e}\}}{\operatorname{argmin}} \sum_{k=1}^{K} \|\mathbf{x}_k\|_\infty + \gamma\|\mathbf{e}\|_1 \quad \text{subj. to} \quad A\mathbf{x} + \mathbf{e} = \mathbf{b}.$$

2. Numerical implementation

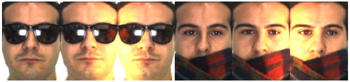
- As an LP, available standard packages include CVX and MOSEK.
- Specialized toolboxes exist: TFOCS and iCAP.
- Other accelerated linear programming algorithms: ALM.

Question: Does the biduality result lead to improved classification in face recognition?

Face Recognition Performance via MOSEK



(a) Unoccluded Images



(b) Occluded Images

Figure: Images from one session of the AR database.

Group Sparsity	ℓ_1	$\ell_{1,2}$	$\ell_{1,\infty}$
unoccluded	92%	93.6%	94.7%
occluded	49.7%	53.6%	57.6%
Total	65.3%	68.3%	69.7%
Speed	53.7s	256.5s	60.9s

Table: 100-subject test set consists of 700 un-occluded images and 1200 occluded images.

Reference:

AYY, et al. *On the Lagrangian biduality of sparsity minimization problems*. UCB Tech Report, 2011.

Capability to implement SRC on parallel computing environments

1. Face Recognition Module [Wright et al. '09]

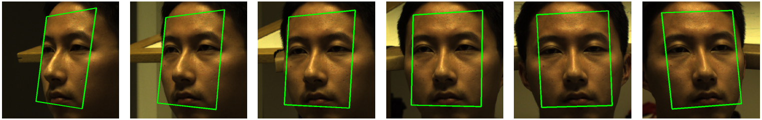
$$\min_{\mathbf{x},\mathbf{e}} \|\mathbf{x}\|_1 + \|\mathbf{e}\|_1 \quad \text{subj. to} \quad \mathbf{b} = A\mathbf{x} + \mathbf{e}. \tag{4}$$

Capability to implement SRC on parallel computing environments

1. Face Recognition Module [Wright et al. '09]

$$\min_{\mathbf{x},\mathbf{e}} \|\mathbf{x}\|_1 + \|\mathbf{e}\|_1 \quad \text{subj. to} \quad \mathbf{b} = A\mathbf{x} + \mathbf{e}. \tag{4}$$

2. Face Alignment Module [Wagner et al. '11]



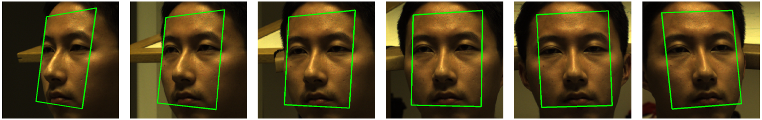
$$\hat{\tau}_i = \arg\min_{\mathbf{x},\mathbf{e},\tau_i} \|\mathbf{e}\|_1 \quad \text{subj. to} \quad \mathbf{b} \circ \tau_i = A_i\mathbf{x} + \mathbf{e}, \tag{5}$$

Capability to implement SRC on parallel computing environments

1. Face Recognition Module [Wright et al. '09]

$$\min_{\mathbf{x},\mathbf{e}} \|\mathbf{x}\|_1 + \|\mathbf{e}\|_1 \quad \text{subj. to} \quad \mathbf{b} = A\mathbf{x} + \mathbf{e}. \tag{4}$$

2. Face Alignment Module [Wagner et al. '11]



$$\hat{\tau}_i = \arg\min_{\mathbf{x},\mathbf{e},\tau_i} \|\mathbf{e}\|_1 \quad \text{subj. to} \quad \mathbf{b} \circ \tau_i = A_i\mathbf{x} + \mathbf{e}, \tag{5}$$

- **Local linearization**

$$\min_{\mathbf{x},\mathbf{e},\Delta\tau_j} \|\mathbf{e}\|_1 \quad \text{subj. to} \quad \mathbf{b} \circ \tau_j + J_j \Delta\tau = A_i\mathbf{x} + \mathbf{e}, \tag{6}$$

  where $J_j \doteq \nabla_{\tau_j}(\mathbf{b} \circ \tau_j)$ is the Jacobian, and $\Delta\tau$ is an iterative update to τ.
- **Per-class alignment** is equivalent to iteratively solving a linear program:

$$\min_{\mathbf{w},\mathbf{e}} \|\mathbf{e}\|_1 \quad \text{subj. to} \quad \mathbf{b}_j = [A_i, -J_j]\mathbf{w} + \mathbf{e}.$$

Demo: Misalignment & Corruption Correction

Alignment Demo

Choice of ℓ_1-Min Algorithm for Parallelization

1. **Primal-Dual Interior-Point**
   - Log-Barrier [Frisch '55, Karmarkar '84, Megiddo '89, Monteiro-Adler '89, Kojima-Megiddo-Mizuno '93]
2. **Homotopy**
   - Homotopy [Osborne-Presnell-Turlach '00, Malioutov-Cetin-Willsky '05, Donoho-Tsaig '06]
   - Polytope Faces Pursuit (PFP) [Plumbley '06]
   - Least Angle Regression (LARS) [Efron-Hastie-Johnstone-Tibshirani '04]
3. **Gradient Projection**
   - Gradient Projection Sparse Representation (GPSR) [Figueiredo-Nowak-Wright '07]
   - Truncated Newton Interior-Point Method (TNIPM) [Kim-Koh-Lustig-Boyd-Gorinevsky '07]
4. **Iterative Thresholding**
   - Soft Thresholding [Donoho '95]
   - Sparse Reconstruction by Separable Approximation (SpaRSA) [Wright-Nowak-Figueiredo '08]
5. **Proximal Gradient** [Nesterov '83, Nesterov '07]
   - FISTA [Beck-Teboulle '09]
   - Nesterov's Method (NESTA) [Becker-Bobin-Candés '09]
6. **Augmented Lagrangian Methods** [Yang-Zhang '09, AY et al '10]
   - Bergman [Yin et al. '08]
   - YALL1 [Yang-Zhang '09]
   - SALSA [Figueiredo et al. '09]
   - Primal ALM, Dual ALM [AY et al '10]

Two Choices for Parallelization

1. Multicore CPU (dual quad-core Intel E5530 processor)
   - CPU Speed: 2.4 GHz
   - Caches on dual CPU

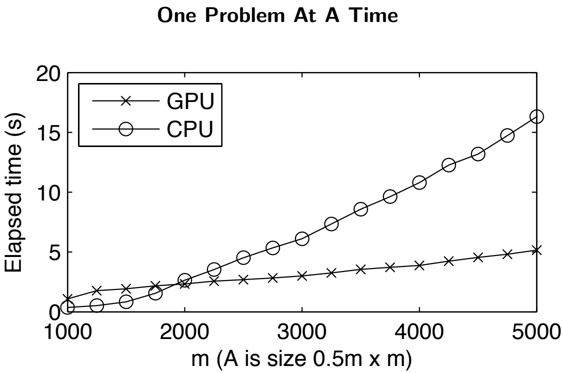
32 KiB L1	32 KiB L1	32 KiB L1	32 KiB L1	32 KiB L1	32 KiB L1	32 KiB L1	32 KiB L1
256 KiB L2	256 KiB L2	256 KiB L2	256 KiB L2	256 KiB L2	256 KiB L2	256 KiB L2	256 KiB L2
8 MiB L3 (Too big to show)				8 MiB L3 (Too big to show)			

   - Memory bandwidth: 25.6 GB/sec
2. Multicore GPU (single Nvidia GTX 480 with 14 SMPs)
   - GPU Speed: 1.4 GHz
   - Caches on a GTX 480

64 KiB L1 + shared	64 KiB L1 + shared	64 KiB L1 + shared	64 KiB L1 + shared	64 KiB L1 + shared	64 KiB L1 + shared	64 KiB L1 + shared	64 KiB L1 + shared	64 KiB L1 + shared	64 KiB L1 + shared	64 KiB L1 + shared	64 KiB L1 + shared	64 KiB L1 + shared	64 KiB L1 + shared	64 KiB L1 + shared
768 KiB L2														

   - Memory bandwidth: 177.4 GB/sec
3. Problem size: 20 images per subject class occupy 384 KiB.

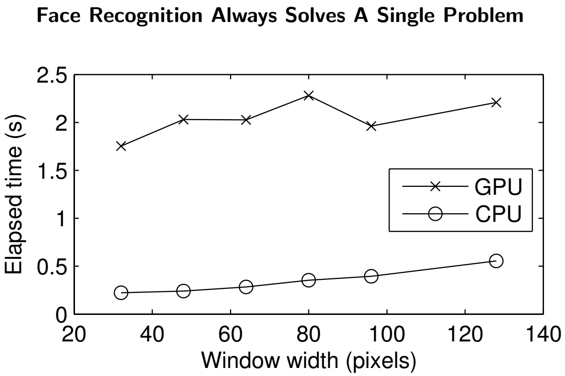
ℓ_1-Min Simulation: Algorithm-Level Parallelism



Trade-off on Random Data

- **CPU:** outperforms on small problems (faster processor speed).
- **GPU:** outperforms on large problems (larger memory bandwidth).

Recognition Module Benchmark on a 10-Subject Training Set

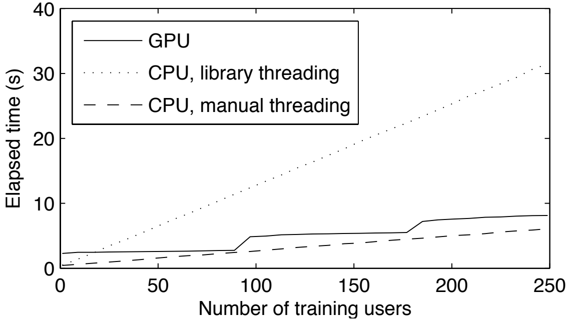


Speed vs Resolution

- With a small data set, CPU outperforms GPU by a wide margin ($4\times$).
- New ALM C implementation accommodates much higher image resolutions **in real time**.

Alignment Module Benchmark: System-Level Parallelism

Batch Parallel Process As Many Alignments As Possible



Speed of Alignment: Each class contains 20 training images at 64 × 64 resolution

- **Sequential:** 600 ms per subject.
- **Parallel:** 40 ms per subject.

Conclusion

Collaborators

- **Berkeley**: Dr. Shankar Sastry, Victor Shia, Dr. Dheeraj Singaraju
- **UIUC**: Dr. Yi Ma, Arvind Ganesh, Zihan Zhou
- **Columbia**: Dr. John Wright

Publications

- Wright, Yang, Ganesh, Sastry, Ma. "Robust face recognition via sparse representation." IEEE PAMI, 2009.
- Wagner, Wright, Ganesh, Zhou, Mobahi, Ma. "Towards a practical face recognition system: robust alignment and illumination by sparse representation." IEEE PAMI, 2011.
- Yang, Ganesh, Zhou, Sastry, Ma. `http://www.eecs.berkeley.edu/~yang/software/l1benchmark/`.
- Singaragu, Elhamifar, Tron, Yang, Sastry. "On the Lagrangian biduality of sparsity minimization problems." UCB Tech Report, 2011.