Dictionary Learning
by $\ell^1$-Minimization

John Wright
Electrical Engineering
Columbia University
Sparse Approximation

Model:

\[ y = A x_0 \]

with \( x_0 \in \mathbb{R}^n \) **sparse** - most of the \( x_0(i) \) are zero.

Good model for many types of imagery data, especially if we can **learn the dictionary** \( A = [a_1 \mid \cdots \mid a_n] \in \mathbb{R}^{m \times n} \):
Motivating Applications (biased sample)

Single image superresolution:

\[ y = Dy_0 \]

reconstruct the original high-resolution image \( y_0 \):

\[ \hat{x} \in \arg \min \|x\|_1 \quad \text{s.t.} \quad y = DA\hat{x} \quad y_0 \approx A\hat{x} \]

Yang, W., Huang and Ma, TIP '10
Motivating Applications (biased sample)

High-resolution hyperspectral imaging for cultural heritage:

Ultra high-res RGB camera
Moshe Ben-Ezra
Microsoft Research

Buddhist Frescos
Dunhuang, China

Can dictionary learning help overcome hardware limitations?

Kawakami, W., Tai, Ikiuchi, Matsushita, Ben-Ezra, CVPR ‘11
When do dictionary learning algorithms succeed?

Huan Wang (Yale)  Quan Geng (UIUC)  Dan Spielman (Yale)
The model problem

Given $Y \approx AX$ with $x_j$ sparse, $(A, X)$ unknown, recover $A$ and $X$.

Ambiguities: $(A, X)$ or $(A \Pi \Lambda, \Lambda^{-1} \Pi^* X)$ ?

Peculiar geometry:

k column subspaces of $A$
When is dictionary learning well-posed?

\( \binom{n}{k} \) k column subspaces of \( \mathbf{A} \)

\( \mathbf{Y} \) k+1 points per subspace

Solution is unique:

**Theorem 1 (ess. Aharon et. al. ’05) (sketch)** There exists k column sparse \( \mathbf{X} = [\mathbf{x}_1 \ldots \mathbf{x}_p] \), of size \( p = (k+1)\binom{n}{k} \) such that if we observe \( \mathbf{Y} = \mathbf{AX} \), (\( \mathbf{A}, \mathbf{X} \)) is essentially the only k-column sparse factorization of \( \mathbf{Y} \).
When does a learned dictionary generalize?

**Theorem 2** (Vainsencher, Mannor and Bruckstein ’11) *(sketch)* If \( y \sim_{\text{iid}} \mu \) on \( \mathbb{S}^{m-1} \), \( p > p_0 \), \( \lambda > \lambda_0 \), then with prob. \( 1 - e^{-t} \) in \( Y \),

\[
\mathbb{E}_y \min_{\|x\|_1 \leq \lambda} \|y - Ax\| \\
\leq \frac{1.1}{p} \sum_i \|y_i - Ax_i\| + \frac{9mn \log(\lambda p) + t}{p}
\]

See also [Maurer and Pontil ’10].
How can we learn a good dictionary?

$Y \approx AX$, $X$ sparse.

**Alternating directions to minimize sparsity surrogate**

[Engan et. al., ‘99, Aharon et. al. ’05, Yaghoobi ‘10]

$$\min \frac{1}{2} \| Y - AX \|_F^2 + J(X)$$

**Recently:** Supervised variants [Mairal et. al. ‘08], structured dictionaries [Rubenstein et. al. ‘10], highly scalable variants [Mairal et. al. ‘10] … and many, many more…
Is the desired solution a local minimum?

\[ Y = AX, \ X \text{ sparse.} \]

\[
\min \|X'\|_1 \quad \text{s.t.} \quad Y = A'X', \ A' \in A
\]

For square \( A \), under probabilistic assumptions on \( X \), \((A, X)\) is a local minimum whp:

**Theorem 3 (Gribonval + Schnass ’10) (sketch)** Let \( X_{ij} = \Omega_{ij}V_{ij} \), with \( \Omega \sim \text{Ber}(\theta), V \sim \mathcal{N}(0,1) \). For square, incoherent \( A \), \((A, X)\) is a local minimum of \( \| \cdot \|_1 \) with high probability, provided \( p = \Omega(n \log n/\theta) \).
Is the desired solution a local minimum?

\[
\min \quad \|X'\|_1 \quad \text{s.t.} \quad Y = A'X', \quad A' \in \mathcal{A}
\]

For general \( \mathcal{A} \), under probabilistic assumptions on \( X \),
\((A, X)\) is a \textbf{local minimum whp}:

**Theorem 4** (Geng, W., ’11). Let \( A \in \mathbb{R}^{m \times n} \), \( k < C/\mu(A) \), and \( X \in \mathbb{R}^{n \times p} \) with random \( k \)-sparse support, independent Gaussian nonzeros. Then \((A, X)\) is a local minimum of the \( \ell^1 \)-norm wp \( \geq 1 - \tilde{O}(n^{3/2}k^{1/2}p^{-1/2}) \).
Is this obvious?

Maybe … but surprisingly resistant to analysis …

\[
\begin{align*}
\min & \quad \|X'\|_1 \\
\text{s.t.} & \quad Y = A'X', \quad A' \in \mathcal{A}
\end{align*}
\]
Is this obvious?
Maybe … but surprisingly resistant to analysis …

$$\min \quad \|X'\|_1 \quad \text{s.t.} \quad Y = A'X', \quad A' \in \mathcal{A}$$

Feasible \((A, X)\)
Is this obvious?

Maybe … but surprisingly resistant to analysis …

$$\min \|X'\|_1 \quad \text{s.t.} \quad Y = A'X', \ A' \in \mathcal{A}$$

Have to analyze an $\ell^1$ problem over an affine space.

RIP ect., fail here ess. sign-permutation ambiguity

Feasible $(A, X)$

Use ideas from low-rank recovery [Gross ‘09], [Candes, Li, Ma, W. ’12].
Uniqueness – square dictionaries

Rows of $X$ are sparse vectors in a known subspace.

If $p > cn \log n$, then whp. rows of $X$ are the sparsest vectors in $\text{row}(Y)$:

\[ e_1^* X = e_2^* X \]

\[ m = n \]
Uniqueness – square dictionaries

**Square:**

\[ \text{row}(Y) \]

\[ e_1^* X \]

\[ e_2^* X \]

\[ m = n \]

**Theorem [Spielman, Wang, W. ‘11]:**

Decomposition essentially unique from \( \Omega(n \log n) \) random observations.

**Overcomplete:**

\[ \text{row}(Y) \]

\[ e_1^* X \]

\[ e_2^* X \]

\[ m < n \]

**Theorem [Aharon, Elad, Bruckstein ‘05]:**

Decomposition is essentially unique from \((k + 1)\binom{n}{k}\) strategically located observations.
Rows of $X$ are sparsest vectors in $\text{row}(Y)$. 

minimize $\|w^*Y\|_0$ subject to $w \neq 0$. 

$e_1^*X \quad m = n \quad e_2^*X \quad \text{row}(Y)$
Algorithms – square dictionaries

Rows of $X$ are sparsest vectors in $\text{row}(Y)$.

\[
\minimize \| w^* Y \|_0 \quad \text{subject to} \quad w \neq 0.
\]

\[
\minimize \| w^* Y \|_1 \quad \text{subject to} \quad r^* w = 1.
\]
minimize $\|w^*Y\|_1$ subject to $r^*w = 1$.

What choice of $r$ will make $\hat{w}^*Y = e_i^*X$?

Change variables $q = A^*w$:

minimize $\|q^*X\|_1$ subject to $(A^{-1}r)^*q = 1$.

If $r = Ae_i$, we’re golden …

Don’t have this; use $y_j = \sum_{i \in I} X_{ij} Ae_i$. 

Algorithms – square dictionaries
ER-SpUD(SC): Exact Recovery of Sparsely-Used Dictionaries using single columns of $Y$ as constraint vectors.

For $j = 1 \ldots p$

Solve $\min_w \|w^T Y\|_1$ subject to $(Ye_j)^T w = 1$, and set $s_j = w^T Y$.

Greedy: A Greedy Algorithm to Reconstruct $X$ and $A$.

1. **REQUIRE:** $S = \{s_1, \ldots, s_T\} \subset \mathbb{R}^p$.
2. For $i = 1 \ldots n$
   
   **REPEAT**
   
   $l \leftarrow \arg\min_{s_l \in S} \|s_l\|_0$, breaking ties arbitrarily
   
   $x_i = s_l$
   
   $S = S \backslash \{s_l\}$
   
   **UNTIL** $\text{rank}([x_1, \ldots, x_i]) = i$

3. Set $X = [x_1, \ldots, x_n]^T$, and $A = YY^T(XY^T)^{-1}$. 

Algorithms – square dictionaries
ER-SpUD(DC): Exact Recovery of Sparsely-Used Dictionaries using the sum of two columns of $Y$ as constraint vectors.

1. Randomly pair columns of $Y$ into $p/2$ groups $g_i = \{Ye_{i1}, Ye_{i2}\}$.
2. For $j = 1 \ldots p/2$
   
   Let $r_j = Ye_{j1} + Ye_{j2}$, where $Ye_{j1}, Ye_{j2} \in g_j$.

   Solve $\min_w \|w^T Y\|_1$ subject to $r_j^T w = 1$, and set $s_j = w^T Y$.

Greedy: A Greedy Algorithm to Reconstruct $X$ and $A$.

1. REQUIRE: $S = \{s_1, \ldots, s_T\} \subset \mathbb{R}^p$.
2. For $i = 1 \ldots n$

   REPEAT
   
   $l \leftarrow \arg \min_{s_l \in S} \|s_l\|_0$, breaking ties arbitrarily
   
   $x_i = s_l$
   
   $S = S \setminus \{s_l\}$
   
   UNTIL $\text{rank}([x_1, \ldots, x_i]) = i$

3. Set $X = [x_1, \ldots, x_n]^T$, and $A = YY^T(XY^T)^{-1}$. 
Recovery guarantee – square dictionaries

If the expected nonzeros per column is smaller than $\sqrt{n}$ the algorithm succeeds whp:

**Theorem 5 (Spielman, Wang, W. ’12) (sketch)** Let $X$ Bernoulli($\theta$)–Rademacher or Bernoulli($\theta$) – Gaussian. If $n > n_0$, $p > c_p n^2 \log^2 n$, and the nonzero probability satisfies

$$\frac{2}{n} \leq \theta \leq \frac{c}{\sqrt{n}},$$

with high probability $\text{ER-SpUD (DC)}$ recovers all $n$ rows of $X$.

**Sample requirement** $p > cn^2 \log^2 n$. 
Does it really work?

Caveat: exact sparse, noiseless setting.
Good news / bad news ...

If the expected nonzeros per column exceeds $\sqrt{n \log n}$ the algorithm fails whp:

**Theorem 6** (Spielman, Wang, W. ’12) *(sketch)* If $n$ large, $p \geq cn$, and the nonzero probability $\theta$ satisfies

$$\theta \geq \sqrt{\frac{\log n}{n}}, \quad (1)$$

then the probability *(in $X$)* that the algorithm correctly recovers one of the rows is at most $n^{-c}$.

**Theory is almost tight** in the sparsity level.

For denser $X$, think about **different constraints**.
Summary and open questions

Two main mathematical results:

- **Local** recovery in the rectangular case
- **Exact** (global) recovery in the square case

Many open questions:

- Past the $\sqrt{n}$ barrier?
- Noise tolerance, multiple vectors?
- Other coefficient structures?
Dictionary Learning by $\ell^1$ Minimization

Thanks to ...

Huan Wang (Yale)  Quan Geng (UIUC)  Dan Spielman (Yale)

Local correctness of $\ell^1$-minimization for dictionary learning, Geng, W., Arxiv
Exact recovery of sparse dictionaries, Spielman, Wang, W., COLT ’12.