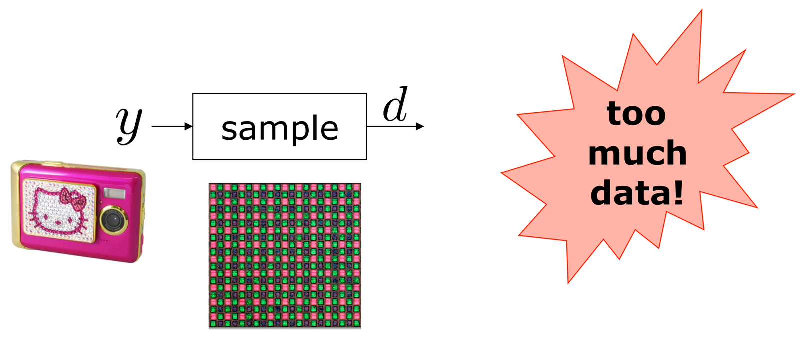
rsrg@caltech
..where theory and practice collide

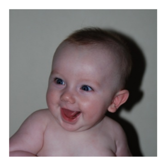
Submodular Dictionary Selection for Sparse Representation

Andreas Krause (Caltech), Volkan Cevher (EPFL, Rice)

Sensing by *Sampling*

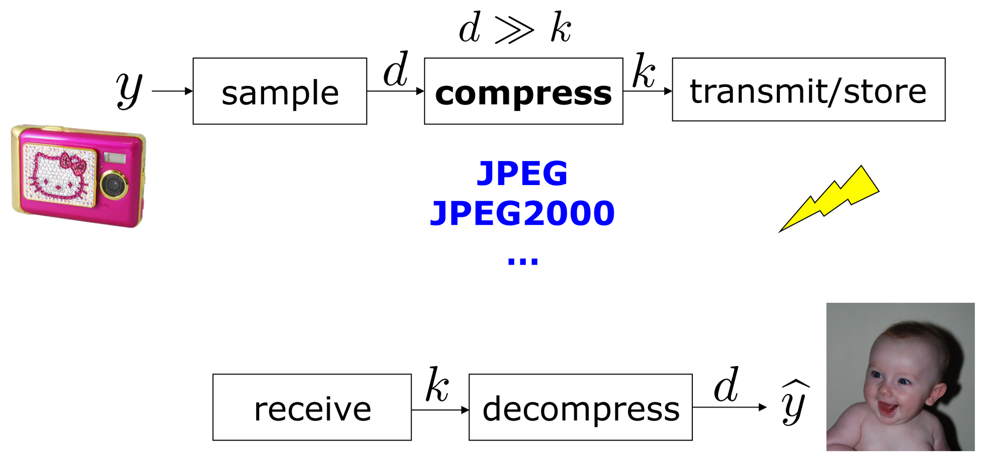
- Long-established paradigm for digital data acquisition
  - uniformly ***sample*** data at Nyquist rate (2x Fourier bandwidth)



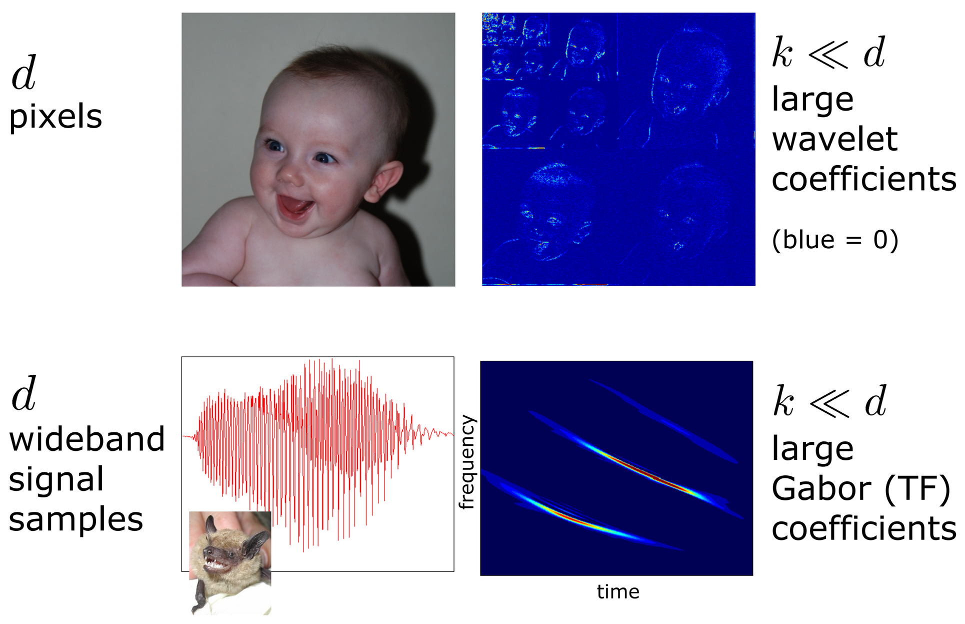


Sensing by *Sampling*

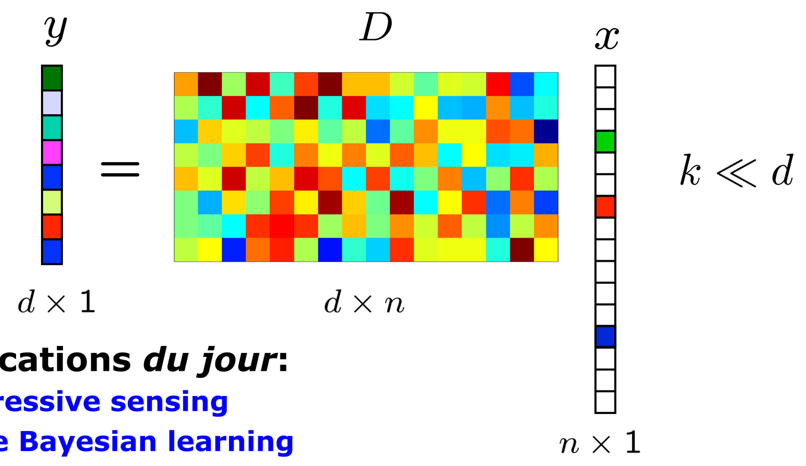
- Long-established paradigm for digital data acquisition
  - uniformly ***sample*** data at Nyquist rate (2x Fourier bandwidth)
  - ***compress*** data



Sparsity / Compressibility



Sparse Representations



Applications *du jour*:
Compressive sensing
Sparse Bayesian learning
Inpainting, denoising, …
Data streaming
Information theory, theoretical computer science, …

Key question: Which dictionary D should we use?

Existing Solutions

- **Dictionary *design***
  - functional space assumptions <> Besov, Sobolev, Triebel…
    ex. natural images <> smooth regions + edges
  - induced norms/designed bases & frames
    ex. wavelets, curvelets, etc.

- **Dictionary *learning***
  - regularization $D^* = \arg\min_{D, x_1, \ldots, x_M} \sum_i \{\|y_i - Dx_i\|_2^2 + \lambda \|x_i\|_{1,\mathrm{TV}}\}$
  - clustering <> identify clustering of data (k-SVD)
  - Bayesian <> non-parametric approaches
    Indian buffet processes

Dictionary *Selection* Problem—*DiSP* NEW!

- **Given**
  - candidate columns $\mathcal{V} = \{\phi_1, \ldots \phi_N\}, \phi_i \in \mathrm{R}^d$
  - sparsity and dictionary size $k \& n$
  - training data $\mathcal{Y} = \{y_1, \ldots, y_M\} \in \mathrm{R}^{d \times M}$
- **Choose D to maximize variance reduction:**
  - reconstruction accuracy $L_s(\mathcal{A}) = \min_x \|y_s - \Phi_{|\mathcal{A}} x\|_2^2$
  - var. reduct. for s-th data: $F_s(\mathcal{D}) = L_s(\emptyset) - \min_{\mathcal{A} \subseteq \mathcal{D}, |\mathcal{A}| \le k} L_s(\mathcal{A})$
  - overall var. reduction: $F(\mathcal{D}) = \frac{1}{m} \sum_s F_s(\mathcal{D})$

Want to solve: $\mathcal{D}^* = \arg\max_{|\mathcal{D}| \le n} F(\mathcal{D})$

Combines Dictionary Design and Learning

Combinatorial Challenges in DiSP

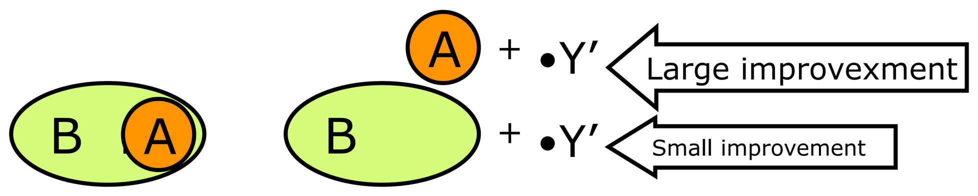
1. Evaluation of $F_s(\mathcal{D}) = L_s(\emptyset) - \min_{\mathcal{A} \subseteq \mathcal{D}, |\mathcal{A}| \le k} L_s(\mathcal{A})$
   $L_s(\mathcal{A}) = \min_x \|y_s - \Phi_{|\mathcal{A}} x\|_2^2$
   Sparse reconstruction problem!
2. Finding D* (NP-hard)
   $\mathcal{D}^* = \arg\max_{|\mathcal{D}| \le n} F(\mathcal{D})$ $F(\mathcal{D}) = \frac{1}{m} \sum_s F_s(\mathcal{D})$

Key observations:

- F(D) <> **approximately submodular**
- submodularity <> **efficient algorithms with provable guarantees**

(Approximate) Submodularity

- Set function F *submodular*, if $\mathcal{D} \subseteq \mathcal{D}' \subseteq \mathcal{V}$ and $v \in \mathcal{V} \setminus \mathcal{D}'$ (*)
  $$F(\mathcal{D} \cup \{v\}) - F(\mathcal{D}) \ge F(\mathcal{D}' \cup \{v\}) - F(\mathcal{D}')$$



- Set function F *approximately submodular*, if
  $$F(\mathcal{D} \cup \{v\}) - F(\mathcal{D}) \ge F(\mathcal{D}' \cup \{v\}) - F(\mathcal{D}') - \varepsilon$$

A Greedy Algorithm

Greedy algorithm: $\mathcal{A}^* = \arg\max_{\mathcal{A}} F(\mathcal{A})$

- Start with $\mathcal{A} \leftarrow \emptyset$
- For i = 1:k do
  - Choose $y^* = \arg\max_y F(\mathcal{A} \cup \{y\}) - F(\mathcal{A})$
  - Set $\mathcal{A} \leftarrow \mathcal{A} \cup \{y^*\}$

How well does this greedy algorithm do?

Submodularity and the Greedy Algorithm

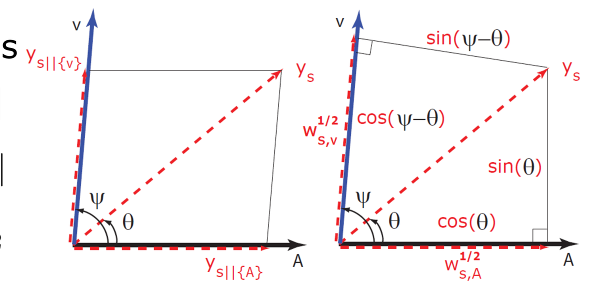
Theorem [Nemhauser et al, '78]
For the greedy solution A_G, it holds that
$$F(\mathcal{A}_G) \ge (1 - 1/e) \max_{|\mathcal{A}| \le k} F(A)$$

Krause et al '08: For approximately submodular F:
$$F(\mathcal{A}_G) \ge (1 - 1/e) \max_{|\mathcal{A}| \le k} F(A) - k\varepsilon$$

Key question:
Is the variance reduction F approximately submodular?
Answer: Not always, but in many practical settings…

A Sufficient Condition: *Incoherence*

Incoherence of columns
$\mu = \max_{\forall (i,j), i \ne j} |\langle \phi_i, \phi_j \rangle|$
$= \max_{\forall (i,j), i \ne j} |\cos \psi_{i,j}|$



Define: $w_{s,v} = \langle y_s, \phi_v \rangle^2$
$\widehat{F}_s(\mathcal{D}) = \max_{|\mathcal{A}| \le k|, \mathcal{A} \subseteq \mathcal{D}} \sum_{v \in \mathcal{A}} w_{s,v}$ *(modular approximation)*

Proposition: $\widehat{F}_s(\mathcal{D})$ is submodular! Furthermore,
$$|\widehat{F}_s(\mathcal{D}) - F_s(\mathcal{D})| \le k\mu$$
Thus $F_s(\mathcal{D})$ **is approximately submodular!**

An Algorithm for DiSP: $\mathbf{SDS_{OMP}}$

Algorithm $\mathbf{SDS_{OMP}}$
- Use *Orthogonal Matching Pursuit* to evaluate F for fixed D
- Use greedy algorithm to select columns of D

Theorem:
$\mathbf{SDS_{OMP}}$ will produce a dictionary such that
$$F(\mathcal{D}_{OMP}) \ge (1 - 1/e) \max_{|\mathcal{D}| \le n} F(\mathcal{D}) - k(6n + 2 - 1/e)\mu$$

- Need *n k* to be much less than *d*

Improved Guarantees: $\mathbf{SDS_{MA}}$

Algorithm $\mathbf{SDS_{MA}}$
- Optimize modular approximation $\widehat{F}$ instead of F
- Use greedy algorithm to select columns of D

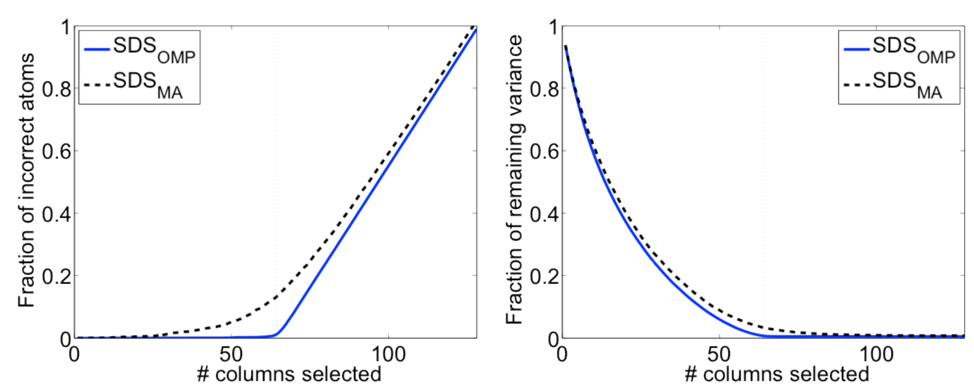
Theorem:
$\mathbf{SDS_{MA}}$ will produce a dictionary such that
$$F(\mathcal{D}_{MA}) \ge (1 - 1/e) \max_{|\mathcal{D}| \le n} F(\mathcal{D}) - (2 - 1/e)k\mu$$

SDS_{MA} is **much faster**, and has **better guarantees**
SDS_{OMP} empirically performs better in some settings (in particular when μ is large)

Experiment: Finding a Basis in a Haystack

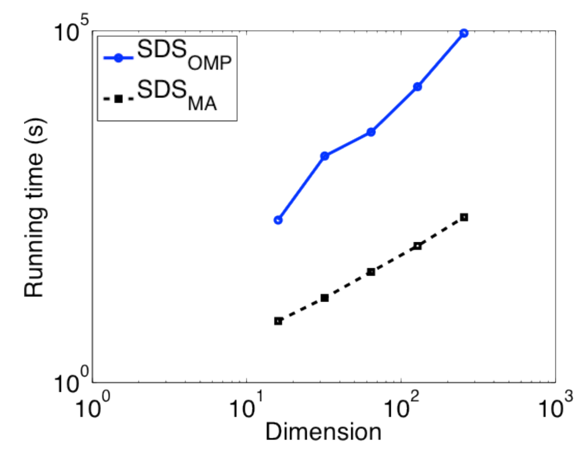
- V union of 8 orthonormal, 64-dimensional bases
  (Discrete Cosine Transform, Haar, Daubechies 4 & 8, Coiflets 1, 3, 5 and Discrete Meyer)
- Pick dictionary $\mathcal{D}^*$ of size n=64 at random
- Generate 100 sparse (k=5) signals at random
- Use SDS to pick dictionary $\mathcal{D}$ of increasing size
- Evaluate
  - fraction of correctly recovered columns
  - variance reduction
  - running time

Reconstruction Performance



SDS_{OMP} has perfect reconstruction accuracy for this data
SDS_{MA} comparable the variance reduction performance

Algorithm Running Times

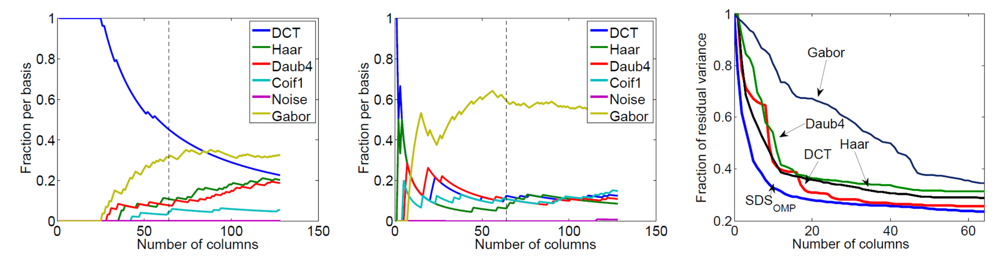


SDS_{MA} is a few orders of magnitude faster!

Battle of Bases on Natural Image Patches

Seek a dictionary among existing bases

discrete cosine transform (DCT), wavelets (Haar, Daub4), Coiflets, noiselets, and Gabor (frame)

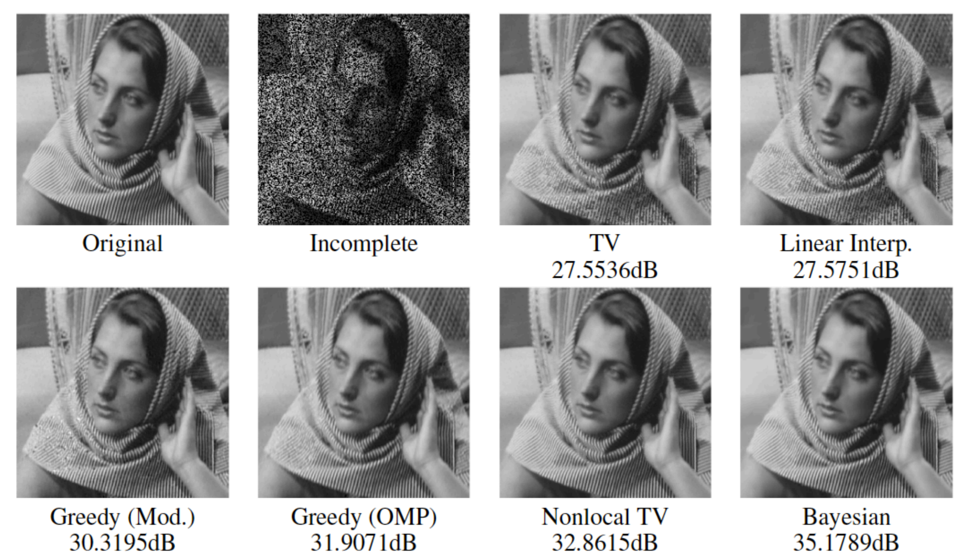


SDS_{OMP} prefers DCT+Gabor
SDS_{MA} chooses Gabor (predominantly)
Optimized dictionary improves compression

Experiment: Inpainting

- Dictionary selection from dimensionality-reduced measurements
- Take Barbara with 50% pixels missing at random
- Partition the image into 8x8 patches
- Optimize dictionary based on observed pixel values
- "Inpaint" the missing pixels via sparse reconstruction

Results: Inpainting



Comparable to state-of-the art nonlocal TV;
Orders of magnitude faster!

Conclusions

- Dictionary Selection <> **new problem**
  dictionary learning + dictionary design
- Incoherence assumptions <> **approximate submodularity**
- Two algorithms <> $\mathbf{SDS_{MA}}$ and $\mathbf{SDS_{OMP}}$ with guarantees
- Extensions to structured sparsity in the paper
- *Novel connection between sparsity and submodularity*