MMSE Estimation for Sparse Representation Modeling*

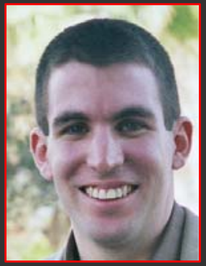
Michael Elad

The Computer Science Department

The Technion – Israel Institute of technology

Haifa 32000, Israel

Joint work with



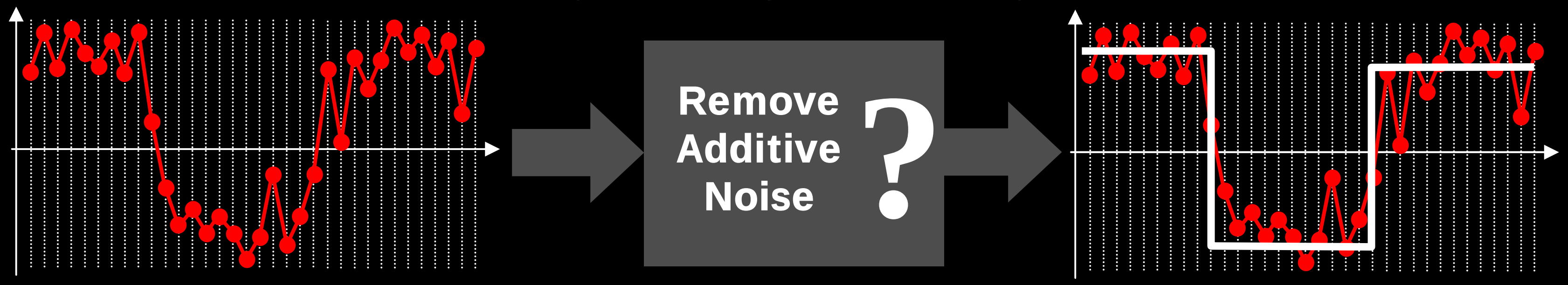
Irad Yavneh & Matan Protter

École Polytechnique

April 6th, 2009

Noise Removal?

In this talk we focus on signal/image denoising ...



Remove Additive Noise ?

- Important: (i) Practical application; (ii) A convenient platform for testing basic ideas in signal/image processing.

- Many Considered Directions: Partial differential equations, Statistical estimators, Adaptive filters, Inverse problems & regularization, Wavelets, Example-based techniques, Sparse representations, ...

- Main Massage Today: Several sparse representations can be found and used for better denoising performance – we introduce, motivate, discuss, demonstrate, and explain this new idea.

Agenda

1. Background on Denoising with Sparse Representations
2. Using More than One Representation: Intuition
3. Using More than One Representation: Theory
4. A Closer Look At the Unitary Case
5. Summary and Conclusions

Part I
Background on Denoising with Sparse Representations

Denoising By Energy Minimization

Many of the proposed signal denoising algorithms are related to the minimization of an energy function of the form

$$f(\underline{x}) = \frac{1}{2}\left\|\underline{x}-\underline{y}\right\|_2^2 + \Pr(\underline{x})$$

Relation to measurements

Prior or regularization

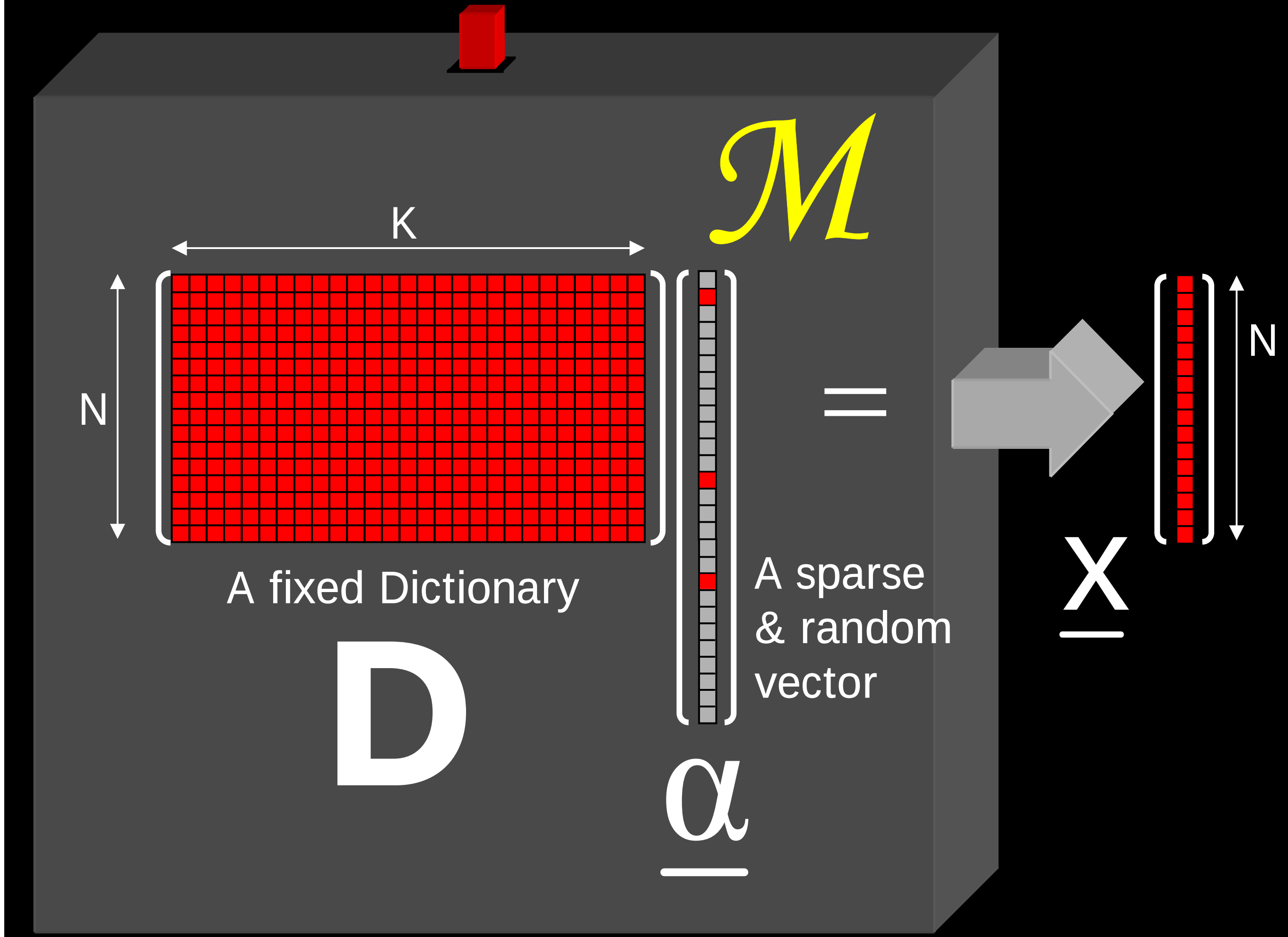
$\underline{y}$: Given measurements

$\underline{x}$: Unknown to be recovered

- ❑ This is in-fact a Bayesian point of view, adopting the Maximum-A-posteriori Probability (MAP) estimation.
- ❑ Clearly, the wisdom in such an approach is within the choice of the prior – **modeling the signals** of interest.

Thomas Bayes
1702 - 1761

Sparse Representation Modeling



- ❑ Every column in **D** (dictionary) is a prototype signal (atom).
- ❑ The vector $\underline{\alpha}$ is generated randomly with few (say L for now) non-zeros at random locations and with random values.

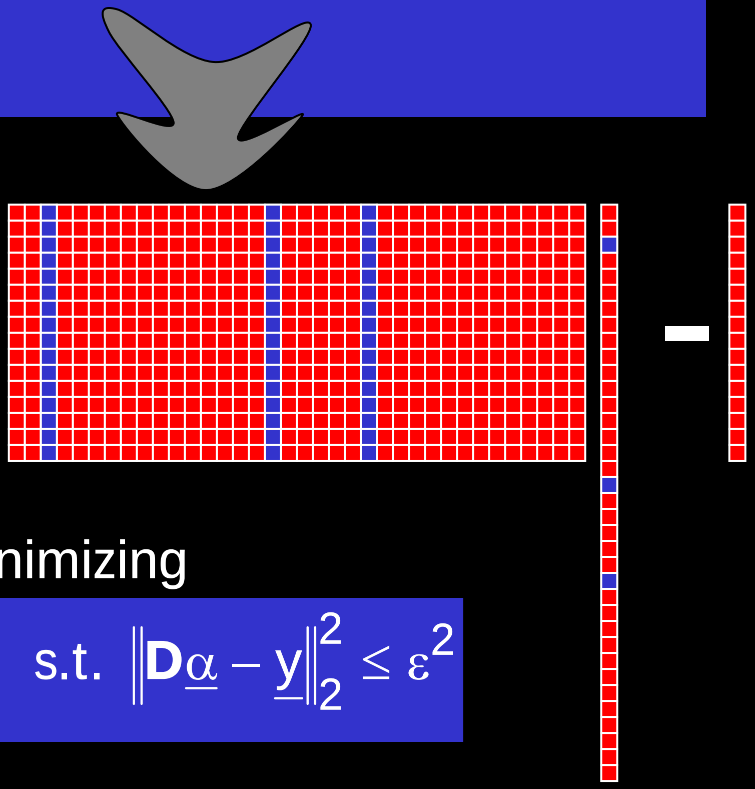
Back to Our MAP Energy Function

- □ The L_0 “norm” is effectively counting the number of non-zeros in $\underline{\alpha}$.

- □ The vector $\underline{\alpha}$ is the representation (**sparse**/**redundant**).

$$\frac{1}{2}\left\|\underline{x}-\underline{y}\right\|_2^2$$

$$\mathbf{D}\underline{\alpha}-\underline{y}=$$



- □ Bottom line: Denoising of $\underline{y}$ is done by minimizing

$$\min_{\underline{\alpha}} \left\|\mathbf{D}\underline{\alpha}-\underline{y}\right\|_2^2 \text{ s.t. } \left\|\underline{\alpha}\right\|_0^0 \le L$$

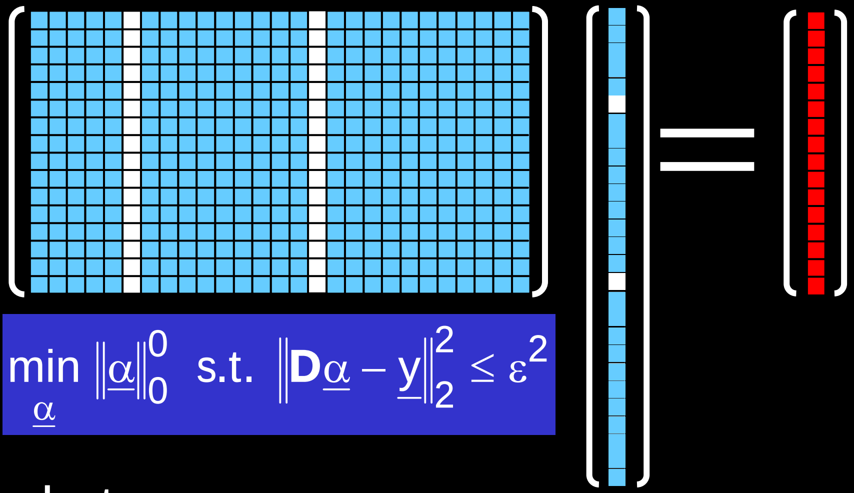
or

$$\min_{\underline{\alpha}} \left\|\underline{\alpha}\right\|_0^0 \text{ s.t. } \left\|\mathbf{D}\underline{\alpha}-\underline{y}\right\|_2^2 \le \varepsilon^2$$

The Solver We Use: Greed Based

- ❑ The MP is one of the greedy algorithms that finds one atom at a time [Mallat & Zhang ('93)].

- ❑ Step 1: find the one atom that best matches the signal.



$$\min_{\underline{\alpha}} \|\underline{\alpha}\|_0^0 \ \text{s.t.}\ \|\mathbf{D}\underline{\alpha}-\underline{y}\|_2^2 \le \varepsilon^2$$

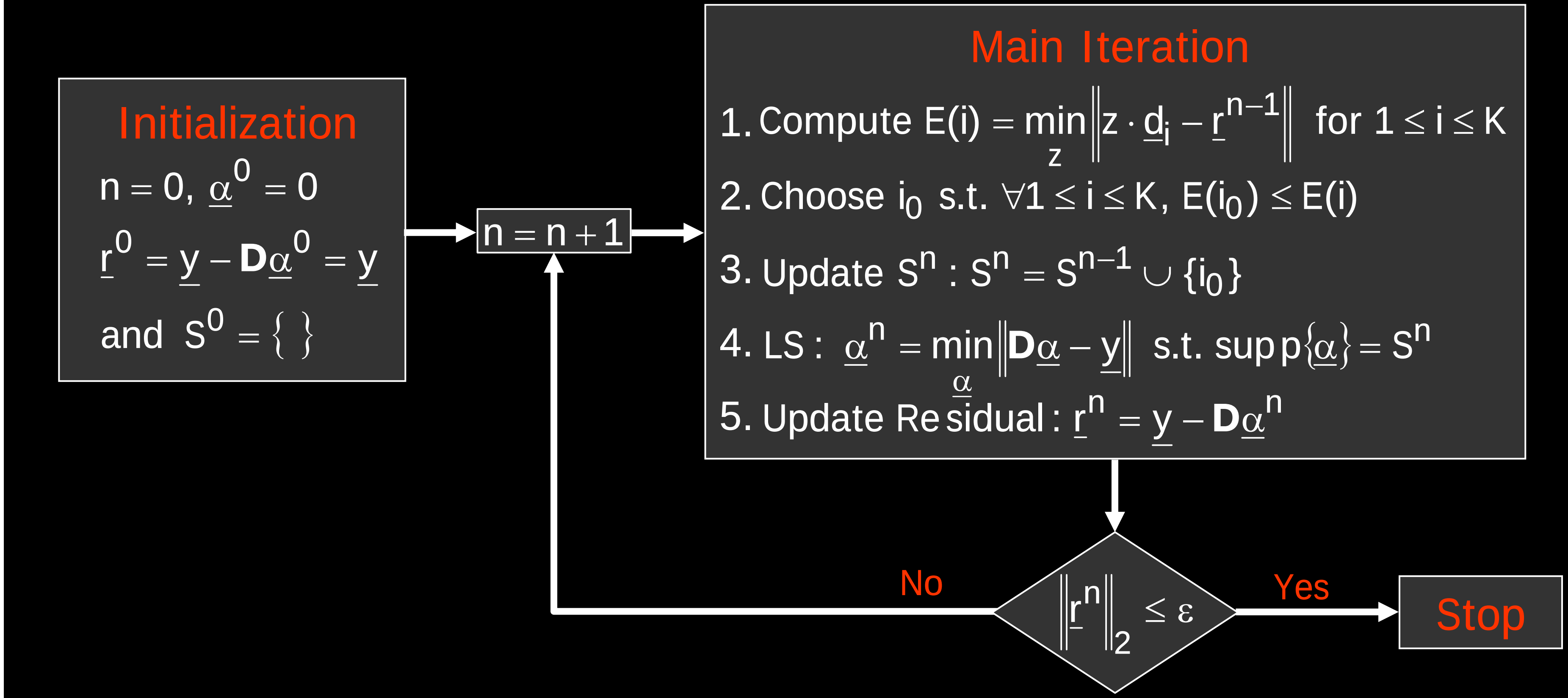
- ❑ Next steps: given the previously found atoms, find the next **<u>one</u>** to best fit the residual.

- ❑ The algorithm stops when the error $\|\mathbf{D}\underline{\alpha}-\underline{y}\|_2$ is below the destination threshold.

- ❑ The Orthogonal MP (OMP) is an improved version that re-evaluates the coefficients by Least-Squares after each round.

Orthogonal Matching Pursuit

OMP finds one atom at a time for approximating the solution of $\min_{\underline{\alpha}} \|\underline{\alpha}\|_0^0$ s.t. $\|\mathbf{D}\underline{\alpha}-\underline{y}\|_2^2 \le \varepsilon^2$



Initialization

$n=0,\ \underline{\alpha}^0=0$

$\underline{r}^0=\underline{y}-\mathbf{D}\underline{\alpha}^0=\underline{y}$

and $S^0=\{\ \}$

$n=n+1$

Main Iteration

1. Compute $E(i)=\min_{z}\left\|z\cdot\underline{d}_i-\underline{r}^{n-1}\right\|$ for $1\le i\le K$
2. Choose i_0 s.t. $\forall 1\le i\le K,\ E(i_0)\le E(i)$
3. Update S^n : $S^n=S^{n-1}\cup\{i_0\}$
4. LS : $\underline{\alpha}^n=\min_{\underline{\alpha}}\|\mathbf{D}\underline{\alpha}-\underline{y}\|$ s.t. $\mathrm{supp}\{\underline{\alpha}\}=S^n$
5. Update Residual : $\underline{r}^n=\underline{y}-\mathbf{D}\underline{\alpha}^n$

$\|\underline{r}^n\|_2\le\varepsilon$

No

Yes

Stop

Part II

Using More than One Representation: Intuition

Back to the Beginning. What If …

Consider the denoising problem

$$\min_{\underline{\alpha}} \|\underline{\alpha}\|_0^0 \quad \text{s.t.} \quad \|\mathbf{D}\underline{\alpha} - \underline{y}\|_2^2 \le \varepsilon^2$$

and suppose that we can find a group of J candidate solutions

$$\left\{\underline{\alpha}_j\right\}_{j=1}^{J}$$

such that

$$\forall j \quad \left\{ \begin{array}{c} \|\underline{\alpha}_j\|_0^0 << N \\ \|\mathbf{D}\underline{\alpha}_j - \underline{y}\|_2^2 \le \varepsilon^2 \end{array} \right\}$$

Basic Questions:

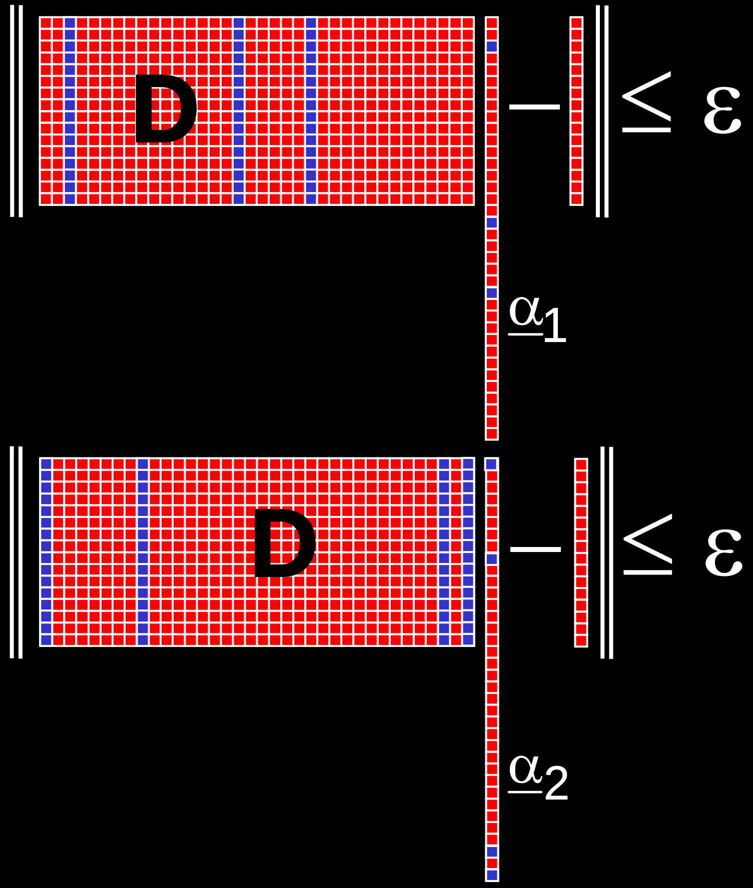
- ❑ **What** could we do with such a set of competing solutions in order to better denoise $\underline{y}$?
- ❑ **Why** should this help?
- ❑ **How** shall we practically find such a set of solutions?

Relevant work: [Larsson & Selen ('07)]
[Schintter et. al. (`08)]
[Elad and Yavneh ('08)]

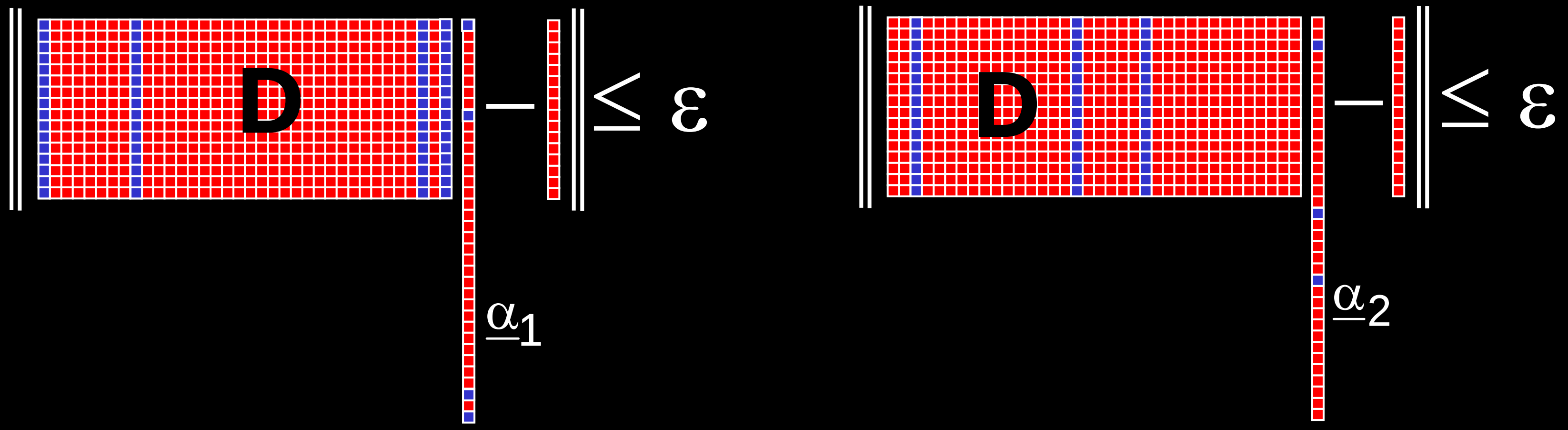
Motivation – General

Why bother with such a set?

- ❑ Because each representation conveys a different story about the desired signal.
- ❑ Because pursuit algorithms are often wrong in finding the sparsest representation, and then relying on their solution is too sensitive.
- ❑ ... Maybe there are “deeper” reasons?



Our Motivation



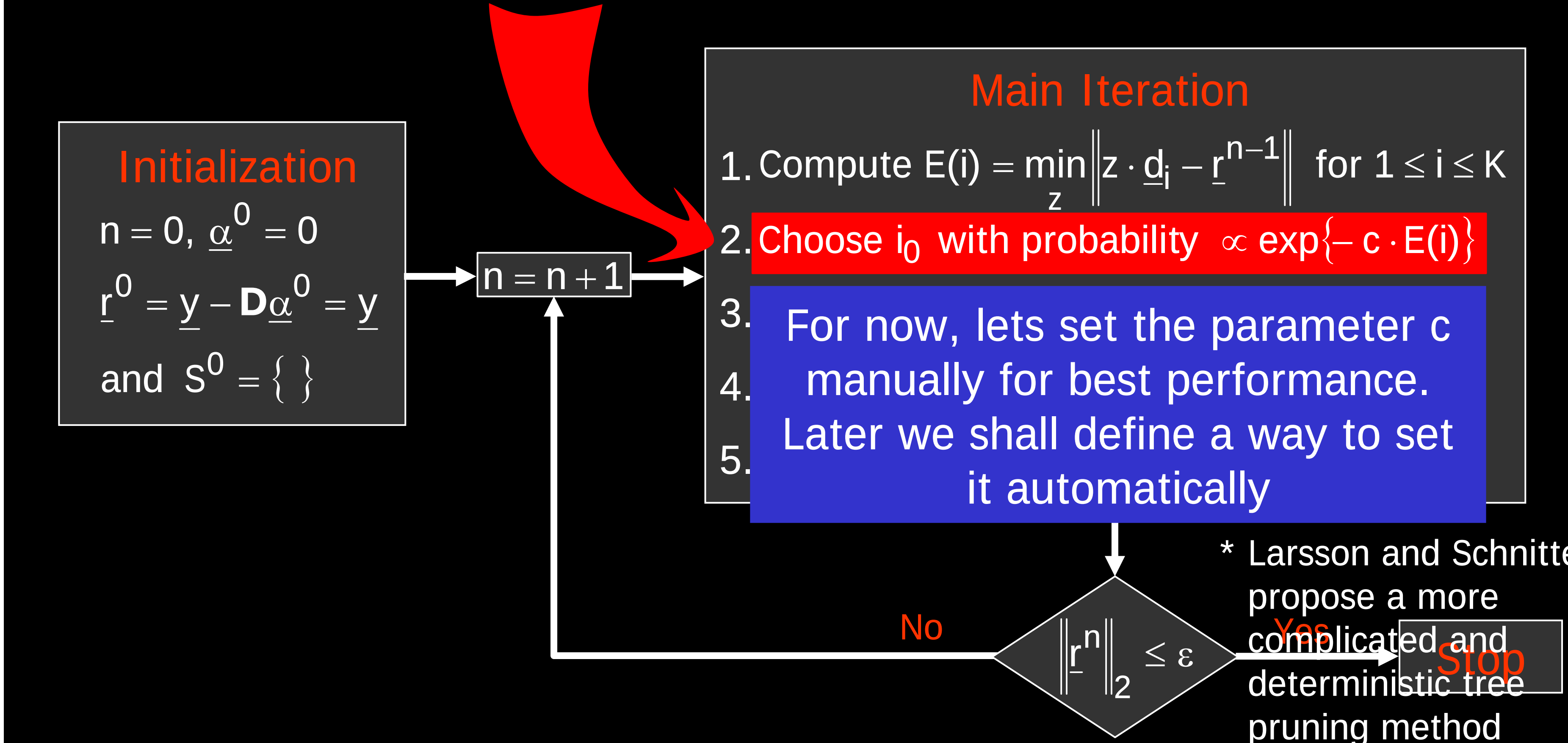
- An intriguing relationship between this idea and the common-practice in example-based techniques, where several examples are merged.

- Consider the Non-Local-Means [Buades, Coll, & Morel ('05)]. It uses
  (i) a local dictionary (the neighborhood patches),
  (ii) it builds several sparse representations (of cardinality 1), and
  (iii) it merges them.

- Why not take it further, and use general sparse representations?

Generating Many Representations

Our[*] Answer: Randomizing the OMP



Initialization

$n=0,\ \underline{\alpha}^0=0$

$\underline{r}^0=\underline{y}-\mathbf{D}\underline{\alpha}^0=\underline{y}$

and $S^0=\{\ \}$

$n=n+1$

Main Iteration

1. Compute $E(i)=\min_{z}\left\|z\cdot\underline{d}_i-\underline{r}^{n-1}\right\|$ for $1\le i\le K$
2. Choose i_0 with probability $\propto\exp\{-c\cdot E(i)\}$
3.
4.
5.

For now, lets set the parameter c manually for best performance. Later we shall define a way to set it automatically

$\left\|\underline{r}^n\right\|_2\le\varepsilon$

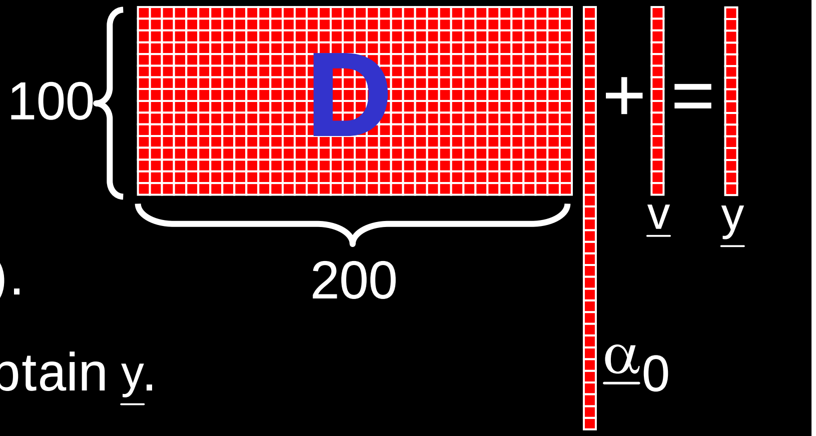
No

Yes

Stop

[*] Larsson and Schnitter propose a more complicated and deterministic tree pruning method

Lets Try



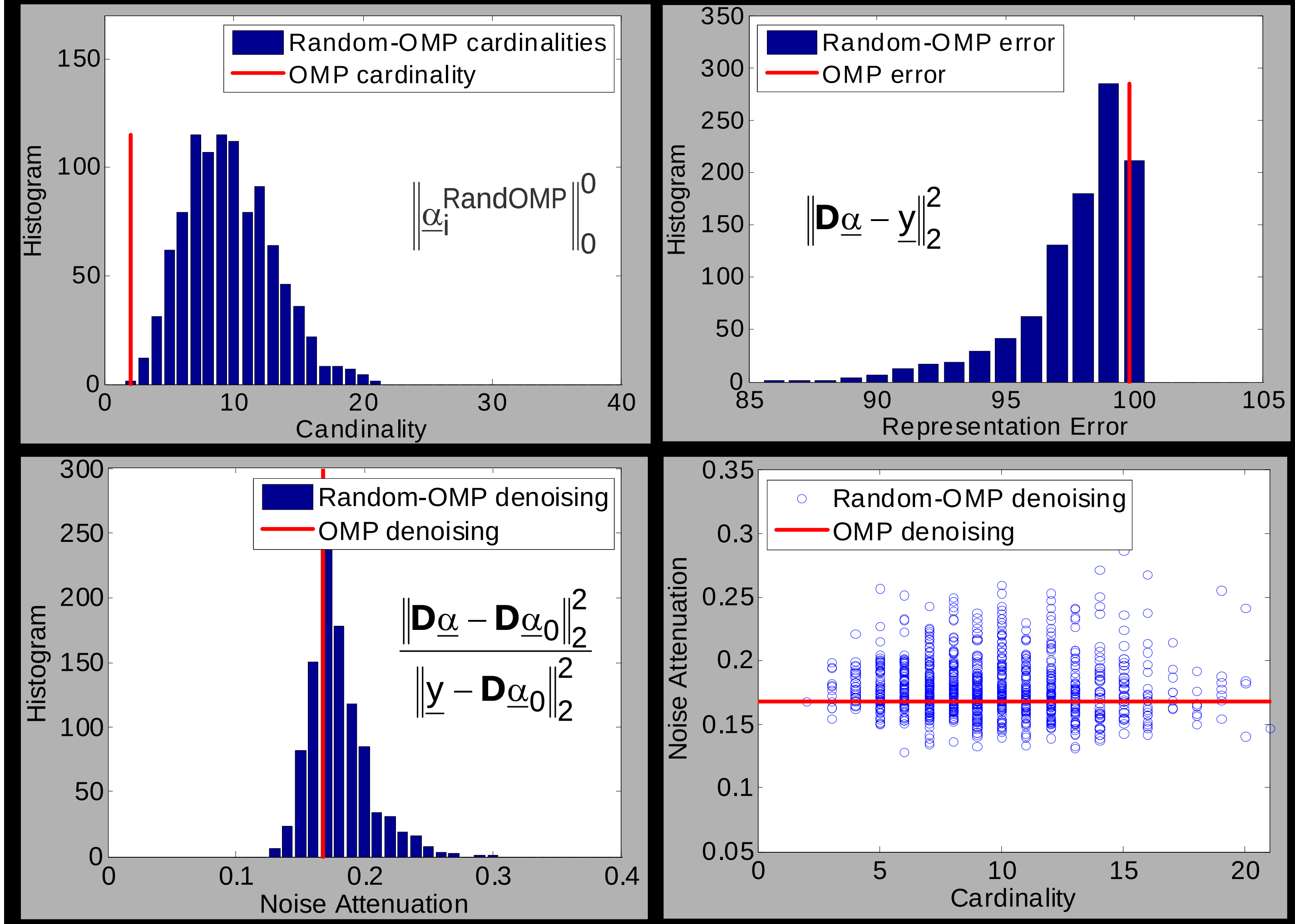
Proposed Experiment :

- ❑ Form a random dictionary **D**.
- ❑ Multiply by a sparse vector $\underline{\alpha}_0$ ($\|\underline{\alpha}_0\|_0^0 = 10$).
- ❑ Add Gaussian iid noise $\underline{v}$ with σ=1 and obtain $\underline{y}$.
- ❑ Solve the problem

  $$\min_{\underline{\alpha}} \|\underline{\alpha}\|_0^0 \text{ s.t. } \|\mathbf{D}\underline{\alpha} - \underline{y}\|_2^2 \le 100$$

  using OMP, and obtain $\underline{\alpha}^{OMP}$.
- ❑ Use Random-OMP and obtain $\left\{\underline{\alpha}_j^{RandOMP}\right\}_{j=1}^{1000}$.
- ❑ Lets look at the obtained representations ...

Some Observations



We see that

- The OMP gives the sparsest solution
- Nevertheless, it is not the most effective for denoising.
- The cardinality of a representation does not reveal its efficiency.

The Surprise (at least for us) ...

Lets propose the average

$$\underline{\hat{\alpha}} = \frac{1}{1000}\sum_{j=1}^{1000} \underline{\alpha}_j^{RandOMP}$$

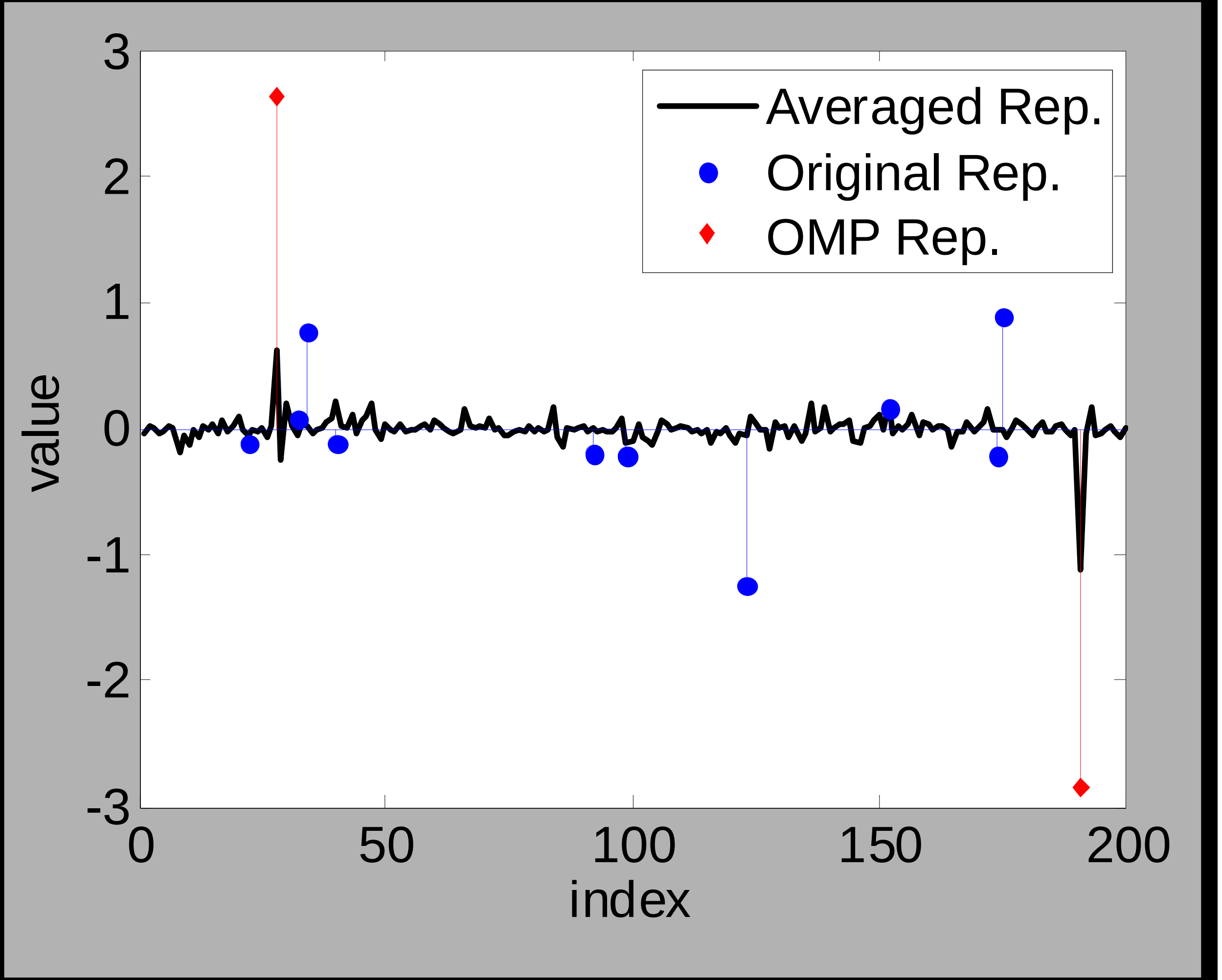
as our representation

This representation

IS NOT SPARSE AT ALL

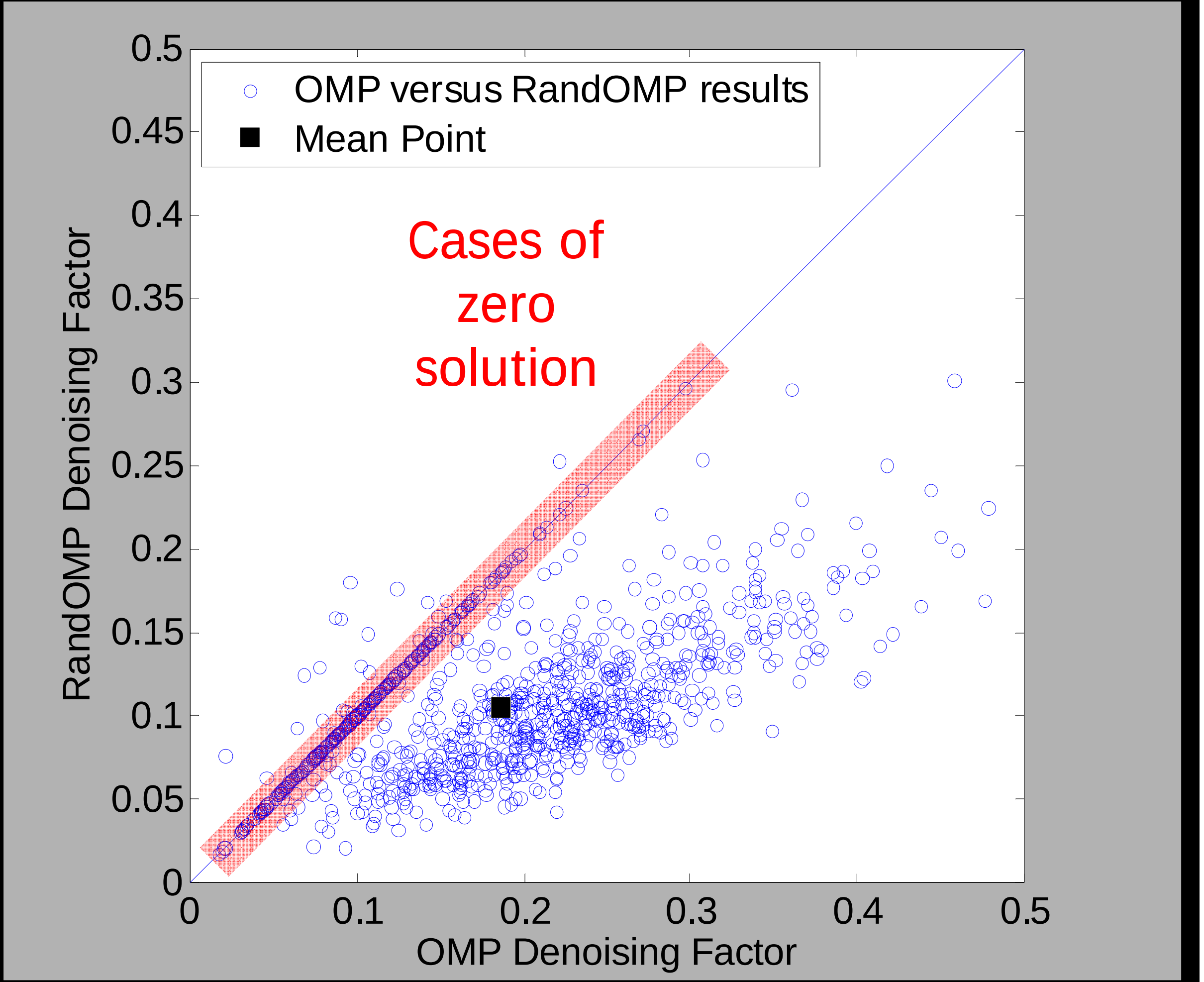
but it gives

$$\frac{\left\| \mathbf{D}\underline{\hat{\alpha}} - \mathbf{D}\underline{\alpha}_0 \right\|_2^2}{\left\| \underline{y} - \mathbf{D}\underline{\alpha}_0 \right\|_2^2} = 0.06$$



Is It Consistent? … Yes!

Here are the results of 1000 trials with the same parameters …

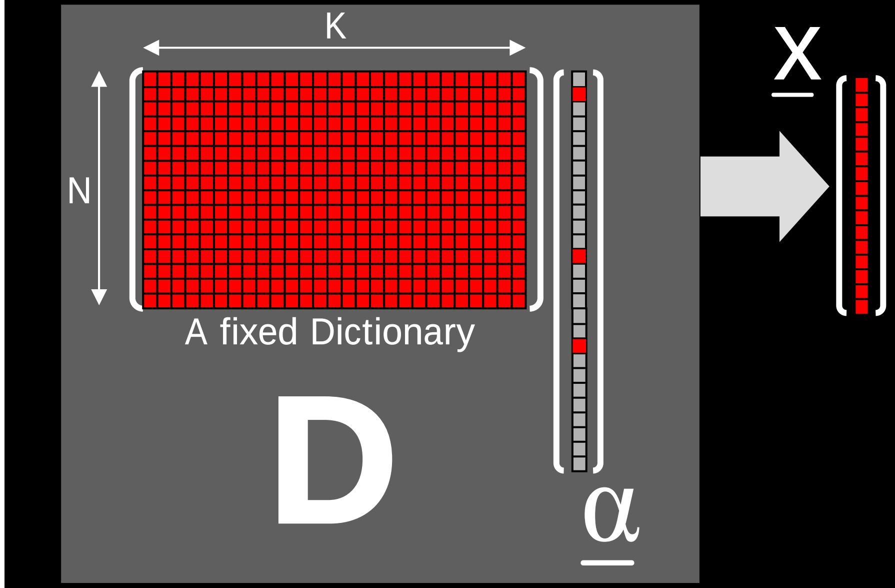


Cases of zero solution

Part III

Using More than One Representation: Theory

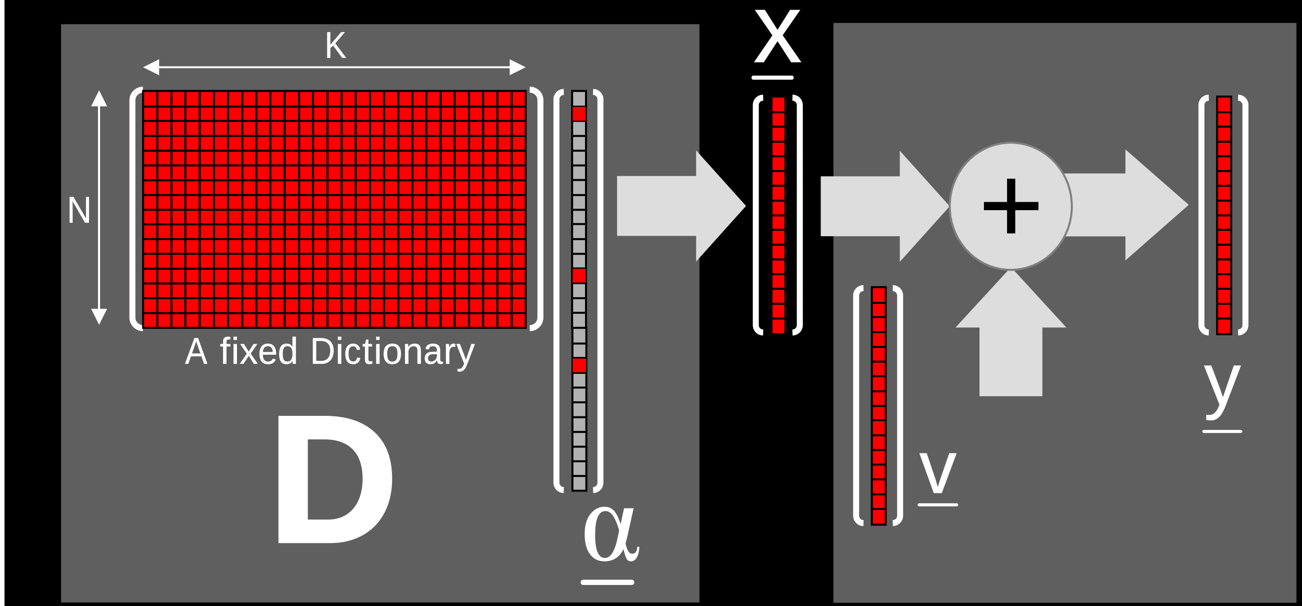
Our Signal Model



- ❑ **D** is fixed and known.
- ❑ The vector $\underline{\alpha}$ is built by:
  - Choosing the support s with probability P(s) from all the 2^K possibilities Ω.
  - **For simplicity, assume that |s|=k is fixed and known.**
  - Choosing the $\underline{\alpha}_s$ coefficients using iid Gaussian entries $N(0,\sigma_x)$.
- ❑ The ideal signal is $\underline{x}=\mathbf{D}\underline{\alpha}=\mathbf{D}_s\underline{\alpha}_s$.

The p.d.f. $P(\underline{\alpha})$ and $P(\underline{x})$ are clear and known

Adding Noise



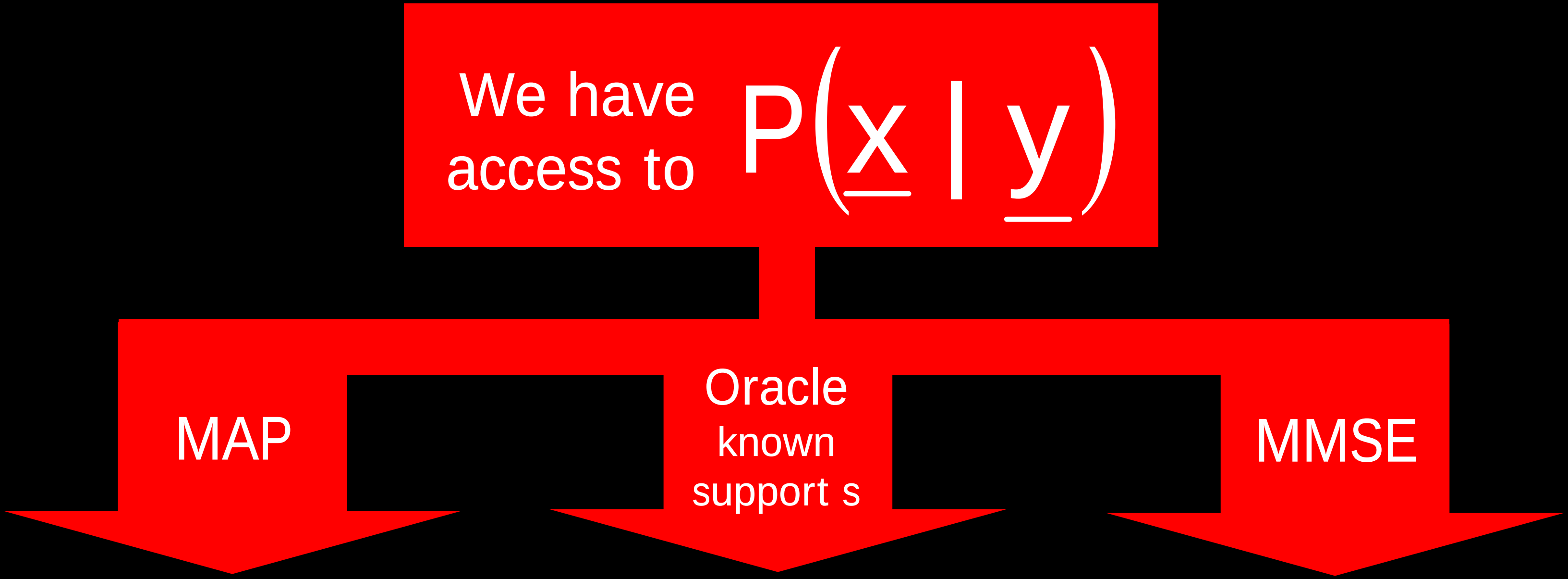
Noise Assumed:

The noise $\underline{v}$ is additive white Gaussian vector with probability $P_v(\underline{v})$

$$P(\underline{y}|\underline{x}) = C \cdot \exp\left\{-\frac{\|\underline{x}-\underline{y}\|^2}{2\sigma^2}\right\}$$

The conditional p.d.f.'s P($\underline{y}$|s), P(s|$\underline{y}$), and even also P($\underline{x}$|$\underline{y}$) are all clear and well-defined (although they may appear nasty).

The Key – The Posterior $P(\underline{x}|\underline{y})$



We have access to $P\left(\underline{x}\,|\,\underline{y}\right)$

MAP

Oracle known support s

MMSE

$$\hat{\underline{x}}^{MAP}=\underset{\underline{x}}{\text{ArgMax}}P(\underline{x}\,|\,\underline{y})$$

$$\hat{\underline{x}}^{oracle}$$

$$\hat{\underline{x}}^{MMSE}=E\left\{\underline{x}\,|\,\underline{y}\right\}$$

- ❑ The estimation of $\underline{\alpha}$ and multiplication by **D** is equivalent to the above.
- ❑ These two estimators are impossible to compute, as we show next.

Lets Start with The Oracle[*]

$$P\left(\underline{\alpha}\,|\,\underline{y},s\right)=P\left(\underline{\alpha}_s\,|\,\underline{y}\right)$$

$$P\left(\underline{y}\,|\,\underline{\alpha}_s\right)\propto\exp\left\{-\frac{\left\|\underline{y}-\mathbf{D}_s\underline{\alpha}_s\right\|^2}{2\sigma^2}\right\}$$

$$P\left(\underline{\alpha}_s\right)\propto\exp\left\{-\frac{\left\|\underline{\alpha}_s\right\|^2}{2\sigma_x^2}\right\}$$

$$P\left(\underline{\alpha}_s\,|\,\underline{y}\right)\propto\exp\left\{-\frac{\left\|\underline{y}-\mathbf{D}_s\underline{\alpha}_s\right\|^2}{2\sigma^2}-\frac{\left\|\underline{\alpha}_s\right\|^2}{2\sigma_x^2}\right\}$$

$$\boxed{\hat{\underline{\alpha}}_s=\left[\frac{1}{\sigma^2}\mathbf{D}_s^T\mathbf{D}_s+\frac{1}{\sigma_x^2}\mathbf{I}\right]^{-1}\frac{1}{\sigma^2}\mathbf{D}_s^T\underline{y}}$$

Comments:

- This estimate is both the MAP and MMSE.
- The oracle estimate of $\underline{x}$ is obtained by multiplication by $\mathbf{D}_s$.

[*] When s is known

The MAP Estimation

$$\hat{\underline{\alpha}}^{MAP} = \underset{\underline{\alpha}_s, s\in\Omega}{\text{ArgMax}}\, P(\underline{\alpha}_s, s \mid \underline{y}) = \underset{s\in\Omega}{\text{ArgMax}}\, P(s \mid \underline{y}) \cdot P(\underline{\alpha}_s \mid \underline{y}, s)$$

$$P(s \mid \underline{y}) \propto P(s) \cdot P(\underline{y} \mid s) = \ldots$$

$$\propto P(s) \cdot \exp\left\{ \frac{\underline{h}_s^T \mathbf{Q}_s^{-1} \underline{h}_s}{2} + \frac{\log(\det(\mathbf{Q}_s^{-1}))}{2} \right\}$$

the oracle's

upport s:

$$\ldots \frac{\ldots \underline{\alpha}_s \|^2}{\ldots 2} - \frac{\| \underline{\alpha}_s \|^2}{2\sigma_x^2} \Bigg\}$$

$$\hat{\underline{s}}^{MAP} = \underset{s\in\Omega}{\text{ArgMax}}\, P(s) \cdot \exp\left\{ \frac{\underline{h}_s^T \mathbf{Q}_s^{-1} \underline{h}_s}{2} + \frac{\log(\det(\mathbf{Q}_s^{-1}))}{2} \right\}$$

The MAP Estimation

Implications:

$$\boxed{\underline{\hat{s}}^{\text{MAP}} = \underset{s\in\Omega}{\text{ArgMax}}\, P(s)\cdot\exp\left\{\frac{\underline{h}_s^T \mathbf{Q}_s^{-1}\underline{h}_s}{2}+\frac{\log(\det(\mathbf{Q}_s^{-1}))}{2}\right\}}$$

❑ The MAP estimator requires to test all the possible supports for the maximization. In typical problems, this is impossible as there is a combinatorial set of possibilities.

❑ This is why we rarely use exact MAP, and we typically replace it with approximation algorithms (e.g., OMP).

The MMSE Estimation

$$\hat{\underline{\alpha}}^{\mathrm{MMSE}} = E\{\underline{\alpha} \mid \underline{y}\} = \sum_{s\in\Omega} P(s \mid \underline{y}) \cdot E\{\underline{\alpha} \mid \underline{y}, s\}$$

$$P(s \mid \underline{y}) \propto P(s) \cdot P(\underline{y} \mid s) = \ldots$$

$$\propto P(s) \cdot \exp\left\{ \frac{\underline{h}_s^T \mathbf{Q}_s^{-1} \underline{h}_s}{2} + \frac{\log(\det(\mathbf{Q}_s^{-1}))}{2} \right\}$$

cle for s, as we

en before

$$= \hat{\underline{\alpha}}_s = \mathbf{Q}_s^{-1} \underline{h}_s$$

$$\boxed{\hat{\underline{\alpha}}^{\mathrm{MMSE}} = \sum_{s\in\Omega} P(s \mid \underline{y}) \cdot \underline{\alpha}_s}$$

The MMSE Estimation

$$\hat{\underline{\alpha}}^{MMSE}=E\left\{\underline{\alpha}\,|\,\underline{y}\right\}=\sum_{s\in\Omega}P(s\,|\,\underline{y})\cdot E\left\{\underline{\alpha}\,|\,\underline{y},s\right\}$$

Implications:

$$\boxed{\hat{\underline{\alpha}}^{MMSE}=\sum_{s\in\Omega}P(s\,|\,\underline{y})\cdot\underline{\alpha}_s}$$

- ❑ The best estimator (in terms of L_2 error) is a weighted average of many sparse representations!!!
- ❑ As in the MAP case, in typical problems one cannot compute this expression, as the summation is over a combinatorial set of possibilities. We should propose approximations here as well.

The Case of |s|=k=1

$$P(s\,|\,\underline{y}) \propto P(s)\cdot\exp\left\{\frac{\underline{h}_s^T \mathbf{Q}_s^{-1}\underline{h}_s}{2}+\frac{\log(\det(\mathbf{Q}_s^{-1}))}{2}\right\}$$

This is our c in the Random-OMP

The k-th atom in **D**

- ❑ Based on this we can propose a greedy algorithm for both MAP and MMSE:
  - MAP – choose the atom with the largest inner product (out of K), and do so one at a time, while freezing the previous ones (almost OMP).
  - MMSE – draw at random an atom in a greedy algorithm, based on the above probability set, getting close to P(s|y) in the overall draw.

Bottom Line

- The MMSE estimation we got requires a sweep through all supports (i.e. combinatorial search) – impractical.
- Similarly, an explicit expression for $P(\underline{x}/\underline{y})$ can be derived and maximized – this is the MAP estimation, and it also requires a sweep through all possible supports – impractical too.
- The OMP is a (good) approximation for the MAP estimate.
- The Random-OMP is a (good) approximation of the Minimum-Mean-Squared-Error (MMSE) estimate. It is close to the Gibbs sampler of the probability $P(s|\underline{y})$ from which we should draw the weights.

Back to the beginning: Why Use Several Representations?

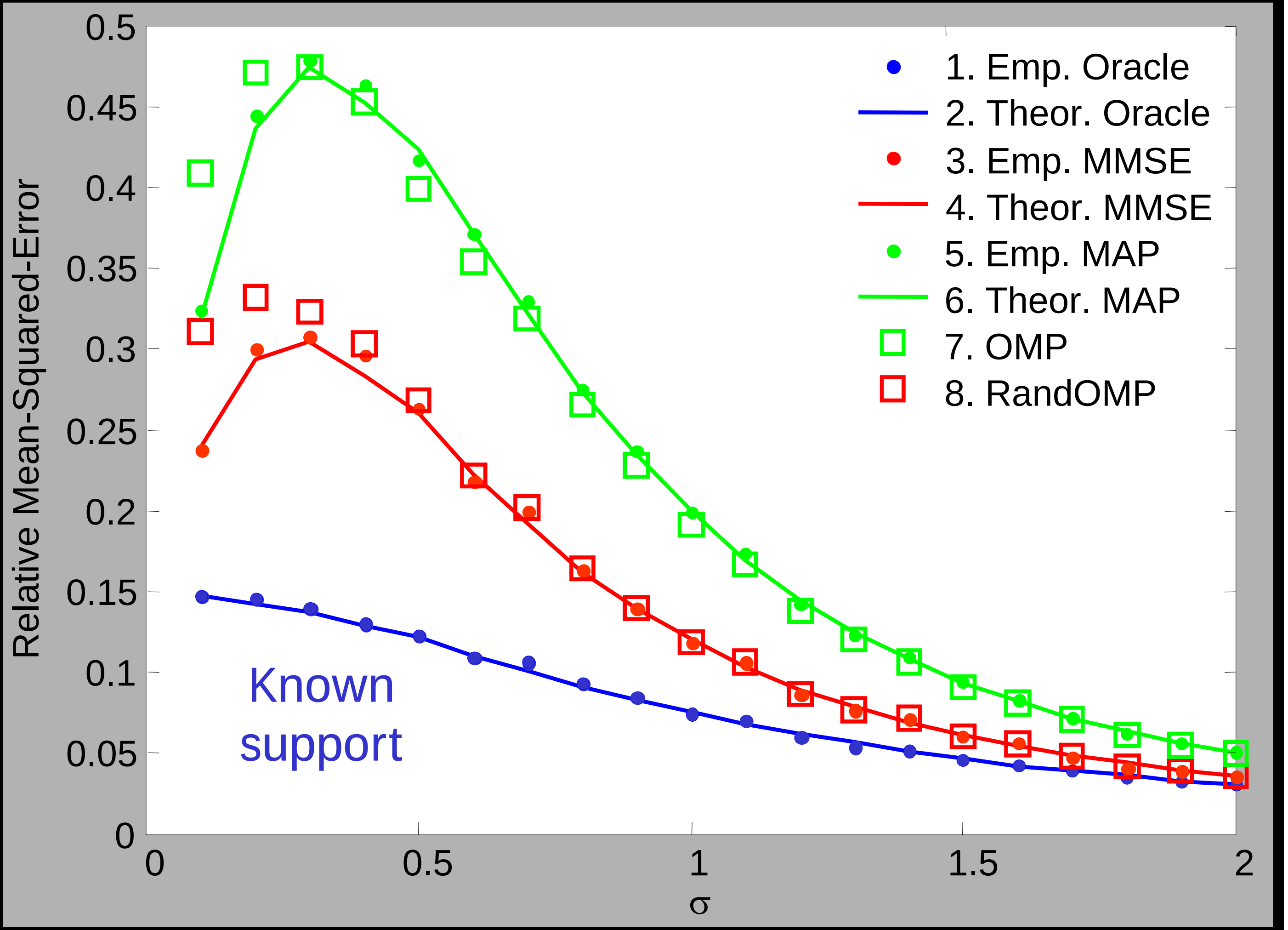
Because their average leads to a provable better noise suppression.

Comparative Results

The following results correspond to a small dictionary (20×30), where the combinatorial formulas can be evaluated as well.

Parameters:

- N=20, K=30
- True support=3
- $\sigma_x=1$
- J=10 (RandOMP)
- Averaged over 1000 experiments



Part IV

A Closer Look At the Unitary Case

$$DD^T = D^T D = I$$

Few Basic Observations

Let us denote $\underline{\beta} = \mathbf{D}^T \underline{y}$

$$\mathbf{Q}_s = \frac{1}{\sigma^2}\mathbf{D}_s^T\mathbf{D}_s + \frac{1}{\sigma_x^2}\mathbf{I} = \frac{\sigma^2 + \sigma_x^2}{\sigma^2\sigma_x^2}\mathbf{I}$$

$$\underline{h}_s = \frac{1}{\sigma^2}\mathbf{D}_s^T\underline{y} = \frac{1}{\sigma^2}\underline{\beta}_s$$

$$\boxed{\underline{\hat{\alpha}}_s} = \mathbf{Q}_s^{-1}\underline{h}_s = \frac{\sigma^2\sigma_x^2}{\sigma^2 + \sigma_x^2}\cdot\frac{1}{\sigma^2}\underline{\beta}_s = \boxed{c\cdot\underline{\beta}_s} \quad \text{(The Oracle)}$$

Back to the MAP Estimation*

$$\underline{\hat{s}}^{MAP} = \underset{s\in\Omega}{\text{ArgMax}}\ \exp\left\{\frac{\underline{h}_s^T \mathbf{Q}_s^{-1} \underline{h}_s}{2} + \frac{\log(\det(\mathbf{Q}_s^{-1}))}{2}\right\}$$

$$\underline{h}_s^T \mathbf{Q}_s^{-1} \underline{h}_s = \frac{c}{\sigma^2}\cdot\left\|\underline{\beta}_s\right\|_2^2$$

This part becomes a constant, and thus can be discarded

This means that MAP estimation can be easily evaluated by computing $\underline{\beta}$, sorting its entries in descending order, and choosing the k leading ones!

*We assume |s|=k fixed with equal probabilities

Closed-Form Estimation

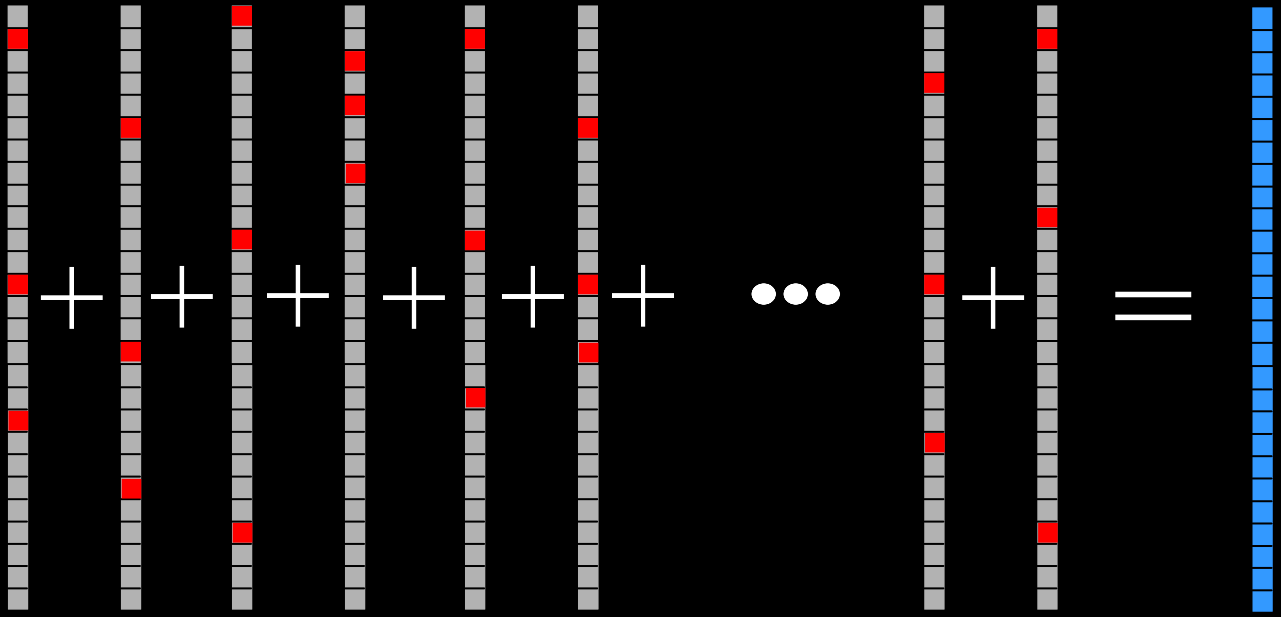
- ❑ It is well-known that MAP enjoys a closed form and simple solution in the case of a unitary dictionary **D**.
- ❑ This closed-form solution takes the structure of thresholding or shrinkage. The specific structure depends on the fine details of the model assumed.
- ❑ It is also known that OMP in this case becomes exact.

What about the MMSE?
Could it have a simple
closed-form solution too ?

The MMSE … Again

This is the formula we got:

$$\underline{\hat{\alpha}}^{MMSE} = c \cdot \sum_{s \in \Omega} P(s \mid \underline{y}) \cdot \underline{\beta}_s$$



The result is one effective representation (not sparse anymore)

We combine linearly many sparse representations (with proper weights)

The MMSE … Again

This is the formula we got: $\boxed{\underline{\hat{\alpha}}^{\text{MMSE}} = c \cdot \sum_{s \in \Omega} P(s \mid \underline{y}) \cdot \underline{\beta}_s}$

❑ We change the above summation to

$$\underline{\hat{\alpha}}^{\text{MMSE}} = \sum_{j=1}^{K} q_j^k \cdot \beta_j \cdot \underline{e}_j$$

where there are K contributions (one per each atom) to be found and used.

❑ We have developed a closed-form recursive formula for computing the q coefficients.

Towards a Recursive Formula

We have seen that the governing probability for the weighted averaging is given by

$$P(s\,|\,\underline{y}) = \ldots \propto \exp\left\{\frac{c}{2\sigma^2}\cdot\left\|\underline{\beta}_s\right\|_2^2\right\}$$

$$\underbrace{\qquad\qquad\qquad}_{q_i}$$

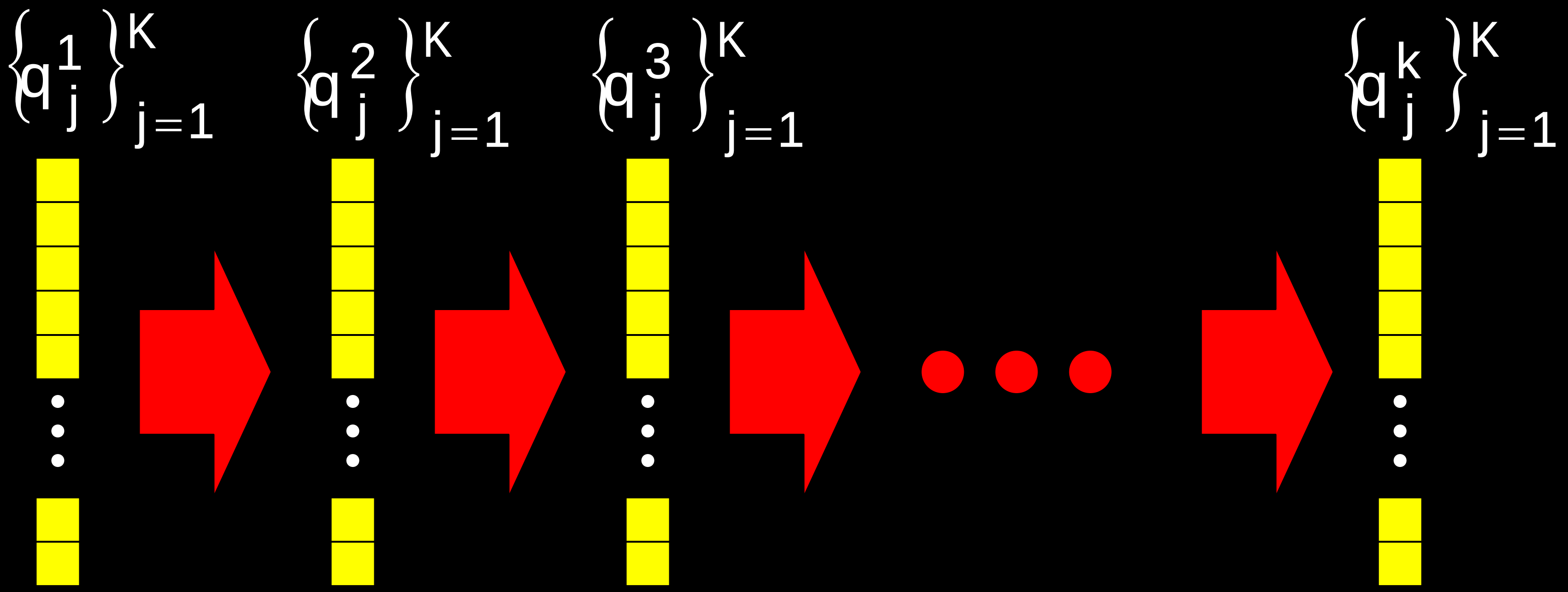
$$\underline{\hat{\alpha}}^{MMSE} = c\cdot\sum_{s\in\Omega} P(s\,|\,\underline{y})\cdot\underline{\beta}_s$$

q_j^k

Indicator function stating if j is in s

The Recursive Formula

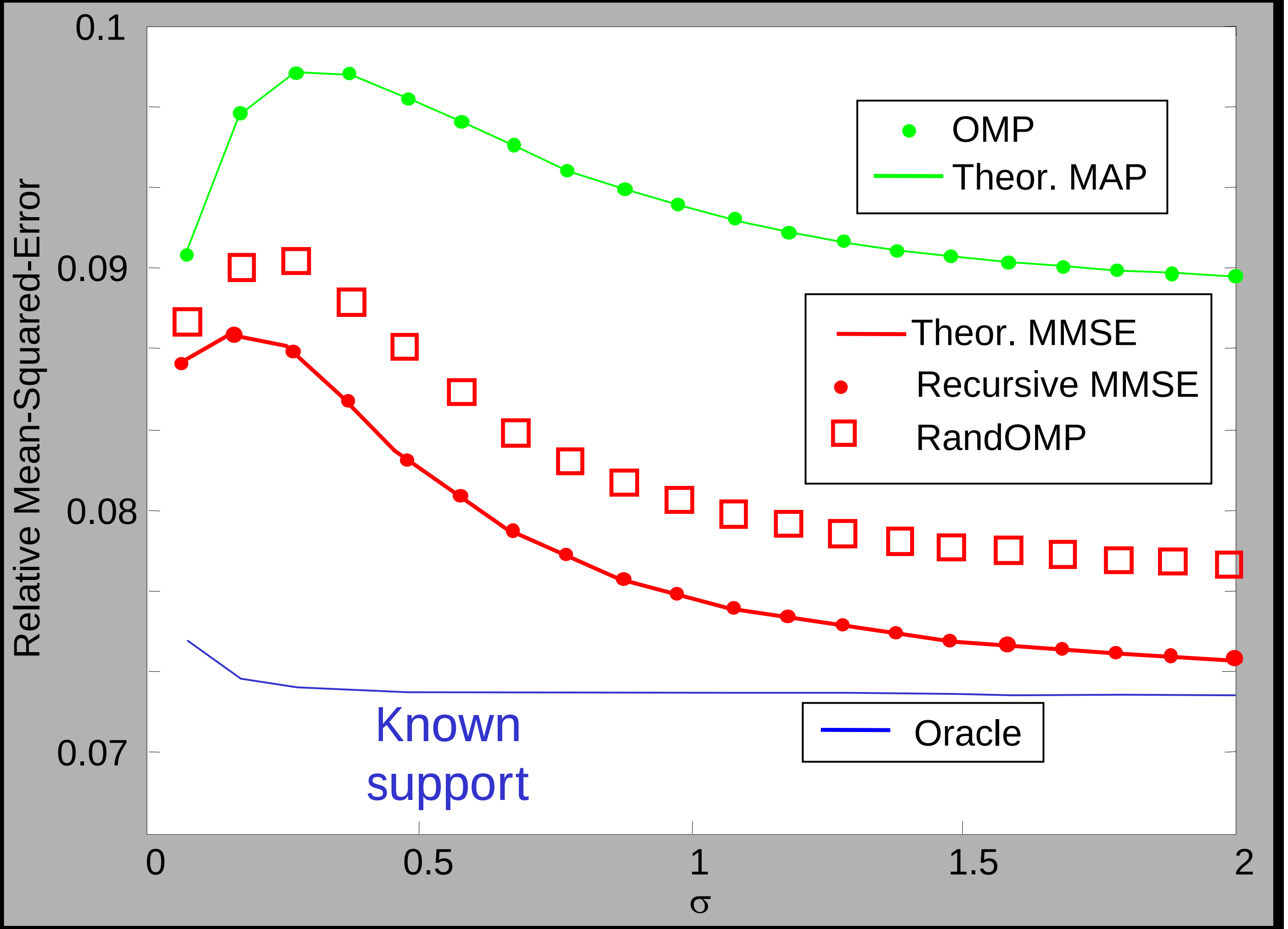
$$q_j^k = \sum_{s\in\Omega}\left(\prod_{i\in s} q_i\right)\cdot I_s(j) = \ldots = k\cdot\frac{q_j^1(1-q_j^{k-1})}{1-\sum_{\ell=1}^{K} q_\ell^1 q_\ell^{k-1}} \quad \text{where} \quad q_j^1 = q_j$$



An Example

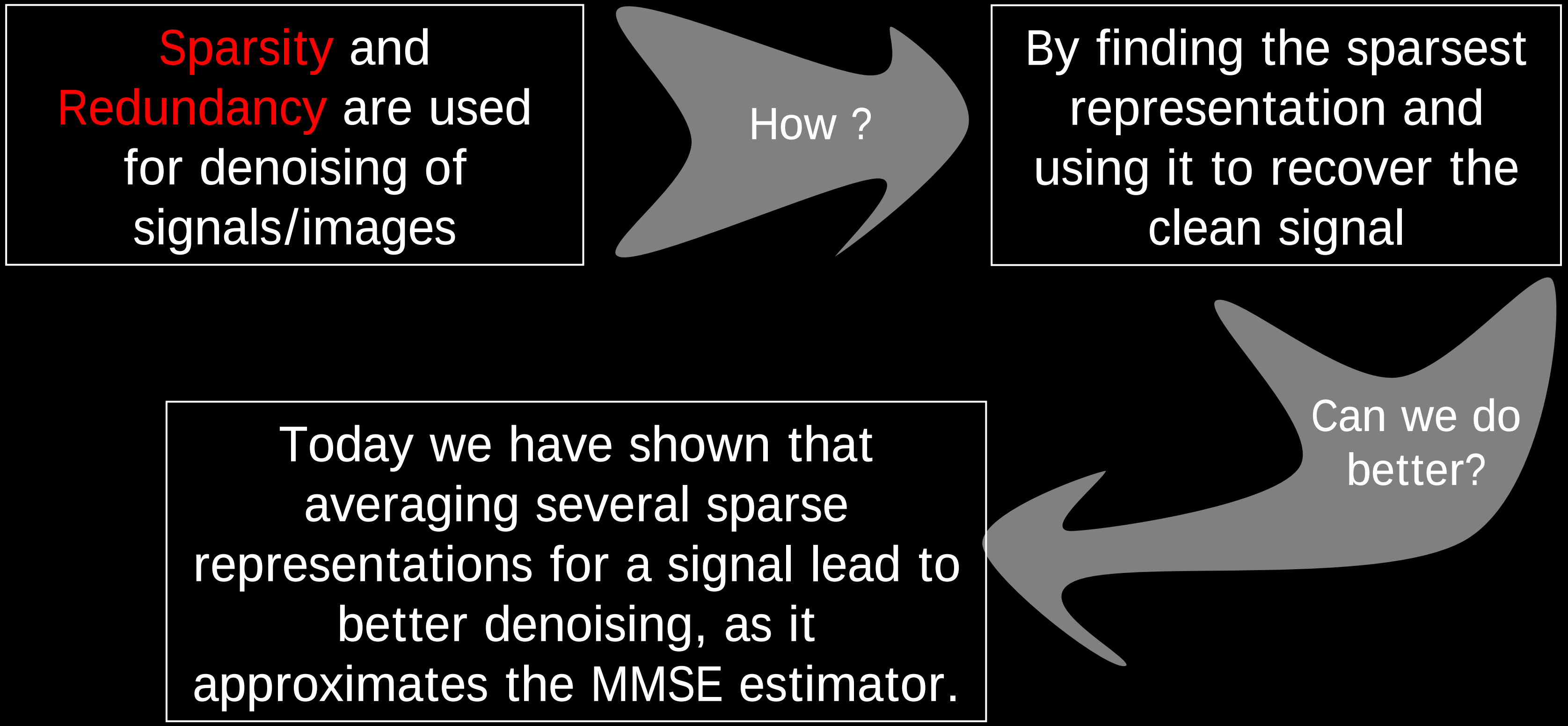
This is a synthetic experiment resembling the previous one, but with few important changes:

- **D** is unitary
- The representation's cardinality is 5 (the higher it is, the weaker the Random-OMP becomes)
- Dimensions are different: N=K=64
- J=20 (RandOMP runs)



Part V
Summary and Conclusions

Today We Have Seen that …



Sparsity and Redundancy are used for denoising of signals/images

How ?

By finding the sparsest representation and using it to recover the clean signal

Can we do better?

Today we have shown that averaging several sparse representations for a signal lead to better denoising, as it approximates the MMSE estimator.

More on these (including the slides and the relevant papers) can be found in http://www.cs.technion.ac.il/~elad