ITERATED SRINKAGE ALGORITHM FOR BASIS PURSUIT MINIMIZATION*

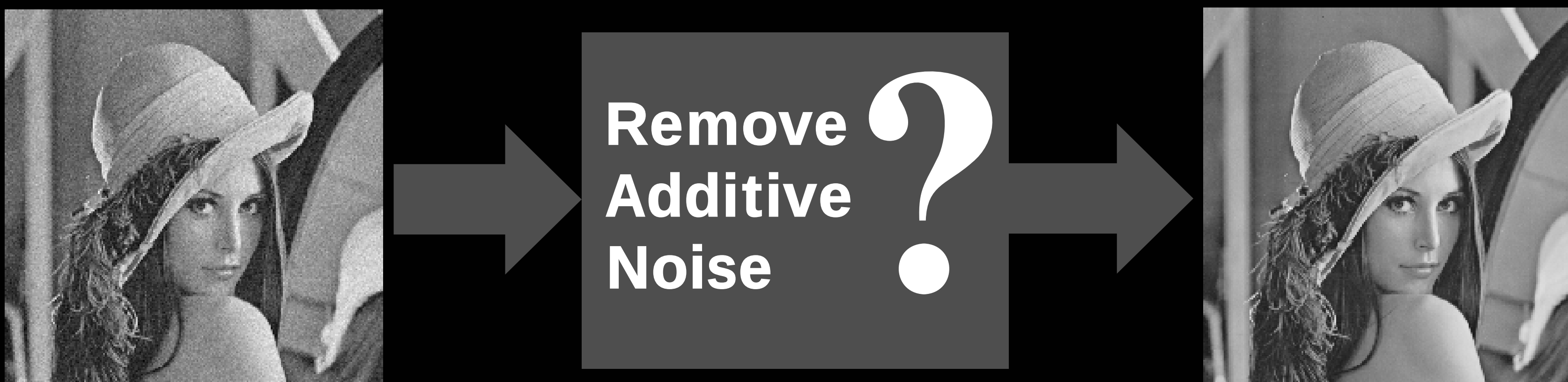
Michael Elad
The Computer Science Department
The Technion – Israel Institute of technology
Haifa 32000, Israel

SIAM Conference on Imaging Science
May 15-17, 2005 – Minneapolis, Minnesota
Sparse Representations – Theory and Applications in Image Processing

* Joint work with Michael Zibulevsky and Boaz Matalon

Noise Removal

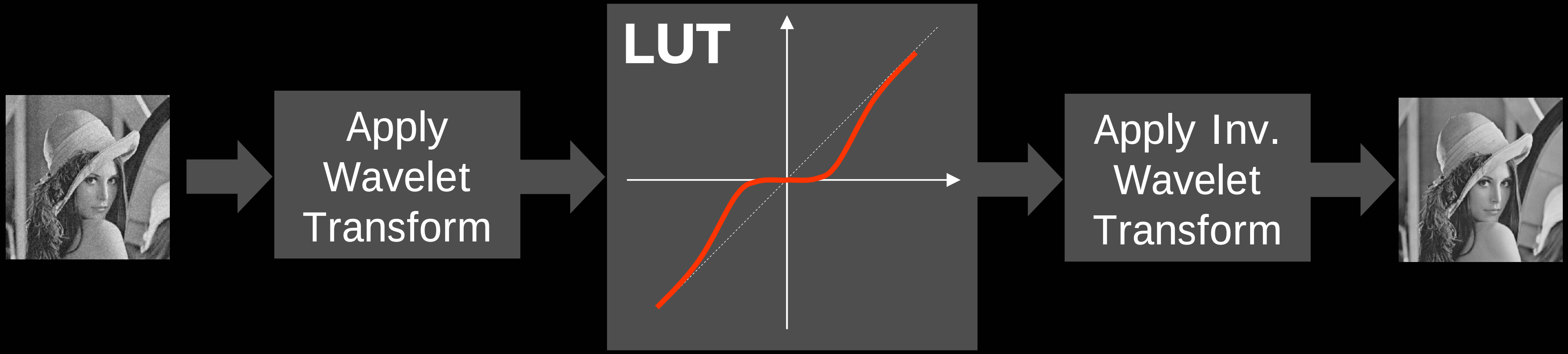
Our story begins with signal/image denoising ...



- 100 years of activity – numerous algorithms.
- Considered Directions include: PDE, statistical estimators, adaptive filters, inverse problems & regularization, example-based restoration, sparse representations, ...

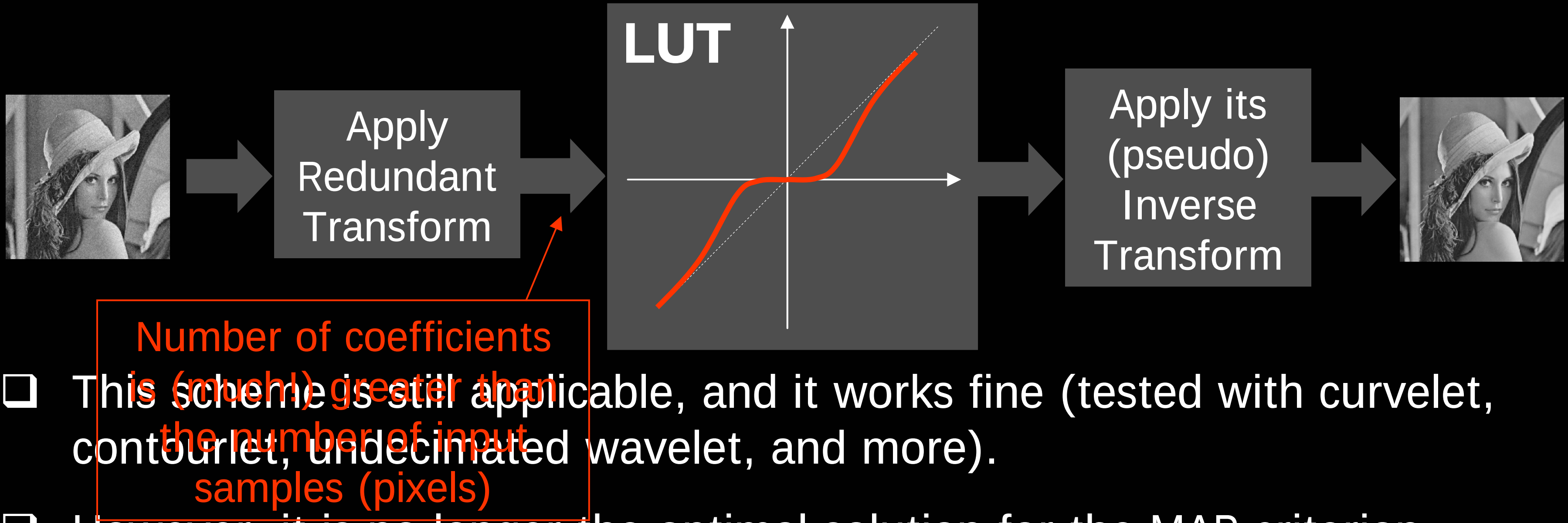
Shrinkage For Denoising

- Shrinkage is a simple yet effective sparsity-based denoising algorithm [Donoho & Johnstone, 1993].
- Justification 1: minimax near-optimal over the Besov (smoothness) signal space (complicated!!!!).



- Justification 2: Bayesian (MAP) optimal [Simoncelli & Adelson 1996, Moulin & Liu 1999].
- In both justifications, an additive Gaussian white noise and a unitary transform are crucial assumptions for the optimality claims.

Redundant Transforms?



- This scheme is still applicable, and it works fine (tested with curvelet, contourlet, undecimated wavelet, and more).
- However, it is no longer the optimal solution for the MAP criterion.

WE SHOW THAT THE ABOVE SHRINKAGE METHOD IS THE FIRST ITERATION IN A VERY EFFECTIVE AND SIMPLE ALGORITHM THAT MINIMIZES THE BASIS PURSUIT, AND AS SUCH, IT IS A NEW PURSUIT TECHNIQUE.

Agenda

1. **Bayesian Point of View – a Unitary Transform**

   Optimality of shrinkage

2. What About Redundant Representation?

   Is shrinkage is relevant? Why? How?

3. Conclusions

Thomas Bayes
1702 - 1761

The MAP Approach

Minimize the following function with respect to $\underline{x}$:

Log-Likelihood term

Prior or regularization

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

Unknown to be recovered

Given measurements

Image Prior?

During the past several decades we have made all sort of guesses about the prior $\Pr(\underline{x})$:

$$\Pr(\underline{x}) = \lambda \|\underline{x}\|_2^2$$

Energy

$$\Pr(\underline{x}) = \lambda \|\mathbf{L}\underline{x}\|_2^2$$

Smoothness

$$\Pr(\underline{x}) = \lambda \|\mathbf{L}\underline{x}\|_{\mathbf{W}}^2$$

Adapt+ Smooth

$$\Pr(\underline{x}) = \lambda \rho\{\mathbf{L}\underline{x}\}$$

Robust Statistics

$$\Pr(\underline{x}) = \lambda \|\nabla \underline{x}\|_1$$

Total-Variation

$$\Pr(\underline{x}) = \lambda \|\mathbf{W}\underline{x}\|_1 \qquad \Pr(\underline{x}) = \lambda \|\mathbf{T}\underline{x}\|_1$$

Today's Focus

- Mumford & Shah formulation,
- Compression algorithms as priors,
- …

(Unitary) Wavelet Sparsity

$$f(\underline{x}) = \frac{1}{2}\left\|\underline{x}-\underline{y}\right\|_2^2 + \lambda \cdot \left\|\mathbf{W}\underline{x}\right\|_1$$

Define

$$\underline{\hat{x}} = \mathbf{W}\underline{x}$$

$$\underline{x} = \mathbf{W}^H\underline{\hat{x}}$$

$$f(\underline{\hat{x}}) = \frac{1}{2}\left\|\mathbf{W}^H(\underline{\hat{x}}-\underline{\hat{y}})\right\|_2^2 + \lambda \cdot \left\|\underline{\hat{x}}\right\|_1$$

L_2 is unitarily invariant

$$f(\underline{x}) = \frac{1}{2}\left\|\underline{\hat{x}}-\underline{\hat{y}}\right\|_2^2 + \lambda \cdot \left\|\underline{\hat{x}}\right\|_1$$

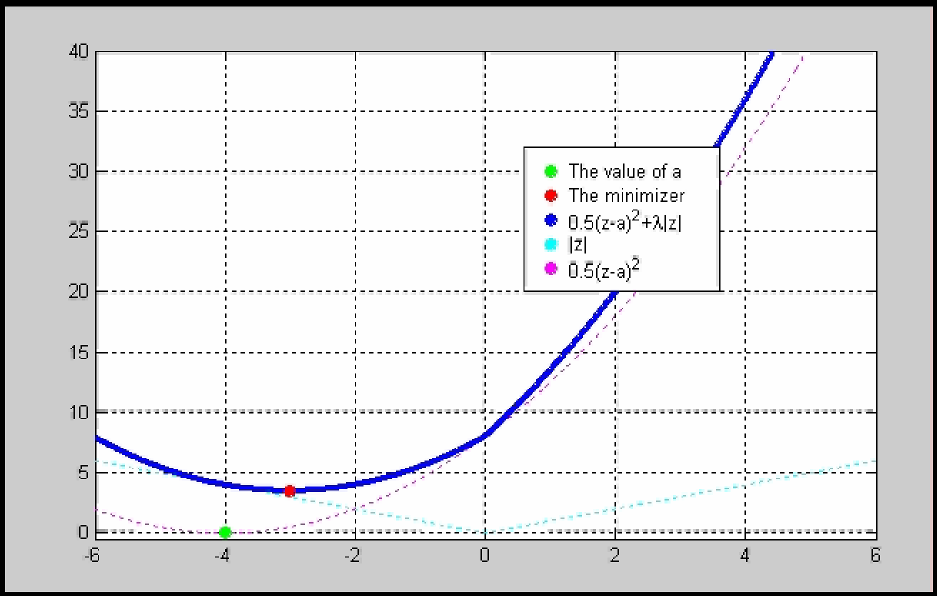
$$= \boxed{\sum_k \left[\frac{1}{2}\left(\hat{x}_k - \hat{y}_k\right)^2 + \lambda\left|\hat{x}_k\right|\right]}$$

We got a separable set of 1D optimization problems

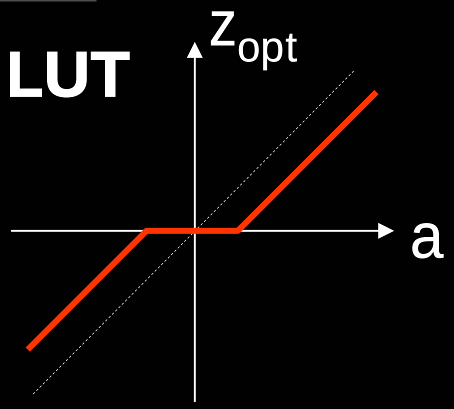
Why Shrinkage?

Want to minimize this 1-D function with respect to z

$$f(z) = \frac{1}{2}(z-a)^2 + \lambda|z|$$



LUT



$$z_{opt} = \begin{cases} a-\lambda & a \geq \lambda \\ 0 & |a| < \lambda \\ a+\lambda & a \leq -\lambda \end{cases}$$

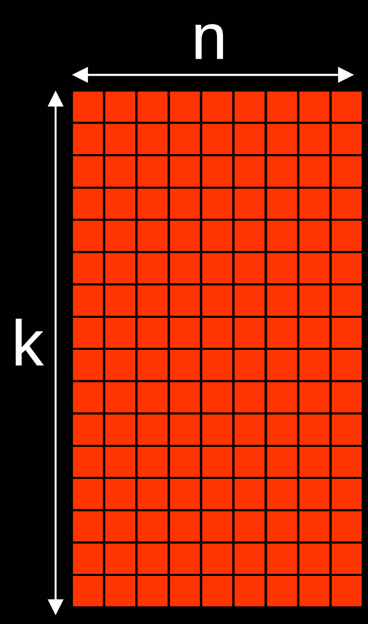
A LUT can be built for any other robust function (replacing the |z|), including non-convex ones (e.g., L_0 norm)!!

Agenda

1. Bayesian Point of View – a Unitary Transform
   Optimality of shrinkage

2. **What About Redundant Representation?**
   Is shrinkage is relevant? Why? How?

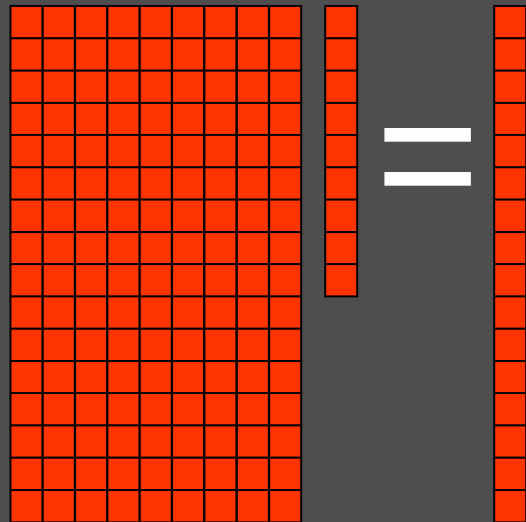
3. Conclusions



$$\mathbf{T} \in \Re^{k \times n}$$

An Overcomplete Transform

$\mathbf{T}\underline{x} =$



$= \underline{\alpha}$

$$f(\underline{x}) = \frac{1}{2}\left\|\underline{x}-\underline{y}\right\|_2^2 + \lambda \cdot \left\|\mathbf{T}\underline{x}\right\|_1$$

Redundant transforms are important because they can

(i) Lead to a shift-invariance property,

(ii) Represent images better (because of orientation/scale analysis),

(iii) Enable deeper sparsity (when used in conjunction with the BP).

Analysis versus Synthesis

Analysis Prior:

$$\boxed{f(\underline{x})} = \frac{1}{2}\left\|\underline{x}-\underline{y}\right\|_2^2 + \lambda \cdot \left\|\mathbf{T}\underline{x}\right\|_1$$

Define

$$\underline{\alpha} = \mathbf{T}\underline{x}$$

$$\underline{x} = \mathbf{T}^{+}\underline{\alpha}$$

$$\tilde{f}(\underline{\alpha}) = \frac{1}{2}\left\|\mathbf{T}^{+}\underline{\alpha}-\underline{y}\right\|_2^2 + \lambda \cdot \left\|\underline{\alpha}\right\|_1$$

Synthesis Prior:

$$\boxed{\tilde{f}(\underline{\alpha})} = \frac{1}{2}\left\|\mathbf{D}\underline{\alpha}-\underline{y}\right\|_2^2 + \lambda \cdot \left\|\underline{\alpha}\right\|_1$$

Basis Pursuit

However

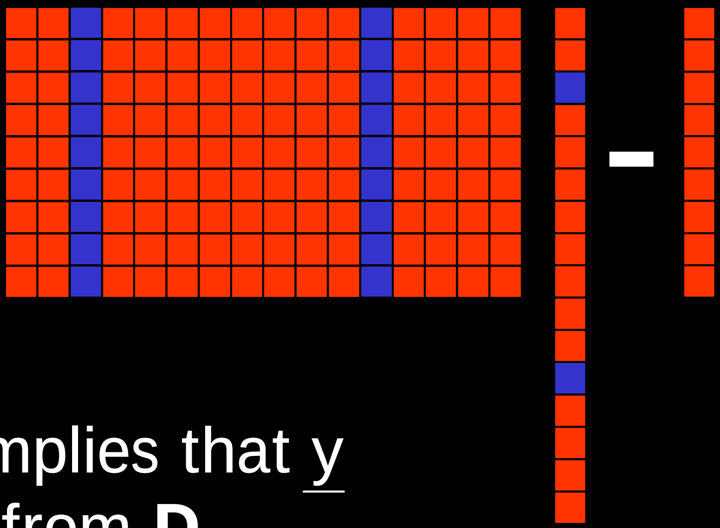
$$\mathbf{D} \cdot \underset{\underline{\alpha}=\mathbf{TT}^{+}\underline{\alpha}}{\text{Arg min}}\ \tilde{f}(\underline{\alpha}) = \underset{\underline{x}}{\text{Arg min}}\ f(\underline{x})$$

!

Basis Pursuit As Objective

Our Objective: $\tilde{f}(\underline{\alpha}) = \frac{1}{2}\left\|\boxed{\mathbf{D}\underline{\alpha} - \underline{y}}\right\|_2^2 + \lambda \cdot \|\underline{\alpha}\|_1$

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



Getting a sparse solution implies that $\underline{y}$ is composed of few atoms from **D**

Sequential Coordinate Descent

Our objective

$$\tilde{f}(\underline{\alpha}) = \frac{1}{2}\left\| \mathbf{D}\underline{\alpha} - \underline{y} \right\|_2^2 + \lambda \cdot \left\| \underline{\alpha} \right\|_1$$

- The unknown, $\underline{\alpha}$, has k entries.
- How about optimizing w.r.t. each of them sequentially?
- The objective per each becomes

$$\tilde{f}(z) = \frac{1}{2}\left\| z\underline{d}_j - \underline{\tilde{y}} \right\|_2^2 + \lambda |z|$$

Set j=1

Fix all entries of $\underline{\alpha}$ apart from the j-th one

Optimize with respect to α_j

j=j+1 mod k

We Get Sequential Shrinkage

BEFORE: We had this 1-D function to minimize

$$f(z) = \frac{1}{2}(z-a)^2 + \lambda|z|$$

and the solution was

$$z_{opt} = \mathbf{S}\{a, \lambda\} = \begin{cases} a - \lambda & a \geq \lambda \\ 0 & |a| < \lambda \\ a + \lambda & a \leq -\lambda \end{cases}$$

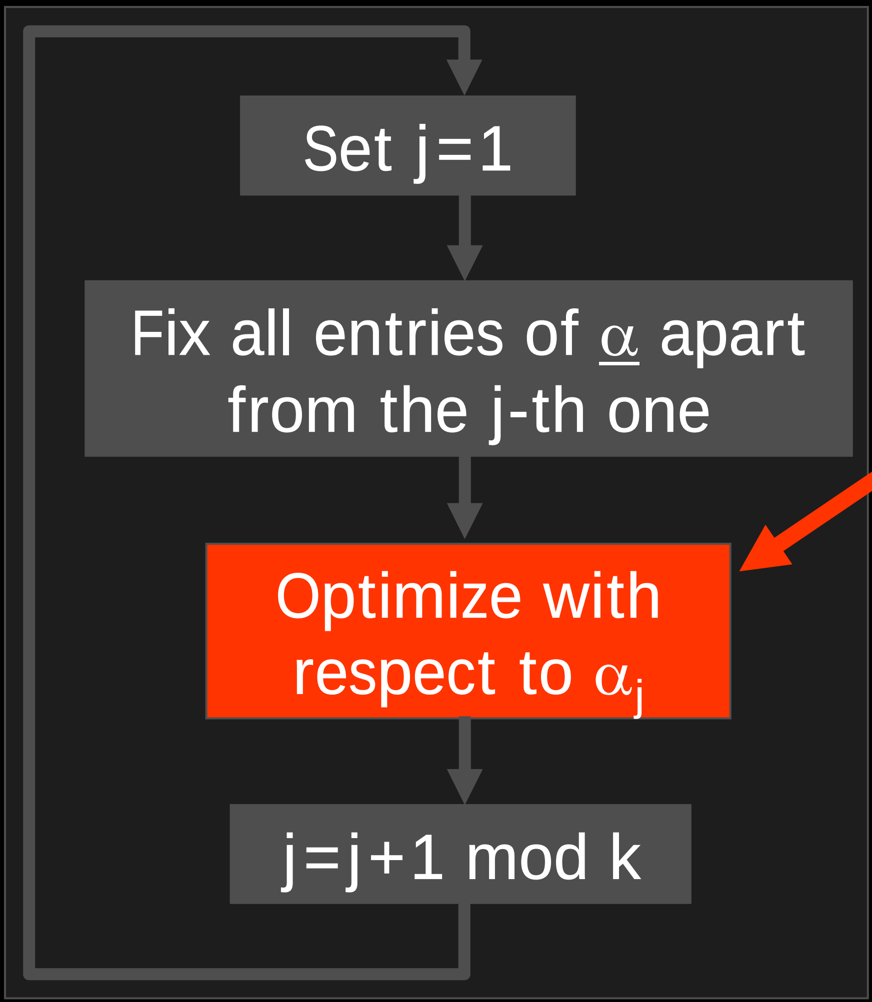
NOW: Our 1-D objective is

$$\tilde{f}(z) = \frac{1}{2}\left\| z\underline{d}_j - \underline{\tilde{y}} \right\|_2^2 + \lambda|z|$$

and the solution now is

$$z_{opt} = \mathbf{S}\left\{ \frac{\underline{d}_j^H \underline{\tilde{y}}}{\left\|\underline{d}_j\right\|_2^2}, \frac{\lambda}{\left\|\underline{d}_j\right\|_2^2} \right\}$$

Sequential? Not Good!!



$$\tilde{\underline{y}} = \underline{y} - \mathbf{D}\underline{\alpha} + \underline{d}_j \alpha_j \quad \text{and}$$

$$\alpha_j^{opt} = \frac{1}{\left\|\underline{d}_j\right\|_2^2} \cdot \mathbf{S}\left\{\underline{d}_j^H \tilde{\underline{y}}, \lambda\right\}$$

- This method requires drawing one column at a time from **D**.
- In most transforms this is not comfortable at all !!!
- This also means that MP and its variants are inadequate.

How About Parallel Shrinkage?

Our objective

$$\tilde{f}(\underline{\alpha}) = \frac{1}{2}\left\| \mathbf{D}\underline{\alpha} - \underline{y} \right\|_2^2 + \lambda \cdot \left\| \underline{\alpha} \right\|_1$$

- Assume a current solution $\underline{\alpha}_n$.
- Using the previous method, we have k descent directions obtained by a simple shrinkage.
- How about taking all of them at once, with a proper relaxation?
- Little bit of math lead to …

For j=1:k

Compute the descent direction per α_j : $\underline{v}_j$.

Update the solution by

$$\underline{\alpha}_{n+1} = \underline{\alpha}_n + \mu \sum_{j=1}^{k} \underline{v}_j$$

Parallel Coordinate Descent (PCD)

Back-projection to the signal domain

Shrinkage operation

The synthesis error

$$\underline{e}_k = \mathbf{S}\left\{\underline{\alpha}_k + \mathbf{Q}\mathbf{D}^H\left(\underline{y} - \mathbf{D}\underline{\alpha}_k\right), \lambda \mathbf{Q} \cdot \underline{1}\right\} - \underline{\alpha}_k$$

$$\underline{\alpha}_{k+1} = \underline{\alpha}_k + \mu_k \underline{e}_k$$

Normalize by a diagonal matrix $\mathbf{Q} = \mathrm{diag}^{-1}\left\{\mathbf{D}^H\mathbf{D}\right\}$

Update by exact line-search

At all stages, the dictionary is applied as a whole, either directly, or via its adjoint

PCD – The First Iteration

Assume: Zero initialization $\underline{\alpha}_0 = \underline{0}$

$\mathbf{D}$ is a tight frame with normalized columns ($\mathbf{Q}=\mathbf{I}$)

Line search is replaced with $\mu_1 = 1$

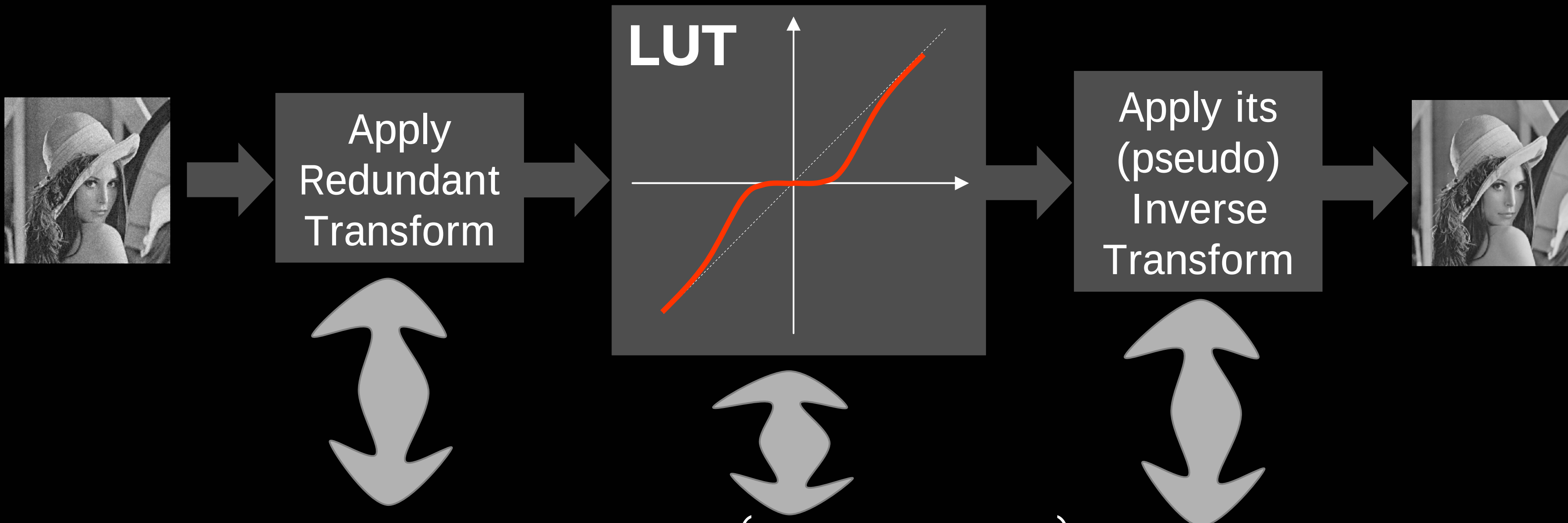
$$\underline{e}_k = \mathbf{S}\left\{\underline{\alpha}_k + \mathbf{Q}\mathbf{D}^H\left(\underline{y} - \mathbf{D}\underline{\alpha}_k\right), \lambda\mathbf{Q}\cdot\underline{1}\right\} - \underline{\alpha}_k$$

$$\underline{\alpha}_{k+1} = \underline{\alpha}_k + \mu_k \underline{e}_k$$

$$\underline{\alpha}_1 = \underline{e}_k = \mathbf{S}\left\{\mathbf{D}^H\underline{y}, \lambda\underline{1}\right\}$$

$$\underline{x}_1 = \mathbf{D}\cdot\mathbf{S}\left\{\mathbf{D}^H\underline{y}, \lambda\underline{1}\right\}$$

Relation to Simple Shrinkage?



$$\mathbf{D}^{H}\underline{y} \qquad \underline{\alpha} = \mathbf{S}\left\{\mathbf{D}^{H}\underline{y}, \lambda \cdot \underline{1}\right\} \qquad \mathbf{D}\underline{\alpha}$$

The first iteration in our algorithm = the intuitive shrinkage !!!

PCD – Convergence Analysis

- We have proven convergence to the global minimizer of the BPDN objective function (with smoothing): $\underline{\alpha}_k \rightarrow \underline{\alpha}^*$
- Approximate asymptotic convergence rate analysis yields:

$$\left[f\left(\underline{\alpha}_{k+1}\right) - f\left(\underline{\alpha}^*\right)\right] \leq \left(\frac{M-m}{M+m}\right)^2 \left[f\left(\underline{\alpha}_k\right) - f\left(\underline{\alpha}^*\right)\right]$$

  where M and m are the largest and smallest eigenvalues of $\mathbf{Q}^{0.5}\mathbf{H}\mathbf{Q}^{0.5}$ respectively (**H** is the Hessian).

- This rate equals that of the Steepest-Descent algorithm, preconditioned by the Hessian's diagonal.
- Substantial further speed-up can be obtained using the subspace optimization algorithm (SESOP) [Zibulevsky and Narkis 2004].

Image Denoising

Minimize

$$\tilde{f}(\underline{\alpha}) = \frac{1}{2}\left\| \mathbf{D}\underline{\alpha} - \underline{y} \right\|_2^2 + \lambda \cdot \left\| \mathbf{M}\underline{\alpha} \right\|_1$$

- The Matrix **M** gives a variance per each coefficient, learned from the corrupted image.
- **D** is the contourlet transform (recent version).
- The length of $\underline{\alpha}$: ~1e+6.
- The Seq. Shrinkage algorithm cannot be simulated for this dim..

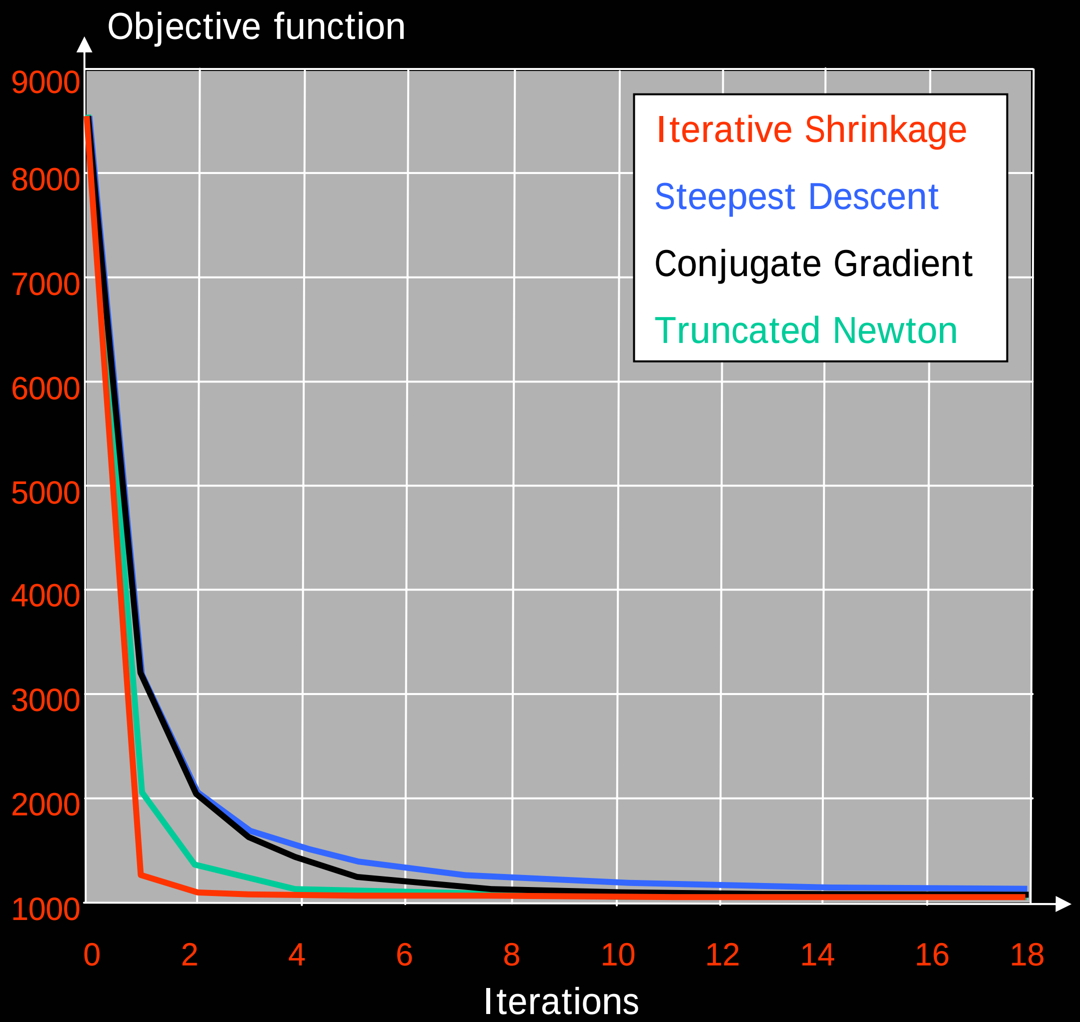


Image Denoising

Minimize

$$\tilde{f}(\underline{\alpha}) = \frac{1}{2}\left\| \mathbf{D}\underline{\alpha} - \underline{y} \right\|_2^2 + \lambda \cdot \left\| \mathbf{M}\underline{\alpha} \right\|_1$$

Evaluate $\left\| \mathbf{D}\hat{\underline{\alpha}} - \underline{x}_{\text{True}} \right\|_2^2$

Even though one iteration of our algorithm is equivalent in complexity to that of the SD, the performance is much better

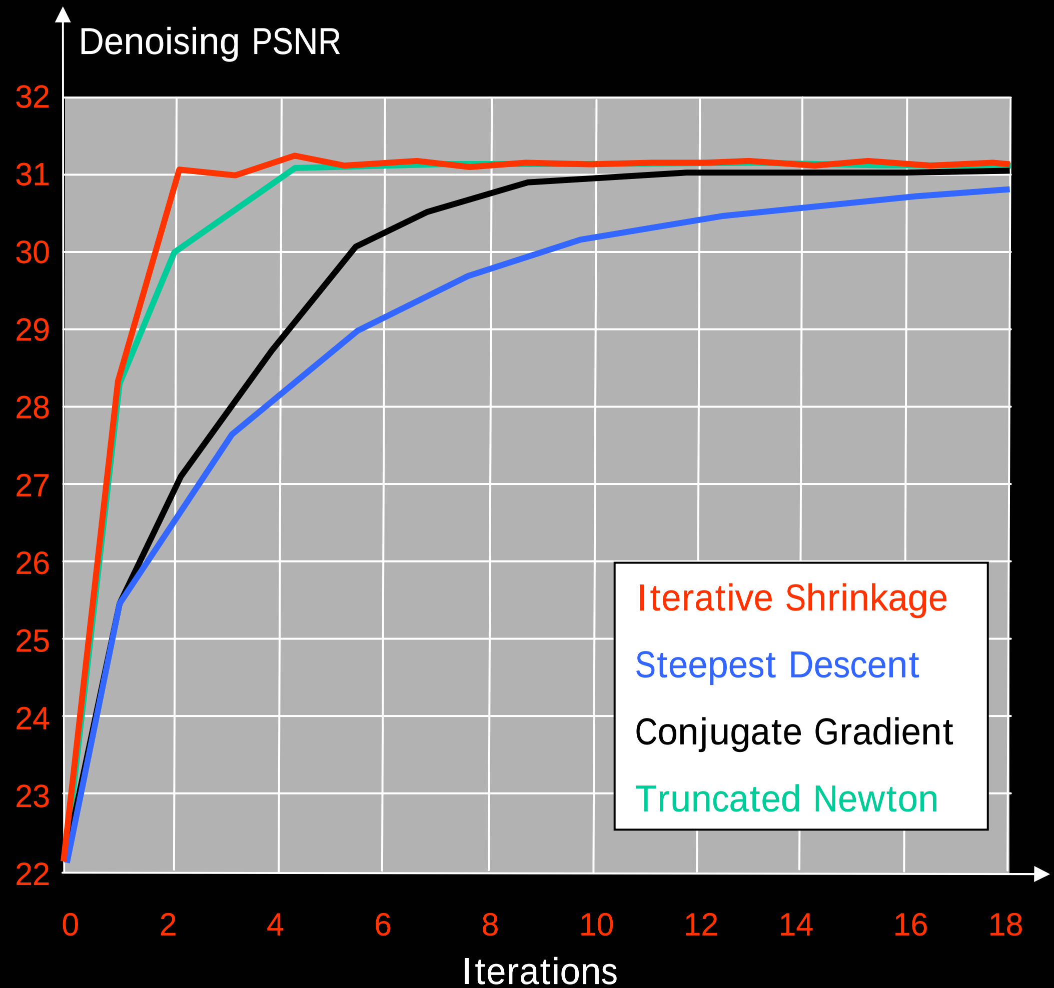
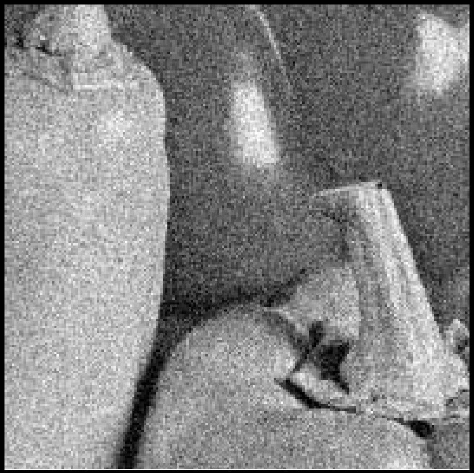


Image Denoising

Original Image



Noisy Image with σ=20

Iterated Shrinkage – First Iteration PSNR=28.30dB

Iterated Shrinkage – second iteration PSNR=31.05dB

Closely Related Work

- Several recent works have devised iterative shrinkage algorithms, each with a different motivation:
  - E-M algorithm for image deblurring - Minimize $\left\|\mathbf{KW}\underline{\alpha}-\underline{y}\right\|_2^2+\lambda\cdot\left\|\underline{\alpha}\right\|_1$ [Figueiredo & Nowak 2003].
  - Surrogate functionals for deblurring as above [Daubechies, Defrise, & De-Mol, 2004] and [Figueiredo & Nowak 2005].
  - PCD minimization for denoising (as shown above) [Elad, 2005].
- While these algorithms are similar, they are in fact different. Our recent work have shown that:
  - PCD gives faster convergence, compared to the surrogate algorithms.
  - All the above methods can be further improved by SESOP, leading to

$$\left[f\left(\underline{\alpha}_{k+1}\right)-f\left(\underline{\alpha}^*\right)\right]\le\left(\frac{\sqrt{M}-\sqrt{m}}{\sqrt{M}+\sqrt{m}}\right)^2\left[f\left(\underline{\alpha}_k\right)-f\left(\underline{\alpha}^*\right)\right]$$

Agenda

1. Bayesian Point of View – a Unitary Transform
   Optimality of shrinkage
2. What About Redundant Representation?
   Is shrinkage is relevant? Why? How?
3. **Conclusions**

Conclusion

Shrinkage is an appealing signal denoising technique

When optimal?

For additive Gaussian noise and unitary transforms

What if the transform is redundant?

Option 1: apply sequential coordinate descent which leads to a sequential shrinkage algorithm

How to avoid the need to extract atoms?

Go Parallel

How?

Compute all the CD directions, and use the average

Getting what?

We obtain an easy to implement iterated shrinkage algorithm (PCD). This algorithm has been thoroughly studied (convergence, rate, comparisons).

THANK YOU!!

These slides
and accompanying papers
can be found in

http://www.cs.technion.ac.il/~elad