Iterative Methods in Statistical Estimation

Mohsen Bayati, David Donoho, Adel Javanmard
Iain Johnstone, Marc Lelarge, Arian Maleki, Andrea Montanari

Stanford University

September 6, 2012

Statistical estimation

$$y = f(\theta; \text{noise})$$

$$\begin{aligned} \theta &\rightarrow \text{Unknown object} \\ y &\rightarrow \text{Observations} \\ f(\,\cdot\,; \text{noise}) &\rightarrow \text{Parametric model} \end{aligned}$$

Problem: Estimate θ from observations y.

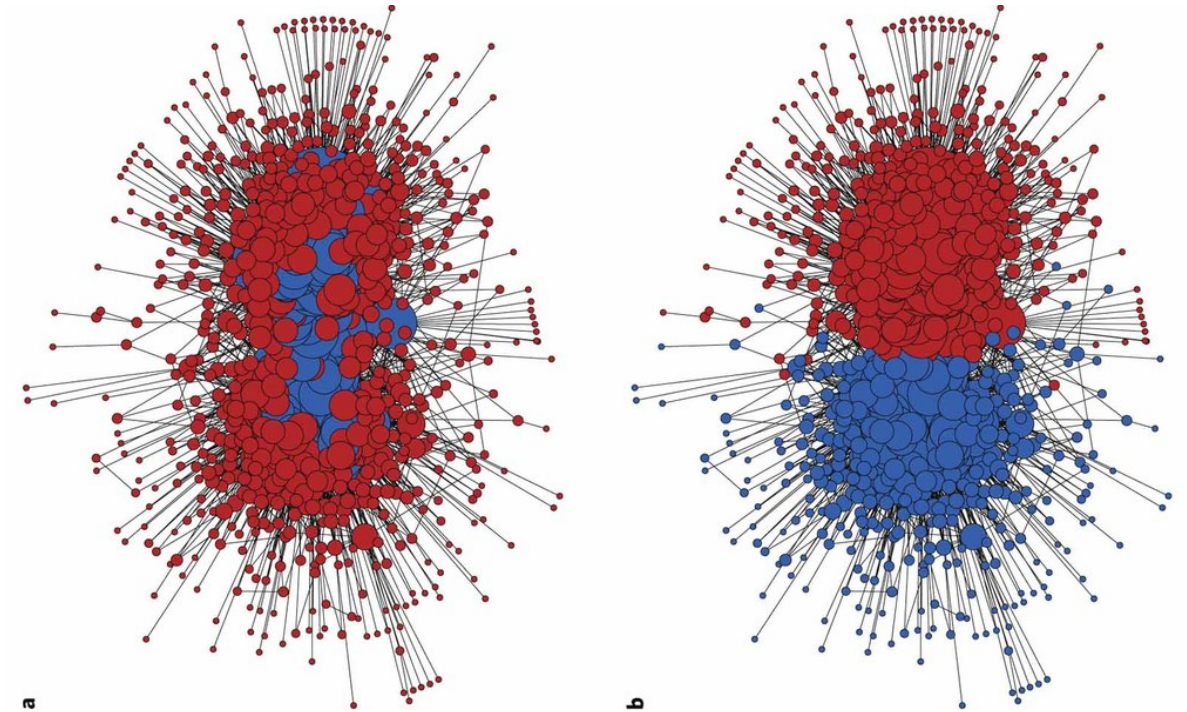
Statistical estimation

$$y = f(\theta; \text{noise})$$

$$\begin{aligned} \theta &\rightarrow \text{Unknown object} \\ y &\rightarrow \text{Observations} \\ f(\,\cdot\,; \text{noise}) &\rightarrow \text{Parametric model} \end{aligned}$$

Problem: Estimate θ from observations y.

Example: Statistical network analysis

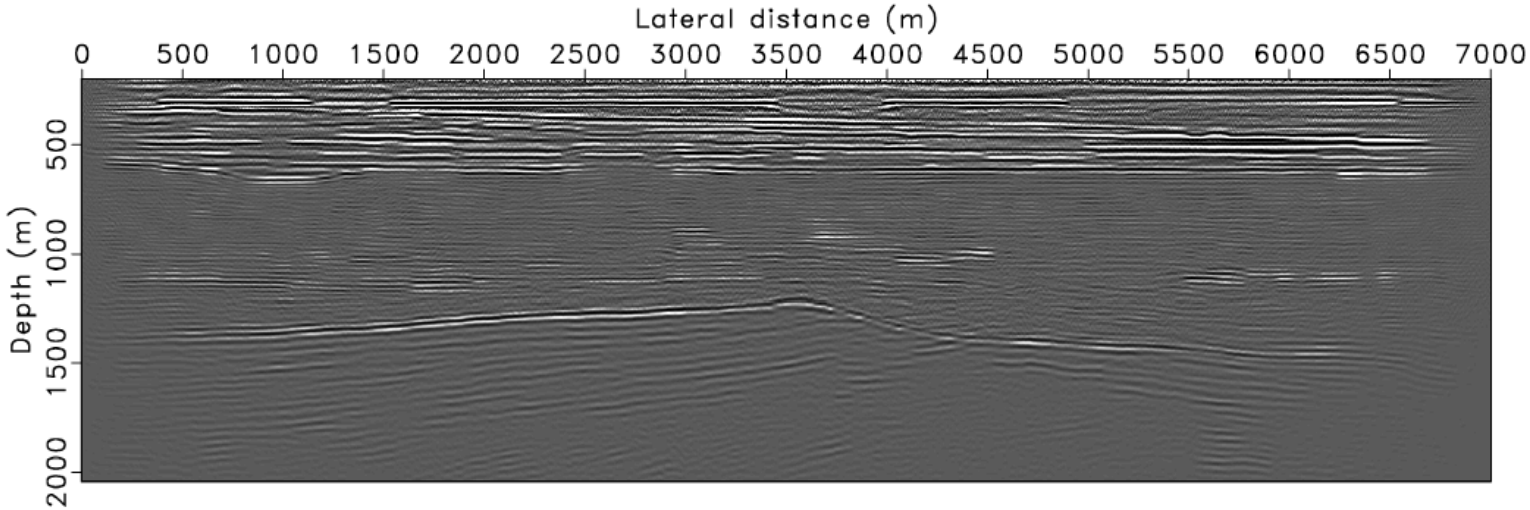


$\theta \quad \rightarrow \quad$ Membership of nodes to 'communities'

$y \quad \rightarrow \quad$ Graph

[Newman, 2012]

Example: Exploration seismology



$\theta \quad \rightarrow \quad$ Density field in the earth

$y \quad \rightarrow \quad$ Seismographic measurements

[Herrmann, 2012]

A broad convergence

- **Statistics**
  [Genomics, . . .]
- **Data mining**
  [Collaborative filtering, Predictive analytics, . . .]
- **Signal processing**
  [Compressive sampling, . . .]
- **Inverse problems**
  [Medical imaging, Seismographic imaging,. . .]

+Data, + Computation, Exploit hidden structure

A broad convergence

- **Statistics**
  [Genomics, . . .]
- **Data mining**
  [Collaborative filtering, Predictive analytics, . . .]
- **Signal processing**
  [Compressive sampling, . . .]
- **Inverse problems**
  [Medical imaging, Seismographic imaging,. . .]

+Data, + Computation, Exploit hidden structure

How should we think about these problems?

How should we think about these problems?

Information theory?

$$\theta \to \text{NOISY CHANNEL} \to y = f(\theta; \text{noise})$$

Fundamental limits, No algorithm

How should we think about these problems?

Information theory?

$$\theta \to \text{NOISY CHANNEL} \to y = f(\theta; \text{noise})$$

Fundamental limits, No algorithm

How should we think about these problems?

Optimization?

$$\text{maximize} \quad \text{Likelihood}(\theta|y) - \text{Complexity}(\theta)$$

Efficient (convex) algorithms, Difficult statistical theory

How should we think about these problems?

Optimization?

$$\text{maximize} \quad \text{Likelihood}(\theta|y) - \text{Complexity}(\theta)$$

Efficient (convex) algorithms, Difficult statistical theory

How should we think about these problems?

Iterative methods?

$$y \to \hat{\theta}^1 \to \hat{\theta}^2 \to \hat{\theta}^3 \to \dots$$

- Each step $\to$ One matrix-vector multiplication
- A few steps (say ≤ 20)
- What can we achieve?

Outline

- An example (algorithm + heuristics)
- A couple of theorems
- Generalizations and open problems

A long example

What type of example?

- Image processing (because they make nice figures)
- Compressed sensing (simpler/cleaner)

What type of example?

- Image processing (because they make nice figures)
- Compressed sensing (simpler/cleaner)

What type of example?

- Image processing (because they make nice figures)
- Compressed sensing (simpler/cleaner)

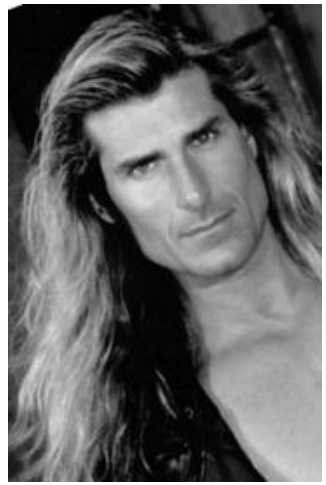
Which image?

Examples appearing in the literature

Lena

Cameraman

Barbara



Fabio

Better someone who is familiar to everybody

Better someone who is familiar to everybody

Suhas

Claude

Rüdi

Better someone who is familiar to everybody

Suhas

Claude

Rüdi

Who's the most handsome?

Better someone who is familiar to everybody

$\theta =$  $\in \mathbb{C}^n$

Unknown object ($n = 512^2 \approx 2.5 \cdot 10^5$)

Noiseless linear measurements

$$y = A\theta = A \cdot$$

Want to reconstruct θ

Noiseless linear measurements

$$y = A\theta = A \cdot$$

Want to reconstruct θ

Measurement structure

$$A = SFR$$

F = Fourier transform

$$S = \begin{bmatrix} 1 & & & & & \\ & 0 & & & & \\ & & 1 & & & \\ & & & 0 & & \\ & & & & 0 & \\ & & & & & 1 \end{bmatrix} = \text{random subsampling matrix (rate } \delta = 0.15)$$

$$R = \begin{bmatrix} +1 & & & & & \\ & -1 & & & & \\ & & -1 & & & \\ & & & +1 & & \\ & & & & +1 & \\ & & & & & -1 \end{bmatrix} = \text{random modulation}$$

$$\to y \in \mathbb{C}^m,\ m = 0.15\, n$$

Measurement structure

$$
\begin{aligned}
A &= \widetilde{F}\, R \\
\widetilde{F} &= \text{subsampled Fourier matrix} \\
R &= \begin{bmatrix} +1 & & & & & \\ & -1 & & & & \\ & & -1 & & & \\ & & & +1 & & \\ & & & & +1 & \\ & & & & & -1 \end{bmatrix} = \text{random modulation}
\end{aligned}
$$

$$\to y \in \mathbb{C}^m,\ m = 0.15\, n$$

Constructing a first estimate

$$y = A\theta$$

Matched filter ($\sim$ pseudoinverse)

$$\begin{aligned} \hat{\theta}^1 &= N^{-1} A^{\dagger} y \\ N &= \text{diag}\left(N_{ii} = \|i\text{-th col of } A\|_2^2\right) \end{aligned}$$

Constructing a first estimate

$$y = A\theta$$

Matched filter

$$\hat{\theta}^1 = \frac{1}{m} A^\dagger y$$

How good is this?

$$
\begin{aligned}
\mathbb{E}\,\hat{\theta}^1 &= \frac{1}{m}\mathbb{E}\{A^{\dagger}y\} \\
&= \frac{1}{m}\mathbb{E}\{A^{\dagger}A\}\theta \\
&= \frac{1}{m}\mathbb{E}\{RF^{\dagger}S\,SFR\}\theta \\
&= \frac{1}{m}\mathbb{E}\{RF\,\delta\mathrm{I}\,F^{\dagger}R\}\theta \\
&= \frac{1}{m}\mathbb{E}\{R\,n\delta\mathrm{I}\,R\}\theta = \frac{n\delta}{m}\theta = \theta
\end{aligned}
$$

Will redefine $A \leftarrow A/\sqrt{m}$

How good is this?

$$
\begin{aligned}
\mathbb{E}\,\hat{\theta}^1 &= \frac{1}{m}\mathbb{E}\{A^{\dagger} y\} \\
&= \frac{1}{m}\mathbb{E}\{A^{\dagger} A\}\theta \\
&= \frac{1}{m}\mathbb{E}\{R F^{\dagger} S\, S F R\}\theta \\
&= \frac{1}{m}\mathbb{E}\{R F\, \delta \mathrm{I}\, F^{\dagger} R\}\theta \\
&= \frac{1}{m}\mathbb{E}\{R\, n\delta \mathrm{I}\, R\}\theta = \frac{n\delta}{m}\theta = \theta
\end{aligned}
$$

Will redefine $A \leftarrow A/\sqrt{m}$

How good is this?

$$
\begin{aligned}
\mathbb{E}\,\hat{\theta}^1 &= \frac{1}{m}\mathbb{E}\{A^{\dagger}y\} \\
&= \frac{1}{m}\mathbb{E}\{A^{\dagger}A\}\theta \\
&= \frac{1}{m}\mathbb{E}\{RF^{\dagger}S\,SFR\}\theta \\
&= \frac{1}{m}\mathbb{E}\{RF\,\delta\mathrm{I}\,F^{\dagger}R\}\theta \\
&= \frac{1}{m}\mathbb{E}\{R\,n\delta\mathrm{I}\,R\}\theta = \frac{n\delta}{m}\theta = \theta
\end{aligned}
$$

Will redefine $A \leftarrow A/\sqrt{m}$

How good is this?

$$
\begin{aligned}
\mathbb{E}\,\hat{\theta}^1 &= \frac{1}{m}\mathbb{E}\{A^{\dagger}y\} \\
&= \frac{1}{m}\mathbb{E}\{A^{\dagger}A\}\theta \\
&= \frac{1}{m}\mathbb{E}\{RF^{\dagger}S\,SFR\}\theta \\
&= \frac{1}{m}\mathbb{E}\{RF\,\delta\mathrm{I}\,F^{\dagger}R\}\theta \\
&= \frac{1}{m}\mathbb{E}\{R\,n\delta\mathrm{I}\,R\}\theta = \frac{n\delta}{m}\theta = \theta
\end{aligned}
$$

Will redefine $A \leftarrow A/\sqrt{m}$

How good is this?

$$
\begin{aligned}
\mathbb{E}\,\hat{\theta}^1 &= \frac{1}{m}\mathbb{E}\{A^{\dagger} y\} \\
&= \frac{1}{m}\mathbb{E}\{A^{\dagger} A\}\theta \\
&= \frac{1}{m}\mathbb{E}\{R F^{\dagger} S\, S F R\}\theta \\
&= \frac{1}{m}\mathbb{E}\{R F\, \delta \mathrm{I}\, F^{\dagger} R\}\theta \\
&= \frac{1}{m}\mathbb{E}\{R\, n\delta \mathrm{I}\, R\}\theta = \frac{n\delta}{m}\theta = \theta
\end{aligned}
$$

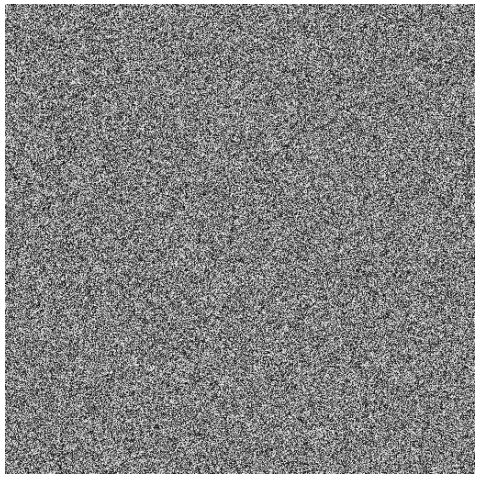
Will redefine $A \leftarrow A/\sqrt{m}$

How good is this?

$$
\begin{aligned}
\mathbb{E}\,\hat{\theta}^1 &= \frac{1}{m}\mathbb{E}\{A^{\dagger}y\} \\
&= \frac{1}{m}\mathbb{E}\{A^{\dagger}A\}\theta \\
&= \frac{1}{m}\mathbb{E}\{RF^{\dagger}S\,SFR\}\theta \\
&= \frac{1}{m}\mathbb{E}\{RF\,\delta\mathrm{I}\,F^{\dagger}R\}\theta \\
&= \frac{1}{m}\mathbb{E}\{R\,n\delta\mathrm{I}\,R\}\theta = \frac{n\delta}{m}\theta = \theta
\end{aligned}
$$

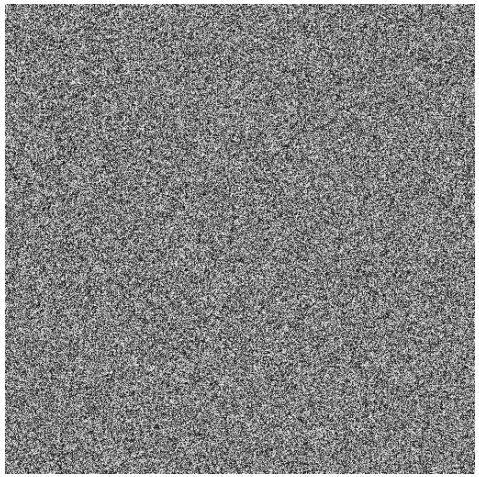
Will redefine $A \leftarrow A/\sqrt{m}$

Check it out

$\hat{\theta}^1 = A^\dagger y =$  $\theta =$ 

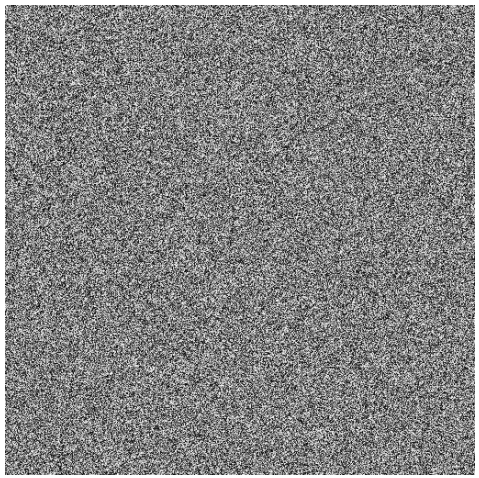
Does not look that good!

Check it out

$\hat{\theta}^1 = A^\dagger y =$  $\theta =$ 

Does not look that good!

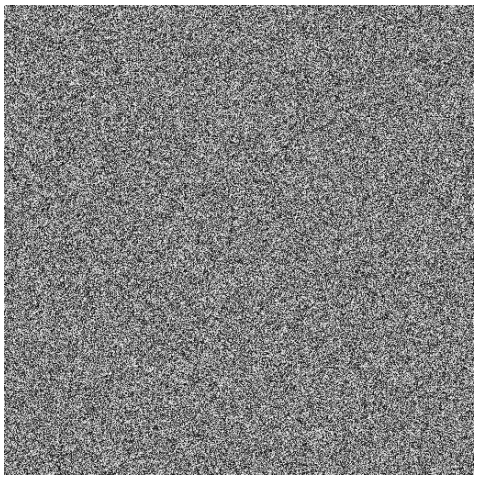
Idea

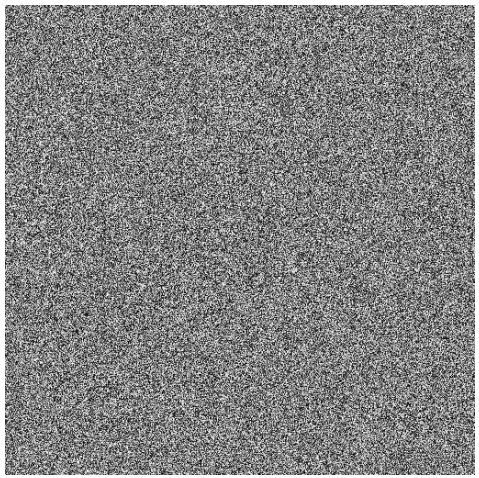


$=$

$+$ 'noise'

Idea





How big is the 'noise'? (take wlog $R = \mathrm{I}$)

$$\hat{\theta}^1 - \theta = (A^{\dagger}A - \mathrm{I})\theta = \Big(\frac{1}{m}F^{\dagger}SSF - \mathrm{I}\Big)\theta = \frac{1}{n\delta}F^{\dagger}(S - \mathbb{E}S)F\theta$$

Hence

$$\mathbb{E}\{\|\hat{\theta}^1 - \theta\|_2^2\} = \frac{1}{\delta^2}\,\mathbb{E}\{\theta^{\dagger}F^{\dagger}(S - \mathbb{E}S)^2F\theta\} = \frac{1}{\delta^2}\,\delta(1-\delta)\,\|F\theta\|_2^2$$
$$= \frac{1-\delta}{\delta}\,\|\theta\|_2^2$$

How big is the 'noise'? (take wlog $R = \mathrm{I}$)

$$\hat{\theta}^1 - \theta = (A^{\dagger}A - \mathrm{I})\theta = \Big(\frac{1}{m}F^{\dagger}SSF - \mathrm{I}\Big)\theta = \frac{1}{n\delta}F^{\dagger}(S - \mathbb{E}S)F\theta$$

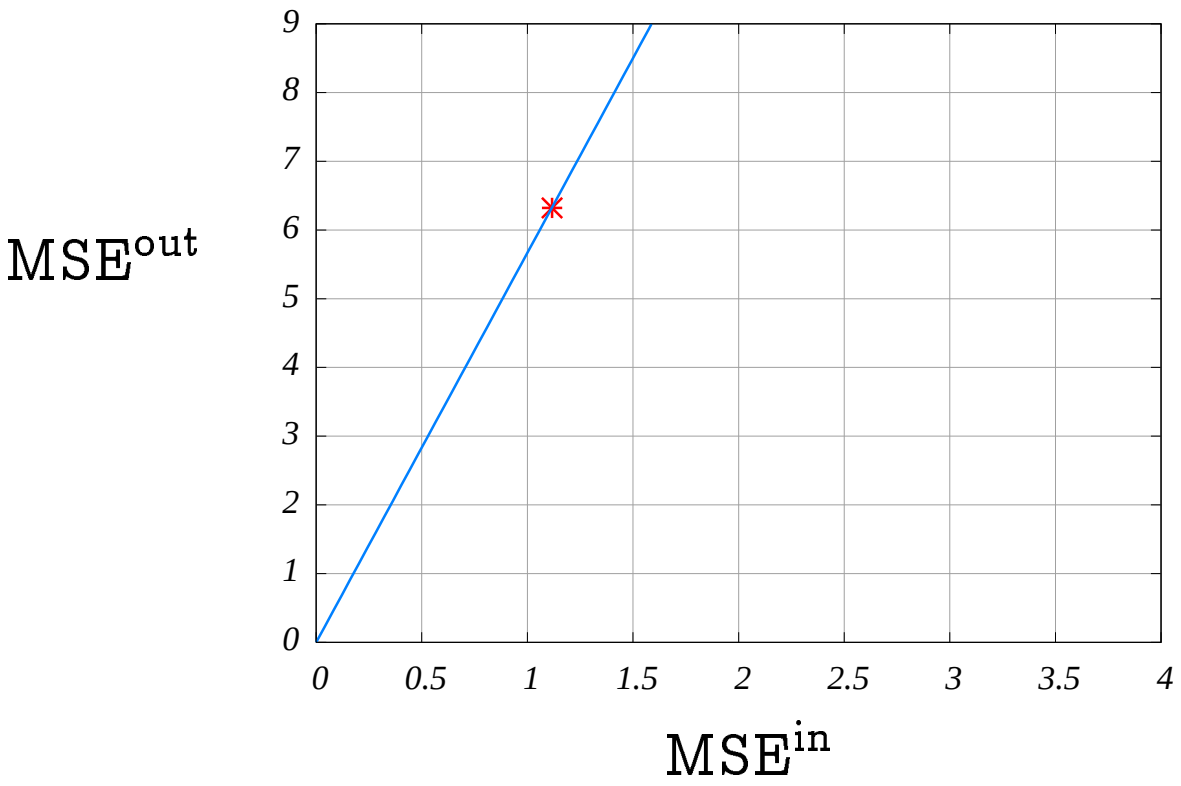
Hence

$$\mathbb{E}\{\|\hat{\theta}^1 - \theta\|_2^2\} = \frac{1}{\delta^2}\,\mathbb{E}\{\theta^{\dagger}F^{\dagger}(S - \mathbb{E}S)^2F\theta\} = \frac{1}{\delta^2}\,\delta(1-\delta)\,\|F\theta\|_2^2$$
$$= \frac{1-\delta}{\delta}\,\|\theta\|_2^2$$

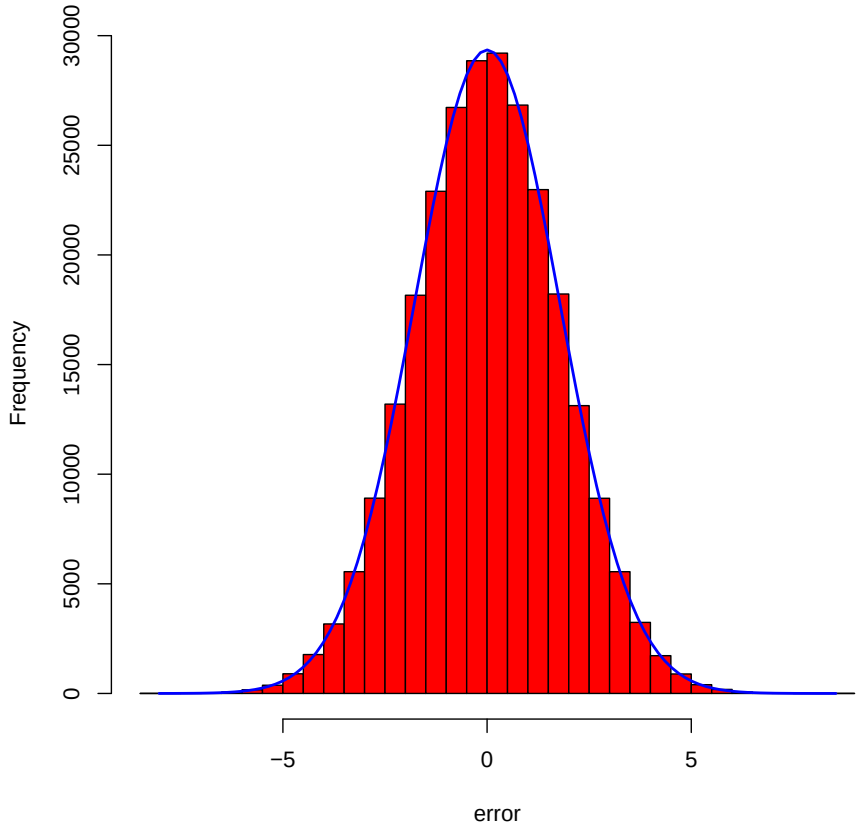
Matched filter blows up noise

$$\mathrm{MSE}^{\mathrm{out}} = \frac{1-\delta}{\delta}\,\mathrm{MSE}^{\mathrm{in}}$$

Let's check



Noise distribution?



Denoising

Statistical estimation with

$$y = f(\theta; \text{noise}) = \theta + \sigma\, z\,, \qquad z_i \sim \mathrm{N}(0,1)$$

Idea: Treat $\hat{\theta}^1$ as effective observations in denoising

Denoising

Statistical estimation with

$$y = f(\theta; \text{noise}) = \theta + \sigma\, z\,, \qquad z_i \sim \mathrm{N}(0,1)$$

Idea: Treat $\hat{\theta}^1$ as effective observations in denoising

Denoising by nonlocal means

$$y = \theta + \sigma z\,,$$

$$\hat{\theta}_i = \frac{\sum_j W(i;j) y_j}{\sum_j W(i;j)}\,,$$

$$W(i;j) = \begin{cases} 1 & \text{if } \|\text{Patch}(i;y) - \text{Patch}(j;y)\|_2^2 \le \tau\,\sigma^2, \\ 0 & \text{otherwise} \end{cases}$$

[Buades, Coll, Morel, 2005]

$\hat{\theta} \equiv \eta(y)$

Denoising by nonlocal means

$$y = \theta + \sigma z\,,$$

$$\hat{\theta}_i = \frac{\sum_j W(i;j) y_j}{\sum_j W(i;j)}\,,$$

$$W(i;j) = \begin{cases} 1 & \text{if } \|\text{Patch}(i;y) - \text{Patch}(j;y)\|_2^2 \le \tau\,\sigma^2, \\ 0 & \text{otherwise} \end{cases}$$

[Buades, Coll, Morel, 2005]

$\hat{\theta} \equiv \eta(y)$

Denoising by nonlocal means

$$y = \theta + \sigma z\,,$$

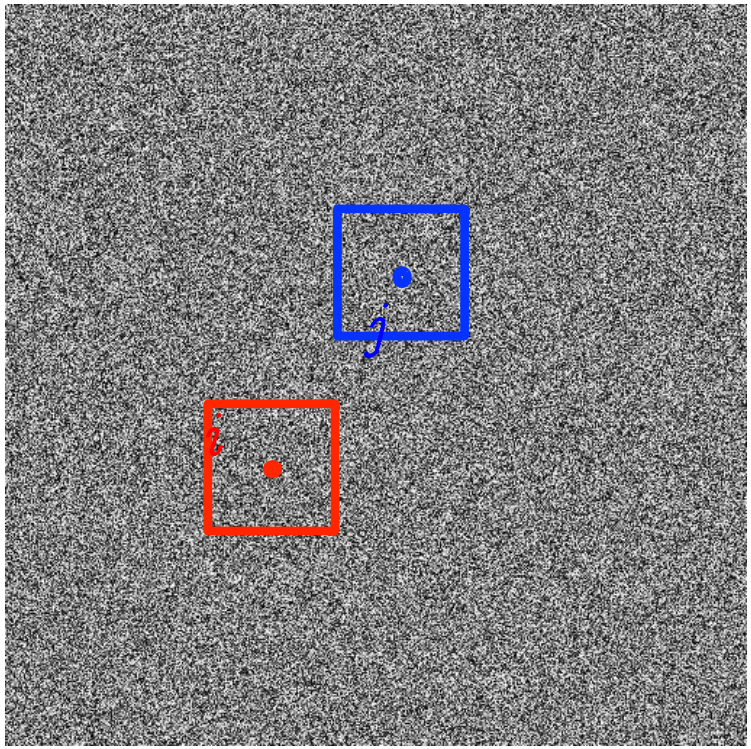
$$\hat{\theta}_i = \frac{\sum_j W(i;j) y_j}{\sum_j W(i;j)}\,,$$

$$W(i;j) = \begin{cases} 1 & \text{if } \|\text{Patch}(i;y) - \text{Patch}(j;y)\|_2^2 \le \tau\,\sigma^2, \\ 0 & \text{otherwise} \end{cases}$$

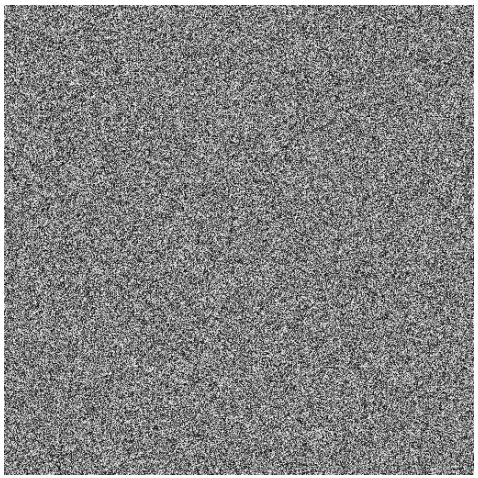
[Buades, Coll, Morel, 2005]

$\hat{\theta} \equiv \eta(y)$

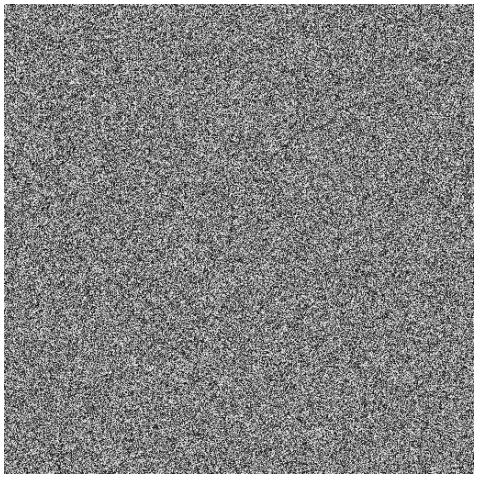
Patches



Will it work?

$$\hat{\theta}^2 = \eta(\hat{\theta}^1) = \eta(A^\dagger y) = \eta\Big($$  $$\Big)$$

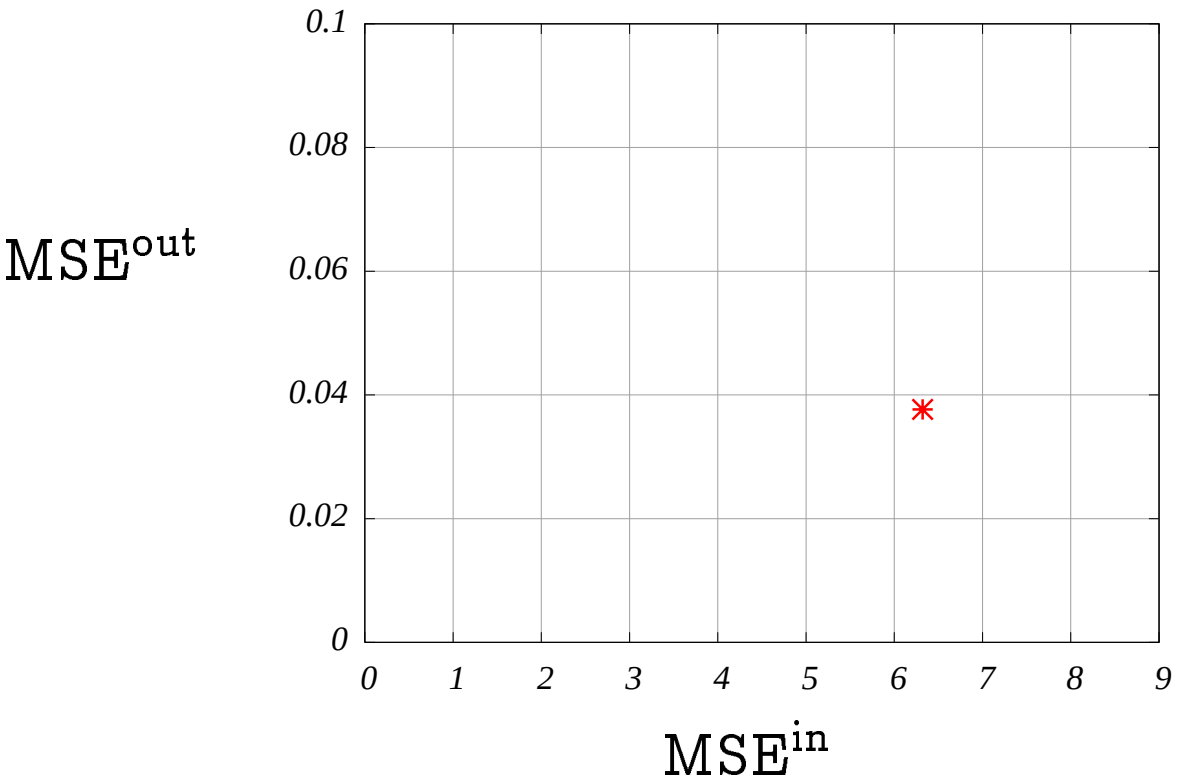
Let's try

$\hat{\theta}^1 = A^\dagger y =$ 

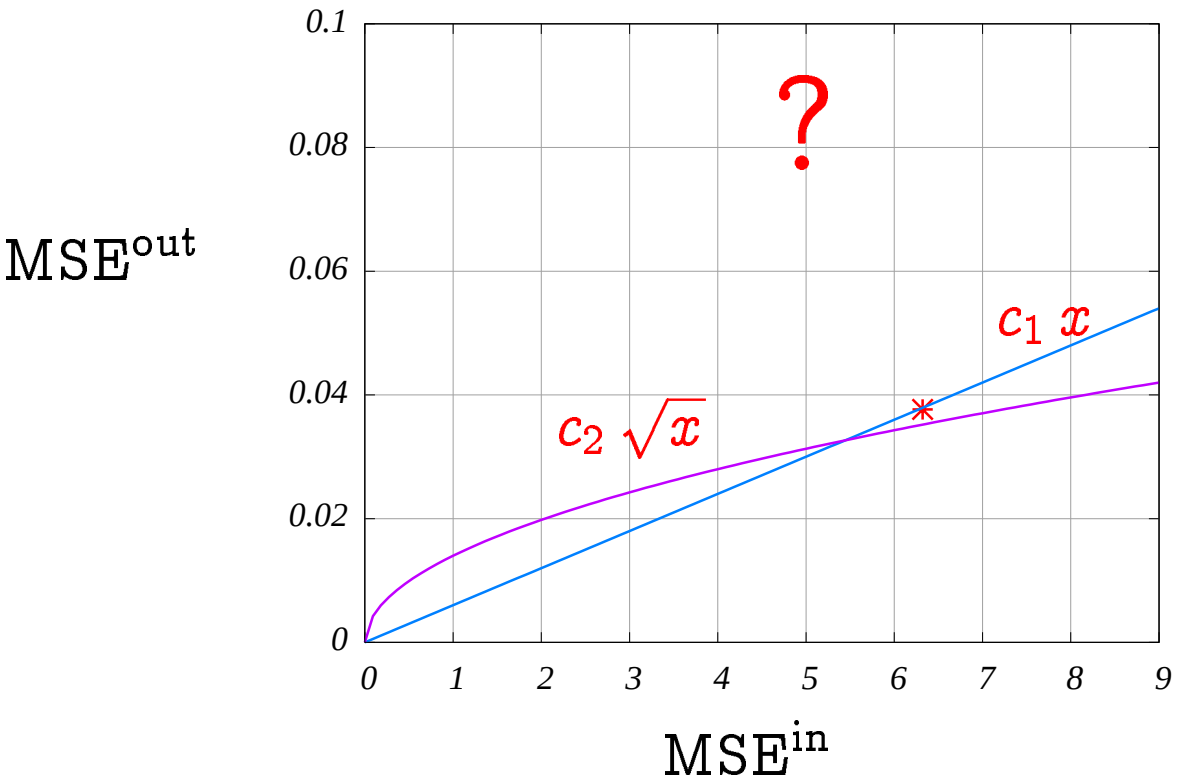
$\hat{\theta}^2 = \eta(A^\dagger y) =$ 

Better than garbage!

How much better?

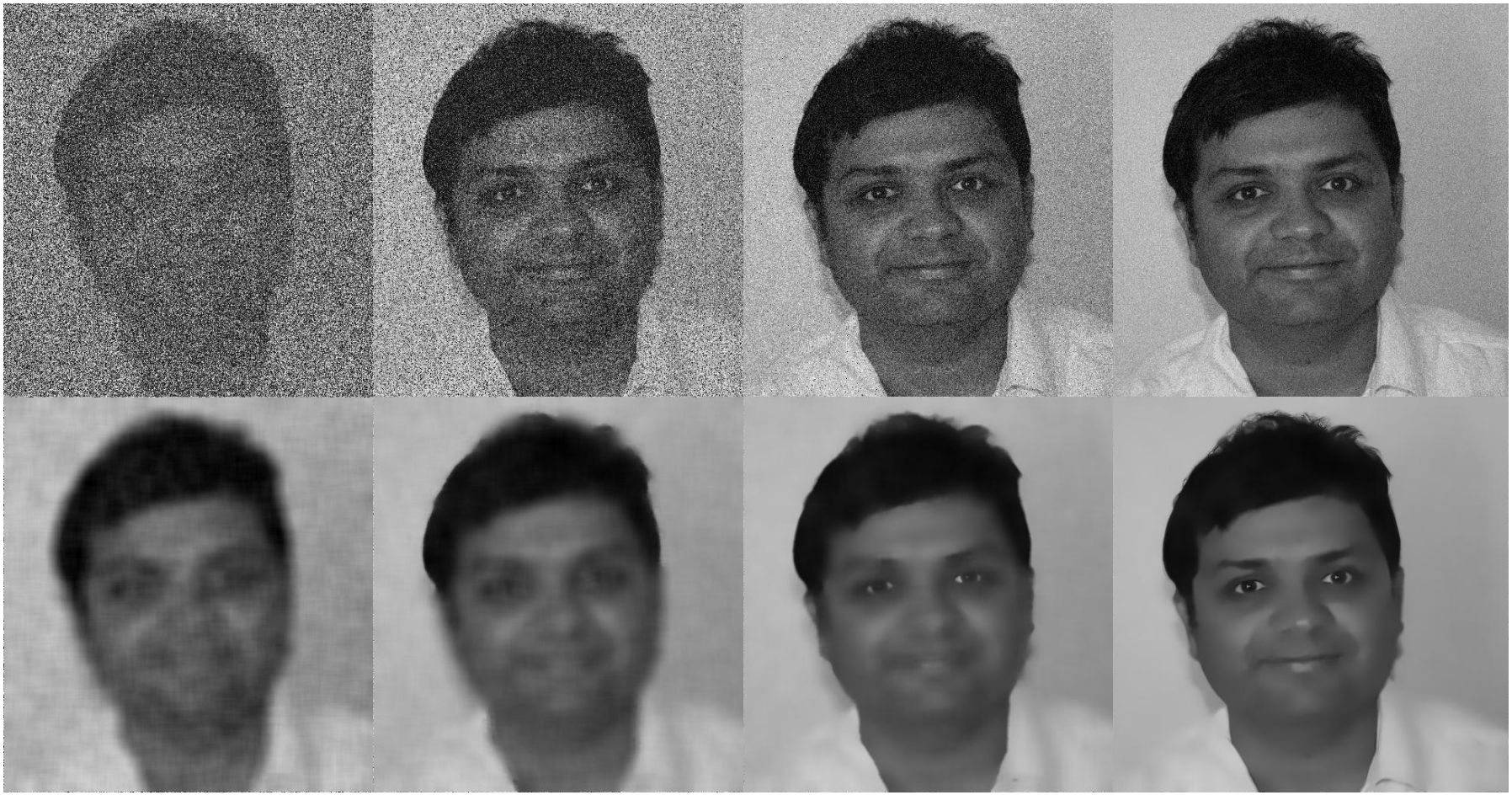


How much better?

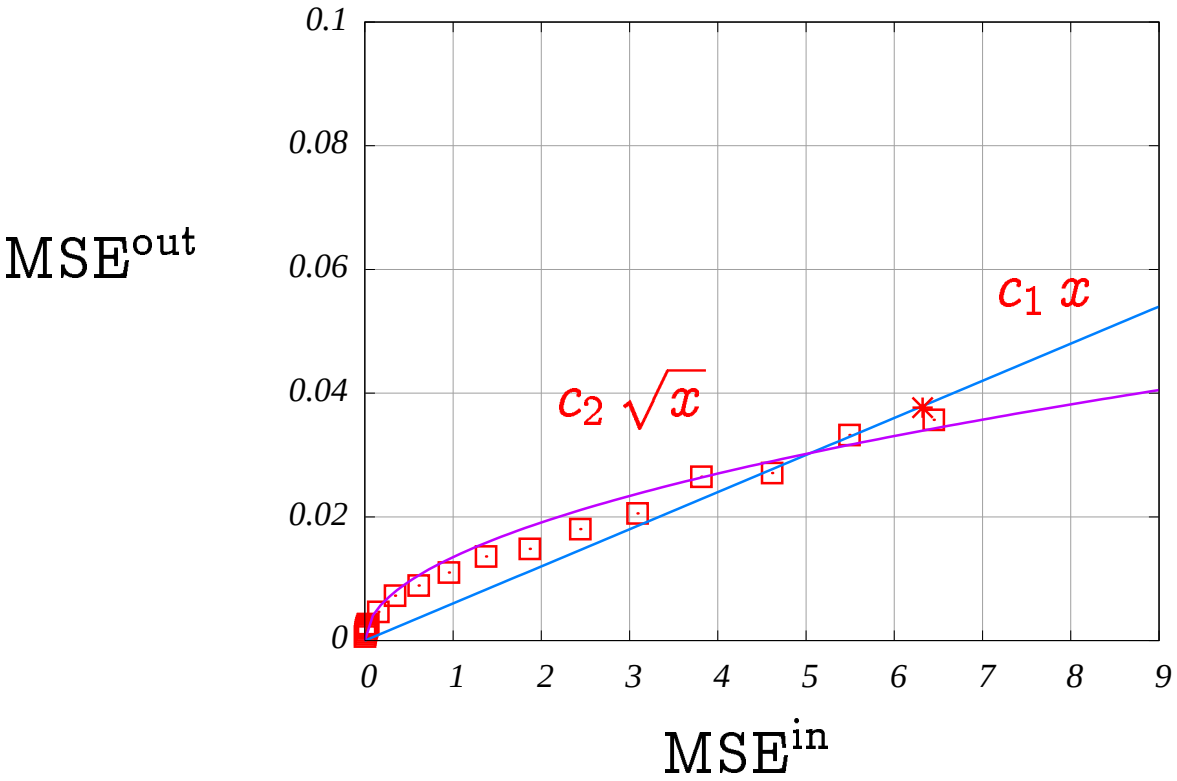


Let us repeat the denoising experiment

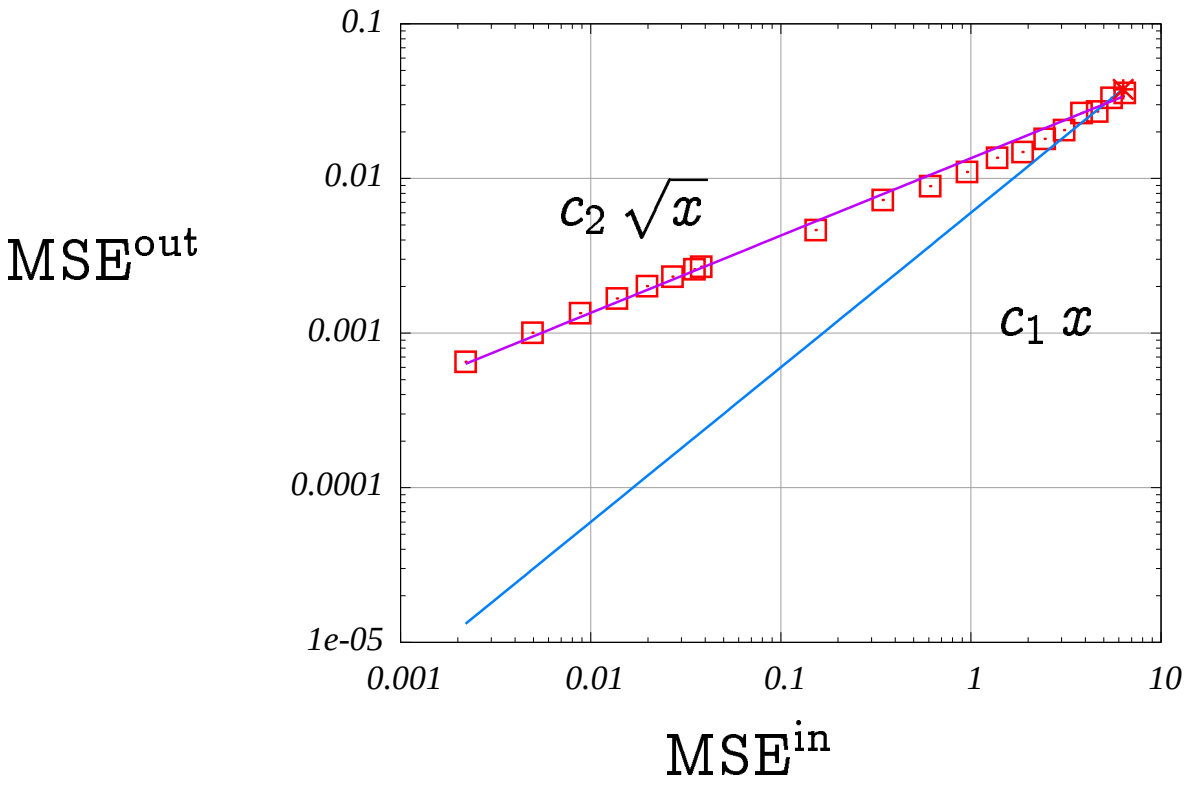
Let us repeat the denoising experiment: $y = \theta + \sigma\, z$



Quantitatively



How much better?



Approximate denoiser characterization

$$\mathrm{MSE}^{\mathrm{out}} = c\sqrt{\mathrm{MSE}^{\mathrm{in}}}$$

(enough for our purposes)

Theorem (Maleki, Baraniuk, Narayan, 2012, informal)

The minimax risk of nonlinear means satisfies

$$\inf_{\text{tuning params}} \sup_{\text{images}} \mathrm{MSE} = \sigma\,\mathrm{Poly}(\log\sigma)$$

Approximate denoiser characterization

$$\mathrm{MSE}^{\mathrm{out}} = c\sqrt{\mathrm{MSE}^{\mathrm{in}}}$$

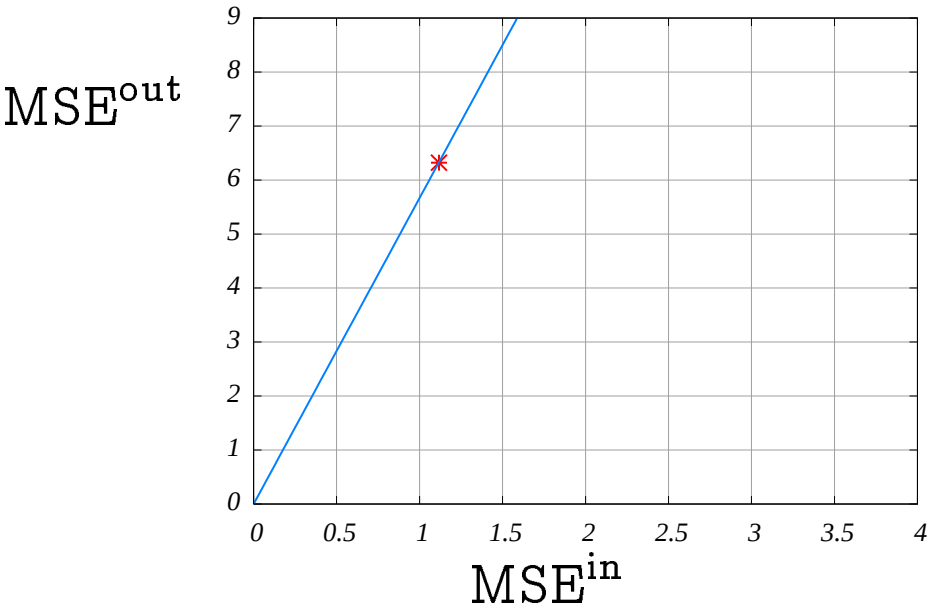
(enough for our purposes)

Theorem (Maleki, Baraniuk, Narayan, 2012, informal)

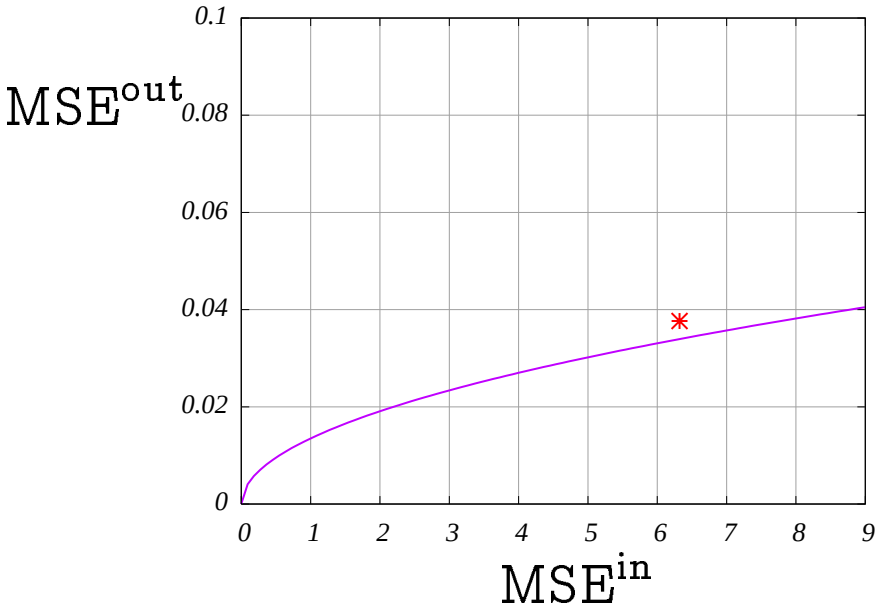
The minimax risk of nonlinear means satisfies

$$\inf_{\text{tuning params}} \sup_{\text{images}} \mathrm{MSE} = \sigma\,\mathrm{Poly}(\log\sigma)$$

What we achieved so far

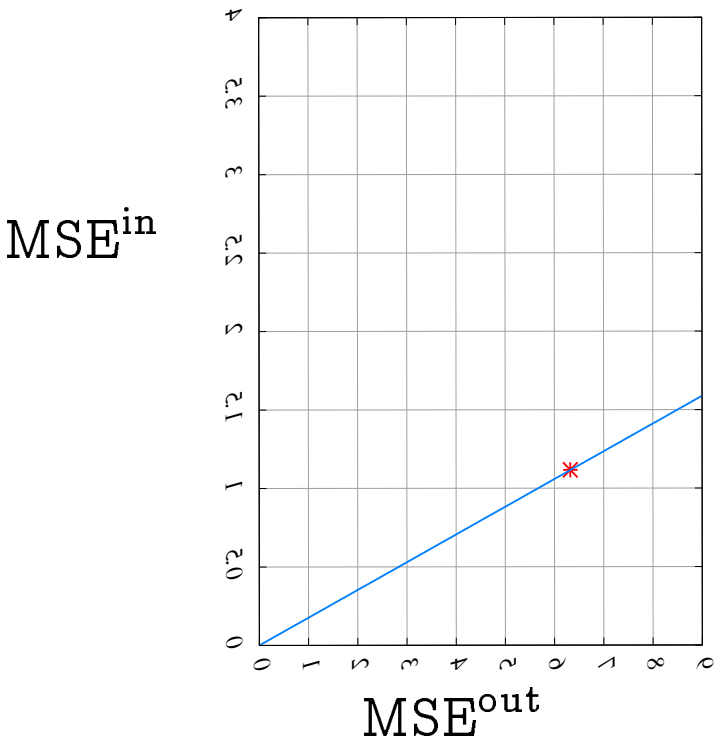


Matched filter

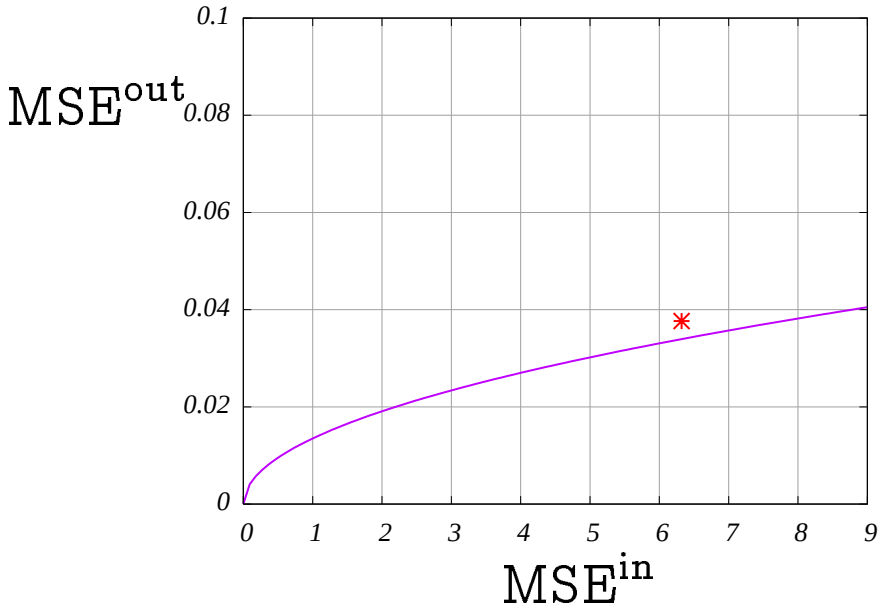


Denoiser

What we achieved so far

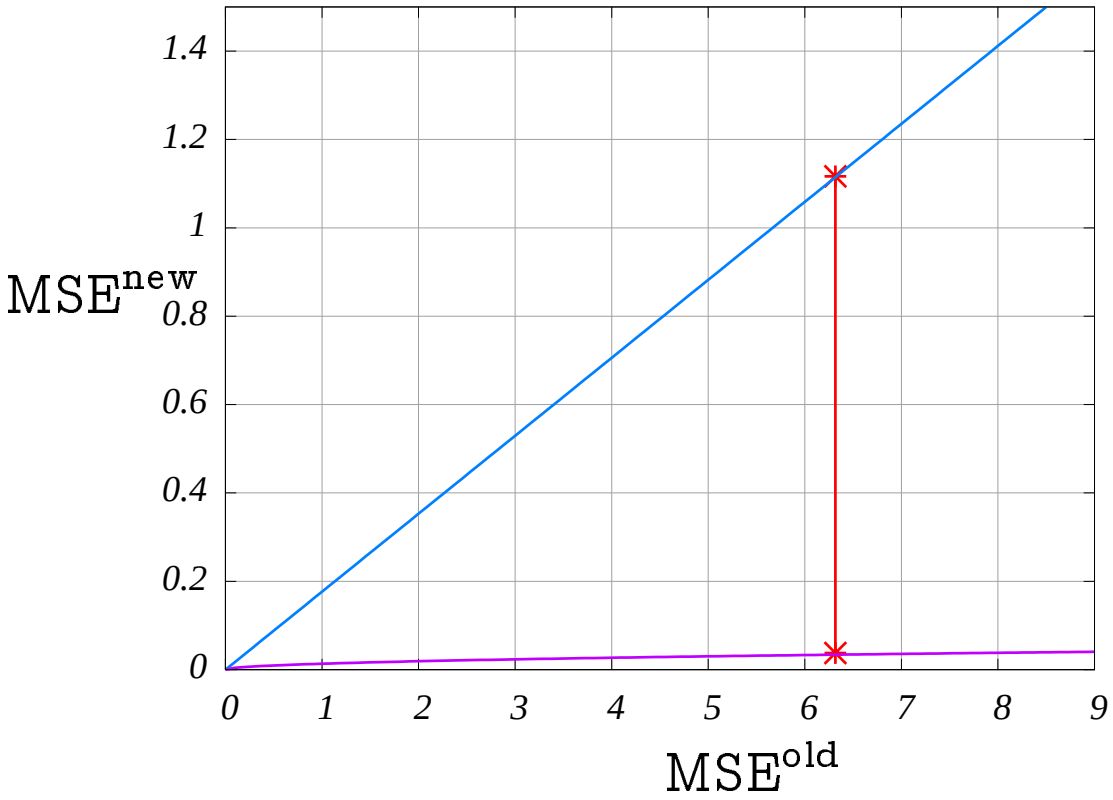


Matched filter

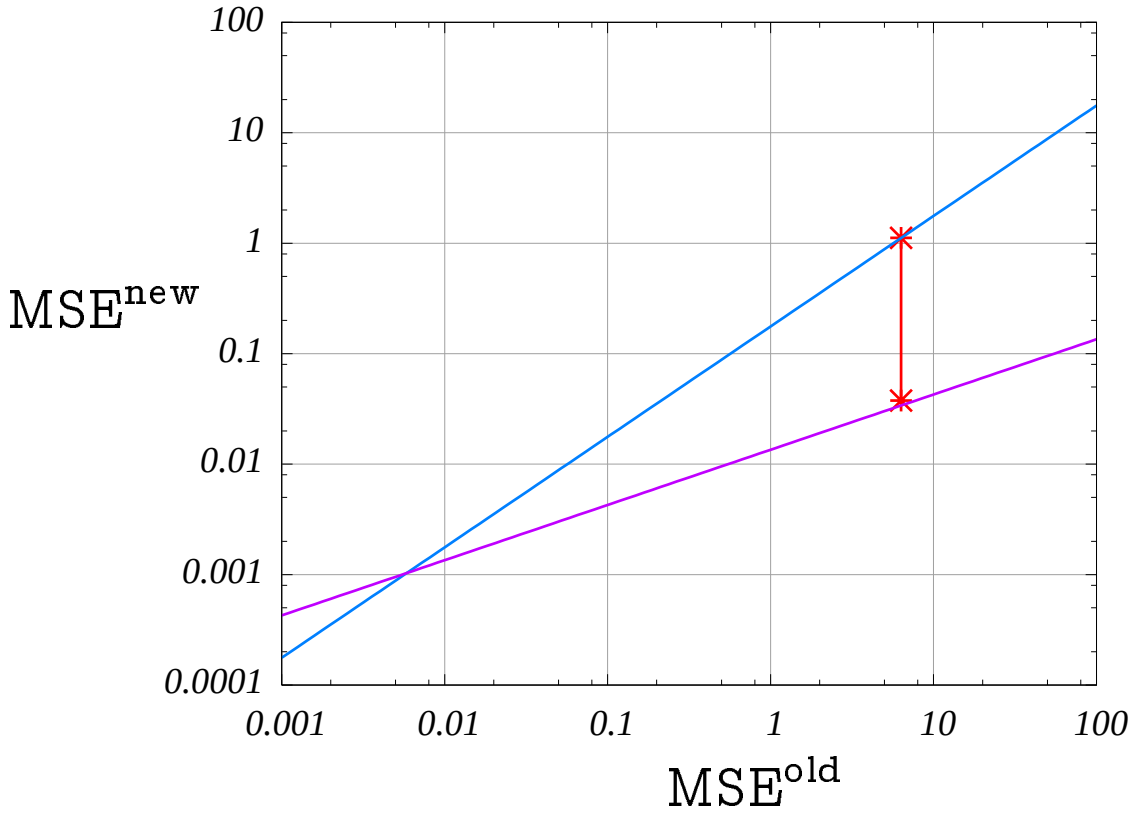


Denoiser

What we achieved so far

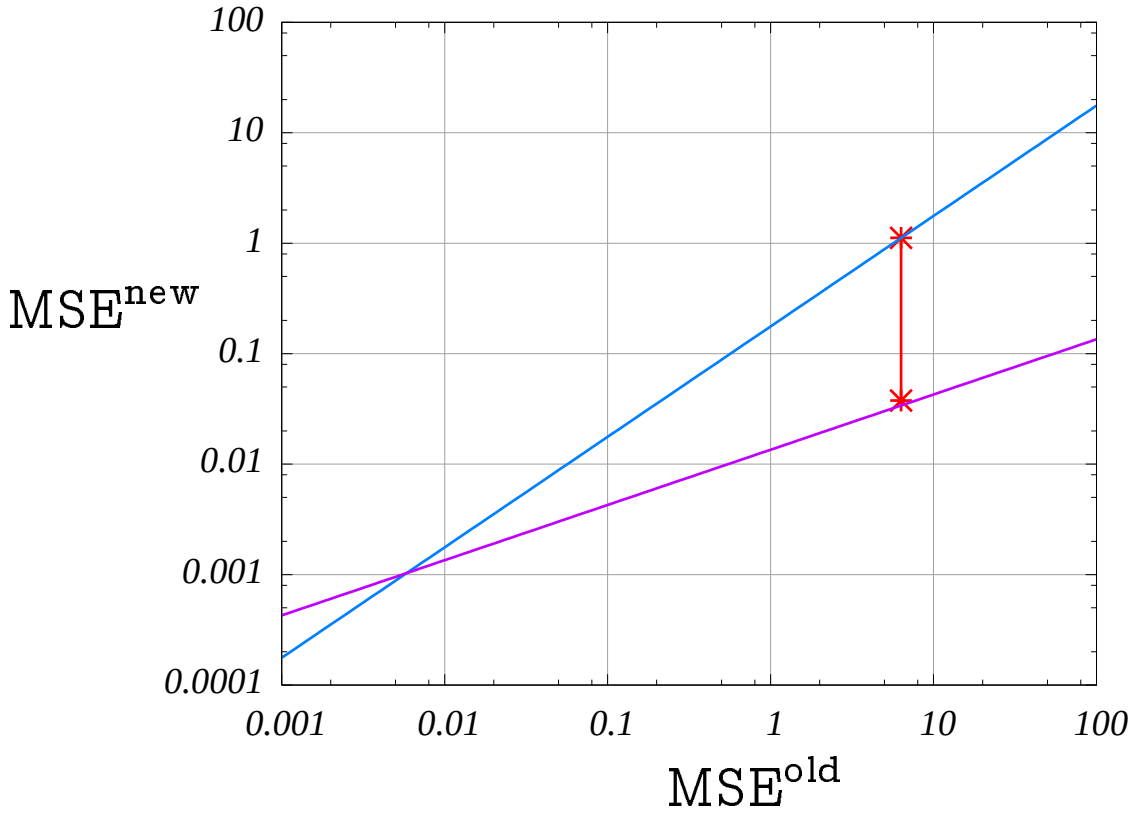


What we achieved so far



What about iterating?

What we achieved so far



What about iterating?

How do we iterate?

$$\begin{aligned} \hat{\theta}^1 &= A^\dagger y \\ \hat{\theta}^2 &= \eta(\hat{\theta}^1) \\ \hat{\theta}^3 &= ??? \end{aligned}$$

$$\begin{aligned} A(\theta - \hat{\theta}^2) &= y - A\hat{\theta}^2 \\ \theta - \hat{\theta}^2 &\approx A^\dagger(y - A\hat{\theta}^2) \end{aligned}$$

How do we iterate?

$$\begin{aligned} \hat{\theta}^1 &= A^\dagger y \\ \hat{\theta}^2 &= \eta(\hat{\theta}^1) \\ \hat{\theta}^3 &= ??? \end{aligned}$$

$$\begin{aligned} A(\theta - \hat{\theta}^2) &= y - A\hat{\theta}^2 \\ \theta - \hat{\theta}^2 &\approx A^\dagger(y - A\hat{\theta}^2) \end{aligned}$$

How do we iterate?

$$\begin{aligned} \hat{\theta}^1 &= A^\dagger y \\ \hat{\theta}^2 &= \eta(\hat{\theta}^1) \\ \hat{\theta}^3 &= ??? \end{aligned}$$

$$\begin{aligned} A(\theta - \hat{\theta}^2) &= y - A\hat{\theta}^2 \\ \theta - \hat{\theta}^2 &\approx A^\dagger(y - A\hat{\theta}^2) \end{aligned}$$

How do we iterate?

$$
\begin{aligned}
\hat{\theta}^1 &= A^\dagger y \\
\hat{\theta}^2 &= \eta(\hat{\theta}^1) \\
\hat{\theta}^3 &= \hat{\theta}^2 + A^\dagger (y - A\hat{\theta}^2) \\
\hat{\theta}^4 &= \eta(\hat{\theta}^3) \\
\hat{\theta}^5 &= \hat{\theta}^4 + A^\dagger (y - A\hat{\theta}^4) \\
\hat{\theta}^6 &= \eta(\hat{\theta}^5) \\
\dots & \quad \dots
\end{aligned}
$$

How do we iterate?

$$
\begin{aligned}
\hat{\theta}^1 &= A^\dagger y \\
\hat{\theta}^2 &= \eta(\hat{\theta}^1) \\
\hat{\theta}^3 &= \hat{\theta}^2 + A^\dagger (y - A\hat{\theta}^2) \\
\hat{\theta}^4 &= \eta(\hat{\theta}^3) \\
\hat{\theta}^5 &= \hat{\theta}^4 + A^\dagger (y - A\hat{\theta}^4) \\
\hat{\theta}^6 &= \eta(\hat{\theta}^5) \\
\dots & \quad \dots
\end{aligned}
$$

How do we iterate?

$$
\begin{aligned}
\hat{\theta}^1 &= A^\dagger y \\
\hat{\theta}^2 &= \eta(\hat{\theta}^1) \\
\hat{\theta}^3 &= \hat{\theta}^2 + A^\dagger(y - A\hat{\theta}^2) \\
\hat{\theta}^4 &= \eta(\hat{\theta}^3) \\
\hat{\theta}^5 &= \hat{\theta}^4 + A^\dagger(y - A\hat{\theta}^4) \\
\hat{\theta}^6 &= \eta(\hat{\theta}^5) \\
\dots & \quad \dots
\end{aligned}
$$

How do we iterate?

$$
\begin{aligned}
\hat{\theta}^1 &= A^\dagger y \\
\hat{\theta}^2 &= \eta(\hat{\theta}^1) \\
\hat{\theta}^3 &= \hat{\theta}^2 + A^\dagger (y - A\hat{\theta}^2) \\
\hat{\theta}^4 &= \eta(\hat{\theta}^3) \\
\hat{\theta}^5 &= \hat{\theta}^4 + A^\dagger (y - A\hat{\theta}^4) \\
\hat{\theta}^6 &= \eta(\hat{\theta}^5) \\
\dots & \qquad \dots
\end{aligned}
$$

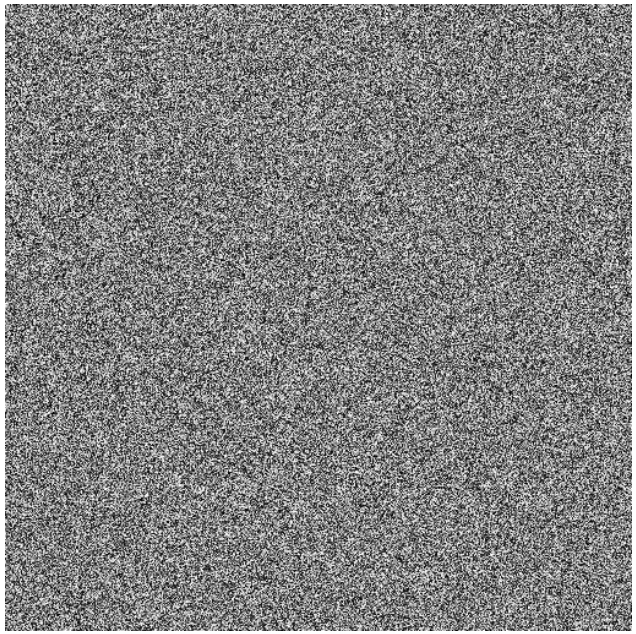
How do we iterate?

$$
\begin{aligned}
\hat{\theta}^1 &= A^\dagger y \\
\hat{\theta}^2 &= \eta(\hat{\theta}^1) \\
\hat{\theta}^3 &= \hat{\theta}^2 + A^\dagger(y - A\hat{\theta}^2) \\
\hat{\theta}^4 &= \eta(\hat{\theta}^3) \\
\hat{\theta}^5 &= \hat{\theta}^4 + A^\dagger(y - A\hat{\theta}^4) \\
\hat{\theta}^6 &= \eta(\hat{\theta}^5) \\
\dots &\quad \dots
\end{aligned}
$$

For $t = 1, 2, 3, \dots, 20$

$$\begin{aligned} \hat{\theta}^{2t} &= \eta(\hat{\theta}^{2t-1}) \\ \hat{\theta}^{2t+1} &= \hat{\theta}^{2t} + A^{\dagger}(y - A\hat{\theta}^{2t}) \end{aligned}$$

$t = 1$

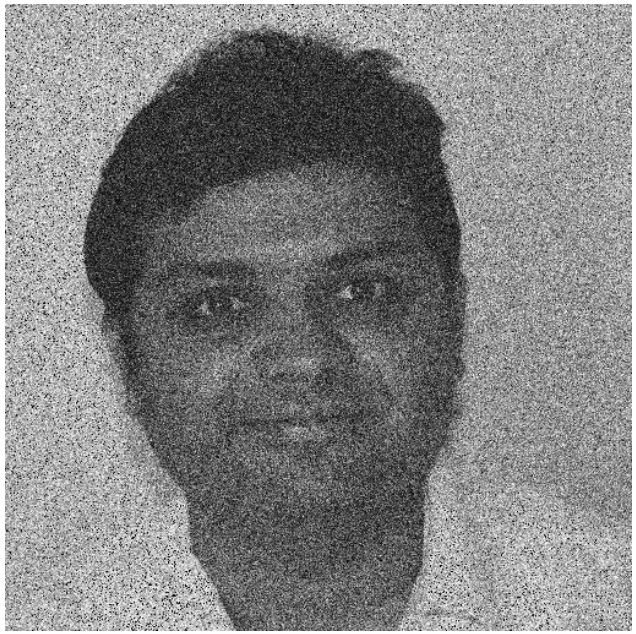
$\hat{\theta}^1 =$ 

$t = 2$

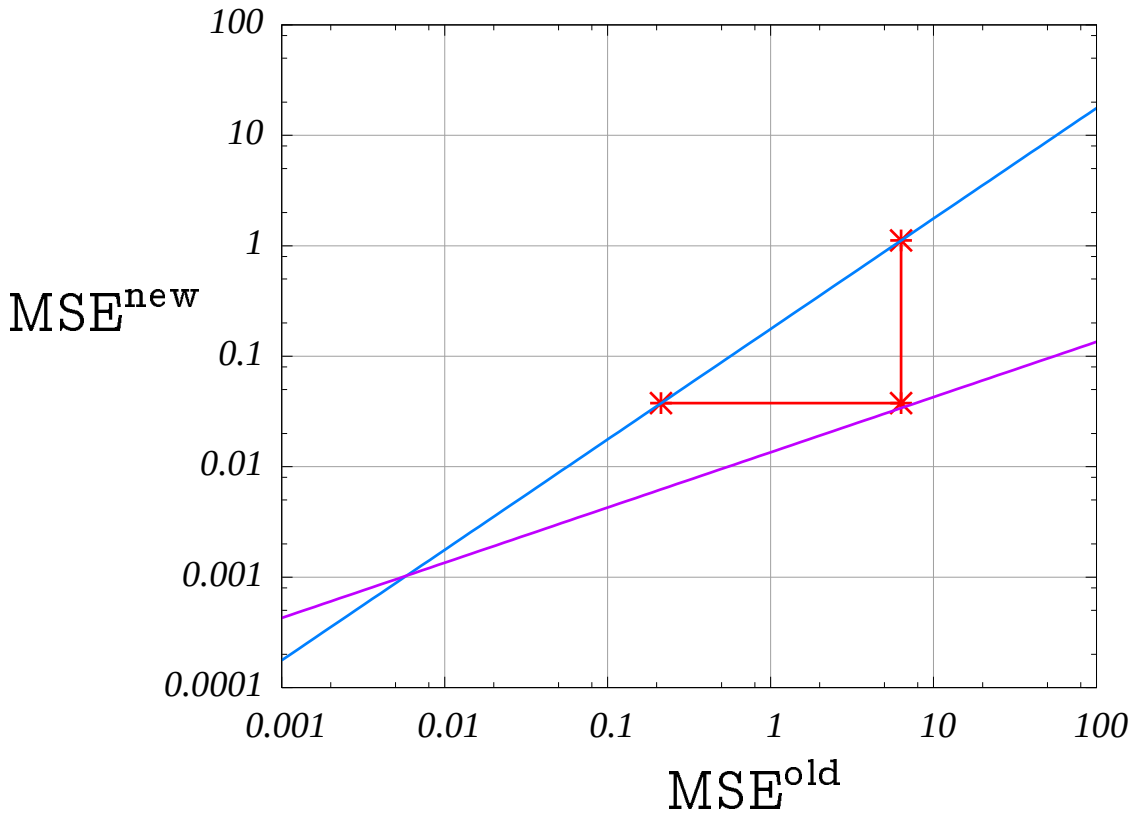
$\hat{\theta}^2 =$ 

$t = 3$

$\hat{\theta}^3 =$



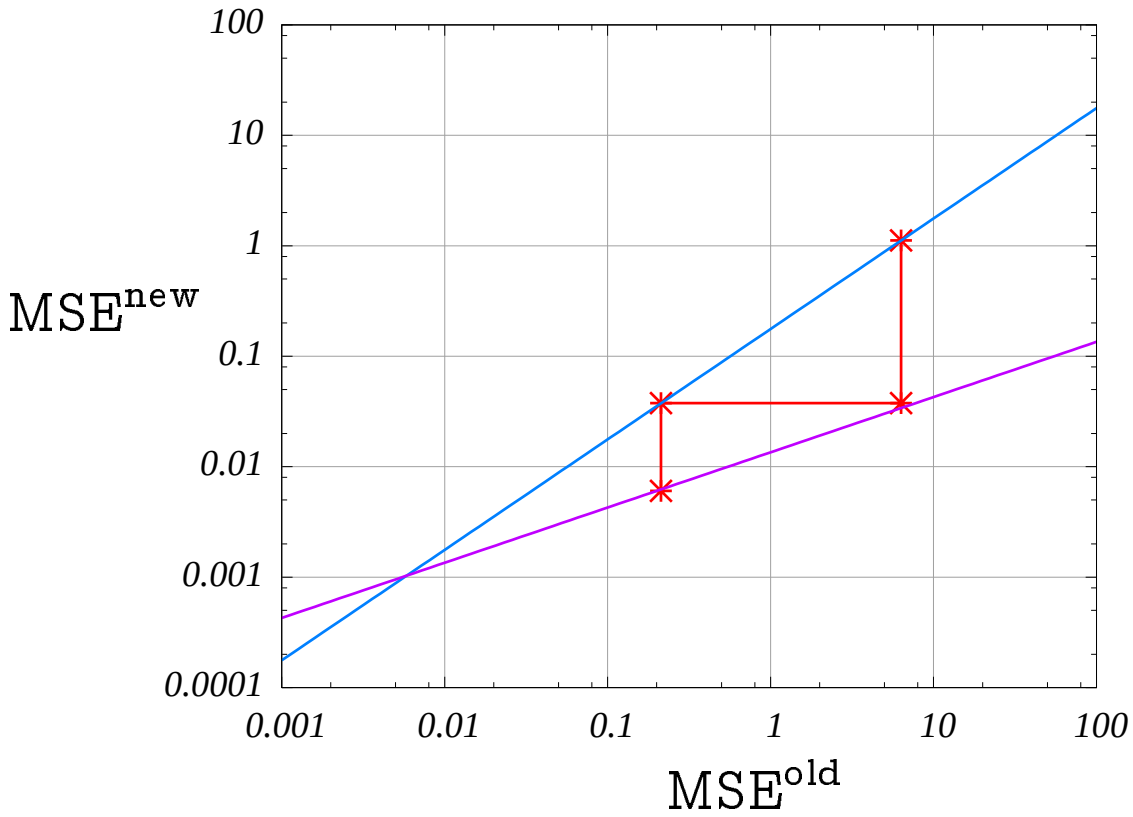
$t = 3$



$t = 4$

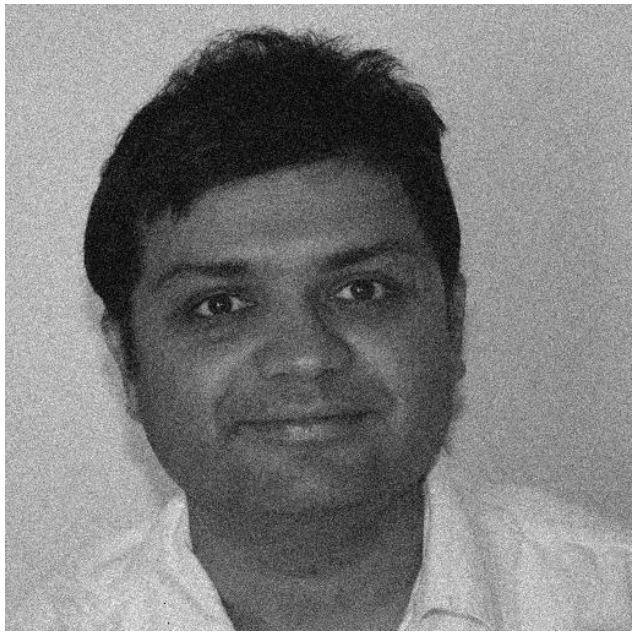
$\hat{\theta}^4 =$

$t = 3$

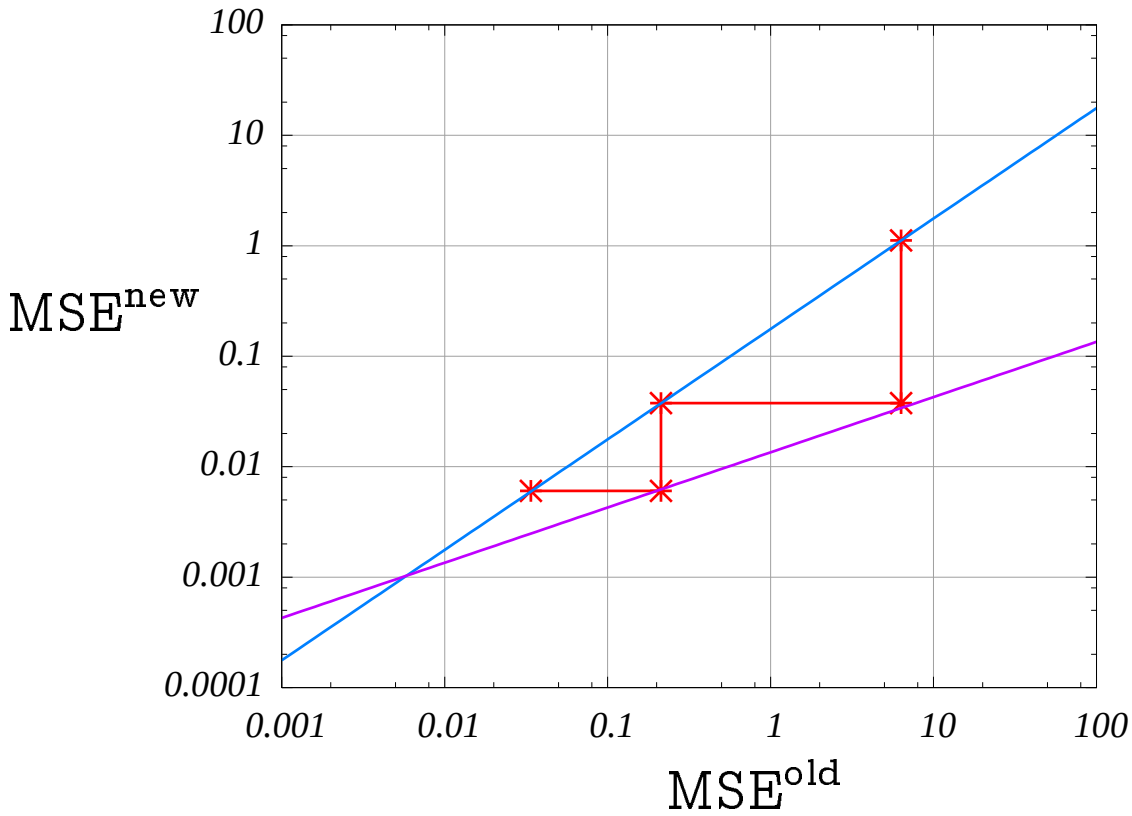


$t = 5$

$\hat{\theta}^5 =$



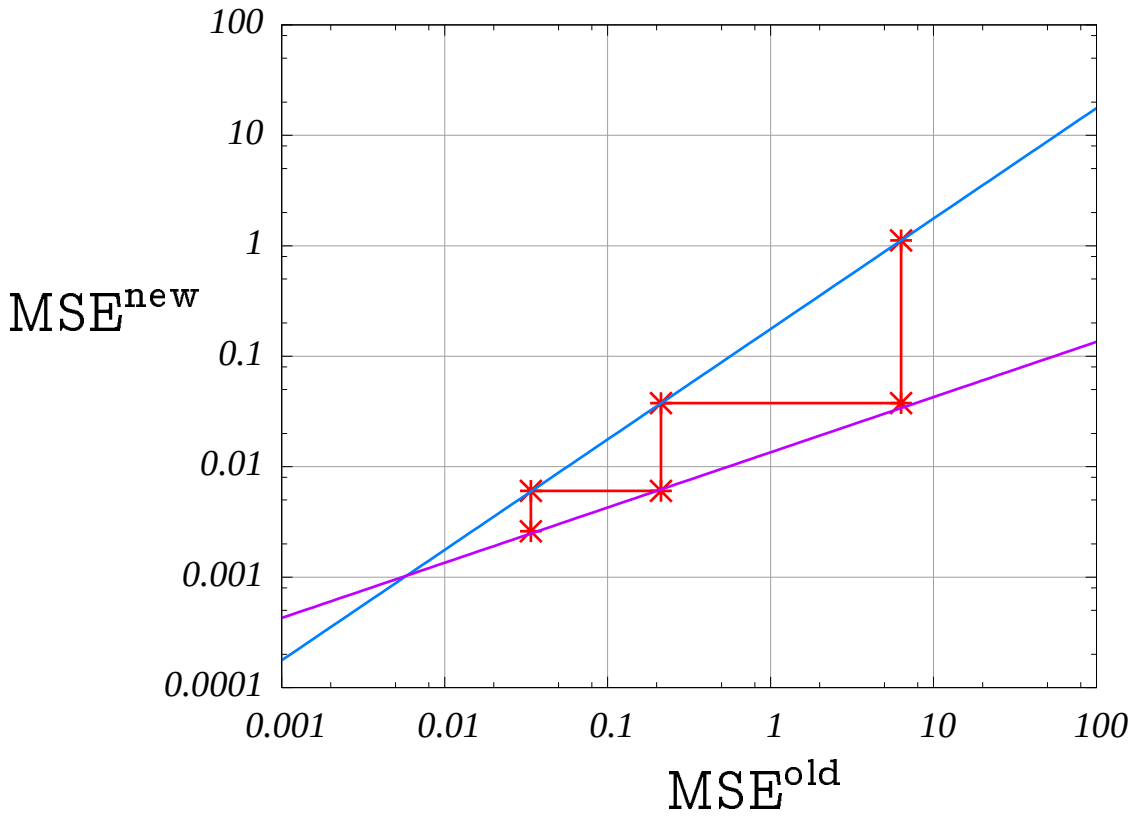
$t = 5$



$t = 6$

$\hat{\theta}^6 =$

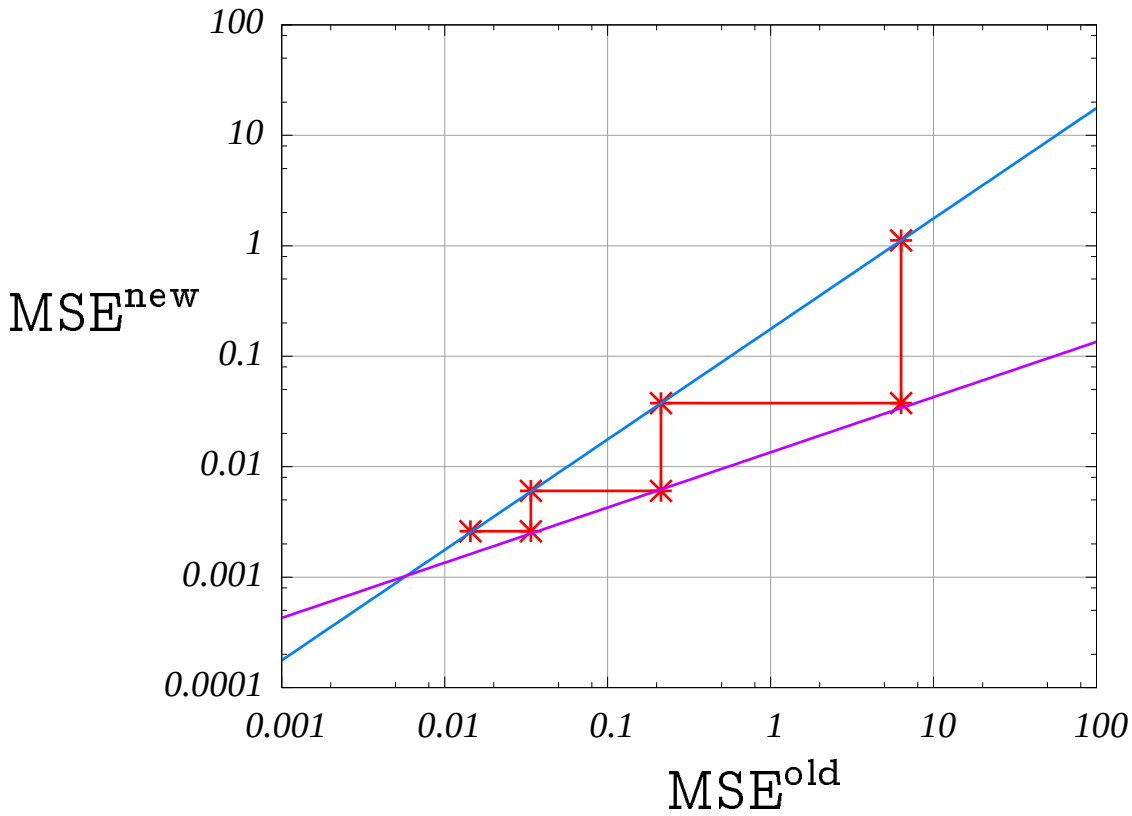
$t = 6$



$t = 7$

$\hat{\theta}^7 =$

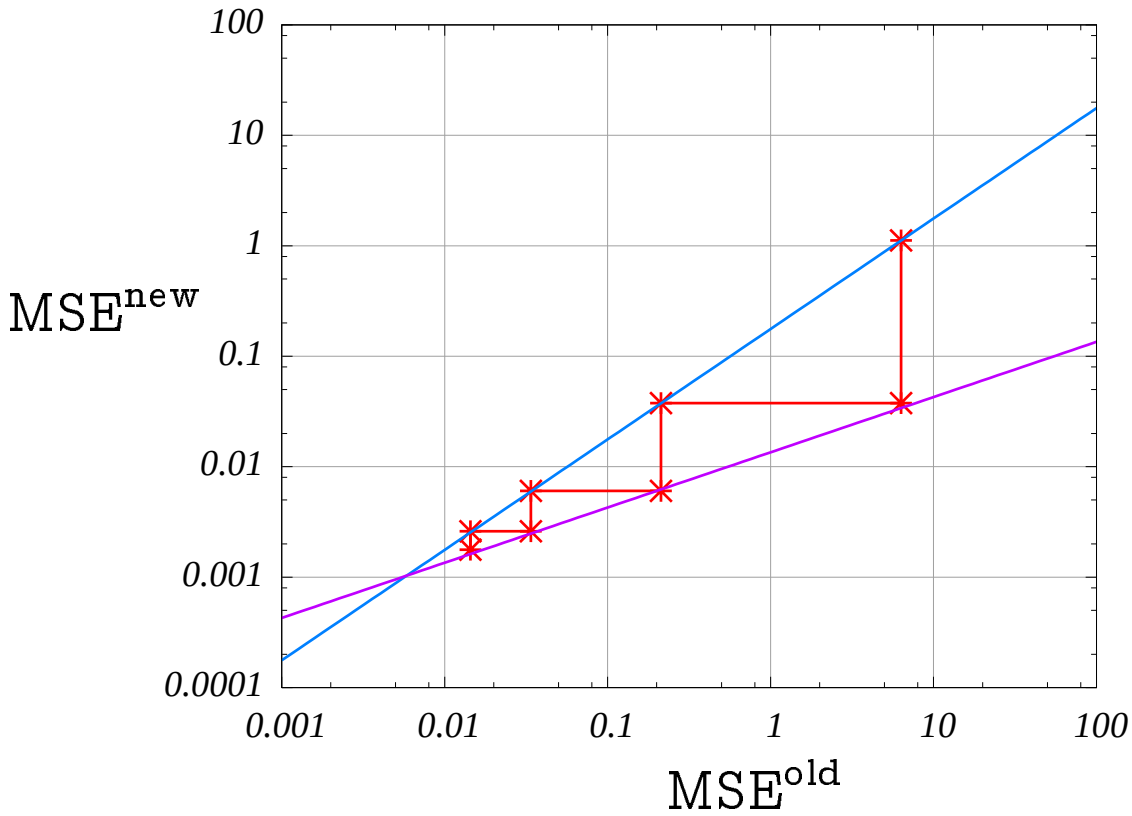
$t = 7$



$t = 8$

$\hat{\theta}^8 =$ 

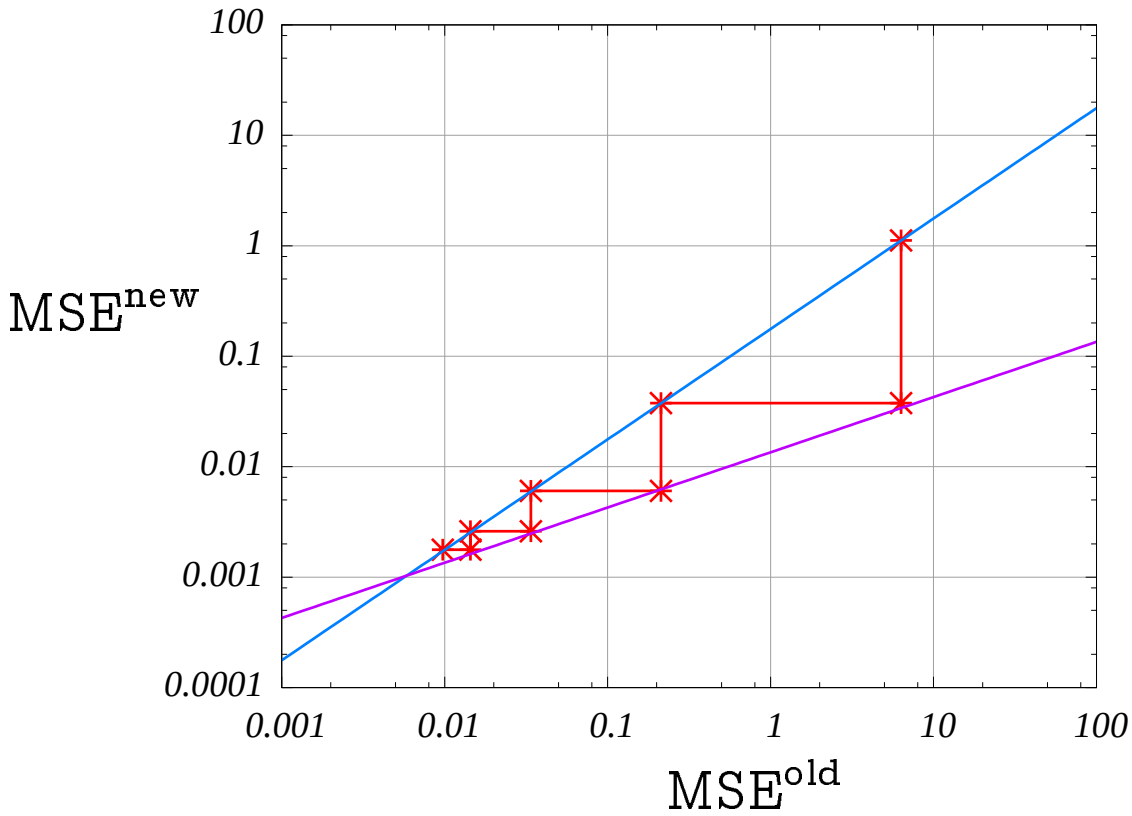
$t = 8$



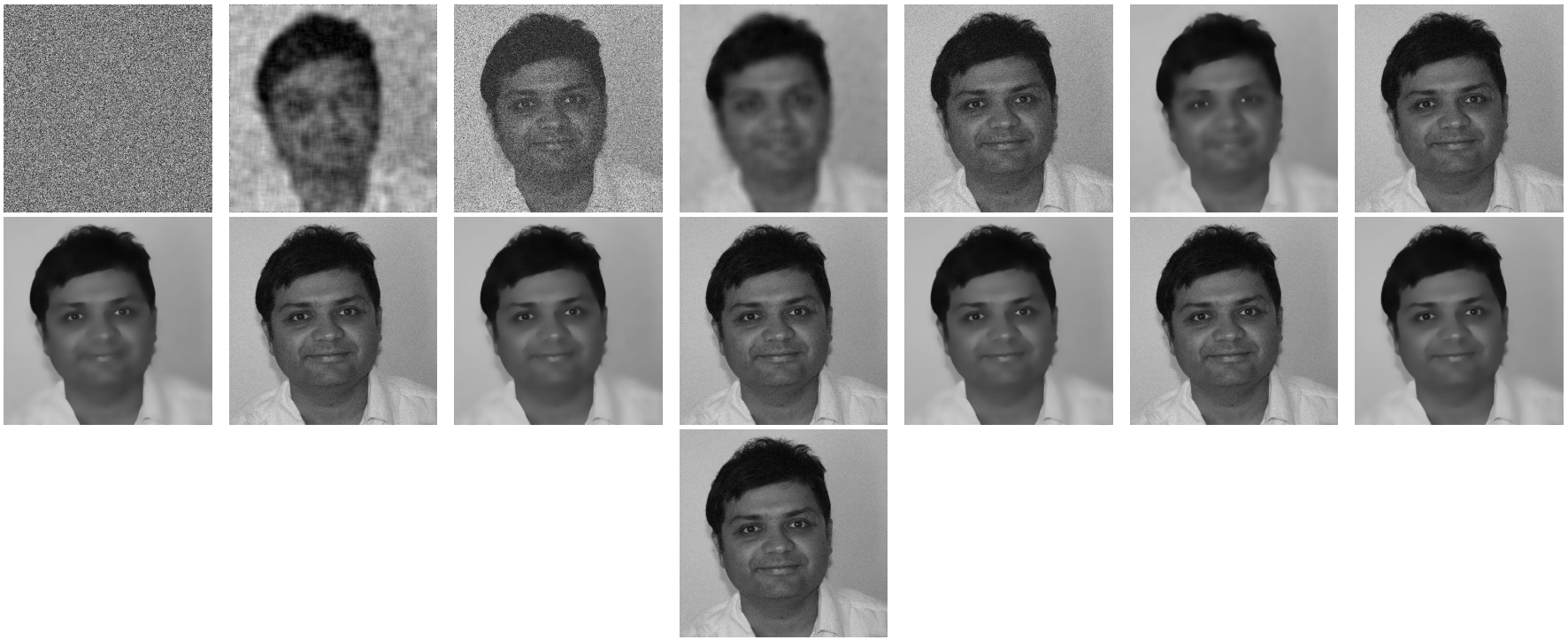
$t = 9$

$\hat{\theta}^9 =$

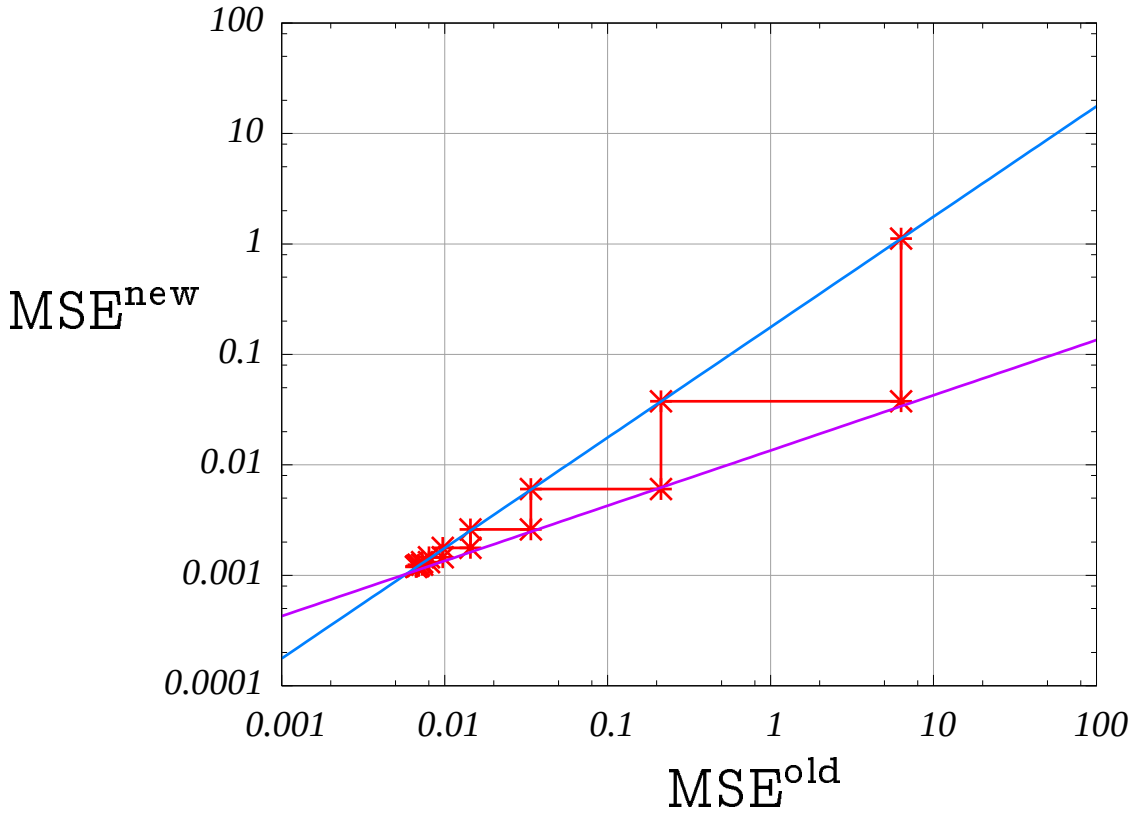
$t = 9$



$t = 0, 1, 2, 3, \ldots, 20$



$t = 0, 1, 2, 3, \ldots$



Well in reality I cheated

Well in reality I cheated

Instead of this:

$$\hat{\theta}^{2t} = \eta(\hat{\theta}^{2t-1})$$

$$\hat{\theta}^{2t+1} = \hat{\theta}^{2t} + A^{\dagger}(y - A\hat{\theta}^{2t})$$

I used this (for $\mathsf{b}_t \in \mathbb{C}$)

$$\hat{\theta}^{2t} = \eta(\hat{\theta}^{2t-1})$$

$$\hat{\theta}^{2t+1} = \hat{\theta}^{2t} + A^{\dagger} r^t$$

$$r^t = y - A\hat{\theta}^{2t} + \mathsf{b}_t\, r^{t-1}$$

Approximate Message Passing (AMP)

$$\hat{\theta}^{2t} = \eta(\hat{\theta}^{2t-1})$$

$$\hat{\theta}^{2t+1} = \hat{\theta}^{2t} + A^{\dagger}\, r^{t}$$

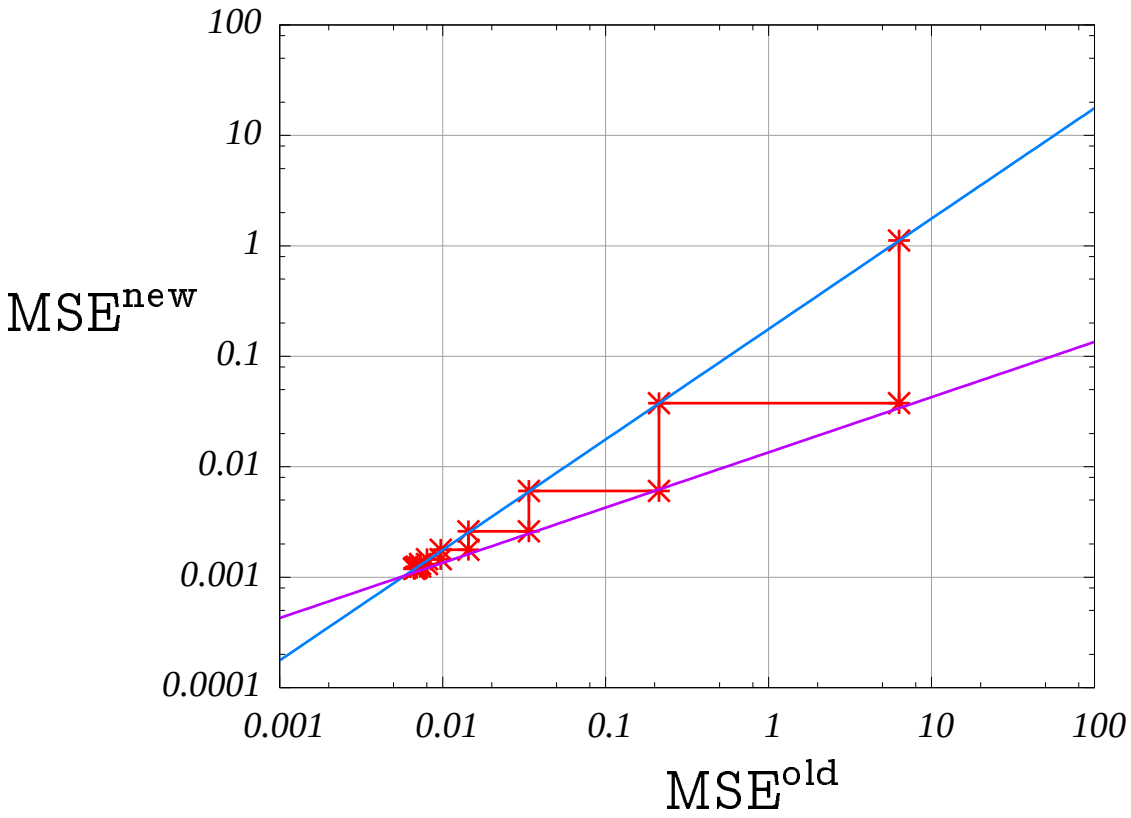
$$r^{t} = y - A\hat{\theta}^{2t} + \mathsf{b}_t\, r^{t-1}$$

$$\mathsf{b}_t = \frac{1}{m}\mathrm{div}\eta(\hat{\theta}^{2t-1})$$

(can be computed explicitly)

[Thouless, Anderson, Palmer, 1977, Kabashima, 2003, Donoho, Maleki, Montanari, 2009, Donoho, Johnstone, Montanari, 2009]

State Evolution

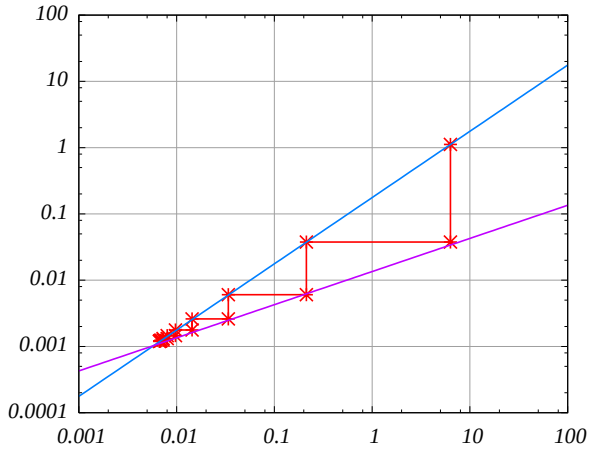


A few theorems

Things to play with

- Matrix $A \in \mathbb{R}^{m \times n}$.
- Denoiser $\eta : \mathbb{R}^n \to \mathbb{R}^n$.
- Additive noise.

State Evolution



Theorem (Bayati, Montanari 2010)

Assume A has i.i.d. Gaussian entries, and η is separable

$$\eta(v) = (\eta_1(v_1), \eta_2(v_2), \dots, \eta_n(v_n))$$

Then state evolution holds asymptotically as $n \to \infty$.

[Proof uses a very nice technique by Erwin Bolthausen]

State Evolution: More theorems

Bayati, Montanari 2010: A more general class of iterations.

Bayati, Lelarge, Montanari 2012: A with non-Gaussian i.i.d. entries; polynomial separable denoiser.

Still far from the example of the first part

State Evolution: More theorems

Bayati, Montanari 2010: A more general class of iterations.

Bayati, Lelarge, Montanari 2012: A with non-Gaussian i.i.d. entries; polynomial separable denoiser.

Still far from the example of the first part

State Evolution: More theorems

Bayati, Montanari 2010: A more general class of iterations.

Bayati, Lelarge, Montanari 2012: A with non-Gaussian i.i.d. entries; polynomial separable denoiser.

Still far from the example of the first part

State Evolution: More theorems

Bayati, Montanari 2010: A more general class of iterations.

Bayati, Lelarge, Montanari 2012: A with non-Gaussian i.i.d. entries; polynomial separable denoiser.

Still far from the example of the first part

Connection with convex optimization

$J : \mathbb{R}^n \to \mathbb{R}$ convex regularizer

Proximal operator

$$\eta(y) = \arg\min_{x \in \mathbb{R}} \left\{ \frac{1}{2} \|y - x\|_2^2 + J(x) \right\}$$

Examples:

$$\begin{aligned} J(x) &= \|x\|_1 , && (\Rightarrow \eta \text{ separable}) \\ J(x) &= \|x\|_{TV}, && (\Rightarrow \text{total variation denoising}) \end{aligned}$$

Connection with convex optimization

Lemma

If $\eta(\cdot)$ is the proximal operator of $J(\cdot)$, and $\hat{\theta}$ is a fixed point of AMP, then

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2} \|y - A\theta\|_2^2 + \lambda J(\theta) \right\},$$

for $\lambda = (1 - \mathsf{b}_\infty)^{-1}$.

(But theory applies to more general denoisers!)

Does AMP converge to a minimizer?

Connection with convex optimization

Lemma

If $\eta(\cdot)$ is the proximal operator of $J(\cdot)$, and $\hat{\theta}$ is a fixed point of AMP, then

$$\hat{\theta} \in \arg\min_{\theta \in \mathbb{R}^n} \left\{ \frac{1}{2} \|y - A\theta\|_2^2 + \lambda J(\theta) \right\},$$

for $\lambda = (1 - \mathsf{b}_\infty)^{-1}$.

(But theory applies to more general denoisers!)

Does AMP converge to a minimizer?

Does AMP converge to a minimizer?

Theorem (Bayati, Montanari 2011)

If $J(x) = \gamma \|x\|_1$ and A is Gaussian with i.i.d. entries, then (for n large enough) AMP converge within relative distance ε from a minimizer in $t = O(\log(1/\varepsilon))$ iterations.

Corollary

Asymptotic distributional characterization of the minimizer.

Does AMP converge to a minimizer?

Theorem (Bayati, Montanari 2011)

If $J(x) = \gamma \|x\|_1$ and A is Gaussian with i.i.d. entries, then (for n large enough) AMP converge within relative distance ε from a minimizer in $t = O(\log(1/\varepsilon))$ iterations.

Corollary

Asymptotic distributional characterization of the minimizer.

Connection with convex optimization

- Asymototic characterization of the minimizer through the (non-rigorous) replica method.
  [Tanaka 2002, Guo, Verdú 2005, Kabashima, Tanaka 2009, Rangan, Fletcher, Goyal 2009, Caire, Tulino, Shamai, Verdú 2012, Javanmard, Montanari 2012...]

- Bayati, Lelarge, Montanari 2012:
  Partial result for $J(x) = \gamma \|x\|_1$ and A with non-Gaussian entries.

- Bean, Bickel, El Karoui, Lim, Yu 2012:
  Alternative argument for robust regression (e.g. $\min_\theta \|y - A\theta\|_1$)

One proof idea: Universality

For simplicity $A \in \mathbb{R}^{n\times n}$ symmetric

$$\hat{\theta}^{t+1} = A f(\hat{\theta}^t) + \mathsf{b}_t f(\hat{\theta}^{t-1})$$

$$f(v) = (f(v_1), f(v_2), \dots, f(v_n))$$

Lemma

If f is a polynomial, then $\mathcal{L}(\hat{\theta}^t_i)$ is asymptotically universal for A with i.i.d. entries with $\mathbb{E}(A_{ij}) = 0$, $\mathbb{E}(A^2_{ij}) = 1/n$.

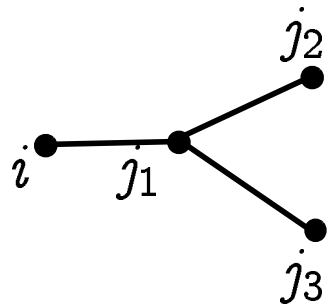
One proof idea: Universality

For simplicity $A \in \mathbb{R}^{n\times n}$ symmetric

$$\hat{\theta}^{t+1} = A f(\hat{\theta}^t)\,, \quad \hat{\theta}^0 = 1$$

$$f(v) = ((v_1)^2, (v_2)^2, \dots, (v_n)^2)$$

$$\hat{\theta}_i^2 = \sum_{i,j_1,j_2,j_3} A_{ij_1} A_{j_1 j_2} A_{j_1 j_3} = \sum$$



One proof idea: Universality

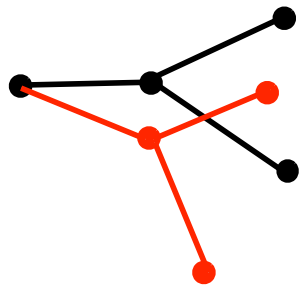
For simplicity $A \in \mathbb{R}^{n\times n}$ symmetric

$$\hat{\theta}^{t+1} = A f(\hat{\theta}^t)\,, \quad \hat{\theta}^0 = 1$$

$$f(v) = ((v_1)^2, (v_2)^2, \dots, (v_n)^2)$$

Prove universality of 2nd moment

$$\mathbb{E}\{(\hat{\theta}_i^2)^2\} = \sum$$



Prove that the only terms that 'survive' as $n \to \infty$ have each edge A_{kl} appearing zero or two times.

One proof idea: Universality

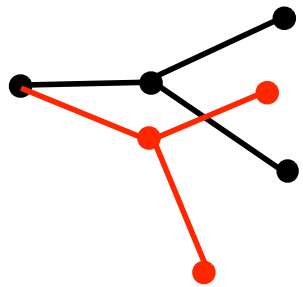
For simplicity $A \in \mathbb{R}^{n\times n}$ symmetric

$$\hat{\theta}^{t+1} = A f(\hat{\theta}^t)\,, \quad \hat{\theta}^0 = 1$$

$$f(v) = ((v_1)^2, (v_2)^2, \dots, (v_n)^2)$$

Prove universality of 2nd moment

$$\mathbb{E}\{(\hat{\theta}_i^2)^2\} = \sum$$



Prove that the only terms that 'survive' as $n \to \infty$ have each edge A_{kl} appearing zero or two times.

Generalizations and open problems

More general matrices

- Partial Fourier matrices, random unitary matrices.
  [Caire, Shamai, Tulino, Verdú, 2012 (non-rigorous)]

- Independent Gaussian rows
  [Javanmard, Montanari, 2012 (non-rigorous)]

- Spatially coupled matrices
  [Krzakala, Mézard, Sausset, Sun, Zdeborova, 2011;
  Donoho, Javanmard, Montanari, 2011]

More general models

- Generalized linear models [Rangan 2011]
- Graphical model priors [Schniter et al. 2010-...]
- Low-rank matrices [Rangan, Fletcher 2012]
- ...

Optimal estimation under limited computation

?

Conclusion

Information theory:
Simple probabilistic models, Sharp asymptotics, Surprising insights

Thanks!

Conclusion

Information theory:
Simple probabilistic models, Sharp asymptotics, Surprising insights

Thanks!