Approximate Message Passing

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

Stanford University

December 8, 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

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

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

Problem: Estimate θ from observations y.

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?

Optimization?

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

'Separation principle'

- Modeler/statistician proposes convex cost function.
- Optimization expert proposes simple iterative algorithm.
- Run for 20 iterations and hope for the best.

How should we think about these problems?

Beyond separation?

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

- Constrained complexity per iteration
- Fixed number of iterations (say 20)
- What is minimum MSE achievable?

Outline

- A long example (algorithm + heuristics)
- A list of theorems/pointers

A long example

What type of example?

- Image processing (because they make nice figures)
- Compressed sensing (simple)

WILL NOT MENTION SPARSITY!

What type of example?

- Image processing (because they make nice figures)
- Compressed sensing (simple)

WILL NOT MENTION SPARSITY!

What type of example?

- Image processing (because they make nice figures)
- Compressed sensing (simple)

WILL NOT MENTION SPARSITY!

An image

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

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

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

$\rightarrow y \in \mathbb{C}^m, \; m = 0.17\, n$

An approach popular in this community

$$y = A\theta + z\,, \qquad z \sim \mathrm{N}(0, \sigma^2 \mathrm{I}_{m\times m})$$

$$\begin{aligned} p_{\Theta|Y}(\theta|y) &\propto \exp\Big\{-\frac{1}{2\sigma^2}\|y - A\theta\|_2^2\Big\} p_{\Theta}(\theta) \\ &\propto \prod_{a=1}^{m} \exp\Big\{-\frac{1}{2\sigma^2}(y_a - A_a^{\mathsf{T}}\theta)^2\Big\} \prod_{i=1}^{n} p_{\Theta_i}(\theta_i) \end{aligned}$$

An approach popular in this community

$$y = A\theta + z\,, \qquad z \sim \mathrm{N}(0, \sigma^2 \mathrm{I}_{m\times m})$$

$$p_{\Theta|Y}(\theta|y) \;\propto\; \exp\Big\{-\frac{1}{2\sigma^2}\|y - A\theta\|_2^2\Big\}\, p_\Theta(\theta)$$

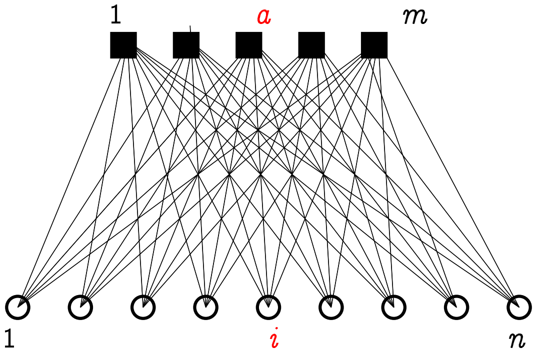
$$\propto\; \prod_{a=1}^{m} \exp\Big\{-\frac{1}{2\sigma^2}(y_a - A_a^{\mathsf{T}}\theta)^2\Big\} \prod_{i=1}^{n} p_{\Theta_i}(\theta_i)$$

An approach popular in this community

$$y = A\theta + z\,, \qquad z \sim \mathrm{N}(0, \sigma^2 \mathrm{I}_{m\times m})$$

$$\begin{aligned} p_{\Theta|Y}(\theta|y) &\propto \exp\Big\{-\frac{1}{2\sigma^2}\|y - A\theta\|_2^2\Big\} p_{\Theta}(\theta) \\ &\propto \prod_{a=1}^{m} \exp\Big\{-\frac{1}{2\sigma^2}\big(y_a - A_a^{\mathsf{T}}\theta\big)^2\Big\} \prod_{i=1}^{n} p_{\Theta_i}(\theta_i) \end{aligned}$$

Factor graph!



$$p_{\Theta|Y}(\theta|y) \propto \prod_{a=1}^{m} \exp\Big\{-\frac{1}{2\sigma^2}\big(y_a - A_a^{\mathsf{T}}\theta\big)^2\Big\} \prod_{i=1}^{n} p_{\Theta_i}(\theta_i)$$

Use BP!

Many issues

- Anyone knows the prior distribution of natural images?
- Computation per iteration, memory $\Theta(mn)$.
- Very loopy graph.
- ...

Let us try something simpler!

Many issues

- Anyone knows the prior distribution of natural images?

- Computation per iteration, memory $\Theta(mn)$.

- Very loopy graph.

- ...

Let us try something simpler!

Constructing a first estimate

$$y = A\theta$$

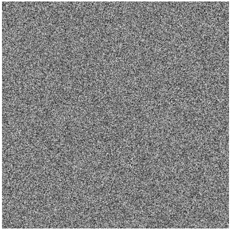
Matched filter

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

How good is this?

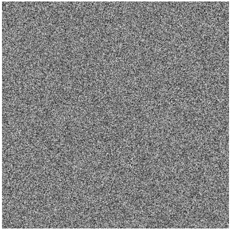
$$\mathbb{E}\,\hat{\theta}^1 = (\text{one line calculation}) = \theta$$

Check it out

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

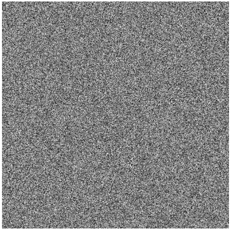
Does not look that good!

Check it out

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

Does not look that good!

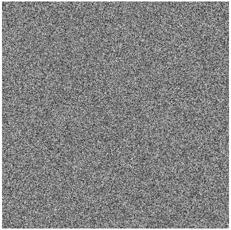
Idea



=

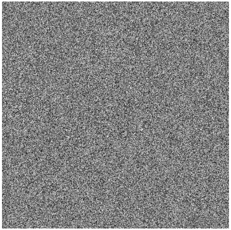
+ 'noise'

Idea



$=$

$+$



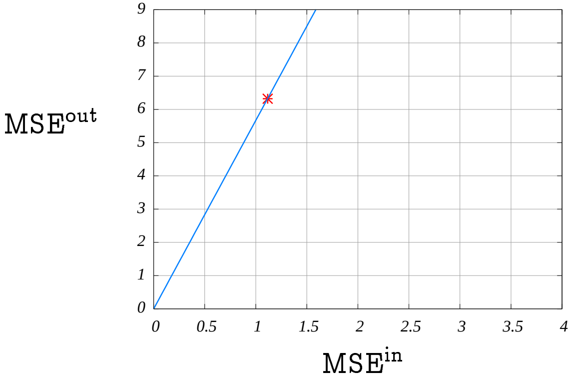
How big is the ‘noise’?

$$\mathbb{E}\{\|\hat{\theta}^1 - \theta\|_2^2\} = \text{(two lines calculation)} = \frac{1-\delta}{\delta}\,\|\theta\|_2^2$$

Matched filter blows up noise

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

Let's check



Denoising

$$\hat{\theta}^1 \approx \theta + \sigma\, z\,, \qquad z_i \sim \mathrm{N}(0,1)$$

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

Denoising

$$\hat{\theta}^1 \approx \theta + \sigma\, z\,, \qquad z_i \sim \mathrm{N}(0,1)$$

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

Denoising by nonlocal means

$$\hat{y} = \theta + \sigma z,$$

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

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

[Buades, Coll, Morel, 2005]

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

Denoising by nonlocal means

$$\widehat{y} = \theta + \sigma z ,$$

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

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

[Buades, Coll, Morel, 2005]

$\hat{\theta} = \eta(\widehat{y})$

Denoising by nonlocal means

$$\widehat{y} = \theta + \sigma z ,$$

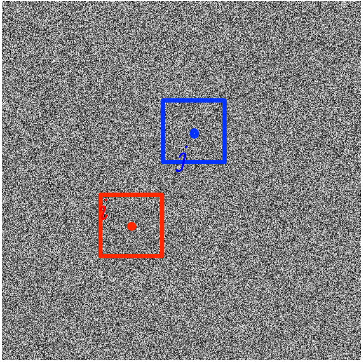
$$\hat{\theta}_i = \frac{\sum_j W(i;j)\widehat{y}_j}{\sum_j W(i;j)} ,$$

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

[Buades, Coll, Morel, 2005]

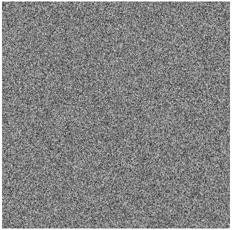
$\hat{\theta} \equiv \eta(\widehat{y})$

Patches



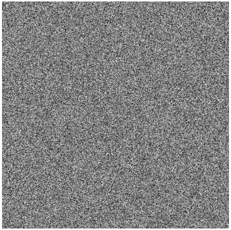
Will it work?

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



$$\Big)$$

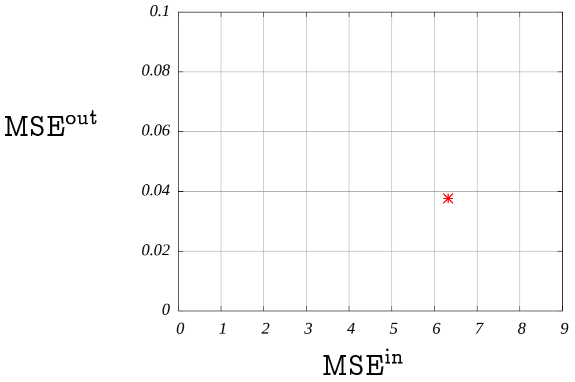
Let's try

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

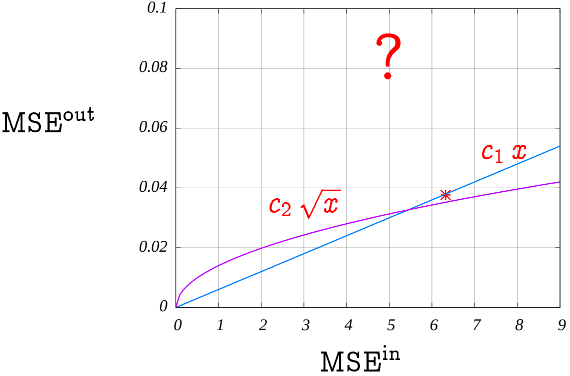
$\hat{\theta}^2 = \eta(A^\dagger y) =$ 

Better than garbage!

How much better?



How much better?



Let us repeat the denoising experiment

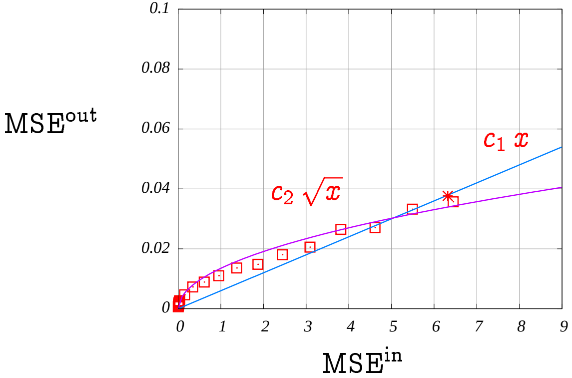
Let us repeat the denoising experiment: $y = \theta + \sigma\, z$

y

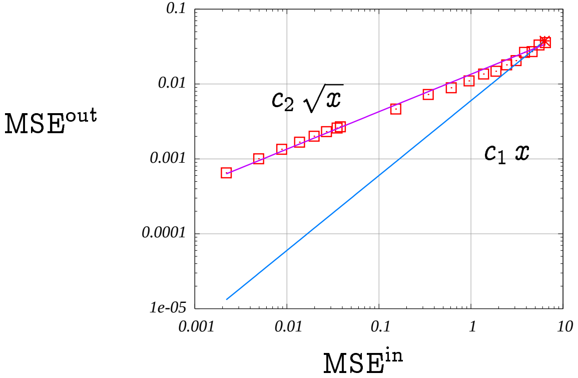
$\eta(y)$

$\sigma = 1$ $\sigma = 0.5$ $\sigma = 0.25$ $\sigma = 0.12$

Quantitatively



Quantitatively



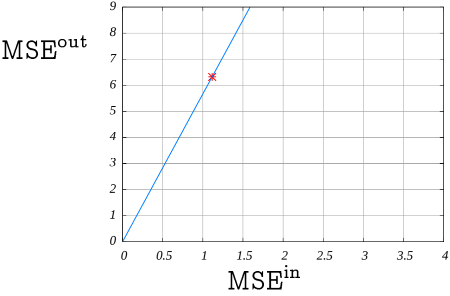
Approximate denoiser characterization

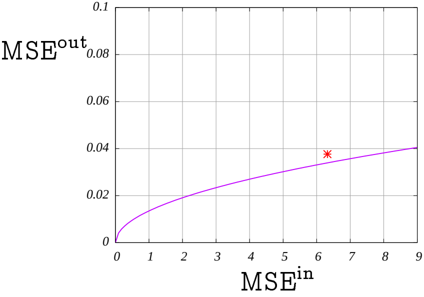
$$\mathrm{MSE}^{\mathrm{out}} = c\,\sqrt{\mathrm{MSE}^{\mathrm{in}}}$$

(enough for our purposes)

(see also Maleki, Baraniuk, Narayan, 2012)
(Arias-Castro, Willett, 2012)

What we achieved so far

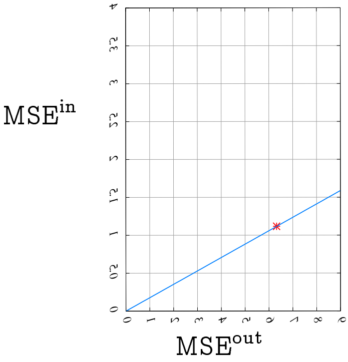




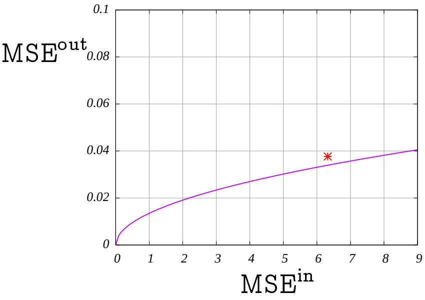
Matched filter

Denoiser

What we achieved so far

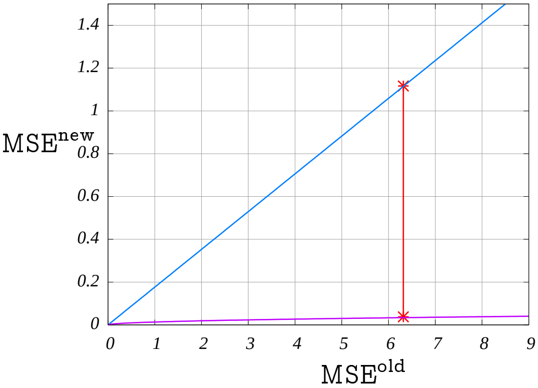


Matched filter

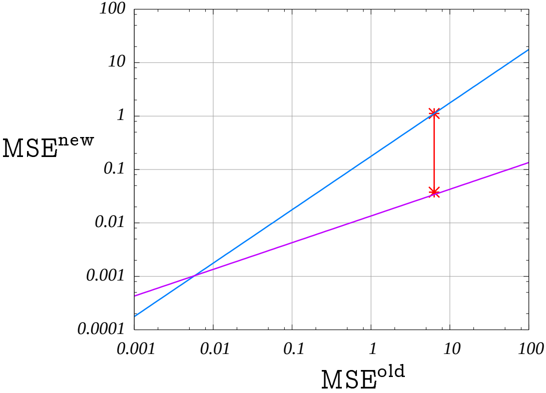


Denoiser

What we achieved so far

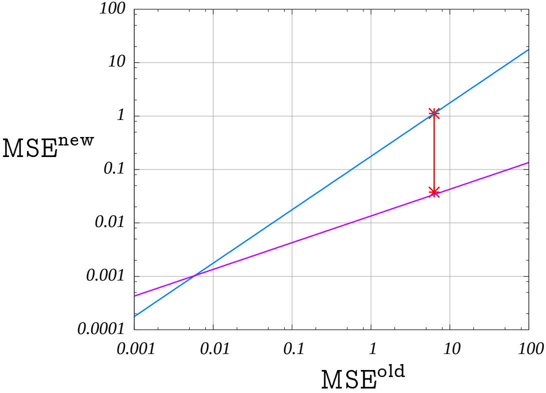


What we achieved so far



What about iterating?

What we achieved so far

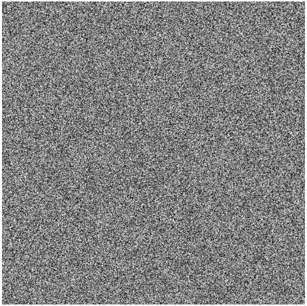


What about iterating?

How do we iterate?

Will tell you later!

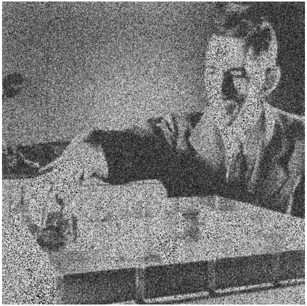
$t = 1$

$\hat{\theta}^1 =$ 

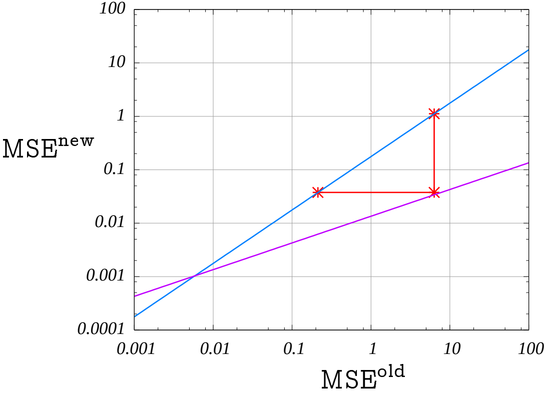
$t = 2$

$\hat{\theta}^2 =$ 

$t = 3$

$\hat{\theta}^3 =$ 

$t = 3$

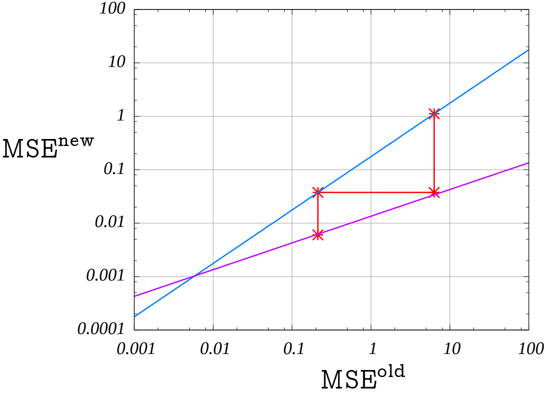


Non obvious!

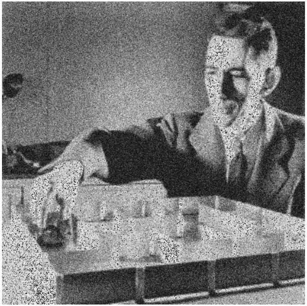
$t = 4$

$\hat{\theta}^4 =$ 

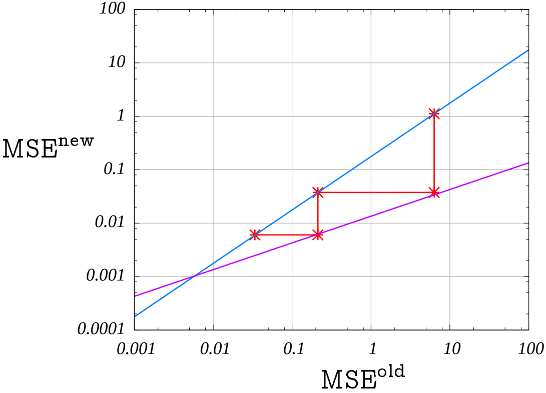
$t = 4$



$t = 5$

$\hat{\theta}^5 =$ 

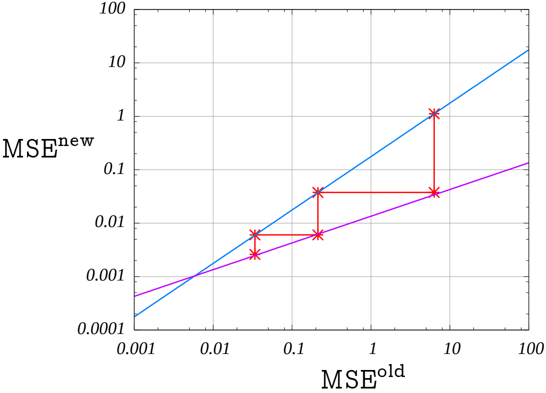
$t = 5$



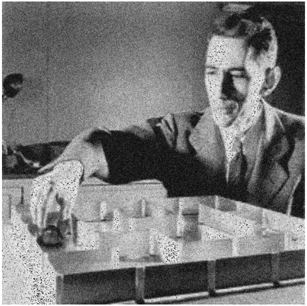
$t = 6$

$\hat{\theta}^6 =$ 

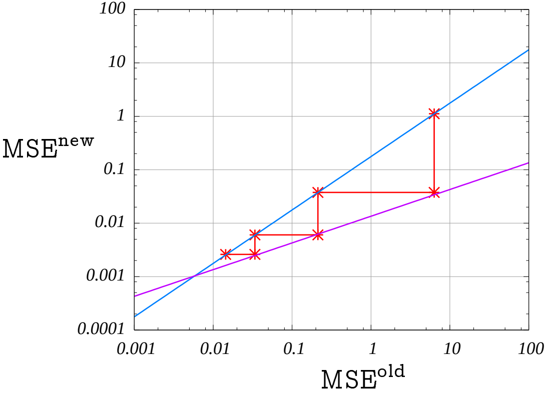
$t = 6$



$t = 7$

$\hat{\theta}^7 =$ 

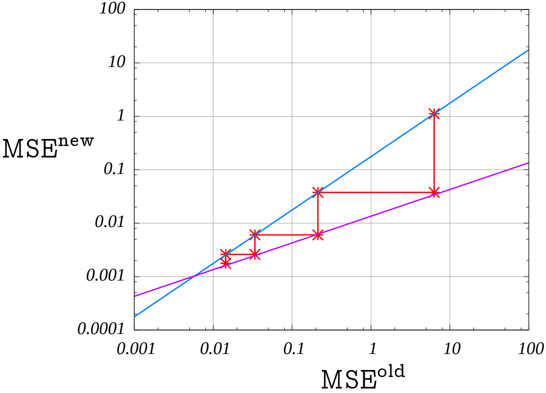
$t = 7$



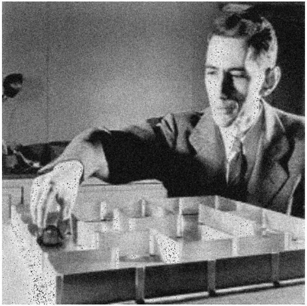
$t = 8$

$\hat{\theta}^8 =$ 

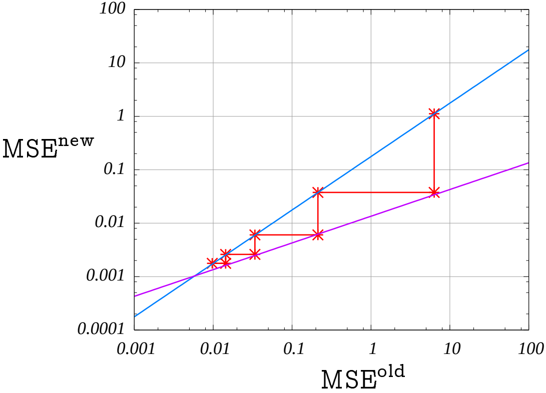
$t = 8$



$t = 9$

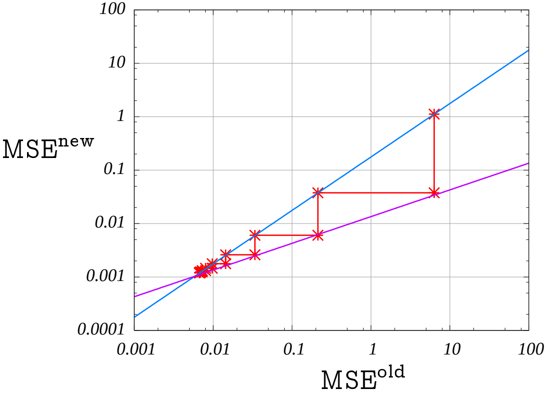
$\hat{\theta}^9 =$ 

$t = 9$



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

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



How do we iterate?

Approximate Message Passing (AMP)

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

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

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

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

(can be computed explicitly)

[Thouless, Anderson, Palmer, 1977, Kabashima, 2003,
Donoho, Maleki, Montanari, 2009, Donoho, Johnstone, Montanari, 2012]

Connection with Belief Propagation

$$m_{i\to j} = m_i + \varepsilon_{i\to j}$$

Linearize in $\varepsilon_{i\to j}$

Very different from naive mean field!

Connection with Belief Propagation

$$m_{i\to j} = m_i + \varepsilon_{i\to j}$$

Linearize in $\varepsilon_{i\to j}$

Very different from naive mean field!

Connection with Perturbation, Optimization, Statistics?

- Robustness wrt $p_X \in$DistributionClass
  ($\eta =$ minimax denoider in DistributionClass)

- Can rigorously track evolution over A random
  (What about $A = A_{\rm det} + \epsilon\, A_{\rm rand}$?)

Connection with Perturbation, Optimization, Statistics?

- Robustness wrt $p_X \in$DistributionClass
  ($\eta =$ minimax denoider in DistributionClass)

- Can rigorously track evolution over A random
  (What about $A = A_{\rm det} + \epsilon\, A_{\rm rand}$?)

A list of theorems/pointers

A list of theorems/pointers

- Connection with optimization [Bayati, Montanari, 2011]
- Minimax theory for sparse/block-sparse/TV [Donoho, Johnstone, Montanari, 2012]
- Bayesian reconstruction up to information dimension [Donoho, Javanmard, Montanari, 2012]
- Universality [Bayati, Lelarge, Montanari, 2012]
- Analysis of Generalized AMP [Javanmard, Montanari, 2012]
- Application to sparse PCA [In preparation, 2013]

Related work

- (Non-rigorous) replica method.
  [Tanaka 2002, Guo, Verdú 2005, Kabashima, Tanaka 2009, Rangan, Fletcher, Goyal 2009, Caire, Tulino, Shamai, Verdú 2012...]

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

- Generalized linear models [Rangan 2011]

- Graphical model priors [Schniter et al. 2010-...]

- Low-rank matrices [Rangan, Fletcher 2012]

Conclusion

Conclusion

- Can do message passing/BP without Bayesian assumptions!
- There is something between naive mean field and BP

Thanks!

Noise distribution?