Approximate Message Passing

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

Stanford University

December 8, 2012
Statistical estimation

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

- \( \theta \rightarrow \text{Unknown object} \)
- \( y \rightarrow \text{Observations} \)
- \( f(\cdot; \text{noise}) \rightarrow \text{Parametric model} \)

**Problem:** Estimate \( \theta \) from observations \( y \).
Statistical estimation

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

\[ \theta \rightarrow \text{Unknown object} \]
\[ y \rightarrow \text{Observations} \]
\[ f(\cdot; \text{noise}) \rightarrow \text{Parametric model} \]

**Problem:** Estimate \( \theta \) 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?

\[
\text{maximize } \text{Likelihood}(\theta|y) - \text{Complexity}(\theta)
\]

‘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?

\[ y \rightarrow \hat{\theta}^1 \rightarrow \hat{\theta}^2 \rightarrow \hat{\theta}^3 \rightarrow \ldots \]

- 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!
\[ \theta = \begin{array}{c}
\end{array} \in \mathbb{C}^n \]

Unknown object \((n = 512^2 \approx 2.5 \cdot 10^5)\)
Noiseless linear measurements

\[ y = A \theta = A. \]

Want to reconstruct \( \theta \)
Noiseless linear measurements

\[ y = A\theta = A. \]

Want to reconstruct \( \theta \)
Measurement structure

\[ A = \tilde{F} R \]
\[ \tilde{F} = \text{subsampled Fourier matrix} \]
\[ R = \begin{bmatrix} +1 & \quad -1 \\ -1 & \quad +1 \\ +1 & \quad -1 \end{bmatrix} = \text{random modulation} \]

→ \( y \in \mathbb{C}^m, m = 0.17n \)
An approach popular in this community

\[ y = A\theta + z, \quad z \sim N(0, \sigma^2 I_{m \times m}) \]

\[ p_{\theta|y}(\theta|y) \propto \exp \left\{ - \frac{1}{2\sigma^2} \|y - A\theta\|_2^2 \right\} p_\theta(\theta) \]

\[ \propto \prod_{a=1}^{m} \exp \left\{ - \frac{1}{2\sigma^2} (y_a - A_a^T \theta)^2 \right\} \prod_{i=1}^{n} p_{\Theta_i}(\Theta_i) \]
An approach popular in this community

\[ y = A\theta + z , \quad z \sim N(0, \sigma^2 I_{m \times m}) \]

\[
p_{\Theta | Y}(\theta | y) \propto \exp \left\{ -\frac{1}{2\sigma^2} \| y - A\theta \|_2^2 \right\} p_\Theta(\theta) \]

\[
\propto \prod_{a=1}^{m} \exp \left\{ -\frac{1}{2\sigma^2} (y_a - A_a^T \theta)^2 \right\} \prod_{i=1}^{n} p_{\Theta_i}(\theta_i) \]
An approach popular in this community

\[ y = A\theta + z , \quad z \sim N(0, \sigma^2 I_{m \times m}) \]

\[ p_{\Theta | Y}(\theta | y) \propto \exp \left\{ - \frac{1}{2\sigma^2} \| y - A\theta \|_2^2 \right\} p_{\Theta}(\theta) \]

\[ \propto \prod_{a=1}^{m} \exp \left\{ - \frac{1}{2\sigma^2} (y_a - A_T a \theta)^2 \right\} \prod_{i=1}^{n} p_{\Theta_i}(\theta_i) \]
\[
p_{\Theta | Y}(\theta | y) \propto \prod_{a=1}^{m} \exp \left\{ -\frac{1}{2\sigma^2} (y_a - A_a^T \theta)^2 \right\} \prod_{i=1}^{n} p_{\Theta_i}(\theta_i)
\]

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?

\[ \mathbb{E} \hat{\theta}^1 = \text{(one line calculation)} = \theta \]
\[ \hat{\theta}^1 = A^\dagger y = \]

\[ \theta = \]

Does not look that good!
Check it out

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

Does not look that good!
Idea

\[ \text{_noise} \]
Idea

\[ \text{signal} = \text{clean} + \text{noise} \]
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

\[ \text{MSE}^{\text{out}} = \frac{1 - \delta}{\delta} \text{MSE}^{\text{in}} \]
Let's check

\[ \text{MSE}^{\text{out}} \]

\[ \text{MSE}^{\text{in}} \]
Denoising

\[ \hat{\theta}^1 \approx \theta + \sigma z, \quad z_i \sim N(0, 1) \]

Idea: Treat \( \hat{\theta}^1 \) as effective observations in denoising
Denoising

\[ \hat{\theta}^1 \approx \theta + \sigma z, \quad z_i \sim N(0, 1) \]

**Idea:** Treat \( \hat{\theta}^1 \) as effective observations in denoising
Denoising by nonlocal means

\[ \hat{y} = \theta + \sigma z , \]

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

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

[Buades, Coll, Morel, 2005]

\[ \hat{\theta} \equiv \eta(\hat{y}) \]
Denoising by nonlocal means

\[ \hat{y} = \theta + \sigma z, \]

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

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

[Buades, Coll, Morel, 2005]

\[ \hat{\theta} \equiv \eta(\hat{y}) \]
Denoising by nonlocal means

\[ \hat{y} = \theta + \sigma z, \]

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

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

[Buades, Coll, Morel, 2005]

\[ \hat{\theta} \equiv \eta(\hat{y}) \]
Will it work?

$$\hat{\theta}^2 = \eta(\hat{\theta}^1) = \eta(A^\dagger y) = \eta(\quad)$$
Let's try

\[ \hat{\theta}^1 = A^\dagger y = \]

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

Better than garbage!
How much better?

\[ \text{MSE}^{\text{out}} \]

\[ \text{MSE}^{\text{in}} \]
How much better?

\[ \text{MSE}^{\text{out}} \]

\[ \text{MSE}^{\text{in}} \]

\[ c_1 x \]

\[ c_2 \sqrt{x} \]

\[ ? \]
Let us repeat the denoising experiment
Let us repeat the denoising experiment: $y = \theta + \sigma z$
Quantitatively
Quantitatively

![Graph showing the relationship between \( \text{MSE}^{\text{in}} \) and \( \text{MSE}^{\text{out}} \). The graph displays a linear relationship with two lines: \( c_1 x \) and \( c_2 \sqrt{x} \).]
Approximate denoiser characterization

\[ \text{MSE}^{\text{out}} = c \sqrt{\text{MSE}^{\text{in}}} \]

(see also Maleki, Baraniuk, Narayan, 2012)

(Arias-Castro, Willett, 2012)
What we achieved so far

\[ \text{Matched filter} \quad \text{Denoiser} \]
What we achieved so far

**Matched filter**

**Denoiser**
What we achieved so far

Andrea Montanari (Stanford)
What we achieved so far

What about iterating?
What we achieved so far

What about iterating?

MSE\textsuperscript{new} vs. MSE\textsuperscript{old}
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!
\( t = 4 \)

\[ \hat{\theta}^4 = \]
\( t = 4 \)
\[ t = 5 \]

\[ \hat{\theta}^5 = \]
$t = 5$
\[ t = 6 \]

\[ \hat{\theta}^6 = \]
$t = 6$
\[ t = 7 \]

\[ \hat{\theta}^7 = \]
$t = 8$

$\hat{\theta}^8 =$
\[ t = 8 \]
\[ t = 9 \]

\[ \hat{\theta}^9 = \]
\( t = 9 \)
\( t = 0, 1, 2, 3, \ldots, 20 \)
\( t = 0, 1, 2, 3, \ldots, 20 \)
How do we iterate?
Approximate Message Passing (AMP)

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

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

\[ r^t = y - A\hat{\theta}^{2t} + b_t r^{t-1} \]

\[ b_t = \frac{1}{m} \text{div} \eta(\hat{\theta}^{2t-1}) \]

(can be computed explicitly)

Connection with Belief Propagation

$m_{i \rightarrow j} = m_i + \epsilon_{i \rightarrow j}$

Linearize in $\epsilon_{i \rightarrow j}$

Very different from naive mean field!
Connection with Belief Propagation

\[ m_{i \rightarrow j} = m_i + \varepsilon_{i \rightarrow j} \]

Linearize in \( \varepsilon_{i \rightarrow j} \)

Very different from naive mean field!
Connection with Perturbation, Optimization, Statistics?

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

- Can rigorously track evolution over $A$ random 
  (What about $A = A_{\text{det}} + \epsilon A_{\text{rand}}$?)
Connection with Perturbation, Optimization, Statistics?

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

- Can rigorously track evolution over $A$ random
  (What about $A = A_{\text{det}} + \epsilon A_{\text{rand}}$?)
A list of theorems/pointers
A list of theorems/pointers

- Connection with optimization [Bayati, Montanari, 2011]
- Bayesian reconstruction up to information dimension [Donoho, Javanmard, Montanari, 2012]
- Universality [Bayati, Lelarge, Montanari, 2012]
- Application to sparse PCA [In preparation, 2013]
Related work

- (Non-rigorous) replica method.

- 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?

Andrea Montanari (Stanford)