Generalized Approximate Message Passing and Interference Coordination

Sundeep Rangan, Polytechnic Institute of NYU

Joint work with Alyson Fletcher (Berkeley), Vivek Goyal (MIT), Ulugbek Kamilov (EPFL/MIT), Phil Schniter (Ohio State) and Ritesh Madan (Qualcomm)

SuRI, EPFL
22 June 2011
Outline

Generalized approximate messaging (GAMP)
- Graphical model approach for estimation with linear mixing
- Analysis of GAMP
  - Global convergence and optimality with large random matrices
- Numerical examples:
  - Compressed sensing, quantization, …
- Turbo-GAMP
- More applications
  - HetNets and Inter-Cellular Interference Coordination
  - Neural connectivity mapping
Estimation with Linear Mixing

- **Problem**: Estimate $\mathbf{x}$ and $\mathbf{z}$ given $\mathbf{q}$, $\mathbf{y}$ and $\mathbf{A}$,
- Linear mixing arises in many estimation problems:
  - Linear regression, communication channels, source coding transforms, …
  - Many examples in this talk.
- **Challenge**: Generically, optimal estimation is hard
Linear Mixing and Coupling

- When $A = I$: estimation decouples into $n$ scalar problems.
  - Posterior distribution and MMSE estimates can be easily computed
  - Computations involve $n$ one-dimensional integrals
- For general $A$, how can we “decouple” the estimation problem?
Decoupling via Graphical Models

- Graphical models: A general framework for “decoupling” large vector-valued problems.
- Exploit that many distributions factor:

\[
p(x | y) = \frac{1}{Z(y)} \prod_{i=1}^{m} f_i(x_{\alpha(i)}, y),
\]

- \(x_{\alpha(i)} = \) Sub-vector with a small number of terms

Factor graph representation

- Variable nodes
- Factor nodes

Wireless Research Lab
Graphical Models and Belief Propagation

- **Loopy belief propagation:**
  - Approximate “global” inference through sequence of “local” problems.
  - Pass “messages” between factors representing estimates of marginal distribution of each variable.

- Significant computational savings when each factor depends on small number of terms.

Factor Graph for Linear Mixing Estimation

\[
p(x | y) = \frac{1}{Z(y)} \prod_{i=1}^{m} p(y_i | z_i) \prod_{j=1}^{n} p(x_j)
\]

- Posterior \( p(x | y) \) factors due to separability assumptions
- Output factors and variables **coupled** by matrix \( A \)
- Can apply BP when coupling is sparse.
Sum-Product BP for Linear Mixing

Output nodes

Compute scalar likelihood function

\[ p_{i \rightarrow j}(x_j) = E\left[ p_{Y|Z}(y_i | z_i) | y_i, x_j \right] \]

\[ z_i = a_i^T x, \quad x_r \sim p_{i \leftarrow r}(x_r) \]

Input nodes

Compute scalar posterior distribution

\[ p_{i \leftarrow j}(x_j) = p(x_j) \prod_{k \neq j} p_{k \rightarrow j}(x_j) \]

- For a sparse graph, BP decouples vector estimation into iterative sequence of low-dimensional estimation problems
- But, computational complex for dense graphs
  - Difficulty is at the output node
Approximate Message Passing Concept

\[
p_{i \rightarrow j}(x_j) = E \left[ p(y_i | z_i) | x_j, y_i \right]
\]

\[
= E \left[ \frac{y_i}{A_{ij} x_j + \sum_{k \neq j} A_{ik} x_k} \right] | x_j, y_i
\]

\[
= EV \left[ \frac{y_i}{A_{ij} x_j + V} \right], \quad V \sim N(0, \sigma^2)
\]

\[
= D_1 A_{ij} x_j + \frac{1}{2} D_2 A^2_{ij} x_j^2
\]

- Suppose \( A \) is large dense matrix
  - Individual terms \( A_{ij} \) are relatively small
- Simplify computation at output nodes via Central Limit Theorem and quadratic approximation
GAMP Iterations

\[ \hat{p}(t) = A\hat{x}(t) - \tau^P(t)\hat{s}(t-1) \]

Output nodes

\[ \hat{z}(t) = E \left[ z \mid y, \hat{p}(t), \tau^P(t) \right] \]

Linear transform

\[ \hat{s}(t) = \frac{1}{\tau^P(t)}(\hat{z}(t) - \hat{p}(t)) \]

m scalar AWGN estimation problems:

\[ y_i \sim p(y_i \mid z_i), \quad z_i \sim N\left(\hat{p}_i(t), \tau^P(t)\right) \]

Input nodes

\[ \hat{r}(t) = \hat{x}(t) + \tau^S(t)A^*\hat{s}(t) \]

Linear transform transpose

\[ \hat{x}(t+1) = E \left[ x \mid \hat{r}(t), \tau^R(t) \right] \]

n scalar AWGN estimation problems:

\[ \hat{r}_j(t) = x_j + v_j, \quad x_j \sim p(x_j), \quad v_j \sim N\left(0, \tau^R(t)\right) \]

- Resulting GAMP algorithm is computationally very simple.
- Iterative decoupling to scalar problems
History

- Gaussian approximations of BP originally used in CDMA multi-user detection with many extensions:
  - Boutros & Caire (02), Montanari & Tse (06), Guo & Wang (06), Tanaka & Okada (06), Donoho, Maleki & Montanari (09).
  - Many names: Approximate message passing (AMP), Approx BP, relaxed BP, parallel interference cancellation (PIC), ….

- Similar structure in iterative separable approximation:
  - IST (Daubechies et al 04, Figurero, Nowak 03, Combettes, Wajs 05), SpaRSA (Wright, Nowak, Figuero 09), …

- Generalized AMP: Extension of AMP methods to arbitrary (non-AWGN) output channels [Rangan 10]
Outline

- Generalized Approximate Messaging (GAMP)
  - Graphical model approach to estimation with linear mixing

Analysis

- Global convergence with large random matrices

Numerical examples:
  - Compressed sensing, quantization, …

- Turbo-GAMP

- More applications
  - HetNets and Inter-Cellular Interference Coordination
  - Neural connectivity mapping
Analysis with Large Random Matrices

- Analyze estimator algorithms for large, random i.i.d. matrices $\mathbf{A}$
- Large random i.i.d. $\mathbf{A}$ arise in many contexts:
  - Estimation with random input data
  - Communication and coding with random spreading
  - Model for high-dimensional data
  - Analytically tractable test case to compare algorithms
- “Blessing of dimensionality”: Many large random problems become easy to analyze / solve in high-dimensions
Asymptotic Scalar Equivalent Model

Vector estimation

\[ z = Ax \]

\[ p(y|z) \]

\[ \hat{x}(t) \]

GAMP vector estimate

Scalar equivalent estimation

AWGN power

\[ \tau^r(t) \]

Scalar MMSE estimate

**Theorem:** For large, i.i.d. Gaussian transform matrices:

- Distribution of each component of the vector GAMP estimate converges to random variables from a scalar equivalent model.
- Enables exact computation of MSE or any other separable performance metric of GAMP estimate at each iteration \( t \).
- Rigorously justifies asymptotic decoupling.
Intuition via a Naïve Algorithm

Scalar output estimation

\[ N(0, \tau^p(t)) \]

\[ p_i(t) \xrightarrow{y_i} z_i \xrightarrow{p_Y|Z} \hat{z}_i(t) \xrightarrow{\text{Scalar MMSE}} \]

\[ p_i(t) = \sum_j A_{ij} \hat{x}_j(t) = z_i + N(0, \tau^p(t)) \]

\[ \tau^p(t) = E\left[ (z_i - \hat{z}_i(t))^2 \mid p_i(t) \right] \]

Scalar input estimation

\[ N(0, \tau^r(t)) \]

\[ x_j \xrightarrow{r_j(t)} r_j(t) \xrightarrow{\text{Scalar MMSE}} \hat{x}_j(t+1) \]

\[ r_j(t) = \sum_i A_{ij} \hat{z}_i(t) = x_j + N(0, \tau^r(t)) \]

\[ \tau^r(t) = E\left[ (x_j - \hat{x}_j(t))^2 \mid r_j(t) \right] \]

- GAMP performance matches naïve analysis of simple iterative algorithm
  - Scalar state evolution equations.
- Actual GAMP algorithm = iterative algorithm + correction term.
## Related Scalar Equivalent Results

<table>
<thead>
<tr>
<th>Analysis</th>
<th>Estimator</th>
<th>Scalar equivalent characterization</th>
</tr>
</thead>
<tbody>
<tr>
<td>Marcenko-Pastur</td>
<td>Linear estimators</td>
<td>MSE predicted by soln to quadratic fixed-point (FP) equation (Verdu-Shamai 99, Tse-Hanly 99)</td>
</tr>
<tr>
<td>Replica analysis (non-rigorous)</td>
<td>Opt. MMSE, postulated MAP &amp; MMSE</td>
<td>Equiv. AWGN noise soln to a nonlinear FP equation. [Tanaka05, Guo-Verdu05, Rangan, Fletcher, Goyal 09]</td>
</tr>
<tr>
<td>Large sparse random matrices</td>
<td>GAMP for MMSE with correct prior</td>
<td>Equiv. AWGN noise is the largest (worst) solution to the FP equation.</td>
</tr>
<tr>
<td></td>
<td>Optimal MMSE estimator</td>
<td>Smallest FP soln is lower bound on any estimator. Corollary: When FP soln is unique, AMP is optimal [Boutros-Caire02, Montanari-Tse05, Guo-Wang06,07, Rangan 10]</td>
</tr>
<tr>
<td>Large dense i.i.d. Gaussian</td>
<td>GAMP MAP or MMSE, possibly mismatched prior</td>
<td>Equiv. AWGN noise is the largest (worst) solution to the FP equation. [Bayati-Montanari 09, Rangan 10]</td>
</tr>
</tbody>
</table>
Outline

- Generalized Approximate Messaging (GAMP)
  - Graphical model approach to estimation with linear mixing
- Analysis
  - Global convergence with large random matrices
- Numerical examples:
  - Compressed sensing, quantization, …
- Generalizations
- More applications:
  - HetNets and Inter-Cellular Interference Coordination
  - Neural connectivity mapping
Example 1: Compressed Sensing

[Gauss-Bernoulli input (10% sparsity), AWGN output. SNR=10 dB

- G-AMP outperforms Lasso

- < 0.8 dB from optimal at n=100.

- <0.2 dB at n=500 [not shown]
MSE Prediction of LASSO

Replica method also provides excellent prediction of LASSO and linear performance at moderate dimensions.

- Much tighter than deterministic bounds
- Related work for partial support recovery [Reeves & Gastpar 10]

RIP-based analysis applies for $m > 7595$ [Candès & Tao (2005)]

Conventional analyses require *oversampling*!

$n = 100$

$\rho = 0.1$
Ex 2: Bounded Noise Estimation

- Gaussian input with bounded noise output
- Arises in quantization
- NP-hard problem
- Relaxed BP close to optimal at \( n=50 \) and outperforms best known reconstruction methods
Further Applications to Quantization

- Reconstruction of sparse vectors from quantized random transforms (see right).

- Practical Slepian-Wolf codes via transforms and binning.
  - Possibly integrate G-AMP with turbo code (similar to Schniter 2011)

[Kamilov, Goyal & Rangan, 2011]
Outline

• Generalized Approximate Messaging (GAMP)
  • Graphical model approach to estimation with linear mixing
• Analysis
  • Global convergence with large random matrices
  • Local convergence for arbitrary matrices
• Numerical examples:
  • Compressed sensing, quantization, …

Turbo-GAMP
• More applications:
  • HetNets and Inter-Cellular Interference Coordination
  • Neural connectivity mapping
Turbo-GAMP [Rangan, Fletcher, Goyal & Schniter 2011]

- GAMP can be incorporated in graphical models via partitioning:
  - Weak, linearizable edges
  - Strong, nonlinear edges
- Run GAMP messages on weak edges
- Combine with regular BP updates in “turbo” manner
- Complexity-performance tradeoff
Example: Group Sparsity

Group sparse prior
Non-zero components of $x$ are in groups

$$x = [x_1 \cdots x_n], \quad x_j = [x_{j1} \cdots x_{jL}]$$

- Suppose sparse vector $x$ has a group sparse structure:
  $$p(x_{j\ell} | \xi_j) = \begin{cases} 
0 & \text{if } \xi_j = 0 \\
N\left(0, \sigma^2_x\right) & \text{if } \xi_j = 1
\end{cases}$$
  $\xi_j = 0$ or $1$ indicating whether $j$-th group is active

- Overall distribution can be represented in a factor graph with variables $\xi_j$ added as latent variables
Turbo-GAMP for Group Sparsity

n=100 groups of d=4 variables
each group nonzero w.p. 0.1
nonzero groups jointly Gaussian
i.i.d. Gaussian A

Costs:
Grp OMP: \(O(kmnd^2)\)
Grp LASSO: \(O(mnd)\) per iter
Turbo-GAMP: \(O(mnd)\) per iter

Turbo-GAMP has best performance at lowest cost and is greatly more general than this application
Turbo-GAMP for General Models

- Similar procedure for group sparsity can be used for more general graphical models
- Can incorporate correlations between variables or noise, unknown parameters in distributions
- Other applications in wavelet image denoising and turbo equalization (Schniter et al 10)

Graphical model where linear relationships are combined with general factor graphs at input and output.
Outline

• Generalized Approximate Messaging (GAMP)
  • Graphical model approach to estimation with linear mixing

• Analysis
  • Global convergence with large random matrices

• Numerical examples:
  • Compressed sensing, quantization, …

• Turbo-GAMP

More applications:

• HetNets and Inter-Cellular Interference Coordination
• Neural connectivity mapping
Femtocells: A Personal Base Station

- Small, personal base station in customer’s premises, but in operator’s spectrum.
- Possibility for greatly increased capacity at low cost:
  - Replace large macrocell tower and related RF equipment with small access point.
  - No operator deployment or maintenance.
  - Offload traffic to subscriber’s backhaul.

The Changing Interference Environment

Loss from randomness (~2dB)

Very bad links (restricted assoc)

Very good links (SNR>10 dB)
Mitigating Strong Interference
Lessons from Information Theory

- Reuse 1 and power control has been the dominant model for cellular systems, esp. since CDMA
- But, stronger interference conditions in Het Nets requires more sophisticated methods.
GAMP Interference Coordination

[Rangan & Madan 2011]

TX nodes

\[ x_1 \]

\[ x_2 \]

\[ x_n \]

TX vectors

RX nodes

\[ f_1(x_1, z_1) \]

\[ f_2(x_2, z_2) \]

\[ f_n(x_n, z_n) \]

Utility functions

Linear interference:
\[ z = Ax \]

Can apply turbo-GAMP procedure with a natural partitioning
- Strong edges: Direct link & strong interferers
- Weak edges: Weak interferes
Linear Interference Models

- Linear mixing interference model provides for generality to handle complex cellular systems
- Encompasses many complex interference models
  - Power control with flat fading:
  - Multiple subbands
  - Beamforming, linear precoding
- Utility $f_i(x_i, z_i)$ can capture arbitrary functions of SINR
- Related BP work for scheduling: Sanghavi, Malioutov & Willsky (09), Bayati, Shah and Sharma (09)
GAMP-Based Multi-Round Protocol

**Round 0**
- TX vector $x_1(0)$
- Sensitivity $D_2(0)$
- Interference $z_1(0)$ and sensitivity $D_1(0)$

**Round 1**
- TX vector $x_1(1)$
- Sensitivity $D_2(1)$
- Interference $z_1(1)$ and sensitivity $D_1(1)$

Data transmission
- Data scheduled along TX vector $x_1$
Example 1: On-off channels, flat fading

- 3GPP femto apartment model, 10 dB wall loss, 50% penetration, 5MHz
- PF scheduler (log utility) with fast flat fading.
- BP rerun in each time slot
- BP shows significant improvement over reuse 1.
  - Improves low rates via orthogonalization
  - Multi-user diversity gain
- Close to optimal performance
Example 2: Subband scheduling

- Subband scheduling with 4 subbands
  - Subband allocation one of the main features of LTE
  - Complex nonlinear problem
- Single static allocation
- BP shows significant improvement over reuse 1.
- Close to optimal performance (but still room for improvement)
Neural Mapping via Multi-Neuron Excitation

[Fletcher, Rangan, Varshney & Bhargava 2011]

- Estimating neural connectivity is typically performed one neuron at a time
- Most measurements wasted:
  - Even nearby neurons unlikely to be connected (say, 10%)
  - Long-distance connections rare but important
- Hu & Chklovskii (NIPS2009) suggested multi-neuron excitation and compressed sensing recovery

C. elegans connectome, Varshney, et al. (‘09)
LNP Model for Multi-Neuron Excitation

- Synaptic weights can be estimated via GAMP-based iteration:
  - Model unknowns synaptic weights $x$ as prior
  - Given nonlinearity, can estimate $x$ from excitation and counts $y$
  - Update estimate of nonlinearity from $z$ and iterate
Simulation: Connectivity Detection

- Compared GAMP with:
  - RC: reverse correlation, non-sparse
  - CoSAMP: linear compressed sensing method
- In each algorithm, can tradeoff missed detections and false alarms by varying threshold level on estimates
- GAMP results in much lower missed detection rate for same false alarm probability.

Comparison of methods at $m = 300$ measurements
Conclusions

• GAMP algorithm provides computationally simple algorithm for large class of non-linear estimation problems.

• Asymptotically exact analysis for large, i.i.d. transforms:
  • Scalar equivalent model

• Among the fastest algorithm computationally

• Modular (can be integrated into general graphical models)

• Many applications
  • Compressed sensing, quantization, interference coordination, neural mapping, …

• Try it out: All code is available at
  http://gampmatlab.sourceforge.net
Related Papers