Performance Trade-Offs in Multi-Processor Approximate Message Passing

Junan Zhu*, Ahmad Beirami+, & Dror Baron*

*North Carolina State University
+MIT and Duke University

International Symposium on Information Theory
Barcelona, Spain; July 10-15, 2017

The work was supported by the National Science Foundation under the Grant CCF-1217749
Motivation
Big Data and Distributed Systems

- Ever increasing data size
- Impossible to use single machine
- Distributed systems needed [Dean & Ghemawat ’08]
Costs of Solving Big Problems

• What costs do we encounter?
  – Computation
  – Communication
  – Quality of result
  – Others?

• How are costs related?

• Study distributed reconstruction algo as example

• May extend to other iterative algos
• Input signal $x \in \mathbb{R}^N$
• Measurements taken by matrix $A \in \mathbb{R}^{M \times N}$
• Additive noise $z$ with variance $\sigma_z^2$
• Measurement rate $\kappa = \frac{M}{N} > 0$
• Estimate $x$ from $y$, $A$, and statistical info on $x$, $z$
Applications of Linear Models

- Medical Imaging
- CDMA
- Seismic Imaging
- Financial Prediction
Approximate Message Passing (AMP)
Approximate Message Passing (AMP)
[Donoho et al. 2009]

\[
\begin{align*}
\mathbf{y} &= \mathbf{A} \in \mathbb{R}^{M \times N} \mathbf{x} + \mathbf{z} \\
&\sim \mathcal{N}(0, \sigma_Z^2 \mathbb{I})
\end{align*}
\]

Iterate:

Residual \quad \mathbf{r}^t = \mathbf{y} - \mathbf{A}\mathbf{x}^t + \frac{\mathbf{r}^{t-1}}{\kappa} < d\eta_t(\mathbf{x}^{t-1} + \mathbf{A}^T \mathbf{r}^{t-1}) >

Pseudo-data \quad \mathbf{f}^t = \mathbf{x}^t + \mathbf{A}^T \mathbf{r}^t = \mathbf{x} + \mathbf{w}^t

Denoising \quad \mathbf{x}^{t+1} = \eta_t(\mathbf{f}^t)

Equivalent scalar channel
Approximate Message Passing (AMP)

[Donoho et al. 2009]

\[ y = Ax + z \sim \mathcal{N}(0, \sigma^2_z I) \]

Iterate:

Residual \[ r^t = y - Ax^t + \frac{r^{t-1}}{\kappa} < d\eta_t(x^{t-1} + A^T r^{t-1}) > \]

Pseudo-data \[ f^t = x^t + A^T r^t = x + w^t \]

Denoising \[ x^{t+1} = \eta_t(f^t) \]
Approximate Message Passing (AMP)  
[Donoho et al. 2009]

\[ y = Ax + z \]

Iterate:

Residual
\[ r^t = y - Ax^t + \frac{r^{t-1}}{\kappa} < d\eta_t(x^{t-1} + A^T r^{t-1}) > \]

Pseudo-data
\[ f^t = x^t + A^T r^t = x + w^t \]

Denoising
\[ x^{t+1} = \eta_t(f^t) \]

State evolution
\[ \sigma_{t+1}^2 = \sigma_Z^2 + \frac{1}{\kappa} \text{MSE}(\eta_t, \sigma_t^2) \]
Multi-Processor AMP (MP-AMP)
Multi-Processor Linear System

[Patterson et al. 2013, Han et al. 2014]

- Matrix $\mathbf{A}$ could be big
- $\mathbf{A}$ stored in distributed nodes
- Node $p$ processes $y_p = z_p + \mathbf{A}_p \mathbf{x}$
Multi-Processor AMP (MP-AMP)

[Han et al. 2014]

Centralized AMP
- Calculate residual $r_t$
- Calculate pseudo-data $f_t$
- Denoise $f_t$

MP-AMP
- $P$ distribute nodes
- $r^p_t$: residual in node $p$
- $f^p_t$: pseudo-data in node $p$

\[
f_t = \sum_{p=1}^{P} f^p_t
\]
- Denoise $f_t$
- Fusion center
MP-AMP

- Messages: uplink $f_t^p$ and downlink $x_{t+1}$
- Compress messages to reduce communication
- Focus on lossy compression of $f_t^p$
- Lossy compression of $x_{t+1}$ - future work
Lossy MP-AMP
Rate-Distortion Theory
[Berger, 1971; Cover & Thomas, 2006]

\[ f_t^p \in \mathbb{R}^N \quad \text{Quantize} \quad Q(f_t^p) \in \mathbb{R}^N \]

- \( R \): Rate (bits/entry) to encode \( Q(f_t^p) \)
- \( D \): Distortion between \( f_t^p \) and \( Q(f_t^p) \)
- \( R \) and \( D \) related through R-D function
- Allowing modest \( D \) can save lots of \( R \)!
Lossy MP-AMP

\[ Q(f^p_t) = \frac{1}{P} \mathbf{x} + w^p_t + n^p_t \] encode w/ \( R \)

State evolution (SE)
\[ \sigma_{t+1}^2 = \sigma_Z^2 + \frac{1}{\kappa} \text{MSE}(\eta_t, \sigma_t^2) \]

Lossy SE
\[ \sigma_{t+1}^2 = \sigma_Z^2 + \frac{1}{\kappa} \text{MSE}(\eta_t, \sigma_t^2 + PD) \]

RD relation \( D = D(R) \)
Trade-offs

- Can use different rates $R_t$ in each iteration

- Three key quantities
  - Number of iterations: $T$
  - Aggregate coding rate: $R_{agg} = \sum_{t=1}^{T} R_t$
  - Quality of the estimate: MSE

- Cannot be minimized simultaneously!
Study of Trade-offs
Achievable Set

• All MSE values achieved by \((T, R_{agg})\) pair:
  \[\mathcal{E}(T, R_{agg})\]

• Achievable set
  \[\mathcal{C} = \{(T, R_{agg}, MSE) \in \mathbb{R}^3_+: MSE \in \mathcal{E}(T, R_{agg})\}\]
Pareto Optimality [Das & Dennis 1998]

• Points: \( \chi_1 = \left( T_1, R_{agg_1}, MSE_1 \right) \in C \)
  \( \chi_2 = \left( T_2, R_{agg_2}, MSE_2 \right) \in C \)

• Point \( \chi_1 \) dominates \( \chi_2 \) if
  \( T_1 \leq T_2, R_{agg_1} \leq R_{agg_2}, EMSE_1 \leq EMSE_2 \)

• Pareto optimal point \( \chi \): no other points dominate \( \chi \)
Achievable Set vs. Pareto Optimality

• Set of all Pareto optimal points:
  \[ P = \{ \chi \in C : \chi \text{ Pareto optimal} \} \]

• Points in \( P \) belong to boundary of \( C \)
Main Result: Achievable Set is Convex!

- Let \((T^{(1)}, R^{(1)}_{agg}, MSE^{(1)}) \in C, (T^{(2)}, R^{(2)}_{agg}, MSE^{(2)}) \in C\)

- Need to show for \(0 < \lambda < 1\), linear combination in set
  \((\lambda T^{(1)} + (1 - \lambda)T^{(2)}, \lambda R^{(1)}_{agg} + (1 - \lambda)R^{(2)}_{agg}, \lambda MSE^{(1)} + (1 - \lambda)MSE^{(2)}) \in C\)

- Proof ideas:
  - Time-sharing
  - Linearity of \(R_{agg}\)
Numerical Demo

Pareto optimal [Das & Dennis 1998]

Surface obtained from dynamic programming (DP) [Zhu & Baron 2016]
Convexity hints at trade-offs among $T, R_{agg}, MSE$

Pareto optimal [Das & Dennis 1998] points on boundary of $C$

$f(x) = 0.1\mathcal{N}(0,1) + 0.9\delta(x)$
$P = 100$ distributed nodes
Measurement rate $\frac{M}{N} = 0.4$
Noise variance $\sigma_Z^2 = \frac{1}{400}$
Interpretation of Convexity

Communication costly → More iterations, less coding rate
Computation costly → More coding rate, few # iterations
Limiting Behavior
- Optimal rate indeed approx. linear when EMSE→0
- Optimal rate obtained from DP [Zhu & Baron 2016]

\[ f(x) = 0.1 \mathcal{N}(0,1) + 0.9 \delta(x) \]
\[ P = 100 \text{ distributed nodes} \]
\[ \text{Measurement rate } \frac{M}{N} = 0.4 \]
\[ \text{Noise variance } \sigma_Z^2 = \frac{1}{400} \]
\[ \text{Comm. more expensive than computation} \]
Define excess MSE:

\[ EMSE_t = MSE_t - MMSE \]

When \( EMSE_t \to 0 \):
- \( D_t \) decays geometrically (implies linear rate)
- \( EMSE_t \) decays geometrically
- \( D_t \) decay as fast as \( EMSE_t \)

Conjecture: \( R_t = C_1 + C_2 t + o_t(1) \) linear!
Costs

- Computation cost $C_3T$

- Communication cost $C_4R_{agg}$

- Combined cost $\Psi = C_3T + C_4R_{agg}$
Scaling of Cost in Low-EMSE Limit

• Fact (1): Due to $R_t = C_1 + C_2 t + o_t(1)$,
  $\Psi = C_3 T + C_4 R_{agg} = O(T) + O(R_{agg}) = O(T^2)$

• Fact (2): EMSE shrinks geometrically

• Combining (1) and (2): $\Psi = O(\log^2(1/\text{EMSE}^*))$
Thank you!
References