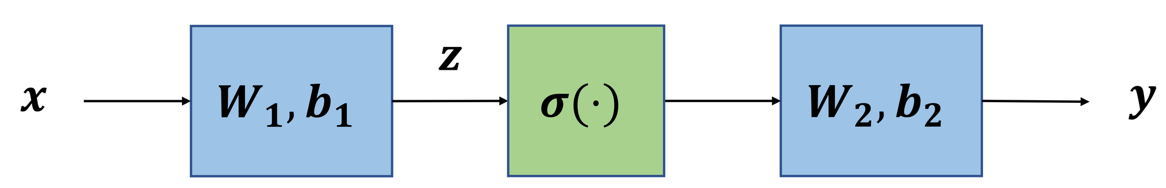
High-Dimensional Analysis of Learning in Two-Layer Models

Parthe Pandit, Mojtaba Sahraee-Ardakan, Alyson K. Fletcher
University of California, Los Angeles

Philip Schniter
Ohio State University

Sundeep Rangan
New York University

Learning Two-Layer Neural Networks



- Two-layer fully connected neural network

$$z = W_1^T x + b_1, \qquad y = W_2^T \sigma(z) + b_2$$

- W_2, b_2 known
- **Goal**: Learn W_1, b_1 from samples $(x_i, y_i), i = 1, \ldots, N$ generated by a ground truth network

Empirical Risk Minimization

Regularized empirical risk minimization

- Learn parameters by minimizing regularized empirical risk

$$\min_{W_1, b_1} \phi(W_1) + \sum_{i=1}^{N} \mathcal{L}(\hat{y}_i, y_i)$$

  - $\phi(\cdot)$ is a regularizer
  - $\mathcal{L}(\cdot,\cdot)$ is a loss function
  - $\hat{y}_i = W_2^T \sigma(z_i) + b_2$ where $z_i = W_1^T x_i + b_1$

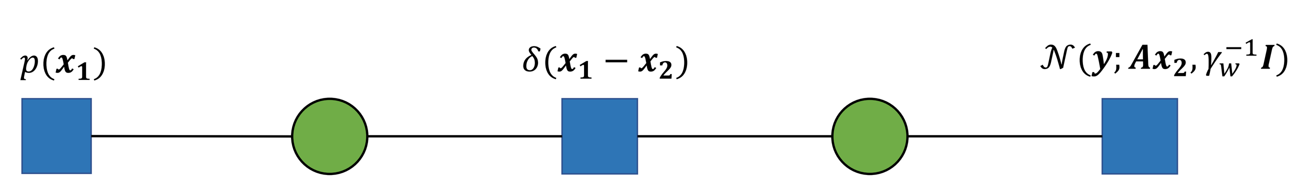
Main Questions

- Optimization typically performed using some variant of SGD
  - Works well in practice
  - But hard to analyze
  - Error bounds not known to be optimal
  - When does learning succeed?
  - What is the generalization error?

This Work: Analysis via Message Passing

- Approximate message passing (AMP) algorithms:
  - Efficiently solve inverse problems
  - Originally developed for linear inverse problems
  - Extended to multi-layer networks (ML-VAMP)
- Key property: State evolution provides performance guarantees in high dimensional regime
- Main contributions:
  - Learning parameters of a two-layer network using AMP framework
  - State evolution provides estimation error at each iteration
  - Predicts when learning will or will not work

Background on Vector AMP



Linear inverse problem

- Model: $x \sim p(x), \quad y = Ax + w, \quad w \sim \mathcal{N}(0, \gamma_w^{-1} I)$
- Posterior density factorizes:

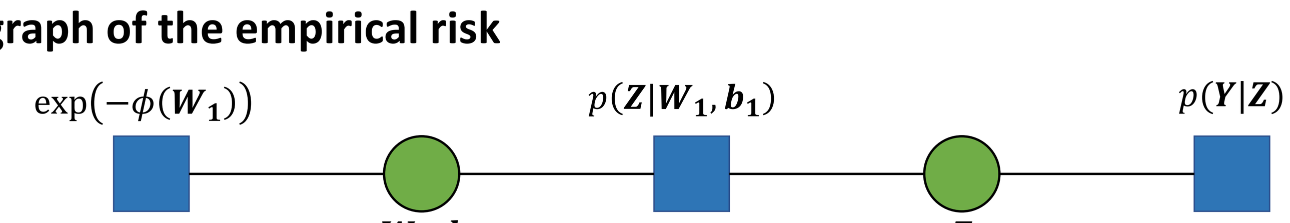
$$p(x|y) \propto \exp\left(-\frac{\gamma_w}{2}\|y - Ax\|^2\right) p(x)$$

- MAP estimation: $\hat{x} = \text{argmin}_x \frac{\gamma_w}{2}\|y - Ax\|^2 - \log p(x)$
- Apply expectation propagation-like inference over factor graph
  - Gaussian approximation of messages
- In each iteration:
  - Linear Gaussian estimation
  - Estimation with prior; separable for separable priors

Properties of VAMP

- State evolution
  - Behavior in each iteration exactly explained for large random A
  - Provides MSE in each iteration
- Provably Bayes optimal in certain cases
  - Including non-convex cases
- Relates to ADMM with carefully chosen adaptive step size

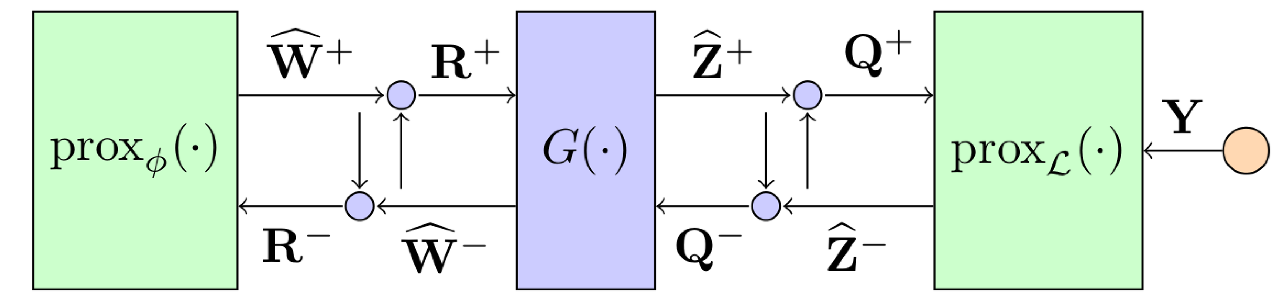
Factor graph of the empirical risk



- Apply expectation propagation over this graph to get matrix AMP

Acknowledgements:
The work of P. Pandit, M. Sahraee-Ardakan, A. K. Fletcher was supported in part by the NSF Grants 1254204 and 1738286, and the Office of Naval Research Grant N00014-15-1-2677. S. Rangan was supported in part by the NSF Grants 1116589, 1302336, and 1547332, NIST, the industrial affiliates of NYU WIRELESS, and the SRC. The work of P. Schniter was supported in part by the National Science Foundation Grant 1716388.

Matrix AMP



Algorithm 1 Matrix Approximate Message Passing

Require: K_{it}, $\text{prox}_\phi(\cdot)$, $\text{prox}_\mathcal{L}(\cdot)$, and linear solver: $\_, \_ = G(\cdot)$

Select initial $\mathbf{R}_0^- \in \text{Range}(\mathbf{X})$ and $\Gamma_0^- \succ 0$.
for $k = 0, 1, \ldots, K_{\text{it}}$ **do**
  // Forward updates
  $\widehat{\mathbf{W}}^+ \leftarrow \text{prox}_\phi(\mathbf{R}^-, \Gamma^-) = \underset{\mathbf{W}}{\text{argmin}}\ \phi(\mathbf{W}) + \frac{1}{2}\text{tr}((\mathbf{W} - \mathbf{R}^-)\Gamma^-(\mathbf{W} - \mathbf{R}^-)^\top)$
  $\Phi^+ \leftarrow \langle \nabla_\mathbf{R}\ \text{prox}_\phi(\mathbf{R}^-, \Gamma^-)\rangle^{-1}\Gamma^-$, $\Gamma^+ \leftarrow \Phi^+ - \Gamma^-$
  $\mathbf{R}^+ \leftarrow (\widehat{\mathbf{W}}^+\Phi^+ - \mathbf{R}^-\Gamma^-)(\Gamma^+)^{-1}$
  $\_, \hat{\mathbf{Z}}^+ \leftarrow G(\mathbf{R}^+, \mathbf{Q}^-, \Gamma^+, \Upsilon^-)$
  $\Lambda^+ \leftarrow \langle \nabla_\mathbf{Q}\, G(\mathbf{R}^+, \mathbf{Q}^-, \Upsilon^+, \Upsilon^-)\rangle^{-1}\Upsilon^-$, $\Upsilon^+ \leftarrow \Lambda^+ - \Upsilon^-$
  $\mathbf{Q}^+ \leftarrow (\hat{\mathbf{Z}}^+\Lambda^+ - \mathbf{Q}^-\Upsilon^-)(\Upsilon^+)^{-1}$

  // Backward updates
  $\hat{\mathbf{Z}}^- \leftarrow \text{prox}_\mathcal{L}(\mathbf{Q}^+, \Upsilon^+) = \underset{\mathbf{Z}}{\text{argmin}}\ \mathcal{L}(\mathbf{Z}|\mathbf{Y}) + \frac{1}{2}\text{tr}((\mathbf{Z} - \mathbf{Q}^+)\Upsilon^+(\mathbf{Z} - \mathbf{Q}^+)^\top)$
  $\Lambda^- \leftarrow \langle \nabla_\mathbf{Q}\ \text{prox}_\mathcal{L}(\mathbf{Q}^+, \Upsilon^+)\rangle^{-1}\Upsilon^+$, $\Upsilon^- \leftarrow \Lambda^- - \Upsilon^+$
  $\mathbf{Q}^- \leftarrow (\hat{\mathbf{Z}}^+\Lambda^+ - \mathbf{Q}^+\Upsilon^+)(\Upsilon^-)^{-1}$
  $\widehat{\mathbf{W}}^-, \_ \leftarrow G(\mathbf{R}^+, \mathbf{Q}^-, \Gamma^+, \Upsilon^-)$
  $\Phi^- \leftarrow \langle \nabla_\mathbf{R}\, G(\mathbf{R}^+, \mathbf{Q}^-, \Gamma^+, \Upsilon^-)\rangle^{-1}\Gamma^+$, $\Gamma^- \leftarrow \Phi^- - \Gamma^+$
  $\mathbf{R}^- \leftarrow (\widehat{\mathbf{W}}^-\Phi^- - \mathbf{R}^+\Gamma^+)(\Gamma^-)^{-1}$
end for

$$\widehat{\mathbf{W}}, \hat{\mathbf{Z}} \leftarrow G(\mathbf{R}, \mathbf{Q}, \Gamma, \Upsilon) \text{ solves } \quad \mathbf{X}^\top\mathbf{X}\widehat{\mathbf{W}}\Upsilon + \widehat{\mathbf{W}}\Gamma = \mathbf{X}^\top\mathbf{Q}\Upsilon + \mathbf{R}\Gamma, \quad \text{and} \quad \hat{\mathbf{Z}} = \mathbf{X}\widehat{\mathbf{W}}$$

High-Dimensional Analysis

If a sequence of problems indexed by N is solved by Algorithm 1, with $\lim_{N\to\infty}\frac{N_{in}}{N} = \beta \in (0, \infty)$

- **Data**: $X(N) = USV^T$ with U, V are square Haar distributed and $S_n(N)$ are bounded
- **True weights**: $W^*(N) \in \mathbb{R}^{N_{in}\times N_{hid}}$ such that rows are i.i.d. samples of $\mathcal{W} \in \mathbb{R}^{1\times N_{hid}}$
- **Prox**: Act row-wise, have symmetric Jacobians which are uniformly Lipschitz

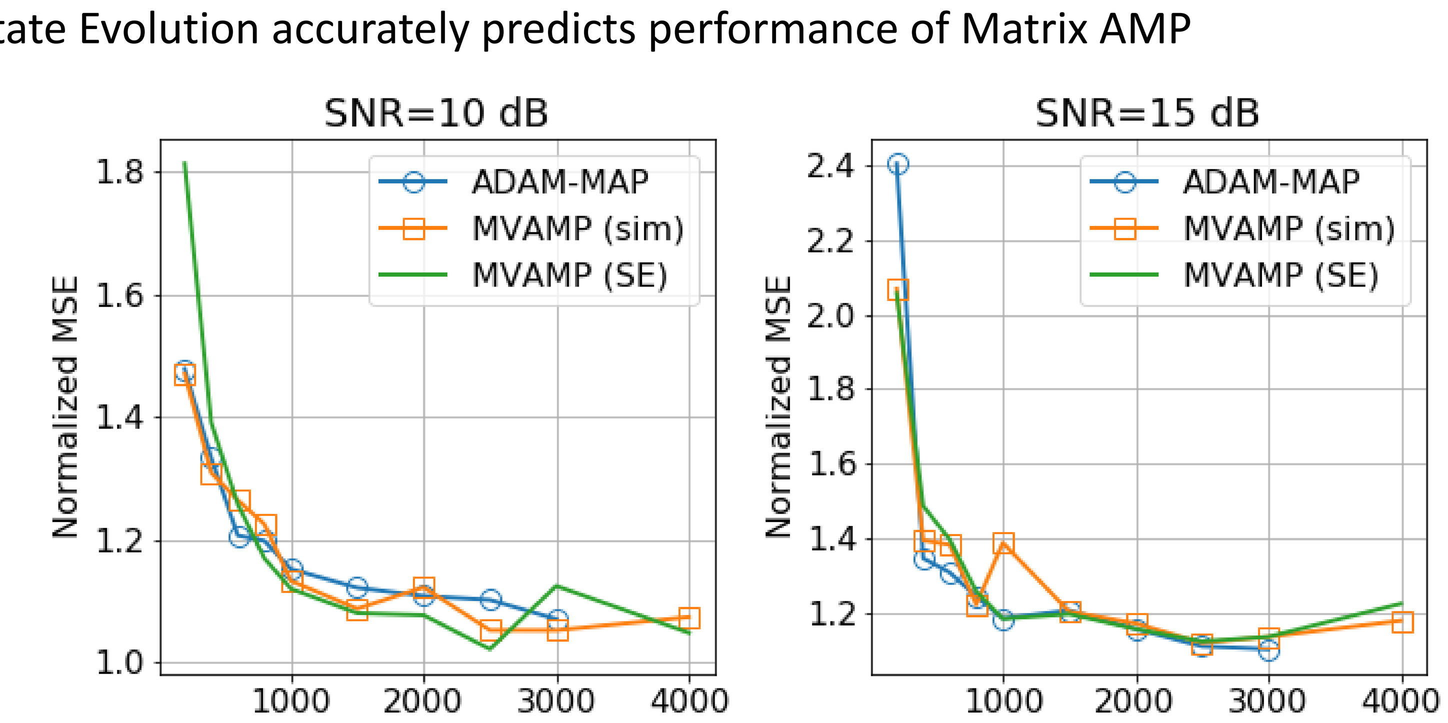
Then there exists deterministic $\bar{\Gamma}_t$ and $\bar{\Sigma}_t$ obtained recursively, $\mathcal{W} \perp \mathcal{Z} \sim \mathcal{N}(0, I_{N_{hid}})$ such that for all iterations t

$$(\widehat{W}_{t,n:}^+, W_{n,:}^*) \xrightarrow{d} (\text{prox}_\phi(\mathcal{W} + \mathcal{Z}\bar{\Sigma}_t, \bar{\Gamma}_t), \mathcal{W})$$

- Joint distribution of estimate and true vectors converge
- Recursive formula for $(\widehat{W}_{t,n:}^+, W_{n,:}^*)$
- Provides exact prediction of parameter and generation error

Predicting Performance for a Synthetic Network

- Two-layer ground truth network
  - $N_{in} = 100, d = 4, N_{out} = 1$
  - Input: iid Gaussian with variance $1/N_{in}$
  - Noise added to output to get different SNR levels
- Key takeaways:
  - ADAM optimizer achieves similar results to Matrix AMP
  - State Evolution accurately predicts performance of Matrix AMP



References:
[1] Pandit, P., Sahraee, M., Rangan, S., & Fletcher, A. K. (2019). Asymptotics of map inference in deep networks. IEEE International Symposium on Information Theory (ISIT), 842-846
[2] Rangan, S., Schniter, P., & Fletcher, A. K. (2019). Vector approximate message passing. IEEE Transactions on Information Theory.
[3] Bayati, M., & Montanari, A. (2011). The dynamics of message passing on dense graphs, with applications to compressed sensing. IEEE Transactions on Information Theory, 57(2), 764-785.