A Unified Approximate Bayesian Inference Framework for Generalized Linear Models

Xiangming Meng
Huawei Technologies Co. Ltd., Shanghai, China
mengxm11@gmail.com
Collaborators: Sheng Wu (Beijing University of Post and Telecommunications) and Jiang Zhu (Zhejiang University)

Physics, Inference, and Learning (PIL) 2018
Nov. 3 2018, Beijing

Outline

- **Problem Statement**
- **AMP: Review and EP perspective**
- **A Unified Approximate Inference Framework**
  - Generalized AMP, SBL and VAMP
  - Bilinear Adaptive Generalized VAMP
- **Conclusion**

Problem Statement

☐ Generalized Linear Models (GLM)

Signal of interest — linear mixing — probabilistic mapping — observations

$$\mathbf{x} \sim p_0(\mathbf{x}) \xrightarrow{\mathbf{x} \in \mathbb{R}^N} \boxed{\mathbf{A} \in \mathbb{R}^{M \times N}} \xrightarrow{\mathbf{z} \in \mathbb{R}^M} \boxed{p(y_a \mid z_a)} \longrightarrow \mathbf{y} \in \mathbb{R}^M$$

- **Goal**

  To infer the input x (and/or z) given the output y and A, assuming the distributions of x and *p*(y|z) are known

Problem Statement

☐ Generalized Linear Models (GLM)

Signal of interest | linear mixing | probabilistic mapping | observations

$$\mathbf{x} \sim p_0(\mathbf{x}) \xrightarrow{\mathbf{x} \in \mathbb{R}^N} \boxed{\mathbf{A} \in \mathbb{R}^{M \times N}} \xrightarrow{\mathbf{z} \in \mathbb{R}^M} \boxed{p(y_a \mid z_a)} \longrightarrow \mathbf{y} \in \mathbb{R}^M$$

- **Goal**

  To infer the input x (and/or z) given the output y and A, assuming the distributions of x and *p*(y|z) are known

- **Applications**

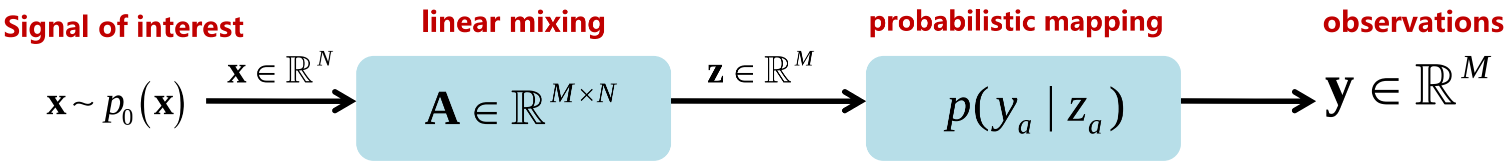
  ✓ Compressed sensing (CS), quantized or 1-bit CS

  ✓ Wireless signal detection: code division multiple access (CDMA), multiple input multiple output (MIMO) in 5G communications, channel estimation, etc.

  ✓ linear regression or classification and a variety of linear inverse problems

Problem Statement

□ Generalized Linear Models (GLM)



- **Goal**

  To infer the input x (and/or z) given the output y and A, assuming the distributions of x and *p*(y|z) are known

- **Applications**

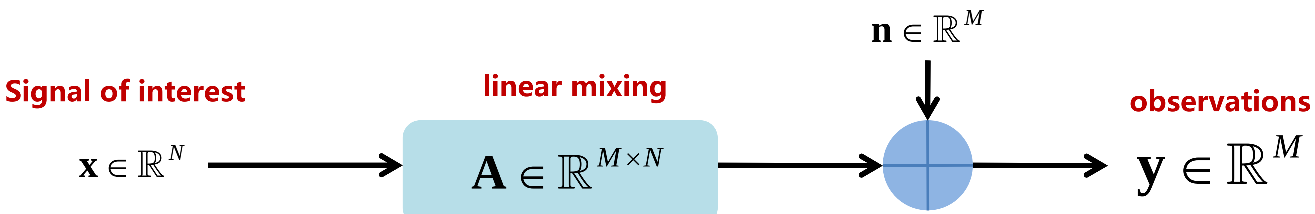
  ✓ Compressed sensing (CS), quantized or 1-bit CS

  ✓ Wireless signal detection: code division multiple access (CDMA), multiple input multiple output (MIMO) in 5G communications, channel estimation, etc.

  ✓ linear regression or classification and a variety of linear inverse problems

- **Special case: standard linear models**

  In particular, if p(y|z) is Gaussian, GLM reduces to the common standard linear models (SLM)



$\mathbf{x} \sim p_0(\mathbf{x})$

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n}$$

$\mathbf{n} \sim \mathcal{N}(0, \sigma^2 I)$

Problem Statement

☐ Generalized Linear Models (GLM)

Signal of interest · linear mixing · probabilistic mapping · observations

$$\mathbf{x}\sim p_0(\mathbf{x}) \xrightarrow{\mathbf{x}\in\mathbb{R}^N} \boxed{\mathbf{A}\in\mathbb{R}^{M\times N}} \xrightarrow{\mathbf{z}\in\mathbb{R}^M} \boxed{p(y_a\,|\,z_a)} \longrightarrow \mathbf{y}\in\mathbb{R}^M$$

- **Optimal Bayesian estimation**

According to the Bayes' rule, the posterior distribution can be computed as

$$p(\mathbf{x}\,|\,\mathbf{y}) = \frac{p_0(\mathbf{x})\,p(\mathbf{y}\,|\,\mathbf{x})}{p(\mathbf{y})} \xrightarrow{\text{marginalize}} p(x_i\,|\,\mathbf{y}) = \int_{\sim x_i} p(\mathbf{x}\,|\,\mathbf{y})\,d\mathbf{x}_{\backslash i}$$

Posterior mean $$\hat{x}_i^{MMSE} = \int x_i\, p(x_i\,|\,\mathbf{y})\,dx_i$$

Posterior variance $$v_i^{MMSE} = \int x_i^2\, p(x_i\,|\,\mathbf{y})\,dx_i - \left(\hat{x}_i^{MMSE}\right)^2$$

MMSE estimates

Problem Statement

☐ Generalized Linear Models (GLM)

Signal of interest | linear mixing | probabilistic mapping | observations

$$\mathbf{x} \sim p_0(\mathbf{x}) \xrightarrow{\mathbf{x}\in\mathbb{R}^N} \boxed{\mathbf{A}\in\mathbb{R}^{M\times N}} \xrightarrow{\mathbf{z}\in\mathbb{R}^M} \boxed{p(y_a \mid z_a)} \longrightarrow \mathbf{y}\in\mathbb{R}^M$$

- **Optimal Bayesian estimation**

According to the Bayes' rule, the posterior distribution can be computed as

$$p(\mathbf{x}\mid\mathbf{y}) = \frac{p_0(\mathbf{x})\,p(\mathbf{y}\mid\mathbf{x})}{p(\mathbf{y})} \xrightarrow{\text{marginalize}} p(x_i\mid\mathbf{y}) = \int_{\sim x_i} p(\mathbf{x}\mid\mathbf{y})\,d\mathbf{x}_{\setminus i}$$

Posterior mean $\hat{x}_i^{MMSE} = \int x_i\, p(x_i\mid\mathbf{y})\,dx_i$

Posterior variance $v_i^{MMSE} = \int x_i^2\, p(x_i\mid\mathbf{y})\,dx_i - \left(\hat{x}_i^{MMSE}\right)^2$

MMSE estimates

Curse of Dimensionality: The optimal Bayesian inference becomes intractable in high dimensional case due to integration (or summation) operation

We have to resort to approximate inference methods

AMP: review and EP perspective

☐ Approximate message passing (AMP)

AMP iteratively decouples the original vector inference problem to scalar inference problems

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n} \longrightarrow \begin{cases} R_1 = x_1 + \tilde{n}_1 \\ \vdots \\ R_N = x_N + \tilde{n}_N \end{cases}$$

- **Evolution of AMP**
  - ✓ Proposed in the field compressed sensing (CS) [DMM09]
  - ✓ Early work in communications [Kabashima 03] [Tanaka 02]
  - ✓ Deeply related to TAP equations and replica methods in statistical physics [KMSSZ12]
  - ✓ Extended with EM learning [KMSSZ12][VS11]
  - ✓ Extended to Generalized AMP (GAMP) [Rangan12] for GLM models
  - ✓ Extended to vector AMP (VAMP) [RSF16] orthogonal AMP (OAMP) [ML17]
  - ✓ Many other extensions….

AMP Algorithm

Initialization

For *t* = 1,…, *T*

$$V_a^t = \sum_i A_{ai}^2 \nu_i^t$$

$$Z_a^t = \sum_i A_{ai}\hat{x}_i^t - \frac{\left(y_a - Z_a^{t-1}\right)}{\sigma^2 + V_a^{t-1}} V_a^t$$

Onsager term

$$\Sigma_i^t = 1 / \sum_a \frac{A_{ai}^2}{\sigma^2 + V_a^t}$$

$$R_i^t = \hat{x}_i^t + \Sigma_i^t \sum_a \frac{A_{ai}\left(y_a - Z_a^t\right)}{\sigma^2 + V_a^t}$$

$$\hat{x}_i^{t+1} = E\left(x_i \mid R_i^t, \Sigma_i^t\right), \hat{\nu}_i^{t+1} = Var\left(x_i \mid R_i^t, \Sigma_i^t\right)$$

end

AMP: review and EP perspective

☐ Approximate message passing (AMP)

AMP iteratively decouples the original vector inference problem to scalar inference problems

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n} \longrightarrow \begin{cases} R_1 = x_1 + \tilde{n}_1 \\ \vdots \\ R_N = x_N + \tilde{n}_N \end{cases}$$

Initialization AMP Algorithm

For *t = 1,…, T*

$$V_a^t = \sum_i A_{ai}^2 \nu_i^t$$

$$Z_a^t = \sum_i A_{ai}\hat{x}_i^t - \frac{\left(y_a - Z_a^{t-1}\right)}{\sigma^2 + V_a^{t-1}} V_a^t$$

$$\Sigma_i^t = 1 / \sum_a \frac{A_{ai}^2}{\sigma^2 + V_a^t}$$

Onsager term

$$R_i^t = \hat{x}_i^t + \Sigma_i^t \sum_a \frac{A_{ai}\left(y_a - Z_a^t\right)}{\sigma^2 + V_a^t}$$

$$\hat{x}_i^{t+1} = E\left(x_i \mid R_i^t, \Sigma_i^t\right), \hat{\nu}_i^{t+1} = Var\left(x_i \mid R_i^t, \Sigma_i^t\right)$$

end

- **Evolution of AMP**
  - ✓ Proposed in the field compressed sensing (CS) [DMM09]
  - ✓ Early work in communications [Kabashima 03] [Tanaka 02]
  - ✓ Deeply related to TAP equations and replica methods in statistical physics [KMSSZ12]
  - ✓ Extended with EM learning [KMSSZ12][VS11]
  - ✓ Extended to Generalized AMP (GAMP) [Rangan12] for GLM models
  - ✓ Extended to vector AMP (VAMP) [RSF16] orthogonal AMP (OAMP) [ML17]
  - ✓ Many other extensions….

- **Properties of AMP**
  - ✓ For i.i.d. Gaussian matrix A, asymptotically optimal and rigorously analyzed via state evolution (SE) [BM11]
  - ✓ For general matrices A, AMP may diverge [BM11]
  - ✓ VAMP converges for right-rotationally invariant matrices [RSF16]

AMP: review and EP perspective

☐ Approximate message passing (AMP)

AMP iteratively decouples the original vector inference problem to scalar inference problems

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{n} \longrightarrow \begin{cases} R_1 = x_1 + \tilde{n}_1 \\ \vdots \\ R_N = x_N + \tilde{n}_N \end{cases}$$

AMP Algorithm

Initialization

For *t = 1,…, T*

$$V_a^t = \sum_i A_{ai}^2 \nu_i^t$$

$$Z_a^t = \sum_i A_{ai} \hat{x}_i^t - \frac{\left(y_a - Z_a^{t-1}\right)}{\sigma^2 + V_a^{t-1}} V_a^t$$

Onsager term

$$\Sigma_i^t = 1 / \sum_a \frac{A_{ai}^2}{\sigma^2 + V_a^t}$$

$$R_i^t = \hat{x}_i^t + \Sigma_i^t \sum_a \frac{A_{ai}\left(y_a - Z_a^t\right)}{\sigma^2 + V_a^t}$$

$$\hat{x}_i^{t+1} = E\left(x_i \mid R_i^t, \Sigma_i^t\right), \hat{\nu}_i^{t+1} = Var\left(x_i \mid R_i^t, \Sigma_i^t\right)$$

end

- **Evolution of AMP**
  - ✓ Proposed in the field compressed sensing (CS) [DMM09]
  - ✓ Early work in communications [Kabashima 03] [Tanaka 02]
  - ✓ Deeply related to TAP equations and replica methods in statistical physics [KMSSZ12]
  - ✓ Extended with EM learning [KMSSZ12][VS11]
  - ✓ Extended to Generalized AMP (GAMP) [Rangan12] for GLM models
  - ✓ Extended to vector AMP (VAMP) [RSF16] orthogonal AMP (OAMP) [ML17]
  - ✓ Many other extensions….

- **Properties of AMP**
  - ✓ For i.i.d. Gaussian matrix A, asymptotically optimal and rigorously analyzed via state evolution (SE) [BM11]
  - ✓ For general matrices A, AMP may diverge [BM11]
  - ✓ VAMP converges for right-rotationally invariant matrices [RSF16]

- **Derivation of AMP**
  - ✓ Originally derived from belief propagation (BP) via central limit theorem and Taylor series expansion [DMM09] [DMM10]
  - ✓ Alternatively derived from expectation propagation (EP) via neglecting high order terms [MWKL15a]

AMP: review and EP perspective

◻ An EP Perspective on AMP

- **Expectation Propagation (EP)** [Minka01] [MO05]

$$p(\mathbf{x}) = \prod_a f_a(\mathbf{x}) \xrightarrow{\text{approximated as}} q(\mathbf{x}) = \prod_a \tilde{f}_a(\mathbf{x})$$

Optimization objective: $\min \; KL(p(\mathbf{x}) \,||\, q(\mathbf{x}))$ $\quad q(\boldsymbol{x}) = h(\boldsymbol{x}) \exp\left\{ \boldsymbol{\theta}^T \phi(\boldsymbol{x}) + g(\boldsymbol{\theta}) \right\}$

AMP: review and EP perspective

An EP Perspective on AMP

- **Expectation Propagation (EP)** [Minka01] [MO05]

$$p(\mathbf{x}) = \prod_a f_a(\mathbf{x}) \xrightarrow{\text{approximated as}} q(\mathbf{x}) = \prod_a \tilde{f}_a(\mathbf{x})$$

Optimization objective: $\min \; KL(p(\mathbf{x}) \,||\, q(\mathbf{x}))$ $\quad q(\boldsymbol{x}) = h(\boldsymbol{x}) \exp\left\{\boldsymbol{\theta}^T \phi(\boldsymbol{x}) + g(\boldsymbol{\theta})\right\}$

Iterative local optimization

Iteratively Refine each factor

$$\tilde{f}_a(\mathbf{x}) = \arg\min_{t(\mathbf{x})\in\Phi} \; KL\Big(f_a(\mathbf{x})\prod_{b\neq a}\tilde{f}_b(\mathbf{x}) \,||\, t(\mathbf{x})\prod_{b\neq a}\tilde{f}_b(\mathbf{x})\Big)$$

AMP: review and EP perspective

An EP Perspective on AMP

- **Expectation Propagation (EP)** [Minka01] [MO05]

$$p(\mathbf{x}) = \prod_a f_a(\mathbf{x}) \xrightarrow{\text{approximated as}} q(\mathbf{x}) = \prod_a \tilde{f}_a(\mathbf{x})$$

Optimization objective: $\min \; KL(p(\mathbf{x}) \,||\, q(\mathbf{x}))$ $\quad q(\boldsymbol{x}) = h(\boldsymbol{x}) \exp\left\{\boldsymbol{\theta}^T \phi(\boldsymbol{x}) + g(\boldsymbol{\theta})\right\}$

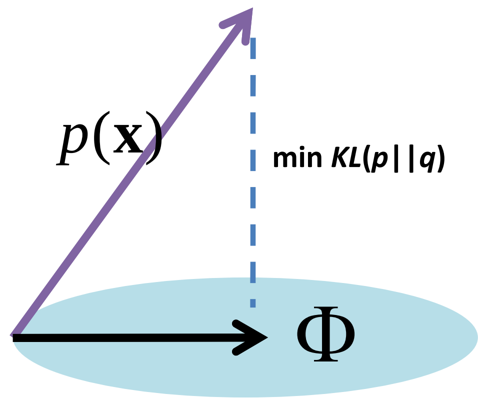
Iterative local optimization

Iteratively Refine each factor

$$\tilde{f}_a(\mathbf{x}) = \arg\min_{t(\mathbf{x}) \in \Phi} \; KL\Big(f_a(\mathbf{x}) \prod_{b \neq a} \tilde{f}_b(\mathbf{x}) \,||\, t(\mathbf{x}) \prod_{b \neq a} \tilde{f}_b(\mathbf{x})\Big)$$

Properties of EP:

✓ **Applicable to both discrete and continuous distributions**
✓ **Equivalent to moment matching**
✓ **Deeply related to the adaptive Cavity Method in statistical physics**



$$q(\mathbf{x}) = \text{Proj}_{\Phi}\big(q(\mathbf{x})\big)$$

AMP: review and EP perspective

An EP Perspective on AMP

Target distribution $p(\mathbf{x}|\mathbf{y}) \propto \prod_{i=1}^{N} p_0(x_i) \prod_{a=1}^{M} \mathcal{N}(y_a; (\mathbf{A}\mathbf{x})_a, \sigma^2)$

Approximate distribution $q(\mathbf{x}) \propto \prod_{i=1}^{N} q_0(x_i) \prod_{a=1}^{M} \prod_{i=1}^{N} q_{ai}(x_i)$ **fully factorized form**

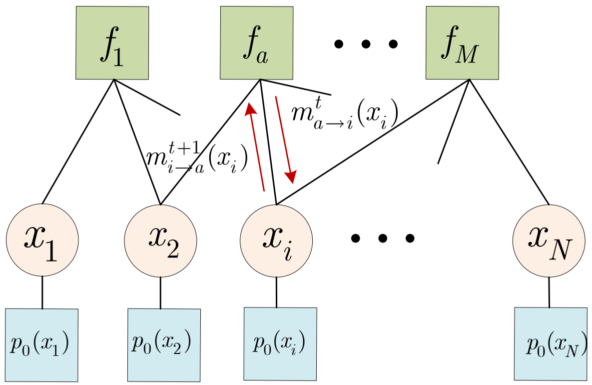
AMP: review and EP perspective

☐ An EP Perspective on AMP

Target distribution $p(\mathbf{x}|\mathbf{y}) \propto \prod_{i=1}^{N} p_0(x_i) \prod_{a=1}^{M} \mathcal{N}\left(y_a; (\mathbf{A}\mathbf{x})_a, \sigma^2\right)$

Approximate distribution $q(\mathbf{x}) \propto \prod_{i=1}^{N} q_0(x_i) \prod_{a=1}^{M} \prod_{i=1}^{N} q_{ai}(x_i)$ **fully factorized form**

- **Then, we can iteratively refine all the approximating factors using EP principle**
- **Intuitively, this optimization process can be realized in an message passing manner as BP**



message passing

Expectation Propagation (EP)

$$m^{t}_{a\to i}(x_i) \propto \frac{\mathrm{Proj}_{\Phi}\left[m^{t}_{i\to a}(x_i) \int \prod_{j\neq i} m^{t}_{j\to a}(x_j) p(y_a|\mathbf{x})\right]}{m^{t}_{i\to a}(x_i)}$$

$$m^{t+1}_{i\to a}(x_i) \propto \frac{\mathrm{Proj}_{\Phi}\left[p_0(x_i) \prod_{b} m^{t}_{b\to i}(x_i)\right]}{m^{t}_{a\to i}(x_i)}$$

- **Choosing the projection set as Gaussian and neglecting high-order terms results in the AMP** [MWKL15a]

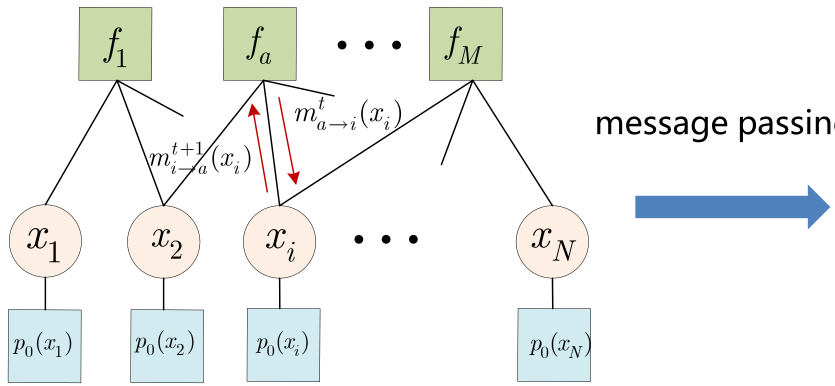
AMP: review and EP perspective

☐ An EP Perspective on AMP

Target distribution $p(\mathbf{x}|\mathbf{y}) \propto \prod_{i=1}^{N} p_0(x_i) \prod_{a=1}^{M} \mathcal{N}\left(y_a; (\mathbf{A}\mathbf{x})_a, \sigma^2\right)$

Approximate distribution $q(\mathbf{x}) \propto \prod_{i=1}^{N} q_0(x_i) \prod_{a=1}^{M} \prod_{i=1}^{N} q_{ai}(x_i)$ **fully factorized form**

- **Then, we can iteratively refine all the approximating factors using EP principle**
- **Intuitively, this optimization process can be realized in an message passing manner as BP**

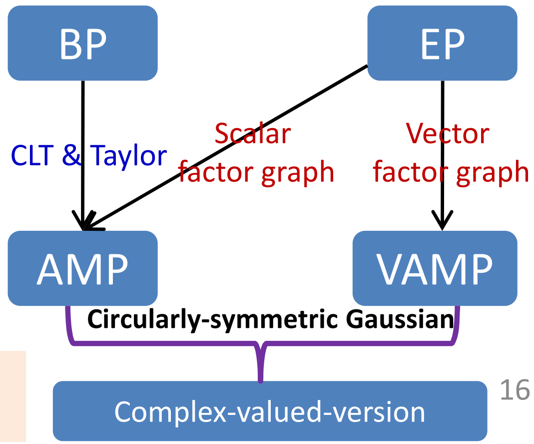


Expectation Propagation (EP)

$$m_{a\to i}^{t}(x_i) \propto \frac{\mathrm{Proj}_{\Phi}\left[m_{i\to a}^{t}(x_i)\int \prod_{j\neq i} m_{j\to a}^{t}(x_j)\, p(y_a|\mathbf{x})\right]}{m_{i\to a}^{t}(x_i)}$$

$$m_{i\to a}^{t+1}(x_i) \propto \frac{\mathrm{Proj}_{\Phi}\left[p_0(x_i)\prod_{b} m_{b\to i}^{t}(x_i)\right]}{m_{a\to i}^{t}(x_i)}$$

- **Choosing the projection set as Gaussian and neglecting high-order terms results in the AMP** [MWKL15a]
- **The EP perspective of AMP:**
  - ✓ Explicitly establishes the relationship between AMP and EP[MWKL15a, MWKL15b, WKNLHDQ14]
  - ✓ Facilitates the extension of AMP to the complex-valued AMP (simply using circularly-symmetric Gaussian) [MWKL15b]
  - ✓ Provides a unified view of AMP (derived from scalar EP [MWKL15a]) and VAMP(derived from vector EP [RSF16])



[MWKL15a] X. Meng, S. Wu, L. Kuang, and J. Lu, "An expectation propagation perspective on approximate message passing," IEEE Signal Process. Lett., vol. 22, no. 8, pp. 1194-1197, Aug. 2015.

A Unified Inference Framework for GLM

☐ Motivations

- **GLM is more general: the measurements are often obtained in a nonlinear way**
  - ✓ quantized measurements, e.g., 1-bit CS, low resolution ADC in wireless communication
  - ✓ incomplete measurements
  - ✓ non-Gaussian and/or non-additive noise
  - ✓ discrete measurements, and so on....

A Unified Inference Framework for GLM

☐ Motivations

- **GLM is more general: the measurements are often obtained in a nonlinear way**
  - ✓ quantized measurements, e.g., 1-bit CS, low resolution ADC in wireless communication
  - ✓ incomplete measurements
  - ✓ non-Gaussian and/or non-additive noise
  - ✓ discrete measurements, and so on….

- **SLM has already been extensively studied**
  - ✓ simple to design and analyze
  - ✓ a variety of Bayesian methods, e.g., AMP and sparse Bayesian learning (SBL) have been proposed
  - ✓ extension of one existing SLM method to GLM needs careful design, which is often *difficult to follow* and *task-specific*, such as the conventional extensions from AMP to GAMP, VAMP to GVAMP

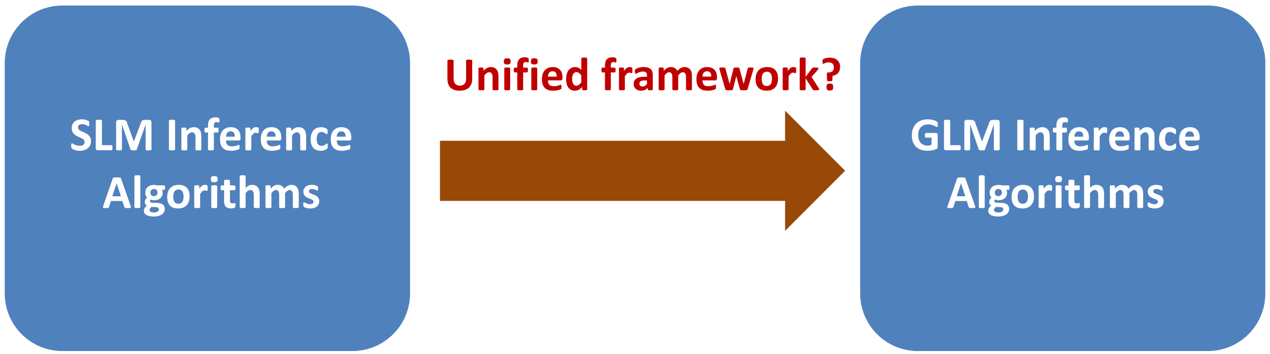
A Unified Inference Framework for GLM

☐ Motivations

- **GLM is more general: the measurements are often obtained in a nonlinear way**
  - ✓ quantized measurements, e.g., 1-bit CS, low resolution ADC in wireless communication
  - ✓ incomplete measurements
  - ✓ non-Gaussian and/or non-additive noise
  - ✓ discrete measurements, and so on….

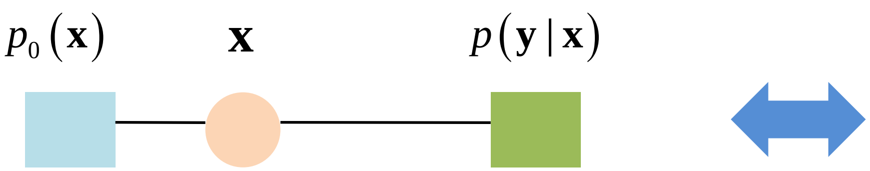
- **SLM has already been extensively studied**
  - ✓ simple to design and analyze
  - ✓ a variety of Bayesian methods, e.g., AMP and sparse Bayesian learning (SBL) have been proposed
  - ✓ extension of one existing SLM method to GLM needs careful design, which is often *difficult to follow* and *task-specific*, such as the conventional extensions from AMP to GAMP, VAMP to GVAMP

- **Natural question: Is there a unified framework for GLM, under which SLM inference methods can be easily extended to GLM ones following a common rule?**

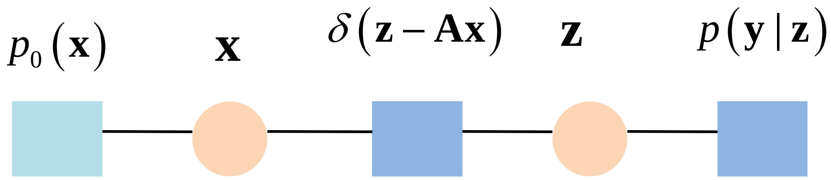


A Unified Inference Framework for GLM

- **Two Equivalent Factor Graphs for GLM**



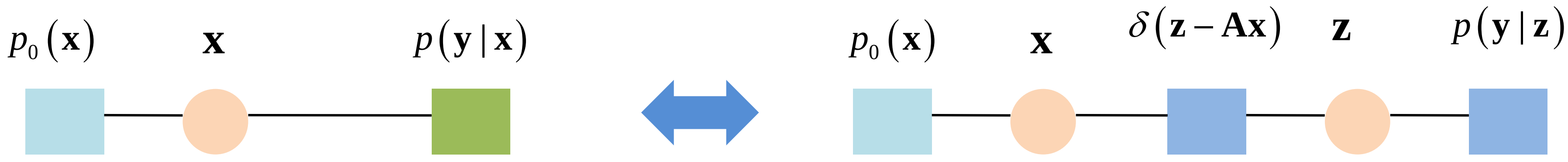
(a) Factor graph v1 (vector form)



(b) Factor graph v2 (vector form)

A Unified Inference Framework for GLM

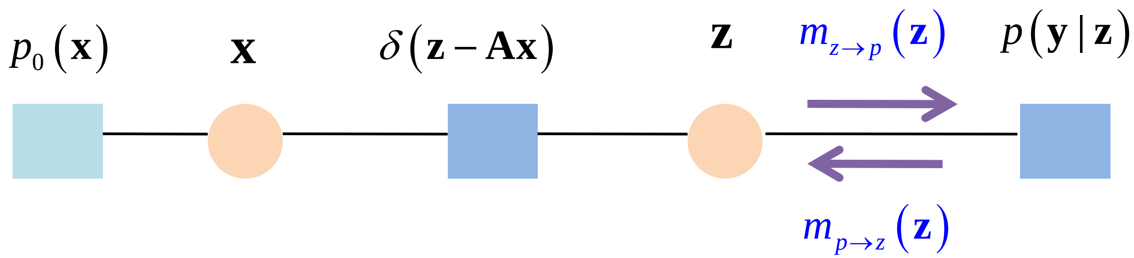
□ Two Equivalent Factor Graphs for GLM



(a) Factor graph v1 (vector form)

(b) Factor graph v2 (vector form)

□ Decoupling GLM into SLM via EP



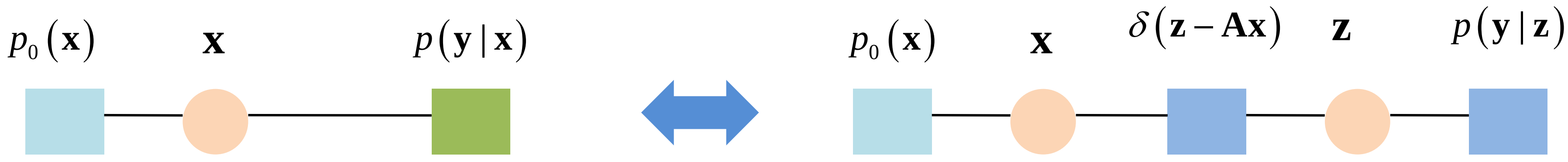
$$m_{z\to p}^{t-1}(\mathbf{z}) \propto \mathcal{N}\left(\mathbf{z}; z_A^{ext}(t-1), v_A^{ext}(t-1)I\right)$$

EP message passing
(t-th iteration)

$$m_{p\to z}^{t}(\mathbf{z}) \propto \frac{\mathrm{Proj}_{\Phi}\left(p(\mathbf{y}|\mathbf{z})\, m_{z\to p}^{t-1}(\mathbf{z})\right)}{m_{z\to p}^{t-1}(\mathbf{z})} \propto \mathcal{N}\left(\mathbf{z}; z_B^{ext}(t), v_B^{ext}(t)I\right)$$

A Unified Inference Framework for GLM

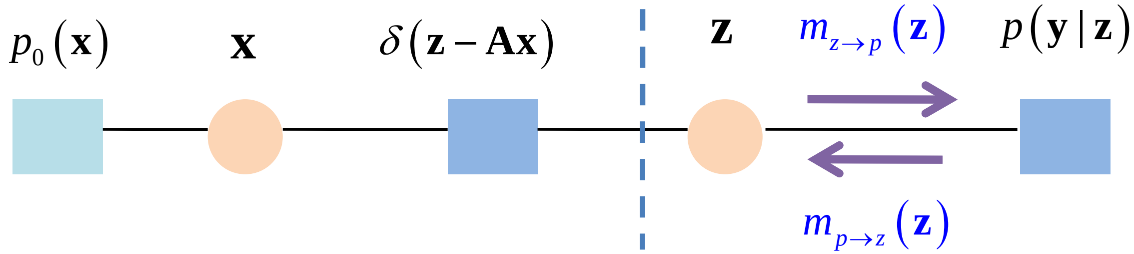
☐ Two Equivalent Factor Graphs for GLM



(a) Factor graph v1 (vector form)

(b) Factor graph v2 (vector form)

☐ Decoupling GLM into SLM via EP



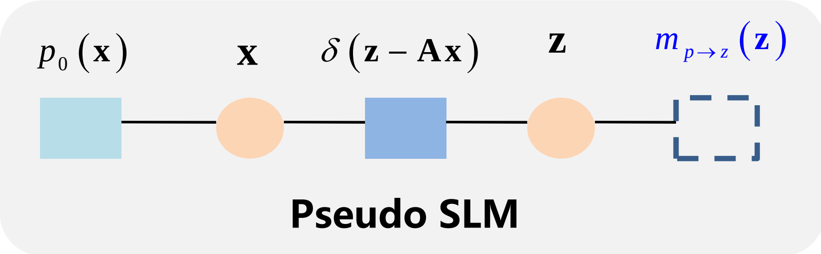
$$m^{t-1}_{z\to p}(\mathbf{z}) \propto \mathcal{N}\left(\mathbf{z}; z_A^{ext}(t-1), v_A^{ext}(t-1)I\right)$$

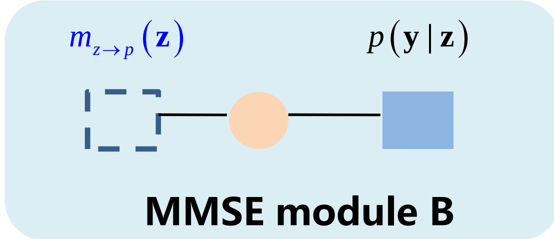
EP message passing
(t-th iteration)

$$m^{t}_{p\to z}(\mathbf{z}) \propto \frac{\mathrm{Proj}_{\Phi}\left(p(\mathbf{y}|\mathbf{z})\, m^{t-1}_{z\to p}(\mathbf{z})\right)}{m^{t-1}_{z\to p}(\mathbf{z})} \propto \mathcal{N}\left(\mathbf{z}; z_B^{ext}(t), v_B^{ext}(t)I\right)$$

A Unified Inference Framework for GLM

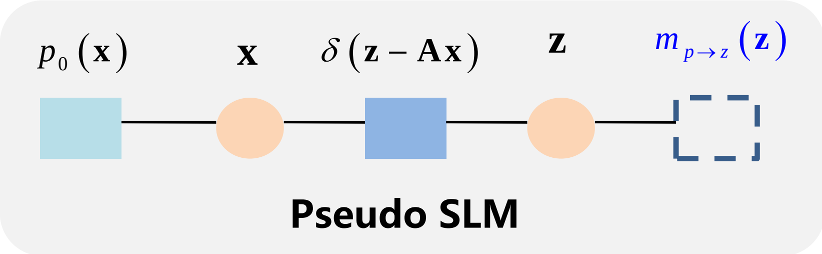
- **Decoupling GLM into SLM via EP**

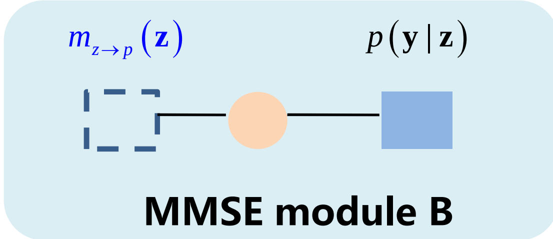




A Unified Inference Framework for GLM

☐ Decoupling GLM into SLM via EP





- **The original GLM is iteratively decoupled into a sequence of SLM problems**

$$\tilde{\mathbf{y}} = \mathbf{A}\mathbf{x} + \tilde{\mathbf{n}}$$

$$\tilde{\mathbf{n}} \sim \mathcal{N}(0, \tilde{\sigma}^2 I)$$

$$\tilde{\mathbf{y}} = \mathbf{z}_B^{ext}(t), \tilde{\sigma}^2 = v_B^{ext}(t)$$

Module A

$\mathbf{z}_A^{ext}(t-1), v_A^{ext}(t-1)$

Component-wise MMSE

$$\mathbf{z}_B^{post} = \mathrm{E}\left[z \mid \mathbf{z}_A^{ext}, v_A^{ext}\right]$$

$$v_B^{post} = \mathrm{Var}\left[z \mid \mathbf{z}_A^{ext}, v_A^{ext}\right]$$

Module B

$\mathbf{z}_B^{ext}(t), v_B^{ext}(t)$

$$\frac{1}{v_A^{ext}(t)} = \frac{1}{v_A^{post}(t)} - \frac{1}{v_B^{ext}(t)}$$

$$\frac{\mathbf{z}_A^{ext}(t)}{v_A^{ext}(t)} = \frac{\mathbf{z}_A^{post}(t)}{v_A^{post}(t)} - \frac{\mathbf{z}_B^{ext}(t)}{v_B^{ext}(t)}$$

$$\frac{1}{v_B^{ext}(t)} = \frac{1}{v_B^{post}(t)} - \frac{1}{v_A^{ext}(t-1)}$$

$$\frac{\mathbf{z}_B^{ext}(t)}{v_B^{ext}(t)} = \frac{\mathbf{z}_B^{post}(t)}{v_B^{post}(t)} - \frac{\mathbf{z}_A^{ext}(t-1)}{v_A^{ext}(t-1)}$$

Unified Inference Framework for GLM
- Initialization $\mathbf{z}_A^{ext}(0), v_A^{ext}(0)$
- For t = 1: T, Do
  1. Perform component-wise MMSE
  2. Update $\mathbf{z}_B^{ext}(t), v_B^{ext}(t)$
  3. Perform SLM inference one or more iterations
  4. Compute $\mathbf{z}_A^{post}(t), v_A^{post}(t)$ and then update $\mathbf{z}_A^{ext}(t), v_A^{ext}(t)$

Note: The computation of posterior mean and variance of z in module A may differ for different SLM inference methods.

[MWZ18] X. Meng, S. Wu and J. Zhu, "A unified Bayesian inference framework for generalized linear model," IEEE Signal Process. Lett., vol. 25, no. 3, Mar. 2018.

A Unified Inference Framework for GLM

☐ From AMP to Gr-AMP [MWZ18]

AMP
(T0 iterations)
Module A

$\mathbf{z}_A^{ext}(t-1), v_A^{ext}(t-1)$ →

← $\mathbf{z}_B^{ext}(t), v_B^{ext}(t)$

Component-wise
MMSE
Module B

The Gr-AMP Algorithm

- Initialization $\mathbf{z}_A^{ext}(0), v_A^{ext}(0)$
- For t = 1: T, Do
  1. Perform component-wise MMSE
  2. Update $\mathbf{z}_B^{ext}(t), v_B^{ext}(t)$
  3. Perform AMP for T0 iterations
  4. Compute $\mathbf{z}_A^{post}(t), v_A^{post}(t)$ and then update $\mathbf{z}_A^{ext}(t), v_A^{ext}(t)$

- **Gr-AMP is a double-loop iterative algorithm**
  - ✓ In the outer-loop, module A and B exchanges extrinsic messages
  - ✓ It is proved for AMP, the output message of module A has already computed within AMP, i.e., $\mathbf{z}_A^{ext}(t) = Z_a(t), v_A^{ext}(t) = V_a(t)$
  - ✓ In each inner-loop, module A performs AMP for T0 iterations, rather than being fixed to 1 as GAMP.

A Unified Inference Framework for GLM

☐ **From AMP to Gr-AMP** [MWZ18]

AMP
(T0 iterations)
Module A

$\mathbf{z}_A^{ext}(t-1), v_A^{ext}(t-1)$ →

← $\mathbf{z}_B^{ext}(t), v_B^{ext}(t)$

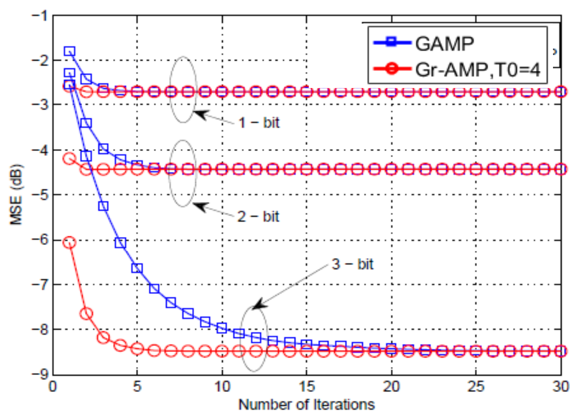
Component-wise
MMSE
Module B

The Gr-AMP Algorithm

- Initialization $\mathbf{z}_A^{ext}(0), v_A^{ext}(0)$
- For t = 1: T, Do
  1. Perform component-wise MMSE
  2. Update $\mathbf{z}_B^{ext}(t), v_B^{ext}(t)$
  3. Perform AMP for T0 iterations
  4. Compute $\mathbf{z}_A^{post}(t), v_A^{post}(t)$ and then update $\mathbf{z}_A^{ext}(t), v_A^{ext}(t)$

- **Gr-AMP is a double-loop iterative algorithm**
  - ✓ In the outer-loop, module A and B exchange extrinsic messages
  - ✓ It is proved for AMP, the output message of module A has already computed within AMP, i.e., $\mathbf{z}_A^{ext}(t) = Z_a(t), v_A^{ext}(t) = V_a(t)$
  - ✓ In each inner-loop, module A performs AMP for T0 iterations, rather than being fixed to 1 as GAMP.

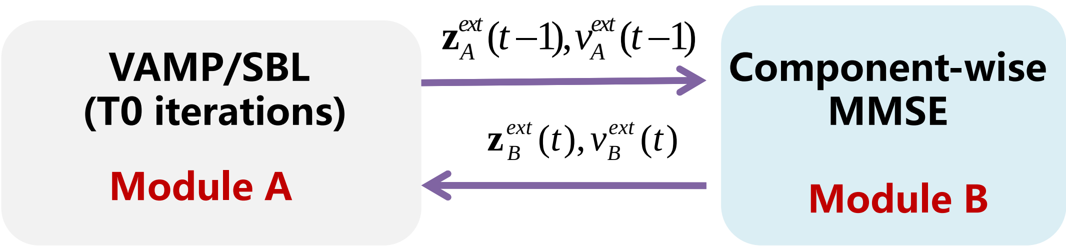
- **Relation of Gr-AMP to GAMP**
  - ✓ Gr-AMP is precisely equivalent to GAMP when T0 = 1 and thus provides an insightful perspective on GAMP: In effect, GAMP performs one iteration of AMP each time after transforming the GLM problem to a pseudo SLM problem. Note that [[QZW18]] also provides an EP derivation of GAMP but in a different way.
  - ✓ A more flexible message passing schedule: double-loop implementation



- Quantized CS for 1,2,3-bit cases: N =1024,M=512,SNR=50dB
- Gr-AMP and GAMP converge to the same performance for i.i.d. Gaussian A
- Total number iterations of AMP are about the same while the number of MMSE operations is reduced for Gr-AMP.
Still needs further study.

A Unified Inference Framework for GLM

☐ From VAMP/SBL to Gr-AMP/Gr-SBL [MWZ18]



The Gr-VAMP/Gr-SBL Algorithm

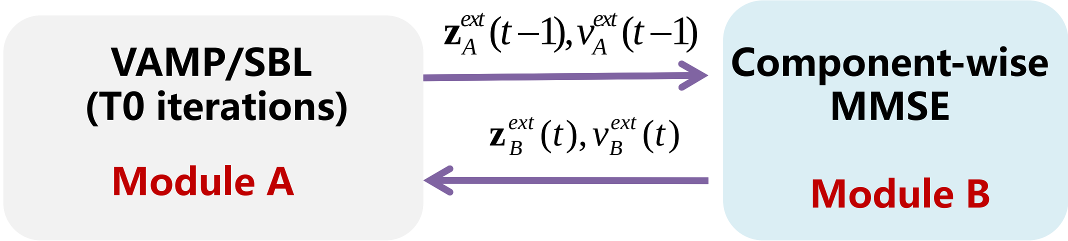
- Initialization $\mathbf{z}_A^{ext}(0), v_A^{ext}(0)$
- For t = 1: T, Do
  1. Perform component-wise MMSE
  2. Update $\mathbf{z}_B^{ext}(t), v_B^{ext}(t)$
  3. Perform VAMP/SBL for T0 iterations
  4. Compute $\mathbf{z}_A^{post}(t), v_A^{post}(t)$ and then update $\mathbf{z}_A^{ext}(t), v_A^{ext}(t)$

- **Gr-VAMP/Gr-SBL**
  - ✓ In the outer-loop, module A and B exchanges extrinsic messages
  - ✓ The posterior mean and covariance of z in module A can be computed as:

    $\mathbf{z}_A^{post} = \mathbf{A}\hat{\mathbf{x}}_A, v_A^{post} = \frac{1}{N}\mathrm{Trace}(\mathbf{A}\Sigma_A\mathbf{A}^T)$ $\hat{\mathbf{x}}_A$ Σ_A are the posterior mean and covariance of $\mathbf{x}$ computed in VAMP/SBL
  - ✓ In each inner-loop, module A performs VAMP/SBL for T0 iterations

A Unified Inference Framework for GLM

☐ From VAMP/SBL to Gr-AMP/Gr-SBL [MWZ18]

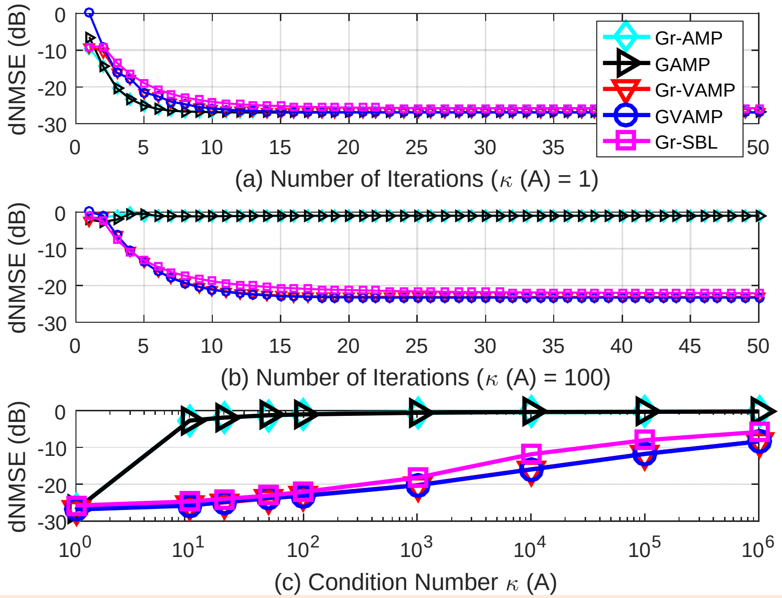


The Gr-VAMP/Gr-SBL Algorithm

- Initialization $\mathbf{z}_A^{ext}(0), v_A^{ext}(0)$
- For t = 1: T, Do
  1. Perform component-wise MMSE
  2. Update $\mathbf{z}_B^{ext}(t), v_B^{ext}(t)$
  3. Perform VAMP/SBL for T0 iterations
  4. Compute $\mathbf{z}_A^{post}(t), v_A^{post}(t)$ and then update $\mathbf{z}_A^{ext}(t), v_A^{ext}(t)$

- **Gr-VAMP/Gr-SBL**
  - ✓ In the outer-loop, module A and B exchange extrinsic messages
  - ✓ The posterior mean and covariance of z in module A can be computed as:

    $\mathbf{z}_A^{post} = \mathbf{A}\hat{\mathbf{x}}_A, v_A^{post} = \frac{1}{N}\mathrm{Trace}(\mathbf{A}\Sigma_A\mathbf{A}^T)$ $\hat{\mathbf{x}}_A$ Σ_A are the posterior mean and covariance of $\mathbf{x}$ computed in VAMP/SBL
  - ✓ In each inner-loop, module A performs VAMP/SBL for T0 iterations



Performance of de-biased NMSE for 1-bit CS

✓ N =512,M=2048,SNR=50dB, sparse ratio 0.1

✓ T0 = 1 for both Gr-VAMP and Gr-SBL

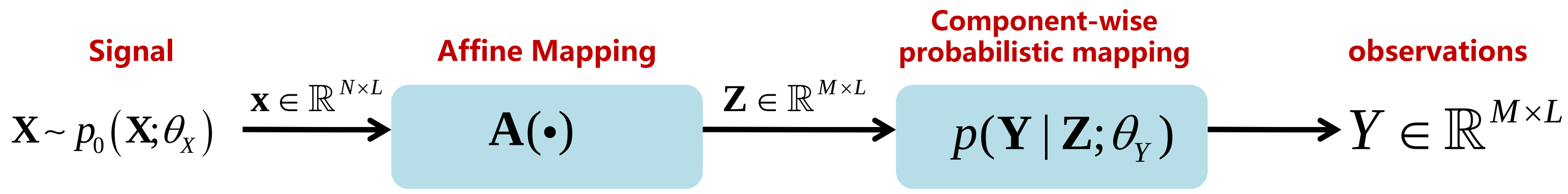
✓ When conditional number is 1, all kinds of algorithms performs nearly the same.

✓ As the condition number increases, the recovery performances degrade smoothly for Gr-VAMP/GVAMP/Gr-SBL while both Gr-AMP and GAMP diverge for even mild condition number, which show the robustness of Gr-VAMP/Gr-SBL/GVAMP for general matrices.

[MWZ18]X. Meng, S. Wu and J. Zhu, “A unified Bayesian inference framework for generalized linear model,” IEEE Signal Process. Lett., vol. 25, no. 3, Mar. 2018.

A Unified Inference Framework for GLM

◻ Bilinear GLM Problems



A(·) is a known affine linear function of unknown vector **b**, i.e., $\mathbf{A}(\mathbf{b}) = \mathbf{A}_0 + \sum_{q=1}^{Q} b_q \mathbf{A}_q$ and known $\mathbf{A}_0, \mathbf{A}_q$

- **Goal**

  To jointly infer X and b, given Y with unknown parameters θ_X, θ_Y

A Unified Inference Framework for GLM

☐ Bilinear GLM Problems

Signal | Affine Mapping | Component-wise probabilistic mapping | observations

$$\mathbf{X} \sim p_0(\mathbf{X};\theta_X) \xrightarrow{\mathbf{x}\in\mathbb{R}^{N\times L}} \boxed{\mathbf{A}(\bullet)} \xrightarrow{\mathbf{Z}\in\mathbb{R}^{M\times L}} \boxed{p(\mathbf{Y}\,|\,\mathbf{Z};\theta_Y)} \longrightarrow Y \in \mathbb{R}^{M\times L}$$

A(·) is a known affine linear function of unknown vector **b**, i.e., $\mathbf{A}(\mathbf{b}) = \mathbf{A}_0 + \sum_{q=1}^{Q} b_q \mathbf{A}_q$ and known $\mathbf{A}_0, \mathbf{A}_q$

- **Goal**

  To jointly infer X and b, given Y with unknown parameters θ_X, θ_Y

- **Applications**
  - ✓ **Quantized Compressed sensing (CS) under matrix uncertainty**
  - ✓ **Self-calibration, dictionary learning, matrix completion from nonlinear measurements**
  - ✓ **Joint signal detection and channel estimation in wireless communications**
  - ✓ **Many others…**

- **Special case: standard bilinear models**

  In particular, if Y = Z + N, where N is i.i.d Gaussian noise, BGLM reduces to the standard bilinear models

A Unified Inference Framework for GLM

◻ Bilinear GLM Problems

$$\mathbf{X} \sim p(\mathbf{X}; \boldsymbol{\theta}_X) = \prod_{i,j} p(x_{ij}; \boldsymbol{\theta}_X) = \prod_{l=1} p(\mathbf{x}_l; \boldsymbol{\theta}_X),$$

$$\mathbf{Z} = \mathbf{A}(\boldsymbol{\theta}_A)\mathbf{X},$$

$$\mathbf{Y} \sim p(\mathbf{Y}|\mathbf{Z}; \boldsymbol{\theta}_Y) = \prod_{i,j} p(Y_{ij}|Z_{ij}; \boldsymbol{\theta}_Y) = \prod_{l=1}^{L} p(\mathbf{y}_l|\mathbf{z}_l; \boldsymbol{\theta}_Y)$$

- **The optimal estimate is the Maximum likelihood (ML) and MMSE estimate as**

$$\hat{\boldsymbol{\Theta}}_{\mathrm{ML}} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}}\ p_{\mathbf{Y}}(\mathbf{Y}; \boldsymbol{\Theta}), \qquad \boldsymbol{\Theta} \triangleq \{\boldsymbol{\theta}_X, \boldsymbol{\theta}_A, \boldsymbol{\theta}_Y\}$$

$$\hat{\mathbf{X}}_{\mathrm{MMSE}} = \mathrm{E}[\mathbf{X}|\mathbf{Y}; \hat{\boldsymbol{\Theta}}_{\mathrm{ML}}],$$

Intractable!

$$p_{\mathbf{Y}}(\mathbf{Y}; \boldsymbol{\Theta}) = \int p(\mathbf{X}; \boldsymbol{\Theta}) p(\mathbf{Y}|\mathbf{X}; \boldsymbol{\Theta}) \mathrm{d}\mathbf{X}$$ **Evidence(partition function)**

$$p(\mathbf{X}|\mathbf{Y}; \hat{\boldsymbol{\Theta}}_{\mathrm{ML}}) = \frac{p(\mathbf{X}, \mathbf{Y}; \hat{\boldsymbol{\Theta}}_{\mathrm{ML}})}{p(\mathbf{Y}; \hat{\boldsymbol{\Theta}}_{\mathrm{ML}})}$$ **Posterior distribution**

A Unified Inference Framework for GLM

□ Bilinear Adaptive VAMP (BAd-VAMP)

• Bad-VAMP is proposed by Sarkar, Flecher, Rangan, Schniter [SFRS18] very recently to address the bilinear recovery from linear measurements using the adaptive VAMP and EM learning framework.

• Early work on bilinear recovery based on AMP methods, e.g., BiGAMP[PSC14a] [PSC14b] , PBGAMP [PS16] and related works by Kabashima, Krzakala and Mézard [KKMSZ16] [KMZ13]

• Compared with AMP based methods, Bad-VAMP shows improved convergence over general matrices.

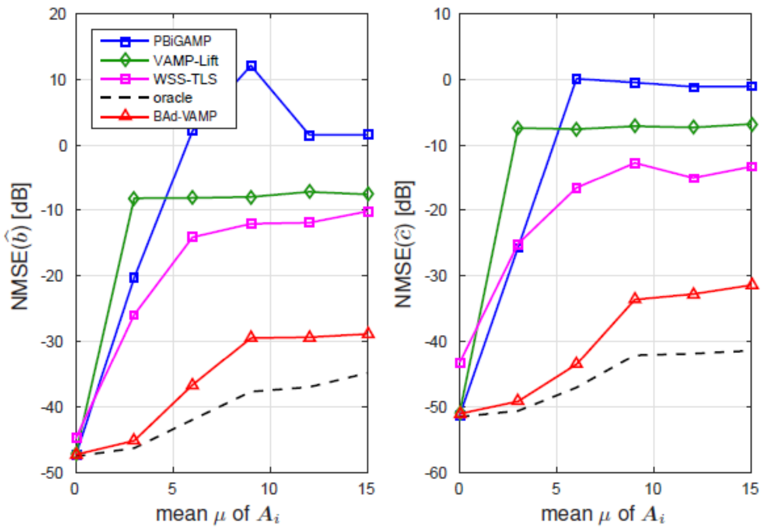


Fig. 2. CS with matrix uncertainty: Median NMSE (over 50 trials) on signal c and uncertainty parameters b versus mean of matrices A_i at $M/N = 0.6$.

Figure 2 copied from [SFRS18]

Algorithm 3 Bilinear Adaptive VAMP [SFRS18]

initialize:
$\forall l: \boldsymbol{r}_{1,l}^0, \gamma_{1,l}^0, \boldsymbol{\theta}_x^0, \boldsymbol{\theta}_A^0, \gamma_w^0$
for $t = 0, \ldots, T_{\max}$ **do**
  for $\tau = 0, \ldots, \tau_{1,\max}$ **do**
    $\forall l: \boldsymbol{x}_{1,l}^t \leftarrow \boldsymbol{g}_1(\boldsymbol{r}_{1,l}^t, \gamma_{1,l}^t; \boldsymbol{\theta}_x^t)$
    $\forall l: 1/\eta_{1,l}^t \leftarrow \langle \boldsymbol{g}_1'(\boldsymbol{r}_{1,l}^t, \gamma_{1,l}^t; \boldsymbol{\theta}_x^t)\rangle / \gamma_{1,l}^t$
    $\forall l: 1/\gamma_{1,l}^t \leftarrow \frac{1}{N}\|\boldsymbol{x}_{1,l}^t - \boldsymbol{r}_{1,l}^t\|^2 + 1/\eta_{1,l}^t$
    $q_1^t(\boldsymbol{X}) \propto \prod_{l=1}^L p_{\mathbf{x}}(\boldsymbol{x}_l; \boldsymbol{\theta}_x^t) e^{-\frac{1}{2}\gamma_{1,l}^t \|\boldsymbol{x}_l - \boldsymbol{r}_{1,l}^t\|^2}$
    $\boldsymbol{\theta}_x^t \leftarrow \arg\max_{\boldsymbol{\theta}_x} \mathbb{E}[\ln p_{\mathbf{X}}(\boldsymbol{X}; \boldsymbol{\theta}_x) | q_1^t]$
  end for
  $\boldsymbol{\theta}_x^{t+1} = \boldsymbol{\theta}_x^t$
  $\forall l: \gamma_{2,l}^t = \eta_{1,l}^t - \gamma_{1,l}^t$
  $\forall l: \boldsymbol{r}_{2,l}^t = (\eta_{1,l}^t \boldsymbol{x}_{1,l}^t - \gamma_{1,l}^t \boldsymbol{r}_{1,l}^t)/\gamma_{2,l}^t$
  for $\tau = 0, \ldots, \tau_{2,\max}$ **do**
    $\forall l: \boldsymbol{x}_{2,l}^t \leftarrow \boldsymbol{g}_{2,l}(\boldsymbol{r}_{2,l}^t, \gamma_{2,l}^t; \boldsymbol{\theta}_A^t, \gamma_w^t)$
    $\forall l: 1/\eta_{2,l}^t \leftarrow \langle \boldsymbol{g}_{2,l}'(\boldsymbol{r}_{2,l}^t, \gamma_{2,l}^t; \boldsymbol{\theta}_A^t, \gamma_w^t)\rangle / \gamma_{2,l}^t$
    $\forall l: 1/\gamma_{2,l}^t \leftarrow \frac{1}{N}\|\boldsymbol{x}_{2,l}^t - \boldsymbol{r}_{2,l}^t\|^2 + 1/\eta_{2,l}^t$
    $q_2^t(\boldsymbol{X}) \propto \prod_l p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y}_l | \boldsymbol{x}_l; \boldsymbol{\theta}_A^t, \gamma_w^t) e^{-\frac{1}{2}\gamma_{2,l}^t \|\boldsymbol{x}_l - \boldsymbol{r}_{2,l}^t\|^2}$
    $\boldsymbol{\theta}_A^t \leftarrow \arg\max_{\boldsymbol{\theta}_A} \mathbb{E}[\ln p_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{Y}|\boldsymbol{X}; \boldsymbol{\theta}_A, \gamma_w^t) | \boldsymbol{Y}, q_2^t]$
    $\gamma_w^t \leftarrow \arg\max_{\gamma_w} \mathbb{E}[\ln p_{\mathbf{Y}|\mathbf{X}}(\boldsymbol{Y}|\boldsymbol{X}; \boldsymbol{\theta}_A^t, \gamma_w) | \boldsymbol{Y}, q_2^t]$
  end for
  $\boldsymbol{\theta}_A^{t+1} = \boldsymbol{\theta}_A^t$
  $\gamma_w^{t+1} = \gamma_w^t$
  $\forall l: \gamma_{1,l}^{t+1} = \eta_{2,l}^t - \gamma_{2,l}^t$
  $\forall l: \boldsymbol{r}_{1,l}^{t+1} = (\eta_{2,l}^t \boldsymbol{x}_{2,l}^t - \gamma_{2,l}^t \boldsymbol{r}_{2,l}^t)/\gamma_{1,l}^{t+1}$
end for

A Unified Inference Framework for GLM

◻ Bilinear Adaptive VAMP (BAd-VAMP)

• Bad-VAMP is proposed by Sarkar, Flecher, Rangan, Schniter [SFRS18] very recently to address the bilinear recovery from linear measurements using the adaptive VAMP and EM learning framework.

• Early work on bilinear recovery based on AMP methods, e.g., BiGAMP[PSC14a] [PSC14b] , PBGAMP [PS16] and related works by Kabashima, Krzakala and Mézard [KKMSZ16] [KMZ13]

• Compared with AMP based methods, Bad-VAMP shows improved convergence over general matrices.

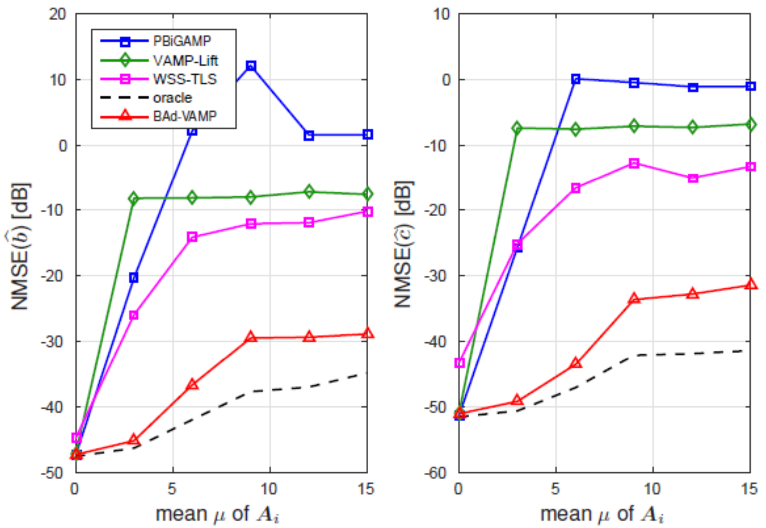


Fig. 2. CS with matrix uncertainty: Median NMSE (over 50 trials) on signal c and uncertainty parameters b versus mean of matrices A_i at $M/N = 0.6$.

Figure 2 copied from [SFRS18]

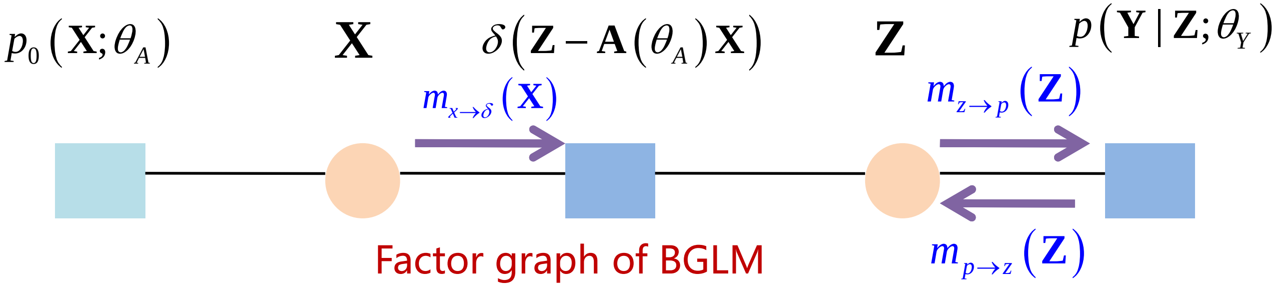
Algorithm 3 Bilinear Adaptive VAMP [SFRS18]

initialize:
  $\forall l : \boldsymbol{r}^0_{1,l}, \gamma^0_{1,l}, \boldsymbol{\theta}^0_x, \boldsymbol{\theta}^0_A, \gamma^0_w$
for $t = 0, \ldots, T_{\max}$ **do**
  for $\tau = 0, \ldots, \tau_{1,\max}$ **do**
    $\forall l : \boldsymbol{x}^t_{1,l} \leftarrow \boldsymbol{g}_1(\boldsymbol{r}^t_{1,l}, \gamma^t_{1,l}; \boldsymbol{\theta}^t_x)$
    $\forall l : 1/\eta^t_{1,l} \leftarrow \langle \boldsymbol{g}'_1(\boldsymbol{r}^t_{1,l}, \gamma^t_{1,l}; \boldsymbol{\theta}^t_x)\rangle / \gamma^t_{1,l}$
    $\forall l : 1/\gamma^t_{1,l} \leftarrow \frac{1}{N}\|\boldsymbol{x}^t_{1,l} - \boldsymbol{r}^t_{1,l}\|^2 + 1/\eta^t_{1,l}$
    $q^t_1(\boldsymbol{X}) \propto \prod_{l=1}^{L} p_{\mathbf{x}}(\boldsymbol{x}_l; \boldsymbol{\theta}^t_x) e^{-\frac{1}{2}\gamma^t_{1,l}\|\boldsymbol{x}_l - \boldsymbol{r}^t_{1,l}\|^2}$
    $\boldsymbol{\theta}^t_x \leftarrow \arg\max_{\boldsymbol{\theta}_x} \mathbb{E}[\ln p_{\boldsymbol{X}}(\boldsymbol{X}; \boldsymbol{\theta}_x) | q^t_1]$
  end for
  $\boldsymbol{\theta}^{t+1}_x = \boldsymbol{\theta}^t_x$
  $\forall l : \gamma^t_{2,l} = \eta^t_{1,l} - \gamma^t_{1,l}$
  $\forall l : \boldsymbol{r}^t_{2,l} = (\eta^t_{1,l}\boldsymbol{x}^t_{1,l} - \gamma^t_{1,l}\boldsymbol{r}^t_{1,l})/\gamma^t_{2,l}$
  for $\tau = 0, \ldots, \tau_{2,\max}$ **do**
    $\forall l : \boldsymbol{x}^t_{2,l} \leftarrow \boldsymbol{g}_{2,l}(\boldsymbol{r}^t_{2,l}, \gamma^t_{2,l}; \boldsymbol{\theta}^t_A, \gamma^t_w)$
    $\forall l : 1/\eta^t_{2,l} \leftarrow \langle \boldsymbol{g}'_{2,l}(\boldsymbol{r}^t_{2,l}, \gamma^t_{2,l}; \boldsymbol{\theta}^t_A, \gamma^t_w)\rangle / \gamma^t_{2,l}$
    $\forall l : 1/\gamma^t_{2,l} \leftarrow \frac{1}{N}\|\boldsymbol{x}^t_{2,l} - \boldsymbol{r}^t_{2,l}\|^2 + 1/\eta^t_{2,l}$
    $q^t_2(\boldsymbol{X}) \propto \prod_l p_{\mathbf{y}|\mathbf{x}}(\boldsymbol{y}_l | \boldsymbol{x}_l; \boldsymbol{\theta}^t_A, \gamma^t_w) e^{-\frac{1}{2}\gamma^t_{2,l}\|\boldsymbol{x}_l - \boldsymbol{r}^t_{2,l}\|^2}$
    $\boldsymbol{\theta}^t_A \leftarrow \arg\max_{\boldsymbol{\theta}_A} \mathbb{E}[\ln p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{Y}|\boldsymbol{X}; \boldsymbol{\theta}_A, \gamma^t_w) | \boldsymbol{Y}, q^t_2]$
    $\gamma^t_w \leftarrow \arg\max_{\gamma_w} \mathbb{E}[\ln p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{Y}|\boldsymbol{X}; \boldsymbol{\theta}^t_A, \gamma_w) | \boldsymbol{Y}, q^t_2]$
  end for
  $\boldsymbol{\theta}^{t+1}_A = \boldsymbol{\theta}^t_A$
  $\gamma^{t+1}_w = \gamma^t_w$
  $\forall l : \gamma^{t+1}_{1,l} = \eta^t_{2,l} - \gamma^t_{2,l}$
  $\forall l : \boldsymbol{r}^{t+1}_{1,l} = (\eta^t_{2,l}\boldsymbol{x}^t_{2,l} - \gamma^t_{2,l}\boldsymbol{r}^t_{2,l})/\gamma^{t+1}_{1,l}$
end for

Question: How to extend BAd-VAMP to general nonlinear measurements ?

A Unified Inference Framework for GLM

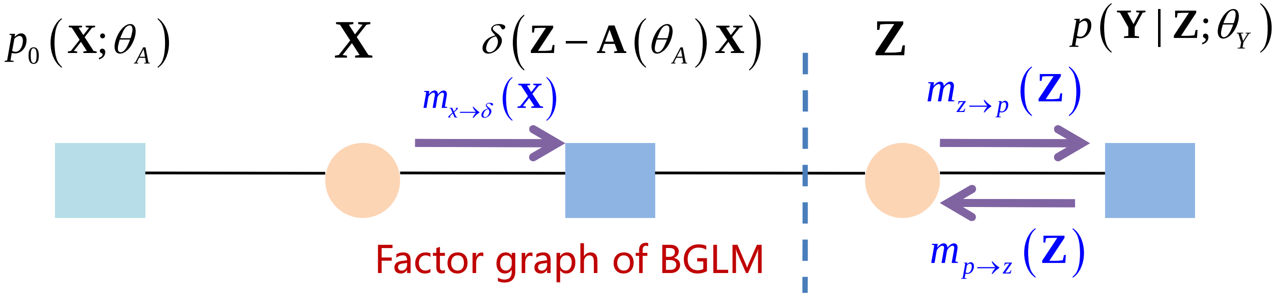
- **Bilinear Adaptive Generalized VAMP (BAd-GVAMP)** [MZ18]



Factor graph of BGLM

A Unified Inference Framework for GLM

◻ Bilinear Adaptive Generalized VAMP (BAd-GVAMP) [MZ18]



Factor graph of BGLM

- **Similar to GLM, using EP, the BGLM can be decoupled into two modules**

$$m_{z\to p}(\mathbf{Z})\propto\prod_{l=1}^{L}\mathcal{N}\left(\mathbf{z}_l;\mathbf{z}_{A,l}^{ext},v_A^{ext}I\right)\triangleq\prod_{l=1}^{L}m_{z\to p}(\mathbf{z}_l)$$

$$m_{p\to z}(\mathbf{Z})\propto\prod_{l=1}^{L}\mathcal{N}\left(\mathbf{z}_l;\mathbf{z}_{B,l}^{ext},v_B^{ext}I\right)$$

$$\frac{1}{v_B^{ext}(t)}=\frac{1}{v_B^{post}(t)}-\frac{1}{v_A^{ext}(t-1)}$$

$$\frac{\mathbf{z}_{B,l}^{ext}(t)}{v_B^{ext}(t)}=\frac{\mathbf{z}_{B,l}^{post}(t)}{v_B^{post}(t)}-\frac{\mathbf{z}_{A,l}^{ext}(t-1)}{v_A^{ext}(t-1)}$$

$$\mathbf{z}_{B,l}^{post}=\mathrm{E}\left[z_l\mid\mathbf{z}_{A,l}^{ext},v_A^{ext}\right]$$

$$v_B^{post}=\left\langle\mathrm{Var}\left[z_l\mid\mathbf{z}_{A,l}^{ext},v_A^{ext}\right]\right\rangle$$

$$m_{z\to p}(\mathbf{z}_l)\propto\frac{\mathrm{Proj}_{\Phi}\left(m_{p\to z}(\mathbf{z}_l)\int\delta\left(\mathbf{z}_l-\mathbf{A}(\theta_A)\mathbf{x}_l\right)m_{x\to\delta}(\mathbf{x}_l)\,d\mathbf{x}_l\right)}{m_{p\to z}(\mathbf{z}_l)}\triangleq\frac{\mathrm{Proj}_{\Phi}\left(q_A(\mathbf{z}_l)\right)}{m_{p\to z}(\mathbf{z}_l)}$$

$m_{x\to\delta}(\mathbf{x}_l)$ has already been computed within BAd-VAMP as $m_{x\to\delta}(\mathbf{x}_l)\propto\mathcal{N}\left(\mathbf{x}_l;\mathbf{r}_{2,l},1/\gamma_{2,l}\mathbf{I}\right)$, then

$$q_A(\mathbf{z}_l)\propto\mathcal{N}\left(\mathbf{z}_l;\mathbf{z}_{A,l}^{post},\Xi_{A,l}^{post}\right)$$

$$\Xi_{A,l}^{post}=\mathbf{A}(\theta_A)\left[\gamma_{2,l}\mathbf{I}+\tilde{\gamma}_w\mathbf{A}^T(\theta_A)\mathbf{A}^T(\theta_A)\right]^{-1}\mathbf{A}^T(\theta_A)$$

$$\mathbf{z}_{A,l}^{post}=\mathbf{A}(\theta_A)\left[\gamma_{2,l}\mathbf{I}+\tilde{\gamma}_w\mathbf{A}^T(\theta_A)\mathbf{A}^T(\theta_A)\right]^{-1}\left(\gamma_{2,l}\mathbf{r}_{2,l}\tilde{\gamma}_w\mathbf{A}^T(\theta_A)\tilde{\mathbf{y}}_l\right)$$

Gaussian with

Scalar covariance matrix

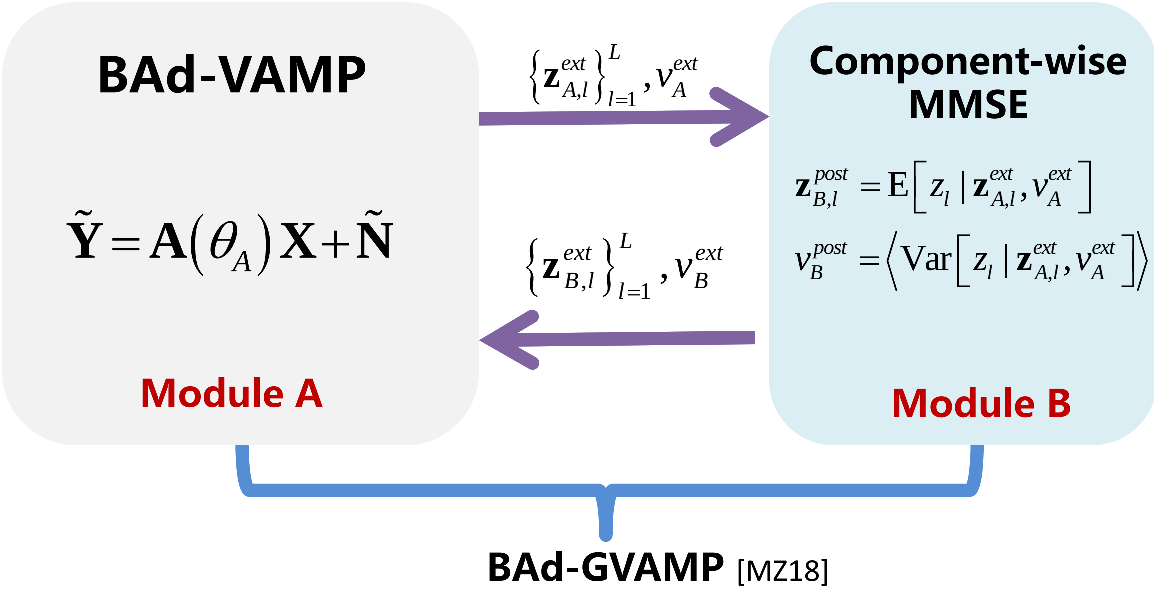
$$\mathrm{Proj}_{\Phi}\left(q_A(\mathbf{z}_l)\right)\propto\mathcal{N}\left(\mathbf{z}_l;\mathbf{z}_{A,l}^{post},v_A^{post}\mathbf{I}\right)$$

$$v_A^{post}=\left\langle\frac{1}{M}\mathrm{Trace}(\Xi_{A,l}^{post})\right\rangle$$

A Unified Inference Framework for GLM

☐ **Bilinear Adaptive Generalized VAMP (BAd-GVAMP)** [MZ18]

- **Similar to GLM, using EP, the BGLM can be decoupled into two modules**



- **Relation of BAd-GVAMP to BAd-VAMP**
  - ✓ An extension of the BAd-VAMP [SFRS18] from linear measurements to nonlinear measurements.
  - ✓ The BAd-GVAMP iteratively reduces the original generalized bilinear recovery problem to a sequence of standard bilinear recovery problems
  - ✓ Note that the message passing schedule within BAd-VAMP module of BAd-GVAMP is slightly different from that of the original BAd-VAMP
  - ✓ In the special case of linear measurements, BAd-GVAMP reduces to BAd-VAMP

[MZ18] X. Meng, and J. Zhu, “Bilinear Adaptive Generalized Vector Approximate Message Passing,” arXiv preprint arXiv:1810.08129, 2018

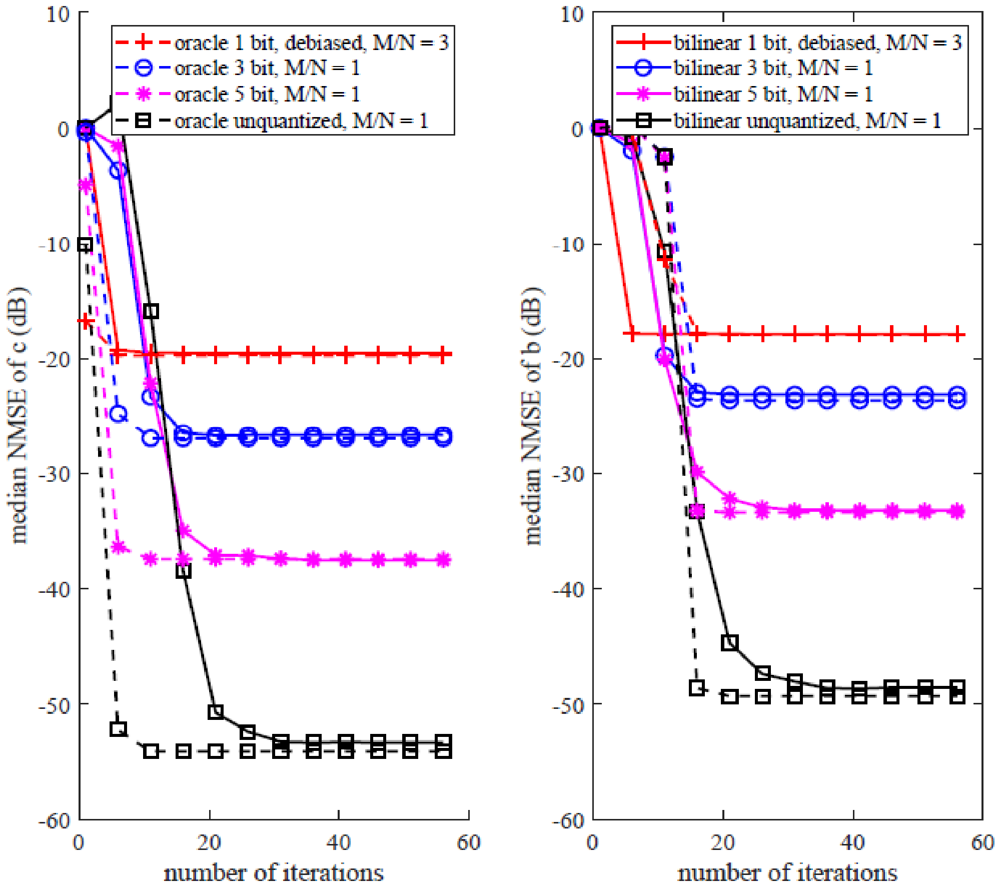
A Unified Inference Framework for GLM

☐ Bilinear Adaptive Generalized VAMP (BAd-GVAMP) [MZ18]

- **Experiment 1: Quantized Compressed Sensing with matrix uncertainty**

$$\mathbf{y} = Q(\mathbf{A}(\mathbf{b})\mathbf{c}+\mathbf{w}) \qquad \mathbf{A}(\mathbf{b}) = \mathbf{A}_0 + \sum_{i=1}^{G} b_i \mathbf{A}_i$$

$\{\mathbf{A}_i\}_{i=0}^{G} \in \mathbb{R}^{M\times N}$ are known, $\mathbf{b}$ are the unknown uncertainty parameters.



$$\text{SNR} \triangleq 10\log\frac{\text{E}\|\mathbf{A}\mathbf{c}\|^2}{\text{E}\|\mathbf{w}\|^2} = 40 \text{ dB}$$

✓ c is generated with uniformly random support with K nonzero elements from i.i.d N(0,1), we set N = 256, G = 10, K = 10

✓ For M/N = 1, the NMSE in dB is shown in left figure:

-- converges fast (20-30 iterations)

-- the same as the oracle performance.

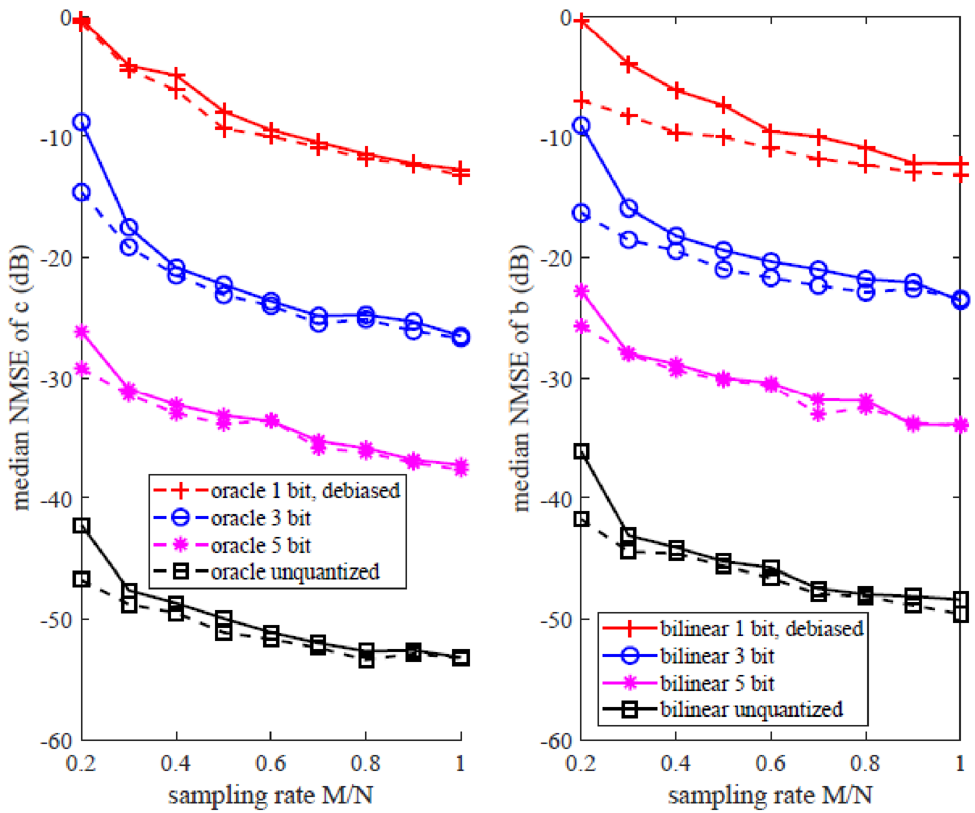
A Unified Inference Framework for GLM

□ Bilinear Adaptive Generalized VAMP (BAd-GVAMP) [MZ18]

- **Experiment 1: Quantized Compressed Sensing with matrix uncertainty**

$$\mathbf{y} = Q(\mathbf{A}(\mathbf{b})\mathbf{c} + \mathbf{w}) \qquad \mathbf{A}(\mathbf{b}) = \mathbf{A}_0 + \sum_{i=1}^{G} b_i \mathbf{A}_i$$

$\{\mathbf{A}_i\}_{i=0}^{G} \in \mathbb{R}^{M\times N}$ are known, $\mathbf{b}$ are the unknown uncertainty parameters.



$$\mathrm{SNR} \triangleq 10\log\frac{\mathrm{E}\|\mathbf{A}\mathbf{c}\|^2}{\mathrm{E}\|\mathbf{w}\|^2} = 40\ \mathrm{dB}$$

✓ c is generated with uniformly random support with K nonzero elements from i.i.d N(0,1), we set N = 256, G = 10, K = 10

✓ Then, the performance vs. ratio M/N is evaluated:

-- as the increase of M/N, the recovery performance improves

-- approaches the oracle in a wide range of M/N values

A Unified Inference Framework for GLM

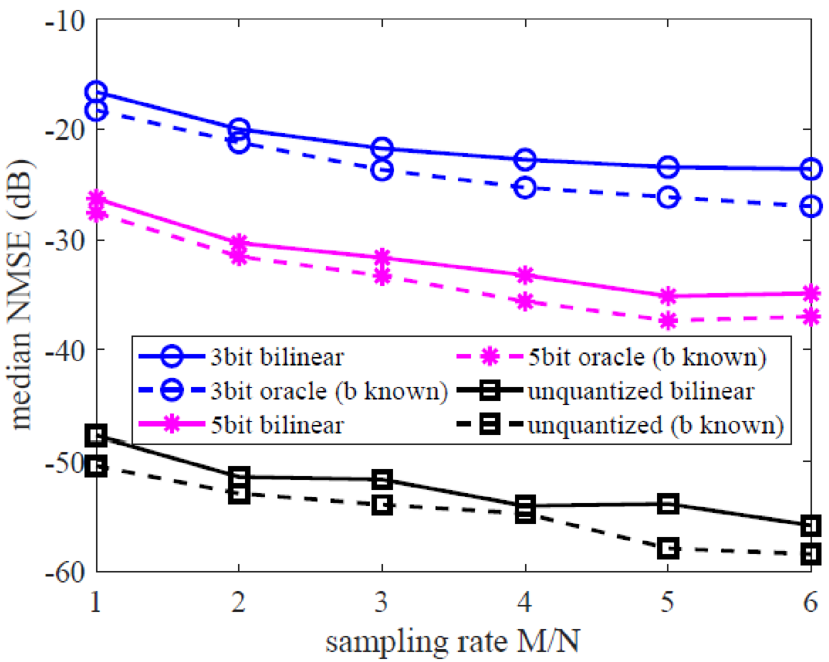
☐ Bilinear Adaptive Generalized VAMP (BAd-GVAMP) [MZ18]

- **Experiment 2: Self-Calibration from quantized measurements**

$$\mathbf{y} = Q(\mathrm{diag}(\mathbf{Hb})\mathbf{\Psi}\mathbf{c} + \mathbf{w}) = Q\left(\left[\sum_{i=1}^{G} b_i \mathrm{diag}(\mathbf{h}_i)\mathbf{\Psi}\right]\mathbf{c} + \mathbf{w}\right)$$

$$\text{with known } \mathbf{H} \in \mathbb{R}^{M\times G} \text{ and } \mathbf{\Psi} \in \mathbb{R}^{M\times N}$$

Aim: to recover the K-sparse signal vector c and the calibration parameters b



✓ K = 10, G = 8, M = 128 and SNR = 40 dB.
✓ H is constructed using Q randomly selected columns of the Hadamard matrix, the elements of **b** and $\mathbf{\Psi}$ are i.i.d. drawn from N(0; 1), and c is generated with *K* nonzero elements i.i.d. drawn from N(0; 1).

$$\mathrm{NMSE} = 10\log\frac{\|\hat{\mathbf{b}}\hat{\mathbf{c}}^{\mathrm{T}} - \mathbf{b}\mathbf{c}^{\mathrm{T}}\|_{\mathrm{F}}^2}{\|\mathbf{b}\mathbf{c}^{\mathrm{T}}\|_{\mathrm{F}}^2}$$

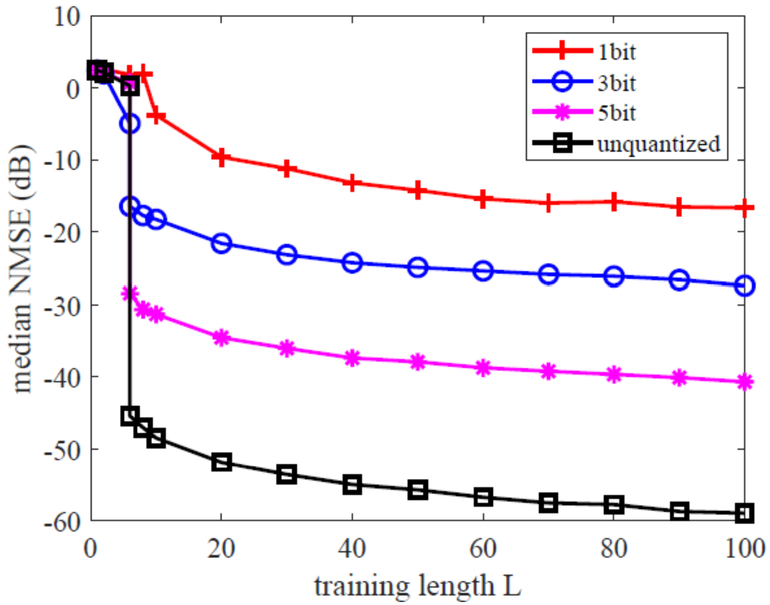
✓ As the sampling rate increases, the median NMSE decreases. Also, the reconstruction performance improves as the bit-depth increases.

A Unified Inference Framework for GLM

☐ Bilinear Adaptive Generalized VAMP (BAd-GVAMP) [MZ18]

- **Experiment 3: Structured dictionary learning from quantized measurements**

The goal of dictionary learning is to find a dictionary matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ and a sparse matrix $\mathbf{X} \in \mathbb{R}^{N \times L}$ such that $\mathbf{Y} \approx \mathbf{AX}$ for a given matrix $\mathbf{Y} \in \mathbb{R}^{M \times L}$. We consider structured dictionary $\mathbf{A}$ such that $\mathbf{A} = \sum_{i=1}^{G} b_i \mathbf{A}_i$ with known $\{\mathbf{A}_i\}_{i=1}^{G}$, where the elements of $\mathbf{A}_i$ and b_i are i.i.d. drawn from $\mathcal{N}(0,1)$ with $G = M = N = 64$ in the structured case. Then the measurements are obtained as $\mathbf{Y} = Q(\mathbf{AX} + \mathbf{W})$



✓ G=M=N=64 and SNR = 40 dB.

$$\mathrm{NMSE}(\hat{\mathbf{A}}) \triangleq \min_{\lambda \in \mathbb{R}} \frac{\|\mathbf{A} - \lambda \hat{\mathbf{A}}\|_F^2}{\|\mathbf{A}\|_F^2}$$

✓ As the training length L increases, the NMSE decreases. And the structured dictionary can be learned from quantized measurements.

Conclusions

- **Considers the design of efficient GLM inference algorithms**
- **Review the AMP algorithm and provides an EP perspective**
- **Present a unified approximate inference framework for GLM**
  - **Facilitates the extension of various SLM inference algorithms to GLM inference in a simple and unified manner**
  - **Provides some new insights on some well-known algorithms, e.g., GAMP, thus offering a flexible way in the message scheduling of practical implementation**
  - **Extend further to the bilinear GLM problem and propose the BAd-GVAMP, extending BAd-VAMP to nonlinear measurements**

Conclusions

- **Considers the design of efficient GLM inference algorithms**
- **Review the AMP algorithm and provides an EP perspective**
- **Present a unified approximate inference framework for GLM**
  - **Facilitates the extension of various SLM inference algorithms to GLM inference in a simple and unified manner**
  - **Provides some new insights on some well-known algorithms, e.g., GAMP, thus offering a flexible way in the message scheduling of practical implementation**
  - **Extend further to the bilinear GLM problem and propose the BAd-GVAMP, extending BAd-VAMP to nonlinear measurements**

- **Possible future work**
  - **Theoretical analysis of this unified framework**
  - **Evaluate or analyze the effect of different ways of message scheduling for GLM**
  - **Design other efficient GLM inference algorithms from SLM ones**
  - **Extend to multi-layer neural network to see if it helps in the learning and/or inference process in deep neural network.**

References

- [DMM09]Donoho, Maleki, Montanari. "Message-passing algorithms for compressed sensing." Proceedings of the National Academy of Sciences 106.45 (2009): 18914-18919.
- [DMM10] Donoho, Maleki, Montanari, "Message passing algorithms for compressed sensing: I. motivation and construction." Proc IEEE ITW, 20110
- [BM11] Bayati, Montanari. "The dynamics of message passing on dense graphs, with applications to compressed sensing." IEEE Transactions on Information Theory 57.2 (2011): 764-785.
- [Rangan10] Rangan, Sundeep. "Estimation with random linear mixing, belief propagation and compressed sensing." Information Sciences and Systems (CISS), 2010 44th Annual Conference on. IEEE, 2010.
- [Tanaka 02] Tanaka, Toshiyuki. "A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors." IEEE Transactions on Information theory 48.11 (2002): 2888-2910.
- [Minka01] Minka, Thomas P. "Expectation propagation for approximate Bayesian inference." 2001
- [OW05] Opper, Manfred, and Ole Winther. "Expectation consistent approximate inference." Journal of Machine Learning Research 6.Dec (2005): 2177-2204.
- [Rangan11] Rangan, "Generalized approximate message passing for estimation with random linear mixing." Proc IEEE ISIT 2011
- [RSF16] Rangan, Schniter, Fletcher, "Vector approximate message passing", 2016
- [SRF16] P. Schniter, S. Rangan, and A. K. Fletcher, "Vector approximate message passing for the generalized linear model," in Proc. 50th Asilomar Conf. Signals, Syst. Comput., Nov. 2016, pp. 1525-1529.
- [Kabashima 03] Kabashima Y. A , " CDMA multiuser detection algorithm on the basis of belief propagation" , Journal of Physics A: Mathematical and General, 2003, 36(43)
- [KMSSZ12] F. Krzakala, M. Mézard, F. Sausset, Y. Sun, and L. Zdeborová, "Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices," Journal of Statistical Mechanics: Theory and Experiment, vol. 2012, no. 08, p. P08009, 2012.
- [MWKL15a] X. Meng, S. Wu, L. Kuang, and J. Lu, "An expectation propagation perspective on approximate message passing," IEEE Signal Process. Lett., vol. 22, no. 8, pp. 1194-1197, Aug. 2015.
- [MWKL15b] X. Meng, S. Wu, L. Kuang, and J. Lu, " Concise derivation of complex Bayesian approximate message passing via expectation propagation," arXiv preprint arXiv:1509.08658, 2015.
- [WKNLHDQ14] S. Wu, L. Kuang, Z. Ni, J. Lu, D. Huang, and Q. Guo, "Low-complexity iterative detection for large-scale multiuser MIMO-OFDM systems using approximate message passing," *IEEE J. Sel. Topics Signal Process.,* vol. 8, no. 5, pp. 902–915, Oct. 2014.
- [MWZ18]X. Meng, S. Wu and J. Zhu, "A unified Bayesian inference framework for generalized linear model," IEEE Signal Process. Lett., vol. 25, no. 3, Mar. 2018.

References

- [MZ18] X. Meng, and J. Zhu, “Bilinear Adaptive Generalized Vector Approximate Message Passing,” arXiv preprint arXiv:1810.08129, 2018
- [PSC14a] J. T. Parker, P. Schniter, and V. Cevher, “Bilinear generalized approximate message passing-Part I: Derivation,” IEEE Trans. Signal Process., vol. 62, no. 22, pp. 5839-5853, Nov. 2014.
- [PSC14b] J. T. Parker, P. Schniter, and V. Cevher, “Bilinear generalized approximate message passing-Part II: Applications,” IEEE Trans. Signal Process., vol. 62, no. 22, pp. 5854-5867, Nov. 2014.
- [PS16] J. T. Parker and P. Schniter, “Parametric bilinear generalized approximate message passing,” IEEE J. Sel. Topics Signal Process., vol. 10, no. 4, pp. 795-808, 2016.
- [KKMSZ16] Y. Kabashima, F. Krzakala, M. M´ezard, A. Sakata and L. Zdeborov´a, “Phase transitions and sample complexity in bayesoptimal matrix factorization,” IEEE Trans. Inf. Theory, vol. 62, no. 7, pp. 4228-4265, 2016.
- [SRF] P. Schniter, S. Rangan, and A. K. Fletcher, “Vector approximate message passing for the generalized linear model,” in Proc. 50th Asilomar Conf. Signals, Syst. Comput., Nov. 2016, pp. 1525-1529.
- [ML17] J. Ma and L. Ping, “Orthogonal AMP,” IEEE Access, vol. 5, pp. 2020-2033, 2017.
- [HWJ17] H. He, C. K. Wen, and S. Jin, “Generalized expectation consistent signal recovery for nonlinear measurements,” in Proc. IEEE Int. Symp. Inf. Theory, Jun. 2017, pp. 2333-2337.
- [QZW18] Zou Q, Zhang H, Wen C K, et al. , “ Concise Derivation for Generalized Approximate Message Passing Using Expectation Propagation ,” IEEE Signal Processing Letters, 2018.
- [SFRS18] S. Sarkar, A. K. Fletcher, S. Rangan and P. Schniter, “Bilinear recovery using adaptive vector-AMP,” avaliable at
- https://arxiv.org/pdf/1809.00024.pdf.
- [VS13] J. P. Vila and P. Schniter, “Expectation-maximization Gaussian-mixture approximate message passing,” IEEE Trans. Signal Process., vol. 61, no. 19, pp. 4658-4672, Oct. 2013.
- [KMZ13] F. Krzakala, M. Mezard, and L. Zdeborova, “Compressed sensing under matrix uncertainty: Optimum thresholds and robust approximate message passing,” ICASSP, 2013.

Thanks！