Estimation for High-Dimensional Multi-Layer Generalized Linear Model – Part II: The ML-GAMP Estimator
Qiuyun Zou, Haochuan Zhang*, and Hongwen Yang

Abstract—This is Part II of a two-part work on the estimation for a multi-layer generalized linear model (ML-GLM) in large system limits. In Part I, we had analyzed the asymptotic performance of an exact MMSE estimator, and obtained a set of coupled equations that could characterize its MSE performance. To work around the implementation difficulty of the exact estimator, this paper continues to propose an approximate solution, ML-GAMP, which could be derived by blending a moment-matching projection into the Gaussian approximated loopy belief propagation. The ML-GAMP estimator is then shown to enjoy a great simplicity in its implementation, where its per-iteration complexity is as low as GAMP. Further analysis on its asymptotic performance also reveals that, in large system limits, its dynamical MSE behavior is fully characterized by a set of simple one-dimensional iterating equations, termed state evolution (SE). Interestingly, this SE of ML-GAMP share exactly the same fixed points with an exact MMSE estimator whose fixed points were obtained in Part I via a replica analysis. Given the Bayes-optimality of the exact implementation, this proposed estimator (if converged) is optimal in the MSE sense.

Index Terms—multi-layer generalized linear model (ML-GLM), minimal mean square error (MMSE), multi-layer generalized approximate message passing (ML-GAMP), state evolution (SE)

I. INTRODUCTION

A. Problem Statement and Recap on Part I

This two-part work considers the estimation of high-dimensional random signal in a multi-layer generalized linear model (ML-GLM), which is illustrated as Fig. 1. In the figure, $x_0$ denotes the random initial input, whose distribution is factorable and known perfectly by the estimator, i.e., $x_0 \sim P_X(x_0) = \prod_{i=1}^{N_1} P_X(x_{0i})$. $y$ denotes the observation attained from the ML-GLM network of $L$ layers, and $\langle x \rangle$ is the MMSE estimator’s output, either exact or approximate. Particularly, the $\ell$-th layer expands as (1 $\leq \ell \leq L$)

$$
\begin{align*}
    & (x^{(\ell)}) \rightarrow H^{(\ell)} x^{(\ell)} \rightarrow z^{(\ell)} \rightarrow P(z^{(\ell+1)}|z^{(\ell)}) \rightarrow x^{(\ell+1)} \rightarrow (1)
\end{align*}
$$

where $x^{(\ell)} \in \mathbb{R}^{N_\ell}$ is its input, and $H^{(\ell)} \in \mathbb{R}^{N_{\ell+1} \times N_\ell}$ is a deterministic weighting matrix that linearly mixes up the input to yield $z \in \mathbb{R}^{N_{\ell+1}}$. This weighted result $z$ is then activated by a random mapping, whose transitional/conditional probability density function (p.d.f.) is also factorable: $P(x^{(\ell+1)}|z^{(\ell)}) = \prod_{a=1}^{N_{\ell+1}} P(x^{(\ell+1)}|z_a^{(\ell)})$. The weighting matrix above is known perfectly to the estimator, and in each experiment, the elements of this matrix are drawn independently from the same Gaussian ensemble of zero mean and $1/N_{\ell+1}$ variance (to ensure a unit row norm). To maintain notational consistency, we also initialize: $x^{(1)} := x_0$, and $x^{(L+1)} := y$. Since we consider exclusively the limiting performance of the MMSE estimators, the following assumptions are made throughout the paper: $N_{\ell} \rightarrow \infty$, while $N_{\ell+1}/N_{\ell} \rightarrow \alpha_{\ell}$, i.e., all weighting matrices are sufficiently large in size, but the ratios of their row numbers to column numbers are fixed and bounded.

The target of an exact MMSE estimator is to generate an estimate $\langle x_k \rangle$ for every input element $x_{0k}$ using $(k = 1, \cdots, N_1)$

$$
\langle x_k \rangle = \arg \min_{\hat{x}_k} \mathbb{E} \left[ \| \hat{x}_k - x_k \|^2 \right] = \mathbb{E} [x_k | y, \{H^{(\ell)}\}] (2)
$$

where the last expectation is taken over a marginal posterior

$$
P(x_{0k}|y, \{H^{(\ell)}\}) = \int P(x_{0}|y, \{H^{(\ell)}\})dx_{0\backslash k}, (3)
$$

whose integration is $(N_1 - 1)$-fold, $x_{0\backslash k}$ equals $x_0$ except its $k$-th element moved, and the joint p.d.f. $P(x_{0}|y, \{H^{(\ell)}\})$ is:

$$
P(x_{0}|y, \{H^{(\ell)}\}) = \frac{P_X(x_{0})P(y|x_0, \{H^{(\ell)}\})}{\int P_X(x_{0})P(y|x_0, \{H^{(\ell)}\})dx_0}.
$$

Due to the difficulty in the evaluation of the $(N_1 - 1)$-fold integrals above, an exact implementation of the MMSE estimator is infeasibility (1), and as a consequence, this Part II paper considers its effective approximation, and proposes a new estimator called multi-layer generalized approximate message passing (ML-GAMP), which is later shown MSE-optimal and computationally efficient.

But before proceeding further, a short recap on Part I’s findings from replica analysis is given first, considering exclusively an exact MMSE estimator:
1) In terms of joint input-and-estimate distribution, the original estimation problem of MIMO nature was identical to that of a simple SISO estimation problem facing no self-interference (caused by the linear weighting), no nonlinear distortion, (caused by the random mapping), but only an effective AWGN. That is,

\[(x_{0k}, x_k) \sim (X_0, X), \quad \forall k,\] (4)

As an illustration, Fig. 2 depicts the MIMO-to-SISO decoupling principle for a relatively simple case of two layers.

2) The noise level of this AWGN could be further determined by solving a set of coupled equations (given in Claim 1 of Part I), whose dependency on the linear weighting and the random mapping was given explicitly (also in Claim 1).

3) As a byproduct of the replica analysis, the average MSE of the exact MMSE estimator could be computed directly from the fixed-point results, in which case time-consuming simulations in the Mote Carlo manner are no longer needed.

B. Contributions of Part II

Given the infeasibility of an exact MMSE implementation in large system limit, we propose in Part II (this paper) a new estimator called ML-GAMP, which is asymptotically optimal in MSE sense but still enjoys a great computational simplicity. For its derivation, we, instead of approximating directly the exact MMSE estimation, resort to the message passing in random graphical models, where a method developed by us in an earlier study [2] could be applied to find new algorithms for the posterior mean estimation (PME). The method developed was in general under the framework of the celebrated loopy belief propagation (LBP) [3], but some of its message updating had already been a more powerful variational inference tool, the expectation propagation (EP) [4, 1], or the expectation consistent (EC) [5]. This blended method was shown [2] capable of recovering the GAMP algorithm [6, 7] in GLM (generalized linear model), and will continue to be demonstrated applicable to the more general ML-GLM setting.

In what follows, we summarize the contributions of Part II:

1) A new estimator called ML-GAMP is proposed for the estimation of the ML-GLM problem. Comparing to prior works that are closely related, our proposed estimator provides a better tradeoff between performance and complexity. To be specific, although both could attain an MSE-optimal performance, our estimator converges in a speed much faster than the ML-AMP [8] proposed by Manoel et al., which we believe was the first to put forward a solution for the challenging problem of ML-GLM estimation. An investigation to this difference in convergence rate is also carried out, which suggests that our iteration could make full use of the messages most recently updated while the ML-AMP [8] put more faith in the outdated ones, so the two diverges and ours becomes superior (especially in the correct-updating-dominant cases). Another solution that solved the ML-GLM problem was ML-VAMP [9] proposed by Fletcher et al., which was a significant extension to VAMP [10], developed by Rangan et al. and tightly connected with EP [4] and EC [5]. This ML-VAMP [9], however, required during its iterations (at least in the very first) to perform a singular value decomposition (SVD) of \( O(N^3) \) complexity on each weighting matrix, which inevitably compromised the computational efficiency, as opposed to the scalar-only operations in ML-AMP [8] and ours. In this context, the proposed estimator has less computational burden than the ML-VAMP [9].

2) An set of simple one-dimensional iterative equations, termed the state evolution (SE), is derived from ML-GAMP, which could characterize fully the proposed estimator’s dynamical MSE behavior under a mild assumption on weighting matrices of zero-mean and bounded-variance Gaussian distributions. Given the SE, the average MSE of ML-GAMP could be easily obtained by iterating the SE. The accuracy of this SE iteration is then validated via numerical experiments, which clearly show that not only the overall MSE at the final step of the estimation is accurate, but also those in the middle (algorithm not-yet converged) steps are accurately captured by the SE.

3) A perfect agreement is observed between the above SE and the fixed point equations of an exact MMSE estimator (given in Claim 1 of Part I). Since the fixed point equations were analyzed from an exact (rather than approximate) MMSE implementation and the average MSE of the exact MMSE estimator could be computed directly from the fixed equations, the above agreement provides an interesting and very important piece of evidence that the proposed ML-GAMP (if converged) could attain the same MSE performance as an exact MMSE estimator, and thus is Bayes-optimal in the MSE sense.

II. ML-GAMP: A MESSAGE PASSING ALGORITHM

This section provides a description for the proposed ML-GAMP, together with its detailed derivation via a moment-matching-based LBP method, and a simplified version to further improve its computational efficiency.

A. Description of the Algorithm

**Proposal 1.** To estimate \( x \) for the ML-GLM problem depicted as [7] we propose a new algorithm, termed ML-GAMP, whose detail is given in Algorithm [7].

On ML-GAMP, we have the following remarks.

**Remark 1.** The ML-GAMP degenerates smoothly to the celebrated AMP [11] and GAMP [6]. Thanks to the generality of the ML-GLM model, the proposed estimator embraces many
The Proposed ML-GLM Estimator

**Algorithm 1:** The Proposed ML-GLM Estimator

**Input:** observed signal $y$, measure matrix $H^{(l)}$, transition distribution $P(x^{(l+1)}|z^{(l)})$ for $(l = 1, \ldots, L)$ and prior distribution $P(x)$.

**Initiation:** $t = 1$, $Z_a^{(1)}(1) = \prod_{k=1}^L \alpha_k$, $V_a^{(1)}(1) = 1$ for $(l = 1, \ldots, L)$ and $(a = 1 \ldots, N_{\ell+1})$.

**Output:** Estimator $m_n^{(1)}(l)$ for $t = 1, \ldots, T$ do

- for $l = 1, \ldots, L$ do
  - $z_a^{(l)}(t) = \mathbb{E}_{\tilde{z}_a^{(l)}(t)} \left[ \tilde{z}_a^{(l)}(t) \right]$ (5a)
  - $v_a^{(l)}(t) = \text{Var}_{\tilde{z}_a^{(l)}(t)} \left[ \tilde{z}_a^{(l)}(t) \right]$ (5b)
  - $s_a^{(l)}(t) = (z_a^{(l)}(t) - Z_a^{(l)}(t))/V_a^{(l)}(t)$ (5c)
  - $\tau_a^{(l)}(t) = (V_a^{(l)}(t) - \tilde{v}_a^{(l)}(t))/V_a^{(l)}(t)^2$ (5d)
  - $\Sigma_i^{(l)}(t) = \sum_a |H_{ai}|^2 r_a^{(l)}(t)$ (5e)
  - $R_i^{(l)}(t) = m_i^{(l)}(t) + \Sigma_i^{(l)}(t) \sum_a |H_{ai}|^2 s_a^{(l)}(t)$ (5f)

end for $l = 1, \ldots, L$ do

- $m_i^{(l)}(t + 1) = \mathbb{E}_{\xi_i^{(l)}(t+1)} \left[ \xi_i^{(l)}(t+1) \right]$ (6a)
- $\tilde{v}_i(t + 1) = \text{Var}_{\xi_i^{(l)}(t+1)} \left[ \xi_i^{(l)}(t+1) \right]$ (6b)
- $V_a^{(l)}(t + 1) = \sum_i |H_{ai}|^2 \tilde{v}_i^{(l)}(t + 1)$ (6c)
- $Z_a^{(l)}(t + 1) = \sum_i H_{ai} m_i^{(l)}(t + 1) - V_a^{(l)}(t + 1) s_a^{(l)}(t)$ (6d)

end for $t = 1, \ldots, T$

where

$$
\zeta_a(t) \sim P(y_a|z_0)N(z_a|Z_a(t), V_a(t))
$$

$$
\xi_i(t + 1) \sim p_X(x_i)|N(x_i|R_i(t), \Sigma_i(t))|
$$

Also in the single-layer case, if we continue to parameterize the random activation $P(y_a|z_a)$ as Gaussian, say, $N(z_a|y_a, \sigma^2_a)$, the ML-GAMP algorithm will further degenerate to AMP after substituting the results below into $m_i(t + 1), \tilde{v}_i(t + 1), V_a(t + 1), Z_a(t + 1)$.

$$
\Sigma_i(t) = \left( \sum_a \frac{|H_{ai}|^2}{\sigma^2_a + V_a} \right)^{-1}
$$

$$
R_i(t) = m_i(t) + \Sigma_i(t) \sum_a \frac{H_{ai}^2(y_a - Z_a)}{\sigma^2_a + V_a}
$$

**Remark 2:** Comparing to the existing results, the proposed algorithm offers a better tradeoff between estimation performance and computational efficiency. As seen from Fig. 3 in the experiment section, the proposal noticeably outperforms ML-AMP in the early stage of its iterations. ML-AMP was a competing state-of-the-art estimator proposed by Manoel, Krzakala, Mézard, and Zdeborová, and, to the best of our knowledge, it was the first (in the AMP family, at least) to solve the ML-GLM estimation problem. Comparing to the ML-AMP, our proposal here converges in a much faster speed, and thus, it could offer an estimate far more accurate than ML-AMP in scenarios of stringent time budget, e.g., in interactive video communications, where the maximum number of iterations is strictly limited. When it comes to computational complexity, the proposed algorithm resembles ML-AMP in a sense that both are having a complexity dominated by some matrix-vector multiplications of $O(N^{(l+1)}N^{(l)})$ in the worst case, c.f. (5c), (5f), (6c), and (6d). That is, both estimators are as efficient as the celebrated GAMP, whose computational simplicity is a key appeal. Comparing to ML-VAMP, another state-of-the-art estimator developed by Fletcher, Rangan, and Schniter, the proposed algorithm is computationally more efficient, because it does not rely on singular value decomposition (SVD), an expansive operation of $O(N^3)$ complexity, while in contrast, ML-VAMP required to perform SVD (at least once) for each of its weighting matrices.

**Remark 3:** To differentiate our proposal further from the existing ML-AMP, we compare the two in more details and find out that the proposed algorithm relies on the most recently updated messages, while ML-AMP put more faith in some older ones. Fig. 3 provides a more detailed illustration, upon which we make these comments:

- The proposed algorithm updates its messages in a direction-by-direction manner, i.e., first backward, then forward, after that backward again, and repeat it until converged, but the ML-AMP adopted a different layer-by-layer manner that first updates (both directions in) the 1st layer, and then (both directions in) the 2nd, after that the 3rd, and so on.
- To see how they differ in more detail, we consider a particular block of equations, (6a)-(6d). In the proposed algorithm

Existing results as special cases. For instance, by particularizing $L = 1$, it recovers GAMP which reads

$$
\hat{z}_a(t) = \mathbb{E}_{\hat{z}_a(t)} \left[ \hat{z}_a(t) \right]
$$

$$
\hat{v}_a(t) = \text{Var}_{\hat{z}_a(t)} \left[ \hat{z}_a(t) \right]
$$

$$
\hat{s}_a(t) = (\hat{z}_a(t) - Z_a(t))/\hat{V}_a(t)
$$

$$
\hat{\tau}_a(t) = (\hat{V}_a(t) - \hat{v}_a(t))/\hat{V}_a(t)^2
$$

$$
\hat{\Sigma}_i(t) = \sum_a |H_{ai}|^2 \hat{\tau}_a(t)
$$

$$
\hat{R}_i(t) = \hat{m}_i(t) + \hat{\Sigma}_i(t) \sum_a |H_{ai}|^2 \hat{s}_a(t)
$$

$$
\hat{m}_i(t + 1) = \mathbb{E}_{\hat{\xi}_i(t+1)} \left[ \hat{\xi}_i(t+1) \right]
$$

$$
\hat{\tilde{v}}_i(t + 1) = \text{Var}_{\hat{\xi}_i(t+1)} \left[ \hat{\xi}_i(t+1) \right]
$$

$$
\hat{V}_a(t + 1) = \sum_i |H_{ai}|^2 \hat{\tilde{v}}_i(t + 1)
$$

$$
\hat{Z}_a(t + 1) = \sum_i H_{ai} \hat{m}_i(t + 1) - \hat{V}_a(t + 1) \hat{s}_a(t)
$$
The projection was inspired by EP [4]/EC[5], and it uses a Kullback-Leibler (KL) divergence to measure the "distance" between two distribution functions. To be specific, let \( p(x) \) and \( q(x) \) denote two densities, and \( D(q(x)||p(x)) \) denote the KL divergence from \( p(x) \) to \( q(x) \). Then, the moment-matching projection given below could offer an approximate function that is Gaussian with a mean and a variance precisely match that of the target function \( q(x) \), i.e.,

\[
\text{Proj}[q(x)] = \arg \min_{p(x) \in \Phi} D(q(x)||p(x)) = \mathcal{N}(x|m, v),
\]

where \( \Phi \) denotes all possible Gaussian densities, \( m = \int x q(x) \, dx \), and \( v = \int |x - m|^2 q(x) \, dx \).

Given the message updating rules, the backbone of the ML-GAMP can now be established, which reads

\[
\begin{align*}
\mu_{i\rightarrow a}(x_i^{(t)}, t) &\propto \frac{\text{Proj}_a[f_{i\rightarrow a}(x_i^{(t)}, t), f_{i\rightarrow a}(x_i^{(t-1)}, t)]}{\mu_{a\rightarrow i}(x_i^{(t-1)}, t)} \quad (12a) \\
\mu_{i\rightarrow a}(x_i^{(t)}, t) &\propto \prod_b \mu_{i\rightarrow b}(x_i^{(t)}, t) \quad (12b) \\
\mu_{i\rightarrow a}(x_i^{(t)}, t + 1) &\propto \frac{\text{Proj}_a[f_{i\rightarrow a}(x_i^{(t)}, t + 1), f_{i\rightarrow a}(x_i^{(t-1)}, t)]}{\mu_{a\rightarrow i}(x_i^{(t-1)}, t)} \quad (12c) \\
\mu_{i\rightarrow a}(x_i^{(t)}, t + 1) &\propto \mu_{i\rightarrow a}(x_i^{(t)}, t + 1) \prod_{b \neq a} \mu_{i\rightarrow b}(x_i^{(t)}, t) \quad (12d)
\end{align*}
\]

where \( \propto \) denotes "is proportional to \( x^t \), \( \mu_{i\rightarrow a}(x_i^{(t)}, t) \) is the message from the factor node \( \mathcal{P}(x_i^{(t+1)}|x_i^{(t)}) \) to the variable node \( x_i^{(t)} \), and \( \mu_{a\rightarrow i}(x_i^{(t-1)}, t) \) is that in the opposite direction. Similarly, \( \mu_{i\rightarrow a}(x_i^{(t)}, t) \) is the message from \( x_i^{(t)} \) to \( \mathcal{P}(x_i^{(t)}|x_i^{(t-1)}) \), while \( \mu_{i\rightarrow a}(x_i^{(t)}, t + 1) \) is the opposite. The following notations are also used in the backbone iteration:

\[
\begin{align*}
&\int \mathcal{P}(x_i^{(t+1)}|x_i^{(t)}) \mu_{a\rightarrow i}(x_i^{(t+1)}, t) \\
&\prod_{j \neq a} \mu_{j \rightarrow a}(x_j^{(t)}, t) \, dx_j^{(t)} \, dx_a^{(t+1)} \quad (13) \\
&\mu_{a\rightarrow i}(x_a^{(t+1)}, t) \propto \prod_{k=1}^{N_i-2} \mu_{a\rightarrow k}(x_a^{(t+1)}, t) \quad (14) \\
&\int \mathcal{P}(x_i^{(t)}|x_i^{(t-1)}) \prod_{k=1}^{N_i-1} \mu_{k \rightarrow i}(x_k^{(t)}, t + 1) \, dx_i^{(t-1)} \quad (15)
\end{align*}
\]

Then, we combine (12) to get a new compacted set of iteration

\[
\begin{align*}
\mu_{i\rightarrow a}(x_i^{(t)}, t) &\propto \frac{\text{Proj}_a[f_{i\rightarrow a}(x_i^{(t)}, t), f_{i\rightarrow a}(x_i^{(t-1)}, t)]}{\mu_{a\rightarrow i}(x_i^{(t-1)}, t)} \quad (16a) \\
\mu_{i\rightarrow a}(x_i^{(t)}, t + 1) &\propto \frac{\text{Proj}_a[f_{i\rightarrow a}(x_i^{(t)}, t + 1), \prod_{b \neq a} \mu_{i\rightarrow b}(x_i^{(t)}, t)]}{\mu_{a\rightarrow i}(x_i^{(t)}, t)} \quad (16b)
\end{align*}
\]

where the approximate marginal of \( x_i \) at the \( T \)-th iteration is:

\[
\hat{p}(x_i|y) \propto \mu_{i\rightarrow a}(x_i, T) \cdot \mu_{i\rightarrow a}(x_i, T) \quad (17)
\]

As LBP could start up in any part of the network, we begin ours with the last layer, i.e., the rightmost block of the factor graph in Fig. 4. Our equations then iterate like this: the
and apply the Gaussian reproduction property \[^{1}\] to \[^{14}\] to get

\[
\mu_{a\leftarrow k}^{(t)} (x_{a}^{(t+1)}, t) = \mathcal{N}(\mu_{a}^{(t+1)}, \Sigma_{a}^{(t+1)}(t), \tau_{a}^{(t)}(t))
\]

\[
\Sigma_{a}^{(t+1)}(t) = \frac{1}{N_{s} + 2} \sum_{k=1}^{N_{s} + 2} \frac{m_{a\leftarrow k}^{(t+1)}(t)}{v_{b\leftarrow k}^{(t+1)}(t)}
\]

\[
R_{a}^{(t+1)}(t) = \Sigma_{a}^{(t+1)}(t) \cdot \sum_{k=1}^{N_{s} + 2} \frac{m_{a\leftarrow k}^{(t+1)}(t)}{v_{b\leftarrow k}^{(t+1)}(t)}
\]

After that, we start the evaluation of \( f_{i\rightarrow a}^{(t)}(x_{i}^{(t)}, t) \), and an interim result is given in \[^{2}\] at the top of next page. To proceed further, we also notice that although the original LBP treats the \( f_{i\rightarrow a}^{(t)}(x_{i}^{(t)}, t) \) as a likelihood function in \( y \) conditioned on \( x_{i}^{(t)} \), this function in our context could be treated as a p.d.f. in \( x_{i}^{(t)} \) with a difference up to some scaling factor depending on \( y \). To be specific, the function as seen in \[^{13}\] will eventually take a form like \( \alpha(y) \cdot \mathcal{N}(x_{i}^{(t)}, \beta(y), \gamma(y)) \), thanks to the extraordinary property of self-conjugation of the Gaussian distribution, see \[^{1}\] for more information on exponential family, exponential conjugate, and the Gaussian as a unique case in the former two. Treating this as a p.d.f. in \( x_{i}^{(t)} \) substantially facilitated\[^{2}\] the analysis that follows, where the evaluation will be shifted to an r.v.s perspective (from the p.d.f.'s), as many well-known results in high dimensional random theories were given in a r.v. point of view. To this end, we let \( \sim \) stand for “follows the distribution of”, and introduce new r.v.s to associate with all messages (distributions) involved in the evaluation, including: \( \xi_{i\rightarrow a}^{(t)}(t) \sim f_{i\rightarrow a}^{(t)}(x_{i}^{(t)}, t) \), \( \zeta_{a}^{(t)}(t) \sim \int \mathcal{P}(x_{a}^{(t+1)} | z_{a}^{(t)}, \mu_{a}, \Sigma_{a}) m_{a\leftarrow a}^{(t+1)}(x_{a}^{(t+1)}) dx_{a}^{(t+1)} \), and \( \xi_{i\rightarrow a}^{(t)}(t) \sim f_{i\rightarrow a}^{(t)}(x_{i}^{(t)}, t) \). By that, we could express the distribution of a compound vector \( \langle \xi_{i}^{(t)}(t), \xi_{a}^{(t)}(t) \rangle_{i\neq a} \) as below

\[
\mathcal{P}_{\xi_{i}^{(t)}(t), \xi_{a}^{(t)}(t)}(\xi_{a}^{(t)}, \xi_{j}^{(t)})_{j \neq i} \propto \int \mathcal{P}(x_{a}^{(t+1)} | z_{a}^{(t)}, \mu_{a}, \Sigma_{a}) m_{a\leftarrow a}^{(t+1)}(x_{a}^{(t+1)}) \prod_{j \neq i} \mu_{j\rightarrow a}^{(t)}(x_{j}^{(t)}, t)
\]

Comparing this to \[^{13}\], we find, after noticing a generic relation\[^{3}\] between an r.v. and its p.d.f., that the r.v. \( \xi_{i\rightarrow a}^{(t)}(t) \) associated with \( f_{i\rightarrow a}^{(t)}(x_{i}^{(t)}, t) \) could be reexpressed as a function of the compound random vector above, i.e.,

\[
\xi_{i\rightarrow a}^{(t)}(t) = \frac{1}{H_{a}^{(t)}} \left( \mathcal{N}_{a}^{(t)}(t) - \sum_{j \neq i} H_{aj}^{(t)} \mathcal{N}_{aj}^{(t)}(t) \right).
\]

1. Gaussian reproduction property \[^{13}\]: \( \mathcal{N}(x_{a} | \mu_{a}, \Sigma_{a}) \mathcal{N}(x_{b} | \mu_{b}, \Sigma_{b}) = \mathcal{N}(x_{a} | \mu_{a} - b, A + B) \mathcal{N}(x_{c} | C, C) \) with \( C = \left( \frac{1}{A} + \frac{1}{B} \right)^{-1} \) and \( c = C \left( \frac{1}{A} + \frac{1}{B} \right) \).

2. It is possible to evaluate (approximate) \( f_{i\rightarrow a}^{(t)}(x_{i}^{(t)}, t) \) without treating it as a p.d.f. in \( x_{i}^{(t)} \). For instance, in our previous paper \[^{2}\], we adopted a method that was similar to that in the Bi-GAMP case \[^{16}\], which also relied on the central limit theorem but did not require it to be a p.d.f.

3. Given two r.v.s, \( u, v \in \mathbb{C}^{k} \) and \( u \in \mathbb{C} \), where \( \mathbb{w} \sim \mathcal{P}_{\mathbb{w}}(w) \), and for any mapping \( g : \mathbb{C}^{k} \rightarrow \mathbb{C}^{k} \), it holds: \( \mathbb{w} \sim \mathcal{P}_{\mathbb{w}}(u) \) if and only if \( \mathbb{w} \sim \mathcal{P}_{\mathbb{w}}(w) \).
Given two independent r.v.s, $X \sim p_X(x)$ and $Y \sim p_Y(y)$, the p.d.f. of a third one, $Z = X + Y$, is then given by $p_Z(z) = \int p_X(x)p_Y(z-x)dx$. 

As $N_i \to \infty$, the central limit theorem comes into play, thus

$$
\sum_{j \neq i} H_{aj} \xi_j \sim N(Z_{i,a}, V_{i,a})
$$

(25)

$$
Z_{i,a}(t) = \sum_{j \neq i} H_{aj} m_{j,a}(t)
$$

(26)

$$
V_{i,a}(t) = \sum_{j \neq i} \left| H_{aj} \right|^2 v_{j,a}(t)
$$

(27)

with $m_{j,a}(t)$ and $v_{j,a}(t)$ denoting the mean and the variance of the Gaussian message $\mu_{j,a}(x_{i,a}(t))$. Going back to the r.h.s. of (24), since its 2nd term could be approximated by a Gaussian r.v. distributed as (25) and its 1st term was given by $\zeta_a(t) \sim \mathcal{P}(x_{i,a}(t) | m_{i,a}(t), v_{i,a}(t), t)d\alpha_{i,a}$, we could simply convolved the p.d.f.'s of these two and get $\xi_i(t)$'s

$$
f_{i,a}(x_{i,a}(t), t) \propto \mathcal{N}(x_{i,a}(t) | m_{i,a}(t), V_{i,a}(t))
$$

(28)

Step 2: Evaluate $\text{Proj}_\Phi[\mu_{i,a}(u_{i,a}(t), t) \cdot f_{i,a}(u_{i,a}(t), t)]$ in (16a).

$$
\text{Proj}_\Phi \left[ \mu_{i,a}(x_{i,a}(t), t), f_{i,a}(x_{i,a}(t), t) \right] = \text{Proj}_\Phi \left[ \mu_{i,a}(x_{i,a}(t), t), \mathcal{N}(x_{i,a}(t) | m_{i,a}(t), V_{i,a}(t)) \right]
$$

(29)

We first notice that the term $\text{Proj}_\Phi \{ \cdot \}$ is playing the role of a joint p.d.f. of the input element and the observation vector. In the presence of the moment-matching projection, the term inside $\text{Proj}_\Phi \{ \cdot \}$ could be treated as a marginal p.d.f. in $x_i(t)$. In this context, we denote the Gaussian message above as $\mu_{i,a}(x_{i,a}(t), t) \triangleq \mathcal{N}(x_{i,a}(t) | m_{i,a}(t), v_{i,a}(t))$, and use a Gaussian reproduction property to simplify the marginal as

$$
\mu_{i,a}(x_{i,a}(t), t) \cdot \mathcal{N}(x_{i,a}(t) | m_{i,a}(t), V_{i,a}(t))
$$

(30)

where we drop all the indices $(t)$ from the iterating variables in (29) as they could be recovered easily from the context, and

$$
\bar{z}_{i,a}(t) = \left( \frac{1}{v_{i,a}} + \frac{\left| H_{ai} \right|^2}{V_{i,a}} \right)^{-1}
$$

(31)

$$
\bar{m}_{i,a}(t) = \bar{z}_{i,a}(t) \left( m_{i,a}(t) + \frac{(H_{ai})^* (\zeta_a(t) - Z_{i,a})}{V_{i,a}} \right)
$$

(32)

$$
Z_{i,a} = H_{ai} m_{i,a} + Z_{i,a}
$$

(33)

$$
V_{i,a} = |H_{ai}|^2 v_{i,a} + V_{i,a}
$$

(34)

As just mentioned, Eq. (29) could be treated as an p.d.f. in $x_{i,a}$. Since an p.d.f. integrates to unity, we recover the normalizing factor (a.k.a. the partition function) omitted in (30) to provide below a more complete form of (29)

$$
\begin{align*}
\mathcal{E}_{\zeta_a} \left[ \mathcal{N}(\zeta_a(t) | Z_a(t), V_a(t)) \right] & \cdot \mathcal{E}_{\bar{z}_{i,a}} \left[ \mathcal{N}(\bar{z}_{i,a}(t) | Z_{i,a}(t), V_{i,a}(t)) \right] \\
& = \int \mathcal{N}(\zeta_a(t) | Z_a(t), V_a(t)) \mathcal{N}(\bar{z}_{i,a}(t) | Z_{i,a}(t), V_{i,a}(t)) dz_a(t) dz_{i,a}(t)
\end{align*}
$$

(35)

For (35), we still need to know its mean and variance before evaluating the moment-matching projection $\text{Proj}_\Phi \cdot$. This starts by introducing yet another r.v.-to-p.d.f. association

$$
\bar{z}_{i,a}(t) \sim \mathcal{P}(\bar{z}_{i,a}(t) | Z_a(t), V_a(t)) \mathcal{N}(\bar{z}_{i,a}(t) | Z_{i,a}(t), V_{i,a}(t))
$$

(36)

$$
\mathcal{N}(\zeta_a(t) | Z_a(t), V_a(t)) \mathcal{N}(\bar{z}_{i,a}(t) | Z_{i,a}(t), V_{i,a}(t))
$$

(37)

where $(\cdot) = (R_0^{(a+i)}(t), \Sigma_0^{(a+i)}(t), Z_a(t), V_a(t))$, and the mean and variance of $\zeta_a(t)$ are denoted as $\bar{z}_{i,a}(t)$ and $\bar{v}_{i,a}(t)$:

$$
\bar{z}_{i,a}(t) = \mathbb{E}_{\bar{z}_{i,a}(t)}[\zeta_a(t)]
$$

(38)

$$
\bar{v}_{i,a}(t) = \text{Var}_{\bar{z}_{i,a}(t)}[\zeta_a(t)]
$$

(39)

Particularly, if $\ell = L$, the r.h.s. of (36) becomes

$$
\int \mathcal{P}(y_2 | z_{2i})(\mathcal{N}(z_{2i} | Z_i(t), V_i(t)) dz_{2i})
$$

In these notations, we are able to express the mean and the variance of (35) as

$$
m_{i,a} = \int x(t) \int \mathcal{P}(z_{2i} | y_2)(\mathcal{N}(z_{2i} | m_{i,a}(t), v_{i,a}(t)) dz_{2i})
$$

(40)

$$
\begin{align*}
& = \int \mathcal{P}(z_{2i} | y_2)(\mathcal{N}(z_{2i} | m_{i,a}(t), v_{i,a}(t)) dz_{2i})
& = \left( \frac{1}{v_{i,a}} + \frac{\left| H_{ai} \right|^2}{V_{i,a}} \right)^{-1}
\end{align*}
$$

(41)

$$
\begin{align*}
& = \int \mathcal{P}(z_{2i} | y_2)(\mathcal{N}(z_{2i} | m_{i,a}(t), v_{i,a}(t)) dz_{2i})
& = \left( \frac{1}{v_{i,a}} + \frac{\left| H_{ai} \right|^2}{V_{i,a}} \right)^{-1}
\end{align*}
$$

(42)

$$
\begin{align*}
& = \int \mathcal{P}(z_{2i} | y_2)(\mathcal{N}(z_{2i} | m_{i,a}(t), v_{i,a}(t)) dz_{2i})
& = \left( \frac{1}{v_{i,a}} + \frac{\left| H_{ai} \right|^2}{V_{i,a}} \right)^{-1}
\end{align*}
$$

(43)

$$
\begin{align*}
& = \int \mathcal{P}(z_{2i} | y_2)(\mathcal{N}(z_{2i} | m_{i,a}(t), v_{i,a}(t)) dz_{2i})
& = \left( \frac{1}{v_{i,a}} + \frac{\left| H_{ai} \right|^2}{V_{i,a}} \right)^{-1}
\end{align*}
$$

(44)

where all iterating variables have the same iteration index $(t)$. So far, we could express the projection as

$$
\text{Proj}_\Phi[\mu_{i,a}(x_{i,a}(t), t), f_{i,a}(x_{i,a}(t), t)] = \mathcal{N}(x_{i,a}(t) | m_{i,a}(t), v_{i,a}(t))
$$
Step 3: Evaluate \( \mu_{i-a}(x_i^{(t)}+1) \) in (16a).
By Gaussian reproduction property, the mean and variance of \( \mu_{i-a}(x_i^{(t)}+1) \) could be expressed as
\[
v_{i-a}(t+1) = \frac{(V_{a}(t))^2 - v_{i-a(a)}[H_{ai}^{(t)}]_2(V_{a}(t) - \tilde{v}_{a}(t))}{|H_{ai}^{(t)}|^2(V_{a}(t) - \tilde{v}_{a}(t))}
\]
(40)
\[
m_{i-a}(t+1) = \frac{(H_{ai}^{(t)})^*z_a(t) - v_a(t)}{|H_{ai}^{(t)}|^2(v_a(t) - \tilde{v}_{a}(t))}
\]
(41)
where the index \((t)\) is omitted from all iterating variables. Further ignoring the high-order items, we get
\[
v_{i-a}(t) = \frac{1}{|H_{ai}^{(t)}|^2(t)}
\]
(42)
\[
m_{i-a}(t) = \frac{H_{ai}^{(t)}s_a(t) + |H_{ai}^{(t)}|^2\tau_a(t)m_{i-a}(t)}{|H_{ai}^{(t)}|^2\tau_a(t)}
\]
(43)
where the followings are defined
\[
\tau_a(t) \triangleq \frac{V_a(t) - \tilde{v}_a(t)}{(V_a(t))^2(t)}
\]
(44)
\[
s_a(t) \triangleq \frac{z_a(t) - \tilde{z}_a(t)}{v_a(t)}
\]
(45)
Step 4: Do Proj\( [f_{i-1}^{(t)}(x_{i}^{(t)}+1) \prod_b \mu^{(t)}_{i-b}(x_{i}^{(t)}+1)] \) in (16b).
We first evaluate \( f_{i-1}^{(t)}(x_{i}^{(t)}+1) \) in a way similar to (23)-(27),
\[
f_{i-1}^{(t)}(x_{i}^{(t)}+1) \propto \int dz_i^{(-1)} (z_i^{(-1)})^{-1} \sum_{j=1}^{N_{i-1}} N(z_i^{(-1)}|\hat{z}_i^{(-1)}(t+1),V_i^{(-1)}(t+1))
\]
(46)
By Gaussian reproduction property, we further get
\[
\text{Proj} \left[ f_{i-1}^{(t)} \prod_b \mu_{i-b}^{(t)} \right] \propto \text{Proj} \left[ \int N(z_i^{(-1)}|\hat{z}_i^{(-1)}(t+1),\tilde{v}_i^{(-1)}(t+1)) \right]
\]
where the following definitions are used
\[
\Sigma_i^{(t)}(t) = \sum_{a=1}^{N_{i-1}} |H_{ai}^{(t)}|^2 \tau_a^{(t)}(t))^{-1}
\]
(48)
\[
R_i^{(t)}(t) = \sum_{a=1}^{N_{i-1}} (H_{ai}^{(t)}s_a(t) + |H_{ai}^{(t)}|^2\tau_a(t)m_{i-a}(t))
\]
(49)
\[
\tilde{m}_i^{(t)}(t+1) \triangleq \mathbb{E}[\xi_i^{(t)}(t+1)]
\]
(50)
\[
\tilde{v}_i^{(t)}(t+1) \triangleq \text{Var} \xi_i^{(t)}(t+1)
\]
(51)
with the expectation and variance in the last two equalities being taken from this p.d.f. of \( \xi_i^{(t)}(t+1) \)
\[
\mathcal{P}_{\xi_i^{(t+1)}(x_i^{(t)})} = \frac{\int N(z_i^{(-1)}|\hat{z}_i^{(-1)}(t+1))dz_i^{(-1)}}{\int N(z_i^{(-1)}|\hat{z}_i^{(-1)}(t+1))dz_i^{(-1)}}
\]
(52)
In the particular case of \( \ell = 1 \), we initialize for \( x_1^{(1)} := x_1 \) and \( \mathcal{P}_{\xi_1^{(1)}(t+1)(x_1^{(1)})} = \frac{\mathcal{P}_X(x_1^{(1)}|N(x_1^{(1)})R_1^{(1)}_{x_1^{(1)}} s_1^{(1)})}{\int \mathcal{P}_X(x_1^{(1)}|N(x_1^{(1)})R_1^{(1)}_{x_1^{(1)}} s_1^{(1)})dx_1^{(1)}}
\]
Step 5: Evaluate \( \mu_{i-a}(x_i^{(t)}+1) \) in (16b).
Applying the Gaussian reproduction property and then ignoring the high-order infinitesimals, we get
\[
v_{i-a}(t+1) = \left( \frac{1}{v_a(t+1)} - \frac{1}{v_{i-a}(t+1)} \right)^{-1}
\]
(53)
\[
m_{i-a}(t+1) = \frac{m_{i-a}(t+1) - m_{i-a}(t+1)}{v_{i-a}(t+1)}
\]
(54)
Step 6: Evaluate \( (Z_{ai}^{(t)}(t), \tilde{V}_{ai}(t), R_i^{(t)}(t)) \) and gather all.
We substitute (54)-(55) into (33)-(35) to get
\[
V_a^{(t)}(t+1) = \sum_{i=1}^{N} |H_{ai}^{(t)}|^2 v_{i-a}^{(t)}(t+1)
\]
\[
Z_{ai}^{(t)}(t+1) = \sum_{i=1}^{N} |H_{ai}^{(t)}|^2 m_{i-a}^{(t)}(t+1) - s_a^{(t)}(t)V_{a}^{(t)}(t+1)
\]
(56)
Substituting (48) and (55) into (49) further yields
\[
R_i^{(t)}(t) = m_{i-a}^{(t)}(t) + \Sigma_i^{(t)}(t) \sum_{a=1}^{M} |H_{ai}^{(t)}|^2 s_a^{(t)}(t)
\]
So far, we have finished the evaluation of the two messages given by (16a)-(16b). From their results we see that these messages are fully characterized by a set of eight iterating variables, i.e., \((z_a^{(t)}(t), \tilde{v}_a^{(t)}(t), \tilde{m}_a^{(t)}(t), \tilde{V}_a^{(t)}(t), \tilde{m}_a^{(t)}(t+1), \tilde{v}_a^{(t)}(t+1), V_a^{(t)}(t+1), Z_{ai}^{(t)}(t+1))\), each depending on either \(a\) or \(i\). Comparing to each original message in (16a)-(16b) which depends on \(a\) and \(i\), the new iteration could effectively reduce the number of messages per layer from \(O(N_{i+1} \times N_t^2)\) to \(O(N_{i+1} + N_t)\). So, we re-arrange them in a backward-first manner, and finally recover ML-GAMP as Algorithm. 1

C. Convergence of the Algorithm

The original AMP [11] since its proposal by Donoho et al. had been observed to exhibit a fast convergence in the case of large i.i.d. zero-mean Gaussian weighting matrices. However, its convergence under general transform matrices was not fully understood for a long time until recently Rangan et al. [18] provided some sufficient conditions, considering damped GAMP (i.e., an extension to AMP) with quadratic cost functions. It was shown there, with sufficient damping, the GAMP algorithm was guaranteed to converge, although the amount of damping grows with peak-to-average ratio of the squared singular values of the transform matrices. Their result explains not only the good performance of the original AMP on i.i.d. Gaussian transform matrices, but also the difficulties in applying the original AMP to ill-conditioned or non-zero-mean transform matrices.

Inherited from AMP and GAMP, our proposed ML-GAMP also exhibits some fast convergence in large i.i.d. zero-mean Gaussian matrices, as we will see later in the experiment section. This is one important aspect in common, but differences also exist. For example, we do find that as the layer
number increases, the need for a damping quickly ramps up. Due to time and space limitation, here we are not able to provide a thorough investigation into this problem, but below we summarize some rules observed in our earlier experience with the algorithm:

- Among them, these two variables could be damped first, i.e., $Z_a^{(t)}$ in (6c) and $V_a^{(t)}$ in (6d), and their damping reads

$$Z_a^{(t)}(t + 1) = \rho^{(t)}[\sum_i H_{ai}^{(t)} \tilde{v}_i^{(t)}(t + 1) - V_a^{(t)}(t + 1)]s_a^{(t)}(t) + (1 - \rho^{(t)})Z_a^{(t)}(t)$$

$$V_a^{(t)}(t + 1) = \rho^{(t)}[\sum_i |H_{ai}^{(t)}|^2 \tilde{v}_i^{(t)}(t + 1)] + (1 - \rho^{(t)})V_a^{(t)}(t)$$

- The variable $\rho^{(t)}$ is a damping factor and could be chosen freely from (0, 1] with $\rho^{(t)} = 1$ meaning no damping at all. From our early experience with the algorithm, we find it better to limit the factor within [0.7, 1].

- If more damping is desired, two more variables could be considered, i.e., $\tilde{m}_k^{(t)}(t + 1)$ and $\tilde{z}_i^{(t)}(t)$, but be aware that excessive damping certainly slows down the convergence rate.

D. Simplification to the Algorithm

To further ease ML-GAMP’s burden in computation, we provide in Algorithm 2 a simplified version, scalar-variance ML-GAMP, where the original variance vectors $\{V_a^{(t)}\}_{a=1,2,\ldots}$ and $\{\Sigma_i^{(t)}\}_{i=1,2,\ldots}$ are averaged to be scalars (i.e., $V_i^{(t)}$ and $\Sigma_i^{(t)}$). The averaging is indeed inspired and supported by a self-adjusting property that had long been observed in high dimensional studies and was also applied to the design of S-AMP [19] and simplified GAMP [6].

To derive the scalar-variance ML-GAMP, we first replace the two variances $\tilde{v}_i^{(t)}(t)$ and $\tilde{v}_a^{(t)}(t)$ with their own averages:

$$\tilde{v}_i^{(t)}(t) \leftarrow \frac{1}{N_t} \sum_{i=1}^{N_t} \tilde{v}_i^{(t)}(t)$$

$$\tilde{v}_a^{(t)}(t) \leftarrow \frac{1}{N_t} \sum_{a=1}^{N_t} \tilde{v}_a^{(t)}(t)$$

$$V_a^{(t)}(t + 1) \leftarrow V_i^{(t)}(t + 1) \leftarrow \frac{1}{\alpha_t} \tilde{v}(t + 1),$$

where (59) is due to $V_a^{(t)} \approx |H_{ai}^{(t)}|^2 \sum_i \tilde{v}_i^{(t)} \approx \frac{1}{N_t} \sum_i \tilde{v}_i^{(t)}$. Substituting the above back into (5c) and (5e) further yields

$$s_a^{(t)}(t) \leftarrow \frac{\tilde{z}_a^{(t)}(t) - Z_a^{(t)}(t)}{V_i^{(t)}}$$

$$\Sigma_i^{(t)}(t) \leftarrow \frac{(V_i^{(t)})^2}{V_i^{(t)} - \tilde{v}^{(t)}(t)}$$

After that, we collect all results and come up with

$$R_i^{(t)}(t) = \tilde{m}_i^{(t)}(t) + \Sigma_i^{(t)}(t) \sum_a (H_{ai}^{(t)})^2 s_a^{(t)}(t)$$

$$Z_a^{(t)}(t + 1) = \sum_i H_{ai}^{(t)} \tilde{m}_i^{(t)}(t + 1) - V_a^{(t)}(t + 1) s_a^{(t)}(t)$$

which, after a rearranging in matrix form, recovers Algo. 2.

Algorithm 2: A Simplified ML-GAMP (Scalar-Variance)

```
for $t = 1, \ldots, T$ do
    for $\ell = 1, \ldots, L$ do
        $\tilde{z}_a^{(t)}(t) = \mathbb{E}_{\tilde{z}_a^{(t)}} \left[ \frac{\tilde{z}_a^{(t)}(t)}{V_i^{(t)}} \right]$
        $\tilde{v}^{(t)}(t) \leftarrow 1 - N_t \sum_{a=1}^{N_t} \text{Var}_{\tilde{z}_a^{(t)}} \left[ \frac{\tilde{z}_a^{(t)}(t)}{V_i^{(t)}} \right]$
        $\tilde{v}^{(t)}(t) \leftarrow 1 - N_t \sum_{a=1}^{N_t} \text{Var}_{\tilde{z}_a^{(t)}} \left[ \frac{\tilde{z}_a^{(t)}(t)}{V_i^{(t)}} \right]$
        $s_a^{(t)}(t) = \frac{\tilde{z}_a^{(t)}(t) - Z_a^{(t)}(t)}{V_i^{(t)}}$
        $\Sigma_i^{(t)}(t) = \frac{(V_i^{(t)})^2}{V_i^{(t)} - \tilde{v}^{(t)}(t)}$
        $R_i^{(t)}(t) = \tilde{m}_i^{(t)}(t) + \Sigma_i^{(t)}(t) \sum_a (H_{ai}^{(t)})^2 s_a^{(t)}(t)$
    end
end
```

III. ML-GAMP: ASYMPTOTIC ANALYSIS AND SE

Inherited from AMP [11], variants in this family usually exhibit a good property that their asymptotic behaviors could be accurately characterized (predicted) through a set of simple one-dimensional iterating equations called state evolution (SE). In the literature, Bayati et al. [20] rigorously proved the SE of the original AMP, and later Rangan [6] offered its GAMP extension. More recently, Manoel et al. presented the SE for their ML-GAMP algorithm [8]. Since an SE relies heavily on the implementation of an algorithm (it would capture the per iteration performance of the algorithm), we particularly derive the SE for our ML-GAMP. The derivation, similar to those given in [6] and [8], are heuristic by nature, while a proof rigorous in the mathematical sense is still pending. But fortunately, the SE derived could capture precisely the per iteration MSE behavior of the ML-GAMP algorithm, as one will see in Fig. 8. This section is then divided into two parts: the first part presents the SE result, while the second part elaborates all derivation details.

A. Results of the Analysis

The SE of the ML-GAMP estimator is given in Algorithm 3 under the assumption of a Gaussian weighting matrix $H^{(t)}$ for all $\ell$, whose elements are i.i.d., zero-mean, and $O(\frac{1}{N_t})$-variance (to ensure column unity).
For this SE in Algorithm 3, we have the following remarks:

Remark 1: The this SE of ML-GAMP degenerates to many results reported earlier in the literature. For example, at $L = 1$, it reduces to the SE of GAMP in Algorithm 3 as given in [21]:

$$\tau = \frac{\sigma^2_X}{\alpha} - \int \frac{z P(y | x) N(z) \sqrt{\frac{\sigma^2_X - z(\Sigma)}{\alpha} \xi, \frac{z(\Sigma)}{\alpha} \xi} dz^2}{\int P(y | x) N(z) \sqrt{\frac{\sigma^2_X - z(\Sigma)}{\alpha} \xi, \frac{z(\Sigma)}{\alpha} \xi} dz} \Xi dy$$

$$\Sigma = (\frac{z(\Sigma)}{\alpha} \Xi, \frac{z(\Sigma)}{\alpha} \Xi)$$

where $\Xi$ is the MSE of a SISO MMSE estimation problem: $Y = X + W$, with $W \sim N(0, \Sigma)$. Also at $L = 1$, but consider a particular case of SLM, i.e., $P(y | z) = N(y | z, \sigma^2_w)$ with $\sigma^2_w$ being a noise power, the above SE further degenerates to the following SE of AMP that was rigorously proved (via a different conditioning technique) in [20]:

$$\Sigma = \sigma^2_w + \varepsilon(\Sigma)/\alpha.$$  (63)

We also note that the fixed points of our SE agrees perfectly with those of ML-AMP’s SE [8], and this has a multi-folding meaning. Firstly, a fixed point means that the algorithm has converged and that its averaged MSE has hit the floor. Since the two SE’s are sharing the same fixed points, their error floor must be the same, which is indeed confirmed later by Fig. 8. Secondly, while fixed points define their commonality in the error floor, the differences in the (arrangement of iterating variables of) the two SE’s would sometimes lead to a dramatic change in the dynamical behaviors. This is also evidenced by our later Fig. 8 where, before converged, i.e., before hitting their common error floor, the two algorithms behave dramatically different.

Remark 3: Comparing the above result to Algorithm I of Part I, we find a perfect agreement in fixed points between the SE of ML-GAMP and the replica equations of an exact MMSE estimator. To see this agreement, we rewrite $V^{(t)}$ and $\Sigma^{(t)}$ as

$$V^{(t)} = \frac{X^{(t)} - d^{(t)}}{\alpha t}, \quad \Sigma^{(t)} = \frac{X^{(t)} - d^{(t)}}{2\alpha t}$$

where (a) is due to an earlier equality in the algorithm, i.e., $T^{(t)} = T^{(t)} / \alpha t$, while (b) holds as a result of the following equality, from Part I’s Algorithm 1: $d^{(t)} = \frac{\alpha t (X^{(t)} - d^{(t)})}{2T^{(t)} - d^{(t)}}$.

Incorporating (64) and (65) into the SE given in Algorithm 3 one recovers exactly the fixed point equations of the replica analysis on an exact MMSE estimator, given as Algorithm 1 in Part I. Since an exact MMSE estimator is Bayes-optimal, such an agreement indicates that the proposed ML-GAMP (if converged) is asymptotically optimal in the MSE sense, and that unlike the exact MMSE estimator, its computational complexity is similar to the GAMP [7], and thus affordable.

Remark 4: Since the SE could capture the MSE behavior of an algorithm, one could obtain ML-GAMP’s MSE of by simply evaluating its SE. In other words, there is no further need to carry out the time-consuming Monte Carlo simulations which used to model the entire process of data generation, signal transmission, and data estimation. Moreover, given an average MSE from the SE, the symbol error rate (SER) and many other performance indices could be obtained via closed form transformations. For example, in the case of a QPSK-distributed $x$, the SER could be expressed analytically as [22] p. 269: $SER = 2Q(\sqrt{MSE}) - [Q(\sqrt{MSE})]^2$, where
\[ Q(x) = \int_{\mathbb{R}} p^+ dx \] is the Q-function.

### B. Derivation Details

We dedicate this subsection to the derivation of SE. This SE, as an application to the average MSE behavior, is obtained by averaging out all randomness from the algorithm's vari-

\[ \mathbb{E} \left[ \phi(x) - \phi(y) \right] \leq C \left\| x \right\|^k + \left\| y \right\|^k \]

To solve the equation, we notice for its r.h.s. that in large system limits, where all weighting matrices are drawn from some i.i.d. zero-mean populations with \( N \rightarrow \infty \) and \( N_{\ell+1} + N_\ell \rightarrow \alpha_n \), the r.v. \( z^\ell \) tends to be Gaussian distributed with mean zero and variance \( T^\ell \), according to the central limit theorem. Hence, we have \( \mathbb{P}(z^\ell) = z^\ell \), where

\[ T^\ell = \sum_{k=1}^{N_{\ell}} \mathbb{E} \left[ h_{\ell} x^\ell_i \right]^2 = T^\ell_i/\alpha_\ell \]

In what follows, we evaluate the scalars above in two different steps and get the SE of desire.

**Step 1:** Evaluate \( \overline{\mathbb{E}}[\mathbb{G}] \) in (66) for backward passing (\( \ell < L \)).

\[ \overline{\mathbb{E}}[\mathbb{G}] = \sum_{k=1}^{N_{\ell}} \mathbb{E} \left[ \hat{h}_\ell^\ell \right]^2 \]

and the variance of the posterior distribution

\[ \hat{z}^\ell \]

where the expectation is taken over \( \mathbb{P}(z^\ell+1) \), while \( \hat{z}^\ell \) is the variance of the posterior distribution

\[ \hat{z}^\ell \]

in which \( \mathbb{N}(z^\ell|a, b, \Sigma) \) is defined as

\[ \mathbb{N}(z^\ell|a, b, \Sigma) = \mathbb{P}(z^\ell|a, b) N(z^\ell|a, b, \Sigma) \]

as we mentioned before. Interestingly, this posterior could be interpreted as a one-step Markov chain depicted in Fig 6.

![Markov Chain](image)

**Step 2:** Evaluate \( \overline{\mathbb{E}}[\mathbb{G}] \) in (66) for forward passing (\( \ell > 1 \)).

\[ \overline{\mathbb{E}}[\mathbb{G}] = \sum_{k=1}^{N_{\ell}} \mathbb{E} \left[ \hat{h}_\ell^\ell \right]^2 \]

where \( \mathbb{P}(z^\ell+1) \) is the probability distribution of \( z^\ell+1 \), while \( \hat{z}^\ell \) is the variance of the posterior distribution

\[ \hat{z}^\ell \]

in which \( \mathbb{N}(z^\ell|a, b, \Sigma) \) is defined as

\[ \mathbb{N}(z^\ell|a, b, \Sigma) = \mathbb{P}(z^\ell|a, b) N(z^\ell|a, b, \Sigma) \]

as we mentioned before. Interestingly, this posterior could be interpreted as a one-step Markov chain depicted in Fig 6.

![Markov Chain](image)
Markov Chain
\[ 1 < l < L, \quad \ell = 1 \]

Fig. 7. Markov chains for the forward passing: \( Z^{(l-1)} \rightarrow Z^{(l-1)} \rightarrow x^{(l)} \rightarrow R^{(l)} \) and \( x \rightarrow R^{(1)} \)

| Example | Input | Layers | Each Layer’s Model | Weighting |
|---------|-------|--------|-------------------|-----------|
| Ex. #1  | QPSK  | 2      | AWGN, AWGN+ADC    | \( N_x = N_x - 1 \) |
| Ex. #2  | QPSK  | 2~4    | AWGN, AWGN, ...   | \( N_x = 2N_x - 1 \) |
| Ex. #3  | QPSK  | 2      | AWGN, AWGN+ADC    | \( N_x = 2N_x - 1 \) |

where \( \hat{y}^{(l)} \) is the variance of an approximate posterior
\[
\hat{y}^{(l)} = \frac{1}{T^x_{(l)}} - d^{(l)}
\]

which could also be interpreted via a Markov chain as shown in Fig. 7. We then evaluate (78) in a way similar to the last step and get
\[
\hat{y}^{(1)} = T^x_{(1)} - d^{(1)} = \varepsilon(\Sigma^{(1)})
\]

and if \( l = 1 \), we further have
\[
\hat{y}^{(1)} = T^x_{(1)} - d^{(1)} = \varepsilon(\Sigma^{(1)})
\]

IV. NUMERICAL EXAMPLES

In this section, we verify the performance of the ML-GAMP (Algorithm 1) by the accuracy of the SE characterization (Algorithm 2), and the agreement with the replica fixed point equations (Claim 1 in Part I). A comparison to the existing ML-GAMP (8) is also provided, confirming our findings on the convergence issue.

A. Parameter Setup

Table I summarizes the settings of three examples considered here. An analog-to-digital converter (ADC) of uniform quantization is adopted to exemplify our applicability to non-linear activation, and it is appended to an AWGN channel, resulting in a signal model as below (for a certain layer):

\[
\begin{align*}
    z &= Hx \\
    \tilde{y} &= z + n \\
    y &= Q_e(\tilde{y})
\end{align*}
\]

where \( Q_e(\cdot) \) is the complex-valued quantization that reads
\[
y_m = Q_e(\tilde{y}_m) = Q_e(\Re(\tilde{y}_m)) + jQ_e(\Im(\tilde{y}_m)), \quad m = 1, \ldots, M
\]

with \( Q_e(\cdot) : \Re \rightarrow \Re_B, \) and \( \Re_B \) being the set of \( b \)-bits defined as: \( \Re_B \triangleq \{(-1/2 + b) \triangle; b = -2^{\beta} + 1, \ldots, 2^{\beta} - 1\} \), where \( \triangle \) denotes the uniform quantization step. For uniform quantization, the output of ADC is assigned a value \( y \) when the input \( \tilde{y} \) falls within the range of \( (q^{\text{low}}(y), q^{\text{up}}(y)) \), where
\[
\begin{align*}
    q^{\text{low}}(y_m) &= \left\{ y_m - \frac{\triangle}{2}, \quad \text{if } y_m \geq -\left(\frac{2^\beta}{2} - 1\right) \triangle, \right. \\
    &\quad \left. \quad \text{otherwise}, \right. \\
    q^{\text{up}}(y_m) &= \left\{ y_m + \frac{\triangle}{2}, \quad \text{if } y_m \leq \left(\frac{2^\beta}{2} - 1\right) \triangle, \right. \\
    &\quad \left. \quad \text{otherwise}. \right.
\end{align*}
\]

Adding ADC, the overall transitional probability from a complex \( z \) to a complex \( y \) then becomes
\[
P(y|z) = \Psi(\Re(y)|\Re(z), \sigma^2_1/2) \cdot \Psi(\Im(y)|\Im(z), \sigma^2_2/2) \cdot \mathbf{1}_{y \in \Re_B + j\Re_B}
\]

where \( \Psi(\cdot|\cdot, 
\cdot) = \Phi(\frac{x-\mu}{\sigma}) \) and \( \Phi(x) = \int_{-\infty}^{x} \mathcal{N}(t|0,1)dt \).

B. Experimental Results

In Example #1, we consider a 2-layer GLM case, where the first layer is AWGN, and the second layer is AWGN+ADC. The sizes of the two transform matrices are uniformly set to be 1024, and the signal-to-noise ratios (SNRs) of the two layers are 20dB and 15dB, respectively. We vary the ADC resolution, from 1-bit, 2-bit, 3-bit, and 6-bit to infinity, to verify the agreements of the proposed algorithm’s simulated MSEs with its SE. For the AWGN+ADC case existing in second layer, we have following notes. By MSE, we refer to the normalized version that is obtained via \( \|x - \hat{x}\|^2_2/\|x\|^2_2 \). Then we conclude from Fig 8:

- The SE results (obtained via numerical evaluation) agree perfectly with the algorithm results (obtained through Monte Carlo simulation).

A major difficulty in particularizing the general SE into a specific AWGN+ADC case is the evaluation of \( \hat{v}_{(2)} \), which involves non-linear activation. To handle this, we take measures similar to [23, 24] and obtain
\[
\hat{v}_{(2)} = \hat{v}_{(2)} - \frac{(\hat{v}_{(2)}^2 + \hat{v}_{(2)})^2}{2} \sum_{\Re_B} \frac{\hat{v}_{(1)}(\eta_1(\eta_1)) + \hat{v}_{(2)}(\eta_2(\eta_2))}{2} \xi
\]

where
\[
\begin{align*}
    \eta_1(y) &= q^{\text{up}}(y) - \sqrt{\frac{\hat{v}_{(2)}^2 + \hat{v}_{(2)}^2}{2}}, \quad \eta_2(y) = q^{\text{low}}(y) - \sqrt{\frac{\hat{v}_{(2)}^2 + \hat{v}_{(2)}^2}{2}}
\end{align*}
\]

In addition, the variable \( \hat{v}_{(1)} \) depends on the prior. In case of a QPSK prior, \( \hat{v}_{(1)} = 1 - \tanh((\Sigma^{(1)})^{-1} + \sqrt{(\Sigma^{(1)})^{-1} - 1}\xi) \). Other variables like \( \hat{v}_{(1)} \) and \( \hat{v}_{(2)} \) could be evaluated readily via Gaussian reproduction property [15].
Carlo simulations), confirming that the SE is capable of capturing the dynamics of the ML-GAMP algorithm, and also that the fixed points predicted by the replica analysis for a generic MMSE estimator could indeed be realized by an ML-GAMP estimator.

- We required, during the theoretical analysis, the system dimension should be large enough (usually tends to infinity) so that theories like central limit theorem and large deviation theory would be true. In practice, a $1024 \times 1024$ matrix size seems good enough already. Besides, the ML-GAMP algorithm admit a fast converge speed, i.e., in all cases we have considered, it converged with less 15 iterations.

In Example #2, we consider 3 different settings of ML-GLM, i.e., $L = 2$, $L = 3$, and $L = 4$, all having AWGN in each layer. The SNR of each layer is fixed at 10 dB, and the matrix dimension is in a doubling manner, i.e., $N_{\ell+1} = 2N_{\ell}$ for $\ell = 1, \ldots, L$ with $N_1 = 256$. In this setting, we compare the MSE performance of the proposed ML-GAMP to that from ML-AMP [8]. This result is given as Fig 9 suggesting:

- Both ML-GAMP and ML-AMP converge to the same MSE, which (recall from last example’s result) is indeed the optimal value a generic MMSE estimator would attain.
- The difference is our ML-GAMP converges in a speed faster than the existing ML-AMP. As the number of layers increases, the superior of our algorithm becomes dominant. As we analyzed earlier, this superiority is owing to a more recent update of messages in our algorithm, and the superiority would accumulate with the increase of layer number.

In Example #3, we consider a two-layer GLM, where the 1st layer is a $256 \times 512$ AWGN setting while the 2nd is a $512 \times 1024$ AWGN+ADC setting. We vary the ADC resolution from 1-bit, 2bit, 3-bit, to $\infty$-bit to see the impact of uniform quantization. From Then, we find from Fig 10:

- Increasing the resolution of the ADC, however, only provides a diminishing marginal benefit. In an extreme case as we illustrate, after $B \geq 5$, the benefit become so trivial that is not even noticeable, and this is because its BER has already hit the lower bound, i.e. AWGN without ADC.

V. CONCLUSIONS

As Part II of the two-part work, this paper proposed a new estimator, the ML-GAMP, to work around the implementation difficulty of an exact MMSE estimator in high-dimensional ML-GLM. The new estimator was derived using a method adopted in our previous paper [2], which blended an moment-matching projection into the LBP framework. Comparing to existing results, the new estimator provided a better tradeoff between efficiency and complexity. On one hand, it converged
in a speed much faster than the ML-AMP [8]. while on the other hand, it avoided the use of an expansive SVD operation that was required by the ML-VAMP [12]. Further analysis in the new estimator also revealed that its asymptotic MSE behavior could be fully characterized by a set of iterating equations in scalar form, called SE. This SE shared exactly the same fixed points as the exact MMSE estimator whose fixed points were obtained in Part I via replica analysis. Since the exact MMSE estimator is Bayes-optimal in an MSE sense, such an coincidence in the fixed points suggested that the proposed estimator (if converged) could offer an estimate that is MSE optimal.

VI. ACKNOWLEDGEMENT

This work was supported in part by the Scientific Research Fund of Guangdong 201904010297, Natural Scientific Fund of Guangdong 2016A030313705, Special Fund for Applied Science and Technology of Guangdong 2015B011029001, and Natural Scientific Fund of Guangxi 2018GXNSFDA281013.

REFERENCES

[1] C. M. Bishop, Pattern recognition and machine learning. springer, 2006.
[2] Q. Zou, H. Zhang, C.-K. Wen, S. Jin, and R. Yu, “Concise derivation for generalized approximate message passing using expectation propagation,” IEEE Signal Processing Letters, vol. 25, no. 12, pp. 1835–1839, 2018.
[3] K. Murphy, Y. Weiss, and M. I. Jordan, “Loopy belief propagation for approximate inference: An empirical study,” arXiv preprint arXiv:1301.6725 2013.
[4] T. P. Minka, “A family of algorithms for approximate bayesian inference,” Ph.D. dissertation, Massachusetts Institute of Technology, 2001.
[5] M. Opper and O. Winther, “Expectation consistent approximate inference,” Journal of Machine Learning Research, vol. 6, no. Dec, pp. 2177–2204, 2005.
[6] S. Rangan, “Generalized approximate message passing for estimation with random linear mixing,” CoRR, vol. abs/1010.5141, 2010. [Online]. Available: http://arxiv.org/abs/1010.5141
[7] ——, “Generalized approximate message passing for estimation with random linear mixing,” in 2011 IEEE International Symposium on Information Theory Proceedings. IEEE, 2011, pp. 2168–2172.
[8] A. Manoel, F. Krzakala, M. Mézard, and L. Zdeborová, “Multi-layer generalized linear estimation,” in 2017 IEEE International Symposium on Information Theory (ISIT). IEEE, 2017, pp. 2098–2102.
[9] A. K. Fletcher, S. Rangan, and P. Schniter, “Inference in deep networks in high dimensions,” in 2018 IEEE International Symposium on Information Theory (ISIT). IEEE, 2018, pp. 1884–1888.
[10] S. Rangan, P. Schniter, and A. K. Fletcher, “Vector approximate message passing,” IEEE Transactions on Information Theory, 2019.
[11] D. L. Donoho, A. Maleki, and A. Montanari, “Message-passing algorithms for compressed sensing,” Proceedings of the National Academy of Sciences, vol. 106, no. 45, pp. 18914–18919, 2009.
[12] A. K. Fletcher and S. Rangan, “Inference in deep networks in high dimensions,” CoRR, vol. abs/1706.06549, 2017. [Online]. Available: http://arxiv.org/abs/1706.06549
[13] K. P. Murphy, Y. Weiss, and M. I. Jordan, “Loopy belief propagation for approximate inference: An empirical study,” in Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence. Morgan Kaufmann Publishers Inc., 1999, pp. 467–475.
[14] D. L. Donoho, A. Maleki, and A. Montanari, “Message passing algorithms for compressed sensing: I. motivation and construction,” in 2010 IEEE information theory workshop on information theory (ITW 2010, Cairo). IEEE, 2010, pp. 1–5.
[15] C. E. Rasmussen and C. K. Williams, Gaussian process for machine learning. MIT press, 2006.
[16] J. T. Parker, P. Schniter, and V. Cevher, “Bilinear generalized approximate message passing–part i: Derivation,” IEEE Transactions on Signal Processing, vol. 62, no. 22, pp. 5839–5853, 2014.
[17] H. D. Pfister. (2014) Compressed sensing supplemental material for graphical models and inference. Internet draft.
[18] S. Rangan, P. Schniter, A. K. Fletcher, and S. Sarkar, “On the convergence of approximate message passing with arbitrary matrices,” IEEE Transactions on Information Theory, vol. 65, no. 9, pp. 5339–5351, Sep. 2019.
[19] B. Cakmak, O. Winther, and B. H. Fleury, “S-amp: Approximate message passing for general matrix ensembles,” in 2014 IEEE Information Theory Workshop (ITW 2014). IEEE, 2014, pp. 192–196.
[20] M. Bayati and A. Montanari, “The dynamics of message passing on dense graphs, with applications to compressed sensing,” IEEE Transactions on Information Theory, vol. 57, no. 2, pp. 764–785, 2011.
[21] S. Rangan, “Estimation with random linear mixing, belief propagation and compressed sensing,” in 2010 44th Annual Conference on Information Sciences and Systems (CISS). IEEE, 2010, pp. 1–6.
[22] J. G. Proakis, Digital communications, 2001.
[23] C.-K. Wen, C.-J. Wang, S. Jin, K.-K. Wong, and P. Ting, “Bayes-optimal joint channel-and-data estimation for massive mimo with low-precision adcs,” IEEE Journal on Selected Areas in Communications, vol. 35, no. 9, pp. 1962–1979, 2017.