Learning Stabilization in Deep Unfolding of Generalized Approximate Message Passing

Tomoharu Furudoi *, Takumi Takahashi *, Shinsuke Ibi †, and Hideki Ochiai *
* Graduate School of Engineering, Osaka University, 2-1 Yamada-oka, Suita, 565-0871, Japan
† Faculty of Science and Engineering, Doshisha University, 1–3 Tataramiyakodani, Kyotanabe, 610–0394, Japan
Email: *{furutomo@wcs., takahashi@, ochiai@}comm.eng.osaka-u.ac.jp, †sibi@mail.doshisha.ac.jp

***Abstract*—Generalized approximate message passing (GAMP) achieves near-optimal performance in detecting spatially multiplexed massive multiple-input multiple-output (MIMO) signals with significantly reduced computational complexity under independent and identically distributed (i.i.d.) measurements. However, in spatially correlated MIMO channels where the ideal assumption of large-scale uncorrelated observation does not hold, the detection capability severely deteriorates. This performance degradation can be compensated for by optimizing GAMP with embedded learnable parameters via deep unfolding (DU) techniques, *i.e.*, *data-driven tuning*; however, the learning process becomes quite unstable. To address this issue, we propose a novel method that achieves stable learning by incorporating a monotonic increase constraint on the reliability of propagated messages by learning the differential (incremental) values of the learnable parameters between two consecutive iterations. The efficacy of the proposed method is confirmed through numerical results in terms of loss trajectory in the learning process and bit error rate (BER) of massive MIMO detection.**

***Index Terms*—Generalized approximate message passing, deep unfolding, MIMO signal detection, data-driven tuning**

I. Introduction

A massive multi-user multiple-input multiple-output (MU-MIMO) system, where a base station (BS) equipped with a large number of antenna elements simultaneously serves multiple user equipments (UEs), is one of the promising technologies in the forthcoming wireless communication systems. The spatial degrees of freedom (DoF) obtained by the multiple antennas enable enhanced detection reliability and improved spectral efficiency. In the uplink scenario, spatially multiplexed signals need to be separated at the BS receiver with low computational complexity and high accuracy [1].

As signal detection schemes reasonably balancing these two demands, various message passing algorithms (MPAs) based on belief propagation (BP) have been proposed [2]–[8]. Among them, Gaussian belief propagation (GaBP) [2], [3] and generalized approximate message passing (GAMP) [4], [5] have been proven to asymptotically achieve the Bayes-optimal performance in the large-system limit[1] with highly affordable computational complexity of $\mathcal{O}(MN)$ per iteration, provided that the measurement matrix is composed of independent and identically distributed (i.i.d.) Gaussian random variables with mean zero [9]. However, unlike expectation propagation (EP) [8] and vector AMP (VAMP) [7], which trade computational cost for robustness in detection capability by allowing matrix inversion or singular value decomposition (SVD) operation, its performance deteriorates significantly when the statistical behavior of the measurement matrix deviates from this ideal condition [10], [11].

To mitigate this performance degradation, several heuristic methodologies have been investigated, such as *belief damping* [12] and *belief scaling* [3]. The former enables to suppress the undesirable vibration behaviors in the propagated messages *i.e.*, *beliefs*, by updating the beliefs via weighted average, and the latter to prevent the generation of harmful belief outliers by adaptively regulating the shape of the *denoiser*. However, the analytical optimization of their parameters is hard due to the non-linearity of their algorithmic structures.

Recently, a novel deep learning (DL)-aided approach, called deep unfolding (DU) [13], [14], has attracted a great deal of attention in the physical layer signal processing of wireless communications [15], [16]. DU regards the *signal-flow graph* obtained by unrolling the iterative algorithm as a feed-forward neural network (FFNN), and enables the learning optimization of the embedded parameters via the standard learning techniques, such as backpropagation and stochastic gradient descent (SGD); this process is also referred to as *data-driven tuning*. In [17], [18], it was shown that trainable GaBP (T-GaBP), where the embedded damping and scaling parameters are learned via *data-driven tuning*, dramatically improves signal detection performance under correlated measurements compared to the conventional GaBP. However, since GaBP requires the calculation of MN beliefs, the computational cost is higher than that of GAMP, which only operates with M beliefs, and the number of learnable parameters also increases accordingly; thus, the learning cost of T-GaBP is also higher. In learned AMP (LAMP) [19], which applies DU to GAMP, it is necessary to learn the weight matrix, and problems such as high load during the learning process and poor generalization performance have not yet been circumvented. In addition, if one attempts to train GAMP with fewer learnable parameters, it becomes difficult to maintain consistency with the *sophisticated* update rule derived based on the ideal assumptions described above. In many cases, this leads to significant instability in the learning process.

In light of the above, in this paper, we propose a novel strategy for stabilizing the learning process in optimizing the few parameters (damping and scaling parameters) embedded in GAMP via data-driven tuning to achieve reliable signal detection under highly correlated massive MIMO channels.

[1]The idealized system assumption, where the input and output dimensions, M and N, respectively, are infinity for a given compression rate $\delta \triangleq N/M$.

Specifically, to impose a monotonically increasing constraint on changes in scaling parameters for iterations, the non-negative incremental value of the parameter at each iteration step is set as a learnable parameter, rather than the scaling parameter itself. This is equivalent to incorporating the operating principle of MPAs as additional *domain* knowledge that the reliability of the beliefs should be gradually improved as iterations proceed [3], [20].

Finally, the efficacy of the proposed method is numerically confirmed in terms of loss trajectory in the training process, dynamics of the learned parameters, and bit error rate (BER) in massive MU-MIMO signal detection.

Notation: Vectors and matrices are denoted in lower- and upper-case bold-face fonts. Sets of non-negative real, real, and complex-valued numbers are denoted by $\mathbb{R}_+$, $\mathbb{R}$, and $\mathbb{C}$, respectively. The conjugate and transpose operators are denoted by $(\cdot)^*$ and $(\cdot)^{\mathsf{T}}$, respectively. The real and imaginary parts of a complex quantity are denoted by $\Re[\cdot]$ and $\Im[\cdot]$. In addition, $\mathrm{j} \triangleq \sqrt{-1}$ represents the imaginary unit. The $K \times K$ identity matrix is denoted by $\boldsymbol{I}_K$. The (i,j)-th element of the matrix $\boldsymbol{A}$ is denoted by $[\boldsymbol{A}]_{i,j}$. The complex Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Lambda}$ is denoted by $\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Lambda})$. The notation $a \sim \mathcal{P}$ indicates a random variable a follows a probability distribution $\mathcal{P}$.

II. Preliminaries

A. Signal Model

Consider an uplink MU-MIMO system consisting of M UE devices and one BS, where each UE device is equipped with a single transmit (TX) antenna while the BS is equipped with N receive (RX) antennas in a uniform linear array (ULA) pattern. Each UE device chooses a TX symbol independently and uniformly from the quadrature amplitude modulation (QAM) constellation points $\mathcal{X} = \{\chi_1, \chi_2, \cdots, \chi_Q\}$ of the average energy E_s, where Q denotes the modulation order. The RX vector $\boldsymbol{y} \in \mathbb{C}^{N\times 1}$ can be expressed as

$$\boldsymbol{y} = \boldsymbol{H}\boldsymbol{x} + \boldsymbol{z}, \tag{1}$$

where the m-th element of the TX vector $\boldsymbol{x} \in \mathcal{X}^{M\times 1}$, denoted by x_m for all $m \in \{1, 2, \cdots, M\}$, represents the TX symbol from the m-th UE device, and the measurement matrix $\boldsymbol{H} \in \mathbb{C}^{N\times M}$ denotes a MU-MIMO channel matrix, whose (n,m)-th entry $h_{n,m} \in \mathbb{C}$ corresponds to the fading coefficient of the channel between the m-th UE device and the n-th antenna element at the BS. In addition, $\boldsymbol{z} \sim \mathcal{CN}(\boldsymbol{0}, N_0\boldsymbol{I}_N)$ is an additive white Gaussian noise (AWGN) vector of the noise power density N_0. Assuming that channel estimation on the BS is performed without error, our goal is to infer the unknown vector $\boldsymbol{x}$ in (1) based on the knowledge of $\boldsymbol{y}$, $\boldsymbol{H}$, and N_0.

B. Channel Model

To represent the spatial correlation among fading coefficients, we construct channel matrices based on the Kronecker model [12] as

$$\boldsymbol{H} = \boldsymbol{R}_\mathrm{RX}^{1/2}\boldsymbol{G}\boldsymbol{R}_\mathrm{TX}^{1/2}, \tag{2}$$

Algorithm 1 - GAMP-based Signal Detector [5]

Input: $\boldsymbol{y} \in \mathbb{C}^{N\times 1}, \boldsymbol{H} \in \mathbb{C}^{N\times M}, T$

1: $\forall m : \check{x}_m^{(1)} = \overline{x}_m^{(0)} = 0,\ \check{v}_m^{(1)} = \overline{v}_m^{(0)} = E_\mathrm{s}$

2: $\forall n : s_n^{(0)} = 0$

3: **for** $t = 1$ to T **do**

/* Module A */

4: $\forall n : \gamma_n^{(t)} = \sum_{m=1}^{M} |h_{n,m}|^2 \check{v}_m^{(t)}$

5: $\forall n : p_n^{(t)} = \sum_{m=1}^{M} h_{n,m}\check{x}_m^{(t)} - \gamma_n^{(t)} s_n^{(t-1)}$

6: $\forall n : \psi_n^{(t)} = \gamma_n^{(t)} + N_0$

7: $\forall n : s_n^{(t)} = \frac{y_n - p_n^{(t)}}{\psi_n^{(t)}}$ ▷ Soft IC

8: $\forall m : \acute{v}_m^{(t)} = \left(\sum_{n=1}^{N} \frac{|h_{n,m}|^2}{\psi_n^{(t)}}\right)^{-1}$

9: $\forall m : \acute{x}_m^{(t)} = \check{x}_m^{(t)} + \overline{v}_m^{(t)} \sum_{n=1}^{N} h_{n,m}^* s_n^{(t)}$ ▷ MF

10: $\forall m : \overline{x}_m^{(t)} = \zeta^{(t)}\acute{x}_m^{(t)} + \left(1-\zeta^{(t)}\right)\overline{x}_m^{(t-1)}$

11: $\forall m : \overline{v}_m^{(t)} = \zeta^{(t)}\acute{v}_m^{(t)} + \left(1-\zeta^{(t)}\right)\overline{v}_m^{(t-1)}$ ▷ Damping

/* Module B */

12: $\forall m : \check{x}_m^{(t+1)} = \dfrac{\sum_{\chi_q\in\mathcal{X}} \chi_q \exp\left[-\frac{\left|\chi_q - \overline{x}_m^{(t)}\right|^2}{\overline{v}_m^{(t)}}\right]}{\sum_{\chi'_q\in\mathcal{X}} \exp\left[-\frac{\left|\chi'_q - \overline{x}_m^{(t)}\right|^2}{\overline{v}_m^{(t)}}\right]}$

13: $\forall m : \check{v}_m^{(t+1)} = \dfrac{\sum_{\chi_q\in\mathcal{X}} \left|\chi_q - \check{x}_m^{(t+1)}\right|^2 \exp\left[-\frac{\left|\chi_q - \overline{x}_m^{(t)}\right|^2}{\overline{v}_m^{(t)}}\right]}{\sum_{\chi'_q\in\mathcal{X}} \exp\left[-\frac{\left|\chi'_q - \overline{x}_m^{(t)}\right|^2}{\overline{v}_m^{(t)}}\right]}$

14: **end for**

where each element of $\boldsymbol{G} \in \mathbb{C}^{N\times M}$ representing the instantaneous variation of the channel follows the i.i.d. complex Gaussian distribution $\mathcal{CN}(0,1)$, and $\boldsymbol{R}_\mathrm{TX} \in \mathbb{C}^{M\times M}$ and $\boldsymbol{R}_\mathrm{RX} \in \mathbb{C}^{N\times N}$ denote the spatial correlation matrices on the TX and RX sides, respectively. Each element of $\boldsymbol{R}_\mathrm{RX}$ is generated based on the exponential attenuation model [21] by

$$[\boldsymbol{R}_\mathrm{RX}]_{i,j} = \begin{cases} 1, & i = j, \\ \rho^{|i-j|}, & i \neq j, \end{cases} \tag{3}$$

where $\rho \in [0,1]$ denotes a fading correlation coefficient between two adjacent RX antennas on the BS side, whereas $\boldsymbol{R}_\mathrm{TX}$ is set to $\boldsymbol{I}_M$ in MU-MIMO systems.

C. GAMP-based Signal Detection

The pseudo-code of the GAMP-based signal detector for the linear inference in (1) is given in Alg. 1. For simplicity of notations, we hereafter refer to the l-th line of Alg. 1 as (A1-l). The qualifier $(\cdot)^{(t)}$ denotes the iteration index $t \in \mathcal{T} = \{1, 2, \cdots, T\}$ for every variable, where T denotes the total number of iterations.

Alg. 1 consists of two modules [6]. In Module A, soft interference cancellation (IC) is performed for every RX symbol in (A1-6, 7) based on the tentative estimates, *a.k.a*, *soft replicas*, followed by the linear estimation process based on matched filter (MF) in (A1-8, 9) with the normalized residual interference and noise component $s_n^{(t)}$. Module B then performs symbol-wise nonlinear estimation by considering

the filter output $\overline{x}_m^{(t)}$ and its variance $\overline{v}_m^{(t)}$ as the output of the virtual AWGN channel and its noise variance, respectively [8], *i.e.*,

$$\overline{x}_m^{(t)} = x_m + w_m, \quad w_m \sim \mathcal{CN}\left(0, \overline{v}_m^{(t)}\right). \tag{4}$$

Specifically, calculating the conditional expectation based on (4) yields the updated soft replica $\check{x}_m^{(t+1)}$ and its mean square error (MSE) $\check{v}_m^{(t+1)}$ in (A1-12, 13). The second term in the right-hand side of (A1-5) is called *Onsager correction* term, comprising the fundamental operating principle of GAMP. Onsager correction can be systematically derived as an approximation of the *extrinsic* belief generation in GaBP under i.i.d. measurements in the large-system limit [20] and enables to decouple the self-noise propagation across iterations. Hence, the validity of the resultant message update rule heavily relies on the ideal condition of large-scale uncorrelated observation, and its elaborately designed behavior even becomes a potential disturbance to the training process when applying DU under non-ideal conditions. Finally, belief damping [12] is introduced in (A1-10, 11) where $\zeta^{(t)} \in [0, 1]$ is the damping parameter.

III. Differential Learning for Trainable-GAMP

In highly correlated channels, the belief damping method can no longer handle the modeling error, leading to significant performance degradation. In such scenarios, the belief scaling method, which is originally proposed for GaBP in [3], enables GAMP to achieve robust signal detection even superior to that of GaBP [20]. In this section, after introducing the scaling method to Alg. 1, we aim to simultaneously optimize both damping and scaling parameters via the DU techniques.

A. Belief Scaling

Under correlated observation, the statistical behavior of the Module A output in Alg. 1 deviates from the ideal AWGN-corrupted model in (4), rendering $\overline{v}_m^{(t)}$ no longer trustable in describing the reliability of $\overline{x}_m^{(t)}$. Consequently, in the early iterations, the denoiser operating based on $\overline{v}_m^{(t)}$ in (A1-12, 13) overestimates the reliability of $\overline{x}_m^{(t)}$, leading to the generation of incorrect hard-decision symbols. These errors propagate across iterations and degrade the detection accuracy of other estimates through the soft IC process. The belief scaling is a method for suppressing this error propagation by controlling reliability with a parameter instead of $\overline{v}_m^{(t)}$, which is achieved by replacing (A1-12, 13) with the following denoiser [3]:

$$\check{x}_m^{(t+1)} = \sum_{\chi_q \in \mathcal{X}} \chi_q \frac{\exp\left[-\frac{\left|\chi_q - \overline{x}_m^{(t)}\right|^2}{c^2/a^{(t)}}\right]}{\sum_{\chi_q' \in \mathcal{X}} \exp\left[-\frac{\left|\chi_q' - \overline{x}_m^{(t)}\right|^2}{c^2/a^{(t)}}\right]}, \tag{5a}$$

$$\check{v}_m^{(t+1)} = \sum_{\chi_q \in \mathcal{X}} \left|\chi_q - \check{x}_m^{(t+1)}\right|^2 \frac{\exp\left[-\frac{\left|\chi_q - \overline{x}_m^{(t)}\right|^2}{c^2/a^{(t)}}\right]}{\sum_{\chi_q' \in \mathcal{X}} \exp\left[-\frac{\left|\chi_q' - \overline{x}_m^{(t)}\right|^2}{c^2/a^{(t)}}\right]}, \tag{5b}$$

where the embedded scaling parameters $\left\{a^{(t)}; t \in \mathcal{T}\right\}$ adjust the *softness* of the denoiser function at each iteration step, and the constant $c \triangleq \sqrt{3E_{\mathrm{s}}/\left(2(Q-1)\right)}$ denotes the real/imaginary part of the Q-QAM modulated constellation point closest to the origin. A strategy empirically known to be effective in designing scaling parameters [3], [20] is keeping $a^{(t)}$ small in the early iterations to prevent the generation of incorrect hard-decision symbols and gradually increasing $a^{(t)}$ as the iterations progress to induce convergence.

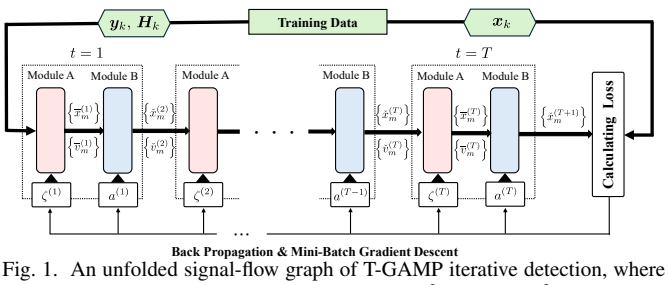


Fig. 1. An unfolded signal-flow graph of T-GAMP iterative detection, where T denotes the number of iterations or layers and $\{\boldsymbol{y}_k, \boldsymbol{H}_k, \boldsymbol{x}_k\}$ representing the training data set.

The use of denoiser in (5) often causes the vanishing gradient problem in back propagation processes due to $\exp(\cdot)$ operations. To overcome this numerical instability, the approximate denoiser was proposed in [3], [17], which can be extended for complex-valued inputs as follows:

$$\check{x}_m^{(t+1)} = \eta\left(\Re\left[\overline{x}_m^{(t)}\right]; \frac{c^2}{2a^{(t)}}\right) + \mathrm{j}\eta\left(\Im\left[\overline{x}_m^{(t)}\right]; \frac{c^2}{2a^{(t)}}\right), \tag{6a}$$

$$\check{v}_m^{(t+1)} = 2c^2\left(\sqrt{Q}-1\right)^2 - \left|\check{x}_m^{(t+1)}\right|^2 + \xi\left(\Re\left[\overline{x}_m^{(t)}\right]; \frac{c^2}{2a^{(t)}}\right) + \xi\left(\Im\left[\overline{x}_m^{(t)}\right]; \frac{c^2}{2a^{(t)}}\right), \tag{6b}$$

where we define

$$\eta(u; v) \triangleq c \sum_{\gamma \in \mathcal{G}_Q} \tanh\left[c \cdot \frac{u - \gamma}{v}\right], \tag{7}$$

$$\xi(u; v) \triangleq 2c \sum_{\gamma \in \mathcal{G}_Q} \gamma \cdot \tanh\left[c \cdot \frac{u - \gamma}{v}\right], \tag{8}$$

with the symbol decision threshold set of

$$\mathcal{G}_Q \triangleq \left\{0, \pm 2c, \pm 4c, \cdots, \pm\left(\sqrt{Q}-2\right)c\right\}. \tag{9}$$

It is worth noting that (6) is exactly the same as (5) in the case of $Q = 4$, and the approximation error in (6) in the case of $Q \geq 16$ can be properly compensated through the data-driven tuning.

We hereafter refer to the algorithm obtained by replacing (A1-12, 13) with (6) as trainable GAMP (T-GAMP), which has the set of learnable parameters $\mathcal{C} \triangleq \{\boldsymbol{a}, \boldsymbol{\zeta}\}$, where we have defined $\boldsymbol{a} \triangleq \left[a^{(1)}, a^{(2)}, \ldots, a^{(T)}\right]^{\mathsf{T}} \in \mathbb{R}_+^{T\times 1}$ and $\boldsymbol{\zeta} \triangleq \left[\zeta^{(1)}, \zeta^{(2)}, \ldots, \zeta^{(T)}\right]^{\mathsf{T}} \in [0, 1]^{T\times 1}$. Fig. 1 shows a signal-flow graph of T-GAMP, whose number of layers is equivalent to the number of iterations T.

B. Differential Learning for T-GAMP

In [17], the scaling parameter $\boldsymbol{a}$ embedded in T-GaBP is given as the output of rectified linear unit (ReLU) function

due to its non-negative constraint, *i.e.*,

$$\boldsymbol{a} = \mathrm{ReLU}\left(\boldsymbol{a}_{\mathrm{target}}\right), \tag{10}$$

where the function $\mathrm{ReLU} : \mathbb{R}^{L\times 1} \to \mathbb{R}_{+}^{L\times 1}$ element-wisely calculates $\max\left(b_l, 0\right), l \in \{1, 2, \cdots, L\}$ for any input vector $\boldsymbol{b} \triangleq [b_1, b_2, \ldots, b_L]^{\mathsf{T}} \in \mathbb{R}^{L\times 1}$ with length L, and the vector $\boldsymbol{a}_{\mathrm{target}} \in \mathbb{R}^{T\times 1}$ is the intermediate parameter to be directly updated via gradient calculation in the training process. However, when the same learning process is performed with T-GAMP, the learning behavior becomes unstable, which suggests the difficulty in optimizing the *sophisticated* message update rule of GAMP, designed to guarantee optimality in the large-system limit, for observations that deviate from the ideal condition.

To stabilize the learning behavior, this paper additionally incorporates the finding that "the reliability of beliefs should gradually increase with iterations" as domain knowledge into the learning process. Specifically, for the purpose of imposing monotonically increasing constraints, *i.e.*, $a^{(1)} \le a^{(2)} \le \cdots \le a^{(T)}$, the scaling parameter $\boldsymbol{a}$ is constructed as

$$\boldsymbol{a} = \mathrm{Cumsum}\left(\mathrm{ReLU}(\boldsymbol{a}_{\mathrm{target}})\right), \tag{11}$$

where the function $\mathrm{Cumsum} : \mathbb{R}^{L\times 1} \to \mathbb{R}^{L\times 1}$ is defined as

$$\mathrm{Cumsum}(\boldsymbol{b}) \triangleq \left[\sum_{l=1}^{1} b_l, \sum_{l=1}^{2} b_l, \ldots, \sum_{l=1}^{L} b_l\right] \in \mathbb{R}^{L\times 1}, \tag{12}$$

for arbitrary input vector $\boldsymbol{b} \in \mathbb{R}^{L\times 1}$. For a more intuitive explanation, the parameter construction in (11) can be interpreted as a method of learning the differential (incremental) values of the scaling parameter between the two successive iterations, *i.e.*,

$$\Delta a_t \triangleq \begin{cases} a^{(1)}\ (\ge 0), & t = 1, \\ a^{(t)} - a^{(t-1)}\ (\ge 0), & t = 2, 3, \ldots, T, \end{cases} \tag{13}$$

rather than the scaling parameter itself. Hereafter, we shall refer to this learning method as "*differential learning*."

IV. Numerical Simulations

To confirm the efficacy of the proposed differential learning in data-driven tuning of T-GAMP, computer simulations were conducted. The average RX power from each TX antenna is assumed to be identical due to the slow TX power control, and the time and frequency synchronization is assumed to be perfect. The training process of T-GAMP is implemented by PyTorch with Adam optimizer. Every learning result in this section is obtained via mini-batch gradient descent method with the mini-batch size of 1×10^2 and the number of parameter updates of 2×10^3. The loss function was set to the MSE between the Module B output in the final iteration (layer) $\breve{x}_m^{(T)}$ and the true symbol x_m for all $m \in \{1, 2, \cdots, M\}$ in one mini-batch. The damping parameters for $t = 2, 3, \ldots, T$ are constructed as the output of the sigmoid function as

$$\zeta^{(t)} = \begin{cases} 1.0, & t = 1, \\ \left(1 + e^{-g\zeta_{\mathrm{target}}^{(t)}}\right)^{-1}, & t = 2, 3, \ldots, T, \end{cases} \tag{14}$$

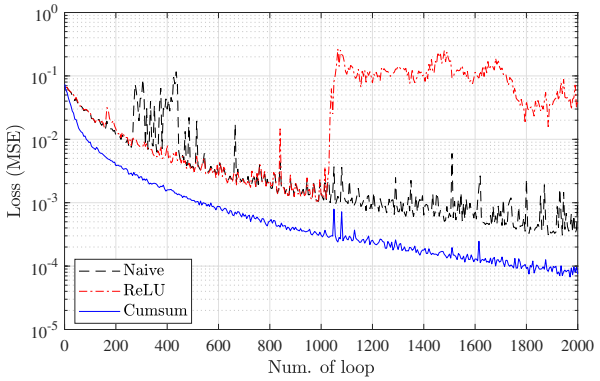


Fig. 2. Loss trajectory of the learning process when $(M, N) = (64, 64)$, $Q = 4$, $\rho = 0.80$, and $E_{\mathrm{s}}/N_0 = 0$ dB.

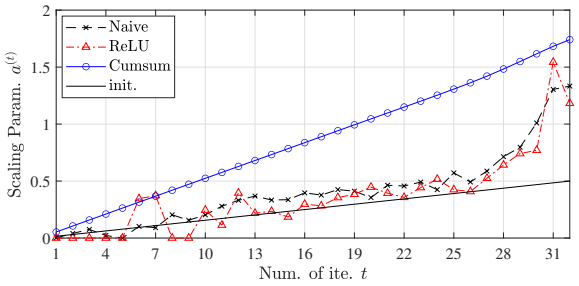
where the hyperparameter $g\ (> 0)$ adjusts the *softness* of the sigmoid function. For $t = 2, 3, \ldots, T$, the damping parameters were initialized as $\zeta^{(t)} = 0.50$. Meanwhile, the scaling parameter $\boldsymbol{a}$ was tested by the following three different approaches: *i.e.*, by $\boldsymbol{a} = \boldsymbol{a}_{\mathrm{target}}$, (10), and (11), each referred to as "Naive," "ReLU," and "Cumsum," respectively. In every case, $\boldsymbol{a}$ was initialized as $a^{(t)} = 0.50 \cdot (t/T), \forall t \in \mathcal{T}$. For "Naive" and "ReLU," the hyperparameter g and the learning rate of Adam were respectively set to 1.0 and 2.0×10^{-3}, while for "Cumsum," 5.0 and 2.0×10^{-4}, respectively[2].

A. Learning Results

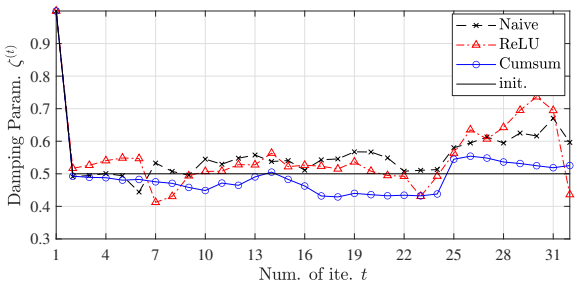
The loss trajectories of the learning process and the learned results are shown in Figs. 2 and 3, respectively, where the MIMO configuration, QAM modulation order, correlation coefficient, and noise power density were respectively set to $(M, N) = (64, 64)$, $Q = 4$, $\rho = 0.80$, and $E_{\mathrm{s}}/N_0 = 0$ dB. In addition, the number of iterations was set to $T = 32$.

In Fig. 2, the "Naive" approach, which does not account for the non-negative constraint of scaling parameters, exhibits violent oscillation, especially during the initial parameter updates, which is due to the back propagation of inconsistent gradient information such that negative scaling parameters are allowed. While the "ReLU" method mitigates this instability to some extent, its loss curve shows a sharp rise around the 1040th update, where the dynamics of the scaling parameters start to exhibit a zigzag pattern, and the loss remains high in the subsequent updates. Intuitively, once the loss value experiences a sudden upward spike, the step size for the parameter updates based on gradient computation can become exaggerated, leading to unintended distortion in the shape of the learned parameters and stagnation in loss reduction. Even worse for "ReLU" case, this stagnation forces repeated drastic parameter updates, which can ultimately disrupt the entire training process. In fact, the shapes of the parameters learned through the "Naive" and "ReLU" approaches in Fig. 3 demonstrate irregular fluctuations. In contrast, the loss trajectory for "Cumsum" learning process in Fig. 2 is remarkably

[2]In differential learning, the learning rate should be set low because the step size of the scaling parameter is determined by the cumulative sum of the step sizes in the incremental values. The parameter g is set to compensate for the learning delay of the damping parameter due to this low learning rate.



(a) Scaling Parameter $\{a^{(t)}; t \in \mathcal{T}\}$.



(b) Damping Parameter $\{\zeta^{(t)}; t \in \mathcal{T}\}$.

Fig. 3. Comparison of the learned parameters when $(M, N) = (64, 64)$, $Q = 4$, $\rho = 0.80$, and $E_{\rm s}/N_0 = 0$ dB.

stable, which facilitates convergence to the solution that well reflects the additional domain knowledge, as shown in Fig. 3. These results suggest the solid efficacy of differential learning in stabilizing the training process of T-GAMP. The further analysis of this phenomenon is not straightforward, and thus is left for our future work.

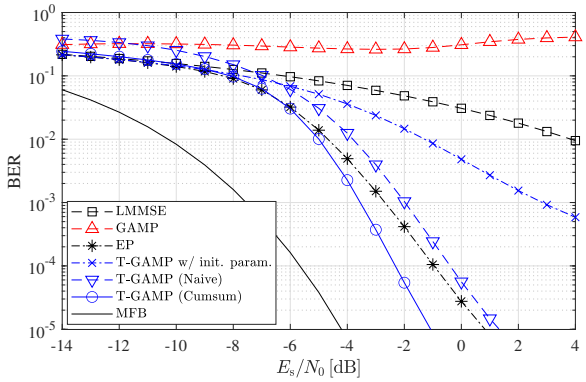
B. BER Performance

Next, BER performance as a function of $E_{\rm s}/N_0$ is presented in Fig. 4 under the same configuration as that in Section IV-A to verify the efficacy of the T-GAMP adopting the parameters obtained by the differential learning method.

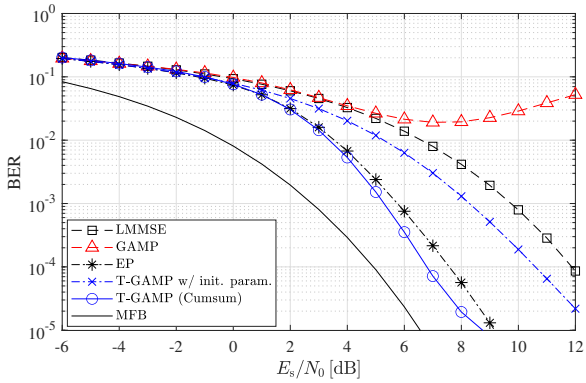
The following schemes are compared:

- *LMMSE*: Baseline performance of linear minimum mean square error (MMSE) filtering.
- *GAMP*: Performance of Alg. 1 where the number of iterations is set to $T = 32$ and the damping parameter to $\zeta^{(1)} = 1.0$ and $\zeta^{(t)} = 0.50$ for $t = 2, 3, \ldots, T$.
- *EP*: A high-complexity but powerful Bayesian signal detector that performs linear MMSE filtering for every iteration. The same belief damping as GAMP is used and the number of iterations is set to $T = 16$.
- *T-GAMP w/ init. param.*: T-GAMP detector using the initialized parameters, with the number of iterations $T = 32$.
- *T-GAMP*: T-GAMP detector adopting the learned parameters, with the number of iterations $T = 32$.
- *MFB*: Matched filter bound (MFB) is the referential lower bound that Bayesian detectors can ideally achieve.

The results of "T-GAMP (ReLU)," whose loss trajectory showed a radical leap in Fig. 2, are omitted.



(a) $(M, N) = (64, 64)$, $Q = 4$.



(b) $(M, N) = (16, 32)$, $Q = 16$.

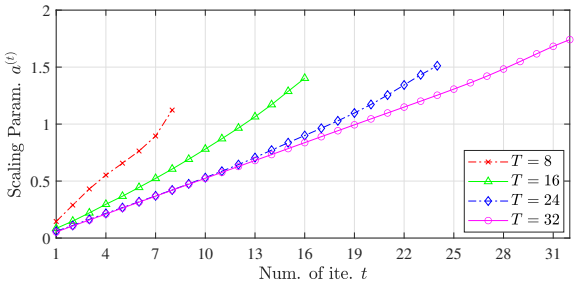
Fig. 4. BER performance versus $E_{\rm s}/N_0$ ($\rho = 0.80$).

As can be observed in Fig. 4(a), due to the strong spatial correlation among fading coefficients, "GAMP" fails to reliably detect spatially multiplexed MIMO signals (BER $> 10^{-1}$) even when belief damping is employed. In contrast, "T-GAMP (Cumsum)," GAMP adopting the scaling and damping parameters optimized through differential learning, shows a significant improvement in detection accuracy. More specifically, it achieves a gain of about 2.0 dB against "EP," and 2.5 dB against "T-GAMP (Naive), with only about 3.0 dB performance degradation from "MFB" at BER $= 10^{-5}$.

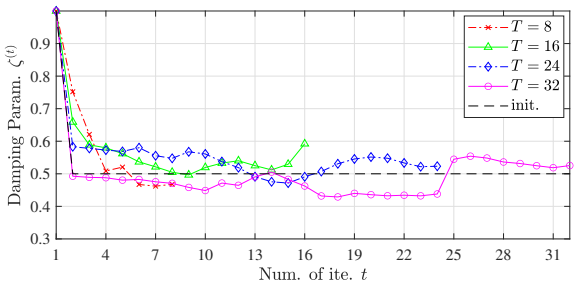
In the case of high-order modulation, Fig. 4(b) shows the BER performance under the system parameters of $(M, N) = (16, 32)$, $Q = 16$, and $\rho = 0.80$, where the learnable parameters for T-GAMP were learned at $E_{\rm s}/N_0 = 8$ dB. The results of "T-GAMP (Naive)," "T-GAMP (ReLU)," whose loss trajectory showed a radical leap, are omitted. Even when using non-orthogonal mapping rules, which tends to make learning unstable, "T-GAMP (Cumsum)" can achieve stable learning, and outperform "EP," with only about 2.0 dB degradation from "MFB" at BER $= 10^{-4}$.

C. Iterative Behavior Analysis

Finally, we evaluate the convergence behavior of iterative detection to confirm the phenomenon of *data-drive acceleration* via differential learning. Figs. 5 and 6 respectively show the learned parameters for different number of iterations ($T = 8, 16, 24,$ and 32) and the resultant BER performance



(a) Scaling Parameter $\{a^{(t)}; t \in \mathcal{T}\}$.



(b) Damping Parameter $\{\zeta^{(t)}; t \in \mathcal{T}\}$.

Fig. 5. Learned parameters for different number of iterations T when $(M, N) = (64, 64)$, $Q = 4$, $\rho = 0.80$, and $E_s/N_0 = 0$ dB.

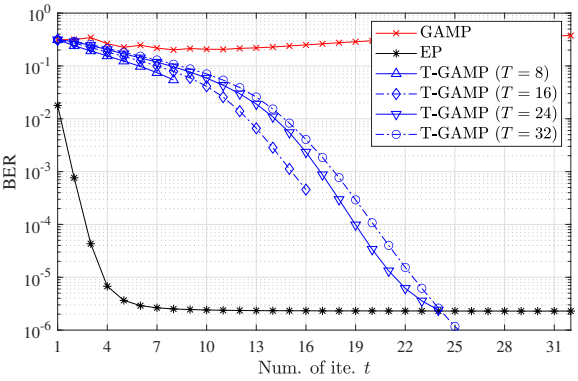


Fig. 6. Comparison of the iterative convergence behavior in terms of BER when $(M, N) = (64, 64)$, $Q = 4$, $\rho = 0.80$, and $E_s/N_0 = 2$ dB.

at each iteration step under the same configuration as that in Section IV-A. Fig. 5 shows that as T decreases, the slope of the change in the scaling parameter increases, and the damping parameter becomes larger in the early iterations, allowing the convergence speed to be increased. Consequently, Fig. 6 shows that the parameters learned with smaller T enable slightly accelerated convergence of the BER, yet at the cost of BER degradation achieved at the final iteration $t = T$. The results suggest that T-GAMP with differential learning under highly correlated channels requires at least about $T = 24$ iterations to achieve sufficient detection accuracy.

V. CONCLUSION

In this paper, we proposed a novel method to stabilize learning behavior while simultaneously optimizing the damping and scaling parameters embedded in the GAMP-based signal detection for separating the spatially multiplexed MU-MIMO signals under highly correlated observation. Numerical results demonstrated that the learning behavior of T-GAMP became dramatically stabilized by imposing a monotonically increasing constraint on the scaling parameters by learning the non-negative incremental values of the scaling parameters across each iteration. The efficacy of the proposed scheme was confirmed in terms of loss trajectory, dynamics of the learned parameters, and BER.

ACKNOWLEDGEMENT

This work was supported in part by JST, CRONOS, Japan Grant Number JPMJCS24N1, and in part by MIC/FORWARD under Grant JPMI240710001.

REFERENCES

[1] W. Chen, X. Lin, J. Lee, A. Toskala, S. Sun, C. F. Chiasserini, and L. Liu, "5G-advanced toward 6G: Past, present, and future," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 6, pp. 1592–1619, 2023.

[2] Y. Kabashima, "A CDMA multiuser detection algorithm on the basis of belief propagation," *J. Phys. A, Math. Gen.*, vol. 36, no. 43, pp. 11 111–11 121, Oct. 2003.

[3] T. Takahashi, S. Ibi, and S. Sampei, "Design of adaptively scaled belief in multi-dimensional signal detection for higher-order modulation," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 1986–2001, 2019.

[4] D. L. Donoho, A. Maleki, and A. Montanari, "Message-passing algorithms for compressed sensing," *Proc. of the National Academy of Sciences*, vol. 106, no. 45, pp. 18 914–18 919, 2009.

[5] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. IEEE ISIT*, 2011, pp. 2168–2172.

[6] X. Meng, S. Wu, and J. Zhu, "A unified Bayesian inference framework for generalized linear models," *IEEE Signal Process. Lett.*, vol. 25, no. 3, pp. 398–402, Mar. 2018.

[7] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector approximate message passing," *IEEE Trans. Inf. Theory*, vol. 65, no. 10, pp. 6664–6684, 2019.

[8] K. Takeuchi, "Rigorous dynamics of expectation-propagation-based signal recovery from unitarily invariant measurements," *IEEE Trans. Inf. Theory*, vol. 66, no. 1, pp. 368–386, 2020.

[9] M. Bayati and A. Montanari, "The dynamics of message passing on dense graphs, with applications to compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 2, pp. 764–785, Feb. 2011.

[10] F. Caltagirone, L. Zdeborová, and F. Krzakala, "On convergence of approximate message passing," in *Proc. ISIT*, 2014, pp. 1812–1816.

[11] T. Takahashi, S. Ibi, and S. Sampei, "Design of criterion for adaptively scaled belief in iterative large MIMO detection," *IEICE Trans. Commun.*, vol. E102.B, no. 2, pp. 285–297, 2019.

[12] A. Chockalingam and B. S. Rajan, *Large MIMO Systems*. Cambridge University Press, 2014.

[13] K. Gregor and Y. LeCun, "Learning fast approximations of sparse coding," in *Proc. 27th Int. Conf. Machine Learning*, 2010, p. 399–406.

[14] V. Monga, Y. Li, and Y. C. Eldar, "Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing," *IEEE Signal Process. Magazine*, vol. 38, no. 2, pp. 18–44, 2021.

[15] D. Ito, S. Takabe, and T. Wadayama, "Trainable ISTA for sparse signal recovery," *IEEE Trans. Signal Process.*, vol. 67, no. 12, pp. 3113–3125, 2019.

[16] N. Shlezinger, J. Whang, Y. C. Eldar, and A. G. Dimakis, "Model-based deep learning," *Proceedings of the IEEE*, vol. 111, no. 5, pp. 465–499, 2023.

[17] D. Shirase, T. Takahashi, S. Ibi, K. Muraoka, N. Ishii, and S. Sampei, "Deep unfolding-aided Gaussian belief propagation for correlated large MIMO detection," in *Proc. IEEE GLOBECOM*, 2020, pp. 1–6.

[18] ——, "Negentropy-aware loss function for trainable belief propagation in coded MIMO detection," in *Proc. IEEE GLOBECOM*, 2021, pp. 1–6.

[19] M. Borgerding, P. Schniter, and S. Rangan, "AMP-inspired deep networks for sparse linear inverse problems," *IEEE Trans. Signal Process.*, vol. 65, no. 16, pp. 4293–4308, 2017.

[20] R. Tamaki, K. Ito, T. Takahashi, S. Ibi, and S. Sampei, "Suppression of self-noise feedback in GAMP for highly correlated large MIMO detection," in *Proc. IEEE ICC*, 2022, pp. 1300–1305.

[21] M. Chiani, M. Win, and A. Zanella, "On the capacity of spatially correlated MIMO rayleigh-fading channels," *IEEE Trans. Inf. Theory*, vol. 49, no. 10, pp. 2363–2371, 2003.