Title	Discrete-Valued Signal Estimation via GAMP under Finite-Sized Highly Correlated GLMs
Author(s)	Furudoi, Tomoharu; Takahashi, Takumi; Ochiai, Hideki
Citation	IEEE International Conference on Communications. 2026
Version Type	AM
URL	https://hdl.handle.net/11094/106088
rights	© 2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.
Note	

Discrete-Valued Signal Estimation via GAMP under Finite-Sized Highly Correlated GLMs

Tomoharu Furudoi *, Takumi Takahashi *, and Hideki Ochiai *
* Graduate School of Engineering, The University of Osaka, 2-1 Yamada-oka, Suita, 565-0871, Japan
Email: *{furutomo@wcs., takahashi@, ochiai@}comm.eng.osaka-u.ac.jp

***Abstract*—This paper proposes a simple yet powerful strategy for generalized approximate message passing (GAMP) to achieve robust discrete-valued signal estimation under generalized linear models (GLMs) with model mismatches. Despite its minimal computational complexity, the design principle of GAMP critically relies on large-scale independent and identically distributed (i.i.d.) measurements, which are often violated in practice. The resulting statistical deviations in the propagating beliefs lead to severe performance degradation, particularly when discrete constraints on the unknown signal induce nonlinear *denoising* behavior. To address this issue, our strategy adaptively manipulates this nonlinearity through an embedded *inverse temperature* mechanism, effectively absorbing analytically intractable model mismatches. Numerical simulations demonstrate that, despite the combined challenges of GLM-induced belief distortions, finite-size effects, and highly correlated measurements, the proposed scheme robustly guides GAMP to approach the absolute performance limit while retaining its minimal computational complexity. Furthermore, we elucidate the underlying operating principle by analyzing the evolution of the message dynamics.**

***Index Terms*—Generalized approximate message passing, generalized linear model, annealed discrete denoiser.**

I. INTRODUCTION

This paper considers a signal estimation problem formulated under the following generalized linear model (GLM) [1]

$$\boldsymbol{y} = g(\boldsymbol{z}, \boldsymbol{w}) \in \mathbb{C}^{N\times 1}, \quad \boldsymbol{z} = \boldsymbol{A}\boldsymbol{x} \in \mathbb{C}^{N\times 1}, \tag{1}$$

which describes the process where the input signal $\boldsymbol{x} \in \mathbb{C}^{M\times 1}$ is linearly mixed by a known measurement matrix $\boldsymbol{A} \in \mathbb{C}^{N\times M}$ to produce the intermediate output $\boldsymbol{z} \in \mathbb{C}^{N\times 1}$. The noise vector $\boldsymbol{w} \in \mathbb{C}^{N\times 1}$ is then nonlinearly superimposed through a known function[1], $g : \mathbb{C}^2 \to \mathbb{C}$, yielding the observation $\boldsymbol{y} \in \mathbb{C}^{N\times 1}$. Note that when $g(z, w) = z + w$, (1) reduces to the well-known standard linear model (SLM): $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{w}$. Except for the special case where g is linear and $\boldsymbol{x}$ is Gaussian distributed, solving the inverse problem of (1) based on the minimum mean square error (MMSE) criterion becomes computationally intractable. Our goal is to approximate the optimal solutions while avoiding such prohibitive complexity, under the assumption that each entry of $\boldsymbol{x}$ is independently drawn from a finite alphabet set $\mathcal{X}$ according to a known distribution.

To that end, generalized approximate message passing (GAMP) [2], rooted in the classical loopy belief propagation (BP) framework, is a powerful iterative algorithm that requires minimal computational complexity of $\mathcal{O}(MN)$. Similarly to the Bayesian approximate message passing (AMP) [3]–[5], its update rules are meticulously designed to ensure its asymptotic convergence to the Bayes-optimal (*i.e.*, MMSE) solution in the large-system limit[2], under an ideal zero-mean independent and identically distributed (i.i.d.) Gaussian measurement matrix ensemble [6]–[8].

The underlying principle behind this design is the *concentration* phenomenon [9], [10], whereby the macroscopic statistical behavior of propagating messages (*i.e.*, *beliefs*) becomes typical and deterministic in the large-system limit. However, this property of GAMP gives rise to two major challenges when applied to practical system configurations. The first one is the *finite-size effect* for realistic system dimensions; the instantaneous behavior of beliefs can deviate from the assumed ideal Gaussian distribution, inevitably producing belief outliers. Particularly when $\boldsymbol{x}$ is discrete-valued, such outliers may result in erroneous and overly decisive tentative estimates with non-negligible probability, leading to inevitable performance degradation. As an additional challenge, practical systems often involve inter-element correlation within $\boldsymbol{A}$, violating the theoretical requirement for GAMP's applicability. Although more robust GLM-based message passing algorithms (MPAs) have been proposed—typically incorporating iterative linear MMSE (LMMSE) filtering [1], [11]—the required condition of *bi-unitarily invariance* of $\boldsymbol{A}$ [10], [12] remains overly restrictive for practical use. Furthermore, information loss in the measurement process g, such as few-bit quantization [13], [14] or clipping [15], increases the estimation burden and further amplifies their reliance on the large-system assumption compared with their SLM counterparts [16], [17]. In short, a steady and reliable scheme for GLM-based signal estimation under realistic system configurations remains absent.

In squarely addressing such intractable modeling errors encountered in practical scenarios, steady bottom-up strategies are often more effective than sophisticated top-down algorithmic designs. One such approach is *belief scaling* [18]–[20], originally proposed for SLM-based discrete-valued signal estimation. More specifically, Bayesian MPAs with discrete priors on the unknown signal exhibit significant nonlinearity in the *denoising* process. The belief scaling method adaptively manipulates this nonlinearity across iterations through an *inverse temperature* mechanism to appropriately adjust the timing to produce confident symbol decisions. Motivated by this perspective, this paper is the first to extend this approach to GLMs, with a particular focus on the uplink multi-user

[1]In this paper, for a function $g : \mathbb{C}^2 \to \mathbb{C}$ and arbitrary vectors $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{C}^{K\times 1}$, we define $g(\boldsymbol{a}, \boldsymbol{b}) = [g([\boldsymbol{a}]_1, [\boldsymbol{b}]_1), \ldots, g([\boldsymbol{a}]_K, [\boldsymbol{b}]_K)]^\mathsf{T} \in \mathbb{C}^{K\times 1}$.

[2]The ideal system assumption where M and N tend to infinity with a fixed ratio $\xi = N/M$, *i.e.*, $N = \xi M \to \infty$.

(MU)-multi-input multi-output (MIMO) system as a representative use case, where the base station (BS) is equipped with low-resolution analog-to-digital converters (ADCs). Through extensive analysis, we demonstrate that, despite the severe model mismatches in $\boldsymbol{A}$ and the detrimental nonlinearity of g, the scaling strategy robustly restores the near-Gaussian behavior of beliefs far beyond the large-scale i.i.d. regime. Consequentially, it guides the algorithm closely toward the performance limits while maintaining minimal complexity.

These findings reveal that the inverse temperature control provides a principled algorithm design framework that bridges the long-standing gap between the theoretical foundations and practical viability of GAMP.

Notation: Sets of real and complex numbers are denoted by $\mathbb{R}$ and $\mathbb{C}$, respectively. Vectors and matrices are represented by lower- and upper-case bold-face letters. The conjugate, transpose, and conjugate transpose are denoted by $(\cdot)^*$, $(\cdot)^{\mathsf{T}}$, and $(\cdot)^{\mathsf{H}}$, respectively. For a complex quantity $c \in \mathbb{C}$, its real and imaginary parts are denoted by $c^{\Re} (= \Re[c])$ and $c^{\Im} (= \Im[c])$. The imaginary unit by $\mathrm{j} \triangleq \sqrt{-1}$. The $K \times K$ identity matrix is denoted by $\boldsymbol{I}_K$, while the zero vector by $\mathbf{0}$. The i-th element of a vector $\boldsymbol{b}$ and the (i,j)-th element of a matrix $\boldsymbol{B}$ are denoted by $[\boldsymbol{b}]_i$ and $[\boldsymbol{B}]_{i,j}$, respectively. The complex Gaussian distribution with mean vector $\boldsymbol{\mu}$ and 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 distribution $\mathcal{P}$. The probability that an event A occurs is denoted by $\mathbb{P}\{A\}$. The probability mass function (PMF), probability density function (PDF), and expectation with respect to (w.r.t.) random variable a are denoted by $P_{\mathsf{a}}[\cdot]$, $p_{\mathsf{a}}(\cdot)$, and $\mathbb{E}_{\mathsf{a}}[\cdot]$, respectively. Likewise, the conditional PMF, PDF, expectation, and variance w.r.t. a given $b = b'$ are denoted by $P_{\mathsf{a}|\mathsf{b}}[\cdot|b']$, $p_{\mathsf{a}|\mathsf{b}}(\cdot|b')$, $\mathbb{E}_{\mathsf{a}}[\cdot|b']$, and $\mathbb{V}_{\mathsf{a}}[\cdot|b']$, respectively. The PDF and cumulative distribution function (CDF) of the standard Gaussian distribution are defined as $\phi(x) \triangleq \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ and $\Phi(x) \triangleq \int_{-\infty}^{x} \phi(t)\,\mathrm{d}t$.

II. Signal Model

This section describes the statistical modeling of the signal in the GLM (1). For notational simplicity, we hereafter define $x_m \triangleq [\boldsymbol{x}]_m, y_n \triangleq [\boldsymbol{y}]_n, z_n \triangleq [\boldsymbol{z}]_n, w_n \triangleq [\boldsymbol{w}]_n, a_{n,m} \triangleq [\boldsymbol{A}]_{n,m}$ for all $m = 1, 2, \ldots, M, n = 1, 2, \ldots, N$.

A. Prior Distribution for Input Signal

We assume that the i.i.d. elements of the input signal vector $\boldsymbol{x} \in \mathcal{X}^{M \times 1}$ follow the distribution

$$p_{\mathsf{x}_m}(x) = \sum_{\chi \in \mathcal{X}} P_{\mathsf{x}_m}[\chi] \cdot \delta(x - \chi), \tag{2}$$

with $\delta(\cdot)$ denoting Dirac's delta. Without loss of generality, we assume $\mathbb{E}_{\mathsf{x}}[\boldsymbol{x}] = \mathbf{0}$ and $\mathbb{E}_{\mathsf{x}}\left[\boldsymbol{x}\boldsymbol{x}^{\mathsf{H}}\right] = \sigma_{\mathrm{x}}^2 \boldsymbol{I}_M$.

B. Probabilistic Representation for Nonlinear Observation

Given that the noise vector $\boldsymbol{w} \in \mathbb{C}^{N \times 1}$ is random and follows $\mathcal{CN}\left(\mathbf{0}, \sigma_{\mathrm{w}}^2 \boldsymbol{I}_N\right)$, the dependency of $\boldsymbol{y} \in \mathbb{C}^{N \times 1}$ on the output signal $\boldsymbol{z} \in \mathbb{C}^{N \times 1}$ can be equivalently expressed in the following probabilistic form:

$$p_{\mathsf{y}|\mathsf{z}}(\boldsymbol{y} \mid \boldsymbol{z}) = \prod_{n=1}^{N} p_{\mathsf{y}_n|\mathsf{z}_n}(y_n \mid z_n), \tag{3}$$

Algorithm 1 Bayes-Optimal GAMP [2]

Input: $\boldsymbol{y} \in \mathbb{C}^{N \times 1}, \boldsymbol{A} \in \mathbb{C}^{N \times M}, T$.
Output: $\{\hat{x}_m(T+1)\}_{m=1}^{M}$.
1: $\forall m : \hat{x}_m(1) = 0, \hat{v}_m^{\mathrm{x}}(1) = \sigma_{\mathrm{x}}^2$
2: $\forall n : s_n(0) = 0.$ ▷ Initialization
3: **for** $t = 1$ to T **do**
/* Outer Module */
4: $\forall n : \overline{v}_n^{\mathrm{z}}(t) = \sum_{m=1}^{M} |a_{n,m}|^2 \hat{v}_m^{\mathrm{x}}(t)$
5: $\forall n : \overline{z}_n(t) = \sum_{m=1}^{M} a_{n,m}\hat{x}_m(t) - \overline{v}_n^{\mathrm{z}}(t) s_n(t-1)$
6: $\forall n : \hat{z}_n(t) = \mathbb{E}_{\mathsf{z}_n}\left[z_n \mid y_n, \overline{z}_n(t), \overline{v}_n^{\mathrm{z}}(t)\right]$
7: $\forall n : \hat{v}_n^{\mathrm{z}}(t) = \mathbb{V}_{\mathsf{z}_n}\left[z_n \mid y_n, \overline{z}_n(t), \overline{v}_n^{\mathrm{z}}(t)\right]$
8: $\forall n : s_n(t) = \frac{\hat{z}_n(t) - \overline{z}_n(t)}{\overline{v}_n^{\mathrm{z}}(t)}$
9: $\forall n : \psi_n(t) = \frac{\overline{v}_n^{\mathrm{z}}(t)}{1 - \hat{v}_n^{\mathrm{z}}(t)/\overline{v}_n^{\mathrm{z}}(t)}$
/* Inner Module */
10: $\forall m : \overline{v}_m^{\mathrm{x}}(t) = \left(\sum_{n=1}^{N} \frac{|a_{n,m}|^2}{\psi_n(t)}\right)^{-1}$
11: $\forall m : \overline{x}_m(t) = \hat{x}_m(t) + \overline{v}_m^{\mathrm{x}}(t) \sum_{n=1}^{N} a_{n,m}^* s_n(t)$
12: $\forall m : \hat{x}_m(t+1) = \mathbb{E}_{\mathsf{x}_m}\left[x_m \mid \overline{x}_m(t), \overline{v}_m^{\mathrm{x}}(t)\right]$
13: $\forall m : \hat{v}_m^{\mathrm{x}}(t+1) = \mathbb{V}_{\mathsf{x}_m}\left[x_m \mid \overline{x}_m(t), \overline{v}_m^{\mathrm{x}}(t)\right]$
14: **end for**

where the equality follows from the independence of $w_1, w_2, \ldots, w_N$ and the component-wise operation of g. For instance, corresponding to the special case of SLM $y_n = g(z_n, w_n) = z_n + w_n$, we have Gaussian PDF $p_{\mathsf{y}_n|\mathsf{z}_n}(y \mid z) = \frac{1}{\pi\sigma_{\mathrm{w}}^2} e^{-|y-z|^2/\sigma_{\mathrm{w}}^2}$.

III. GAMP and Its Model Error-Aware Denoiser Design

A. Generalized Approximate Message Passing (GAMP)

Alg. 1 presents the algorithmic structure of GAMP, expressed in pseudocode, which is applicable to general input signals $\boldsymbol{x}$ (not limited to discrete-valued cases). For later reference, we denote the l-th line of Alg. 1 as (A1-l). The notation $(\cdot)(t)$ represents a variable at the t-th iteration, where $t = 1, 2, \ldots, T$, and T denotes the total number of iterations. Alg. 1 consists of two modules, referred to as the *outer module* and the *inner module* [21], which are briefly described below. For further details, readers are referred to [2].

1) Outer Module: The outer module first constructs the *beliefs* and their estimated mean square error (MSE) for $\boldsymbol{z}$, defined as $\{\overline{z}_n, \overline{v}_n^{\mathrm{z}}\}_{n=1}^{N}$, by combining the tentative estimates of $\boldsymbol{x}$ and their estimated MSEs, denoted as $\{\hat{x}_m, \hat{v}_m^{\mathrm{x}}\}_{m=1}^{M}$, over the input domain $m = 1, 2, \ldots, M$, as indicated in (A1-4, 5). Then, it computes the posterior estimates and their MSEs, $\{\hat{z}_n, \hat{v}_n^{\mathrm{z}}\}_{n=1}^{N}$, based on the obtained beliefs and the given observations $\{y_n\}_{n=1}^{N}$, as shown in (A1-6, 7). Specifically, assuming a Gaussian modeling $\overline{z}_n \sim \mathcal{CN}(z_n, \overline{v}_n^{\mathrm{z}})$ in conformity with the central limit theorem (CLT), these quantities are given as the mean and variance of the following local posterior distribution:

$$\begin{aligned} p_{\mathsf{z}_n|\mathsf{y}_n,\overline{\mathsf{z}}_n,\overline{\mathsf{v}}_n^{\mathrm{z}}}(z_n \mid y_n, \overline{z}_n, \overline{v}_n^{\mathrm{z}}) &= \frac{p_{\mathsf{z}_n,\mathsf{y}_n|\overline{\mathsf{z}}_n,\overline{\mathsf{v}}_n^{\mathrm{z}}}(z_n, y_n \mid \overline{z}_n, \overline{v}_n^{\mathrm{z}})}{p_{\mathsf{y}_n|\overline{\mathsf{z}}_n,\overline{\mathsf{v}}_n^{\mathrm{z}}}(y_n \mid \overline{z}_n, \overline{v}_n^{\mathrm{z}})} \\ &\propto p_{\mathsf{z}_n|\overline{\mathsf{z}}_n,\overline{\mathsf{v}}_n^{\mathrm{z}}}(z_n \mid \overline{z}_n, \overline{v}_n^{\mathrm{z}}) \cdot p_{\mathsf{y}_n|\mathsf{z}_n,\overline{\mathsf{z}}_n,\overline{\mathsf{v}}_n^{\mathrm{z}}}(y_n \mid z_n, \overline{z}_n, \overline{v}_n^{\mathrm{z}}) \end{aligned}$$

$$\overset{\text{(a)}}{=} \frac{1}{\pi \overline{v}_n^{\mathsf{z}}} \exp\left[-\frac{|z_n - \overline{z}_n|^2}{\overline{v}_n^{\mathsf{z}}}\right] \cdot p_{\mathsf{y}_n|\mathsf{z}_n}(y_n \mid z_n), \tag{4}$$

wherein step (a) follows from the statistical independency between y_n and $\{\overline{z}_n, \overline{v}_n^{\mathsf{z}}\}$, and $p_{\mathsf{y}_n|\mathsf{z}_n}(y_n \mid z_n)$ is given in (3). Finally, the corresponding feedback information from each observation, $\{s_n, \psi_n\}_{n=1}^{N}$, is calculated in (A1-8, 9).

2) Inner Module: In a dual manner, the inner module first constructs beliefs and their estimated MSEs for $\boldsymbol{x}$, denoted by $\{\overline{x}_m, \overline{v}_m^{\mathsf{x}}\}_{m=1}^{M}$, by combining the feedback components over the output domain $n = 1, 2, \ldots, N$, as described in (A1-10, 11). Then, using the prior distribution $p_{\mathsf{x}_m}(\cdot)$ and the CLT-based Gaussian modeling $\overline{x}_m \sim \mathcal{CN}(x_m, \overline{v}_m^{\mathsf{x}})$, the mean and variance of the posterior distribution are obtained as

$$\begin{aligned} & p_{\mathsf{x}_m|\overline{\mathsf{x}}_m,\overline{\mathsf{v}}_m^{\mathsf{x}}}(x_m \mid \overline{x}_m, \overline{v}_m^{\mathsf{x}}) = \frac{p_{\mathsf{x}_m,\overline{\mathsf{x}}_m|\overline{\mathsf{v}}_m^{\mathsf{x}}}(x_m, \overline{x}_m \mid \overline{v}_m^{\mathsf{x}})}{p_{\overline{\mathsf{x}}_m|\overline{\mathsf{v}}_m^{\mathsf{x}}}(\overline{x}_m \mid \overline{v}_m^{\mathsf{x}})} \\ & \propto p_{\mathsf{x}_m|\overline{\mathsf{v}}_m^{\mathsf{x}}}(x_m \mid \overline{v}_m^{\mathsf{x}}) \cdot p_{\overline{\mathsf{x}}_m|\mathsf{x}_m,\overline{\mathsf{v}}_m^{\mathsf{x}}}(\overline{x}_m \mid x_m, \overline{v}_m^{\mathsf{x}}) \\ & \overset{\text{(b)}}{=} p_{\mathsf{x}_m}(x_m) \cdot \frac{1}{\pi \overline{v}_m^{\mathsf{x}}} \exp\left[-\frac{|x_m - \overline{x}_m|^2}{\overline{v}_m^{\mathsf{x}}}\right], \end{aligned} \tag{5}$$

where step (b) follows from the statistical independency between x_m and $\overline{v}_m^{\mathsf{x}}$. The resulting statistics, calculated in (A1-12, 13), serve as the updated tentative estimates of x_m.

3) Onsager Correction: A naive construction of the output belief $\overline{z}_n = \sum_{m=1}^{M} a_{n,m}\hat{x}_m$ in (A1-5) would correlate belief errors $(\overline{z}_n - z_n)$ and $(\overline{x}_m - x_m)$ with themselves across successive iterations, thereby violating their asymptotic Gaussianity in the large-system limit, even under the ideal large-scale i.i.d. measurements. The *Onsager correction* term $\overline{v}_n^{\mathsf{z}}(t)s_n(t-1)$ in (A1-5) subtracts the self-feedback components arising from the algorithm's iterative structure, thereby statistically decoupling consecutive iteration steps.

B. Annealed Discrete Denoiser (ADD) for Input Signal

In the particular case where $\boldsymbol{x}$ is defined over a finite alphabet set $\mathcal{X}$, a specific expression of (A1-12) is obtained—often refereed to as the *Bayes-optimal denoiser (BOD)* in its functional form [21]—by substituting (2) into (5):

$$\begin{aligned} \mathbb{E}_{\mathsf{x}_m}\left[x_m \mid \overline{x}_m, \overline{v}_m^{\mathsf{x}}\right] &= \sum_{\chi \in \mathcal{X}} \chi \cdot \frac{P_{\mathsf{x}_m}[\chi] e^{-\frac{|\overline{x}_m - \chi|^2}{\overline{v}_m^{\mathsf{x}}}}}{\sum_{\chi' \in \mathcal{X}} P_{\mathsf{x}_m}[\chi'] e^{-\frac{|\overline{x}_m - \chi'|^2}{\overline{v}_m^{\mathsf{x}}}}} \\ = \sum_{\chi \in \mathcal{X}} \chi \cdot \frac{e^{u_\chi(\overline{x}_m; \overline{v}_m^{\mathsf{x}})}}{\sum_{\chi' \in \mathcal{X}} e^{u_{\chi'}(\overline{x}_m; \overline{v}_m^{\mathsf{x}})}} &= \sum_{\chi \in \mathcal{X}} \chi \cdot \zeta\left(u_\chi\left(\overline{x}_m; \overline{v}_m^{\mathsf{x}}\right)\right), \end{aligned} \tag{6}$$

where we define the log posterior PDF and softmax function respectively as $u_\chi(x; v) \triangleq \ln P_{\mathsf{x}_m}[\chi] - |x - \chi|^2/v$ and $\zeta(\cdot)$.

As described in the previous subsection, Alg. 1 has a mechanism to analytically estimate the reliability (*i.e.*, variance) of $\overline{x}_m$ in (A1-11), denoted as $\overline{v}_m^{\mathsf{x}}$ in (A1-10). However, the estimated MSE $\overline{v}_m^{\mathsf{x}}$ precisely characterizes the statistical behavior of $\overline{x}_m$ *only* under two strong assumptions: an infinitely large-sized, and an ideally randomized $\boldsymbol{A}$. In the presence of severe modeling errors, such as highly correlated or ill-conditioned measurements, these estimated second-order statistics fail to capture the stochastic deviations of beliefs. Consequently, an overly underestimated $\overline{v}_m^{\mathsf{x}}$ causes the softmax function to concentrate an excessively large probability mass on a single

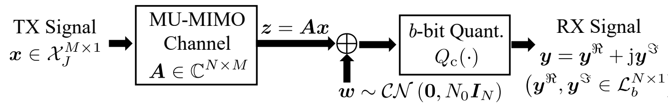


Fig. 1. The assumed MU-MIMO system with b-bit quantization ADCs.

point, leading to premature hard decisions and the propagation of erroneous estimates. Such error propagation severely hinders GAMP from converging to the correct solutions, resulting in significantly degraded performance.

One possible remedy for this issue is to replace the unreliable variance parameters $\{\overline{v}_m^{\mathsf{x}}(t)\}_{t=1}^{T}$ with *inverse temperature* parameters $\{\beta(t)\}_{t=1}^{T}$, whereby we compute $\hat{x}_m(t+1), \hat{v}_m(t+1)$ in (A1-12, 13) using

$$\hat{x}_m(t+1) = \sum_{\chi \in \mathcal{X}} \chi \cdot \zeta\left(u_\chi\left(\overline{x}_m(t); 1/\beta(t)\right)\right), \tag{7a}$$

$$\begin{aligned} & \hat{v}_m^{\mathsf{x}}(t+1) \\ & = \sum_{\chi \in \mathcal{X}} |\chi|^2 \cdot \zeta\left(u_\chi\left(\overline{x}_m(t); 1/\beta(t)\right)\right) - |\hat{x}_m(t+1)|^2, \end{aligned} \tag{7b}$$

We refer to (7) as the *annealed discrete denoiser (ADD)* [20], in analogy to BOD in (6). This formulation can be viewed as a natural extension of the existing belief scaling method for SLM-based discrete-valued signal estimation [18]–[20] to the more general GLM-based settings. The inverse temperature parameters $\{\beta(t)\}_{t=1}^{T}$ control the degree of nonlinearity—or equivalently, the *softness*—of ADD across iterations, and should be designed to prevent overly optimistic convergence and the resulting error propagation in GAMP.

IV. Performance Analysis

We evaluate the performance of the ADD-aided GAMP under an uplink MU-MIMO signal detection scenario, where the BS receiver is equipped with low-resolution ADCs.

A. System Model

Consider an uplink MU-MIMO system consisting of M user equipments (UEs), each equipped with a single transmit (TX) antenna, and one BS equipped with N receive (RX) antennas in a uniform linear array (ULA) pattern. In this setting, $x_m \triangleq [\boldsymbol{x}]_m$ in (1) represents a TX symbol from the m-th UE, which is assumed to be uniformly drawn from the J-ary quadrature amplitude modulation (QAM) constellation:

$$\mathcal{X} = \mathcal{X}_J \triangleq \left\{c_1 + \mathrm{j}c_2 \,\middle|\, c_1, c_2 \in \left\{\pm c, \pm 3c, \cdots, \pm\left(\sqrt{J} - 1\right)c\right\}\right\}, \tag{8}$$

with $c \triangleq \sqrt{3E_{\mathrm{s}}/\left(2(J-1)\right)}$ denoting a constant to normalize the average TX energy to $E_{\mathrm{s}}\,(=\sigma_{\mathsf{x}}^2)$.

The TX signal $\boldsymbol{x} \in \mathcal{X}_J^{M\times 1}$ is spatially multiplexed through channel matrix $\boldsymbol{A} \in \mathbb{C}^{N\times M}$ before reaching the BS. To evaluate the benchmark performance, we construct $\boldsymbol{A}$ based on the exponentially decaying Kronecker correlation model [22], [23], using an RX correlation matrix $\boldsymbol{R}_{\mathrm{RX}} \in \mathbb{C}^{N\times N}$ as

$$\boldsymbol{A} = \boldsymbol{R}_{\mathrm{RX}}^{1/2}\boldsymbol{G}, \quad [\boldsymbol{R}_{\mathrm{RX}}]_{n,n'} = \rho^{|n-n'|}, \tag{9}$$

where $\boldsymbol{G} \in \mathbb{C}^{N\times M}$ represents the small-scale fading coefficients and consists of i.i.d. entries following $\mathcal{CN}(0,1)$, while $\rho \in [0,1]$ denotes the correlation coefficient between adjacent antenna elements at the BS. Note that when $\rho = 0$, $\boldsymbol{A}$ reduces to the standard i.i.d. Gaussian matrix.

The BS observes the output signal $\boldsymbol{z}$ corrupted by an additive white Gaussian noise (AWGN) vector $\boldsymbol{w}$ with a power spectral density $N_0 (= \sigma_{\rm w}^2)$, after feeding it through a b-bit complex-valued quantizer $Q_{\rm c}$, as

$$y_n = \underbrace{Q_{\rm c}(z_n + w_n)}_{=g(z_n, w_n) \text{ in (1)}} \triangleq Q\left(z_n^{\Re} + w_n^{\Re}\right) + {\rm j}Q\left(z_n^{\Im} + w_n^{\Im}\right). \quad (10)$$

The mapping $Q : \mathbb{R} \to \mathcal{L}_b \triangleq \{l_1, l_2, \cdots, l_{2^b}\}$ is defined as a real-valued scalar quantizer satisfying $Q(a) = l_k, \forall a \in (\tau_k, \tau_{k+1}]$ for $k = 1, 2, \ldots, 2^{b+1}$, where $\mathcal{L}_b$ and $\mathcal{T}_b \triangleq \{\tau_1, \tau_2, \cdots, \tau_{2^b+1}\}$, (with $-\infty = \tau_1 < \tau_2 < \cdots < \tau_{2^b} < \tau_{2^b+1} = \infty$) respectively denoting the sets of quantization labels and thresholds. In our simulation, both $\mathcal{L}_b$ and $\mathcal{T}_b$ are obtained using Lloyd-Max algorithm [24], assuming $z_n + w_n \sim \mathcal{CN}(0, ME_{\rm s} + N_0)$ according to the CLT.

B. Tentative Estimates of Output Signal

Given the system model described in the previous subsection, one can obtain the specific expressions of (A1-6, 7). Owing to the independence of the real and imaginary parts in the quantization operation, the likelihood part[3] in (4) can be separated as $P_{{\sf y}_n|{\sf z}_n}[l_i + {\rm j}l_j \mid z_n] = P_{{\sf y}_n^{\Re}|{\sf z}_n^{\Re}}\left[l_i \mid z_n^{\Re}\right] \cdot P_{{\sf y}_n^{\Im}|{\sf z}_n^{\Im}}\left[l_j \mid z_n^{\Im}\right]$, where each factor is given by

$$P_{{\sf y}_n^{\mathfrak{A}}|{\sf z}_n^{\mathfrak{A}}}\left[l_k \mid z_n^{\mathfrak{A}}\right] = \mathbb{P}\left\{\tau_k < z_n^{\mathfrak{A}} + w_n^{\mathfrak{A}} \le \tau_{k+1}\right\}$$
$$= \mathit{\Phi}\left(\frac{\tau_{k+1} - z_n^{\mathfrak{A}}}{\sqrt{N_0/2}}\right) - \mathit{\Phi}\left(\frac{\tau_k - z_n^{\mathfrak{A}}}{\sqrt{N_0/2}}\right), \quad \mathfrak{A} \in \{\Re, \Im\}. \quad (11)$$

Hence, the mean and variance of (4) are respectively computed as [25]

$$\hat{z}_n = \overline{z}_n - \frac{\overline{v}_n^{\rm z}}{\sqrt{2(N_0 + \overline{v}_n^{\rm z})}}\left[f_i\left(\overline{z}_n^{\Re}, \overline{v}_n^{\rm z}\right) + {\rm j}f_j\left(\overline{z}_n^{\Im}, \overline{v}_n^{\rm z}\right)\right], \quad (12)$$
$$\hat{v}_n^{\rm z} = \overline{v}_n^{\rm z} - \frac{(\overline{v}_n^{\rm z})^2 \cdot \left(F_i\left(\overline{z}_n^{\Re}, \overline{v}_n^{\rm z}\right) + F_j\left(\overline{z}_n^{\Im}, \overline{v}_n^{\rm z}\right)\right)}{2(N_0 + \overline{v}_n^{\rm z})} - |\hat{z}_n - \overline{z}_n|^2,$$

where we define

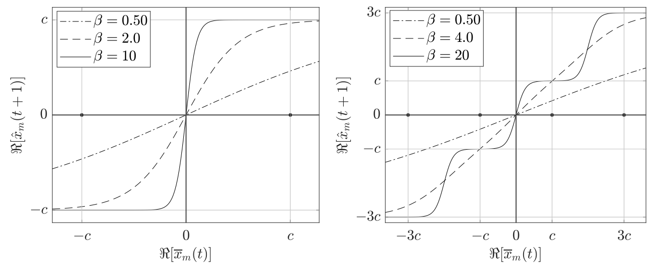
$$f_k(\lambda, \omega) \triangleq \frac{\phi\left(\theta_{k+1}(\lambda, \omega)\right) - \phi\left(\theta_k(\lambda, \omega)\right)}{\mathit{\Phi}\left(\theta_{k+1}(\lambda, \omega)\right) - \mathit{\Phi}\left(\theta_k(\lambda, \omega)\right)}, \quad (13)$$
$$F_k(\lambda, \omega) \triangleq \frac{\theta_{k+1}(\lambda, \omega)\phi\left(\theta_{k+1}(\lambda, \omega)\right) - \theta_k(\lambda, \omega)\phi\left(\theta_k(\lambda, \omega)\right)}{\mathit{\Phi}\left(\theta_{k+1}(\lambda, \omega)\right) - \mathit{\Phi}\left(\theta_k(\lambda, \omega)\right)},$$

for $\theta_k(\lambda, \omega) \triangleq (\tau_k - \lambda)/\sqrt{(N_0 + \omega)/2}$.

C. Inverse Temperature Parameter Design for ADD

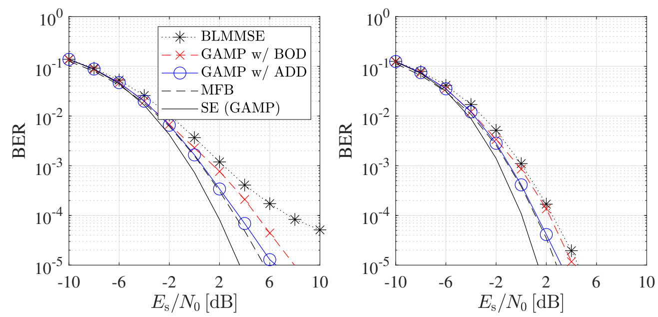
This subsection presents the design of the inverse temperature parameters $\{\beta(t)\}_{t=1}^{T}$ of ADD in (7). For intuitive understanding, Fig. 2 illustrates the mapping relationship between $\overline{x}_m(t)$ and $\hat{x}_m(t+1)$ in (7a) for different values of β, under the assumption that $P_{{\sf x}_m}[\chi] = 1/J, \forall \chi \in \mathcal{X}_J$ in the considered system. It can be observed that a larger β induces a sharper behavior of the softmax function, thereby producing more confident (*i.e.*, hard-valued) outputs. In our simulation,

[3]Since $y_n^{\Re}$ and $y_n^{\Im}$ take discrete values on $\mathcal{L}_b$, the PDF $p_{{\sf y}_n|{\sf z}_n}(y_n \mid z_n)$ in (4) is replaced with the PMF $P_{{\sf y}_n|{\sf z}_n}[y_n \mid z_n]$ [13], [25].



(a) $J = 4$. (b) $J = 16$.

Fig. 2. The mapping dynamics of ADD according to the inverse temperature parameter $\beta \in \{0.50, 2.0, 10\}$.



(a) $b = 2$. (b) $b = 3$.

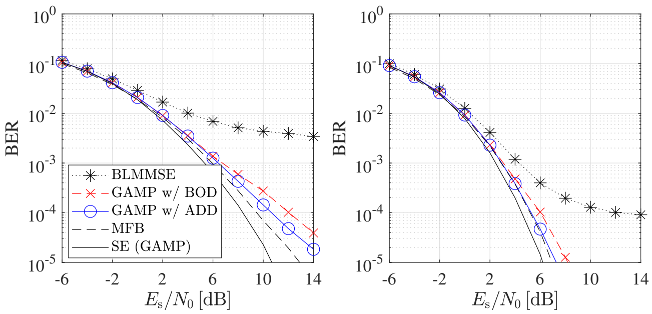
Fig. 3. BER performance comparison under a finite-sized i.i.d. measurement ($\rho = 0.0$) with $(M, N) = (4, 16)$ and $J = 4$.

as a simple design rule, the inverse temperature parameter is linearly scheduled as $\beta(t) = d \cdot (t/T)$ across iteration steps, where $d (> 0)$ denotes the target inverse temperature at the final iteration ($t = T$) and is determined empirically through preliminary simulations. This monotonically increasing strategy prevents unreliable beliefs from generating erroneous hard decisions in the early iterations, while promoting correct convergence in the later iterations [18], [19].

D. BER Performance

We numerically evaluate the bit error rate (BER) performance of Alg. 1 using BOD and ADD, respectively denoted as "GAMP w/ BOD" and "GAMP w/ ADD." For comparison, we also include the following benchmark schemes: 1) "Bussgang LMMSE (BLMMSE)," which applies conventional LMMSE filtering after approximately reformulating (1) into an SLM via Bussgang decomposition [26], and 2) "matched filter bound (MFB)," which represents the absolute performance limit achieved only when perfect knowledge of $\boldsymbol{x}$ is available to Alg. 1 as $\{\hat{x}_m(T)\}_{m=1}^{M}$ [27]. The total number of iterations T for Alg. 1 is set to 32, and typical belief damping [22] is applied to (A1-12) and (A1-13) with a damping coefficient of 0.50.

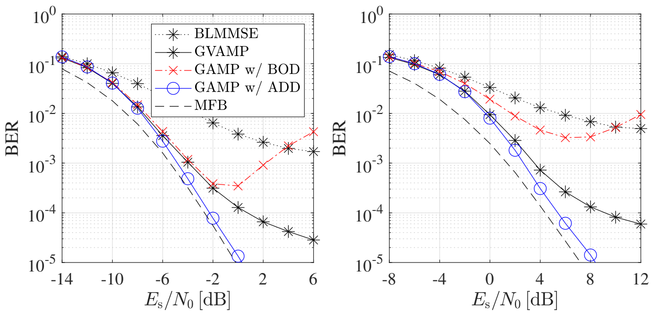
We first show the results under finite-sized i.i.d. measurement conditions ($\rho = 0.0$). Figs. 3 and 4 respectively correspond to the 4-QAM and 16-QAM cases, each obtained for $(M, N) = (4, 16)$ and $(4, 32)$, with $d = 4$ and 20. For reference, the analytically predicted performance of GAMP with BOD, evaluated via state evolution (SE) [2], is also plotted as "SE (GAMP)," serving as a characterization of its asymptotic performance in the large-system limit.



(a) $b = 2$.

(b) $b = 3$.

Fig. 4. BER performance comparison under a finite-sized i.i.d. measurement ($\rho = 0.0$) with $(M, N) = (4, 32)$ and $J = 16$.



(a) $b = 2,\ J = 4$.

(b) $b = 3,\ J = 16$.

Fig. 5. BER performance comparison under a finite-sized correlated measurement ($\rho = 0.80$) with $(M, N) = (16, 64)$.

Due to the severely finite-sized system configuration, the instantaneous dynamics of beliefs in GAMP inevitably deviate from the ideal Gaussian behavior, resulting in a noticeable performance degradation of "GAMP w/ BOD" relative to "SE (GAMP)" across all results. This degradation becomes even more pronounced in the coarse quantization ($b = 2$) cases shown in Figs. 3(a) and 4(a), corroborating the empirical observation that an increased estimation burden amplifies algorithmic sensitivity to finite-size effects. It is also observed that such finite-size effects render the BOD—naturally arising within the Bayesian framework [2], [19], [25] and grounded on the large-system assumption—no longer the optimal choice even under the i.i.d. conditions. In contrast, the deliberately designed "GAMP w/ ADD" successfully regulates the iterative convergence, thereby effectively mitigating the stochastic emergence of erroneous hard decisions. As shown in Fig. 3(a), "GAMP w/ ADD" achieves a gain of about 2 dB against "GAMP w/ BOD" at $\mathrm{BER} = 10^{-5}$, and closely approaches the finite-size-aware performance limit "MFB."

Next, Fig. 5 presents the BER performance when $\boldsymbol{A}$ exhibits severe inter-element correlation ($\rho = 0.80$), with $(M, N) = (16, 64)$. Figs. 5(a) and 5(b) correspond to $(b, J) = (2, 4)$ and $(3, 16)$, respectively, where d is set to 10 and 20. As a state-of-the-art MPA-based signal estimator for GLMs, the performance of generalized vector AMP (GVAMP) [1] using BODs is also included, with $T = 32$. Despite employing powerful iterative LMMSE filtering with singular value decomposition (SVD)-based preprocessing, "GVAMP" fails to achieve $\mathrm{BER} = 10^{-5}$ due to violations of both the large-system assumption ($M, N \gg 10^2$) and the theoretically required bi-unitary invariance of $\boldsymbol{A}$ [10], [12] in (9). Remarkably, "GAMP w/ ADD" reliably achieves $\mathrm{BER} = 10^{-5}$ regardless of modulation order J, requiring neither SVD nor iterative matrix-inversion operation—even far beyond the theoretically validated i.i.d. ($\rho = 0$) regime. In stark contrast, "GAMP w/ BOD" fails to achieve even $\mathrm{BER} = 10^{-4}$ in both figures. The next subsection visualizes the corresponding changes in the belief dynamics for both schemes, thereby elucidating the operating principle of ADD.

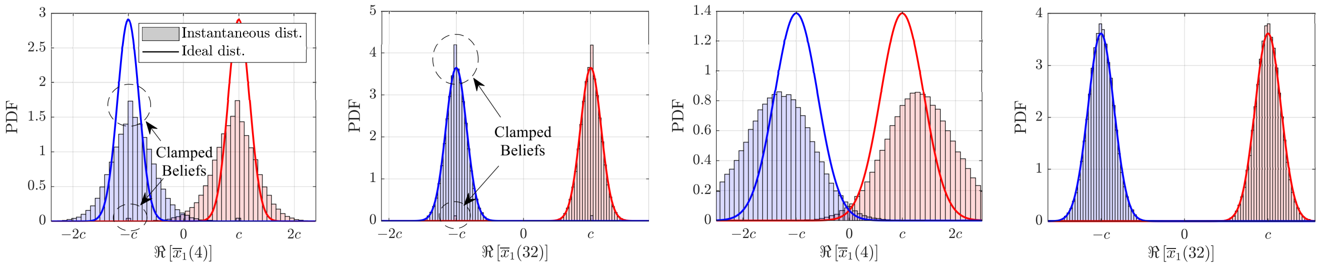
E. Iterative Convergence Analysis

Finally, we analyze the changes in belief dynamics induced by the use of ADD under the same system configuration as in Fig. 5(a), fixed at $E_s/N_0 = 5$ dB. Fig. 6 shows the empirical distributions of $\Re\left[\overline{x}_1(4)\right]$ and $\Re\left[\overline{x}_1(32)\right]$ obtained from 10^4 samples ("Instantaneous dist."), compared with the assumed Gaussian distribution with mean $\pm c$ and variance given by the empirical average of $\overline{v}_1(t)$ over realizations ("Ideal dist."), for both BOD (Figs. 6(a), 6(b)) and ADD (Figs. 6(c), 6(d)). We also illustrate in Fig. 7 the iterative convergence of Alg. 1 in terms of both BER and MSE of $\boldsymbol{z}$, the latter being defined as $\|\hat{\boldsymbol{z}}(t) - \boldsymbol{z}\|_2^2/N$ with $\hat{\boldsymbol{z}}(t) \triangleq [\hat{z}_1(t), \hat{z}_2(t), \ldots, \hat{z}_N(t)]^{\mathsf{T}} \in \mathbb{C}^{N\times 1}$. The MSE values exceeding 0 dB are clipped at every realization prior to averaging.

Due to the severe model mismatches, GAMP with the inherent BOD fails to correctly capture the actual belief dynamics, as evidenced by the deviation of the histogram from the Gaussian curves in Fig. 6(a). The resulting erroneous hard-valued estimates, caused by the overly confident behavior of the inner module based on incorrect $\overline{v}_m^{\mathrm{x}}$, can occasionally trigger fatal divergence of the outer module processing in subsequent iterations. This failure manifests as abnormal masses clamped at the output boundaries of BOD, *i.e.*, $\Re[\overline{x}_1(t)] = \pm c$, in Figs. 6(a) and 6(b), thereby violating the principled Gaussian assumption of belief dynamics. Despite such difficulties arising from the nonlinear nature of g in GLMs, GAMP with ADD effectively suppresses the generation of incorrect decisive estimates and completely prevents the divergence of Alg. 1, as demonstrated by the significantly improved MSE trajectory in Fig. 7. Consequently, ADD successfully restores the ideal Gaussianity of the histogram at $t = 32$ in Fig. 6(d), underscoring the exceptional effectiveness of the belief scaling strategy as a promising remedy for the model mismatches of GAMP, even under the GLMs.

V. Conclusion

We proposed a performance improvement strategy for GAMP to achieve robust GLM-based discrete-valued signal recovery under realistic system constraints. The fundamental operating principle of GAMP and other MPAs tailored for GLMs critically hinges on the idealized, unrealistic assumptions, leading to their inevitable performance degradation in practical usage when naively applied. The proposed scheme, which simply replaces the inherent BOD with ADD and manipulates its annealing rate via the inverse temperature, enabled GAMP to dramatically surpass the existing methods with minimal complexity and approach the performance



(a) GAMP w/ BOD, $t = 4$.
(b) GAMP w/ BOD, $t = 32$.
(c) GAMP w/ ADD, $t = 4$.
(d) GAMP w/ ADD, $t = 32$.

Fig. 6. Empirical distribution of $\Re[\overline{x}_1(t)]$ with comparison to the ideal Gaussian distribution , corresponding to Fig. 5(a) at $E_s/N_0 = 5$ dB.

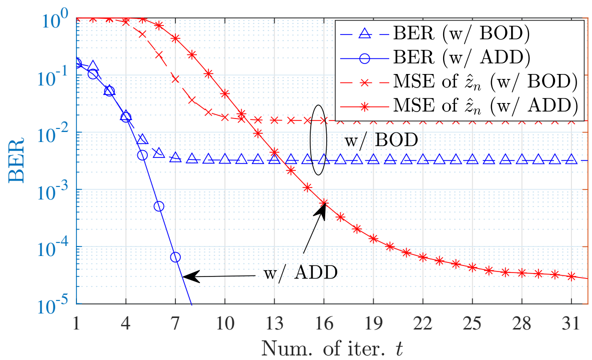


Fig. 7. Trajectories of BER and MSE of output signal $\boldsymbol{z}$ over iteration steps achieved by Alg. 1 corresponding to Fig. 5(a) at $E_s/N_0 = 5$ dB.

limit, even under the combined difficulties of theoretically intractable GLM-induced belief distortions, finite-size effects, and severe inter-element correlation of $\boldsymbol{A}$.

Acknowledgement

This work was supported in part by JSPS KAKENHI Grant Number 25H01111, and in part by MIC/FORWARD under Grant JPMI240710001.

References

[1] P. Schniter, S. Rangan, and A. K. Fletcher, "Vector approximate message passing for the generalized linear model," in *Proc. Conf. Rec. Asilomar Conf. Signals Syst.*, 2016, pp. 1525–1529.
[2] S. Rangan, "Generalized approximate message passing for estimation with random linear mixing," in *Proc. 2011 IEEE Int. Symp. Inf. Theory (ISIT)*, July 2011, pp. 2168–2172.
[3] 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.
[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] 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, 2011.
[6] A. Javanmard and A. Montanari, "State evolution for general approximate message passing algorithms, with applications to spatial coupling," *Inf. Inference: A Journal of the IMA*, vol. 2, no. 2, pp. 115–144, 2013.
[7] J. Barbier, F. Krzakala, N. Macris, L. Miolane, and L. Zdeborová, "Optimal errors and phase transitions in high-dimensional generalized linear models," *Proc. Natl. Acad. Sci.*, vol. 116, no. 12, pp. 5451–5460, 2019. [Online]. Available: https://www.pnas.org/doi/abs/10.1073/pnas.1802705116
[8] P. Pascual Cobo, K. Hsieh, and R. Venkataramanan, "Bayes-optimal estimation in generalized linear models via spatial coupling," *IEEE Trans. Inf. Theory*, vol. 70, no. 11, pp. 8343–8363, 2024.
[9] C. Rush and R. Venkataramanan, "Finite sample analysis of approximate message passing algorithms," *IEEE Trans. Inf. Theory*, vol. 64, no. 11, pp. 7264–7286, 2018.
[10] C. Cademartori and C. Rush, "A non-asymptotic analysis of generalized vector approximate message passing algorithms with rotationally invariant designs," *IEEE Trans. Inf. Theory*, vol. 70, no. 8, pp. 5811–5856, 2024.
[11] F. Tian, L. Liu, and X. Chen, "Generalized memory approximate message passing for generalized linear model," *IEEE Trans. Signal Process.*, vol. 70, pp. 6404–6418, 2022.
[12] P. Pandit, M. Sahraee-Ardakan, S. Rangan, P. Schniter, and A. K. Fletcher, "Inference with deep generative priors in high dimensions," *IEEE J. Sel. Areas Inf. Theory*, vol. 1, no. 1, pp. 336–347, 2020.
[13] U. S. Kamilov, V. K. Goyal, and S. Rangan, "Message-passing dequantization with applications to compressed sensing," *IEEE Trans. Signal Process.*, vol. 60, no. 12, pp. 6270–6281, 2012.
[14] T. Takahashi, H. Iimori, K. Ando, K. Ishibashi, S. Ibi, and G. T. F. de Abreu, "Bayesian receiver design via bilinear inference for cell-free massive MIMO with low-resolution ADCs," *IEEE Trans. Wireless Commun.*, vol. 22, no. 7, pp. 4756–4772, 2023.
[15] C. Yang, X. Liu, Y. L. Guan, and R. Liu, "Fast GAMP algorithm for nonlinearly distorted OFDM signals," *IEEE Commun. Lett.*, vol. 25, no. 5, pp. 1682–1686, 2021.
[16] S. Rangan, P. Schniter, and A. K. Fletcher, "Vector approximate message passing," *IEEE Trans. Inf. Theory*, vol. 65, no. 10, pp. 6664–6684, 2019.
[17] L. Liu, S. Huang, and B. M. Kurkoski, "Memory AMP," *IEEE Trans. Inf. Theory*, vol. 68, no. 12, pp. 8015–8039, 2022.
[18] 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, Mar. 2019.
[19] 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. 2022 IEEE Int. Conf. Commun. (ICC)*, 2022, pp. 1300–1305.
[20] T. Furudoi, T. Takahashi, S. Ibi, and H. Ochiai, "Discrete-valued signal estimation via low-complexity message passing algorithm for highly correlated measurements," 2024. [Online]. Available: https://arxiv.org/abs/2411.07558
[21] K. Takeuchi, "Generalized approximate message-passing for compressed sensing with sublinear sparsity," *IEEE Trans. Inf. Theory*, vol. 71, no. 6, pp. 4602–4636, 2025.
[22] A. Chockalingam and B. S. Rajan, *Large MIMO Systems*. Cambridge University Press, 2014.
[23] 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.
[24] J. Max, "Quantizing for minimum distortion," *IRE Trans. Inf. Theory*, vol. 6, no. 1, pp. 7–12, 1960.
[25] C. Wen, C. Wang, S. Jin, K. Wong, and P. Ting, "Bayes-optimal joint channel-and-data estimation for massive MIMO with low-precision ADCs," *IEEE Trans. Signal Process.*, vol. 64, no. 10, pp. 2541–2556, 2016.
[26] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu, "Channel estimation and performance analysis of one-bit massive MIMO systems," *IEEE Trans. Signal Process.*, vol. 65, no. 15, pp. 4075–4089, 2017.
[27] S. Wu, L. Kuang, Z. Ni, J. Lu, D. Huang, and Q. Guo, "Low-complexity iterative detection for large-scale multiuser MIMO-OFDM systems using approximate message passing," *IEEE J. Sel. Top. Signal Process.*, vol. 8, no. 5, pp. 902–915, 2014.