AMP Inspired Antenna Activity and Signal Detection Algorithm for Generalized Spatial Modulated NOMA

Xiang Li, Yang Huang, Wei Heng$^{(\boxtimes)}$, Jing Wu, Ke Wang,
Gang Wang, and Yuan Zhang

National Mobile Communications Research Laboratory, Southeast University,
Nanjing 210096, People's Republic of China
{230159390, 213151630, wheng, 230169362, 230179413,
wanggang, 101010681}@seu.edu.cn

Abstract. The non-orthogonal multiple access technology has been considered as one of potential technologies for the next generation wireless network. Spatial modulation, which improves both spectral and energy efficiency at the same time, has found its potentials in NOMA system. Spatial modulation, together with multiple-input multiple-output technique, could maintain massive connections and provide low latency at the same time. But it also puts forward challenges for multi-user and signal detection. By exploiting the sparsity nature of generalized spatial modulation system, we formulate the active antenna and user signal detection into a general sparse linear-inverse problem. An approximate message passing based algorithm is proposed to detect the antenna activity and transmitted signal simultaneously in the uplink grant-free NOMA scenario. Expect maximum algorithm is utilized to learn the parameters of activity level and noise variance. Simulation results show that proposed scheme outperforms the CS based schemes over a wide range of SNR and sparsity level. Moreover, proposed algorithm achieves convergency in 15 iterations which makes it very practical.

Keywords: Spatial modulation · NOMA · Approximate message passing · Compressive sensing

1 Introduction

Next generation wireless network is expected to provide low latency and support massive connectivity with a large number of devices. To address these challenges, nonorthogonal multiple access (NOMA), which was proposed to deal with these challenges by efficiently using finite available bandwidth, has been regarded as one of the most promising technologies for the 5G network [1].

In NOMA system, nonorthogonal resources are allocated to different users rather than orthogonal resources distribution in conventional orthogonal multiple access. Therefore, the base station is able to support much more users in resource limited uplink scenario. So far, several NOMA schemes have been investigated. The power domain NOMA utilizes superposition coding at the transmitter and successive

J. Zheng et al. (Eds.): ADHOCNETS 2019, LNICST 306, pp. 262–275, 2019.
https://doi.org/10.1007/978-3-030-37262-0_22

interference cancellation at the receiver [2]. The code domain NOMA takes the forms of low-density spreading CDMA (LDS-CDMA) [3], sparse code multiple access (SCMA) [4] and so on. In order to reduce the control signaling overhead and latency, the grant-free NOMA system where active users transmit data at synchronized time slot without a complex request-grant procedure is investigated here.

Multiple-input multiple-output (MIMO) technology is getting increasing interests by using antennas on the terminals to achieve a better performance. However, a key challenge of future mobile communication network is to strike a compromise between spectral efficiency (SE) and energy efficiency (EE). Fueled by this consideration, the spatial modulation (SM) [5], which uses the spatial constellation to meet the demand of SE and EE, has been established as a promising transmission concept [6]. Figure 1 shows the configuration of different multi-antenna systems. The main distinguishing feature of SM is that it maps additional information on the SM constellation diagram. The generalized SM is a generalization of SM by taking advantage of the whole antenna array without the RF chain limitation. Therefore, the generalized SM fully uses the available antennas to improve SE and EE [7]. This unique characteristic allows the coexistence of high-rate devices and massive connections. Recent analytical and simulation studies have shown that SM outperforms many state-of-art MIMO schemes [8].

According to the statistics, the number of active users is usually much smaller than the number of supported users even in rush hours [9]. This coincides with the sparsity hypothesis of user activity in the NOMA system. Due to the sparsity nature of the SM signals, compressive sensing (CS) based detectors [10, 11] become competitive solutions with low complexity especially in the large-scale scenario.

Recently, approximate message passing (AMP) algorithm [12] was proposed by Donoho to solve CS problems. Despite its low complexity, AMP performs exactly the same as l_1 – norm minimization and it admits rigorous analysis based on the state evolution [13]. AMP has been extended to general linear mixing problems and widely used in various scenarios [14, 15].

Starting from the analysis of generalized spatial modulation system, we develop an AMP-based algorithm to detect the active antennas and the data modulated on each active antenna. To learn the unknown parameters, we present a detailed derivation of the expectation maximization (EM) algorithm iteratively. The simulation results show the effectiveness of proposed algorithm and it performs better than existing CS approaches and converges in 15 iterations.

Notation: Bold lower and upper-case symbols represent vectors and matrices, respectively. The superscript $(\cdot)^{\mathrm{T}}$ denotes the transpose operation and $\mathcal{N}(x;\theta,\phi)$ denotes that x is Gaussian distributed with mean θ and variance ϕ.

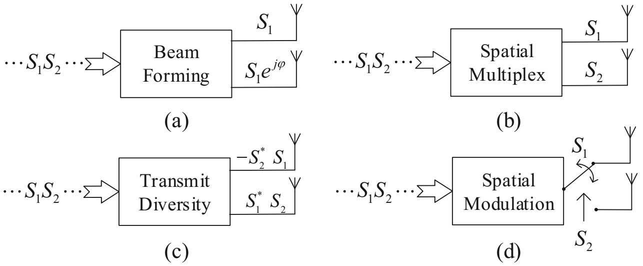

Fig. 1. Illustration of multiple antenna configurations: (a) beam forming (b) spatial multiplex (c) transmit diversity (d) spatial modulation

2 System Model

Here we consider an uplink grant-free NOMA system with one base station and K users of which S are active within one transmission slot. As shown in Fig. 2, the base station is equipped with M antennas and all antennas are set to spatial modulation mode. For simplicity, we assume each user has n_t antennas, however, our algorithm can be easily extended to a more general case. Then the total antenna at the user side is $N = K \times n_t$. The uplink NOMA system can be modelled as

$$\mathbf{y} = \sum_{k=1}^{K} \mathbf{G}_k \mathbf{x}_k + \mathbf{w} \tag{1}$$

where $\mathbf{y} \in \mathbb{C}^{M \times 1}$ is the receive signal at the base station and $\mathbf{G}_k \in \mathbb{C}^{M \times n_t}$ is the channel response matrix of user k. Here we assume the channel is quasi-static, i.e., the coefficients keep constant during one transmission slot. $\mathbf{x}_k \in \mathbb{C}^{M \times 1}$ represents the transmitted symbols of user k. Additive white gaussian noise (AWGN) vector $\mathbf{w} \in \mathbb{C}^{M \times V} \sim \mathcal{CN}(0, \sigma^2 \mathbf{I}_M)$ with $\mathbf{I}_M$ being the identity matrix of size $M \times M$. Then Eq. (1) can be rewritten as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w} \tag{2}$$

where $H = [G_1 \quad G_1 \ldots G_K]$ denotes the equivalent channel matrix. In the generalized spatial modulated NOMA scenario, transmitted signal $\mathbf{x}$ is generated by users and their antennas, which is defined as

$$\mathbf{x} = [\mathbf{x}_1 \quad \mathbf{x}_2 \ldots \mathbf{x}_K]^{\mathrm{T}} \tag{3}$$

Generally, not all the antennas are active at the same time. We assume $n_a \in [0, n_t]$ antennas are activated randomly at one time slot, then $\mathbf{x}_i$ can be expressed as

$$\mathbf{x}_i = \left[0, x_{p_1}, 0, \ldots, x_{p_{na}}, \ldots, 0\right] \tag{4}$$

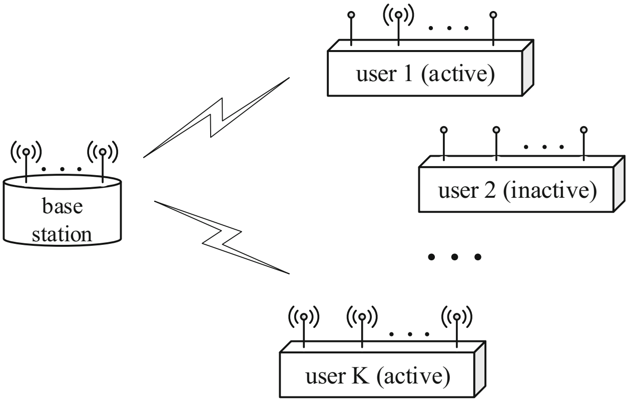

Fig. 2. Generalized spatial modulation in grant-free uplink NOMA system

where $p_j \in [1, n_t]$ is the index of active antenna and $j \in [1, n_a]$. x_{p_j} represents the symbol transmitted on the p_j th antenna which is chosen from a constellation set $\mathbf{\Theta} = \{\mathbf{\Theta}_1, \mathbf{\Theta}_2, \ldots, \mathbf{\Theta}_Q\}$, such as PSK or QAM. Then the sum rate is defined as bit per user.

$$R = \log_2 \binom{n_t}{n_a} + n_a \times \log_2 \|\mathbf{\Theta}\|_0 \tag{5}$$

According to [9], only a small number of users are active simultaneously and their antennas are activated randomly during transmission. We denote 0 for the antenna which is inactive. Then $\mathbf{x}$ takes the following form

$$\mathbf{x} = \left[\ldots, \underbrace{0, \ldots, 0,}_{\text{inactive user}} \ldots, \underbrace{x_{p_1}, 0, \ldots, x_{p_{n_a}},}_{\text{active user}} \ldots \right]^{\mathrm{T}} \tag{6}$$

For a more general case, users have different number of antennas and the number of active antennas is different for each user. Apart from that fact, users may change their spatial constellations between each transmission, so block-sparse hypothesis will not always hold. Active entries are not clustered either, so nearest neighbor sparsity methods do not work well. Therefore, the problem should be modeled as the typical sparse linear inverse problem naturally.

In the uplink grant-free NOMA, BS needs to know the antenna activity before decoding. Therefore, our goal is to estimate the support and value of nonzero element of $\mathbf{x}$ from $\mathbf{y}$. The sparsity level of each antenna is also unknown a prior. Without loss of generality, we assume the prior on each antenna is i.i.d, i.e., having the following marginal pdf

$$p(\mathbf{x}) = \prod_{n=1}^{N} p(x_n) = \prod_{n=1}^{N} [(1 - \lambda_n)\delta(x_n) + \lambda_n f(x_n)] \tag{7}$$

where $\lambda_n \in (0, 1)$ is the sparsity level, $\delta(x_n)$ is the Dirac delta function. It is worth noting that we specify an individual ratio for each antenna rather than a common one. This is one key feature for reconstruction in proposed algorithm. Transmitted symbols on each active antenna are chosen from the modulation constellation set $\mathbf{\Theta}$. Let $p_{n,q}$ represents the probability of transmitting Θ_q of the n^{th} user, then the distribution of nonzero entries can be written as

$$f(x_n) = \sum_{q=1}^{Q} p_{n,q}\delta\left(x_n - \mathbf{\Theta}_q\right) \tag{8}$$

The system considered here is assumed to be well synchronized in each transmission slot and inter-symbol interference is ignored.

3 Proposed Algorithm

3.1 AMP Algorithm

Inspired by approximate message passing algorithms and MAP inference, we detail the proposed detection in Algorithm I. Based on message passing algorithm, we decouple the estimation problem in Eq. (2) into scalar problems:

$$\begin{cases} \mathbf{y} = \mathbf{z} + \mathbf{w} \\ \mathbf{z} = \mathbf{H}\mathbf{x} \end{cases} \rightarrow \begin{cases} \gamma_1 = x_1 + w_1 \\ \cdots \\ \gamma_N = x_N + w_N \end{cases} \tag{9}$$

where the equivalent noise w_n asymptotically follows $\mathcal{CN}(w_n; 0, \phi_n)$. The value of γ_1 and ϕ_n are updated in each iteration. The posterior distribution of x_n is defined as

$$p(x_n|\gamma_n, \phi_n) = \frac{1}{Z(\gamma_n, \phi_n)} p(x_n)\mathcal{CN}(x_n; \gamma_n, \phi_n) \tag{10}$$

where

$$Z(\gamma_n, \phi_n) = \sum_{x_n \in \{\Theta, 0\}} p(x_n)\mathcal{CN}(x_n; \gamma_n, \phi_n) \tag{11}$$

is the normalizing factor and

$$p(x_n) = (1 - \lambda_n)\delta(x_n) + \lambda_n \sum_{q=1}^{Q} p_{n,q}\delta\left(x_n - \Theta_q\right) \tag{12}$$

```
Algorithm I    AMP based detection algorithm
(1) initialization
$\lambda_n^0 = \lambda_0,\ \hat{x}_n^0 = \lambda_0 \sum_{q=1}^{Q} p_{n,q}\Theta_q,\ \upsilon_n^0 = \lambda_0 \sum_{q=1}^{Q} p_{n,q}\left|\Theta_q\right|^2 - \left|\hat{x}_n^0\right|^2$
$Z_m^0 = y_m,\ V_m^0 = 0$
(2) main iteration
for $t = 1 \ldots T$
    for $m = 1:M$  (decoupling step)
        $V_m^t = \sum_{n=1}^{N} \left|H_{m,n}\right|^2 \upsilon_n^t$
        $Z_m^t = \sum_{n=1}^{N} H_{m,n}\hat{x}_n^t - V_m^t \frac{y_m - Z_m^{t-1}}{\sigma^2 + V_m^{t-1}}$
    for $n = 1:N$  (denoising step)
        $\phi_n^t = \left( \sum_{m=1}^{M} \frac{\left|H_{m,n}\right|^2}{\sigma^2 + V_m^{t-1}} \right)^{-1}$
        $\gamma_n^t = \hat{x}_n^t + \phi_n^t \sum_{m=1}^{M} \frac{H_{m,n}^*\left(y_m - Z_m^t\right)}{\sigma^2 + V_m^t}$
        cal $p\left(x_n^t \middle| \gamma_n^t, \phi_n^t\right)$ using (12)
        $\hat{x}_n^t = \sum_{x_n^t \in \{\mathbf{\Theta},0\}} x_n^t p\left(x_n^t \middle| \gamma_n^t, \phi_n^t\right)$
        $\upsilon_n^t = \sum_{x_n^t \in \{\mathbf{\Theta},0\}} \left|x_n^t\right|^2 p\left(x_n^t \middle| \gamma_n^t, \phi_n^t\right) - \left|\hat{x}_n^t\right|^2$
    update $\lambda_k$ and $\sigma^2$ using (18) and (24)
finally $\mathbf{r} = \left[\gamma_1^T, \gamma_2^T, \ldots, \gamma_n^T\right]$
(3) select the active antenna using antenna detection rule
in sectiong III-D
(4) decode symbols on each active antenna
```

From above, the estimates of mean and variance of x_n are

$$\begin{aligned} \hat{x}_n &= \sum_{x_n \in \{\Theta,0\}} x_n p(x_n|\gamma_n, \phi_n) \\ v_n &= \sum_{x_n \in \{\Theta,0\}} |x_n|^2 p(x_n|\gamma_n, \phi_n) - |\hat{x}_n|^2 \end{aligned} \tag{13}$$

The term $V_m^t \frac{y_m - Z_m^{t-1}}{\sigma^2 + V_m^{t-1}}$ in decoupling step is known as the *Onsager correction* which is the heart of the AMP [12]. Under large i.i.d. sub-Gaussian channel matrix configuration, Onsager correction ensures that the input of the denoiser can be modeled as

$$\mathbf{Z} = \mathbf{x} + \mathbf{n}, \text{ where } \mathbf{n} \sim \mathcal{CN}\left(\mathbf{0}, \frac{\mathbf{I}_N}{M} \|\mathbf{V}\|_2^2\right) \tag{14}$$

The gaussian distribution enables the denoiser to work efficiently.

From the detail implement of AMP, full knowledge of prior distribution and noise variance are needed, which is an impractical assumption. Therefore, we resort to EM algorithm to learn the unknown parameters. The EM algorithm we adopt here is an increment update rule [16], i.e., updating one element at a time while others remain fixed. EM increases the likelihood probability at each iteration, guaranteeing convergence to at least local maximum of the likelihood function $p(\mathbf{y}|\lambda_k, \sigma^2)$.

3.2 λ_k Update

Now we resort to EM algorithm to learn the user activity λ_n. Since user may change their spatial modulation at each transmission slot, we estimate λ_n element-wisely. Denoting the estimate parameters by λ_n^t at t^{th} iteration, EM updates can be expressed as

$$\lambda_{\mathrm{k}}^{t+1} = \underset{\lambda_k^k \in [0,1]}{\arg\max} \, \mathrm{E}\left\{\ln p\left(x_n^t | \lambda_n^t\right) | \mathbf{y}\right\} \tag{15}$$

where $\mathrm{E}\{\cdot\}$ denotes expectation conditioned on observation $\mathbf{y}$ and parameter λ_k^t. In order to obtain the maximum value, we differentiate Eq. (15) with respect to λ_k^t. and set it to zero:

$$\sum_{x_n\{\mathbf{\Theta},0\}} p\left(x_n^t|\mathbf{y}\right) \frac{\mathrm{d}}{\mathrm{d}\lambda_n^t} \ln p\left(x_n^t|\lambda_n^t\right) \tag{16}$$

where

$$p\left(x_n^t|\mathbf{y}\right) = p\left(x_n^t|\gamma_n^t, \phi_n^t\right)$$

$$\frac{\mathrm{d}}{\mathrm{d}\lambda_n^t} \ln\, p\left(x_n^t|\lambda_n^t\right) = \frac{\sum_{q=1}^{Q} p_{n,q}\delta\left(x_n^t - \mathbf{\Theta}_q\right) - \delta\left(x_n^t\right)}{\left(1-\lambda_n^t\right)\delta\left(x_n^t\right) + \lambda_t^t \sum_{q=1}^{Q} p_{n,q}\delta\left(x_n^t - \mathbf{\Theta}_q\right)}$$

$$= \begin{cases} \frac{1}{\lambda_n^t - 1} & x_n^t \notin \mathbf{\Theta} \\ \frac{1}{\lambda_n^t} & x_n^t \in \mathbf{\Theta} \end{cases} \tag{17}$$

Then λ_n^{t+1} can be obtained in a direct form

$$\lambda_n^{t+1} = \sum_{x_n \in \boldsymbol{\Theta}} p\left(x_n^t \middle| \gamma_n^t, \phi_n^t\right) \tag{18}$$

3.3 σ^2 Update

Then we derive the update rule for σ^2 given previous parameters. Note that $\mathbf{w}$ is independent of $\mathbf{x}$ and i.i.d, the joint pdf decouples into

$$p(\mathbf{x}, \mathbf{w}) = \prod_{n=1}^{N} p\left(w_m; \sigma^2\right) \tag{19}$$

so

$$\left(\sigma^2\right)^{t+1} = \underset{\sigma^2 > 0}{\arg\max} \sum_{n=1}^{N} \mathrm{E}\left\{\ln p\left(w_m; \sigma^2\right) | \mathbf{y}; \boldsymbol{\theta}^t\right\} \tag{20}$$

The maximizing value of σ^2 can be obtained by zeroing the derivative, i.e.,

$$\sum_{n=1}^{N} \int_{\sigma^2} p\left(\sigma^2 | \mathbf{y}; \boldsymbol{\theta}^t\right) \frac{\mathrm{d}}{\mathrm{d}\sigma^2} \ln p\left(w_m; \sigma^2\right) = 0 \tag{21}$$

where

$$\frac{\mathrm{d}}{\mathrm{d}\sigma^2} \ln p\left(w_m; \sigma^2\right) = \frac{1}{2}\left(\frac{\left(w_m\right)^2}{\left(\sigma^2\right)^2} - \frac{1}{\sigma^2}\right) \tag{22}$$

From (21) and (22), we have

$$\left(\sigma^2\right)^{t+1} = \frac{1}{N} \sum_{n=1}^{N} \int_{w_m} \left|w_m\right|^2 p\left(\sigma^2 | \mathbf{y}; \boldsymbol{\theta}^t\right) \tag{23}$$

since $w_m = y_m - z_m$, we have

$$\begin{aligned} \left(\sigma^2\right)^{t+1} &= \frac{1}{N} \sum_{n=1}^{N} \int_{w_m} \left|y_n - \hat{z}_n\right|^2 p\left(\sigma^2 | \mathbf{y}; \boldsymbol{\theta}^t\right) \\ &= \frac{1}{N} \sum_{n=1}^{N} \left(\left|y_n - \hat{z}_n\right|^2 - \mu_{zm}\right) \end{aligned} \tag{24}$$

Above $\hat{z}_n$ and μ_{2m} are the posterior mean and variance, which can be calculated by

$$\begin{aligned} z_m &= \sum_{n=1}^{N} H_{mn} x_n \mu_{zm} \\ &= \sum_{n=1}^{N} |H_{mn}|^2 \gamma_n \end{aligned} \tag{25}$$

3.4 CFAR Threshold

After several iterations, based on the estimated mean and variance, we design an adaptable threshold x_{TH} to detect the support of $\mathbf{r}$ with constant false alarm rate (CFAR)η. Support detection can be seen as the final layer. From Eq. (2) we have

$$\mathbf{r} = \mathbf{x} + \mathbf{w} \tag{26}$$

where $\mathbf{w} \sim \mathcal{CN}\left(\mathbf{0}, \gamma_{1\text{T}}^{-1}\mathbf{I}\right)$, and $\mathbf{r} \sim \mathcal{CV}\left(\mathbf{x}, \gamma_{1\text{T}}^{-1}\mathbf{I}\right)$. For a given CFAR η, x_{TH} can be derived as

$$x_{\text{TH}} = \sqrt{\gamma_{1\text{T}}}\Phi^{-1}\left(\frac{\eta+1}{2}\right) \tag{27}$$

where $\mathit{\Phi}^{-1}(\cdot)$ is the probit function of the standard normal distribution. Based on this threshold, missing detection can be calculated

$$\Pr(x_{\text{TH}}) = \int_{-x_{\text{TH}}}^{x_{\text{TH}}} p(x_n|\mathbf{r})dx \approx \lambda \int_{-x_{\text{TH}}}^{x_{\text{TH}}} \mathcal{N}(x; \theta, \mathbf{\Phi})dx \tag{28}$$

Once the threshold is selected, the support is detected for $|\mathbf{r}| \geq x_{\text{TH}}$. Then the original symbol pair is restored by applying MAP detection:

$$\begin{aligned} x_n &= \underset{x_e \in \Theta}{\arg\max}\, p\left(\Theta_q | x_n\right) \\ &= \underset{x_x \in \Theta}{\arg\max} \frac{p\left(x_n|\Theta_q\right)p\left(\Theta_q\right)}{\sum\limits_{x_e \in \Theta} p\left(x_n|\Theta_q\right)p\left(\Theta_q\right)} = \underset{x_x, = \Theta}{\arg\max}\, p\left(x_n|\Theta_q\right) \end{aligned} \tag{29}$$

3.5 Parameter Initialization

Since EM algorithm may converge into a local maximum or a saddle point, proper initialization of unknown parameters is essential. The sparsity level is initialized as $\lambda_0 = \frac{M}{N}\rho_{Prc}$ where ρ_{PTC} is the sparsity ratio achieved by Lasso PTC

$$\rho_{PTC} = \max_{a>0} \frac{1 - 2N/M[(1+a^2)\Phi(a) - a\phi(a)]}{1 - a^2 - 2[(1+a^2)\Phi(a) - a\phi(a)]} \tag{30}$$

Due to the fact that the active pdf is symmetric, the active mean is initialized as 0. Then σ^2 is initialized as $(\mathrm{sNR}+1)\|\mathrm{y}\|^2/M$ and SNR is set to 100 if there is no extra knowledge.

3.6 Computational Complexity

The computational complexity is evaluated in terms of floating-point operations (FPOs). The multiplication of a real number and a complex number require 2 FPOs and the multiplication of two complex numbers requires 6 FPOs. The value from operation $\mathcal{CN}(\cdot)$ is implemented by look-up table. In the main iteration of proposed algorithm, the computation of inner decoupling step requires (10 N + 2) FPOs and these computations need M iterations. The computation of denoising step needs $N(22M + 16\|\mathbf{\Theta}\|_0 + 9)$ FPOs. Equations (18) and (24) require $N\|\mathbf{\Theta}\|_0$ and (8 M + 10) $(N-1)+9$ FPOs, respectively. Therefore, the total computation required by proposed algorithm is $T\left(40MN - 6M + 19N + 17N\|\mathbf{\Theta}\|_0 - 1\right)$ FPOs. Modulation order is a factor that affect the computation burden especially when high-order modulations are used.

The computational complexity of proposed algorithm is dominated by matrix-vector multiplications in each iteration, i.e., $\mathcal{O}(MN)$. The number of iterations required to guarantee convergence is not large. From the simulation result shown in Sect. 4, it takes 15 layers to achieve its error floor. Therefore, this linear complexity of proposed algorithm is suitable for large scale antenna MIMO configurations which are encountered in next generation wireless communication system.

4 Performance Evaluation

In this section, we present the simulation results of proposed detection scheme. The base station is equipped with $M = 100$ receive antennas. The number of total supported antennas N is set to 250, thus the overload factor is 250%. The symbols transmitted on the active antenna are QPSK modulated. The channel is modeled as Rayleigh flat fading channel and the elements of channel matrix $\mathbf{H}$ are i.i.d, i.e., $H_{m,n} - \mathcal{CN}(0,1)$. CS based schemes, OMP, SP and CoSaMP are implemented as reference. The performance of user detection and SM demodulation. User detection performance is measured by antenna detection error which is defined as active antenna being detected as inactive. Symbol detection error rate is used to measure signal detection performance.

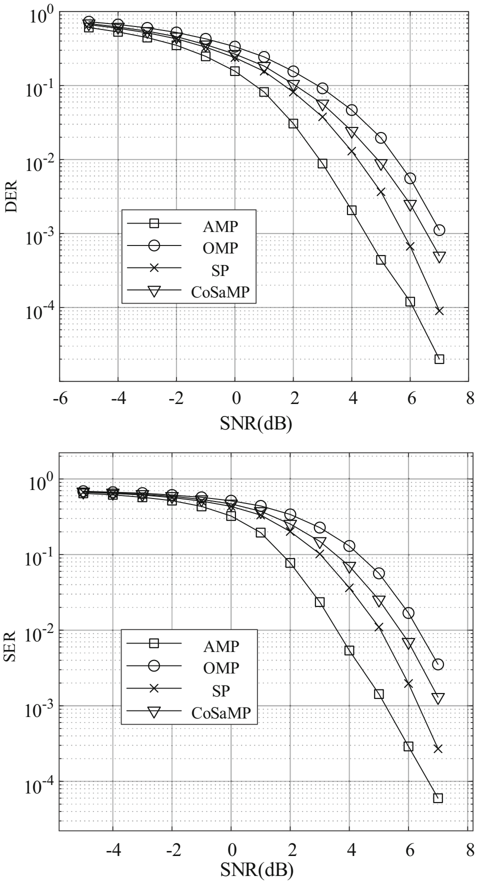

Fig. 3. DER and SER performance versus SNR

The detection error rate (DER) and symbol error rate (SER) of different algorithms is depicted as functions of SNR in Fig. 3. The number of active antennas is set to 10. When measuring the SER, the active antennas is assumed to be known. The proposed detector outperforms SP 1 dB in terms of SER and DER. It is worth mentioning that SP and CoSaMP based methods need the information of the number of active antennas as prior information which is learned iteratively in proposed method.

Figure 4 shows the SER and DER performance versus the sparsity level. It considers the case SNR equals 3 dB and 9 dB. In both configurations, proposed algorithm can achieve a better performance than other methods over a wide range of sparsity level. With up to 15 active antennas, proposed detector still has a 10^{-3} SER and 10^{-4} DER at 9 dB. This means proposed algorithm is robust to a changing number of active antennas. Another interesting observation is that, in low SNR case CoSaMP perform better than OMP, but in high SNR case, OMP is about 1 dB outperform CoSaMP.

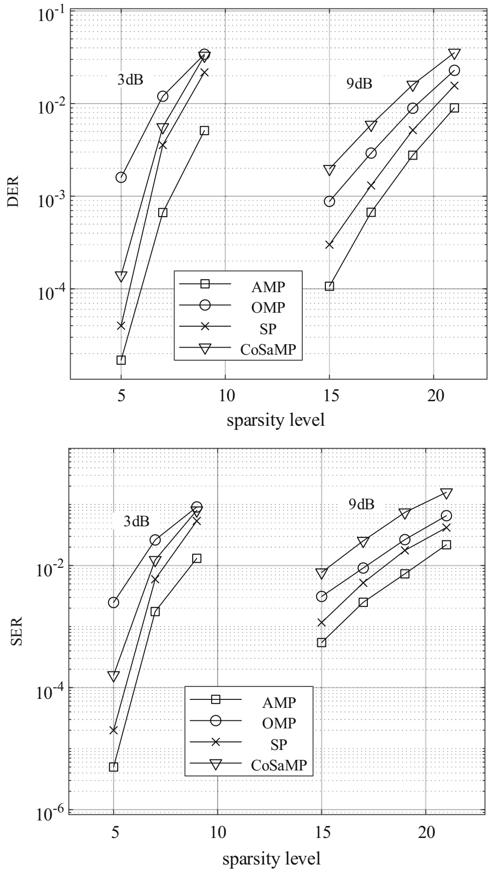

Fig. 4. DER and SER performance versus sparsity level

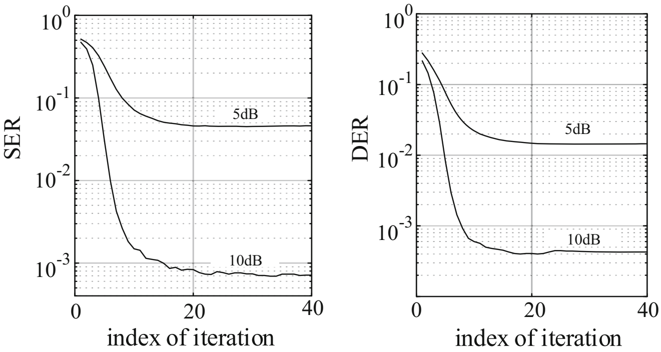

Fig. 5. Convergency performance versus iterations

Figure 5 shows the performance versus iteration indexes with sparsity level being 17 when SNR equals 5 dB and 10 dB. It demonstrates that, proposed detector converges much faster in the first 10 iterations and does not significantly improve after 15 iterations.

5 Conclusion

In this paper, we present an AMP based antenna activity and signal detection algorithm for spatial modulated NOMA. This solution shows an improved performance compared to previous CS approaches. This is mainly because Onsager correction ensures the denoiser input is an AWGN corrupt version of the ground truth. Another major advantage of proposed algorithm lies in the fact that it achieves its convergency in 15 iterations and its linear computational complexity makes it very practical.

Acknowledgment. This work was sponsored by the National Nature Science Foundation of China (NSFC) projects (61771132, 61571111) and the Research Fund of National Mobile Communications Research Laboratory, Southeast University (No. 2019A02).

References

1. Wu, Z., Lu, K., Jiang, C., Shao, X.: Comprehensive study and comparison on 5G NOMA schemes. IEEE Access **6**, 18511–18519 (2018)
2. Islam, S.M.R., Avazov, N., Dobre, O.A., Kwak, K.S.: Power-domain non-orthogonal multiple access (NOMA) in 5G systems: potentials and challenges. IEEE Commun. Surv. Tutor. **12**(2), 721–742 (2016)
3. Hoshyar, R., Wathan, F.P., Tafazolli, R.: Novel low-density signature for synchronous CDMA systems over AWGN channel. IEEE Trans. Signal Proc. **56**(4), 1616–1626 (2008)
4. Nikopour, H., Baligh, H.: Sparse code multiple access. In: Proceedings of the IEEE International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), London, UK, pp. 332–336, September 2013

5. Di Renzo, M., Haas, H., Ghrayeb, A., Sugiura, S., Hanzo, L.: Spatial modulation for generalized MIMO: challenges, opportunities, and implementation. Proc. IEEE **102**(1), 56–103 (2014)
6. Marzetta, T.: Noncooperative cellular wireless with unlimited numbers of base station antennas. IEEE Trans. Wirel. Commun. **9**(11), 3590–3600 (2010)
7. Wang, T., Liu, S., Yang, F., Wang, J., Song, J., Han, Z.: Block-sparse compressive sensing based multi-user and signal detection for generalized spatial modulation in NOMA. In: Proceedings of the IWCMC, Valencia, Spain, pp. 1992–1997, June 2017
8. Di Renzo, M., Haas, H.: On transmit-diversity for spatial modulation MIMO: impact of spatial-constellation diagram and shaping filters at the transmitter. IEEE Trans. Veh. Technol. **62**(6), 2507–2531 (2013)
9. Hong, J.P., Choi, W., Rao, B.D.: Sparsity controlled random multiple access with compressed sensing. IEEE Trans. Wirel. Commun. **14**(2), 998–1010 (2015)
10. Han, Z., Li, H., Yin, W.: Compressive Sensing for Wireless Networks. Cambridge University Press, Cambridge (2013)
11. Gao, Z., Dai, L., Qi, C., Yuen, C., Wang, Z.: Near-optimal signal detector based on structured compressive sensing for massive SM-MIMO. IEEE Trans. Veh. Technol. **66**(2), 1860–1865 (2017)
12. Donoho, D.L., Maliki, A., Montanari, A.: Message-passing algorithms for compressed sensing. Proc. Nat. Acad. Sci. **106**(45), 18914–18919 (2009)
13. Bayati, M., Montanari, A.: The dynamics of message passing on dense graphs, with applications to compressed sensing. IEEE Trans. Inf. Theory **57**(2), 764–785 (2011)
14. Rangan, S.: Generalized approximate message passing for estimation with random linear mixing. In: Proceedings of the IEEE International Symposium on Information Theory (ISIT), pp. 2168–2172, August 2011
15. Schniter, P.: A message-passing receiver for BICM-OFDM over unknown clustered-sparse channels. IEEE J. Sel. Top. Signal Process. **5**(8), 1462–1474 (2011)
16. Neal, R.M., Hinton, G.E.: A view of the EM algorithm that justifies incremental, sparse, and other variants. In: Learning in Graphical Models, vol. 89, pp. 355–368. MIT Press, Cambridge (1998)