Asynchronous Transmission for Multiple Access Channels: Rate-Region Analysis and System Design for Uplink NOMA

Mehdi Ganji, *Student Member, IEEE*, Xun Zou, *Student Member, IEEE*, and Hamid Jafarkhani, *Fellow, IEEE*

Abstract

In this work, we thoroughly analyze the rate-region provided by the asynchronous transmission in multiple access channels (MACs). We derive the corresponding capacity-regions, applicable to a wide range of pulse shaping methods. We analytically prove that asynchronous transmission enlarges the capacity-region of MACs. Although successive interference cancellation (SIC) can achieve the optimal sum-rate for the conventional uplink non-orthogonal multiple access (NOMA) methods, it is unable to achieve the boundary of the capacity-region for the asynchronous transmission. We demonstrate that for the asynchronous transmission, the optimal SIC decoding order to achieve the maximum sum-rate is based on the users' channel strengths. This optimal ordering is in contrast to the conventional uplink NOMA, where various decoding orders can result in the maximum sum-rate. Furthermore, we provide practical transceiver designs to approach the capacity-region. The memory induced by asynchronous transmission enables the use of the trellis-based detection methods which improves the performance. In addition, we propose a transceiver design, based on channel diagonalization to exploit the frequency-selectivity introduced by timing offsets. The proposed transceiver design, joint with the turbo principle, enables us to achieve a rate pair that is not achievable by the synchronous transmission.

I. Introduction

With the increase of mobile users and the higher data demand, the use of non-orthogonal methods in wireless networks is inevitable. Notably, in the uplink, where multiple users attempt to connect to the base station (BS), orthogonal multiple access (OMA) incurs inefficiency in resource utilization and requires more overhead signals. The capacity-region of the multiple

This work was supported in part by the NSF Award CCF-2008786. The authors are with the Center for Pervasive Communications and Computing, University of California, Irvine, CA, 92697 USA (email: {mganji, xzou4, hamidj}@uci.edu).

access channel (MAC) is derived and expressed in [1], [2] and is historically called the Cover-Wyner pentagon. While the performance analysis of the MAC and multiuser detection schemes is not new [3], recently, it has resurfaced under the category of uplink non-orthogonal-multiple-access (NOMA). For example, in the power-domain NOMA, the signals from multiple users are superposed with different power levels exploiting the difference in channel coefficients and a multiuser detection method, such as successive interference cancellation (SIC), is employed at the receiver [4]. The advantages of the NOMA over the OMA have been extensively studied in [5] and the references therein including higher system throughput compared with OMA and supporting massive connectivity.

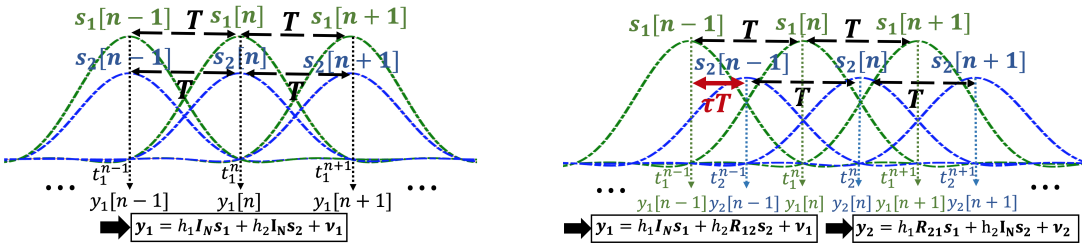
In most of the work in the literature, perfect synchronization in time and frequency is a common presumption. Indeed, the asynchrony is mostly considered an impairment [6], [7], where different synchronization methods are applied to eliminate it [8]. However, by using an appropriate transceiver design, asynchrony can indeed be beneficial. Asynchronous transmission refers to the case where the symbol epochs of the signals transmitted by the users are not aligned at the receiver. The results in [9] show that time asynchrony can increase the MAC's capacity-region in a code division multiple access (CDMA) system model where each user uses rectangular code sequences. By intentionally introducing symbol asynchrony in the transmitted signal, a higher diversity gain can be achieved by zero-forcing detection in spatial multiplexing [10]–[12]. The benefits of asynchrony in CDMA systems with random spreading are analyzed in [13], and it is shown that asynchronous transmission can indeed enhance the spectral efficiency. Besides, asynchronous NOMA (ANOMA) achieves a better throughput performance compared to the conventional (synchronous) NOMA [14]–[17]. Orthogonal differential decoding is improved by utilizing the oversampling technique [18], [19] to achieve the sampling diversity gain. An asynchronous network coding transmission strategy for multiuser cooperative networks is investigated in [20]. In [21], an interference cancellation (IC) technique, exploiting a triangular pattern, is proposed for asynchronous NOMA systems. In [22], a message passing (MP) detection method is proposed for symbol-asynchronous uplink NOMA systems.

The results of [9] are applicable to the pulse shapes whose duration is bounded by the symbol interval which limits its applicability to next-generation wireless networks. Due to the importance of bandwidth-efficient pulse shaping in modern communication, we generalize the results of [9] to band-limited pulse shapes such as raised cosine (r.c.). To derive the capacity region, the concept of asymptotic similarity of Toeplitz and Circular matrices [23], [24] is employed, which

links the capacity region to the Fourier transform of the used pulse shapes. The introduced relation provides insightful results including the effect of pulse shaping on the performance of ANOMA and optimal timing delays. In addition, we prove that while SIC is optimal for synchronous transmission in a capacity-achieving sense, it provides a sub-optimal rate-region for the asynchronous transmission. We analytically show the optimal decoding order of SIC detection for the asynchronous transmission. Moreover, to bridge the gap between theoretical capacity analysis and practical design of communication systems, we provide a transceiver design to approach the capacity-region boundaries. The major contributions of this work are summarized as follows:

- The MAC capacity-region analysis is generalized to a wide range of pulse shapes including the well-known and practically common r.c. pulse shape. In addition, the effect of users' channel coefficients on the capacity-region is analyzed.
- The advantage of asynchronous transmission in enlarging the capacity-region is analytically proven. It is proved that the time delay of half symbol interval provides the largest capacity region for r.c. pulse shape.
- The performance of the well-known SIC method is thoroughly analyzed and compared for the scenarios of synchronous and asynchronous transmissions.
- Due to the memory imposed by asynchronous transmission, the receiver is enabled to use trellis-based detection methods which significantly improve the performance. ANOMA methods are applicable even if the difference in the channel quality is not significant which is required for the power-domain NOMA.
- To further improve the performance and reduce computational complexity, a new transceiver design is proposed, exploiting the introduced frequency-selectivity by asynchronous transmission. The new design in conjunction with the turbo principle provides a performance close to the capacity boundaries, not achievable by conventional synchronous transmission.

The rest of the manuscript is organised as follows: In Section II, the general system model, characteristics, and input-output relation are outlined. In Section III, the capacity-region is derived for a general type of pulse shape. In Section IV, the transceiver design is discussed and various receiver designs are proposed and analyzed, including SIC, trellis-based receivers, and turbo receivers. In Section V, numerical results are provided to analyze the performance of the proposed transceiver designs and finally, the conclusion is presented in Section VI.



(a) Symbol Synchronous (b) Symbol Asynchronous

Fig. 1: Demonstration of sufficient statistics for synchronous and asynchronous transmission.

II. System Model

For a system with K single-antenna users and one common receiver, assuming frame synchronisation and a flat fading channel model, we can write the received signal as

$$y(t) = \sum_{k=1}^{K} h_k \sum_{n=1}^{N} s_k[n] p(t - nT - \tau_k T) + \nu(t). \tag{1}$$

User k transmits a codeword $(s_k[1], \cdots, s_k[N]) \in \mathbb{S}_k^N$, where $\mathbb{S}_k$ represents the input domain, $p(t)$ is the normalized pulse shaping function, i.e., $\int |p(t)|^2 dt = 1$, with the support of T_p, T denotes the symbol interval and $\nu(t)$ is the white Gaussian noise with variance of σ^2. The $\tau_k \in [0, 1)$ accounts for the timing offset of User k, and h_k represents the kth user's channel coefficient and are known to the receiver. For capacity analysis, the symbols are assumed to be chosen from Gaussian processes whose optimal power spectral density (PSD) should be found. However, for practical transceiver design, well-known constellations, e.g., BPSK, and proper coding schemes are assumed. In addition, it is assumed that the pulse shape occupies a frequency bandwidth of B. For example, for a r.c. or root raised cosine (r.r.c) pulse shape with roll-off factor of β, the occupied bandwidth is $B = \frac{1+\beta}{T}$, where T is assumed to be normalized to 1.

If the transmitted signals are not aligned at the receiver, then the channel is symbol asynchronous as shown in Fig. 1. To get the sufficient statistics, a matched filter with the impulse response $p(t)$ is applied to the received signal and its output is sampled at time instants $t_k^n = nT + \tau_k T$, $n = 1, 2, \cdots, N$, $k = 1, \cdots, K$ which results in the output samples $y_k[n]$. One can arrange the output samples in a vector and define $\boldsymbol{y_k} = (y_k[1], \cdots, y_k[N])^T$. Each set of samples, $\boldsymbol{y_k}$, is matched to the corresponding time delay, i.e., τ_k. For synchronous transmission, sufficient statistics include one set of samples while for asynchronous transmission, it includes K set of samples as shown in Fig. 1. For each set of samples we can have the input-output

relationship of $\boldsymbol{y_k} = \sum_{l=1}^{K} h_l \boldsymbol{R_{kl}} \boldsymbol{s_l} + \boldsymbol{\nu_k}$. Matrix $\boldsymbol{R_{kl}}$ is an $N \times N$ Toeplitz matrix whose elements depend on the pulse shape, the corresponding time delay and are calculated by:

$$[\boldsymbol{R_{kl}}]_{n,m} = g(\tau_{kl}T + (m-n)T) \triangleq g_{\tau_{kl}}(m-n), \quad m,n = 1, \cdots, N, \tag{2}$$

where $\tau_{kl} = \tau_l - \tau_k$ and $g(t) = p(t) * p(t)$, where the operation $*$ represents convolution. The vector $\boldsymbol{\nu_k}$ represents the noise vector whose co-variance matrix is defined as $\boldsymbol{Q_{kl}} = \mathbb{E}[\boldsymbol{\nu_k}{\boldsymbol{\nu_l}}^H] = \sigma^2 \boldsymbol{R_{kl}}$. For any square-root Nyquist pulse, e.g., r.r.c., $\boldsymbol{R_{kk}} = \boldsymbol{I_N}$ and $\boldsymbol{R_{kl}^T} = \boldsymbol{R_{lk}}$. Also note that for the synchronous transmission, i.e., $\tau_k = 0 \ \forall k$, $\boldsymbol{R_{lk}} = \boldsymbol{I_N}$. We can put all the received samples together and define $\boldsymbol{y} = \left({\boldsymbol{y_1}}^T, \cdots, {\boldsymbol{y_K}}^T\right)^T$ to get the system model in a matrix form as:

$$\boldsymbol{y} = \begin{pmatrix} \boldsymbol{R_{11}} & \cdots & \boldsymbol{R_{1K}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{R_{K1}} & \cdots & \boldsymbol{R_{KK}} \end{pmatrix} \begin{pmatrix} h_1 \boldsymbol{I_N} & \cdots & \boldsymbol{0_N} \\ \vdots & \ddots & \vdots \\ \boldsymbol{0_N} & \cdots & h_K \boldsymbol{I_N} \end{pmatrix} \begin{pmatrix} \boldsymbol{s_1} \\ \vdots \\ \boldsymbol{s_K} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\nu_1} \\ \vdots \\ \boldsymbol{\nu_K} \end{pmatrix} = \boldsymbol{RHs} + \boldsymbol{\nu}, \tag{3}$$

where $\boldsymbol{R}$ is an $NK \times NK$ matrix constructed by Toeplitz blocks of $\boldsymbol{R_{lk}}$.

III. Capacity-Region Analysis

To analyze the capacity-region of the resulting asynchronous system, we assume a two-user scenario. Let us define $\boldsymbol{s_1}$ and $\boldsymbol{s_2}$ as Gaussian processes with PSD of $\{S_1(f), S_2(f), \ f \in [0,1]\}$ for Users 1 and 2, respectively. Then, the capacity-region of a two-user MAC with memory is [25]:

$$C = \bigcup_{\substack{S_k(f) \geq 0, \ f \in [0,1] \\ \int_0^1 S_k(f) df \leq P_k \\ k=1,2}} \left\{ (R_1, R_2), \begin{array}{c} 0 \leq R_1 \leq \lim_{N\to\infty} \frac{1}{N} I(\boldsymbol{y}; \boldsymbol{s_1} | \boldsymbol{s_2}) \\ 0 \leq R_2 \leq \lim_{N\to\infty} \frac{1}{N} I(\boldsymbol{y}; \boldsymbol{s_2} | \boldsymbol{s_1}) \\ 0 \leq R_1 + R_2 \leq \lim_{N\to\infty} \frac{1}{N} I(\boldsymbol{y}; \boldsymbol{s_1}, \boldsymbol{s_2}) \end{array} \right\}, \tag{4}$$

where R_k and P_k represent the achievable rate and the available power of User k, respectively, and $\boldsymbol{y}$ represents the $2N$ received samples obtained as explained in Eq. (3). In the next theorem, the capacity-region is derived.

Theorem 1: The capacity-region of a two-user asynchronous MAC is:

$$C = \bigcup_{\substack{S_k(f) \geq 0, \ k=1,2 \\ \int_0^1 S_k(f) df \leq P_k}} \left\{ (R_1, R_2), \begin{array}{c} 0 \leq R_1 \leq \frac{1}{2} \int_0^1 \log_2 \left(1 + \frac{S_1(f)}{\sigma_1^2}\right) df \\ 0 \leq R_2 \leq \frac{1}{2} \int_0^1 \log_2 \left(1 + \frac{S_2(f)}{\sigma_2^2}\right) df \\ 0 \leq R_1 + R_2 \leq \frac{1}{2} \int_0^1 \log \left(1 + \frac{S_1(f)}{\sigma_1^2} + \frac{S_2(f)}{\sigma_2^2} + \frac{S_1(f) S_2(f) (1 - G_\tau^2(f))}{\sigma_1^2 \sigma_2^2}\right) df \end{array} \right\}, \tag{5}$$

where $\sigma_1^2 = \frac{\sigma^2}{|h_1|^2}$, $\sigma_2^2 = \frac{\sigma^2}{|h_2|^2}$ and $G_\tau(f)$ depends on the pulse shape and the timing offset and is defined as

$$G_\tau(f) = \left| \frac{1}{T} \sum_{i=-\infty}^{\infty} e^{-j2\pi\tau(f+i)} \hat{g}\left(\frac{f+i}{T}\right) \right|, \tag{6}$$

where $\hat{g}(f)$ is the Fourier transform of the matched filter pulse $g(t)$.

Proof: The proof is presented in Appendix A. ■

Note that with no timing offset, i.e., $\tau = 0$, $G_\tau(f)$ becomes the conventional folded spectrum which is constant and is equal to 1 for Normalized Nyquist pulse shapes. Then, the capacity-region of the MAC turns into the conventional Cover-Wyner region:

$$C_{synch} = \left\{ \begin{array}{c} 0 \le R_1 \le \frac{1}{2}\log_2\left(1+\frac{P_1}{\sigma_1^2}\right) \\ (R_1, R_2), \quad 0 \le R_2 \le \frac{1}{2}\log_2\left(1+\frac{P_2}{\sigma_2^2}\right) \\ 0 \le R_1 + R_2 \le \frac{1}{2}\log\left(1+\frac{P_1}{\sigma_1^2}+\frac{P_2}{\sigma_2^2}\right) \end{array} \right\}. \tag{7}$$

Proposition 1: The asynchronous transmission enlarges the capacity-region of a two-user MAC compared with the synchronous transmission.

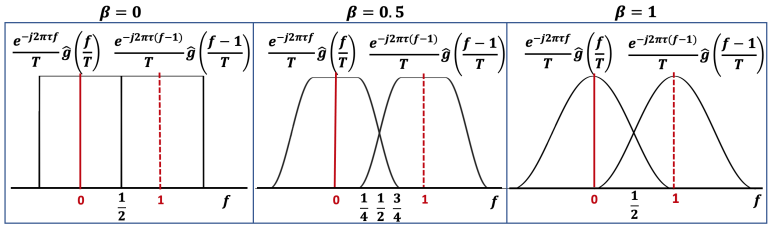
Proof: To prove the proposition, it is enough to prove that for the asynchronous transmission, the function $G_\tau(f)$ is less than or equal to that of the synchronous transmission. Since $\hat{g}(f)$ is a non-negative real-valued function (due to the matched filtering process at the receiver), we have:

$$\begin{aligned} G_{asynch}(f) = \left| \frac{1}{T} \sum_{i=-\infty}^{\infty} e^{-j2\pi\tau(f+i)} \hat{g}\left(\frac{f+i}{T}\right) \right| &\le \frac{1}{T} \sum_{i=-\infty}^{\infty} \left| e^{-j2\pi\tau(f+i)} \right| \left| \hat{g}\left(\frac{f+i}{T}\right) \right|, \\ &= \frac{1}{T} \sum_{i=-\infty}^{\infty} \hat{g}\left(\frac{f+i}{T}\right) = G_{synch}(f), \ \forall f. \end{aligned} \tag{8}$$

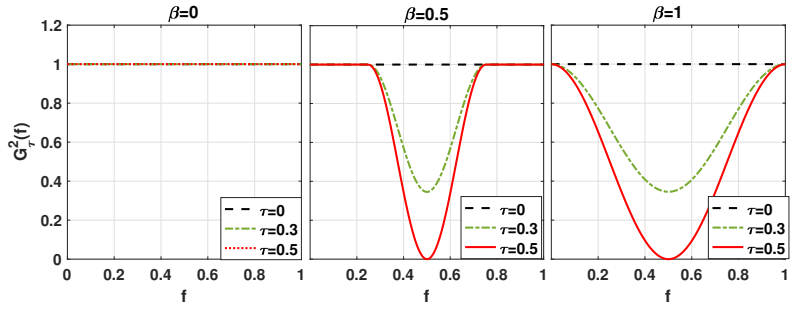
■

Example 1: In this example, practically common pulse shape of r.r.c is considered. After matched filtering, the effective pulse shape is the raised cosine whose frequency spectrum is $\hat{g}(f) = \begin{cases} T & |f| \le \frac{1-\beta}{2T} \\ \frac{T}{2}\left[1+\cos\left(\frac{\pi T}{\beta}\left(|f|-\frac{1-\beta}{2T}\right)\right)\right] & \frac{1-\beta}{2T} < |f| \le \frac{1+\beta}{2T} \\ 0 & o.w. \end{cases}$. The phase-shifted folded spectrum, $G_\tau(f) = \left|\frac{1}{T}\sum_{i=-\infty}^{\infty} e^{-j2\pi\tau(f+i)}\hat{g}\left(\frac{f+i}{T}\right)\right|$ is periodic with the period of 1 and can be derived as:

$$G_\tau(f) = \begin{cases} 1 & 0 \le f < \frac{1-\beta}{2} \\ \left| e^{-j2\pi\tau f} A(f) + e^{-j2\pi\tau(f-1)} A(f-1) \right| & \frac{1-\beta}{2} \le f \le \frac{1+\beta}{2} \\ 1 & \frac{1+\beta}{2} < f \le 1 \end{cases}, \tag{9}$$



(a) Schematic description of phase-shifted folded spectrum for different β



(b) The resulting phase-shifted folded spectrums

Fig. 2: Demonstration of $G_\tau(f)$ for r.r.c. pulse with $\beta = 0, 0.5, 1$.

where $A(f) = \frac{1}{2}\left[1+\cos\left(\frac{\pi T}{\beta}\left(\frac{f}{T}-\frac{1-\beta}{2T}\right)\right)\right]$. Examples of $G_\tau(f)$ is shown with various parameters in Fig. 2.

Proposition 2: The timing offset $\tau = 0.5$ provides the largest capacity region for the r.r.c. pulse shape.

Proof: To show the optimilaty of $\tau = 0.5$, we show that for every frequency f_0 in range $(\frac{1-\beta}{2}, \frac{1+\beta}{2})$, $G_\tau(f_0)$ is minimized by $\tau = 0.5$. Since $A(f_0)$ and $A(f_0-1)$ are real positive values, $G_\tau(f_0)$ can be interpreted as the magnitude of sum of two vectors, namely, $e^{-j2\pi\tau f_0}A(f_0)$ and $e^{-j2\pi\tau(f_0-1)}A(f_0-1)$ in a 2-dimensional space. Consequently, the magnitude of the sum vector is minimized if the individual vectors are aligned in opposite directions. The phase difference between two individual vectors is $2\pi\tau$ which will be equal to π by setting $\tau = 0.5$. Thus, $\tau = 0.5$ minimizes $G_\tau(f_0)$ for every f_0, and provides the largest capacity region. The proof is illustrated in Fig. 3 ■

In simpler words, $\tau = 0.5$ causes the shifted spectrums to add up with opposite phase offsets and results in reduced $G_\tau(f)$. In addition, as observed in Fig. 2, a higher roll-off factor further reduces $G_\tau(f)$. The underlying reason is that the increased out-of-Nyquist-band (ONB) spectrum of the r.r.c. causes more destructive addition in the phase-shifted folded spectrum. With no ONB

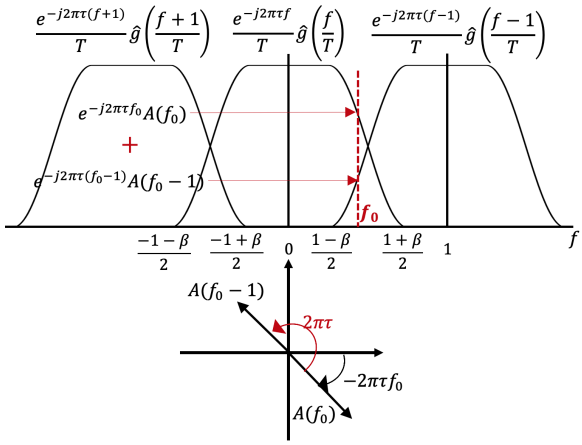


Fig. 3: Phase-shifted folded spectrum description: proof of optimality $\tau = 0.5$.

spectrum, i.e., $\beta = 0$, we have $G_\tau(f) = 1, f \in [0,1], \forall\tau$, and thus, the capacity-region is the conventional Wyner-Cover pentagon for any choice of time delay. In summary, asynchronous transmission exploits the ONB spectrum and is ineffective for the pulse shapes that do not have an ONB spectrum such as the Sinc pulse shape.

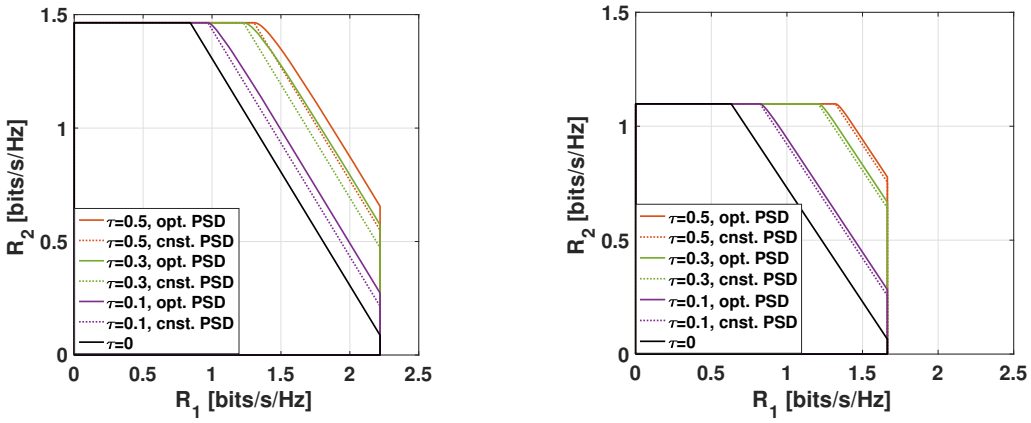
The next step to specify the capacity-region is to find the optimal power allocations. It is worth describing the achievable rate-region using constant PSDs first. Assuming a constant power allocation for each user, i.e., $S_k(f) = P_k$, the achievable rate-region is:

$$C = \left\{ \begin{array}{c} (R_1, R_2), \\ \end{array} \begin{array}{c} 0 \le R_1 \le \frac{1}{2}\log_2\left(1+\frac{P_1}{\sigma_1^2}\right) \\ 0 \le R_2 \le \frac{1}{2}\log_2\left(1+\frac{P_2}{\sigma_2^2}\right) \\ 0 \le R_1+R_2 \le \frac{1}{2}\int_0^1 \log\left(1+\frac{P_1}{\sigma_1^2}+\frac{P_2}{\sigma_2^2}+\frac{P_1P_2(1-G_\tau^2(f))}{\sigma_1^2\sigma_2^2}\right)df \end{array} \right\}. \tag{10}$$

Using Proposition 1, it is obvious that the asynchronous rate-region with constant PSDs is larger than the capacity-region of the synchronous transmission. However, optimizing the input PSDs provides further improvement in achievable rates and results in the asynchronous capacity-region. Following the steps in [9], the optimization problem to find the optimal PSDs can be formulated as:

$$\underset{S_1(f),S_2(f)}{\arg\max} \max_{R_1,R_2} \alpha R_1 + (1-\alpha)R_2 \tag{11}$$

$$s.t. \;\; R_k \in C \; in \; (5), \; \int_0^1 S_k(f)df = P_k, \; S_k(f) \ge 0$$



(a) $\beta = 0.5$, $\sigma_1^2 = 1, \sigma_2^2 = 5$ (b) $\beta = 1$, $\sigma_1^2 = 1, \sigma_2^2 = 5$

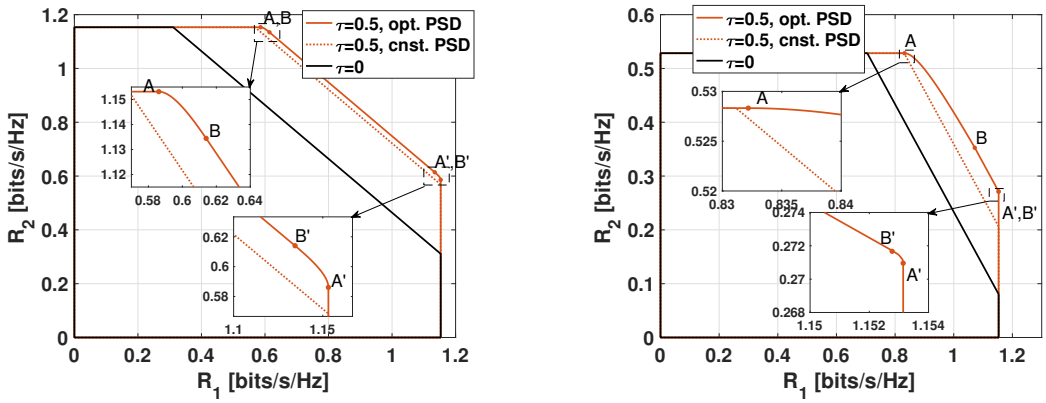
Fig. 4: Capacity-region of the asynchronous MAC with $P_1 = P_2 = 20\ dB$ and r.r.c. pulse shape.

for every $0 \leq \alpha \leq 1$. The inner maximization which describes the Pareto points of the region can be simplified as:

$$\begin{aligned} &\arg\max_{S_1(f),S_2(f)} \max \big\{ \alpha F\big(S_1(f),0\big) + (1-\alpha)[F\big(S_1(f),S_2(f)\big) - F\big(S_1(f),0\big)], \\ &\qquad \alpha[F\big(S_1(f),S_2(f)\big) - F\big(0,S_2(f)\big)] + (1-\alpha)F\big(0,S_2(f)\big) \big\} \qquad (12) \\ &= \arg\max_{S_1(f),S_2(f)} \begin{cases} (2\alpha-1)F\big(S_1(f),0\big) + (1-\alpha)F\big(S_1(f),S_2(f)\big) & 1/2 \leq \alpha \leq 1 \\ (1-2\alpha)F\big(S_1(f),0\big) + \alpha F\big(S_1(f),S_2(f)\big) & 0 \leq \alpha \leq 1/2 \end{cases}, \qquad (13) \end{aligned}$$

where $F(S_1(f), S_2(f)) = \frac{1}{2}\int_0^1 \log\left(1 + \frac{S_1(f)}{\sigma_1^2} + \frac{S_2(f)}{\sigma_2^2} + \frac{S_1(f)S_2(f)(1-G_\tau^2(f))}{\sigma_1^2\sigma_2^2}\right) df$. After solving the nonlinear optimization problem for various values of α, the resulting capacity-regions for $\beta = 0.5$ and $\beta = 1$ are shown in Fig. 4. It is shown that the asynchronous transmission improves the capacity-region compared with the synchronous transmission. In addition, $\tau = 0.5$ provides the largest capacity region. Although constant PSDs are optimal for synchronous transmission, those cannot achieve the capacity for asynchronous transmission. However, even with constant PSDs, the asynchronous transmission provides a significant improvement in achievable rates compared with the synchronous transmission. Fig. 4 shows that power allocation is more effective for the r.r.c. pulse shape with $\beta = 0.5$. In addition, simulation results that are not included here confirm that as the transmit power increases, the provided gain by power optimization reduces.

Note that by increasing the roll-off factor, β, the gain provided by asynchronous transmission is increased. For example, using r.r.c. with $\beta = 0.5$ and $\beta = 1$, asynchronous transmission provides



(a) $\beta = 0.5$, $\sigma_1^2 = 1, \sigma_2^2 = 1$ (b) $\beta = 0.5$, $\sigma_1^2 = 1, \sigma_2^2 = 5$

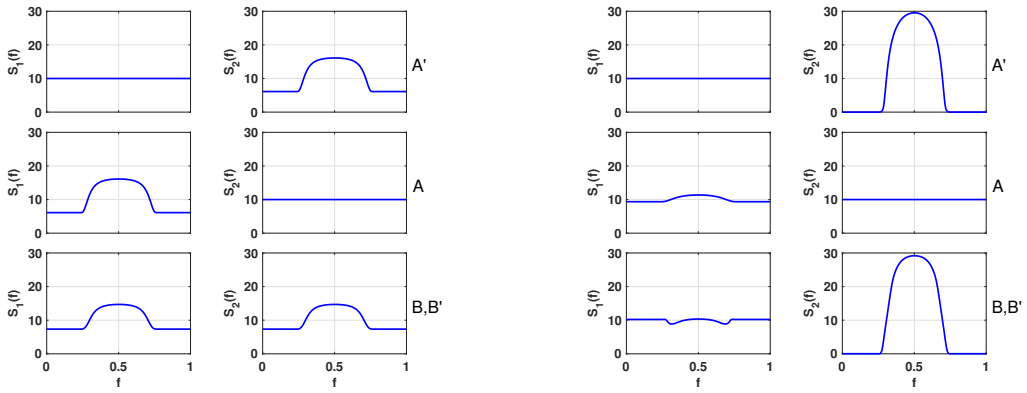
Fig. 5: Capacity-region analysis of asynchronous MAC with $P_1 = P_2 = 10\ dB$ and r.r.c. pulse shape.

up to 25% (R_{sum} : $2.3 \rightarrow 2.87$ $[bits/s/Hz]$) and 41% (R_{sum} : $1.73 \rightarrow 2.44$ $[bits/s/Hz]$) improvements in sum-rate, respectively, compared with the synchronous transmission. The further improvement provided with the roll-off factor of $\beta = 1$ is supported by the fact that $\beta = 1$ results in more reduction in the $G_\tau(f)$ function, shown in Fig. 2. Worth mentioning that using pulse shapes with a greater roll-of factor is not spectrally efficient since the frequency spectrum is not fully utilized. However, asynchronous transmission can restore the spectral efficiency loss partially.

In the capacity-region, four points are of great importance, denoted as A, A', B, B' in Fig. 5. Points A and A' are obtained by maximizing the sum-rate upper-bound over one user's PSD while the other user's PSD is assumed to be constant. By formulating the optimization problem, considering the KKT conditions and some simplifications, we have:

$$S_k^*(f) = \left[\lambda - \frac{\sigma_{\bar{k}}^2 + P_{\bar{k}}}{\frac{\sigma_{\bar{k}}^2}{\sigma_k^2} + \frac{P_{\bar{k}}}{\sigma_k^2}(1 - G_\tau^2(f))}\right]^+, \quad k = 1, 2, \tag{14}$$

where $\bar{k} = \{1, 2\} - k$ and $[x]^+ = \max\{0, x\}$. Eq. (14) shows that the optimal PSDs for points A and A' are not constants and depend on $G_\tau(f)$. In addition, the ratio of channel strengths, i.e., $\frac{\sigma_{\bar{k}}^2}{\sigma_k^2}$, plays an important role in specifying optimal PSDs. Particularly, when one of the users has a much stronger channel strength, the dependence of the stronger user's PSD on $G_\tau(f)$ reduces and the optimal PSD approaches a constant PSD. On the other hand, the weaker user's PSD



(a) $\beta = 0.5$, $\sigma_1^2 = 1, \sigma_2^2 = 1$ (b) $\beta = 0.5$, $\sigma_1^2 = 1, \sigma_2^2 = 5$

Fig. 6: Optimal PSD pairs achieving the critical points, A, A', B, B', in the capacity-region figure.

greatly depends on $G_\tau(f)$ and the optimal PSD asymptotically assigns no power to frequencies for which $G_\tau(f) = 1$. These observations are illustrated in Fig. 6b, where the optimized PSD for the stronger user (row A) is almost constant with insignificant variations while the weaker user (row A') selectively assigns its available power based on $G_\tau(f)$. This fact is the underlying reason for the proximity of point A to the constant PSD corner (dashed line) and the large gap between A' and the constant PSD corner in Fig. 5b.

Points B and B' are obtained by maximizing the sum-rate upper-bound over both users' PSDs. Similarly, the optimized PSDs can be obtained as:

$$S_k^*(f) = \left[\lambda - \frac{\sigma_{\bar{k}}^2 + S_{\bar{k}}^*(f)}{\frac{\sigma_{\bar{k}}^2}{\sigma_k^2} + \frac{S_{\bar{k}}^*(f)}{\sigma_k^2}(1 - G_\tau^2(f))}\right]^+, \quad k = 1, 2. \tag{15}$$

Again, if the difference between channel strengths is large, then, the stronger user's optimal PSD approaches a constant, as shown in Fig. 6b (row B, B'). Therefore, with asynchronous transmission and assuming a near-far scenario, the power optimization is more critical for the weaker user as the stronger user's optimal PSD is close to a constant PSD.

For a synchronous scenario, the pentagon's corner points can be achieved by the SIC method, and the rest of Pareto points on the connecting line are achieved by time-sharing. This optimality is the reason that SIC is widely used in the uplink NOMA literature. Due to the importance of the SIC method, its achievable rates are analyzed in the next subsection.

A. SIC's Achievable Rate-Region

Using SIC and assuming the decoding order of $\{k, j\}$, the information symbols of User k are decoded while User j's signal is considered as noise. After removal of the decoded symbols, the symbols belonged to User j are decoded without interference assuming error-free decoding at the first stage [26]. For the asynchronous system model in Eq. (3), let us rewrite the received samples as $y_k[m] = h_k s_k[m] + h_{\bar{k}} \sum_{n=-\infty}^{\infty} g_\tau(n) s_{\bar{k}}[m+n] + \nu_k[m], k = 1, 2$. Note that due to the time delay, the inter-user interference (IUI) is caused by multiple interfering symbols rather than the one in the synchronous case. Assuming Gaussian signaling with error-free decoding and defining $\eta_\tau = \sum_{n=-\infty}^{\infty} g_\tau^2(n)$, the achievable rate pairs for the ANOMA scheme using SIC are:

$$R_1 = \frac{1}{2}\log_2\left(1 + \frac{P_1|h_1|^2}{\eta_\tau P_2|h_2|^2 + \sigma^2}\right), R_2 = \frac{1}{2}\log_2\left(1 + \frac{P_2|h_2|^2}{\sigma^2}\right), \tag{16}$$

if the decoding order is $\{1, 2\}$ and

$$R_2 = \frac{1}{2}\log_2\left(1 + \frac{P_2|h_2|^2}{\eta_\tau P_1|h_1|^2 + \sigma^2}\right), R_1 = \frac{1}{2}\log_2\left(1 + \frac{P_1|h_1|^2}{\sigma^2}\right), \tag{17}$$

if the decoding order is $\{2, 1\}$. Equivalently, η_τ can be calculated by $\eta_\tau = \int_0^1 G_\tau^2(f)df \leq 1$, where equality is achieved by $\tau = 0$ [27]. The achievable rate pairs for the NOMA scheme using SIC are similarly obtained by inserting $\eta_0 = 1$. Therefore, for every set of power and channel coefficients, the asynchronous transmission enjoys less inter-user interference compared to the synchronous case. Particularly, for the r.r.c. pulse shape $\eta_\tau = 1 - \beta/4 + \beta/4\cos(2\pi\tau)$, which is minimized at $\tau = 0.5$ [27]. The rate pairs for the NOMA scheme using SIC coincide with the corner points of the pentagon capacity-region and the sum-rate is calculated as:

$$R_{sum} = \frac{1}{2}\log_2\left(1 + \frac{P_j|h_j|^2}{P_k|h_k|^2 + \sigma^2}\right) + \frac{1}{2}\log_2\left(1 + \frac{P_k|h_k|^2}{\sigma^2}\right) \quad (k, j) = \{(1, 2), (2, 1)\} \tag{18}$$

$$= \frac{1}{2}\log_2\left(1 + \frac{P_1|h_1|^2}{\sigma^2} + \frac{P_2|h_2|^2}{\sigma^2}\right), \tag{19}$$

which is equal to the maximum sum-rate in capacity-region (7). Hence, assuming error-free decoding, as done in capacity calculations, both decoding orders can provide the maximum sum-rate in uplink NOMA. However, for ANOMA, decoding the stronger user first results in a strictly larger sum-rate. In more details, assuming $|h_1|^2 > |h_2|^2$, and $P_1, P_2, \sigma^2 > 0$, results in $\frac{1}{2}\log_2\left(1 + \frac{P_1|h_1|^2}{\eta_\tau P_2|h_2|^2+\sigma^2}\right) + \frac{1}{2}\log_2\left(1 + \frac{P_2|h_2|^2}{\sigma^2}\right) > \frac{1}{2}\log_2\left(1 + \frac{P_2|h_2|^2}{\eta_\tau P_1|h_1|^2+\sigma^2}\right) + \frac{1}{2}\log_2\left(1 + \frac{P_1|h_1|^2}{\sigma^2}\right)$. In addition, it can be easily concluded that the SIC rate-region achieved by the asynchronous

transmission is larger than the capacity-region of the synchronous transmission. Nevertheless, the SIC method is not capacity-achieving for ANOMA.

Proposition 3: Unlike NOMA, in which SIC is capacity-achieving, the SIC method cannot achieve the boundaries of the capacity region for ANOMA.

Proof: The optimallity of the SIC method for the NOMA scheme was expressed previously. For ANOMA, the maximum sum-rate achieved by SIC, assuming $|h_1|^2 \geq |h_2|^2$, is $R_{sum}^{SIC} = \frac{1}{2}\log_2\left(1+\frac{P_1|h_1|^2}{\eta_\tau P_2|h_2|^2+\sigma^2}\right)+\frac{1}{2}\log_2\left(1+\frac{P_2|h_2|^2}{\sigma^2}\right)$. The sum-rate boundary of the rate-region with constant PSDs is denoted by $R_{sum}^{const.} = \frac{1}{2}\int_0^1 \log\left(1+\frac{P_1|h_1|^2}{\sigma^2}+\frac{P_2|h_2|^2}{\sigma^2}+\frac{P_1P_2|h_1|^2|h_2|^2(1-G_\tau^2(f))}{\sigma^4}\right)df$. To prove the proposition, it is enough to show that $R_{sum}^{SIC} < R_{sum}^{const.}$:

$$R_{sum}^{SIC} = \frac{1}{2}\log_2\left(1+\frac{P_1|h_1|^2}{\eta_\tau P_2|h_2|^2+\sigma^2}\right)+\frac{1}{2}\log_2\left(1+\frac{P_2|h_2|^2}{\sigma^2}\right) \tag{20}$$

$$\overset{(a)}{=} \frac{1}{2}\log_2\left(1+\frac{P_1|h_1|^2}{\int_0^1 G_\tau(f)df\,P_2|h_2|^2+\sigma^2}\right)+\frac{1}{2}\log_2\left(1+\frac{P_2|h_2|^2}{\sigma^2}\right) \tag{21}$$

$$\overset{(b)}{\leq} \frac{1}{2}\int_0^1 \log_2\left(1+\frac{P_1|h_1|^2}{G_\tau(f)P_2|h_2|^2+\sigma^2}\right)df+\frac{1}{2}\log_2\left(1+\frac{P_2|h_2|^2}{\sigma^2}\right) \tag{22}$$

$$\overset{(c)}{\leq} \frac{1}{2}\int_0^1 \log\left(1+\frac{P_1|h_1|^2}{\sigma^2}+\frac{P_2|h_2|^2}{\sigma^2}+\frac{P_1P_2|h_1|^2|h_2|^2(1-G_\tau^2(f))}{\sigma^2\sigma^2}\right)df \tag{23}$$

$$= R_{sum}^{const.}, \tag{24}$$

where (a) is obtained by substituting $\eta_\tau = \int_0^1 G_\tau(f)df$, (b) is the result of applying the Jenson's inequality to the convex function $\log_2(1+\frac{1}{x})$, and finally, (c) can be achieved by simple calculations assuming $G_\tau(f) \leq 1, \forall f$. The equalities in (b) and (c) are achieved by having a constant $G_\tau(f) = 1$. ■

Fig. 7 depicts the rate-regions for ANOMA with the optimized PSD (i.e., capacity-regions), the constant PSD, and the SIC method as well as the NOMA capacity-region with $P_1 = P_2 = 10\ dB, \sigma_1^2 = 1, \sigma_2^2 = 5$, using the r.r.c. pulse shape having $\beta = 0.5$. The figure verifies the results in this subsection.

IV. Transceiver Design

The capacity-region derived in the previous section is obtained by using the well-known Gaussian random coding. Although Gaussian random coding shows that a specific rate is achievable, its practical implementation is cumbersome. Therefore, in practical applications, proper transceiver design and using intelligent coding with manageable complexity are of

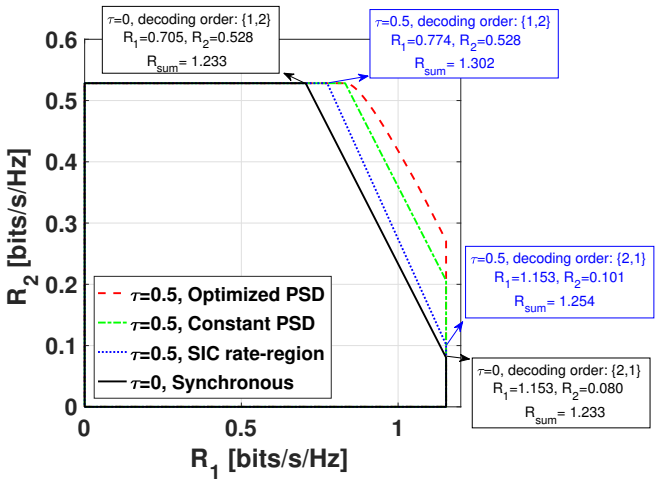


Fig. 7: Different rate-regions for $P_1 = P_2 = 10\ dB, \sigma_1^2 = 1, \sigma_2^2 = 5$, using r.r.c. with $\beta = 0.5$.

great importance. In this section, we focus on the transceiver design. We show that even with practical considerations such as limited frame length and non-Gaussian symbols, asynchronous transmission can provide significant performance gain. Note that some of the explained concepts are well-studied in the literature in various contexts. Nevertheless, the details are presented here for completeness. The highlight of this section includes (1) Introducing a Toeplitz system model for the asynchronous transmission which resembles the Ungerboeck model for ISI channels. Then, applying the trellis-based detection methods to the introduced system model which avoids the complex process of noise whitening. (2) Proposing a transceiver design for asynchronous transmission based on the concept of channel diagonalization. The proposed method outperforms the NOMA method with comparable complexity. (3) Using the proposed transceiver design joint with turbo principle to show that ANOMA can achieve an operational rate pair that is outside of the capacity region of conventional NOMA.

Let us consider the general transmitter structure in Fig. 8. The information bits of User k, denoted as $\boldsymbol{a_k} = (a_k[1], \cdots, a_k[U])$, are first encoded by an error correcting code to generate $\boldsymbol{b_k} = (b_k[1], \cdots, b_k[V])$, where $r = \frac{U}{V}$ is the encoding rate. After interleaving the coded bits, they are mapped to a predefined constellation resulting in symbols $\boldsymbol{s_k} = (s_k[1], \cdots, s_k[N])$, where $N = \frac{V}{\log_2 M}$ and M is the size of the constellation. This scheme is called Bit Interleaved Coded Modulation (BICM) in the literature and is used in many applications and communication standards [28], [29]. Then, the symbols are pulse shaped with the pulse $p(t)$ and transmitted through the channel which deteriorates the signal with channel coefficient of h_k and imposes

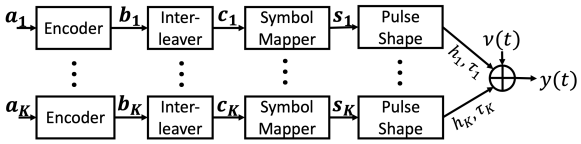
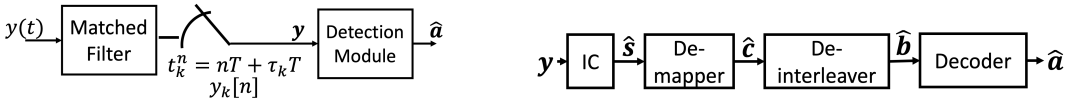


Fig. 8: Transmitter.



(a) Generic Receiver Design

(b) Separate Equalizer and Decoder

Fig. 9: Receiver Structure.

the time delay of τ_k. The received signal is matched filter and sampled properly to generate the discrete samples $y_k[n]$ at the receiver. The obtained samples, then, go through the detection module to find the estimate of the transmitted information bits as shown in Fig. 9a. The MAP decoder computes estimates of the information bits, $\hat{\boldsymbol{a}}$, minimizing bit error rate (BER).

$$\hat{a}_k[n] = \arg\max_{a \in \mathbb{F}_2} P(a_k[n] = a|\boldsymbol{y}) = \arg\max_{a \in \mathbb{F}_2} \sum_{\boldsymbol{a} \in \mathbb{F}_2^U : a_k[n] = a} P(\boldsymbol{y}|\boldsymbol{a}) P(\boldsymbol{a}). \tag{25}$$

Unfortunately, the BER-optimal decoder in Eq. (25) has a computational complexity that is of order $\mathcal{O}\left(2^U\right)$, which becomes intractable as U increases. A conventional approach to reduce the computational burden of the receiver is to split the detection problem into two sub-problems of interference cancellation (IC) and decoding as shown in Fig. 9b. IC refers to canceling the interference caused by the user's own symbols, i.e., inter-symbol interference (ISI), or caused by another interfering user, i.e., IUI. Next, more details on IC methods and their comparison for NOMA and ANOMA are discussed.

A. Successive Interference Cancellation: SIC

One possible solution to remove IUI is SIC which is commonly used in the uplink NOMA systems. As explained in the previous section, in SIC, the detected symbols of each user are reconstructed and subtracted from the received samples. Assuming the IC order of $\{1, 2, \cdots, K\}$ and error-free decoding, the effective SINR of the kth user can be calculated as:

$$\delta_k^{synch} = \frac{|h_k|^2 P_k}{\sum_{j=k+1}^{K} |h_j|^2 P_j + \sigma^2}, \tag{26}$$

where the first term in the denominator is the interference power caused by undecoded users with higher IC orders. Unlike NOMA where only one interfering symbol degrades the performance, using ANOMA, due to timing offset, multiple adjacent symbols cause the interference as well. Assuming the IC order of $\{1, 2, \cdots, K\}$ and error-free decoding, the effective SINR of the kth user with ANOMA is:

$$\delta_k^{asynch} = \frac{|h_k|^2 P_k}{\sum_{j=k+1}^{K} \eta_{\tau_{kj}} |h_j|^2 P_j + \sigma^2}, \tag{27}$$

where $\eta_{\tau_{kj}} = \sum_{n=-u}^{u} g_{\tau_{kj}}^2(n)$ indicates the accumulative interference factor caused by interfering symbols of User j. The factor u is the normalized truncation length, i.e. $u = \lfloor \frac{T_p}{2T} \rfloor$. Because $\eta_\tau < 1$ for $\tau \neq 0$, the asynchronous transmission increases the effective SINR and can improve the performance as verified by simulation results.

B. Trellis-Based Algorithms

Because of the memory and Toeplitz structure imposed by asynchronous transmission, trellis-based equalization methods are applicable to ANOMA. In more details, if we re-order the samples and define $\boldsymbol{y[m]} = (y_1[m], \cdots, y_K[m])^T$ and $\boldsymbol{s[m]} = (s_1[m], \cdots, s_K[m])^T$, the input-output relationship of the system can be presented in a matrix form as:

$$\begin{pmatrix} \boldsymbol{y[1]} \\ \boldsymbol{y[2]} \\ \vdots \\ \boldsymbol{y[N]} \end{pmatrix} = \begin{pmatrix} \boldsymbol{R'_0} & \boldsymbol{R'_{-1}} & \cdots & \boldsymbol{R'_{1-N}} \\ \boldsymbol{R'_1} & \boldsymbol{R'_0} & \cdots & \boldsymbol{R'_{2-N}} \\ \vdots & \ddots & \ddots & \vdots \\ \boldsymbol{R'_{N-1}} & \boldsymbol{R'_{N-2}} & \cdots & \boldsymbol{R'_0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\mathcal{H}} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{\mathcal{H}} & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{\mathcal{H}} \end{pmatrix} \begin{pmatrix} \boldsymbol{s[1]} \\ \boldsymbol{s[2]} \\ \vdots \\ \boldsymbol{s[N]} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\nu[1]} \\ \boldsymbol{\nu[2]} \\ \vdots \\ \boldsymbol{\nu[N]} \end{pmatrix}, \tag{28}$$

where $\boldsymbol{\mathcal{H}} = diag(h_1, \cdots, h_K)$ and $\boldsymbol{R'_i}$ is the $K \times K$ constructing sub-block whose elements are defined as:

$$\boldsymbol{R'_i}(l, k) = g(iT + (\tau_l - \tau_k)T). \tag{29}$$

Matrix $\boldsymbol{R'}$ is a block-Toeplitz matrix whose sub-blocks are not necessarily Toeplitz. Examples of matrix $\boldsymbol{R'}$ for rectangular pulse shapes with $K = 2$, $N = 3$, are shown below ($\tau_1 = 0, \tau_2 = 0.2$ on the left and $\tau_1 = 0, \tau_2 = 0.5$ on the right):

$$\boldsymbol{R'} = \left[\begin{array}{cc|cc|cc} 1 & 0.8 & 0 & 0 & 0 & 0 \\ 0.8 & 1 & 0.2 & 0 & 0 & 0 \\ \hline 0 & 0.2 & 1 & 0.8 & 0 & 0 \\ 0 & 0 & 0.8 & 1 & 0.2 & 0 \\ \hline 0 & 0 & 0 & 0.2 & 1 & 0.8 \\ 0 & 0 & 0 & 0 & 0.8 & 1 \end{array}\right], \quad \boldsymbol{R'} = \left[\begin{array}{cc|cc|cc} 1 & 0.5 & 0 & 0 & 0 & 0 \\ 0.5 & 1 & 0.5 & 0 & 0 & 0 \\ \hline 0 & 0.5 & 1 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & 1 & 0.5 & 0 \\ \hline 0 & 0 & 0 & 0.5 & 1 & 0.5 \\ 0 & 0 & 0 & 0 & 0.5 & 1 \end{array}\right]. \tag{30}$$

Note that if the time delays are equi-spaced, i.e., $\tau_k = (k-1)\frac{T}{K},\ k = 1, \cdots, K$, then the matrix $\boldsymbol{R'}$ will turn into a Toeplitz matrix. The resulting Toeplitz structure caused by equi-spaced timing offsets enables the use of trellis-based algorithms such as the Viterbi and BCJR algorithms. Equivalently, the Toeplitz system model is:

$$y_l = \sum_{n=-u'}^{n=u'} r_n h_{\pi(l-n)} s_{l-n} + \nu_l, \quad l = 1, \cdots, NK, \tag{31}$$

where r_n represents the diagonal elements of Toeplitz matrix $\boldsymbol{R'}$, h_k, s_l and ν_l represent the channel coefficients, transmitted symbols and noise samples, respectively, with a proper mapping. To be more precise, $y_l = y_{\pi(l)}[\varphi(l)]$, $r_n = g(-n\frac{T}{K})$, $s_l = s_{\pi(l)}[\varphi(l)]$ and $\nu_l = \nu_{\pi(l)}[\varphi(l)]$ where $\pi(l) = (l-1 \mod K) + 1$ and $\varphi(l) = \lfloor {}^{(l-1)}/_{K} \rfloor + 1$. In addition, u' depends on the truncation length and the number of users, i.e., $u' = (K-1)(u+1)/2$. The model in (31) is commonly referred to as the Ungerboeck model in which the noise samples are not white, but are correlated according to $E[\nu_{l+n}\nu_l^*] = \sigma^2 r_n$. Normally, the samples are filtered by a noise whitening filter to generate a model with white noise, referred to as the Forney model [30].

To avoid the complex process of noise whitening, we directly use the Ungerboeck model to apply trellis-based algorithms which is novel in the context of asynchronous transmission. By representing the received signal vector $\boldsymbol{y}$ by an orthonormal basis and denoting its vector representation by $\hat{\boldsymbol{y}}$, it is possible to express [31]:

$$p(\hat{\boldsymbol{y}}|\boldsymbol{s}) \propto \prod_l P_l(s_l, s_{l-1}, \cdots, s_{l-u'}), \tag{32}$$

where

$$P_l(s_l, s_{l-1}, \cdots, s_{l-u'}) = \exp\left[\frac{1}{\sigma^2} Re\left\{ y_l h^*_{\pi(l)} s^*_l - \frac{1}{2}|h_{\pi(l)}|^2 |s_l|^2 r_0 - h^*_{\pi(l)} s^*_l \sum_{n=1}^{u'} s_{l-n} h_{\pi(l-n)} r_l \right\}\right]. \tag{33}$$

Therefore, the likelihood function has a recursive factorization that can be expressed in terms of the y_l samples. We observe that only the u' most recent input symbols $s_{l-1}, \cdots, s_{l-u'}$ are required at each time epoch l. Thus, the signal can be described by means of a trellis where each state is defined as $\rho_l = (s_{l-1}, \cdots, s_{l-u'})$. As an example, for BPSK modulation and $u' = 2$, a section of the corresponding trellis between the discrete-time instants l and $l+1$ is shown in Fig. 10. The Viterbi algorithm or BCJR can be applied to the trellis shown in Fig. 10 in which the branch metrics are calculated based on the recursive factorization in (33) [32]. Interested readers can refer to [30] for more details on the Ungerboeck model.

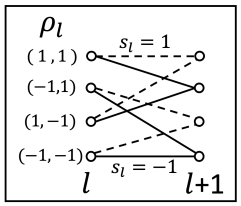


Fig. 10: Trellis Representation.

Although we avoid the noise whitening process by using the Ungerboeck model, the trellis-based algorithms' computational complexity is determined by the number of trellis states, which grows exponentially with the memory of the effective ISI channel. To reduce the receiver's computational complexity, we also propose another transceiver design, which is inspired by the channel diagonalization used in the capacity derivation in Appendix A.

C. Optimal Transceiver

For notational simplicity, we present the optimal transceiver design for a two-user case but the extension to more users is straightforward. Recalling the system model:

$$\begin{pmatrix} \boldsymbol{y_1} \\ \boldsymbol{y_2} \end{pmatrix} = \begin{pmatrix} \boldsymbol{R_{11}} & \boldsymbol{R_{12}} \\ \boldsymbol{R_{21}} & \boldsymbol{R_{22}} \end{pmatrix} \begin{pmatrix} h_1 \boldsymbol{I_N} & \boldsymbol{0_N} \\ \boldsymbol{0_N} & h_2 \boldsymbol{I_N} \end{pmatrix} \begin{pmatrix} \boldsymbol{s_1} \\ \boldsymbol{s_2} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\nu_1} \\ \boldsymbol{\nu_2} \end{pmatrix}, \tag{34}$$

the diagonal sub-blocks, $\boldsymbol{R_{kk}}$s, are identity matrices and off-diagonal sub-blocks are Toeplitz matrices which can be decomposed using the singular value decomposition, i.e., $\boldsymbol{R_{12}} = \boldsymbol{R_{21}^T} = \boldsymbol{U G_\tau V}^T$, where $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices and $\boldsymbol{G_\tau}$ is an $N \times N$ diagonal matrix with singular values g_is as its diagonal elements. Users 1 and 2 can apply precoding to send $\boldsymbol{s_1} = \boldsymbol{U P_1 x_1}$ and $\boldsymbol{s_2} = \boldsymbol{V P_2 x_2}$, respectively. After processing the received samples to construct $\hat{\boldsymbol{y}}_\mathbf{1} = \boldsymbol{U}^T \boldsymbol{y_1}$ and $\hat{\boldsymbol{y}}_\mathbf{2} = \boldsymbol{V}^T \boldsymbol{y_2}$, we have:

$$\begin{pmatrix} \hat{\boldsymbol{y}}_\mathbf{1} \\ \hat{\boldsymbol{y}}_\mathbf{2} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I_N} & \boldsymbol{G_\tau} \\ \boldsymbol{G_\tau} & \boldsymbol{I_N} \end{pmatrix} \begin{pmatrix} h_1 \boldsymbol{P_1} & \boldsymbol{0_N} \\ \boldsymbol{0_N} & h_2 \boldsymbol{P_2} \end{pmatrix} \begin{pmatrix} \boldsymbol{x_1} \\ \boldsymbol{x_2} \end{pmatrix} + \begin{pmatrix} \hat{\boldsymbol{\nu}}_\mathbf{1} \\ \hat{\boldsymbol{\nu}}_\mathbf{2} \end{pmatrix}, \tag{35}$$

where $\hat{\boldsymbol{\nu}}_\mathbf{1} = \boldsymbol{U}^T \boldsymbol{\nu_1}$ and $\hat{\boldsymbol{\nu}}_\mathbf{2} = \boldsymbol{V}^T \boldsymbol{\nu_2}$. Thus, $\mathbb{E}[\hat{\boldsymbol{\nu}}_\mathbf{1} {\hat{\boldsymbol{\nu}}_\mathbf{1}}^H] = \mathbb{E}[\hat{\boldsymbol{\nu}}_\mathbf{2} {\hat{\boldsymbol{\nu}}_\mathbf{2}}^H] = \sigma^2 \boldsymbol{I_N}$ and $\mathbb{E}[\hat{\boldsymbol{\nu}}_\mathbf{1} {\hat{\boldsymbol{\nu}}_\mathbf{2}}^H] = \mathbb{E}[\hat{\boldsymbol{\nu}}_\mathbf{2} {\hat{\boldsymbol{\nu}}_\mathbf{1}}^H] = \sigma^2 \boldsymbol{G_\tau}$. Due to the diagonalization of the channel sub-matrices and noise covariance matrices, to detect $x_1[n]$ and $x_2[n]$, it is sufficient to use the following set of samples:

$$\begin{pmatrix} \hat{y}_1[n] \\ \hat{y}_2[n] \end{pmatrix} = \begin{pmatrix} 1 & g_n \\ g_n & 1 \end{pmatrix} \begin{pmatrix} h_1 P_1[n] & 0 \\ 0 & h_2 P_2[n] \end{pmatrix} \begin{pmatrix} x_1[n] \\ x_2[n] \end{pmatrix} + \begin{pmatrix} \hat{\nu}_1[n] \\ \hat{\nu}_2[n] \end{pmatrix}, \tag{36}$$

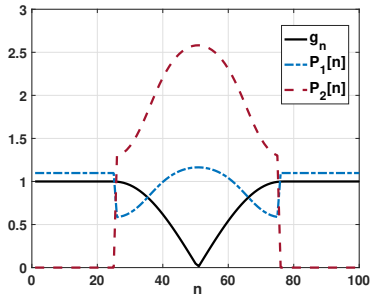


Fig. 11: Power Allocation Functions $\boldsymbol{P_1}$ and $\boldsymbol{P_2}$ for r.r.c. pulse shape with $\beta = 0.5$, $\tau = 0.5$, $P_1 = P_2 = 0\ dB$, $\sigma_1^2 = 2$ and $\sigma_2^2 = 1$.

where the covariance matrix of the noise vector is $\begin{pmatrix} 1 & g_n \\ g_n & 1 \end{pmatrix}$.

Remark 1: Intuitively, the frequency-selective channel of the interfering user is transformed into multiple single-tap subchannels where the strength of each subchannel is denoted by g_n. Equivalently, the system model in (36) can be interpreted as an interference channel with a direct-subchannel gain of 1 and a cross-subchannel gain of $g_n \leq 1$. Notably, the cross-subchannel gain in the synchronous transmission system is always equal to 1. Therefore, depending on the used pulse shape, the cross-subchannel gain can be reduced using asynchronous transmission. An example of cross-subchannel gains are shown in Fig. 11.

Remark 2: In the trellis-based methods presented previously, every transmitted symbol affects up to u' output samples. However, with diagonalization, every transmitted symbol only contributes to two output samples, reducing the detection complexity. The ML criterion can be maximized over two transmitted symbols:

$$\max_{x_1[n],x_2[n]\in\mathbb{S}} 2Re\{\hat{y}_1^*[n]\tilde{x}_1[n] + \hat{y}_2^*[n]\tilde{x}_2[n]\} - |\tilde{x}_1[n]|^2 - |\tilde{x}_2[n]|^2 - 2Re\{\tilde{x}_1^*[n]\tilde{x}_2^*[n]g_n\},\ \forall n, \quad (37)$$

where $\tilde{x}_k[n] = h_k P_k[n] x_k[n]$. The last term in the maximization shows the inter-dependence of interfering symbols. As g_n reduces, the inter-dependency reduces. Particularly, with $g_n = 0$, symbols can be detected separately without any interference. Note that for higher-order constellations, simpler detection methods such as zero-forcing (ZF) are also applicable to ANOMA while they are not relevant for NOMA with single antennas.

Remark 3: Based on the capacity-achieving power allocation scheme, presented in the previous section, more power is allocated to subchannels with smaller cross-subchannel gains, i.e., g_n. Power allocation is particularly important for the weaker user which avoids assigning power

to the subchannels with cross-subchannel gains of $g_n = 1$ and concentrates more power to the subchannels with smaller cross-subchannel gains. As an example, the power allocation functions for a two-user scenario with r.r.c. pulse shape $\beta = 0.5$, $\tau = 0.5$, $P_1 = P_2 = 0\ dB$, $\sigma_1^2 = 2$ and $\sigma_2^2 = 1$ are shown in Fig. 11.

After IC, the soft or hard information is passed to the de-mapping and de-interleaving modules. Finally, the information on coded bits is passed to the decoder to decode the information bits, as shown in Fig. 9b. Although the separation of the IC and decoding processes reduces the receiver complexity, it results in substantial performance loss compared with the optimal receiver formulated in (25). To recover the loss, and approach the capacity boundaries, we employ the turbo principle. The turbo principle is based on the exchange of extrinsic information between the IC and the decoder [33], [34].

D. Complexity Analysis

The complexity order of proposed IC methods including SIC, trellis-based and the ML-based design are compared in Table I. The complexity of the SIC method is $\mathcal{O}(KM)$ per

TABLE I: Comparison of the complexity of IC Methods

Methods	NOMA	ANOMA
SIC	$\mathcal{O}(KM)$	$\mathcal{O}(KM)$
trellis-based	NA	$\mathcal{O}(M^{\frac{(K-1)(u+1)}{2}})$
ML-based	$\mathcal{O}\left(KM^K\right)$	$\mathcal{O}\left(KM^K\right) + \mathcal{O}(KN)$

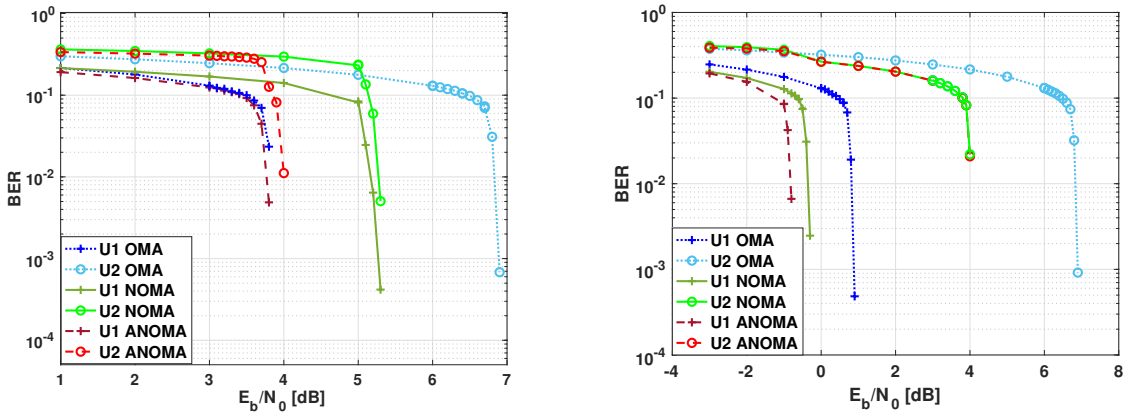
symbol where K represents the number of users and M represents the constellation size. The SIC method has the same complexity for both NOMA and ANOMA schemes. However, the ANOMA scheme with SIC outperforms the NOMA scheme with SIC as explained in Section IV-A. The complexity of the trellis-based algorithms such as Viterbi and BCJR is proportional to the number of trellis states, which grows exponentially with the memory size of the effective ISI channel [34]. The memory size of the effective ISI channel depends on the number of users and the truncation length of the pulse shapes. As explained in Section IV-B, the number of states in the introduced system model is $M^{\frac{(K-1)(u+1)}{2}}$ where u represents the normalized truncation length of the pulse shape, i.e., $u = \lfloor \frac{T_p}{2T} \rfloor$. Thus, the complexity order of the trellis-based methods can be expressed as $\mathcal{O}(M^{\frac{(K-1)(u+1)}{2}})$ per symbol. Trellis-based methods can be applied

to ANOMA due to introduced memory by asynchronous transmission and are not applicable to NOMA. Various implementation techniques exist in the literature to reduce the computational complexity of the trellis-based methods such as reduced state estimation methods [35]. The ML-based methods exhibit complexity orders of $\mathcal{O}\left(KM^K\right)$ and $\mathcal{O}\left(KM^K\right)+\mathcal{O}\left(KN\right)$ for NOMA and ANOMA, respectively. The additional complexity of $\mathcal{O}\left(KN\right)$ for ANOMA corresponds to the channel diagonalization explained in Section IV-C. The channel diagonalization includes K matrix multiplications, which each can be performed with a complexity of $\mathcal{O}\left(N^2\right)$, resulting in the overall complexity of $\mathcal{O}\left(KN\right)$ per symbol.

V. Numerical Results

In this section, we present numerical results to verify our analysis for NOMA and ANOMA systems. We utilize the LDPC codes with length 64,800 from the DVB-S2X standard [36] for channel coding. With the exception of the last set of simulations with a range of coding rates, we use the rate $r=1/2$. To improve the performance and reduce the dependence among transmitted symbols, we use an interleaver to scramble the coded bits, executed by the common Mersenne Twister algorithm [29]. After interleaving, the scrambled coded bits are mapped to constellation symbols, e.g., BPSK and QPSK, to generate transmitted symbols after proper pulse shaping. The r.r.c. with roll-off factor $\beta=0.5$ is considered for pulse shaping and $T=10\ \mu s$ which results in occupied bandwidth of $B=150\ MHz$. The pulse shape is truncated to include 4 significant side lobes and the timing offset is assumed to be $\tau=0.5$.

First, we present results for the sub-optimal separate detection schemes. Different types of IC schemes including the well-known SIC, trellis-based algorithms, and ML methods are considered. In Fig. 12, the SIC methods are compared for OMA, NOMA, and ANOMA. The NOMA method exploits the difference in the channel qualities; as the difference between channel coefficients increases, the provided gain improves which is confirmed by Fig. 12. In Fig. 12a, with $|h_1|^2=2, |h_2|^2=1$, NOMA's performance for the stronger user suffers from interference and is worse than that of the OMA method. However, as the difference between channel coefficients increases, i.e., $|h_1|^2=4, |h_2|^2=1$, the NOMA's performance improves for both users since the interference for decoding User 1's message is reduced and thus the effect of error propagation is decreased for decoding User 2's message. On the other hand, the ANOMA method performs well in both scenarios showing its capability even when the channel coefficients are close. As the difference between channel coefficients increases, the performance gap between NOMA and



(a) $|h_1|^2 = 2, |h_2|^2 = 1$ (b) $|h_1|^2 = 4, |h_2|^2 = 1$

Fig. 12: Comparing SIC method for OMA, NOMA and ANOMA methods.

ANOMA methods is reduced. Particularly, their performance for the weaker user converge since the stronger user's message is decoded and removed with a very small error, reducing the effect of error propagation.

Trellis-based algorithms can exploit the favorable ISI caused by asynchronous transmission. We apply the well-known BCJR algorithm for ANOMA, which is compared with the ML NOMA method in Fig. 13. The trellis-based ANOMA method provides around $1.5\ dB$ gain for both stronger and weaker users compared with the ML NOMA method. The optimal ANOMA transceiver employs the channel diagonalization and power allocation to further improve the performance and reduce complexity. The weaker user only utilizes half of the available subchannels and uses QPSK modulation to produce the same bit rate. To show the strength of the proposed transceiver, we suffice to use the capacity-achieving power allocation scheme, although more efficient BER-minimizing power allocations can be used. Since half of the subchannels are occupied solely by the stronger user, the receiver enjoys interference-free detection with lowered computational complexity. The resulting transceiver design provides around $2.5\ dB$ and $5\ dB$ gains for the stronger user and weaker user, respectively, compared with the NOMA method. In Fig. 14, the performance of the introduced ML-ANOMA is compared with ML-NOMA for a 3-user scenario with $|h_1|^2 = 10$, $|h_2|^2 = 4$ and $|h_3|^2 = 1$. The ANOMA transceiver employs channel diagonalization as explained in Section IV-C, however, with no power allocation. It verifies that even without power allocation, the asynchronous transmission

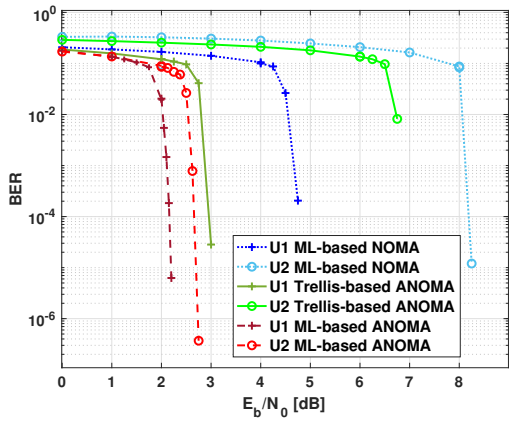


Fig. 13: Comparison of the ML NOMA, trellis-based ANOMA, and optimal ANOMA transceivers with $|h_1|^2 = 2,\ |h_2|^2 = 1$.

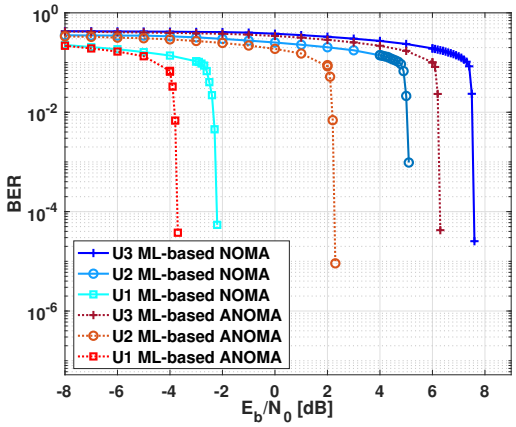


Fig. 14: Comparison of the ML ANOMA and ML NOMA transceivers with $|h_1|^2 = 10,\ |h_2|^2 = 4$ and $|h_3|^2 = 1$.

can be beneficial. The ANOMA transceiver based on channel diagonalization and ML detection provides around $2\ dB$ gain for each user.

In Fig. 15, a comparison between the performance of the ML NOMA and the proposed optimal ANOMA is presented considering the turbo principle. We apply up to 15 turbo iterations and the results show that the turbo principle improves the BER performance, particularly for the weaker user. The reason is that, as soon as the stronger user is decoded, the error-free feedback from the decoder can improve the performance of the weaker user substantially. The turbo iterations provide around $3.5\ dB$ and $1.5\ dB$ gain for the weaker user in the ML NOMA

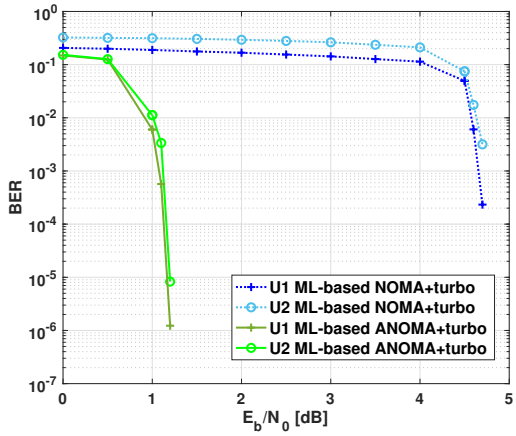


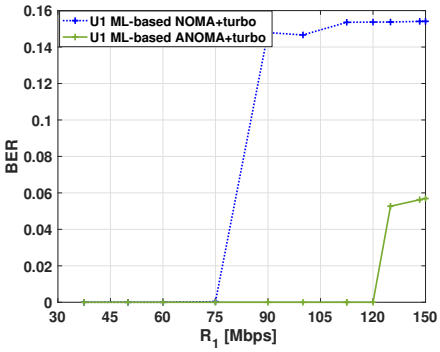
Fig. 15: Comparing the ML NOMA and the optimal ANOMA transceivers using turbo principle with $|h_1|^2 = 2,\ |h_2|^2 = 1$.

and the optimal ANOMA, respectively. Note that the proposed transceiver design for ANOMA substantially improves the performance of SIC ANOMA which verifies the sub-optimality of SIC in ANOMA.

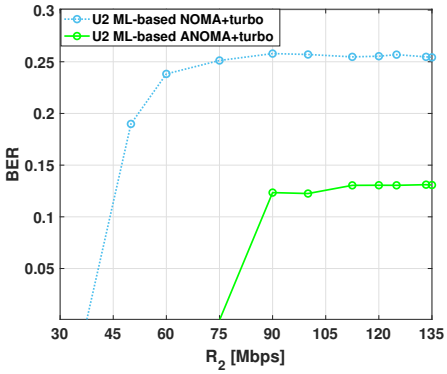
In Fig. 16, the performance of the synchronous and the asynchronous transmission is compared for various transmission rates including $r = [\frac{1}{4}, \frac{1}{3}, \frac{2}{5}, \frac{1}{2}, \frac{3}{5}, \frac{2}{3}, \frac{3}{4}, \frac{4}{5}, \frac{5}{6}, \frac{8}{9}, \frac{9}{10}]$ with $P_1 = P_2 = 1.76\ dB$ and $|h_1|^2 = 2,\ |h_2|^2 = 1$. In Fig. 16a, the rate of the weaker user is set to $R_2 = 75\ Mbps$ and the BER performance of the stronger user is shown with respect to various transmission rates. ANOMA can achieve up to $R_1 = 120\ Mbps$ while NOMA only achieves up to $R_1 = 75\ Mbps$. In Fig. 16b, the stronger user transmits with $R_1 = 120\ Mbps$ and the BER performance of the weaker user is shown with respect to various transmission rates. ANOMA can support around $R_2 = 75\ Mbps$ while NOMA only supports around $R_2 = 45\ Mbps$. Thus, the asynchronous transmission can achieve the rate pair $R_1 = 120\ Mbps,\ R_2 = 75\ Mbps$, shown with a star in Fig. 16c, which is not achievable by the synchronous transmission. These results verify the capacity analysis and shows the effectiveness of the asynchronous transmission.

A. Final Remarks

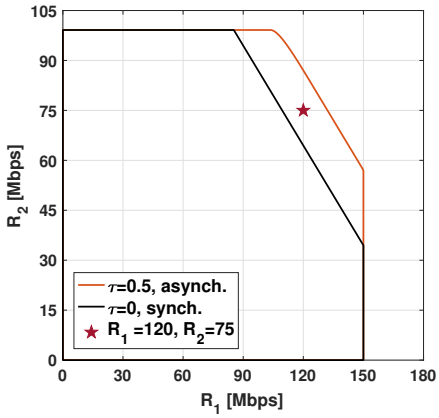
- In this work, we assume flat fading channels, applicable to scenarios with no scattering such as satellite communication [37]. In [27], it is shown that for high scattering environments, the effect of asynchronous transmission is negligible. However, as the power of the dominant path compared with the scattering paths is increased, the effect of asynchronous transmission



(a) $R_2 = 75\ Mbps$



(b) $R_1 = 120\ Mbps$



(c) Operational point achieved by asynchronous transmission which is not achievable by synchronous transmission

Fig. 16: Performance comparison of synchronous and asynchronous transmission for various operational rate points with $P_2 = P_1 = 1.76\ dB$ and $|h_1|^2 = 2,\ |h_2|^2 = 1$.

is intensified. As an alternative, frequency offsets can be intentionally added to induce time-selectivity to improve the performance [17].

- The focus of this work is on two-user scenarios since in the context of NOMA, it is common to cluster users into groups to maintain low complexity. For example, in [38], it is shown that two layers of superposition coding is a good compromise between prospected gains and added complexity.
- The optimal timing offset is shown to be $\tau = 0.5$ for a two-user scenario. It is also verified in [12] that the equi-spaced timing offsets are advantageous for any number of users. Thus,

in a MAC, the timing offsets can be predetermined and be controlled by control channels. In this work, we assume perfect knowledge of timing offsets, however, the effects of timing error in performance are analyzed in [15] where it is shown that the performance of both NOMA and ANOMA degrades comparably by timing errors.

VI. Conclusion

In this work, we thoroughly analyzed the rate-region provided by the asynchronous transmission in multiple access channels. We proved that the capacity-region of asynchronous transmission is larger than that of the synchronous transmission as long as the pulse shape has the ONB spectrum. We studied the well-known SIC method and compared its performance for NOMA and ANOMA. We also demonstrated that asynchronous transmission introduces memory in the system model, which makes trellis-based algorithms applicable. The introduced Toeplitz structure resembles frequency-selective channels, which can be exploited by proper channel diagonalization and power allocation. The proposed optimal transceiver design joint with the turbo principle can provide transmission rates close to the capacity boundary. We showed theoretically and numerically that the proposed asynchronous method could achieve the operational rates that are not achievable by the synchronous transmission.

Appendix A
Proof of Theorem 1

The system model in Eq. (3) can be re-written for a two-user scenario as:

$$\boldsymbol{y}=\begin{pmatrix}\boldsymbol{R_{11}} & \boldsymbol{R_{12}}\\ \boldsymbol{R_{21}} & \boldsymbol{R_{22}}\end{pmatrix}\begin{pmatrix}h_1\boldsymbol{I_N} & \boldsymbol{0_N}\\ \boldsymbol{0_N} & h_2\boldsymbol{I_N}\end{pmatrix}\begin{pmatrix}\boldsymbol{s_1}\\ \boldsymbol{s_2}\end{pmatrix}+\begin{pmatrix}\boldsymbol{\nu_1}\\ \boldsymbol{\nu_2}\end{pmatrix}=\boldsymbol{RHs}+\boldsymbol{\nu}. \tag{38}$$

The mutual information $I(\boldsymbol{y};\boldsymbol{s_1},\boldsymbol{s_2})$ can be upper-bounded by:

$$I(\boldsymbol{y};\boldsymbol{s_1},\boldsymbol{s_2})\leq\frac{1}{2}\log\det[2\pi e\ cov(\boldsymbol{y})]-\frac{1}{2}\log\det[2\pi e\ cov(\boldsymbol{\nu})], \tag{39}$$

where $cov(\boldsymbol{y})=\boldsymbol{RH}\mathbb{E}[\boldsymbol{ss}^H]\boldsymbol{H}^H\boldsymbol{R}+\boldsymbol{R}\sigma^2$ and $cov(\boldsymbol{\nu})=\boldsymbol{R}\sigma^2$. Therefore,

$$I(\boldsymbol{y};\boldsymbol{s_1},\boldsymbol{s_2})\leq\frac{1}{2}\log\det\left[\boldsymbol{I_{2N}}+\frac{1}{\sigma^2}\begin{pmatrix}|h_1|^2\boldsymbol{Q_{s_1}} & \boldsymbol{0_N}\\ \boldsymbol{0_N} & |h_2|^2\boldsymbol{Q_{s_2}}\end{pmatrix}\begin{pmatrix}\boldsymbol{R_{11}} & \boldsymbol{R_{12}}\\ \boldsymbol{R_{21}} & \boldsymbol{R_{22}}\end{pmatrix}\right], \tag{40}$$

where $\boldsymbol{Q_{s_k}}$ is the covariance matrix of the input process for User k. The upper-bound is achieved by Gaussian processes. Matrices $\boldsymbol{R_{11}}$ and $\boldsymbol{R_{22}}$ are equal to identity matrices and $\boldsymbol{R_{12}}$ and $\boldsymbol{R_{21}}$ are Toeplitz matrices. Let us decompose $\boldsymbol{R_{12}}$, that depends on the pulse shape and the timing offset, using the singular value decomposition, i.e., $\boldsymbol{R_{12}}=\boldsymbol{R_{21}^T}=\boldsymbol{UG_\tau V}^T$, where $\boldsymbol{U}$ and $\boldsymbol{V}$

are orthogonal matrices and $\boldsymbol{G_\tau}$ is an $N \times N$ diagonal matrix consisting of the singular values of $\boldsymbol{R_{12}}$. Hence, we can simplify the upper-bound using:

$$\begin{aligned}
&\det\left[\boldsymbol{I_{2N}} + \frac{1}{\sigma^2}\begin{pmatrix} |h_1|^2\boldsymbol{Q_{s_1}} & \boldsymbol{0_N} \\ \boldsymbol{0_N} & |h_2|^2\boldsymbol{Q_{s_2}}\end{pmatrix}\begin{pmatrix}\boldsymbol{R_{11}} & \boldsymbol{R_{12}} \\ \boldsymbol{R_{21}} & \boldsymbol{R_{22}}\end{pmatrix}\right] \qquad (41)\\
&= \det\left[\begin{pmatrix}\boldsymbol{U}^T & \boldsymbol{0_N} \\ \boldsymbol{0_N} & \boldsymbol{V}^T\end{pmatrix}\right]\det\left[\boldsymbol{I_{2N}} + \frac{1}{\sigma^2}\begin{pmatrix} |h_1|^2\boldsymbol{Q_{s_1}} & \boldsymbol{0_N} \\ \boldsymbol{0_N} & |h_2|^2\boldsymbol{Q_{s_2}}\end{pmatrix}\begin{pmatrix}\boldsymbol{I_N} & \boldsymbol{R_{12}} \\ \boldsymbol{R_{21}} & \boldsymbol{I_N}\end{pmatrix}\right]\det\left[\begin{pmatrix}\boldsymbol{U} & \boldsymbol{0_N} \\ \boldsymbol{0_N} & \boldsymbol{V}\end{pmatrix}\right] \\
&= \det\left[\boldsymbol{I_{2N}} + \frac{1}{\sigma^2}\begin{pmatrix} |h_1|^2\boldsymbol{U}^T\boldsymbol{Q_{s_1}} & \boldsymbol{0_N} \\ \boldsymbol{0_N} & \boldsymbol{V}^T|h_2|^2\boldsymbol{Q_{s_2}}\end{pmatrix}\begin{pmatrix}\boldsymbol{U} & \boldsymbol{R_{12}V} \\ \boldsymbol{R_{21}U} & \boldsymbol{V}\end{pmatrix}\right] \\
&= \det\left[\boldsymbol{I_{2N}} + \frac{1}{\sigma^2}\begin{pmatrix} |h_1|^2\boldsymbol{U}^T\boldsymbol{Q_{s_1}U} & |h_1|^2\boldsymbol{U}^T\boldsymbol{Q_{s_1}U}\boldsymbol{G_\tau} \\ |h_2|^2\boldsymbol{V}^T\boldsymbol{Q_{s_2}}\boldsymbol{G_\tau} & |h_2|^2\boldsymbol{V}^T\boldsymbol{Q_{s_2}V}\end{pmatrix}\right] = \det\left[\boldsymbol{I_{2N}} + \frac{1}{\sigma^2}\begin{pmatrix} |h_1|^2\boldsymbol{\mathcal{S}_1} & \boldsymbol{0_N} \\ \boldsymbol{0_N} & |h_2|^2\boldsymbol{\mathcal{S}_2}\end{pmatrix}\begin{pmatrix}\boldsymbol{I_N} & \boldsymbol{G_\tau} \\ \boldsymbol{G_\tau} & \boldsymbol{I_N}\end{pmatrix}\right],
\end{aligned}$$

where $\boldsymbol{\mathcal{S}_1} = \boldsymbol{U}^T\boldsymbol{Q_{s_1}U}$ and $\boldsymbol{\mathcal{S}_2} = \boldsymbol{V}^T\boldsymbol{Q_{s_2}V}$. To further simplify the upper-bound, we use Lemma 2 in [9], which states:

Lemma 1: Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be $N \times N$ non-negative-definite matrices, and let $\boldsymbol{G} = diag[g_1, \cdots, g_N]$ where $|g_n| \leq 1, n = 1, \cdots, N$. Then,

$$\det\left[\boldsymbol{I_{2N}} + \begin{pmatrix}\boldsymbol{A} & \boldsymbol{0_N} \\ \boldsymbol{0_N} & \boldsymbol{B}\end{pmatrix}\begin{pmatrix}\boldsymbol{I_N} & \boldsymbol{G} \\ \boldsymbol{G} & \boldsymbol{I_N}\end{pmatrix}\right] \leq \prod_{n=1}^{N}\left(1 + a_{nn} + b_{nn} + a_{nn}b_{nn}(1 - g_n^2)\right), \qquad (42)$$

where a_{nn} and b_{nn} are the diagonal elements of matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, respectively. The equality is achieved when $\boldsymbol{A}$ and $\boldsymbol{B}$ are diagonal.

As a result, the upper-bound is achieved if $\boldsymbol{\mathcal{S}_1}$ and $\boldsymbol{\mathcal{S}_2}$ are diagonal, or equivalently, $\boldsymbol{Q_{s_1}}$ and $\boldsymbol{Q_{s_2}}$ are eigen-decomposed by $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. Denoting the diagonal elements of $\boldsymbol{\mathcal{S}_1}$ and $\boldsymbol{\mathcal{S}_2}$ as s_{1n} and s_{2n}, respectively, we have:

$$R_1 + R_2 \leq \frac{1}{2}\lim_{N\to\infty}\sum_{n=1}^{N}\log\left(1 + \frac{|h_1|^2}{\sigma^2}s_{1n} + \frac{|h_2|^2}{\sigma^2}s_{2n} + \frac{|h_1|^2|h_2|^2}{\sigma^4}s_{1n}s_{2n}(1 - g_n^2)\right)\frac{1}{N}. \qquad (43)$$

Toeplitz matrices are asymptotically equivalent to circulant matrices as the matrix dimension goes to infinity [23], [39]. The implication of the asymptotic equivalence of Toeplitz matrices with circular matrices is that the values of the singular values of Toeplitz matrices are asymptotically equal to samples of their generating function. In more details, considering a Toeplitz matrix, $\boldsymbol{R}$, its generating function, $R(f), f \in [0, 1]$, and its singular values, $r_n, n = 1, \cdots, N$, we have $r_n = |R(n/N)|, n = 1, \cdots, N$ [24].

Defining $f_n = n/N$, $df_N = 1/N$, $S_k(f_n) = s_{kn}$, and $G_\tau(f_n) = g_n$, we can rewrite the sum-rate upper-bound as $C = \frac{1}{2}\lim_{N\to\infty}\sum_{n=1}^{N}C(f_n)df_N$ where $C(f_n) = \log\left(1 + \frac{|h_1|^2}{\sigma^2}S_1(f_n) + \frac{|h_2|^2}{\sigma^2}S_2(f_n) + \frac{|h_1|^2|h_2|^2}{\sigma^4}S_1(f_n)S_2(f_n)(1 - G_\tau^2(f_n))\right)$. Because

$C(f_n)$ is bounded and almost everywhere continuous on the interval $[0,1]$, then it is Reimann integrable on the interval [40], [41], and therefore:

$$R_1+R_2\leq\frac{1}{2}\int_0^1\log\left(1+\frac{|h_1|^2}{\sigma^2}S_1(f)+\frac{|h_2|^2}{\sigma^2}S_2(f)+\frac{|h_1|^2|h_2|^2}{\sigma^4}S_1(f)S_2(f)(1-G_\tau^2(f))\right)df, \tag{44}$$

where $S_1(f)$ and $S_2(f)$ are PSDs of Users 1 and 2, respectively. In addition, $G_\tau(f)=|R_{12}(f)|=|R_{12}(f)|$ where $R_{12}(f)$ and $R_{12}(f)$ are the generating functions of Toeplitz matrices $\boldsymbol{R_{12}}$ and $\boldsymbol{R_{21}}$, respectively. Therefore, we have:

$$C=\bigcup_{\substack{S_k(f)\geq 0,\ k=1,2\\ \int_0^1 S_k(f)df\leq P_k}}\left\{(R_1,R_2),\ \begin{array}{c}0\leq R_1\leq\frac{1}{2}\int_0^1\log_2\left(1+\frac{S_1(f)}{\sigma_1^2}\right)df\\ 0\leq R_2\leq\frac{1}{2}\int_0^1\log_2\left(1+\frac{S_2(f)}{\sigma_2^2}\right)df\\ 0\leq R_1+R_2\leq\frac{1}{2}\int_0^1\log\left(1+\frac{S_1(f)}{\sigma_1^2}+\frac{S_2(f)}{\sigma_2^2}+\frac{S_1(f)S_2(f)(1-G_\tau^2(f))}{\sigma_1^2\sigma_2^2}\right)df\end{array}\right\},$$

where the union is over all the PSDs that satisfy users' power constraints. The provided capacity-region settles the proof of converse, setting upper-bounds on the achievable rate pairs. On the other hand, the achievability is verified using Lemma 1 which states that the upper-bounds in the capacity-region can be achieved by Gaussian processes with covariance matrices $\boldsymbol{Q_{s_1}}=\boldsymbol{U\mathcal{S}_1U}^T$ and $\boldsymbol{Q_{s_2}}=\boldsymbol{V\mathcal{S}_2V}^T$ where $\boldsymbol{\mathcal{S}_1}$ and $\boldsymbol{\mathcal{S}_2}$ are diagonal. Further discussion on finding the optimal $\boldsymbol{\mathcal{S}_1}$, $\boldsymbol{\mathcal{S}_2}$ or equivalently users' optimal PSDs are provided throughout the manuscript.

To analyze $G_\tau(f)$, recall the structure of $\boldsymbol{R_{12}}$ as:

$$\boldsymbol{R_{12}}=\begin{pmatrix}g_\tau(0) & g_\tau(1) & \cdots & g_\tau(N-1)\\ g_\tau(-1) & g_\tau(0) & \ddots & \vdots\\ \vdots & \ddots & \ddots & g(1)\\ g_\tau(1-N) & \cdots & g_\tau(-1) & g_\tau(0),\end{pmatrix}, \tag{45}$$

where τ is the timing offset between two users and $g(t)$ is the matched filter pulse. Recall that $g_\tau(m)=g(\tau T+mT)$. The generating function of $\boldsymbol{R_{12}}$ is defined as $R_{12}(f)=\sum_{n=1-N}^{N-1}g_\tau(n)e^{-j2\pi nf}$ and is periodic with period 1. Equivalently, $R_{12}(f)$ is:

$$R_{12}(f)=\frac{1}{T}\sum_{i=-\infty}^{\infty}e^{-j2\pi\tau(f+i)}\hat{g}(\frac{f+i}{T}), \tag{46}$$

where $\hat{g}(f)$ is the Fourier transform of $g(t)$. Note that $R_{12}(f)$ can be interpreted as the folded-spectrum. The only difference is that each replica is phase shifted due to the timing offset between users. The concept of phase-shifted folded spectrum is shown in Fig. 2.

REFERENCES

[1] A. Wyner, "Recent results in the Shannon theory," *IEEE Transactions on Information Theory*, vol. 20, no. 1, pp. 2–10, Jan. 1974.

[2] T. M. Cover, "Some advances in broadcast channels," *Advances in Communication Systems, Elsevier*, vol. 4, pp. 229–260, Jan. 1975.

[3] J. Kazemitabar and H. Jafarkhani, "Performance analysis of multiple antenna multi-user detection," *Information Theory and Applications (ITA) Workshop, 2009*, pp. 150–159, 2009.

[4] K. Higuchi and A. Benjebbour, "Non-orthogonal multiple access (NOMA) with successive interference cancellation for future radio access," *IEICE Transactions on Communications*, vol. 98, no. 3, pp. 403–414, Mar. 2015.

[5] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 10, pp. 2181–2195, Jul. 2017.

[6] V. Kotzsch and G. Fettweis, "Interference analysis in time and frequency asynchronous network MIMO OFDM systems," *IEEE Wireless Communication and Networking Conference (WCNC)*, pp. 1–6, Apr. 2010.

[7] S. Abeywickrama, L. Liu, Y. Chi, and C. Yuen, "Over-the-air implementation of uplink NOMA," *IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, Dec. 2017.

[8] A. A. Nasir, S. Durrani, H. Mehrpouyan, S. D. Blostein, and R. A. Kennedy, "Timing and carrier synchronization in wireless communication systems: a survey and classification of research in the last 5 years," *EURASIP Journal on Wireless Communications and Networking*, no. 180, Aug. 2016, available online at: https://link.springer.com/article/10.1186/s13638-016-0670-9.

[9] S. Verdu, "The capacity region of the symbol-asynchronous Gaussian multiple-access channel," *IEEE Transactions on Information Theory*, vol. 35, no. 4, pp. 733–751, Jul. 1989.

[10] S. Shao, Y. Tang, T. Kong, K. Deng, and Y. Shen, "Performance analysis of a modified V-BLAST system with delay offsets using zero-forcing detection," *IEEE Transactions on Vehicular Technology*, vol. 56, no. 6, pp. 3827–3837, Nov. 2007.

[11] A. Das and B. D. Rao, "MIMO systems with intentional timing offset," *EURASIP Journal on Advances in Signal Processing*, vol. 2011, no. 1, pp. 1–14, Dec. 2011.

[12] M. Ganji and H. Jafarkhani, "Interference mitigation using asynchronous transmission and sampling diversity," *IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, Dec. 2016.

[13] L. Cottatellucci, R. R. Muller, and M. Debbah, "Asynchronous CDMA systems with random spreadingPart I: Fundamental limits," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1477–1497, Mar. 2010.

[14] J. Cui, G. Dong, S. Zhang, H. Li, and G. Feng, "Asynchronous NOMA for downlink transmissions," *IEEE Communications Letters*, vol. 21, no. 2, pp. 402–405, Oct. 2016.

[15] X. Zou, B. He, and H. Jafarkhani, "An analysis of two-user uplink asynchronous non-orthogonal multiple access systems," *IEEE Transactions on Wireless Communications*, vol. 18, no. 2, pp. 1404–1418, Jan. 2019.

[16] M. Ganji and H. Jafarkhani, "Time asynchronous NOMA for downlink transmission," *IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, Apr. 2019.

[17] ——, "Improving NOMA multi-carrier systems with intentional frequency offsets," *IEEE Wireless Communications Letters*, vol. 8, no. 4, pp. 1060–1063, Aug. 2019.

[18] M. Avendi and H. Jafarkhani, "Differential distributed space-time coding with imperfect synchronization in frequency-selective channels," *IEEE Transactions on Wireless Communications*, vol. 14, no. 4, pp. 1811–1822, Nov. 2014.

[19] S. Poorkasmaei and H. Jafarkhani, "Asynchronous orthogonal differential decoding for multiple access channels," *IEEE Transactions on Wireless Communications*, vol. 14, no. 1, pp. 481–493, Jan. 2015.

[20] X. Zhang, M. Ganji, and H. Jafarkhani, "Exploiting asynchronous signaling for multiuser cooperative networks with analog network coding," *IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, Mar. 2017.

[21] H. Haci, H. Zhu, and J. Wang, "Performance of non-orthogonal multiple access with a novel asynchronous interference cancellation technique," *IEEE Transactions on Communications*, vol. 65, no. 3, pp. 1319–1335, Jan. 2017.

[22] J. Liu, Y. Li, G. Song, and Y. Sun, "Detection and analysis of symbol-asynchronous uplink NOMA with equal transmission power," *IEEE Wireless Communications Letters*, vol. 8, no. 4, pp. 1069–1072, Mar. 2019.

[23] R. M. Gray, "Toeplitz and circulant matrices: A review," *Foundations and Trends® in Communications and Information Theory*, vol. 2, no. 3, pp. 155–239, 2006.

[24] Z. Zhu and M. B. Wakin, "On the asymptotic equivalence of circulant and Toeplitz matrices," *IEEE Transactions on Information Theory*, vol. 63, no. 5, pp. 2975–2992, Mar. 2017.

[25] S. Verdu, "Multiple-access channels with memory with and without frame synchronism," *IEEE Transactions on Information Theory*, vol. 35, no. 3, pp. 605–619, May 1989.

[26] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," *IEEE Vehicular Technology Conference (VTC)*, pp. 1–5, Jun. 2013.

[27] M. Ganji, X. Zou, and H. Jafarkhani, "Exploiting time asynchrony in multi-user transmit beamforming," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3156–3169, Feb. 2020.

[28] G. Caire, G. Taricco, and E. Biglieri, "Bit-interleaved coded modulation," *IEEE Transactions on Information Theory*, vol. 44, no. 3, pp. 927–946, May 1998.

[29] L. Szczecinski and A. Alvarado, *Bit-interleaved coded modulation: fundamentals, analysis and design*. John Wiley & Sons, 2015.

[30] G. Colavolpe and A. Barbieri, "On MAP symbol detection for ISI channels using the Ungerboeck observation model," *IEEE Communications Letters*, vol. 9, no. 8, pp. 720–722, Aug. 2005.

[31] G. Ungerboeck, "Adaptive maximum-likelihood receiver for carrier-modulated data-transmission systems," *IEEE Transactions on Communications*, vol. 22, no. 5, pp. 624–636, May 1974.

[32] F. Rusek, G. Colavolpe, and C. E. W. Sundberg, "40 years with the Ungerboeck model: A look at its potentialities [lecture notes]," *IEEE Signal Processing Magazine*, vol. 32, no. 3, pp. 156–161, Apr. 2015.

[33] M. Tuchler, R. Koetter, and A. C. Singer, "Turbo equalization: principles and new results," *IEEE Transactions on Communications*, vol. 50, no. 5, pp. 754–767, Aug. 2002.

[34] M. Tuchler and A. C. Singer, "Turbo equalization: An overview," *IEEE Transactions on Information Theory*, vol. 57, no. 2, pp. 920–952, Jan. 2011.

[35] M. V. Eyuboglu and S. U. Qureshi, "Reduced-state sequence estimation with set partitioning and decision feedback," *IEEE Transactions on Communications*, vol. 36, no. 1, pp. 13–20, Jan. 1988.

[36] P.-D. Arapoglou, A. Ginesi, S. Cioni, S. Erl, F. Clazzer, S. Andrenacci, and A. Vanelli-Coralli, "DVB-S2X-enabled precoding for high throughput satellite systems," *International Journal of Satellite Communications and Networking*, vol. 34, no. 3, pp. 439–455, May 2016.

[37] G. Maral, M. Bousquet, and Z. Sun, *Satellite communications systems: systems, techniques and technology*. John Wiley & Sons, 2020.

[38] A. Zafar, M. Shaqfeh, M.-S. Alouini, and H. Alnuweiri, "On multiple users scheduling using superposition coding over Rayleigh fading channels," *IEEE Communications Letters*, vol. 17, no. 4, pp. 733–736, Apr. 2013.

[39] E. E. Tyrtyshnikov, "A unifying approach to some old and new theorems on distribution and clustering," *Linear Algebra and its Applications, Elsevier*, vol. 232, pp. 1–43, Jan. 1996.

[40] W. Rudin, *Principles of mathematical analysis*. McGraw-hill New York, 1964.

[41] A. J. Goldsmith and M. Effros, "The capacity region of broadcast channels with intersymbol interference and colored Gaussian noise," *IEEE Transactions on Information Theory*, vol. 47, no. 1, pp. 219–240, Jan. 2001.