Link Level Performance Evaluation of MIMO

Yang Zhao

I. Introduction

Multiple-input multiple-output (MIMO) has become a popular approach to multiply the capacity, reduce the error rate and improve the spectrum efficiency for modern communication systems. The data rate can be enhanced by spatial multiplexing (SM) that transmits independent streams on different antennas, while the space-time block coding (STBC) encodes message blocks on multiple transmitters over several time slots to reduce the error rate. This article computes the analytical capacity of typical 2-by-2 MIMO channels based on water-filling power allocation, simulates the ergodic capacity of different systems with full (CSIT) and partial channel knowledge (CDIT) at the transmitter, and evaluate the bit error rate (BER) and diversity gain performance of SM with maximum-likelihood (ML), zero-forcing (ZF) and unordered ZF successive interference cancelling (SIC) receivers. It is assumed that the transmission is over i.i.d. Rayleigh fading channels and the channel information at the receiver is fully known (CSIR).

II. Theory and Methods

A. MIMO System

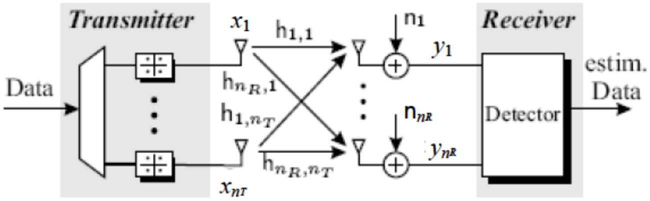

Fig. 1. Block diagram of MIMO [1]

Figure 1 [1] shows a typical diagram of MIMO. The transmitter modulates bits into symbols with proper constellation then performs source and channel coding for robustness. The MIMO channel is composed of multiple single-input single-output (SISO) pairs which can be modelled as i.i.d. if the antenna spacing is larger than half wavelength in practical scattering environments. The receiver detects and decodes the received symbol stream back to bits based on the strategies as ML, ZF, MMSE and SIC that rely on CSIR. At any time instant k, the system equation reads as:

$$\mathbf{y}_k = \sqrt{E_s}\mathbf{H}_k\mathbf{c}'_k + \mathbf{n}_k \tag{1}$$

where $\mathbf{y}_k$ is the received symbol vector, E_s is the symbol power, $\mathbf{H}_k$ is the channel matrix of size $n_r \times n_t$, $\mathbf{c}'_k$ denote the precoded input, and $\mathbf{n}_k$ is the noise vector.

B. Channel Capacity and Water-Filling Algorithm

Channel capacity is defined as the upper bound of mutual information which indicates the maximum available data rate for reliable transmission:

$$C(\mathbf{H}) = \max_{\mathbf{Q}\geq 0, Tr\{\mathbf{Q}\}=1} \log_2 \det(\mathbf{I}_{n_r} + \rho\mathbf{HQH}^H) \tag{2}$$

where C is the capacity, $\mathbf{H}$ is the channel matrix, $\mathbf{Q}$ is the input covariance matrix whose trace is normalised, and ρ is the signal-to-noise ratio (SNR). The maximisation of mutual information depends on $\mathbf{Q}$, and the optimum input covariance matrix is $\mathbf{Q}^\star = \mathbf{V_H} diag\{s_1^\star, \dots s_n^\star\}\mathbf{V_H}^H$ where $\mathbf{V_H}$ is the right-singular vector matrix of $\mathbf{H}$ and $s_k^\star$ is the power allocated to transmission stream k. Denote the singular values as $\Sigma_{\mathbf{H}} = diag\{\sigma_1, \dots \sigma_n\}$ with $\sigma_k^2 \triangleq \lambda_k$, the capacity writes as:

$$C(\mathbf{H}) = \sum_{k=1}^{n} \log_2(1 + \rho s_k^\star \lambda_k) \tag{3}$$

The optimum power allocation is determined by the water-filling algorithm that achieves the capacity with power constraint $\sum_{k=1}^n s_k = 1$ at the transmitter:

$$s_k^\star = \left(\mu - \frac{1}{\rho\lambda_k}\right)^+ \tag{4}$$

where μ is the water level to ensure the allocated power does not exceed the budget. An intuition given in Figure 2 [2] suggests that those streams in good condition are allocated with more power, and some streams may be abandoned with no power because the status is unacceptable.

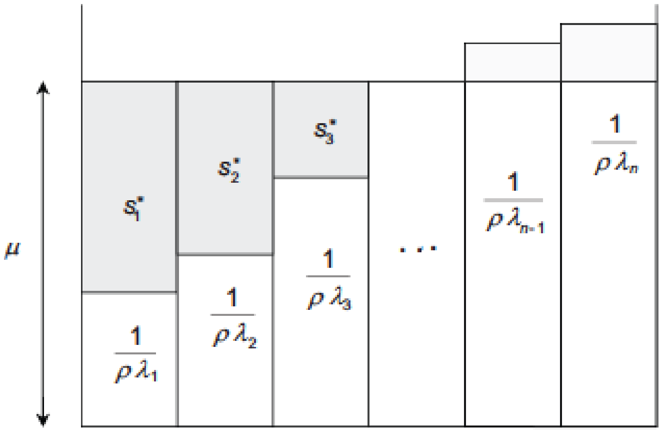
μ
s_1^*
s_2^*
s_3^*
$\frac{1}{\rho\lambda_1}$
$\frac{1}{\rho\lambda_2}$
$\frac{1}{\rho\lambda_3}$
. . .
$\frac{1}{\rho\lambda_{n-1}}$
$\frac{1}{\rho\lambda_n}$

Fig. 2. An intuition of water-filling algorithm [2]

Water-filling can be implemented in different realisations, and the iterative method is used in this coursework. First, the channel matrix $\mathbf{H}$ is decomposed to obtain the eigenvalues λ_k which are sorted in decreasing order. Then, the 'bases' $\frac{1}{\rho\lambda_k}$ that indicates condition of the streams are calculated based on eigenvalues and SNR. The iteration begins next and at iteration i the algorithm updates the water level $\mu(i) = \frac{1}{n-i+1}\left(1+\sum_{k=1}^{n-i+1}\frac{1}{\rho\lambda_k}\right)$ and allocate power to the first i streams. The iteration will stop if the stream power starts to be negative and the last valid result will be used. The complexity of this iterative algorithm is acceptable when the number of streams is not very large.

C. Outage Capacity and Ergodic Capacity

For slow fading channels, a non-zero probability of deep fade exists, and the capacity is zero in the strict sense because no scheme can communicate reliably at a certain fixed data rate [3]. The outage capacity is defined as the transmission rate at which the outage probability is $p\%$.

In comparison, it is possible to code over multiple channel realisations for fast fading channels. Ergodic capacity indicates the maximal rate of reliable communication if the communication duration is long enough to experience all channel states. This mean value averages out the randomness of the channel can be achieved with a particular input distribution. Nevertheless, it is not suitable for applications with short-delay requirements. The ergodic capacity is the expectation of Equation 2.

For CSIT and CDIT, the difference in capacity comes from the input covariance matrix $\mathbf{Q}$. With CSIT, water-filling power allocation can be used to maximise the instantaneous capacity for all channel realisation and therefore the mean value. For CDIT, uniform power allocation is demonstrated globally optimum to maximise the ergodic capacity [2].

D. Space-Time Block Coding (STBC) and Spatial Multiplexing (SM)

STBC encode Q symbols over T time slots to a codeword $\mathbf{c}$ of size $n_t \times T$. It spreads information over symbol and space to enhance the diversity gain $g_d = -\log_2 P_e / \log_2 \rho$ and the spatial multiplexing rate $r_s = Q/T$. If the coding is orthogonal (O-STBC), the complexity in detection and decoding becomes much lower because it can be decoupled into several single-input multiple-output (SIMO) streams. Also, a full diversity gain of $n_r \times n_t$ is achieved, but the spatial multiplexing rate r_s tends to be much smaller than SM. Although there exist full-rate full-rank codes as Dayal [4] which benefits both diversity and rate, the complexity in decoding can be much higher than conventional schemes. The most popular STBC code may be the Alamouti code which achieves a diversity gain of $n_r \times n_t$ but no spatial multiplexing gain.

SM transmits independent streams on different antennas to achieve full rate. Every single symbol is only encoded once in space and time domains, and the transmitted blocks boil down to vectors that write as:

$$\mathbf{C} = \frac{1}{\sqrt{n_t}}[c_1 \dots c_{n_t}]^T \tag{5}$$

where c_i is the modulated symbol on antenna i. Compared with STBC, it enhances the data rate but increases the BER since the resource employed to create redundancy is used to transmit new symbols. With CDIT, equal power allocation on all streams is performed to ensure ergodic capacity, and there is no transmit diversity.

E. Typical Receivers

1) Maximum-Likelihood (ML) Receiver: ML receiver divides the received signal into individual components, enumerates all possible candidates, and estimates the transmitted components based on ML estimation. It regards the hypothesis that maximises the conditional probability as the output, which can be interpreted as selecting the codeword that minimises the error:

$$\hat{\mathbf{C}} = \arg\min_{\hat{\mathbf{C}} \in \{\mathbf{C}\}} \left\| \mathbf{y} - \sqrt{\frac{\rho}{n_t}}\mathbf{H}\mathbf{c}' \right\| \tag{6}$$

where the $\{\mathbf{C}\}$ is the candidate set of transmitted block. For SM the blocks reduces to vectors, and in our 2-by-2 quadrature phase shift keying (QPSK) system, every antenna has 4 possible input that leads to 16 candidates each of size 2×1. The diversity gain of ML is n_r.

2) Zero-Forcing (ZF) Receiver: ZF receiver reduces the impact of the channel on the transmitted signal to zero by channel inversion. The ZF filter is given by:

$$\mathbf{G}_{ZF} = \sqrt{\frac{n_r}{E_s}}\mathbf{H}^\dagger \tag{7}$$

It decouples the channel into n_t independent channels and normalises the gain to suppress the interference and enable individual decoding. With SM transmitter, the output of the filter is $\mathbf{z} = \mathbf{G}_{ZF}\mathbf{y} = [c_1 \dots c_{n_t}]^T + \mathbf{G}_{ZF}\mathbf{n}$. The diversity gain is $n_r - n_t + 1$.

3) Successive Interference Canceller (SIC): SIC receiver decodes the signal stream by stream. In every iteration, ordered SIC chooses the layer with the highest signal-to-interference-plus-noise ratio (SINR), decodes that stream, then subtracts the influence from the received signal. In comparison, unordered SIC selects the layer randomly, and the main problem is error propagation. If the previous streams are wrongly decoded, the result of the latter tends to be false because the process is based on the remainder. Therefore, the BER of unordered SIC is larger than the ordered one. The diversity of the i-th layer is $n_r - n_t + i$, and the overall diversity gain is slightly larger than ZF.

III. Results and Analysis

A. Capacity of Typical Deterministic Channels

For a general $\mathbf{H}$ known at the transmitter, dominant eigenmode transmission (DET) is preferred at low SNR, and multiple eigenmode transmission (MET) achieves higher rate at high SNR. To achieve the capacity, the water-filling algorithm should be implemented to determine the power of all available streams. However, the optimum power allocation of the two special cases below can be observed directly.

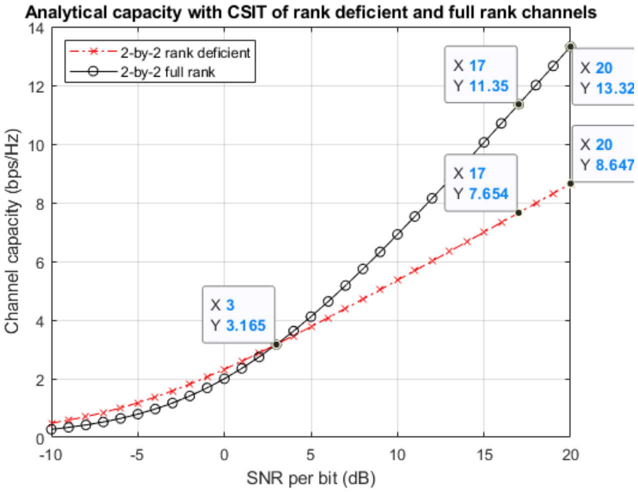

Fig. 3. Analytical capacity with CSIT of rank deficient and full rank channels

1) Rank-deficient Channel: The first channel is $\mathbf{H}_1 = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$. The rank of the channel matrix determines the number of possible transmission streams which is 1 for $\mathbf{H}_1$. Therefore, all power budget is given to the only layer which is similar to DET. Performing SVD, it can be decomposed to

$$\mathbf{u} = \begin{pmatrix} -\sqrt{2} & -\sqrt{2} \\ -\sqrt{2} & \sqrt{2} \end{pmatrix}, \Sigma = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix}, \mathbf{v} = \begin{pmatrix} -\sqrt{2} & \sqrt{2} \\ -\sqrt{2} & -\sqrt{2} \end{pmatrix},$$

and the transmit beamforming direction is $\mathbf{w} = \mathbf{v}_{\max}$ while the receive matched filter is $\mathbf{g} = \mathbf{u}_{\max}^H$. Equation 3 reduces to $\log_2(1 + 4\rho)$. The simulation result is shown in Figure 3 by the red curve.

2) Full-rank Channel: $\mathbf{H}_2 = \begin{pmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{2} \end{pmatrix}$ is a full rank channel. With the rank equals 2, there are two available streams and MET is available. $\mathbf{H}_2$ is a special case where both layers are in the same condition and the power should be uniformly distributed. SVD gives $\mathbf{u} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma = \begin{pmatrix} \sqrt{2} & 0 \\ 0 & \sqrt{2} \end{pmatrix}, \mathbf{v} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$ and $\mathbf{w} = \mathbf{v}$, $\mathbf{g} = \mathbf{u}^H$. Equation 3 indicates $C = 2\log_2(1 + \rho)$ and the plot is given by the black curve in Figure 3.

3) Capacity Comparison: $\mathbf{H}_1$ and $\mathbf{H}_2$ have the same power, but the capacity is different. It is obvious that the capacity depends on the SNR in both cases. At low SNR, the difference between the eigenvalues has a significant impact on the capacity because the log function is sensitive even to a little change when the input is small. In other words, allocating more power to the dominant stream to maximise the product of eigenvalue and power is more effective than two identical streams with intermediate input. Therefore, the layer in good condition dominates the rate. In this case, transmitting on the only stream of $\mathbf{H}_1$ with eigenvalue 2 is better than using both streams of $\mathbf{H}_1$ with eigenvalue $\sqrt{2}$.

In comparison, when the SNR is relatively large, transmitting over multiple streams provides a larger rate because the increasing speed of the log function becomes slower when the input is considerable. In this case, the sum of the rate of independent streams can be higher than the rate of a better stream. As SNR approaches infinity, the rate of $\mathbf{H}_2$ is proportional to $2\log_2\rho$ but that of $\mathbf{H}_1$ depends on $\log_2\rho$ only.

The threshold can be determined by the equation $\log_2(1 + 4\rho) = 2\log_2(1 + \rho)$ which suggest $\rho = 2$ that corresponds to 3 dB. As a result, $\mathbf{H}_1$ has a larger capacity when SNR is smaller than 3 dB and $\mathbf{H}_2$ can achieve a higher rate for SNR larger 3 dB. Figure 3 demonstrates the conclusion. Also, an 3 dB increase provides roughly 1 bps/Hz for $\mathbf{H}_1$ and 2 bps/Hz for $\mathbf{H}_2$ in high SNR region (17 to 20 dB).

B. Ergodic Capacity of Typical MIMO Systems

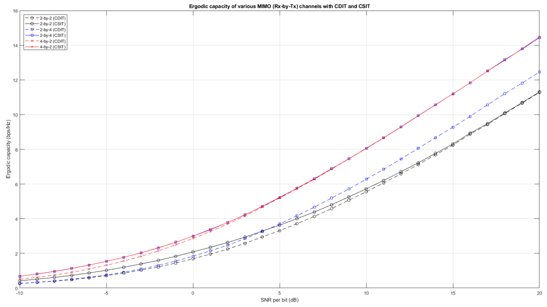

Fig. 4. Ergodic capacity of various MIMO (Rx-by-Tx) channels with CDIT and CSIT

It has been mentioned in the theory section that for the fast-fading channel, the water-filling algorithm can adjust the power allocated to each stream in real time with perfect CSIT to achieve the maximum rate for all channel realisations. Nevertheless, with CDIT where the transmitter only knows the channel distribution rather than actual value, uniform power allocation is the strategy to achieve the capacity.

Figure 4 presents the ergodic capacity of 2-by-2, 2-by-4, and 4-by-2 MIMO systems with CDIT and CSIT (note the notation is Rx-by-Tx). For the 2-by-2 system, the ergodic capacity with CDIT and CSIT are very close, which coincide in the high SNR region. The reason is that when SNR is high, the water-filling algorithm allocates power uniformly between two available streams as discussed above. However, when SNR is relatively low, the strategy allocates more power to the dominant mode and suppress the other stream to enhance the data rate with CSIT. Therefore, the ergodic capacity with CSIT is a bit larger than that with CDIT when SNR is low.

The 2-by-4 and 4-by-2 systems are similar. If CSIT is known, the two cases are the same. With CSIT and CSIR, the transmitter and receiver can be exchanged as the system is entirely symmetrical. Therefore, the red and blue solid lines overlap all the time. Nevertheless, with CDIT only, the power is uniformly distributed on the two transmit antennas for the 4-by-2 system. Similar to the previous case, there is a small gap to the CSIT case which is narrowed as SNR increases, and the reason is the water-filling algorithm evolves from DET to uniform-power MET. In sharp contrast, the 2-by-4 CDIT system requires the power to be uniformly allocated to four transmit antennas. Nevertheless, the number of available streams is the rank of the channel matrix which is two only. It means that the power allocated to two of the antennas are not fully utilised. Also, in the high SNR region, Equation 3 is approximately:

$$\bar{C}_{CDIT} \approx n\log_2\frac{\rho}{n_t} + \varepsilon\left\{\sum_{k=1}^{n}\log_2\lambda_k\right\} \tag{8}$$

It indicates using more transmit antennas can reduce the ergodic capacity. Hence, in the CDIT case, the ergodic capacity of SIMO is always greater than MISO when the number of antennas are fixed.

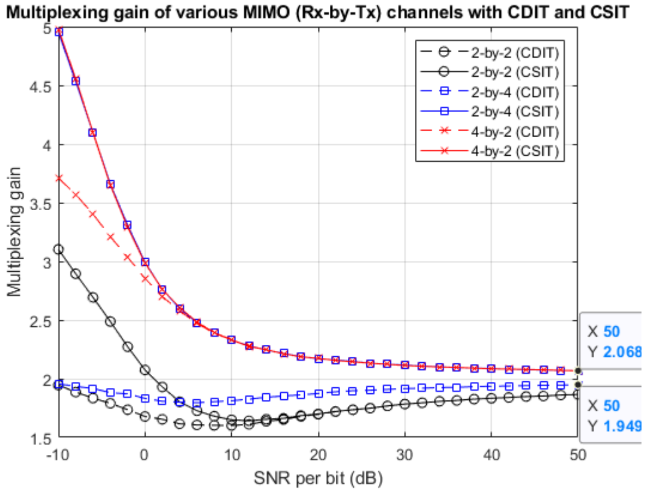

Fig. 5. Multiplexing gain of various MIMO (Rx-by-Tx) channels with CDIT and CSIT

The multiplexing gain defined as the pre-log factor of the rate as SNR approaches infinity is shown in Figure 5. Although the ergodic capacity varies for these systems, as SNR increases, the influence of CSIT disappears because the coincide of uniform power allocation and the water-filling algorithm. It indicates that the multiplexing gain converges to 2 for all the systems, which verifies the theory that the multiplexing gain equals the minimum of the transmit and receive antennas.

C. BER of Different Receivers

1) ML Receiver: As the optimum decoding scheme, ML has the minimum error rate among the three strategies. It has been observed in Figure 6 that the slope of ML is sharper than the others especially in the high SNR region, where the BER gap becomes wider. It suggests a larger diversity gain over the rest. With a 10 dB increase of SNR (10 to 20), the error rate of ML reduces around 100 times from 0.01025 to 7.95×10^{-5} implying a diversity gain approaching 2. Figure 7 shows that as SNR increases, the diversity gain decreases steadily to 2 which equals the number of receive antennas 2 times the rank of error matrix 1. Therefore, the ML receiver utilises the system diversity effectively to reduce the error rate.

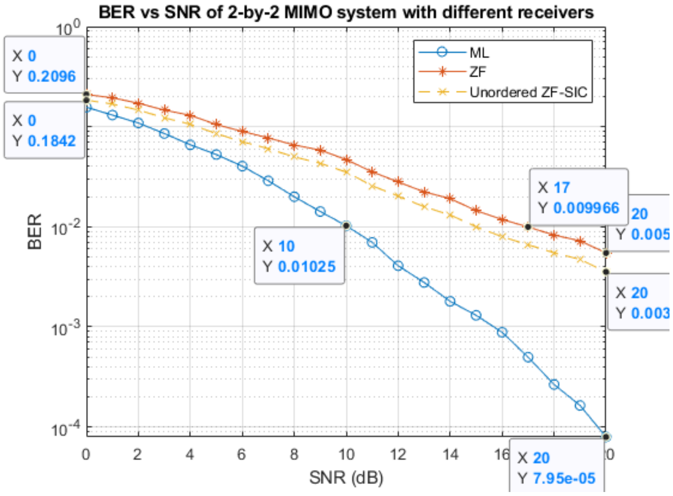

Fig. 6. BER vs SNR of 2-by-2 MIMO system with different receivers

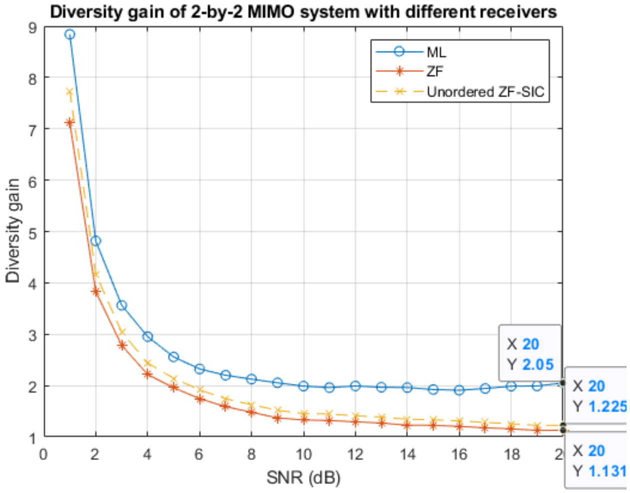

Fig. 7. Diversity gain of 2-by-2 MIMO system with different receivers

Nevertheless, as a block-by-block matching scheme, ML is optimum in accuracy but the complexity is high especially when the transmit antenna number, or the constellation size is large. In such cases, there are numerous possible combinations to be examined and the time cost can increase exponentially. Moreover, ML can handle the case where the number of transmit antennas is more than receiver antennas because the algorithm examines all possible input combinations directly.

2) ZF Receiver: The performance of ZF is worse than the others. At a high SNR of 20 dB, the BER is still around 5.5×10^{-3} for ZF which is about 1.5 and 70 times the value of ZF-SIC and ML. Moreover, with an improvement of 3 dB in the high SNR region (17 to 20), the BER is reduced by half which indicates the diversity gain denoted by the slope is 1. It is as expected by the discussion in the theory section, and the plot is shown in Figure 7. As SNR grows to infinity, the diversity gain converges to 1. The reason is that it decouples the received signal into multiple SISO streams where the diversity is not fully utilised. Hence, the complexity is low, but the cost is the robustness from system diversity. Also, the gap between ZF and ML reaches 0.025 when SNR equals 0 dB. It comes from the channel inversion step that brings in noise enhancement especially when SNR is low. Finally, the system based on the ZF receiver will be underdetermined if the number of transmit antennas is more than receive antennas. In other words, the n_r receive antennas cannot cancel the $n_t - 1$ interferences and the channel capacity cannot be achieved.

3) Unordered ZF SIC Receiver: SIC is an optimisation approach based on the existing decoding results layer by layer to mitigate the impact of interference. The yellow line in Figure 6 proves its improvement over plain ZF receiver. In the 2-by-2 case, one stream is detected first that sees a diversity gain of one. Then, the receiver reduces to the maximum ratio combining (MRC) for the second stream with a diversity gain of 2, because the interference of the previous layer is removed. Compared with the plain ZF, as the decoding process went on, the layers peeled later are with larger diversity gain rather than all equal one. The diversity gain of SIC is different for individual streams which depends on the decoding order. When detecting the i-th layer, there are $n_r - n_t + i$ remained streams producing interferences and the diversity of the i-th layer is $n_r - n_t + i$. Overall, the error performance is dominated by the stream in the worst condition and the diversity gain of non-ordered SIC is approximately $n_r - n_t + 1$. Therefore, the slope is slightly shaper than ZF because the error rate and the diversity gain are primarily dominated by the weakest stream. Therefore, even there are some improvements on the BER, it mainly depends on the first decoded layer with unit diversity and the overall diversity gain is slightly larger than 1 as Figure 7 presents. It suggests that the sequence of decoding is significant and the error rate can be reduced if starting from the most confident one. The enhancement of SIC is expected to be larger for the systems with more antennas especially in the high SNR region, but the error propagation can be more severe and ordered SIC is recommended. An interesting phenomenon is that according to probability, the layers decoded latter tend to be with a larger error rate than the previous ones due to error propagation, but the boost in diversity gain compensates the disadvantage.

IV. Conclusion

In this article, we investigated the Shannon capacity when the channel matrix is given and explored the ergodic capacity with the knowledge of channel distribution. It has been demonstrated that the water-filling algorithm that allocates power based on the stream status and SNR can achieve the capacity with CSIT. This method tends to allocate more power to the better stream, which boils down to DET when the SNR is relatively low and MET with uniform power distribution in the high SNR region. With CDIT, uniform power allocation is proved to be the capacity-achieving strategy although the ergodic capacity with CSIT is a bit larger when SNR is low. Regarding the BER performance, ML receiver that utilises the full diversity at the receiver produces the lowest BER with high complexity, while the low-complexity ZF receiver decouples the signal into independent SISO streams but the BER is high, and the diversity is not exploited. The unordered ZF-SIC receiver improves the BER by reducing interferences and brings a larger diversity gain. The result can be extended to ordered SIC with the code attached.

V. Appendix: MATLAB code

The source code can be retrieved from https://github.com/SnowzTail/.

References

[1] C. Ling, “Multi-input multi-output communication,” September 2018.

[2] B. Clerckx, “Wireless communications,” January 2019.

[3] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[4] P. Dayal and M. K. Varanasi, “An optimal two transmit antenna space-time code and its stacked extensions,” *IEEE Transactions on Information Theory*, vol. 51, no. 12, pp. 4348–4355, 2005.