INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.
Abstract

Delay-limited Throughput Maximization in Fading Channels

by

Nadeem Ahmed

Fading channels, seen in many wireless systems, provide a hostile environment for reliable communication. Conventional analysis of fading channels has been performed from the single-attempt paradigm. That is, the amount of information that can be reliably communicated with a single codeword transmission attempt is quantified. This works well for idealized, delay-unconstrained systems that always transmit a single, infinite-length codeword. However, practical systems are delay-limited since they must use finite-length codewords. Therefore, the conventional performance metrics based on the single-attempt paradigm have drawbacks for delay-limited systems: \(e\)-capacity does not provide a measure of error-free performance, while delay-limited capacity underestimates performance.

For delay-limited systems, transmitters need not restrict themselves to a single transmission attempt per codeword. In fact, practical communication protocols, such as TCP or ARQ, retransmit data when errors occur. Clearly, there is a disconnect in the design of delay-limited systems (multi-attempt) and the conventional measures
used to quantify their performance (single-attempt). In this thesis we provide a new analysis framework for delay-limited systems based on the multi-attempt paradigm. We maximize the average communications throughput by optimizing system parameters and use the maximum throughput as a measure of delay-limited communication performance. We consider two common scenarios, the first being only when the receiver has channel state information (CSI-R), while in the second both transmitter and receiver it (CSI-RT). With CSI-R, the average transmit power is held constant and throughput is maximized by performing optimal rate selection. With CSI-RT, the transmitter knows the condition of the channel at the time of transmission and can vary the power accordingly. Our analysis is done for an average power constraint on the transmitted signal. We also consider the scenario if an additional peak power constraint on the transmitted signal is added. Therefore throughput is maximized by performing optimal rate selection and power control. As a pre-requisite for throughput maximization, we also solved the outage minimization problem for signals with both peak and average power constraints.

We propose maximum ε-throughput (MεT) and maximum zero-outage throughput (MZT) as measures of best-case communications performance when there is, and is not, a restriction on the maximum number of transmission attempts per codeword, respectively. We show that a far greater throughput is achieved with the multi-attempt approach than the single-attempt approach. The increased throughput comes at the cost of queueing delays that are not present when transmitters are limited to
a single transmission attempt. Therefore, we also consider the important situation in
which throughput is maximized with a constraint on the queueing delay.

In this thesis we provide the procedure to maximize communications throughput
for systems and give some non-intuitive design guidelines for delay-limited communi-
cation systems in fading channels. Our novel analysis shows that some conventionally
held wisdom for delay-unconstrained systems does not hold for delay-limited systems.
Acknowledgments

First, I would like to thank my advisor Dr. Richard Baraniuk whose enthusiasm, encouragement and guidance helped make this work possible. Second, I would also like to thank my committee members, Dr. Don Johnson, Dr. Bart Sinclair, Dr. Rudolf Riedi and Dr. Dinesh Rajan for their useful comments and suggestions. Third would also like to thank members of the DSP and Comm groups, specifically Tarik Muharemovic, Amir Khojastepour, Nasir Ahmed, Jaber Borran, and Dror Baron, for their valuable insight. Fourth, I would like to thank Texas Instruments for their support under the Texas Instruments Fellowship and the TI Leadership University Program. Most importantly, I would like to thank my family, my wife, parents and brothers, for their support and encouragement during the course of my studies.
Contents

Abstract ii
Acknowledgments v
List of Illustrations x

1 Introduction 1

1.1 Layered view of communication systems and the single-attempt paradigm 3
1.2 Multi-attempt communication paradigm 4
1.3 Throughput and fading channels 6
1.4 Related Work 7
1.5 Contributions 8
1.5.1 Scenario 1: CSI-R 10
1.5.2 Scenario 2: CSI-RT 14
1.6 Outline 16

2 Background 17

2.1 Notation and Definitions 17
2.2 Block fading channel model 18
2.2.1 Power Constraints 20
2.3 Instantaneous capacity 22
2.3.1 Ergodic capacity ($K = \infty$) .................................................. 23
2.3.2 $\epsilon$-capacity and delay-limited capacity ($K < \infty$) ............. 26
2.3.3 Outage minimization for delay-limited systems ..................... 28

3 Multi-attempt throughput maximization ........................................ 33
  3.1 Cross-layer queueing model .................................................... 34

4 Throughput maximization with optimal rate selection .................... 37
  4.1 Maximum zero-outage throughput with scheme RT (MZT_{RT}) .......... 42
     4.1.1 Mathematical formulation ............................................... 42
     4.1.2 Uniqueness of MZT_{RT} ............................................... 43
     4.1.3 Properties of MZT_{RT} ............................................... 46
     4.1.4 Simulation results ....................................................... 50
     4.1.5 Behavior of MZT and $R_{MZT}^*$ under Ricean fading .......... 58
  4.2 Maximum zero-outage throughput with scheme ID (MZT_{ID}) .......... 65
     4.2.1 Mathematical formulation ............................................... 65
     4.2.2 Uniqueness of MZT_{ID} when $K = 1$ ................................ 66
     4.2.3 Properties of MZT_{ID} ............................................... 70
     4.2.4 Simulation results ....................................................... 71
  4.3 Maximum $\epsilon$ throughput with scheme RT (McT_{RT}) .............. 76

5 Outage minimization under a peak and average power

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
constraint

5.1 Short-term average and peak power constraints .................................. 80
5.2 Long-term average and peak power constraints .................................. 83
5.3 Simulation Results ............................................................................. 88

6 Throughput maximization with optimal rate selection and power control

6.1 Maximum zero-outage throughput with scheme DT (MZT DT ) under
different power constraints ................................................................. 99
6.2 Special cases of MZT DT ................................................................ 102
6.3 Examples and discussion .................................................................. 105
   6.3.1 Increased throughput with the multi-attempt approach ............ 105
   6.3.2 Importance of power control ...................................................... 106
   6.3.3 Importance of rate selection ...................................................... 110
   6.3.4 Effect of a peak power constraint ............................................. 112

7 Throughput maximization with queuing delay constraints

7.1 Mathematical formulation ................................................................. 121
7.2 Simulation Results ........................................................................... 124
7.3 Near-optimal throughput maximization with queuing delay constraints for Bernoulli sources ..................................................... 133
7.4 The queueing delay vs. coding delay tradeoff ............................................ 136

8 Conclusions and Future Work ...................................................................... 139
8.1 Conclusions .................................................................................................. 139
8.2 Future Work .................................................................................................. 142

A Proof of $M_{ZTDT}$ for different power constraints ...................................... 145
A.1 Proof of Theorem 6.1.1 ............................................................................... 145
A.2 Proof of Theorem 6.1.2 ............................................................................... 146
A.3 Proof of Theorem 6.1.3 ............................................................................... 147
A.4 Proof of Theorem 6.1.4 ............................................................................... 148

Bibliography ........................................................................................................ 149
Illustrations

1.1 Model of a wireless communications system. The transmitter encodes data which is distorted by the communications channel. Noise is added prior to reception by the receiver. A feedback link relays retransmission requests and channel state information to the transmitter. .................................................. 3

1.2 Queueing model of a wireless communication system. Codewords arrive into the queue with average rate $\lambda$ and are served by the server with average rate $\mu$. The server encompasses the details of the physical layer including the transmitter and receiver and it perform retransmission until codewords are successfully decoded. .................. 7

2.1 Ergodic capacity, with and without power control, as a function of SNR. Since the two curves converge we see that power control provides diminishing capacity gains as SNR increases. .................. 26

2.2 Minimum outage probability as a function of transmitted power for a transmission rate of 1 nat/sec/Hz with constant power and under the long-term power constraint. Power control provides significant outage probability gains. .............................. 29
4.1 Codeword error probability for (a) $K = \infty$; (b) $K = 1$. In each case
the (rate, error) pairs that are achievable and not are indicated. For
$K = \infty$, and rate less than ergodic capacity is achievable with
arbitrarily small error. For $K = 1$ and for any $K < \infty$ the codeword
error probability will follow the outage probability curve. .................. 38

4.2 The throughput vs. transmission rate when $P_{av} = 10$dB for $K = 1$
and $K = \infty$ is indicated. The transmission rate should be selected
carefully to maximize the throughput for any $K$. The optimal
operating point is different for each $K$. The maximum throughput
achievable when $K = \infty$ is $C_{erg}$. .......................... 39

4.3 Throughput vs. transmission rate for various values of $K$ and
$P_{av} = 10$dB. MZT$_{RT}(P_{av}, K)$ is the peak of each of these curves. The
optimal transmission rate for each $K$ is different ......................... 51

4.4 Outage Probability vs. transmission rate for various values of $K$ and
$P_{av} = 10$dB. As $K \rightarrow \infty$ we see that $P_{out}(R, P_{av}, K) \rightarrow I_F(R, C_{erg})$. 52

4.5 MZT($P_{av}, K$) vs. $P_{av}$ for different $K$. For larger $K$, MZT$_{RT}(P_{av}, K)$ is
closer to ergodic capacity. ..................................................... 53

4.6 MZT$_{RT}(P_{av}, K)$ as a function of coding delay $K$ for
$P_{av} \in \{0, 5, 10\}$dB. Ergodic capacity for each of these values of $P_{av}$ is
plotted for reference. ......................................................... 54
4.7 $R_{MZR_{RT}}^*$ as a function of coding delay $K$ for $P_{av} \in \{0, 5, 10\}$dB.

Ergodic capacity for each of these values of $P_{av}$ is plotted for reference.  

4.8 Throughput vs. $K$ achieved with an SNR of 10dB for transmission rates set to $R = \beta C_{erg}$ with $\beta \in \{0.5, 0.7, 0.99\}$. MZT($P_{av}, K$) is also indicated. Transmitting at the optimal transmission rate of $R^*$ yields significant gains in throughput vs. $R = \beta C_{erg}$. 

4.9 Throughput vs. Outage Probability vs. for various values of $K$ and $P_{av} = 10$dB. The optimal outage probability for each $K$ is different. 

4.10 The optimal outage probability $P_{out_{R_{MZR_{RT}}, K, P_{av}}}$ as a function of $K$ for $P_{av} \in \{0, 5, 10\}$dB. For small values of $K$ the optimal outage probability can be very high, over 0.53 for $P_{av} = 0$ dB. 

4.11 MZT$_{RT}$ plotted as a function of coding delay $K$ for $P_{av} = 10$dB. The simulations are for Rayleigh fading and Ricean fading with a LOS component 9dB stronger than the other fading components. In both cases MZT$_{RT}$ converges to ergodic capacity. 

4.12 Optimal transmission rate $R_{MZR_{RT}}^*$ plotted as a function of coding delay $K$ for $P_{av} = 10$dB. The simulations are for Rayleigh fading and Ricean fading with a LOS component 9dB stronger than the other fading components. In both cases the optimal transmission rate converges to ergodic capacity.
4.13 \( \log[R] \) and \( \log[1 - P_{\text{out}}(R, P_{\text{av}}, K)] \) for \( P_{\text{av}}, K \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100 \} \) and Rayleigh fading. The optimal \( R_{\text{MRT}}^* \) are plotted for all \( K \) and the tangent lines to \( \log[R] \) and \( \log[1 - P_{\text{out}}(R, P_{\text{av}}, K)] \) at \( R_{\text{MRT}}^* \) for \( K = 1 \). The optimal transmission rate is not a monotonic function of \( K \). ................................................ 63

4.14 \( \log[R] \) and \( \log[1 - P_{\text{out}}(R, P_{\text{av}}, K)] \) for \( P_{\text{av}}, K \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100 \} \) and Ricean fading. The LOS component of the fading are 9dB stronger than the non-LOS components. The optimal \( R_{\text{MRT}}^* \) are plotted for all \( K \) and the tangent lines to \( \log[R] \) and \( \log[1 - P_{\text{out}}(R, P_{\text{av}}, K)] \) at \( R_{\text{MRT}}^* \) for \( K = 1 \). For Ricean fading the optimal transmission rate is a monotonic non-decreasing function of \( K \). ....................... 64

4.15 With \( \chi_2^2 \) fading Throughput vs. transmission rate for an SNR of 10dB using simple retransmission and incremental diversity. Incremental diversity provides significant gains for larger rates for which outages are more frequent .......................... 72
4.16 With $\chi^2$ fading $\text{MZT}_{\text{RT}}(P_{\text{av}}, K)$ and $\text{MZT}_{\text{ID}}(P_{\text{av}}, K)$ vs SNR for $K = 1, K = 10$, and $K = 100$. When $K = 1$ incremental diversity provides a gain of approximately 1dB. For $K = 5$, when outages are less likely, the gains are much smaller and become negligible for $K = 100$.

4.17 $\text{MZT}_{\text{ID}}(P_{\text{av}}, K)$ and $\text{MZT}_{\text{RT}}(P_{\text{av}}, K)$ plotted against coding delay $K$ for $P_{\text{av}} \in \{0, 5, 10\}$dB. For any coding delay the maximum throughput for both schemes can be predicted. Scheme ID outperforms scheme RT for small $K$ due to the fact that outages are more frequent resulting in more opportunity to perform codeword combining.

4.18 $R_{\text{MZTID}}^*(P_{\text{av}}, K)$ and $R_{\text{MZTRT}}^*(P_{\text{av}}, K)$ plotted against coding delay $K$ for $P_{\text{av}} \in \{0, 5, 10\}$dB. For any coding delay the maximum throughput for both schemes can be predicted. We see that for scheme ID the optimal transmission rate is larger and can even be higher than $C_{\text{erg}}$. This is due to the fact that codeword combining at the receiver allows the transmitter to use a higher rate for the same average service time.
4.19 MeT_{RT} with P_{out}^{eff} = 0.01 for L = \{1, 2, 3, 4, 5\} and P_{av} = 10dB. Also indicated is the throughput vs. transmission rate curve with no limit to the number of transmission attempts — MZT_{RT} is the peak of this curve. We see that allowing multiple transmission attempts can increase the average spectral efficiency for the same effective outage probability.

5.1 We illustrate the power allocate for a particular channel under the short-term and long-term power allocation strategies for R = 2 nats/sec/Hz. Here the average and peak powers are 10dB and 12dB, respectively. Exceeding the average power constraint for this particular codeword allows the power allocated under the long-term strategy to prevent outage.

5.2 For R = 2 nats/sec/Hz, a coding delay of K = 5, long-term average power constraint P_{av} = 10dB, and the transmission of 10000 codewords, we show histograms of the transmitted power. The top figure is without a peak power constraint and the bottom with a peak power constraint of 12dB. The peak power constraint significantly changes the distribution the transmitted power; without a constraint on the peak power, the maximum power allocated for these 10000 codewords is 22.6dB.
5.3 For \( R = 2 \) nats/sec/Hz, a coding delay of \( K = 5 \), short-term average power constraint \( P_{av} = 10\) dB, and for a the transmission of 10000 codewords, we show histograms of the transmitted power. The top figure is without a peak power constraint and the bottom with a peak power constraint of 12dB. We notice that the peak power constraint does not change the distribution of the transmitted power as much as under the long-term power constraint. This is due to the fact that the with the short-term power constraint the distribution of the power is not as 'peaky' as under the long-term constraint. For example without a peak power constraint, the maximum power allocated for any block is 14.8dB as compared to the 22.6dB for the long-term power constraint.

5.4 Minimum outage probability vs. \( P_{av} \) for a transmission rate of \( R = 1 \) nats/sec/Hz. We illustrate situations when there is no peak power constraint, when \( P_p = \infty \), the peak to average power ratio (PAR) is fixed at 2, \( \frac{P_p}{P_{av}} = 2 \), and when the peak power is fixed at 7 dB, \( P_p = 7\) dB. The peak power constraint results in a drop in outage performance for both average power scenarios. However, the performance degradation is most significantly under the long-term average power constraints.
5.5 Minimum outage probability vs. $R$ for a long-term power constraint of $P_{av} = 10$dB. We illustrate the situations where $P_p \in \{10, 13, 16, \infty\}$dB. With the peak power constraint the outage probability is larger than without a peak power constraint for large $R$ and/or for small $P_p$. ............................................. 94

5.6 Minimum outage probability vs. $R$ for a short-term power constraint of $P_{av} = 10$dB. We illustrate the situations where $P_p \in \{10, 12, \infty\}$dB. The outage probability with a peak power constraint is higher than without. However, the difference is not as large as with the long-term power scenario as the transmitted signal is less affected by the peak power constraint. .......................... 95

6.1 Delay-limited capacity and $MZT^H_{DT}$ for $K=2$ vs. SNR. Increased throughput is possible by allowing multiple transmission attempts per codeword rather than a single attempt. For example, for a throughput of 0.5 nats/sec/Hz, $MZT^H$ provides a 2.4dB gain over delay-limited capacity................................. 106
6.2 $C_{\text{erg-const}}, C_{\text{erg-pc}}, MZT_{\text{DT}}^{\text{const}}$ and $MZT_{\text{DT}}^{h}$ for $K = 1$ as a function of SNR. We see that when $K = \infty$ power control provides negligible capacity gains, as the difference between $C_{\text{erg-const}}$ and $C_{\text{erg-pc}}$ shrinks as SNR increases. When $K < \infty$ power control clearly provides significant throughput gains as the difference between $MZT_{\text{DT}}^{\text{const}}$ and $MZT_{\text{DT}}^{h}$ remains large for all SNR. .......................... 108

6.3 MZT vs. $K$ for constant, short-term and long-term power allocation strategies. For each power allocation strategy, MZT converges to ergodic capacity. The more flexible the power control control policy, the faster the convergence. ............................. 109

6.4 Throughput vs. transmission rate for various $K$ and $P_{av} = 10$dB. MZT for each power allocation strategy corresponds to the peaks of each curve. The optimal transmission rate is different for each power allocation strategy and for each $K$. ................................. 111

6.5 Optimal transmission rate vs. $K$ for constant, short-term and long-term power allocation strategies. Under the long-term power constraint, the optimal transmission rate is higher than ergodic capacity, counter to conventional practice. ............................. 112

6.6 Optimal outage probability vs. coding delay $K$ for $P_{av} = 10$. The optimal outage probability under the constant, short-term and long-term average power constraints are 0.4, 0.37 and 0.27, respectively. 113

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
6.7 MZT<sub>DT</sub> vs. \( P_{av} \) with \( K = 5 \) and for both \( P_p \in \{10, 13, 16, \infty\} \) dB and for fixed PAR's of 1.5 and 2.0. Ergodic capacity is shown for reference. For a fixed \( P_p \), MZT<sub>DT</sub> becomes limited by the peak power constraint for \( P_{av} \) near \( P_p \). At a target spectral efficiency of 1 nat/sec/Hz MZT<sub>DT</sub> is 0.9 dB and 2.1 dB away from ergodic capacity for \( \text{PAR} = 2.0 \) and \( \text{PAR} = 1.5 \), respectively. ........................................................................... 114

6.8 MZT<sup>st</sup> vs. \( P_{av} \) with \( K = 5 \) and for both \( P_p \in \{10, 12, \infty\} \) dB and for fixed PAR’s of 1.5 and 2.0. Ergodic capacity is shown for reference. 115

6.9 MZT<sub>DT</sub> vs. \( K \) for an long-term average power \( P_{av} = 10 \) dB for peak power \( P_p = \{10, 13, 16, \infty\} \) dB. The effect of the peak power constraint diminishes for larger coding delays, \( K \). ........................................................................................................... 116

6.10 MZT<sub>DT</sub> vs. \( K \) under short-term average power \( P_{av} = 10 \) dB and for peak power \( P_p = \{10, 12, \infty\} \) dB. The effect of the peak power constraint is not as pronounced as under the long-term power constraint. ........................................................................................................... 116

6.11 Throughput vs. transmission rate under long-term and peak power constraints for \( K = 5, P_{av} = 10 \) dB and \( P_p = \in \{10, 13, 16, \infty\} \) dB. The system is considered peak-limited for those \( R \) and \( P_p \) for which the throughput drops from the case when \( P_p = \infty \). For \( P_p = 10, 13 \) dB the throughput is limited throughout the range of transmission rates. For \( P_p = 16 \) dB the throughput is peak-limited for \( R > 3.5 \) nats/sec/Hz. 117
6.12 Throughput vs. transmission rate under short-term and peak power constraints for \( K = 5, \mathcal{P}_{av} = 10\text{dB} \) and \( \mathcal{P}_p = \in \{10, 12, \infty\}\text{dB} \). The system is considered peak-limited for those \( R \) and \( \mathcal{P}_p \) for which the throughput drops from the case when \( \mathcal{P}_p = \infty \). .......................... 118

6.13 The optimal transmission rate \( R_{lp}^{st} \) vs. \( K \) under a long-term average power of \( \mathcal{P}_{av} = 10\text{dB} \) and peak powers \( \mathcal{P}_p = \{10, 13, 16, \infty\}\text{dB} \). For larger values of \( \mathcal{P}_p \) the optimal transmission rate can actually be higher than the ergodic capacity of the channel. ....................... 119

6.14 The optimal transmission rate \( R_{sp}^{st} \) vs. \( K \) under a short-term average power of \( \mathcal{P}_{av} = 10\text{dB} \) and peak power \( \mathcal{P}_p = \{10, 13, 16, \infty\}\text{dB} \). The difference in the transmission rates is not as pronounced as under the long-term power constraint. ........................................... 119

6.15 The optimal outage probability vs. coding delay \( K \) under the long-term average and peak power constraints with \( \mathcal{P}_{av} = 10\text{dB} \). We can see that the optimal outage probability that maximizes the communications throughput can be quite high. ............................... 120

6.16 The optimal outage probability vs. coding delay \( K \) under the short-term average and peak power constraints with \( \mathcal{P}_{av} = 10\text{dB} \). We can see that the optimal outage probability that maximizes the communications throughput can be quite high. ............................... 120
7.1 The maximum throughput $M_{\text{ZT}}^D(P_{av}, K)$ is plotted as a function of $D$ for $K = 1$ and $P_{av} = 10$. The throughput for a finite coding delay of $K = 1$ and finite waiting-time $D$ can be predicted. MMPP1 represents the more bursty MMPP process with $p_{low} = 0.95$, while MMPP2 represents the scenario with $p_{low} = 0.85$.

7.2 The maximum throughput $M_{\text{ZT}}^D(P_{av}, K)$ is plotted as a function of $D$ for $K = 1$ and $P_{av} = 10$. We also plot the maximum throughput corresponding to the upper and lower 95% confidence intervals of the expected waiting time.

7.3 Optimal transmission rate $R_{\text{ZT}}^D$ as function of the average waiting time $D$ for $K = 1$ and $P_{av} = 10$. For small $D$, $R_{\text{ZT}}^D$ can be quite far from the asymptotic value, suggesting a reduction in the transmission rate to satisfy a waiting-time constraint. MMPP1 represents the less bursty MMPP process with $p_{low} = 0.85$, while MMPP2 represents the scenario with $p_{low} = 0.95$.

7.4 Optimal arrival rate $a_{\text{ZT}}^D$ as a function of $D$ for $K = 1$ and $P_{av} = 10$. We see that for non-bursty sources, that the optimal arrival rate remains relatively constraint. However, for the bursty MMPP source, the arrival rate drops for smaller $D$. Note that MMPP1 represents the more bursty MMPP process with $p_{low} = 0.95$, while MMPP2 represents the scenario with $p_{low} = 0.85$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
7.5 $\text{MZT}^D_{\text{KT}}(P_{av}, K)$ and $\text{nMZT}^D_{\text{KT}}(P_{av}, K)$ are plotted as a function of $D$ for $K = 1$ and $P_{av} = 10$ for the Bernoulli arrival process. Clearly, the suboptimal solution performs nearly as well as the optimal solution.  

7.6 $\text{MZT}^D_{\text{KT}}(P_{av}, K)$ as a function of $K$ for a waiting time of $D = 20$ and for $P_{av} = 10\text{dB}$. Clearly there is a tradeoff between coding delay and queueing delay. In this example a coding delay of $K = 16$ achieves maximum throughput for $D = 20$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Chapter 1

Introduction

The channels encountered by many wireless communication systems often scatter the transmitted signal along its transmission path. Time variation of the channel results in random fluctuations of the received power level, or fading, making reliable communications difficult [1]. The widespread and growing use of wireless systems necessitates understanding the limits of communication over fading channels.

Transmitters typically employ channel coding techniques that map sequences of input data to codewords that add redundancy to combat the effects of fading and noise prior to transmission [1]. Codewords consist of a number of symbols carrying data at the transmission rate, the number of information bits communicated with each symbol. The channel coherence time is the amount of time the time-varying channel is assumed constant; signals transmitted within the coherence time are affected by a single fading state. During transmission, each codeword is affected by one or more fading states with the specific number greatly affecting the communications performance. The coding delay is proportional to the codeword length and is often quantified in terms of the number of fading states affecting each codeword; it significantly affects a system’s reliable communications performance. A system is considered delay unconstrained if it uses infinite-length codewords resulting in infi-
nite coding delay. Practical communication systems are delay-limited; they must use
finite-length codewords and therefore have a finite coding delay.

Communication systems designers are interested in the throughput, or reliable data rate, between the transmitter and receiver. For example a system transmitting at a rate of 100 Kilobits/second (Kbps) and losing 20% of its codewords to errors has a throughput of 80 Kbps. The maximum throughput more accurately reflects communication performance than the raw transmission rate. The purpose of our work is to maximize the communications throughput of delay-limited communication systems in fading channels.

We consider two common scenarios. The first is the situation in which only the receiver has knowledge of the channel state information (CSI) and the transmitter always uses the same average transmit power — we maximize throughput by performing optimal rate selection. The second is the situation in which the transmitter additionally has CSI and therefore can vary the transmit power based on the condition of the channel — we maximize throughput by performing optimal rate selection and power control. As we shall see maximizing the performance of delay-limited systems often requires selecting system parameters in a rather non-intuitive manner. Moreover, many classic assumptions for delay-unconstrained systems do not hold for delay-limited systems.
1.1 Layered view of communication systems and the single-attempt paradigm

Historically, communication systems have been examined and designed using a layered approach. The well-known Open Systems Interconnection (OSI) model separates communications systems into seven layers, including the physical, data-link, network, and upper layers [2]. The physical layer deals with the transmission of unstructured data across the physical medium, while the data-link layer is responsible for creating a reliable data pipe between transmitter and receiver. This separation works well for analyzing idealized communication systems; however, in practical systems there can be significant coupling between layers [3]. This suggests that \textit{cross-layer optimization}, rather than optimizing each layer independently, should be performed to maximize the performance of practical communication systems.

The field of information theory has concerned itself primarily with understanding
the performance of the physical layer. Information theoretic measures traditionally characterize the amount of information that can be transmitted reliably with a single transmission attempt for any codeword. Single-attempt measures, for delay-limited and delay-unconstrained systems, are motivated by the fact that the upper layers will ensure reliable delivery of the data if there are errors in the physical link. For delay-unconstrained systems the communications performance is quantified by the ergodic capacity, the ultimate reliable data rate over a fading channel [4]. The concept of outage has been introduced for delay-limited systems [5]. If the transmission rate exceeds what the channel condition will reliably allow then an outage occurs, resulting in a decoding error at the receiver. The outage concept leads to $\epsilon$-capacity (or outage capacity) [5] and delay-limited capacity [6, 7] as measures of delay-limited communication performance. $\epsilon$-capacity is the highest transmission rate that can be supported with outage probability no greater than $\epsilon$, while delay-limited capacity is simply $\epsilon$-capacity when outages cannot be tolerated, that is, when $\epsilon = 0$.

1.2 Multi-attempt communication paradigm

The single-attempt paradigm works well, theoretically, for delay-unconstrained systems. Such systems buffer an infinite amount of data and then transmit a single infinite-length codeword. Here, error-free communication is possible as long as the transmission rate is less than the ergodic capacity of the channel [8]. Since error-free communication is possible, data retransmission is unnecessary, making the purely
physical-layer, single-attempt approach perfectly suited for delay-unconstrained systems. For delay-limited systems, the single-attempt approach makes error-free communication difficult. Traditional communication measures for delay-limited systems reflect this: $\epsilon$-capacity does not provide a measure of error-free communications performance [5], while delay-limited capacity tends to underestimate communication performance [6, 7].

The multi-attempt paradigm is more suitable for delay-limited systems than the single-attempt paradigm. Delay-limited systems need not restrict themselves to a single transmission attempt for each codeword — multiple transmission attempts can be performed since codewords are finite length. In practical systems upper layers will often retransmit data to ensure reliable communication. For example variants of the link-layer ARQ [2] or transport layer TCP [2] protocols are often used in real-world systems. There is a clear disconnect between how delay-limited systems are designed and used (practical, multi-attempt) and the measures (idealized, single-attempt) used to quantify their performance. Characterizing the maximum communications throughput, when multiple transmission attempts per codeword is permitted [9, 10, 11, 12], would lead to a more accurate reflection of communications performance of delay-limited systems than any of the single-attempt measures used today.
1.3 Throughput and fading channels

Within the single attempt paradigm, zero-outage (error-free) communications is often viewed as an all-or-nothing phenomenon. For delay-unconstrained systems it is possible to transmit reliably at rates approaching ergodic capacity, while for delay-limited systems delay-limited capacity is zero for many fading distributions of interest [7].

We develop a new analysis framework for delay-limited systems in fading channels. By modelling the communications systems as a queue (see Figure 1.2) it is possible to equate the throughput of the system with the amount of information passing through the queue. The server in our queueing model encompasses the details of both the physical and data-link layers. The server takes codewords that arrive in the queue and attempts transmission repeatedly until the channel condition allows successful transmission. The service time for a codeword is the number of transmission attempts required and can vary from system to system based on the particular retransmission scheme. Using this approach the throughput is simply the transmission rate divided by the service time, the amount of data in each codeword divided by the number of transmission attempts required for successful decoding. Maximizing the throughput through the queue is equivalent to maximizing the throughput of the delay-limited communication system.
Figure 1.2: Queueing model of a wireless communication system. Codewords arrive into the queue with average rate $\lambda$ and are served by the server with average rate $\mu$. The server encompasses the details of the physical layer including the transmitter and receiver and it performs retransmission until codewords are successfully decoded.

1.4 Related Work

Our work is an example of the recent focus on the merging of concepts from information theory and networking [3]. In particular, the use of queuing models in the cross layer optimization and analysis of communication systems is gaining momentum. We highlight some of the most similar work below.

In [13], a queuing model for a multi-access AWGN channel is used. Here, the service time of a transmitted codeword is proportional to its transmission time. Once enough coded symbols have been transmitted, and the receiver has accumulated enough “error exponent”, it tells the transmitter to terminate transmission of the codeword. Using this idea, the tradeoff between delay and the probability of error is explored.

The “timing” capacity of continuous-time queues is investigated in [14]. In this model, information is passed between transmitter and receiver with the interarrival times of packets. The random service time of each packet distorts the timing information and is analogous to a noisy channel in traditional communication systems.
The analysis is extended to discrete-time queueing systems in [15, 16].

Another significant example of cross-layer optimization is the work in [17, 18]. A queuing model is used to minimize the packet loss rate with constraints on the average transmit power and delay. The dual problem of minimum power transmission with constraints on average delay and packet loss is studied in [19]. In both cases, the transmission rate and power are according to the condition of the queue. A similar problem is studied in [20]. In this work, a queuing model is used to explore delay-constrained communication over fading channels. The transmission rate and power use per codeword are adapted dynamically based on the condition of the fading channel and condition of the queue to ensure reliable communication and prevent excessive queuing delays. This work was extended to consider causal power allocation strategies in [21].

Along these same lines, the queuing behavior of wireless systems, using a model similar to [13], under a two-state fading channel is studied in [22]. Here, the average packet delay is minimized with adaptive rate and power control. The tradeoff between the average data rate and the burstiness of the data rate, with respect to queuing delay, is explored.

1.5 Contributions

Throughput, a key measure of the quality of service (QoS) of a communications system, is commonly used in the networking literature for characterizing the performance
of the data-link layer. The idea of using a QoS metric to characterize the performance
of a fading channel \([9, 10, 11, 12]\), emphasizes the close relationship between information theory and networking. By maximizing the throughput within the multi-attempt framework, we jointly maximize the communications throughput of the physical and data-link layers. We provide an analysis framework that can quantify the maximum throughput for any coding delay. In this thesis we identify several multi-attempt schemes for delay-limited communication systems and for each scheme maximize the average throughput. This allows us to define several throughput measures for each multi-attempt scheme.

**Definition 1.5.1. Maximum zero-outage throughput (MZT):** If codewords are retransmitted until successfully decoded at the receiver, then zero-outage (error-free) communications is possible. Under such a scheme MZT represents the highest error-free throughput for a particular multi-attempt scheme.

**Definition 1.5.2. Maximum \(\epsilon\)-throughput (MeT):** If codewords are retransmitted until successfully decoded or a maximum number of transmission attempts is reached, then zero-outage communication is not possible. Under such a scheme we define MeT as the highest throughput with \(\epsilon\) probability of outage after a maximum number of transmission attempts for a particular multi-attempt scheme.

In this thesis, we consider two main CSI scenarios: when only the receiver has CSI (CSI-R) and when both transmitter and receiver have CSI (CSI-RT). With CSI-R,
throughput is maximized using optimal rate selection, while with CSI-RT it is maximized by optimal rate selection and power control. For both scenarios we illustrate that the maximum throughput under the multi-attempt paradigm exceeds that under the single-attempt paradigm. That is, for the same coding delay a higher throughput is possible by allowing multiple transmission attempts per codeword, rather than a single attempt.

A key difference in our work and much of the related work in Section 1.4 lies in the transmission rate. In [17, 18, 20, 21, 22] both the transmission rate and power are continuously adapted based on a number of criteria. In practical systems adapting the transmission power can be accomplished easily by adjusting the gain of the transmitted signal. However, continuously adapting the transmission rate requires the multiplexing of (possibly infinitely) many codebooks, one for every possible transmission rate, making such a system prohibitively expensive. In our work, we consider rate selection rather than rate adaptation. By selecting a single rate based on statistics of the fading channel, only the hardware for a single codebook would need to be used in practice making the system amenable to implementation.

1.5.1 Scenario 1: CSI-R

With CSI-R we propose two multi-attempt schemes that are used to reduce or eliminate the occurrence of outage events.
Definition 1.5.3. Scheme RT: When a decoding error occurs after a transmission attempt, the receiver discards the codeword in error and requests codeword retransmission from the transmitter. This process is repeated until the codeword is successfully decoded or a maximum number of transmission attempts has been reached. This Re-Transmission scheme is denoted RT.

Definition 1.5.4. Scheme ID: When a decoding error occurs after a transmission attempt, the receiver requests retransmission from the transmitter. Retransmitted codewords are combined optimally with received codewords previously in error using maximal ratio receive combining (MRRC) [23] to make a joint decoding decision. Since the number of codewords combined increases with each retransmission we term this technique as Incremental Diversity and abbreviate it as ID.

Our analysis framework is quite general — the maximum throughput for any coding delay and multi-attempt scheme can be predicted. We illustrate the utility of the multi-attempt approach by analyzing MZT metric in detail for schemes RT and ID, denoted $MZT_{RT}$ and $MZT_{ID}$, respectively. We also analyze the limited transmission approach for scheme RT and study $McT_{RT}$ in detail. In each case MZT and McT can be used as measures of communication performance of delay-limited systems. Some of our contributions include:

- Illustrating that the throughput of a communications system is intimately related to the transmission rate and the coding delay. In fact we show that there
is often a unique transmission rate that corresponds to maximum throughput. Transmitting at a transmission rate other than this maximizer can result in a significant loss in throughput (Chapter 4).

- A conventional performance measure, ergodic capacity, is a special case of our MZT definition (Chapter 4, Theorem 4.1.6).

- The optimal operating point in terms of outage probability can actually be quite high, say $\epsilon = 0.5$. This runs counter to conventional analysis, which normally targets a small outage probability, say $\epsilon = 0.01$ (Chapter 4, Section 4.1.4).

- Often system designers assume that for a "large enough" coding delay the ergodic nature of the fading channel can be captured. Such a situation would result in a negligible outage probability and a transmission rate close to ergodic capacity can be used. Our analysis illustrates the difficulty in making the ergodic assumption (Chapter 4, Section 4.1.4).

- In some cases, the optimal transmission rate, that maximizes throughput, can actually be higher than the ergodic capacity of the channel. This does not violate the channel capacity theorem since the resulting throughput is always less than ergodic capacity. (Chapter 4, Section 4.2.4)

- Our analysis can be used to predict the best-case performance of practical retransmission schemes, such as ARQ [2], in fading channels.
The benefits of the multi-attempt approach over the single-attempt approach do not occur without a cost. Since the number of transmission attempt per codeword is random a queueing delay is introduced. Thus, it is instructive to determine the maximum throughput when the average waiting-time (delay) of codewords is bounded.

Definition 1.5.5. Maximum zero-outage throughput with an average waiting time constraint (MZT\textsuperscript{D}): If codewords are retransmitted until successfully decoded at the receiver, then zero-outage (error-free) communications is possible. Then MZT\textsuperscript{D} represents the highest error-free throughput for a particular multi-attempt scheme with an average waiting-time less than D.

We study the effect of the queueing delay for scheme RT and quantify MZT\textsubscript{RT}. Our analysis produced several contributions:

- The nature of the queueing delay in fading channels is quite different than in systems with a reliable transmission medium. By scaling back the coding rate rather than the frequency of packet arrivals (as is done by conventional congestion control techniques) queueing delay can be controlled without significant loss in throughput (Chapter 7, Section 7.2).

- Since end users only care about the total waiting-time, they are indifferent as to whether the delay is mostly coding delay or queueing delay. We examine the effect on throughput if queueing delay is traded for coding delay (Chapter 7, Section 7.4).
1.5.2 Scenario 2: CSI-RT

With CSI-RT, the transmitter knows the condition of the channel at the time of transmission and knows if the if it is sufficient to allow reliable communication, or if an outage event will occur. Therefore, with CSI at both the transmitter and receiver, the transmitter can simply delay transmission, to prevent outages and retransmit when the channel condition will allow successful transmission.

**Definition 1.5.6. Scheme DT:** If the transmitter realizes that the condition of the fading channel at the time of codeword transmission will result in an outage, then it delays transmission until the channel condition changes. This process is repeated until the codeword is successfully decoded at the receiver or the maximum number of transmission attempt is reached. This Delayed Transmission scheme is denoted DT.

We maximize throughput for scheme DT to achieve $M_{Z_TDT}$. Our analysis for systems with CSI-RT has lead to several contributions:

- We develop a novel outage-minimizing power allocation strategy, necessary to maximize the throughput, for delay-limited systems under both peak and average power constraints. Using this strategy we quantify the effect of a peak power constraint on the minimum outage probability and maximum throughput (Chapter 5).

- We also show that delay-limited capacity does not represent the highest throughput for delay-limited systems. By allowing multiple transmission attempts
per codeword, and performing rate selection and power control, a throughput, $MZT_{DT}$, higher than delay-limited capacity is possible. This illustrates the power of the multi-attempt communications approach which achieves a higher throughput than the single-attempt approach (Chapter 6, Theorem 6.0.2).

- It is well known that for delay-unconstrained systems that power control provides “negligible capacity gains” [4]. That is, for error-free communications power control does not provide significant capacity gains. The significant of power control in the context of outage was shown in [24]. We add to this by showing that for delay-limited systems power control becomes increasingly important for error-free communications (Chapter 6, Section 6.3.2).

- We also illustrate that a throughput near ergodic capacity, achievable only with infinite coding delay, is possible for very small coding delays using optimal rate selection and power control (Chapter 6, Section 6.3.2).

- We show that the optimal transmission rate for a delay-limited system can be non-intuitive. In fact, the transmission rate that maximizes throughput can actually be higher than ergodic capacity. Of course, the resulting throughput is always less than ergodic capacity and capacity theorems are not violated by selecting such a high rate. The use of such high rates is counter to conventional practice (Chapter 6, Section 6.3.3).
1.6 Outline

This thesis is organized as follows. Chapter 2 overviews relevant background information and notation. Next in Chapter 3 the concept of throughput maximization for delay-limited systems is introduced. Chapter 4 discusses throughput maximization with optimal rate selection. Outage minimization under peak and average power constraints, a prerequisite for throughput maximization, is dealt with in Chapter 5. Chapter 6 deals with throughput maximization with rate selection and power control. Throughput maximization with consideration for queueing delay constraints is considered in Chapter 7. Finally we provide conclusions and directions for future work in Chapter 8.
Chapter 2

Background

2.1 Notation and Definitions

Let $Z$, $Z_1$, and $Z$ represent a scalar, vector, and matrix, respectively. Then $\text{diag}(Z) = Z$ is a diagonal matrix with diagonal elements $Z$, and $I^L = \text{diag}(1, 1, \ldots, 1)$ is the $L \times L$ identity matrix. Denote $\mathbb{E}[g(z)]$ as the expected value of $g(z)$. Let $f(\alpha)$ and $F(\alpha)$ represent the probability density function (PDF) and cumulative distribution function (CDF) of the random vector $\alpha$, respectively. Let $\mathbb{R}$ and $\mathbb{R}_+$ represent the real line and the positive real line. Then $\mathbb{R}^L$ and $\mathbb{R}_+^{L \times M}$ are the set of length-$L$ vectors and $L \times M$ matrices with elements in $\mathbb{R}$, respectively. Similarly, $\mathbb{R}_+^L$ and $\mathbb{R}_+^{L \times M}$ are the set of length-$L$ vectors and $L \times M$ matrices with elements in $\mathbb{R}_+$, respectively.

For $\{a, b\} \in \mathbb{R}$, let $\mathcal{I}_F(a, b)$ be the indicator function, which is 1 if $a > b$ and 0 if $a < b$. Let $w \sim \mathcal{N}(m, V)$ represent a jointly Gaussian random vector with mean $m$ and covariance matrix $V$. Similarly let $x \sim \chi^2_a$ with $a = 1, 2, 3, \ldots$ represent a chi-squared random variable with $a$ degrees of freedom. Finally let $W(b)$ be Lambert’s $W$ function [25], the solution to $xe^x = b$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
2.2 Block fading channel model

In many important applications the condition of the fading channel changes on a time scale that is much slower than the communications signalling. This motivates modelling the channel as a discrete-time, block-fading, additive white Gaussian noise (BF-AWGN) channel [26]. In this model, each “block” of \( N \) symbols corresponds to the amount of time the channel remains constant, the channel coherence time. The system in the \( k^{th} \) block can be written

\[
y_k = x_k \sqrt{\alpha_k} + w_k, \tag{2.1}
\]

with \( x_k, y_k \in \mathbb{R}^N \) representing the system input and output. We assume a Gaussian noise process, \( w_k \sim \mathcal{N}(0, I^N) \). Scattering by the environment results in reflections of the transmitted signal that add constructively or destructively with the original signal. The multipath interference due to scattering is represented by a random multiplicative gain \( \sqrt{\alpha_k} \in \mathbb{R} \) on the transmitted signal. In the rest of the paper we will refer to \( x, y, w \) and \( \sqrt{\alpha} \) as the channel input, output, noise and gain when the relative position in the codeword is not important. Figure 1.1 provides a block diagram of the wireless communication system. The model contains a delay-less, error-free feedback link used to relay acknowledgements of codewords (whether they were successfully decoded or not) back to the transmitter.

Codewords span \( K \) blocks of the BF-AWGN channel, contain \( KN \) symbols, and correspond to a \( K \) block coding delay. Each of the \( KN \) symbols contain information.
encoded at the transmission rate $R$ nats/sec/Hz (nat := $\frac{\text{bit}}{\log_2(2)}$). The time-variations of the channel are assumed to be independent and identically distributed (i.i.d.) from block to block. Blocks can physically correspond to slots in time, frequency, or both. The $K$ i.i.d. channel fades affecting each codeword are

$$\alpha := [\alpha_0, \alpha_1, \ldots, \alpha_{K-1}] .$$  \hspace{1cm} (2.2)

This model is often used and applies, for example, to wireless multicarrier modulated systems with $K$ parallel subchannels [24, 7].

In this work, we assume that the fading states, $\alpha_k$, follow a $\chi^2_2$ (chi-squared with 2 degrees of freedom) distribution with

$$f(\alpha) = e^{-\alpha},$$  \hspace{1cm} (2.3)

and

$$F(\alpha) = 1 - e^{-\alpha}$$ \hspace{1cm} (2.4)

the PDF and CDF, respectively. Such a distribution results when the $\sqrt{\alpha_k}$ are Rayleigh distributed. This model is commonly used for wireless communication systems without line-of-sight between transmitter and receiver [1]. Constructive interference results in a large $\alpha$ and thus a large received signal power that is conducive to communication; we term this situation a "good" fade. Destructive interference results in a small $\alpha \approx 0$ and thus a small received signal power that is not conducive to communication; we term this situation a "bad" fade.
2.2.1 Power Constraints

A system's capacity is normally measured with an average power constraint on the input, denoted $P_{av}$. Without such a restriction the capacity of the channel is infinite since the cardinality of the input distribution is infinite, that is $x \in \mathbb{R}^N$ [8]. The transmitted power in the $k^{th}$ block of codeword is

$$\gamma_k := \frac{1}{N} \sum_{n=0}^{N-1} |x(n)|^2. \quad (2.5)$$

Random fading results in a received power of $\alpha_k \gamma_k$. Since we assume a unit variance noise process $\mathbb{E}[w] = 1$, $\alpha_k \gamma_k$ also equals the received signal-to-noise ratio (SNR) in the block. Additionally, since $\mathbb{E}[\alpha] = 1$ the average received SNR is also $\gamma$.

We consider two CSI scenarios. The first is when only the receiver has perfect, delay-less and error-free, CSI (CSI-R). In this case the transmitter cannot vary the average power based on the condition of the channel since it is unknown. Therefore, performance is maximized by always transmitting at the average power. That is $\gamma_k = P_{av}, \forall k \{0, 1, \ldots, K - 1\}$.

For the second scenario when both transmitter and receiver have perfect CSI (CSI-RT), the average transmit power need not be constant; it can be varied in different blocks of the codeword based on the condition of the channel. Let $\gamma$ represent a power allocation policy, a strategy that assigns the power allocation vector

$$\gamma(\alpha) := [\gamma_0(\alpha), \gamma_2(\alpha), \ldots, \gamma_{K-1}(\alpha)] \quad (2.6)$$

given the channel $\alpha$. When performing power control, the transmitter must be careful.
not to violate the specified power constraint. A common example is the short-term average power constraint

$$\langle \gamma(\alpha) \rangle := \frac{1}{K} \sum_{k=0}^{K-1} \gamma_k \leq \mathcal{P}_{av}. \quad (2.7)$$

Here the average power in any block of the codeword can exceed $\mathcal{P}_{av}$ while the average within the entire codeword cannot. Another widely used example is the long-term average power constraint

$$\mathbb{E}_\alpha \left[ \langle \gamma(\alpha) \rangle \right] \leq \mathcal{P}_{av}. \quad (2.8)$$

This is a more relaxed condition since it allows the average power for any particular codeword to exceed $\mathcal{P}_{av}$ as long as the average long-term power across all codewords does not.

In practical communication systems there is often a peak power constraint on the channel input in addition to the average power constraint. Non-linearities in power amplifiers force transmitters to limit the peak power to avoid distortion of the transmitted signal. Similarly, peak power may be limited to comply with communication standards that limit the interference to other communication systems. We define the peak power constraint as

$$\gamma_k \leq \mathcal{P}_p, \quad \forall k \in \{0, 1, \ldots, K - 1\} \quad (2.9)$$

which limits the maximum average power that can be allocated in any block of a codeword. While not a constraint on the absolute peak, our approach allows us
to constrain the peak power of the transmitted signal while remaining in the class of capacity achieving Gaussian channel inputs. We also define the peak-to-average power ratio (PAR) as

$$\text{PAR} = \frac{P_p}{P_{av}}.$$  \hspace{1cm} (2.10)

Note that $P_p = \infty$ corresponds to no peak power constraint on the channel input.

With these constraints defined, we denote

$$\mathcal{O}_K^\text{st}(P_{av}) = \{ \gamma : \langle \gamma(\alpha) \rangle \leq P_{av} \}$$  \hspace{1cm} (2.11)

$$\mathcal{O}_K^\text{lt}(P_{av}) = \{ \gamma : \mathbb{E}_\alpha[\langle \gamma(\alpha) \rangle] \leq P_{av} \}$$  \hspace{1cm} (2.12)

$$\mathcal{O}_K^\text{stlt}(P_{av}, P_p) = \{ \gamma : \langle \gamma(\alpha) \rangle \leq P_{av}, \gamma_k \leq P_p \ \forall k = 0, 1, \ldots, K - 1 \}$$  \hspace{1cm} (2.13)

$$\mathcal{O}_K^{\text{ltlt}}(P_{av}, P_p) = \{ \gamma : \mathbb{E}_\alpha[\langle \gamma(\alpha) \rangle] \leq P_{av}, \gamma_k \leq P_p \ \forall k = 0, 1, \ldots, K - 1 \}$$  \hspace{1cm} (2.14)

as the set of all $K$-block power allocation policies that satisfy the short-term average (2.11), long-term average (2.12), short-term average and peak (2.13), and long-term average and peak (2.14) power constraints.

### 2.3 Instantaneous capacity

The *instantaneous capacity* [24] (spectral efficiency actually, but we use the terms interchangeably), the highest reliable data rate for a codeword, is found by maximizing the mutual information over a frame of $K$ i.i.d. channel fades, $\alpha$. Assuming a Gaussian noise process and with CSI-R, and a constant average transmit power of $P_{av}$, the
instantaneous capacity is given by

\[ C_K(\alpha, P_{av}) := \frac{1}{K} \sum_{k=0}^{K-1} \log(1 + \alpha_k P_{av}). \tag{2.15} \]

With CSI-RT and for power allocation vector \( \gamma(\alpha) \), it is given by

\[ C_K^{\text{CSI-RT}}(\alpha, \gamma(\alpha)) := \frac{1}{K} \sum_{k=0}^{K-1} \log(1 + \alpha_k \gamma_k(\alpha)). \tag{2.16} \]

In both cases it is achieved using random coding at the transmitter, with the elements of \( x_k \) drawn from a Gaussian codebook \( \sim \mathcal{N}(0,1) \). Prior to transmission, each of the \( K \) blocks in the codeword is scaled by \( \sqrt{P_{av}} \) (CSI-R) or \( \sqrt{\gamma_k(\alpha)} \) (CSI-RT), respectively.

Maximum a posteriori (MAP) detection is used at the receiver[8]. Instantaneous capacity is an asymptotic quantity that is achieved as \( N \to \infty \).

Communication performance measures based on the instantaneous capacity depend on the coding delay \( K \). In the following subsections we overview the delay unconstrained \((K = \infty)\) and delay-limited \((K < \infty)\) cases.

### 2.3.1 Ergodic capacity \((K = \infty)\)

If the sequence of fading states \( \alpha_k \) for \( k \in \{0, 1, \ldots, K-1\} \) is asymptotically ergodic as \( K \to \infty \), then the channels indexed by the block length \( N \) form a family that have the same capacity [24]. This quantity is known as *ergodic capacity* and with CSI-R is given by [4]

\[ C_{\text{erg}} = \lim_{K \to \infty} C_K(\alpha, P_{av}) = \mathbb{E}_{\alpha} [\log(1 + \alpha P_{av})], \tag{2.17} \]
for an average power constraint on the channel input. The expectation is performed with respect to the distribution of the channel fading process $f(\alpha)$. It is found by taking $K \to \infty$ in (2.15).

With CSI-RT ergodic capacity is given by [4]

$$C_{\text{erg-PC}}(P_{av}) := \sup_{\gamma \in C_{\text{PC}}} \mathbb{E}_\alpha [\log(1 + \alpha\gamma)].$$

for an average power constraint $P_{av}$ on the channel input. Again the expectation is performed with respect to $f(\alpha)$. It is found by taking $K \to \infty$ in (2.16) and selecting the optimal power allocation strategy that satisfies the average power constraint. The capacity achieving power allocation strategy

$$\gamma^C(\alpha) = \left[ \frac{1}{\lambda^C} - \frac{1}{\alpha} \right]_+.$$  (2.19)

assigns power $\gamma^C(\alpha)$ to any block affected by fading state $\alpha$. Here, $\lambda^C$ chosen such that the power constraint is satisfied,

$$\int_{\lambda^C}^{\infty} \left( \frac{1}{\lambda^C} - \frac{1}{\alpha} \right) dF(\alpha) = P_{av}. \quad (2.20)$$

For our model in which $\alpha \sim \chi^2_2$, ergodic capacity can be written

$$C_{\text{erg}} = e^{-P_{av}} \int_{\frac{1}{P_{av}}}^{\infty} \frac{e^{-t}}{t} dt \quad (2.21)$$

with CSI-R and as

$$C_{\text{erg-PC}} = \int_{\lambda^C}^{\infty} \frac{e^{-t}}{t} dt \quad (2.22)$$
with CSI-RT, with $\lambda^C$ as the solution to

$$
\frac{e^{-\lambda^C}}{\lambda^C} - \int_{\lambda^C}^{\infty} \frac{e^{-t}}{t} dt = P_{av}.
$$

For both CSI scenarios codewords are drawn from an infinite-length codebook with i.i.d. symbols $\sim \mathcal{N}(0,1)$. Prior to transmission, the $N$ symbols in each block are scaled by either $\sqrt{P_{av}}$ (CSI-R) or $\sqrt{\gamma^C(\alpha)}$ (CSI-RT) [24]. Since codewords are affected by infinitely many fading states, the effect of the fading channel can be “averaged out” and reliable transmission at ergodic capacity is possible. Contrary to what (2.17) and (2.18) seem to suggest, ergodic capacity is not actually an average capacity, but rather the highest rate that can be sustained on all channel states with arbitrarily small probability of error [24].

Figure 2.1 compares ergodic capacity, $C_{\text{erg}-pc}$, with power control (2.18) to ergodic capacity for constant power allocation, $C_{\text{erg}}$, as a function of the average power $P_{av}$ for $\chi_2^2$ fading. For small transmit powers, the capacity with power control is larger than capacity with constant power. For larger transmit powers, the difference between variable and constant power transmission shrinks, leading to the well-known conclusion that power control “yields negligible capacity gains” over constant power transmission [4]. This occurs because the power allocated for each fading state (2.19) differs very little when $P_{av}$ is large.
2.3.2 $e$-capacity and delay-limited capacity ($K < \infty$)

For finite $K < \infty$, the sequence of fading states $\alpha_k$ for $k \in \{0, 1, \ldots, K - 1\}$ cannot be considered asymptotically ergodic. As such the instantaneous capacity becomes a random quantity. When the channel condition is good, a number of the $K$ channel fades affecting a codeword are good and a large amount of information can be transmitted per codeword. Conversely, when the channel condition is bad only a small amount of information can be reliably transmitted. An outage is declared if the transmission rate is larger than the instantaneous capacity, $R > C_K(\alpha, P_{av})$ (CSI-R) or $R > C_K(\alpha, \gamma(\alpha))$ (CSI-RT). For large $N$ the outage probability closely approximates the codeword error probability [27].
Since the instantaneous capacity is a random quantity, outages can occur no matter how small or how large the transmission rate. The outage probability, the likelihood of outage events, is given by

\[
P_{\text{out}}(R, \mathcal{P}_{av}, K) := \text{Prob}[R > C_K(\alpha, \mathcal{P}_{av})]
\]

\[
= \mathbb{E}_x[I_F(R, C_K(\alpha, \mathcal{P}_{av}))],
\]

(2.24)

with CSI-R and by

\[
P_{\text{out}}(R, \gamma, K) := \text{Prob}[R > C^\text{pc}_K(\alpha, \gamma(\alpha))]
\]

\[
= \mathbb{E}_x[I_F(R, C^\text{pc}_K(\alpha, \gamma(\alpha)))],
\]

(2.25)

with CSI-RT. That is, for any transmission rate \( R \) and power allocation policy (including constant power transmission) there is an associated outage probability \( P_{\text{out}}(R, \mathcal{P}_{av}, K) \) (CSI-R) or \( P_{\text{out}}(R, \gamma, K) \) (CSI-RT). Using this, \( \epsilon \)-capacity is defined as

\[
C_\epsilon(\mathcal{P}_{av}, K) := \sup_R \{R : P_{\text{out}}(R, \mathcal{P}_{av}, K) \leq \epsilon\},
\]

(2.26)

with CSI-R and by

\[
C^\text{pc}_\epsilon(\mathcal{P}_{av}, K) := \sup_R \{R : P_{\text{out}}(R, \gamma, K) \leq \epsilon : \gamma \in \mathcal{O}_K\}.
\]

(2.27)

with CSI-RT, where \( \mathcal{O}_K \) is the set of all valid power allocation strategies over which the optimization is performed and can represent either \( \mathcal{O}^\text{a}_K(\mathcal{P}_{av}) \), \( \mathcal{O}^\text{st}_K(\mathcal{P}_{av}, \mathcal{P}_p) \), \( \mathcal{O}^\text{lt}_K(\mathcal{P}_{av}) \), or \( \mathcal{O}^\text{lt}_K(\mathcal{P}_{av}, \mathcal{P}_p) \). \( \epsilon \)-capacity represents the highest rate that can be supported with
outage probability less than $\epsilon$ and is commonly used to quantify the communications performance of delay-limited communications systems in fading channels.

The need for an error-free measure of performance led to the notion of delay-limited capacity [6, 7]. With CSI-R it is

$$C_{dl}(P_{av}, K) := C_{\epsilon} |_{\epsilon=0}$$

$$= \log(1 + \alpha_{\text{min}} P_{av})$$

and with CSI-RT it is

$$C_{\text{op}}(P_{av}, K) := C_{\epsilon} |_{\epsilon=0}$$.

When the minimum channel gain

$$\alpha_{\text{min}} := \min\{\alpha\} = 0,$$

delay-limited capacity is 0 for all $K < \infty$ [7] with CSI-R. Similarly, delay-limited capacity is 0 for $K = 1$ [6] with CSI-RT. However, it is possible in this case to have non-zero delay-limited capacity for $K > 1$ [24].

2.3.3 Outage minimization for delay-limited systems

Since the transmit power can be varied based on the condition of the channel with CSI-RT, the power allocation policy used greatly affects performance. The policy of particular importance is the one that minimizes the outage probability. This policy can also be used to maximize the transmission rate for a target outage probability;
that is it can be used to achieve $C_{fc}^C$. The outage minimization problem can be stated as

$$
\min_{\gamma} \{ P_{out}(R, \gamma, K) : \gamma \in \mathcal{O}_K \}.
$$

The solution to (2.32) is known as the \textit{outage minimizing power allocation strategy} and has been found for $\mathcal{O}_K = \mathcal{O}^{st}_K(P_{av})$ and $\mathcal{O}_K = \mathcal{O}^{lt}_K(P_{av})$, the short-term and long-term average power constraints, in [24]. We now briefly overview the solutions for these cases.

Under the short-term average power constraint $\mathcal{O}_K = \mathcal{O}^{st}_K(P_{av})$ in (2.32) and the
outage minimizing power allocation policy is [24]

\[ \gamma^*_{\text{out}}(\alpha) = \left[ \lambda^*_{\text{out}}(\alpha) - \frac{1}{\alpha_k} \right]_+ , \quad (2.33) \]

with

\[ \lambda^*_{\text{out}}(\alpha) = \frac{1}{\mu} \sum_{i=0}^{\mu(\alpha)-1} \frac{1}{\alpha(i)} + \frac{K}{\mu(\alpha)} P_{\text{av}} \quad (2.34) \]

for \( \mu(\alpha) \in \{1, 2, \ldots, K\} \) and \( \alpha(0) \geq \alpha(1) \geq \ldots \geq \alpha(K-1) \) an ordered permutation of the fading states affecting the codeword.

Under the long-term average power constraint \( O_K = O_K^L(P_{\text{av}}) \) in (2.32), and the outage minimizing power allocation policy takes the form [24]

\[ \gamma^L(\alpha) = \begin{cases} \hat{\gamma}(\alpha), & \text{w/ prob 1} \quad \text{if } \alpha \in R_1(s^*_1) \\ \hat{\gamma}(\alpha), & \text{w/ prob } w^* \quad \text{if } \alpha \in \overline{R}_1(s^*_1) - R_1(s^*_1) \\ 0, & \text{w/ prob } (1 - w^*) \quad \text{if } \alpha \in \overline{R}_1(s^*_1) - R_1(s^*_1) \\ 0, & \text{w/ prob 1} \quad \text{if } \alpha \notin R_1(s^*_1) \cup \overline{R}_1(s^*_1) \end{cases} \quad (2.35) \]

where

\[ R_1(s) = \{ \alpha : \langle \overline{\gamma}(\alpha) \rangle < s \} \quad (2.36) \]

\[ \overline{R}_1(s) = \{ \alpha : \langle \overline{\gamma}(\alpha) \rangle \leq s \} \quad (2.37) \]

\[ \overline{R}_1(s) - R_1(s) = \{ \alpha : \langle \overline{\gamma}(\alpha) \rangle = s \} \quad (2.38) \]

represent sets of fading states differentiated by the amount power allocated for each
fading state. Then

$$P_1(s) = \int_{\mathcal{R}_1(s)} \langle \gamma(\alpha) \rangle dF(\alpha)$$  \hfill (2.39)

$$\overline{P}_1(s) = \int_{\overline{\mathcal{R}}_1(s)} \langle \gamma(\alpha) \rangle dF(\alpha).$$  \hfill (2.40)

is the average power allocated over these sets. Then

$$s_1^* = \sup\{ s : P_1(s) < P_{av} \}$$  \hfill (2.41)

is maximum average power allocated for any fading state and

$$w_1^* = \frac{P_{av} - P_1(s^*)}{P_1(s^*) - P_1(s^*)}$$ \hfill (2.42)

is the probability that the codeword is transmitted when this maximum is achieved. Both $s_1^*$ and $w^*$ ensure the average transmitted power across all fading states is $P_{av}$ as desired. Finally,

$$\hat{\gamma}_k(\alpha) = \left[ \lambda^h(\alpha) - \frac{1}{\alpha_k} \right]_+$$ \hfill (2.43)

is the form of the power allocated for fading state $\alpha$, with

$$\lambda^h(\alpha) = \left( \frac{e^{KR}}{\prod_{l=0}^{\mu(\alpha)-1} \alpha(l)} \right)$$ \hfill (2.44)

and $\mu(\alpha) \in \{1, 2, \ldots, K\}$.

Figure 2.2 plots the minimum outage probability for $K = 1$ as a function of $P_{av}$ under constant power allocation and under the long-term average power constraint. We can clearly see the gain of power control — for a target outage probability the average power required is far less than when using constant power transmission.
We provide the solutions for outage minimization for the remaining cases, $\mathcal{O}_K = \mathcal{O}_K^{\text{ex}}(P_{av}, P_p)$ and $\mathcal{O}_K = \mathcal{O}_K^{\text{in}}(P_{av}, P_p)$, in Chapter 5.
Chapter 3

Multi-attempt throughput maximization

The conventional single-attempt approach to communications performance has a significant drawback for delay-limited systems in fading channels. Error-free communications proves very difficult to achieve for systems with finite coding delay. Single-attempt communication measures for delay-limited systems reflect this fact: $\epsilon$-capacity is not an estimate of error-free performance while delay-limited capacity tends to be overly conservative.

For delay-limited systems, codeword transmission can be attempted more than once since codewords are finite-length. There are many communications protocols, ARQ for example, in use today that retransmit data in the face of errors to ensure reliable communication. The disconnect between how practical systems are designed (which allow multiple transmission attempts) and the traditional measures used to quantify their performance (which assume only a single transmission attempt) highlights the need for more accurate analysis framework for delay-limited communications performance. The maximum average throughput with the multi-attempt approach is such a framework.

In this section we propose a procedure for maximizing the throughput for any multi-attempt scheme, which can then be used as a measure of delay-limited com-
munications performance. In our analysis we restrict ourselves to the transmission techniques used for the single-attempt measures outlined in Chapter 2. This allows systems to use the same well-known channel coding and decoding techniques used to achieved ergodic capacity, $\epsilon$-capacity, and delay-limited capacity. However, by allowing multiple transmission attempts in the face of outages, an error-free communications link can be established for delay-limited systems in fading channels.

In Section 3.1 we describe the queueing interpretation and the mathematical formulation for the throughput maximization.

3.1 Cross-layer queueing model

By maximizing the communications throughput within the multi-attempt framework we jointly maximize the communications throughput of the physical layer (which is responsible for selecting the transmission rate $R$) and data-link layer (which is responsible for data retransmission in the face of errors). *This joint optimization can be used to predict the best case performance for any retransmission scheme in fading channels.* In subsequent chapters we specialize our results for CSI-R and CSI-RT.

The physical and data-link layers can be modelled jointly as a queue. Codewords arrive into the queue encoded at rate $R$, and therefore contain $RKN$ nats. The server takes a codeword from the queue and attempts transmission. When an outage occurs, the codeword is retransmitted until successful transmission or until a maximum number of transmission attempts is reached. The number of transmission attempts
for each codeword, the *service time*, is a random quantity due to the random nature of the fading channel. We use the number of transmission attempts to quantify the service time, since each transmission attempt corresponds to $K$ blocks and therefore corresponds to the channel coherence time scaled by a factor of $K$. The service time distribution, the probability that $s$ attempts are required for successful transmission, depends on the nature of the retransmission scheme, the transmission rate and power, and the statistics of the fading channel. In general, the probability that a codeword’s service time, $S$, will be $s$ attempts for successful transmission is

$$
\text{Prob}(S = s) = \text{Prob}\left(\bigcap_{i=1}^{s-1} \text{out}_i\right) \left[ 1 - \text{Prob}\left(\text{out}_s | \bigcap_{i=1}^{s-1} \text{out}_i\right)\right]
$$

$$
= \text{Prob}\left(\bigcap_{i=1}^{s-1} \text{out}_i\right) - \text{Prob}\left(\bigcap_{i=1}^{s} \text{out}_i\right),
$$

(3.1)

which is the probability of outage events on the first $s - 1$ attempts multiplied by the probability of successful transmission on the $s^{th}$ attempt given that it was previously in error.

The service time distribution can be used to determine the expected service time $\mathbb{E}[S]$ and the expected service rate $\frac{1}{\mathbb{E}[S]}$ of a codeword. The average amount of data passing through the queue with each transmission attempt is $\frac{R}{\mathbb{E}[S]}$ (nats/sec/Hz), the amount of data contained in each codeword divided by the average number of attempts for successful transmission. For example if the transmission rate is $R = 10$ nats/sec/Hz and takes on average $\mathbb{E}[S] = 2$ transmission attempts per codeword, then the average throughput is 5 nats/sec/Hz. Using this idea we define the maximum
throughput of the system as

\[
T_{\text{max}}(\mathcal{P}_{av}, K, \mathcal{P}_p) := \sup_{R} \sup_{\gamma} \left\{ \frac{R}{\mathbb{E}[S(R, \gamma)]} : \gamma \in \mathcal{O}_K \right\}. 
\]

(3.2)

where the supremum is taken over all transmission rates and power allocation strategies in \( \mathcal{O}_K \). We consider either constant power transmission, \( \gamma = \mathcal{P}_{av} \), for a system with CSI-R, or the transmitter performing power control with \( \mathcal{O}_K \in \{ \mathcal{O}_K^{\text{L}}(\mathcal{P}_{av}), \mathcal{O}_K^{\text{H}}(\mathcal{P}_{av}), \mathcal{O}_K^{\text{M}}(\mathcal{P}_{av}, \mathcal{P}_p), \mathcal{O}_K^{\text{L}}(\mathcal{P}_{av}, \mathcal{P}_p) \} \), for a system with CSI-RT. We note that \( T_{\text{max}}(\mathcal{P}_{av}, K, \mathcal{P}_p) \) predicts the best case performance for a particular multi-attempt scheme, coding delay \( K \), average power constraint \( \mathcal{P}_{av} \) and peak power constraint \( \mathcal{P}_p \). By matching the multi-attempt scheme used in the analysis to one that is used in practice, this analysis can be used to predict the best case communication performance of practical retransmission algorithms, i.e. ARQ, in fading channels.

If the transmission rate is \( R \), then the amount of data successfully decoded with any transmission attempt is either 0 or \( R \), depending on whether an outage does or does not occur, respectively. The maximum average throughput is a very representative measure of communications performance. **Our goal is to maximize the average throughput for delay-limited systems using a particular multi-attempt scheme by optimally selecting the transmission rate and power control policy.**
Chapter 4

Throughput maximization with optimal rate selection

When only the receiver has CSI the transmitter cannot vary the transmit power level based on the condition of the channel. As such for this scenario the transmitter always uses the average power, \( \gamma_k = P_{\text{av}}, \forall k \in \{0, 1, \ldots, K - 1\} \). In this case the optimization in (3.2) is only over the transmission rate and

\[
T_{\text{max}}(P_{\text{av}}, K) := \sup_{R} \frac{R}{\mathbb{E}[S(R, P_{\text{av}}, K)]}.
\] (4.1)

\( T_{\text{max}}(P_{\text{av}}, K) \) represents the optimal balance between the amount of information in each codeword and the frequency at which codewords pass through the queueing system. As \( R \to 0 \) the amount of information carried per codeword shrinks and the throughput approaches 0. Similarly, as \( R \to \infty \) outages become frequent and \( \mathbb{E}[S] \to \infty \), resulting in a throughput that approaches 0.

The optimal transmission rate depends significantly on the coding delay \( K \). Figure 4.1(a)(b) illustrates the codeword error probability when \( K = \infty \) and \( K = 1 \) (for scheme RT), respectively. We can see that optimal operating point when \( K = \infty \) is obvious, transmit at a rate as close to ergodic capacity as possible with codeword error probability close to zero. However, for \( K = 1 \) the optimal transmission rate is not immediately obvious. When we examine the system from a throughput perspective.
Figure 4.1: Codeword error probability for (a) $K = \infty$; (b) $K = 1$. In each case the (rate, error) pairs that are achievable and not are indicated. For $K = \infty$, and rate less than ergodic capacity is achievable with arbitrarily small error. For $K = 1$ and for any $K < \infty$ the codeword error probability will follow the outage probability curve.
Figure 4.2: The throughput vs. transmission rate when $P_{av} = 10$dB for $K = 1$ and $K = \infty$ is indicated. The transmission rate should be selected carefully to maximize the throughput for any $K$. The optimal operating point is different for each $K$. The maximum throughput achievable when $K = \infty$ is $C_{erg}$.
in Figure 4.2 we see that both systems have a transmission rate that maximizes throughput. For \( K = \infty \) we verify that \( R = C_{\text{erg}} \) is the unique throughput maximizing transmission rate. For \( K = 1 \) we see that for scheme RT that there is also a unique throughput maximizing transmission rate. For delay-limited systems, the optimal transmission rate depends on the particular retransmission scheme being used and its expected service time. In general, it is possible to specify conditions on the expected service time, for a particular retransmission scheme, that guarantee the existence of a unique throughput maximizing transmission rate.

**Theorem 4.0.1.** If \( \frac{1}{g(S(R))} \) is a log-concave function of \( R \), then (3.2) has a unique global maximum.

**Proof.** Let \( T(R) = \frac{R}{g(S(R)))} \); then \( f(R) = \log T(R) = \log R + \log \frac{1}{g(S(R))} \). If \( \log \frac{1}{g(S(R))} \) is a concave function then \( f(R) \) is also concave, since \( \log R \) is concave and the sum of two concave functions is also concave. Then from convex optimization theory[28], \( f(R) \) has a unique maximizer on the convex set \( \mathbb{R}_+ \). Let \( R^* \) be the argument that maximizes \( f(R) \). If we compose \( f(R) \) with the monotonically increasing function \( e^x \), then \( e^{f(R)} = T(R) \) has the same maximizer \( R^* \). Hence (3.2) has a unique maximum. \( \square \)

Note that this is a sufficient, but not necessary, condition for the existence of a unique solution. It is possible for \( T(R) = \frac{R}{g(S)} \) to be log-concave without \( \frac{1}{g(S)} \) being log-concave. This scenario would also have a unique maximizer for the throughput. The uniqueness of the optimal transmission rate is of great practical importance. Often
(3.2) cannot be solved explicitly and numerical techniques must be used. Fortunately, if \( \frac{1}{\log \frac{1}{1}} \) is log-concave, we can be assured that any numerical solution to (3.2) is globally optimal.

We will consider schemes RT and ID and maximize the communications throughput for each. Since only the receiver has CSI, the transmitter has no way of knowing an outage has occurred unless it receives feedback from the receiver which can be relayed to the transmitter in the form of retransmission requests. For both schemes a single bit of feedback is required for each codeword to relay (un)successful decoding acknowledgements back to the transmitter. The amount of feedback per block is \( \frac{1}{K} \) bits, which approaches to 0 as the coding delay increases \( K \to \infty \).

When the transmitter is allowed to retransmit each codeword as many times as necessary, then zero-outage, or error-free, communications is possible. The maximum throughput is termed maximum zero-outage throughput (MZT). The name is particularly appropriate as it quantifies the maximum error-free throughput of a communications system for a particular retransmission scheme. Sections 4.1 and 4.2 below deal with retransmission schemes RT and ID, respectively. In Section 4.3, we also consider the situation in which the transmitter limits the maximum number of transmission attempts. Here, successful transmission of each codeword cannot be guaranteed, and there is \( \epsilon \) probability of outage after a maximum number of transmission attempts is reached. We term the maximum throughput under this scenario as the maximum \( \epsilon \) throughput (MeT). We consider MeT only for scheme RT, due to the analytical
4.1 Maximum zero-outage throughput with scheme RT (MZT$_{RT}$)

4.1.1 Mathematical formulation

For scheme RT all transmission attempts have the same probability of success or failure, and (3.1) becomes

$$\text{Prob}(S = s) = [P_{\text{out}}(R, \mathcal{P}_{av}, K)]^{s-1}[1 - P_{\text{out}}(R, \mathcal{P}_{av}, K)].$$

(4.2)

The service time distribution becomes geometric on the positive integers with parameter $[1 - P_{\text{out}}(R, \mathcal{P}_{av}, K)]$ and having the well-known mean [29]

$$E[S] = \frac{1}{1 - P_{\text{out}}(R, \mathcal{P}_{av}, K)}.$$  

(4.3)

Using (3.2), we define the maximum zero-outage throughput for scheme RT as

$$\text{MZT}_{RT}(\mathcal{P}_{av}, K) = \sup_{R} R \cdot [1 - P_{\text{out}}(R, \mathcal{P}_{av}, K)].$$

(4.4)

When the channel fading is good a rate $R$ is achieved; when the channel fading is bad 0 rate is achieved due to outage. By optimizing over the transmission rate, the maximum average throughput across all channel fading states is $\text{MZT}_{RT}(\mathcal{P}_{av}, K)$.

Note that (4.4) is simply the transmission rate $R$ multiplied by the success probability $[1 - P_{\text{out}}(R, \mathcal{P}_{av}, K)]$ and that this same throughput can be achieved without
any feedback to the transmitter. This occurs because the feedback only ensures that codewords in error are retransmitted and is not used to improve the throughput. Without such feedback, codewords in error are discarded by the receiver, and the transmitter sends a new codeword with the next transmission attempt. \( \text{MZT}_{\text{RT}} \) can also be thought of as selecting the best rate and outage probability pair \((R, \epsilon)\) based on the statistics of the channel that maximizes the throughput. Typically, communications performance in fading channels is measured with \( \epsilon \)-capacity, the highest rate for a given outage probability \( \epsilon \) and a small value of \( \epsilon \) is normally chosen such as \( \epsilon = 0.01 \). However, fixing \( \epsilon \) may yield a low throughput. \( \text{MZT}_{\text{RT}} \) finds the best \((R, \epsilon)\) pair that maximizes the communications throughput.

4.1.2 Uniqueness of \( \text{MZT}_{\text{RT}} \)

In general (4.4) does not have a closed form due to the difficulty of obtaining exact expressions for the outage probability for common fading distributions. However, it is possible use properties of the fading distribution to show that a unique global maximizer exists.

**Theorem 4.1.1.** *If the probability density \( f_C(R) \) of the instantaneous capacity over a single block, \( C = \log(1 + \alpha P_{sv}) \), is log-concave, then there is a unique transmission rate that achieves \( \text{MZT}_{\text{RT}}(P_{sv}, K) \).*
Proof. The instantaneous capacity over a codeword spanning \( K \) blocks is

\[
C_K = \frac{C_{(1)} + C_{(2)} + \ldots + C_{(K)}}{K},
\]

(4.5)

where \( C_{(k)} \) is the instantaneous capacity in the \( k^{th} \) block having distribution \( f_C(R) \).

The outage probability \( P_{\text{out}}(R, \mathcal{P}_{\text{av}}, K) = \operatorname{Prob}(C_K < R) \) is the CDF of the random variable \( C_K \) evaluated at \( R \). The PDF of \( C_K \) is then

\[
f_{C_K}(R) = \frac{f_C(R) * f_C(R) * \ldots * f_C(R)}{K}
\]

(4.6)

where * is convolution. Since \( f_C(R) \) is log-concave, (4.6) is also log-concave since the convolution of log-concave functions is also log-concave[30, 31]. Then both

\[
P_{\text{out}}(R, \mathcal{P}_{\text{av}}, 1) = \int_0^R f_{C_K}(x) \, dx
\]

(4.7)

and

\[
(1 - P_{\text{out}}(R, \mathcal{P}_{\text{av}}, 1)) = \int_R^\infty f_{C_K}(x) \, dx
\]

(4.8)

are log-concave (see Section 2 in [30] and Theorems 3 and 4 in [31]). Since \( \frac{1}{\mathbb{E}[S]} = [1 - P_{\text{out}}(R, \mathcal{P}_{\text{av}}, 1)] \) is log-concave, by Theorem 4.0.1 there is a unique transmission rate corresponding to \( \text{MZT}_{RT}(\mathcal{P}_{\text{av}}, K) \).

This result is quite general and holds for any fading distribution corresponding to a log-concave instantaneous capacity over a single block. We now consider the \( \chi_2^2 \) fading model.

**Proposition 4.1.2.** If the channel fading \( \alpha \) follows a \( \chi_2^2 \) distribution, then the PDF of the instantaneous capacity \( f_C(R) \) is log-concave.
Proof. If the fading process follows a $\chi_2^2$ distribution, then

$$P_{\text{out}}(R, \mathcal{P}_{av}, 1) = \text{Prob}[R > \log (1 + \alpha \mathcal{P}_{av})]$$

$$= \text{Prob} \left[ \left( \frac{e^R - 1}{\mathcal{P}_{av}} \right) > \alpha \right]. \quad (4.9)$$

The CDF of a $\chi_2^2$ random variable is $F(x) = 1 - e^{-x}$ and therefore

$$F_C(R) = P_{\text{out}}(R, \mathcal{P}_{av}, 1) = 1 - e^{-\left( \frac{e^R - 1}{\mathcal{P}_{av}} \right)}. \quad (4.10)$$

Then the PDF and its derivatives are given by

$$f_C(R) = \frac{e^R}{\mathcal{P}_{av}} e^{-\left( \frac{e^R - 1}{\mathcal{P}_{av}} \right)}, \quad (4.11)$$

$$f'_C(R) = \frac{e^R}{\mathcal{P}_{av}^2} e^{-\left( \frac{e^R - 1}{\mathcal{P}_{av}} \right)} (\mathcal{P}_{av} - e^R), \quad (4.12)$$

$$f''_C(R) = \frac{e^R}{\mathcal{P}_{av}^3} e^{-\left( \frac{e^R - 1}{\mathcal{P}_{av}} \right)} (\mathcal{P}_{av}^2 - 3e^R + e^{2R}). \quad (4.13)$$

After some algebraic manipulations it can be shown that

$$f_C(R)f''_C(R) \leq \left[ f'_C(R) \right]^2, \quad (4.14)$$

which is a necessary and sufficient condition for log-concavity[28]. □

Proposition 4.1.2 implies that with $\chi_2^2$ fading there is a unique transmission rate that maximizes the communications throughput. However, explicit expressions for the transmission rate and the maximum throughput have eluded us. Examining (4.7), we see that determining the outage probability involves integrating (4.6). However, a closed form expression for (4.6), let alone (4.7), may not exist since it is the convolution of one or more complicated functions. Nonetheless, when $K = 1$, (4.4) admits a semi-explicit solution.
Theorem 4.1.3. If $K = 1$ and the channel fading $\alpha$ follows a $\chi^2$ distribution then

$$M_{ZT}(\mathcal{P}_{av}, 1) = V(\mathcal{P}_{av}) e^{-\left(\frac{V(\mathcal{P}_{av})}{\mathcal{P}_{av}}\right)}.$$  (4.15)

Proof. If $\alpha$ is a $\chi^2$ random variable, then $P_{\text{out}}(R, \mathcal{P}_{av}, 1) = 1 - e^{-\left(\frac{R}{\mathcal{P}_{av}}\right)}$. Using this, let $T(R) = R e^{-\left(\frac{R}{\mathcal{P}_{av}}\right)}$. Taking the derivative with respect to $R$ and equating with zero, we see that the transmission rate corresponding to the critical point is the solution to $R e^R = \mathcal{P}_{av}$. The solution to this is the optimal transmission rate $R^* = \mathcal{W}(\mathcal{P}_{av})$. We know that this rate corresponds to a throughput maximum (rather than a minimum) from Theorem 4.1.1. Substituting this back into $T(R)$, we arrive at (4.15). \qed

4.1.3 Properties of $M_{ZT}$

The performance of many communication systems is often maximized with respect to a fixed outage probability. Normally, system designers select the highest transmission rate that supports a predetermined outage probability (or in practice, packet error rate). However, a greater communications throughput is possible if the constraint of a target outage probability is removed.

Theorem 4.1.4. $M_{ZT}$ is always greater than or equal to the throughput achieved by transmitting at $\epsilon$-capacity.

Proof. For a fixed outage probability $\epsilon$, $\epsilon$-capacity is given by

$$C_\epsilon := \sup_{R} \{ R : P_{\text{out}}(R, \mathcal{P}_{av}, K) \leq \epsilon \}.$$  (4.16)
Every transmission rate $R = C_\epsilon$ corresponds to an outage probability $\epsilon$. This results in

$$T_\epsilon = C_\epsilon(1 - \epsilon) \tag{4.17}$$

as the throughput for outage probability $\epsilon$. Therefore (6.5) is a single point on the curve

$$T_{RT}(R) = R[1 - P_{\text{out}}(R, P_{\text{av}}, K)], \tag{4.18}$$

with $P_{\text{out}}(R, P_{\text{av}}, K)$ the outage probability achievable for transmission rate $R$ and coding delay $K$. Since

$$M_{\text{ZT}_{RT}} = \sup_R \{T_{RT}(R)\}, \tag{4.19}$$

we have

$$M_{\text{ZT}_{RT}} \geq T_\epsilon. \tag{4.20}$$

With the single-attempt approach, zero-outage communications must be guaranteed with a single transmission attempt, i.e., $\epsilon = 0$ in the above analysis. When $0$ is in the support of the fading process, as is the case for $\chi_2^2$ fading, we have $T_\epsilon \mid_{\epsilon=0} = C_{\text{dl}} = 0$ as the highest single-attempt throughput. However, with the multi-attempt approach we need not restrict ourselves to $\epsilon = 0$ since codeword retransmission is permitted. Therefore with $0$ in the support of the fading process it is possible to have $M_{\text{ZT}_{RT}} > 0$. This illustrates the power of the multi-attempt approach; zero-outage communica-
tions can be possible with the multi-attempt approach when it is not possible with the single-attempt approach.

It is a well-known phenomenon that the outage probability approximates the codeword error probability when $N$ is large [27]. Since we know that zero-outage communication is possible with $K = \infty$, this suggests that the outage probability converges asymptotically to $I_F(R, C_{erg})$ as $K \to \infty$.

**Theorem 4.1.5.** The outage probability $P_{out}(R, P_{av}, K)$ converges to the indicator function $I_F(R, C_{erg})$ as $K \to \infty$.

**Proof.** We bound (2.24) using Chebyshev's inequality for $R < C_{erg}$ by

$$0 \leq P_{out}(R, P_{av}, K) \leq \frac{\beta}{K} \quad (4.21)$$

and for $R > C_{erg}$ by

$$1 \geq P_{out}(R, P_{av}, K) \geq 1 - \frac{\beta}{K}. \quad (4.22)$$

In both cases $\beta$ is a constant. Taking $K \to \infty$ we have

$$P_{out}(R, P_{av}, K) = I_F(R, C_{erg}). \quad (4.23)$$

Intuitively as $K$ grows, codewords become more immune to the effects of the fading channel; blocks in the codeword that experience a good channel fade can compensate for blocks that suffer from a bad fade.
Theorem 4.1.6. $\text{MZT}_{\text{RT}}(P_{av}, K)$ converges to ergodic capacity as $K \to \infty$.

Proof. Taking $K \to \infty$ and using (4.23), we arrive at

$$M\text{ZT}_{\text{RT}}(P_{av}, \infty) = \sup_{R} R[1 - I_F(R, C_{\text{erg}})] = C_{\text{erg}}.$$ \hspace{1cm} (4.24)

The maximization is trivial since $P_{\text{out}}(R, P_{av}, K)$ takes only two values: 0 or 1. We see that $M\text{ZT}_{\text{RT}}(P_{av}, K)$ does converge to $C_{\text{erg}}$ as $K \to \infty$, and thus the optimal transmission rate $R_{M\text{ZT}_{\text{RT}}} = M\text{ZT}_{\text{RT}}(P_{av}, \infty) = C_{\text{erg}}$. □

When $K = \infty$, ergodic capacity is viewed as a hard-limit on the transmission rate. From the ergodic capacity theorem [4], if the transmission rate is less than ergodic capacity then the codeword error probability can always be driven to 0. On the other hand, if the transmission rate is larger than ergodic capacity, then codeword errors always occur. Thus only transmission rates below ergodic capacity result in non-zero throughput.

However, when $K < \infty$ the situation is different. The outage probability approaches 1 only as $R \to \infty$ and for any finite transmission rate other than $R = 0$ it is possible to transmit data successfully. More specifically, when $K < \infty$ any finite transmission rate other than $R = 0$ results in non-zero throughput. That is, the notion of capacity as a hard-limit on the transmission rate is "softened." This is due to the fact that multiple transmission attempts per codeword is permitted and there is no need to guarantee successful transmission with a single attempt. Note that
using a transmission rate above ergodic capacity does not contradict any information theoretic notions since the resulting throughput is always below ergodic capacity.

**Theorem 4.1.7.** Non-zero throughput is achievable for transmission rates \( R > C_{\text{erg}} \) when \( K < \infty \).

**Proof.** Let \( R = C_{\text{erg}} + \epsilon \). By Theorem 4.1.5 we see that \( P_{\text{out}}(C_{\text{erg}} + \epsilon, \mathcal{P}_{\text{av}}, K) < 1 \) for \( K < \infty \). Using this inequality in (4.4), we see that the throughput \( R [1 - P_{\text{out}}(C_{\text{erg}} + \epsilon, \mathcal{P}_{\text{av}}, K)] > 0 \). \( \square \)

The intuition behind this phenomenon is that for finite \( K < \infty \) the instantaneous capacity (2.15) is a random quantity. For fading distributions that have support on \( \mathbb{R}_+ \), this means that no matter how high the transmission rate there is non-zero probability that the channel state is good enough to support it. Hence non-zero throughput is possible for \( R > C_{\text{erg}} \) if \( K < \infty \). In the limit when \( K \to \infty \), the instantaneous capacity becomes a constant — the ergodic capacity — and it is impossible for the channel to support \( R > C_{\text{erg}} \).

### 4.1.4 Simulation results

We now empirically verify the properties of MZT via Monte Carlo simulation. For the purposes of our simulations we assume that the channel fading follows a \( \chi^2 \) distribution.

Theorem 4.1.1 and Proposition 4.1.2 tell us that if the channel fading is \( \chi^2 \) then
a unique solution for $\text{MZT}_{RT}$ exists. This phenomenon can be easily observed in Figure 4.3, which plots throughput vs. transmission rate for various values of $K$ and $\mathcal{P}_{av} = 10\,\text{dB}$. We see that each curve has a unique maximum corresponding to $\text{MZT}_{RT}(\mathcal{P}_{av}, K)$. Figure 4.3 also empirically verifies Theorem 4.1.7 since it is apparent that for finite $K < \infty$ if $R > C_{\text{erg}}$ then non-zero throughput is possible. We also see that as $K$ increases the throughput achievable for $R > C_{\text{erg}}$ decreases.

Theorem 4.1.5 tells us that the outage probability as a function of $R$, $P_{\text{out}}(R, \mathcal{P}_{av}, K)$, converges to $\mathcal{I}_p(R, C_{\text{erg}})$ as $K \to \infty$. This effect can be seen in Figure 4.4, which plots outage probability vs. transmission rate for $\mathcal{P}_{av} = 10\,\text{dB}$ and for various values of $K$. Clearly the larger the $K$ the closer the outage probability is
Figure 4.4: Outage Probability vs. transmission rate for various values of $K$ and $P_{av} = 10$ dB. As $K \to \infty$ we see that $P_{out}(R, P_{av}, K) \to I_F(R, C_{erg})$.

Theorem 4.1.6 also shows that $MZT_{RT}(P_{av}, K)$ converges to $C_{erg}(P_{av})$ as $K \to \infty$. This is verified in Figure 4.4, which plots $MZT_{RT}(P_{av}, K)$ as a function of transmit power $P_{av}$ for various $K$. As $K$ increases we can see that $MZT_{RT}(P_{av}, K)$ approaches $C_{erg}(P_{av})$, verifying Theorem 4.1.6. Figure 4.5 also demonstrates the performance penalty suffered by delay-limited systems with scheme RT when compared to ergodic capacity. For example, at a target throughput of 1 nats/sec/Hz, $MZT_{RT}(P_{av}, K)$ is about 1.18 dB away from ergodic capacity when $K = 100$, 2.21 dB away when $K = 20$, 2.96 dB away when $K = 10$, and 5.54 dB when $K = 1$.

Figure 4.6 plots $MZT_{RT}(P_{av}, K)$ as a function of coding delay $K$ for $P_{av} \in \{0, 5, 10\}$ dB. Again, this illustrates that the maximum throughput approaches $C_{erg}(P_{av})$ as $K \to \infty$. $MZT_{RT}(P_{av}, K)$ also appears to be a monotonically increasing
function of $K$. This does make sense intuitively; larger coding delays result in more fading states affecting each codeword and more opportunity to "average out" poor channel conditions and therefore reach a higher throughput. Often system designers assume that for a "large enough" coding delay $K$ the ergodic nature of the fading channel can be captured and an outage probability close to zero can be achieved. Such a scenario would result in a throughput equivalent to the transmission rate.

Figure 4.8 plots the throughput achieved if $R = \beta C_{\text{erg}}$ for various $\beta \in [0, 1]$. We see for $\beta = 0.5$ and $\beta = 0.7$ that the throughput is close to the transmission rate when $K \approx 10$ and $K \approx 25$, respectively. For $\beta = 0.99$ the throughput is far below the transmission rate even for $K = 500$. Clearly, the closer the transmission rate is to ergodic capacity the harder it is to capture the ergodic nature of the channel.
The transmission rate of the system should be selected to achieve the MZT_Rt rather than attempting to achieve $C_{\text{erg}}$, which is unattainable for finite $K$. Figure 4.7 plots the optimal transmission rate $R_{\text{MZT}_R}^*$ as a function of $K$ for $P_{av} \in \{0, 5, 10\}$dB. The relationship between $R_{\text{MZT}_R}^*$ and $K$ is not as obvious. $R_{\text{MZT}_R}^*$ does converge to ergodic capacity as $K \rightarrow \infty$, but not monotonically and can fluctuate a great deal for small $K$. This can be attributed to the behavior of the tail of $f_{C_K}(R)$, the distribution of the instantaneous capacity, as a function of $K$ with $\chi_2^2$ fading. For different fading processes the behavior of $R_{\text{MZT}_R}^*$ may be different. This highlights the need to properly select $R_{\text{MZT}_R}^*$ by solving (4.4).

Selecting a transmission rate that overshoots (is larger than) $R_{\text{MZT}_R}^*$ will result in a loss in throughput when compared to MZT_Rt, as seen in Figure 4.3. The severity of

Figure 4.6: MZT_Rt($P_{av}, K$) as a function of coding delay $K$ for $P_{av} \in \{0, 5, 10\}$dB. Ergodic capacity for each of these values of $P_{av}$ is plotted for reference.
Figure 4.7: $R_{\text{MZT}_\text{RT}}$ as a function of coding delay $K$ for $P_{av} \in \{0, 5, 10\}$ dB. Ergodic capacity for each of these values of $P_{av}$ is plotted for reference.

The throughput loss depends on $K$. For larger $K$, the throughput vs. transmission rate curve is narrower, which tells us that the throughput loss is more severe if the optimal transmission rate is overshot. In the limit when $K = \infty$, selecting a transmission rate infinitesimally larger than the optimal one yields zero throughput. Therefore, system designers must be very careful not to overshoot the optimal transmission rate for larger $K$. This phenomenon also suggests a trade off: throughput vs. system robustness. For larger $K$, MZT$_{RT}$ is higher but the loss in throughput if the optimal transmission rate is overshot is more severe. For smaller $K$, MZT$_{RT}$ is lower but the loss in throughput if the optimal rate is overshot is less severe.

Underestimating the transmission rate also yields similar losses in throughput. This is seen in Figure 4.8 where we compare MZT$_{RT}(P_{av}, K)$, achieved using $R_{M^\text{MZT}_\text{RT}}$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 4.8: Throughput vs. $K$ achieved with an SNR of 10dB for transmission rates set to $R = \beta C_{\text{avg}}$ with $\beta \in \{0.5, 0.7, 0.99\}$. MZT($P_{\text{av}}, K$) is also indicated. Transmitting at the optimal transmission rate of $R^*$ yields significant gains in throughput vs. $R = \beta C_{\text{avg}}$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 4.9: Throughput vs. Outage Probability vs. for various values of $K$ and $P_{av} = 10$dB. The optimal outage probability for each $K$ is different.

to the throughput achieved using $R = \beta C_{erg}$ for various $\beta \in (0, 1)$. For $\beta = 0.5$ and $\beta = 0.7$ the throughput curves plateau as a function of $K$ and are far below $MZT_{RT}(P_{av}, K)$. This occurs because the transmission rate is underestimated and a larger rate, yielding a larger throughput, can be supported for that $K$. These examples clearly illustrate the importance of properly selecting the transmission rate.

The uniqueness of $MZT_{RT}$ implies that there is a unique outage probability $P_{out}(R^{*}_{MZT_{RT}}, P_{av}, K)$ that corresponds to the maximum throughput. Figure 4.9 plots throughput vs. the optimal outage probability for various values of $K$ and $P_{av} = 10$dB. $MZT_{RT}(P_{av}, K)$ corresponds to the peak of each curve. Figure 4.10 plots the optimal outage probability as a function of $K$ for $P_{av} \in \{0, 5, 10\}$dB. From both figures we can clearly see that the optimal outage probability can be quite high, especially for
small coding delays. For example if $P_{av} = 0$dB and $K = 1$ then the optimal outage probability is $P_{out}(R_{MT_{RT}}^*, P_{av}; 1) = 0.53$. This is very interesting, because it suggests that in order to maximize the throughput it is necessary to lose over half of the transmitted codewords to outage. When compared to the conventional practice of constraining the outage probability to be rather small, say $\epsilon = 0.01$, the result is rather striking. The penalty of outages means zero rate for $K$ consecutive blocks, which is small for small $K$ and large for large $K$. For small $K$ the ergodic nature of the channel cannot be captured, and the instantaneous capacity is highly variable. Throughput is maximized by exploiting this variability and transmitting codewords with a high rate and therefore a large outage probability, since the penalty for outage is small. For larger but finite $K$, the instantaneous capacity is still a random quantity but is not highly variable and codewords begin to see "average" channels. Since the penalty for outage is large, throughput is optimized by selecting rates that the "average" channel can support. Intuitively, this makes sense since the optimal outage probability decreases as $K \to \infty$ and at the extreme $K = \infty$ the optimal outage is 0, corresponding to a maximum throughput of $C_{erg}$.

4.1.5 Behavior of MZT and $R_{MT_{RT}}^*$ under Ricean fading

We now examine the effect of the fading distribution on both the optimal transmission rate and the maximum throughput. Figure 4.11 illustrates $M_{RT}$ as a function of $K$ for Rayleigh fading and $P_{av} = 10$dB. It also shows $M_{RT}$ for a Ricean fading sce-
Figure 4.10: The optimal outage probability $P_{out_{MZT_{RT}}, K, P_{av}}$ as a function of $K$ for $P_{av} \in \{0, 5, 10\}$ dB. For small values of $K$ the optimal outage probability can be very high, over 0.53 for $P_{av} = 0$ dB.

Ricean distributions are commonly used to model communication systems that have line of sight (LOS) between transmitter and receiver. In this particular example, the power in the LOS component is 9 dB higher than the average of other scattering of the transmitted signal. As with the Rayleigh distribution, $MZT_{RT}$ converges monotonically to $C_{erg}$ as $K \to \infty$ for Ricean fading. However, since the distribution of fading states for the Rayleigh and Ricean scenario are different, the ergodic capacity in each case is also different.

Previously we saw that the optimal transmission can behave in a rather non-intuitive manner. For Rayleigh fading, the $R_{MZT_{RT}}$ did not increase monotonically with $K$. We see this illustrated in Figure 4.12 which plots the optimal transmission
Figure 4.11: $\text{MZT}_{\text{RT}}$ plotted as a function of coding delay $K$ for $P_{av} = 10$dB. The simulations are for Rayleigh fading and Ricean fading with a LOS component 9dB stronger than the other fading components. In both cases $\text{MZT}_{\text{RT}}$ converges to ergodic capacity.
Figure 4.12: Optimal transmission rate $R^*_{\text{MzT}}$ plotted as a function of coding delay $K$ for $P_{av} = 10$dB. The simulations are for Rayleigh fading and Ricean fading with a LOS component 9dB stronger than the other fading components. In both cases the optimal transmission rate converges to ergodic capacity.

against coding delay for $P_{av} = 10$dB and Rayleigh fading. Here $R^*_{\text{MzT}}$ at first decreases with $K$ before increasing and converging to ergodic capacity. However, this is not necessarily the case for all fading distributions. Figure 4.12 also shows $R^*_{\text{MzT}}$ for a Ricean fading distribution. Clearly in this case the optimal transmission rate behaves differently, it is a non-decreasing function of $K$ which also converges to ergodic capacity.

We can gain insight into the behavior of the optimal transmission rate by examining the form of $\text{MzT}_{\text{RT}}$. Since,

$$\text{MzT}_{\text{RT}} = \sup_R R[1 - P_{\text{out}}(R, P_{av}, K)]$$  \hspace{1cm} (4.25)
we know that

$$\log [\text{MZT}_{\text{RT}}] = \sup_R \{\log [R] + \log [1 - P_{\text{out}}(R, P_{\text{av}}, K)]\}. \quad (4.26)$$

Since $\log(\cdot)$ is a monotonically increasing function, the optimal transmission rates that maximizes both (4.25) and (4.26) are the same. By examining (4.26) we realize the optimal transmission rate occurs when the slope of $\log [\text{MZT}_{\text{RT}}]$ equals zero. This implies that the slope of $\log [R]$ and $\log [1 - P_{\text{out}}(R, P, K)]$ have the same amplitude but different signs or

$$\frac{d}{dR} \log [R] = -\frac{d}{dR} \log [1 - P_{\text{out}}(R, P, K)]. \quad (4.27)$$

In Figures 4.13 and 4.14 we plot $\log [R]$ and $\log [1 - P_{\text{out}}(R, P_{\text{av}}, K)]$ for various $K$ and $P_{\text{av}} = 10\text{dB}$ for Rayleigh and Ricean fading, respectively. Here the LOS component of the fading process is again 9dB stronger than the other components. The optimal transmission rate occurs when the slope of $\log [1 - P_{\text{out}}(R, P_{\text{av}}, K)]$ begins to decrease faster than the slope of $\log [R]$ increases. Since we have previously shown that for scheme RT that $R[1 - P_{\text{out}}(R, P_{\text{av}}, K)]$ is a log-concave function, there is a unique optimal transmission rate. For any $R < R^{*}_{\text{MZT}_{\text{RT}}}$, $|\frac{d}{dR} \log [R]| > |\frac{d}{dR} \log [1 - P_{\text{out}}(R, P, K)]|$ while for $R > R^{*}_{\text{MZT}_{\text{RT}}}$, $|\frac{d}{dR} \log [R]| < |\frac{d}{dR} \log [1 - P_{\text{out}}(R, P, K)]|$. For $R = R^{*}_{\text{MZT}_{\text{RT}}}$, the two slopes are equal in amplitude but have opposite signs. As an example we plot the tangent line of both $\log [R]$ and $\log [1 - P_{\text{out}}(R, P_{\text{av}}, K)]$ at the optimal transmission rate $R^{*}_{\text{MZT}_{\text{RT}}}$ for $K = 1$. Figures 4.13 and 4.14 also plot the optimal $R^{*}_{\text{MZT}_{\text{RT}}}$ for various $K$ for Rayleigh and
Figure 4.13: \( \log[R] \) and \( \log[1 - P_{\text{out}}(R, P_{\text{av}}, K)] \) for \( P_{\text{av}}, K \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100\} \) and Rayleigh fading. The optimal \( R^*_\text{MZT}_{\text{RT}} \) are plotted for all \( K \) and the tangent lines to \( \log[R] \) and \( \log[1 - P_{\text{out}}(R, P_{\text{av}}, K)] \) at \( R^*_\text{MZT}_{\text{RT}} \) for \( K = 1 \). The optimal transmission rate is not a monotonic function of \( K \).

Ricean fading, respectively. Clearly, the optimal transmission rates have the non-monotonic behavior for Rayleigh fading, but not for Ricean, due to the behavior of \( \log[1 - P_{\text{out}}(R, P_{\text{av}}, K)] \).
Figure 4.14: $\log[R]$ and $\log[1 - P_{out}(R, P_{av}, K)]$ for $P_{av}$, $K \in \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 50, 60, 70, 80, 90, 100\}$ and Ricean fading. The LOS component of the fading are 9dB stronger than the non-LOS components. The optimal $R_{MT}^*$ are plotted for all $K$ and the tangent lines to $\log[R]$ and $\log[1 - P_{out}(R, P_{av}, K)]$ at $R_{MT}^*$ for $K = 1$. For Ricean fading the optimal transmission rate is a monotonic non-decreasing function of $K$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
4.2 Maximum zero-outage throughput with scheme ID (MZT_ID)

4.2.1 Mathematical formulation

Scheme ID is a more complex retransmission scheme than scheme RT. Feedback is used not only to guarantee that codewords are successfully received but also to improve communications performance. This is accomplished by a more intelligent receiver design. Since the receiver saves, rather than discards, codewords that are in outage and optimally combines them with subsequent retransmitted versions (prior to making a decoding decision), the outage probability decreases with each retransmission.

In general when combining J codewords with a K block coding delay using MRRC the instantaneous capacity is given by

$$C_K(P_{av}, K, J) = \frac{1}{K} \sum_{k=0}^{K-1} \log \left( 1 + \sum_{j=1}^{J} \alpha_{k,j} P_{av} \right).$$  \hspace{1cm} (4.28)

The inner summation is the SNR gain provided with MRRC while the outer summation is due to the fact that the codeword spans K blocks of the BF-AWGN channel. The associated outage probability is

$$P_{out}(R, P_{av}, K, J) = \text{Prob}[R > C_K(P_{av}, K, J)].$$  \hspace{1cm} (4.29)

As a result, the probability of outage after s consecutive transmission attempts be-
comes
\[
\text{Prob}\left(\bigcap_{j=1}^{s} \text{out}_j\right) = \text{Prob}\left[\left( R > \frac{1}{K} \sum_{k=0}^{K-1} \log(1 + \sum_{j=1}^{s} \alpha_{k,j} P_{av})\right) \cap \left( R > \frac{1}{K} \sum_{k=0}^{K-1} \log(1 + \sum_{j=1}^{s-1} \alpha_{k,j} P_{av})\right) \cap \ldots \cap \left( R > \frac{1}{K} \sum_{k=0}^{K-1} \log(1 + \alpha_{k,1} P_{av})\right) \right] 
\]
(4.30)

for scheme ID. In general, (4.30) is difficult to solve analytically and must be computed numerically. The numerical solution to (4.30) can be used to determine the service-time distribution (3.1), the expected service time \( \mathbb{E}[S] \), and the associated maximum zero-outage throughput for scheme ID \( \text{MZT}_{ID} \).

4.2.2 Uniqueness of \( \text{MZT}_{ID} \) when \( K = 1 \)

The difficulty of determining a closed form expression for the service time is apparent from (4.30). However, when \( K = 1 \) and the channel fading is \( \chi_2^2 \), (4.30) admits a closed form expression and therefore the service time distribution and expected service-time can be determined analytically.

**Theorem 4.2.1.** If \( K = 1 \) and the channel fading is \( \chi_2^2 \) then

\[
\mathbb{E}[S] = \frac{e^R + P_{av} - 1}{P_{av}}. 
\]
(4.31)

**Proof.** For simplicity, let \( x = \frac{e^R - 1}{P_{av}}. \) Then since \( K = 1 \) the probability of outage on
successive transmission attempts (4.30) is given by

\[
\text{Prob}(\text{out}_1) = \text{Prob}[x > \alpha_1],
\]

\[
\text{Prob}\left(\bigcap_{j=1}^{2} \text{out}_j\right) = \text{Prob}\left[(x > \alpha_1 + \alpha_2) \cap (x > \alpha_1)\right],
\]

\[
\vdots
\]

\[
\text{Prob}\left(\bigcap_{j=1}^{s} \text{out}_j\right) = \text{Prob}\left[(x > \sum_{j=1}^{s} \alpha_j) \cap (x > \sum_{j=1}^{s-1} \alpha_j) \cap \ldots \cap (x > \alpha_1)\right],
\]

where \(\sum_{j=1}^{s} \alpha_j\) is a \(\chi_{2\alpha}^2\) random variable. This can be determined by integrating the joint distribution of the \(\alpha_j\)’s over the appropriate region

\[
\text{Prob}\left(\bigcap_{j=1}^{s} \text{out}_j\right) = \int_{0}^{x} \int_{0}^{x-\alpha_1} \ldots \int_{0}^{x-\sum_{j=1}^{s-1} \alpha_i} f(\alpha_1, \alpha_2, \ldots, \alpha_s) \, d\alpha_s d\alpha_{s-1} \ldots d\alpha_1.
\]

(4.33)

Since the channel gains are assumed to be i.i.d.,

\[
f(\alpha_1, \alpha_2, \ldots, \alpha_s) = \prod_{i=1}^{s} f(\alpha_i).
\]

Then (4.33) becomes

\[
\text{Prob}\left(\bigcap_{j=1}^{s} \text{out}_j\right) = 1 - e^{-x} \sum_{j=0}^{s-1} \frac{x^j}{j!}.
\]

(4.34)

Using this we find the service time distribution (3.1) to be

\[
\text{Prob}(S = s) = \text{Prob}\left(\bigcap_{k=1}^{s-1} \text{out}_k\right) - \text{Prob}\left(\bigcap_{k=1}^{s} \text{out}_k\right)
\]

\[
= e^{-x} \sum_{j=0}^{s-1} \frac{x^j}{j!}.
\]

(4.35)
The expected service time can then be computed as

\[ E[S] = \sum_{s=1}^{\infty} s \cdot \text{Prob}(S = s) \]

\[ = e^{-x} \sum_{s=1}^{\infty} \frac{sx^{s-1}}{(s-1)!} \]

\[ = e^{-x} \left( \sum_{s=1}^{\infty} \frac{(s-1)x^{s-1}}{(s-1)!} + \sum_{s=1}^{\infty} \frac{x^{s-1}}{(s-1)!} \right) \]

\[ = e^{-x} \left( \sum_{s=0}^{\infty} \frac{sx^{s}}{s!} + \sum_{s=1}^{\infty} \frac{x^{s-1}}{(s-1)!} \right) \]

\[ = e^{-x} \left( x \sum_{s=1}^{\infty} \frac{x^{s-1}}{(s-1)!} + e^{x} \right) \]

\[ = e^{-x} (xe^{x} + e^{x}) \]

\[ = 1 + x. \quad (4.36) \]

By substituting the value of \( x \) we arrive at \((4.31)\) as desired. □

Using the form of \( E[S] \) described above, \( \text{MZT}_{\text{ID}}(P_{av}, 1) \) can be written as

\[ \text{MZT}_{\text{ID}}(P_{av}, 1) = \sup_{R} \frac{RP_{av}}{e^{R} + P_{av} - 1}. \quad (4.37) \]

Note that this equation is quite different from \((4.4)\). For scheme ID, the throughput is no longer the transmission rate multiplied by the success probability. The difference is due to the receiver performing MRRC with the retransmitted codewords.

As was the case for scheme RT, the special case when \( K = 1 \) for scheme ID also admits a semi-explicit solution for the optimal transmission rate and therefore for \( \text{MZT}_{\text{ID}}(P_{av}, 1) \).
Theorem 4.2.2. If the channel fading is $\chi^2_2$ then (4.37) has a unique maximizer.

Proof. Let $T(R) = \frac{RP_{av}}{e^{R+P_{av}-1}}$. Its first two derivatives are given by

$$T'(R) = \frac{P_{av}(e^R + P_{av} - 1 - Re^R)}{(e^R + P_{av} - 1)^2}, \quad \text{(4.38)}$$
$$T''(R) = \frac{e^R P_{av}[R(e^R - P_{av} + 1) - 2(e^R + P_{av} - 1)]}{(e^R + P_{av} - 1)^3}. \quad \text{(4.39)}$$

After some algebraic manipulations it can be shown that

$$T(R)T''(R) \leq \left[T'(R)\right]^2 \quad \text{(4.40)}$$

is satisfied, which means that $T(R)$ is a log-concave function [28]. Then from convex optimization theory, $\log T(R)$ has a unique maximizer $R_{\text{MZTID}}^*$ on the convex set $\mathbb{R}_+$. If we compose $T(R)$ with the monotonic increasing function $e^x$, then $e^{T(R)}$ has the same maximizer $R_{\text{MZTID}}^*$, completing the proof. $\Box$

Theorem 4.2.3. If $K = 1$ and the channel gains follow a $\chi^2_2$ distribution then

$$\text{MZTID}(P_{av}, 1) = \frac{(\mathcal{W}(\frac{P_{av} - 1}{e}) + 1) P_{av}}{e^\mathcal{W}(\frac{P_{av} - 1}{e}) + P_{av} - 1}. \quad \text{(4.41)}$$

Proof. Let $T(R) = \frac{RP_{av}}{e^{R+P_{av}-1}}$. Let $f(R) = \log[T(R)]$, which is concave in $R$ since $T(R)$ is log-concave as shown in Theorem 4.2.2. Taking the derivative of $f(R)$ with respect to $R$ and equating with zero, we see that transmission rate corresponding to the critical point is the solution to $e^R(1 - R) + P_{av} - 1 = 0$, which turns out to be $R_{\text{MZTID}}^* = \mathcal{W}(\frac{P_{av} - 1}{e}) + 1$. We know that this rate corresponds to a throughput maximum (rather than a minimum) from Theorem 4.2.2. Substituting this back into $T(R)$ we arrive at (4.41). $\Box$
Note that Theorem 4.2.2 was proved by illustrating that $T(R)$ directly is a log-concave function, rather than by showing that $\frac{1}{S_i^1}$ is log-concave. Indeed, $\frac{1}{S_i^1}$ is not log-concave in this case, highlighting the fact that Theorem 4.1.1 is a sufficient, but not necessary, condition for uniqueness.

### 4.2.3 Properties of $\text{MZT}_{\text{ID}}$

Since it optimally combines multiple codewords to make a decoding decision, scheme ID must perform at least as well as scheme RT. Intuitively, if discarding codewords in error is optimal, then the optimal combining scheme would adopt this strategy.

We can prove this explicitly for $K = 1$.

**Theorem 4.2.4.** If $K = 1$ and the channel fading follows a $\chi^2$ distribution then $\text{MZT}_{\text{ID}}(P_{\text{av}}, 1) \geq \text{MZT}_{\text{RT}}(P_{\text{av}}, 1)$.

**Proof.** Let $\text{MZT}_{\text{RT}}(P_{\text{av}}, 1) = R_1[1 - P_{\text{out}}(R_1, P_{\text{av}}, 1)] = R_1 e^{-\frac{(eR^*_1-1)}{P_{\text{av}}}}$ and let $\text{MZT}_{\text{ID}}(P_{\text{av}}, 1) = \frac{R_2P_{\text{av}}}{e^{R_2}+P_{\text{av}}-1}$, where $R_1 = R^*_1$ and $R_2 = R^*_2$. Then

\[
\log \text{MZT}_{\text{RT}}(P_{\text{av}}, 1) = \log R_1 - \left(\frac{eR^*_1-1}{P_{\text{av}}}\right)
\leq \log R_2 - \log\left(1 + \frac{eR^*_1-1}{P_{\text{av}}}\right)
= \log R_2 - \log\left(\frac{eR^*_1-1 + P_{\text{av}}}{P_{\text{av}}}\right)
\leq \log R_2 - \log\left(\frac{eR^*_2-1 + P_{\text{av}}}{P_{\text{av}}}\right)
= \log \text{MZT}_{\text{ID}}(P_{\text{av}}, 1).
\]
The first inequality comes from the fact that \( x \geq \log(1 + x), \forall x \geq 0 \) as well as \( R_1, R_2 > 0 \) and \( P_{\text{av}} > 0 \). The second inequality occurs since \( R_2 \) is the optimizer for scheme ID. Finally, since \( \log(x) \) is a monotonically increasing function of \( x \), \( \text{MZT}_{\text{RT}}(P_{\text{av}}, 1) \leq \text{MZT}_{\text{ID}}(P_{\text{av}}, 1) \), completing the proof. □

The gain in the throughput from scheme ID over scheme RT is due to the fact that the feedback is implicitly used to optimize the transmission rate. Incremental diversity reduces the outage probability on each retransmission attempt. This allows the transmitter to more aggressively select the transmission rate resulting in a larger throughput than with scheme RT.

4.2.4 Simulation results

We now empirically verify some of the Theorems and properties of \( \text{MZT}_{\text{ID}} \) via Monte Carlo simulation. In all of the simulations we assume that the channel fading is \( \chi^2 \).

Throughput is plotted against transmission rate in Figure 4.15 with \( K = 1 \) and \( P_{\text{av}} = 10\text{dB} \) for both schemes RT and ID. \( \text{MZT}_{\text{ID}} \) and \( \text{MZT}_{\text{RT}} \) correspond to the peaks of each curve. We see that for Scheme ID there is also a single peak in the throughput vs. rate curve and a unique optimal transmission rate. This empirically validates Theorem 4.2.2. We also see that the throughput for scheme ID is clearly higher than that using scheme RT, verifying Theorem 4.2.4. Moreover, we see that the gap between Scheme ID and Scheme RT is larger for large \( R \). This is due to the fact that for large \( R \) there are frequent outages and more retransmission attempts. This
Figure 4.15: With $\chi^2_2$ fading Throughput vs. transmission rate for an SNR of 10dB using simple retransmission and incremental diversity. Incremental diversity provides significant gains for larger rates for which outages are more frequent.

results in more opportunities for codeword combining yielding greater throughput.

As a means to estimate the performance penalty for having a finite coding delay $K$, both $\text{MZT}_{\text{ID}}(P_{av},K)$ and $\text{MZT}_{\text{RT}}(P_{av},K)$ are plotted as a function of the transmitted power $P_{av}$ for various $K$ in Figure 4.16. For a target throughput of 1 nat/sec/Hz and a coding delay $K = 1$, scheme ID provides a 1.22dB gain over scheme RT. When $K = 10$ and $K = 100$ the gain shrinks to 0.18dB and 0.04dB, respectively. The decreasing difference between the two schemes can be explained as follows. For a given transmission rate outage events are more likely when $K$ is small. Therefore, there are more retransmissions and more opportunities for codeword combining, resulting in higher throughput for scheme ID vs. scheme RT. In the limit as $K \to \infty$ the
Figure 4.16: With $\chi^2$ fading $\text{MZT}_{\text{RT}}(P_{av}, K)$ and $\text{MZT}_{\text{ID}}(P_{av}, K)$ vs SNR for $K = 1, K = 10$, and $K = 100$. When $K = 1$ incremental diversity provides a gain of approximately 1dB. For $K = 5$, when outages are less likely, the gains are much smaller and become negligible for $K = 100$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 4.17: $M Z T_{D}(P_{av}, K)$ and $M Z T_{RT}(P_{av}, K)$ plotted against coding delay $K$ for $P_{av} \in \{0, 5, 10\}$dB. For any coding delay the maximum throughput for both schemes can be predicted. Scheme ID outperforms scheme RT for small $K$ due to the fact that outages are more frequent resulting in more opportunity to perform codeword combining.

Figure 4.17 plots $M Z T_{D}(P_{av}, K)$ as a function of the coding delay $K$ for various values of $P_{av}$. The maximum throughput for scheme RT is also plotted for reference. As is the case for scheme RT we see that the maximum throughput with scheme ID can be quite far from ergodic capacity for finite $K$. Also, as $K \to \infty$, $M Z T_{D}(P_{av}, K) \to C_{erg}(P_{av})$. As with scheme RT, the convergence appears to be monotonic.

As is the case for scheme RT, the transmission rate for scheme ID should be selected carefully. Rather than trying to achieve $C_{erg}(P_{av})$ the transmitter should
Figure 4.18: \( R_{\text{MZTID}}^* (P_{av}, K) \) and \( R_{\text{M Zhu}}^* (P_{av}, K) \) plotted against coding delay \( K \) for \( P_{av} \in \{0, 5, 10\} \)dB. For any coding delay the maximum throughput for both schemes can be predicted. We see that for scheme ID the optimal transmission rate is larger and can even be higher than \( C_{\text{erg}} \). This is due to the fact that codeword combining at the receiver allows the transmitter to use a higher rate for the same average service time.
select the transmission rate to maximize the throughput. Figure 4.18 plots the optimal transmission rate \( R^*_{\text{MZT-ID}}(P_{av}, K) \) against the coding delay \( K \) for various values of \( P_{av} \). The optimal transmission rate for scheme RT is included for reference. We see that as \( K \to \infty \) that the optimal transmission rate converges to ergodic capacity, \( R^*_{\text{MZT-ID}}(P_{av}, K) \to C_{\text{erg}}(P_{av}) \). Similar to scheme RT, the convergence is not monotonic and the optimal transmission can fluctuate a great deal as a function of \( K \), especially for small \( K \). In fact for scheme ID the optimal transmission rate can actually be higher than the ergodic capacity of the channel. We note that this does not contradict information theoretic capacity theorems as the resulting throughput is always less than ergodic capacity. However, this is rather non-intuitive as in practice transmission rates lower than ergodic capacity are normally selected. This can be explained by the fact that the codeword combining of scheme ID reduces the outage probability with each retransmission attempt allowing the transmitter to more aggressively select the transmission rate — in some cases resulting in rates higher than ergodic capacity.

4.3 Maximum \( \epsilon \) throughput with scheme RT (\( M\epsilon T_{RT} \))

Many applications, including streaming video and voice, are sensitive to delay and jitter, the variance of the delay. These applications may not be compatible with a possibly infinite number of transmission attempts for each codeword. Limiting the number of transmission attempts provides a tighter bound on delay and jitter at the cost of not guaranteeing successful transmission of every codeword. We illustrate by
generalizing scheme RT, due to its analytical tractability, to at most $L$ attempts. We denote this as $\text{RT}_L$.

With scheme $\text{RT}_L$ we have,

$$\text{Prob}(S = s) = \begin{cases} [P_{\text{out}}(R, P_{\text{av}}, K)]^{s-1}[1 - P_{\text{out}}(R, P_{\text{av}}, K)] & s < L \\ [P_{\text{out}}(R, P_{\text{av}}, K)]^{s-1} [1 - P_{\text{out}}(R, P_{\text{av}}, K)] + [P_{\text{out}}(R, P_{\text{av}}, K)]^L & s = L \\ 0 & s > L. \end{cases}$$

(4.43)

as the service time distribution. For $s < L$ it is a geometric distribution with parameter $[1 - P_{\text{out}}(R, P_{\text{av}}, K)]$. Since $L$ is the maximum number of transmission attempts, it is impossible for the service time to exceed $L$ and thus the service time distribution is 0 for $s > L$. Finally, $s = L$ consists of those codewords that are successfully transmitted with exactly $L$ attempts or require more attempts and are in outage after $L$ attempts. This "effective outage" probability can be found by summing the tail of a geometric distribution for $s = L + 1, \ldots, \infty$. That is,

$$P_{\text{out}}^{\text{eff}} = \sum_{s=L+1}^{\infty} [P_{\text{out}}(R, P_{\text{av}}, K)]^{s-1}[1 - P_{\text{out}}(R, P_{\text{av}}, K)]$$

$$= [P_{\text{out}}(R, P_{\text{av}}, K)]^L$$

(4.44)

From (4.43), the expected service time is given by

$$E[S] = \frac{1 - [P_{\text{out}}(R, P_{\text{av}}, K)]^L}{[1 - P_{\text{out}}(R, P_{\text{av}}, K)]}.$$  

(4.45)

We define the maximum $\epsilon$-throughput for scheme $\text{RT}_L$ ($\text{MeT}_{\text{RT}_L}$) as the highest achievable throughput using at most $L$ transmission attempts per codeword, with the
effective outage probability no greater than \( \epsilon \); that is,

\[
M \epsilon T_{RTL}(P_{av}, K, L) = \sup_R \left\{ \frac{R[1 - P_{eff}]}{E[S]} : P_{out} \leq \epsilon \right\} = \sup_R \left\{ R[1 - P_{out}(R, P_{av}, K)] : P_{out} \leq \epsilon \right\}. \tag{4.46}
\]

It is remarkable that \( M \epsilon T_{RTL} \) is found by maximizing the same objective function as the one for \( MZT_{RT} \) in (4.4).

The only difference in finding \( MZT_{RT} \) and \( M \epsilon T_{RTL} \) is that the optimization is performed over different sets of transmission rates. For \( MZT_{RT} \), this set is restricted to those rates that result in an effective outage probability less than a target \( \epsilon \). Clearly as \( L \to \infty \), \( M \epsilon T_{RTL} \to MZT_{RT} \) since the constraint on the transmission rate disappears and \( P_{out}^{eff} \to 0 \).

The effect of a limited number of transmission attempts can be seen in Figure 4.19, which plots \( M \epsilon T_{RTL} \) for various values of \( L \) and \( P_{out}^{eff} = 0.01 \). Clearly as \( L \) increases the set of valid transmission rates over which the optimization in (4.46) is performed increases, resulting in a throughput that approaches \( MZT_{RT} \). As \( L \) increases, and the set of transmission rates over which the optimization in (4.46) is performed includes the transmission rate that achieves \( MZT_{RT} \), then there is no throughput benefit in further increasing \( L \). This can be seen in Figure 4.19 in which \( M \epsilon T_{RTL} \) for \( L = 5 \) is the same as \( MZT_{RT} \).
Figure 4.19: $\text{MeT}_\text{RT}$ with $P_{\text{eff}}^{\text{out}} = 0.01$ for $L = \{1, 2, 3, 4, 5\}$ and $P_{\text{av}} = 10$dB. Also indicated is the throughput vs. transmission rate curve with no limit to the number of transmission attempts — MZT$_{\text{RT}}$ is the peak of this curve. We see that allowing multiple transmission attempts can increase the average spectral efficiency for the same effective outage probability.
Chapter 5
Outage minimization under a peak and average power constraint

Outage minimization for a fixed transmission rate is a pre-requisite step required in order to maximize the throughput of delay-limited communication systems with CSI-RT. The solution under the short-term and long-term average power constraints was given in [24]. We devote this Chapter to the fixed rate outage minimization under both peak and average power constraints, which to the best of our knowledge has yet to be solved in the literature. Those readers interested only in throughput maximization can proceed directly to Chapter 6.

5.1 Short-term average and peak power constraints

Under the short-term average and peak power constraints, the minimum outage probability

$$\min_{\gamma} \{P_{\text{out}}(R, \gamma, K) : \gamma \in \mathcal{O}_K^{\text{st}}(P_{\text{av}}, P_p)\}$$  \hspace{1cm} (5.1)

is achieved by an optimal outage minimizing power allocation strategy $\gamma^{\text{st}} \in \mathcal{O}_K^{\text{st}}(P_{\text{av}}, P_p)$.

Theorem 5.1.1. The power allocation strategy that satisfies the short-term average power...
and peak power constraints is given by

\[
\tilde{\lambda}_{\gamma_k}(\alpha) = \begin{cases} 
\min \left( \max \left( \tilde{\lambda}_{\gamma_k}(\alpha) - \frac{1}{\alpha_k}, 0 \right), \mathcal{P}_p \right) & \text{if } \mathcal{P}_p > \mathcal{P}_{av} \\
\mathcal{P}_p & \text{if } \mathcal{P}_p \leq \mathcal{P}_{av} 
\end{cases}
\]  

(5.2)

with \( \tilde{\lambda}_{\gamma_k}(\alpha) \) the solution to

\[
\frac{1}{K} \sum_{k=0}^{K-1} \min \left( \max \left( \tilde{\lambda}_{\gamma_k}(\alpha) - \frac{1}{\alpha_k}, 0 \right), \mathcal{P}_p \right) = \mathcal{P}_{av}.
\]  

(5.3)

Proof. The power allocation policy that solves (5.1) is the same as that which maximizes the \( K \)-block instantaneous capacity (2.15) for a codeword affected by channel \( \alpha \). If \( \mathcal{P}_p \leq \mathcal{P}_{av} \) then (2.15) is trivially maximized by always transmitting at the peak power, \( \tilde{\gamma}_{\gamma_k}(\alpha) = \mathcal{P}_p \).

If \( \mathcal{P}_p > \mathcal{P}_{av} \), then (2.15) is maximized by solving

\[
\min_{\gamma} \left\{ - \sum_{k=0}^{K-1} \log(1 + \alpha_k \gamma_k) : 0 \leq \gamma_k \leq \mathcal{P}_p, \sum_{k=0}^{K-1} \gamma_k = K \mathcal{P}_{av} \right\}. 
\]  

(5.4)

We first set up the Lagrangian functional in standard form

\[
J = - \sum_{k=0}^{K-1} \log(1 + \alpha_k \gamma_k) - \sum_{k=0}^{K-1} \psi_k \gamma_k + \sum_{k=0}^{K-1} \mu_k (\gamma_k - \mathcal{P}_p) + \nu (\sum_{k=0}^{K-1} \gamma_k - \mathcal{P}_{av}). 
\]  

(5.5)

Since both the objective function and set of feasible points are convex, we know that the Karush-Kuhn-Tucker (KKT) conditions are sufficient for optimality. Therefore any feasible point that satisfies the KKT conditions is the globally optimal point that minimizes the objective function. The optimal power allocation policy \( \gamma^* \) and the
associated $\psi^*_k$, $\mu^*_k$, and $\nu^*$, satisfy [28]

\begin{align}
\gamma^*_k &\geq 0 \quad (5.6a) \\
\gamma^*_k &\leq \mathcal{P}_p \quad (5.6b) \\
\sum_{k=0}^{K-1} \gamma^*_k &= K\mathcal{P}_{av} \quad (5.6c) \\
\psi^*_k &\geq 0 \quad (5.6d) \\
\mu^*_k &\geq 0 \quad (5.6e) \\
\psi^*_k\gamma^*_k &= 0 \quad (5.6f) \\
\mu^*_k(\gamma^*_k - \mathcal{P}_p) &= 0 \quad (5.6g) \\
\frac{\partial J}{\partial \gamma^*_k} &= \frac{-\alpha^*_k}{1 + \alpha^*_k\gamma^*_k} - \psi^*_k + \mu^*_k + \nu^* = 0 \quad (5.6h)
\end{align}

with (5.6a), (5.6b), and (5.6c) the set of feasible points, (5.6d) and (5.6e) the non-negativity of the Lagrange multipliers, (5.6f) and (5.6g) the complimentary slackness condition, and (5.6h) the vanishing gradient of the Lagrangian at the optimal solution [28]. It is clear that a solution of the form

$$\tilde{\gamma}^*_k(\alpha) = \gamma^*_k = \min \left( \max \left( \tilde{\lambda}^*(\alpha) - \frac{1}{\alpha_k}, 0 \right), \mathcal{P}_p \right)$$

(5.7)

with $\tilde{\lambda}^*(\alpha) = \frac{1}{\nu}$ the solution to

$$\frac{1}{K} \sum_{k=0}^{K-1} \min \left( \max \left( \tilde{\lambda}^*(\alpha) - \frac{1}{\alpha_k}, 0 \right), \mathcal{P}_p \right) = \mathcal{P}_{av},$$

(5.8)

satisfies (5.6a-5.6h). Thus, $\tilde{\gamma}^*$ is the power allocation strategy that minimizes outage probability under a peak and short-term average power constraint.  

□
Note that when $P_p > P_{av}$ the optimal solution has three regions. We use a constant power allocation of $\lambda^st(\alpha) = P_p$ when $\lambda^st(\alpha) - \frac{1}{\alpha_k} \geq P_p$. Next, we apply the waterfilling solution $\lambda^st(\alpha) = \lambda^st(\alpha) - \frac{1}{\alpha_k}$ when $0 < \lambda^st(\alpha) - \frac{1}{\alpha_k} < P_p$. Finally, no power is allocated $\lambda^st(\alpha) = 0$ when $0 > \lambda^st(\alpha) - \frac{1}{\alpha_k}$.

The functional implementation of this power control strategy is straightforward. The receiver relays CSI to the transmitter. If the current channel is $\alpha$, the transmitter encodes the current codeword at rate $R$ with the power allocation vector $\lambda^st(\alpha)$. If the transmission rate is higher than what the channel can support, $R > C_K(R, \lambda^st(\alpha))$, then an outage is declared.

### 5.2 Long-term average and peak power constraints

Under the short-term average and peak power constraints, the minimum outage probability

$$
\min_{\gamma} \left\{ P_{out}(R, \gamma, K) : \gamma \in O_K^l(P_{av}, P_p) \right\}
$$

(5.9)

is achieved by an optimal outage minimizing power allocation strategy $\lambda^st(\alpha) \in O_K^l(P_{av}, P_p)$.

**Theorem 5.2.1.** The power allocation policy that minimizes outage under the long-
term average and peak power constraints is

\[
\tilde{\gamma}(\alpha) = \begin{cases} 
\tilde{\gamma}_1(\alpha), & \text{w/ prob 1} \quad \text{if } \alpha \notin \mathcal{G}(P_p) \text{and } (\tilde{\gamma}(\alpha)) < s^* \\
\tilde{\gamma}_2(\alpha), & \text{w/ prob } w^* \quad \text{if } \alpha \notin \mathcal{G}(P_p) \text{and } (\tilde{\gamma}(\alpha)) = s^* \\
0, & \text{w/ prob } (1 - w^*) \quad \text{if } \alpha \notin \mathcal{G}(P_p) \text{and } (\tilde{\gamma}(\alpha)) = s^* \\
0, & \text{w/ prob 1} \quad \text{if } \alpha \in \mathcal{G}(P_p) \text{ and } (\tilde{\gamma}(\alpha)) > s^* \\
0, & \text{w/ prob 1} \quad \text{if } \alpha \in \mathcal{G}(P_p) \text{ and } (\tilde{\gamma}(\alpha)) \leq s^* 
\end{cases}
\] (5.10)

for some subset of fading states \( \mathcal{G}(P_p) \subset \mathbb{R}_+^K \), \( s^* > 0 \) and \( w^* \in [0, 1] \) with

\[
\tilde{\gamma}_k(\alpha) = \min \left( \max \left( \tilde{\lambda}_k(\alpha) - \frac{1}{\alpha_k}, 0 \right), P_p \right)
\] (5.11)

and \( \tilde{\lambda}_k(\alpha) \) the solution to

\[
\frac{1}{K} \sum_{k=0}^{K-1} \log \left[ 1 + \alpha_k \min \left( \max \left( \tilde{\lambda}_k(\alpha) - \frac{1}{\alpha_k}, 0 \right), P_p \right) \right] = R.
\] (5.12)

Proof. Suppose \( \gamma^* \) is the outage minimizing power control policy,

\[
\gamma^* = \arg \min_{\gamma} \{ P_{\text{out}}(R, \gamma, K) : \gamma \in \mathcal{O}_K^k(P_{av}, P_p) \}.
\] (5.13)

For this minimum outage power allocation policy, the outage region

\[
\phi(R, K) = \{ \alpha : C_K(\alpha, \gamma^*(\alpha)) < R \}
\] (5.14)

is the set of channels that cannot support rate \( R \). Let \( \tilde{\gamma} \) represent the power allocation strategy that prevents outage with minimum power. That is,

\[
\tilde{\gamma}(\alpha) = \arg \min_{\gamma} \{ (\gamma(\alpha)) : C_K(\alpha, \gamma(\alpha)) \geq R \}.
\] (5.15)
Then by definition
\[ \int_{\phi(R,K)} \mathcal{I}_F[C_K(R, \gamma(\alpha)) < R]dF(\alpha) \geq \int_{\phi(R,K)} \mathcal{I}_F[C_K(R, \gamma^*(\alpha)) < R]dF(\alpha) \] (5.16)
and
\[ \mathbb{E}_{\alpha \notin \phi(R,K)}[\gamma(\alpha)] \leq \mathbb{E}_{\alpha \notin \phi(R,K)}[\gamma^*(\alpha)]. \] (5.17)
Since \( \gamma^* \) is the optimal solution to (5.13) then the inequalities in (5.16) and (5.17) become equalities and hence \( \gamma \) is also an optimal solution.

Since we are minimizing outage with respect to both a peak and average power constraint, there will be a subset of the outage region for which outage cannot be prevented even if the peak power is used in each of the \( K \) blocks in the codeword. Denote this region by
\[ \mathcal{G}(P_p) = \{ \alpha : \frac{1}{K} \sum_{k=0}^{K-1} \log(1 + \alpha_k P_p) < R \} \] (5.18)
and note that it is a subset of the outage region, \( \mathcal{G}(P_p) \subset \phi(R,K) \). Using this definition, we follow the notation of [24] and define two sets of fading states
\[ \mathcal{R}(s) = \{ \alpha \notin \mathcal{G}(P_p) : \langle \gamma(\alpha) \rangle < s \} \] (5.19)
and
\[ \overline{\mathcal{R}}(s) = \{ \alpha \notin \mathcal{G}(P_p) : \langle \gamma(\alpha) \rangle \leq s \} \] (5.20)
that are differentiated by the average power allocated for each fading state using power allocation policy \( \gamma \). The corresponding average power over these sets are
\[ P(s) = \int_{\mathcal{R}(s)} \langle \gamma(\alpha) \rangle dF(\alpha) \] (5.21)
and

$$\bar{P}(s) = \int_{\mathcal{R}(s)} \langle \bar{\gamma}(\alpha) \rangle dF(\alpha). \quad (5.22)$$

Then by Lemma 3 in [24] the optimal power allocation policy under the peak and long-term average power constraints for all $\alpha \notin \mathcal{G}(P_p)$ is

$$\bar{\gamma}^{th}(\alpha) = \begin{cases} 
\bar{\gamma}(\alpha), \text{ w/ prob 1} & \text{if } \langle \bar{\gamma}(\alpha) \rangle < s^* \\
\bar{\gamma}(\alpha), \text{ w/ prob } w^* & \text{if } \langle \bar{\gamma}(\alpha) \rangle = s^* \\
0, \text{ w/ prob } (1 - w^*) & \text{if } \langle \bar{\gamma}(\alpha) \rangle = s^* \\
0, \text{ w/ prob 1} & \text{if } \langle \bar{\gamma}(\alpha) \rangle > s^* 
\end{cases} \quad (5.23)$$

where

$$s^* = \sup \{ s : \mathcal{P}(s) < \mathcal{P}_{av} \} \quad (5.24)$$

and

$$w^* = \frac{\mathcal{P}_{av} - \mathcal{P}(s^*)}{\bar{P}(s^*) - \mathcal{P}(s^*)}. \quad (5.25)$$

The form of $\bar{\gamma}$, the power allocation policy that prevents outage with minimum power, can be determined by solving

$$\min_{\gamma} \left\{ \frac{1}{K} \sum_{k=0}^{K-1} \gamma_k : \frac{1}{K} \sum_{k=0}^{K-1} \log(1 + \alpha_k \gamma_k) = R, 0 \leq \gamma_k \leq \mathcal{P}_p \right\}. \quad (5.26)$$

We begin by setting up the functional

$$L = \frac{1}{K} \sum_{k=0}^{K-1} \gamma_k - \sum_{k=0}^{K-1} \psi_k \gamma_k + \sum_{k=0}^{K-1} \mu_k (\gamma_k - \mathcal{P}_p) + \nu \left( \frac{\sum_{k=0}^{K-1} \log(1 + \alpha_k \gamma_k)}{K} - R \right) \quad (5.27)$$

and realizing that we have a convex objective function and convex set of feasible points. This implies that the globally optimal power allocation strategy $\bar{\gamma}$, and the
associated $\tilde{\psi}_k, \tilde{\mu}_k$ and $\tilde{\nu}$, satisfy the KKT conditions [28]

$$\tilde{\gamma}_k \geq 0$$  \hspace{1cm} (5.28a)

$$\tilde{\gamma}_k \leq P_p$$  \hspace{1cm} (5.28b)

$$\frac{1}{K} \sum_{k=0}^{K-1} \log(1 + \alpha_k \tilde{\gamma}_k) = R$$  \hspace{1cm} (5.28c)

$$\tilde{\psi}_k \geq 0$$  \hspace{1cm} (5.28d)

$$\tilde{\mu}_k \geq 0$$  \hspace{1cm} (5.28e)

$$\tilde{\psi}_k \tilde{\gamma}_k = 0$$  \hspace{1cm} (5.28f)

$$\tilde{\mu}_k (\tilde{\gamma}_k - P_p) = 0$$  \hspace{1cm} (5.28g)

$$\frac{\partial L}{\partial \tilde{\gamma}_k} = \frac{1}{K} - \tilde{\psi}_k + \tilde{\mu}_k + \frac{\tilde{\nu}}{K} \frac{\alpha_k}{1 + \alpha_k \tilde{\gamma}_k} = 0$$  \hspace{1cm} (5.28h)

with (5.28a-5.28h) corresponding to (5.6a-5.6h). A solution of the form

$$\tilde{\gamma}_k(\alpha) = \min \left( \max \left( \tilde{\lambda}^u(\alpha) - \frac{1}{\alpha_k}, 0 \right), P_p \right)$$  \hspace{1cm} (5.29)

with $\tilde{\lambda}^u(\alpha) = -\tilde{\nu}$ as the solution to

$$\frac{1}{K} \sum_{k=0}^{K-1} \log \left[ 1 + \alpha_k \min \left( \max \left( \tilde{\lambda}^u(\alpha) - \frac{1}{\alpha_k}, 0 \right), P_p \right) \right] = R,$$  \hspace{1cm} (5.30)

satisfies (5.28b-5.28h) and is therefore the power allocation policy that prevents outage with minimum power. Therefore $\tilde{\gamma}_k$ is the power allocation strategy that minimizes outage probability under a peak and long-term average power constraint.

As with the short-term case the optimal solution has three regions. We use a constant power allocation of $\tilde{\gamma}_k^u(\alpha) = P_p$ when $\lambda^u(\alpha) - \frac{1}{\alpha_k} \geq P_p$. The waterfilling
solution \( \lambda_0^k(\alpha) - \frac{1}{\alpha_k} \geq \mathcal{P}_p \tilde{\gamma}_k^0(\alpha) = \lambda_0^u(\alpha) - \frac{1}{\alpha_k} \) is applied when \( 0 < \lambda_0^u(\alpha) - \frac{1}{\alpha_k} < \mathcal{P}_p \).

Finally, no power is allocated, \( \tilde{\gamma}_k^0(\alpha) = 0 \), when \( 0 > \lambda_0^u(\alpha) - \frac{1}{\alpha_k} \).

The implementation of this transmission scheme is relatively simple. The receiver relays the condition of the channel \( \alpha \) back to the transmitter. For the desired transmission rate \( R \) if \( \alpha \in \mathcal{G}(\mathcal{P}_p) \) then outage is immediately declared. If \( \alpha \notin \mathcal{G}(\mathcal{P}_p) \), then the transmitter encodes the codeword at rate \( R \) with power allocation \( \tilde{\gamma}(\alpha) \) and if \( \langle \tilde{\gamma}(\alpha) \rangle > s^* \) an outage is again declared. In the cases where \( \langle \tilde{\gamma}(\alpha) \rangle < s^* \) and \( \langle \tilde{\gamma}(\alpha) \rangle = s^* \), the codeword is transmitted with probability 1 and \( w^* \), respectively.

5.3 Simulation Results

Shown in Figure 5.1 is the power allocated for a particular channel \( \alpha \) under a peak power constraint of 12dB and short-term and long-term average power constraints of 10dB for a transmission rate of 2 nats/sec/Hz. For this \( \alpha \) both situations require the power allocated in several of the blocks to reach the peak power. However, under the short-term power constraint, the average power in the codeword must not exceed \( \mathcal{P}_{av} \) and therefore several of the blocks in the codeword are allocated no power. As a result an outage is unavoidable for this \( \alpha \). Under the long-term power constraint, since the average power in the codeword can exceed \( \mathcal{P}_{av} \), enough power is allocated to prevent outage.

For a fixed transmission rate the power allocated under the long-term average power constraint tends to have a much larger variance than that allocated under...
Figure 5.1: We illustrate the power allocate for a particular channel under the short-term and long-term power allocation strategies for $R = 2$ nats/sec/Hz. Here the average and peak powers are 10dB and 12dB, respectively. Exceeding the average power constraint for this particular codeword allows the power allocated under the long-term strategy to prevent outage.
Figure 5.2: For $R = 2$ nats/sec/Hz, a coding delay of $K = 5$, long-term average power constraint $P_{av} = 10$dB, and the transmission of 10000 codewords, we show histograms of the transmitted power. The top figure is without a peak power constraint and the bottom with a peak power constraint of 12dB. The peak power constraint significantly changes the distribution the transmitted power; without a constraint on the peak power, the maximum power allocated for these 10000 codewords is 22.6dB.
the short-term average power constraint. This is due to the fact that the average power for any particular codeword under the long-term average power constraint can exceed $P_{av}$, while it cannot under the short-term average power constraint. The larger variance results in a larger portion of the transmitted signal exceeding the peak power constraint were it present. Therefore, when a peak power constraint is additionally imposed, the transmitted signal under the long-term average power constraint is affected to a greater degree. This can be seen in Figure 5.2 and Figure 5.3, which illustrates histograms of the allocated power for 10,000 transmitted codewords both with and without a peak power constraint. Clearly, when the additional peak power constraint is imposed the power allocation distribution changes significantly under the long-term average power constraint.

Imposing a peak power constraint limits the ability of a communications system to ensure reliable communication. That is, when a peak power constraint is imposed in addition to an average power constraint, the outage probability will be higher than if only an average power constraint is present. Figure 5.4 plots the outage probability vs. $P_{av}$ for a fixed transmission rate. We see that with only an average power constraint a far lower outage probability is achievable for the same average power under the long-term constraint than under the short-term constraint. However, the performance difference shrinks greatly when a peak power constraint is also imposed. For a fixed PAR, under both the short-term and long-term average power scenarios the outage probability is much higher with the peak power constraint than without. However, we
Figure 5.3: For $R = 2 \text{ nats/sec/Hz}$, a coding delay of $K = 5$, short-term average power constraint $P_{av} = 10\text{dB}$, and for the transmission of 10000 codewords, we show histograms of the transmitted power. The top figure is without a peak power constraint and the bottom with a peak power constraint of $12\text{dB}$. We notice that the peak power constraint does not change the distribution of the transmitted power as much as under the long-term power constraint. This is due to the fact that the with the short-term power constraint the distribution of the power is not as 'peaky' as under the long-term constraint. For example without a peak power constraint, the maximum power allocated for any block is $14.8\text{dB}$ as compared to the $22.6\text{dB}$ for the long-term power constraint.
Figure 5.4: Minimum outage probability vs. $P_{av}$ for a transmission rate of $R = 1$ nats/sec/Hz. We illustrate situations when there is no peak power constraint, when $P_p = \infty$, the peak to average power ratio (PAR) is fixed at 2, $\frac{P_p}{P_{av}} = 2$, and when the peak power is fixed at 7 dB, $P_p = 7$dB. The peak power constraint results in a drop in outage performance for both average power scenarios. However, the performance degradation is most significantly under the long-term average power constraints.

see that the short-term average power constraint is affected to a lesser degree. This occurs because the variance is higher under the long-term power scenario making it more susceptible to the peak power constraint. For a fixed $P_p$, the outage probability curve plateaus as $P_{av}$ increases. The closer $P_{av}$ is to $P_p$, the larger the performance degradation. We see that the long-term average power scenario plateau’s for a smaller $P_{av}$ than the short-term average power scenario, illustrating its sensitivity to the peak power constraint.

Figure 4.4 plots outage probability as a function of $R$ for a fixed $P_{av}$. Here we
Figure 5.5: Minimum outage probability vs. $R$ for a long-term power constraint of $P_{av} = 10$dB. We illustrate the situations where $P_p \in \{10, 13, 16, \infty\}$dB. With the peak power constraint the outage probability is larger than without a peak power constraint for large $R$ and/or for small $P_p$.

again see that when an additional peak power constraint is imposed, the outage probability increases. For large values of $R$ and/or small values of $P_p$ the outage probability is higher than without a peak power constraint. This is most clearly seen for the long-term power scenario with a peak power of $P_p = 16$dB. For $R < 3.5$ nats/sec/Hz the outage probability is nearly the same as that achieved without a peak constraint, since $R$ is relatively small and the power required for any channel state is rarely limited by the peak power constraint. For $R > 3.5$ nats/sec/Hz, the outage probability is higher than that achieved without a peak constraint, since the power required for any channel state is often limited by the peak power constraint.

Figure 5.6 plots results analogous to Figure 5.5 except under the short-term aver-
Figure 5.6: Minimum outage probability vs. $R$ for a short-term power constraint of $P_{av} = 10$dB. We illustrate the situations where $P_{p} \in \{10, 12, \infty\}$dB. The outage probability with a peak power constraint is higher than without. However, the difference is not as large as with the long-term power scenario as the transmitted signal is less affected by the peak power constraint.

...
Chapter 6

Throughput maximization with optimal rate selection and power control

We now consider the scenario in which both the transmitter and receiver have CSI. When this occurs the transmitter knows prior to transmission if an outage will occur. We propose scheme DT that simply delays transmission until the channel condition allows successful decoding at the receiver. Also since the transmitter knows the condition of the channel it can vary the transmit power accordingly. We now maximize the average throughput by optimally selecting the transmission rate and power control strategy.

For scheme DT the outage probabilities are independent from one transmission attempt to the next, due to the fact that the channel states are assumed i.i.d. in the BF-AWGN model. As such the service time distribution, the probability that it will take $s$ attempts for successful transmission, is

$$\text{Prob}(S = s) = [P_{\text{out}}(R, \gamma, K)]^{s-1}[1 - P_{\text{out}}(R, \gamma, K)]$$

(6.1)

for transmission rate $R$, coding delay $K$ and power allocation policy $\gamma$. This implies that the service time distribution is geometric on the positive integers with parameter $[1 - P_{\text{out}}(R, \gamma, K)]$. Then

$$\mathbb{E}[S] = \frac{1}{1 - P_{\text{out}}(R, \gamma, K)}$$

(6.2)
is the expected service time.

Using the form of the expected service time and the fact that throughput is the transmission rate over the expected service time, we define

\[
MZT_{DT}(P_{av}, K, P_p) = \sup_{R} \sup_{\gamma} \{ R [1 - P_{out}(R, \gamma, K)] : \gamma \in O_K \} \tag{6.3}
\]
as the *maximum zero-outage throughput with scheme DT* for a system with coding delay \(K\), average transmit power \(P_{av}\) and peak transmit power \(P_p\). We denote \(MZT_{DT}(P_{av}, K)\) as the maximum throughput without a peak power constraint or when \(P_p = \infty\).

\(MZT_{DT}\) is found by minimizing the outage probability for a given transmission rate and then taking the supremum over all transmission rates. Here the power allocation policy \(\gamma\) belongs to \(O_K\) which can represent any one of \(O_K^{\text{ut}}(P_{av})\), \(O_K^{\text{ut}}(P_{av})\), \(O_K^{\text{ut}}(P_{av}, P_p)\) or \(O_K^{\text{ut}}(P_{av}, P_p)\). For any transmission rate \(R\) there is an associated minimum outage probability \(\epsilon\) that is achieved by using the appropriate outage minimizing power allocation strategy. Then, \(MZT_{DT}\) can be thought of as selecting the throughput maximizing \((R, \epsilon)\) pair. For each power constraint, codewords are encoded using the optimal transmission rate that is the maximizer of (6.3) and power is allocated using the appropriate outage minimizing power allocation strategy. If the transmission rate is larger than the instantaneous capacity, then an outage is declared and the transmission of the codeword is delayed.

Communications performance in fading channels has been quantified historically.
by $\epsilon$-capacity (2.26). Typically the target outage probability is fixed to a small value such as $\epsilon = 0.01$. In practice it may be better from a throughput perspective not to fix the target outage probability. This is illustrated in Theorem 6.0.1.

**Theorem 6.0.1.** MZT$_{DT}$ is always greater than or equal to the throughput achieved by transmitting at $\epsilon$-capacity.

**Proof.** For a fixed outage probability $\epsilon$, the $\epsilon$-capacity

$$C^\text{pc}_\epsilon := \sup_R \sup_{\gamma} \{R : P_{\text{out}}(R, \gamma, K) \leq \epsilon, \gamma \in \mathcal{O}_K \}$$

(6.4)

is found by optimally selecting $R$ and $\gamma$ with $\mathcal{O}_K \in \{\mathcal{O}_K^t(\mathcal{P}_\text{av}), \mathcal{O}_K^r(\mathcal{P}_\text{av}), \mathcal{O}_K^s(\mathcal{P}_\text{av}, \mathcal{P}_p), \mathcal{O}_K^t(\mathcal{P}_\text{av}, \mathcal{P}_p)\}$. For the outage minimizing power allocation strategy, every transmission rate $R = C^\text{pc}_\epsilon$ corresponds to a minimum outage probability $\epsilon$. Conversely, this means every outage probability $\epsilon$ corresponds to a throughput maximizing transmission rate $R = C^\text{pc}_\epsilon$. Transmitting at $R = C^\text{pc}_\epsilon$ results in a throughput

$$T_\epsilon = C^\text{pc}_\epsilon(1 - \epsilon).$$

(6.5)

Therefore (6.5) is a single point on the curve

$$T_{DT}(R) = R[1 - P_{\text{out}}(R, \gamma^*, K)],$$

(6.6)

with $P_{\text{out}}(R, \gamma^*, K)$ the minimum outage probability that achievable for transmission rate $R$ and coding delay $K$. Since

$$\text{MZT}_{DT} = \sup_R \{T_{DT}(R)\}$$

(6.7)
we have

\[ \text{MZT}_{DT} \geq T_e, \quad (6.8) \]

completing the proof. \( \square \)

**Corollary 6.0.2.** \( \text{MZT}_{DT} \) is always greater than or equal to the throughput achieved by transmitting at delay-limited capacity.

*Proof.* This is trivially shown by setting \( \epsilon = 0 \) and applying Theorem 6.0.1. \( \square \)

Corollary 6.0.2 illustrates the power of the multi-attempt approach for delay-limited systems. For the same coding delay \( K \) a higher throughput is achieved by allowing multiple, rather than a single, transmission attempts per codeword. That is, \( \text{MZT}_{DT} \) is larger delay-limited capacity \( (T_e |_{\epsilon=0}) \). The cost of the improved throughput is a queueing delay that is not present if the system is restricted to a single transmission attempt per codeword. For a detailed analysis of the queueing delay under the multi-attempt approach we refer the reader to Chapter 7.

### 6.1 Maximum zero-outage throughput with scheme DT \( (\text{MZT}_{DT}) \) under different power constraints

We now examine \( \text{MZT}_{DT} \) under the short-term average and long-term average power constraints both with and without an additional peak power constraint. For each power constraint we substitute either \( O^t_K (P_{av}) \), \( O^t_K (P_{av}, P_p) \), \( O^t_K (P_{av}, P_p) \) or \( O^t_K (P_{av}, P_p) \)
for $\mathcal{O}_K(\mathcal{P}_{av}, \mathcal{P}_p)$ in (6.3). Then using the form of the outage minimizing power allocation strategy, (6.3) can be reduced to an optimization problem of only a single variable, the transmission rate $R$.

**Theorem 6.1.1.** The maximum zero-outage throughput with the delayed transmission scheme under the short-term average power constraint is

$$MZT_{DT}(\mathcal{P}_{av}, K) = \sup_R \left( RE_{\alpha} \left[ \mathcal{I}_{\mathcal{F}} \left[ R \leq \frac{1}{K} \log \left( 1 + \alpha_k \max \left\{ \lambda^{st}(\alpha) - \frac{1}{\alpha_k}, 0 \right\} \right) \right] \right] \right)$$

(6.9)

with $\lambda^{st}(\alpha)$ as the solution to

$$\sum_{k=0}^{K-1} \max \left( \lambda^{st}(\alpha) - \frac{1}{\alpha_k}, 0 \right) = K\mathcal{P}_{av}. \tag{6.10}$$

**Proof.** See Appendix A. \hfill \square

**Theorem 6.1.2.** The maximum zero-outage throughput with the delayed transmission scheme under both the short-term average and peak power constraints is

$$MZT_{DT}(\mathcal{P}_{av}, K, \mathcal{P}_p) = \begin{cases} \sup_R \left( RE_{\alpha} \left[ \mathcal{I}_{\mathcal{F}} \left[ R \leq \frac{\log(1+\alpha_k)}{K} \right] \right] \right) & \text{if } \mathcal{P}_p > \mathcal{P}_{av} \\ \sup_R \left( RE_{\alpha} \left[ \mathcal{I}_{\mathcal{F}} \left[ R \leq \frac{\log(1+\alpha_k \mathcal{P}_p)}{K} \right] \right] \right) & \text{if } \mathcal{P}_p \leq \mathcal{P}_{av} \end{cases} \tag{6.11}$$

with $\xi = \min \left\{ \max \left[ \lambda^{st}(\alpha) - \frac{1}{\alpha_k}, 0 \right], \mathcal{P}_p \right\}$ and $\tilde{\lambda}^{st}(\alpha)$ as the solution to

$$\sum_{k=0}^{K-1} \min \left( \max \left( \tilde{\lambda}^{st}(\alpha) - \frac{1}{\alpha_k}, 0 \right), \mathcal{P}_p \right) = K\mathcal{P}_{av}. \tag{6.12}$$
Proof. See Appendix A.

\[ \square \]

**Theorem 6.1.3.** The maximum zero-outage throughput with the delayed transmission scheme under the long-term average power constraint is

\[
M_{ZT}^{\text{DT}}(P_{av}, K) = \sup_{R} \left\{ R \left[ \mathbb{E}_{\alpha} \left\{ I_F \left( \frac{1}{K} \sum_{k=0}^{K-1} \max \left( \lambda^u(\alpha) - \frac{1}{\alpha_k}, 0 \right) < s^* \right) \right\} \right] \right. \\
+ w^* \mathbb{E}_{\alpha} \left\{ I_F \left( \frac{1}{K} \sum_{k=0}^{K-1} \max \left( \lambda^u(\alpha) - \frac{1}{\alpha_k}, 0 \right) = s^* \right) \right\} \right\} \tag{6.13}
\]

with \( \lambda^u(\alpha) \) as the solution to

\[
\frac{1}{K} \sum_{k=0}^{K-1} \log \left( 1 + \alpha_k \max \left( \lambda^u(\alpha) - \frac{1}{\alpha_k}, 0 \right) \right) = R, \tag{6.14}
\]

\[
R
\]

Proof. See Appendix A.

\[ \square \]

**Theorem 6.1.4.** The maximum zero-outage throughput with the delayed transmission scheme under both the long-term average and peak power constraints is

\[
M_{ZT}^{\text{DT}}(P_{av}, K, P_p) = \sup_{R} \left\{ R \left[ \mathbb{E}_{\alpha} \left\{ I_F \left( \alpha \notin G_p \right) \right\} \right] \right. \\
+ w^* \mathbb{E}_{\alpha} \left\{ I_F \left( \kappa = s^* \right) \right\} \left. \right\} \tag{6.15}
\]

with \( \kappa = \frac{1}{K} \sum_{k=0}^{K-1} \min \left( \max \left( \lambda^u(\alpha) - \frac{1}{\alpha_k}, 0 \right), P_p \right) \) and \( \lambda^u(\alpha) \) as the solution to

\[
\frac{1}{K} \sum_{k=0}^{K-1} \log \left( 1 + \alpha_k \min \left( \max \left( \lambda^u(\alpha) - \frac{1}{\alpha_k}, 0 \right), P_p \right) \right) = R, \tag{6.16}
\]
Proof. See Appendix A.

For each power constraint, codewords are encoded using the optimal transmission rates that are the optimizers to (6.1.1), (6.1.2), (6.1.3) and (6.1.4), respectively. Using the appropriate outage minimizing power allocation strategy if the transmission rate is larger than the instantaneous capacity then an outage is declared and transmission of the codeword delayed until a more favorable channel state arises.

6.2 Special cases of MZT_{DT}

Since the form of the outage minimizing power allocation policies are complicated functions of the channel state $a$, the expression for MZT_{DT} are even more complex. However, for $K = 1$ and a $\chi^2$ fading process we have found more explicit expressions for three of the four power allocation scenarios.

For the short-term average power constraint we can find the optimal transmission rate and the maximum throughput.

**Theorem 6.2.1.** If $K = 1$ and the fading process $a$ follows a $\chi^2$ distribution, then

$$MZT_{DT}^{st}(P_{av}, 1) = W(P_{av})e^{-\frac{W(P_{av})-1}{P_{av}}}.$$  \hspace{1cm} (6.17)

**Proof.** Since $K = 1$ and the entire codeword spans a single block of the BF-AWGN channel, the outage minimizing power allocation is to use all the power $P_{av}$ within
the codeword. In this case the solution is the same as constant power allocation when
\( K = 1 \) [11].

If \( \alpha \) follows a \( \chi^2 \) distribution, then \( 1 - P_{out}(R, P_{av}, 1) = e^{-\left(\frac{R^{\gamma - 1}}{P_{av}}\right)} \). Using this, we let
\( T(R) = R e^{-\left(\frac{R^{\gamma - 1}}{P_{av}}\right)} \). Taking the derivative with respect to \( R \) and equating with zero, we see that transmission corresponding to the critical point is the solution to \( R e^R = P_{av} \). The solution to this is the optimal transmission rate \( R^* = W(P_{av}) \). Substituting this back into \( T(R) \) we arrive at (6.17). From Theorem 4.1.1 and Proposition 4.1.2 we know that this solution corresponds to a unique maximum. From Theorem 4.1.1 and Proposition 4.1.2 we know that this solution corresponds to a unique maximum. □

When a peak power constraint is imposed in addition to the short-term average power constraint we find a similar result.

**Theorem 6.2.2.** If \( K = 1 \) and the channel fading \( \alpha \) follows a \( \chi^2 \) distribution, then
\[
MZT_{DT}^{est}(P_{av}, 1, P_p) = W(\gamma) e^{-\left(\frac{W(\gamma) - 1}{\gamma}\right)}
\]
(6.18)
with
\[
\gamma = \min(P_{av}, P_p)
\]
(6.19)

**Proof.** If \( K = 1 \) and the entire codeword spans a single block of the fading channel and is affected by only a single channel fade. The situation is then the same as constant power allocation [11]. The instantaneous capacity is maximized by allocating the maximum allowable power to the codeword, which is
\[
\gamma = \min(P_{av}, P_p)
\]
(6.20)
Then, by the procedure of Theorem 6.2.1, the optimal transmission rate is $R^* = W(\gamma)$ and we arrive at (6.18).

Finally, in the case of a long-term average power constraint we can find sufficient conditions that the optimal transmission rate $R^t_l$ and optimal power cutoff $s^*_{R^t_l}$ satisfy.

**Theorem 6.2.3.** If $K = 1$ and the channel gains follow a $\chi^2_2$ distribution, then

$$e^{R_{lt}} E_i \left( 1, \frac{e^{R_{lt}} - 1}{s_{R^t_l}} \right) = \mathcal{P}_{av}, \quad (6.21)$$

$$(s_{R^t_l})^2 - \mathcal{P}_{av} R^t_l e^{R_{lt}} e^{\left( \frac{e^{R_{lt}} - 1}{s_{R^t_l}} \right)} = 0 \quad (6.22)$$

where $E_i(1, x) = \int_{1}^{\infty} \frac{e^{-at}}{t} \, dt$ are sufficient conditions that $R^t_l$ and $s^*_{R^t_l}$ satisfy.

**Proof.** Condition (6.21) is a sufficient condition for the optimal power cutoff $s^*_{R^t_l}$. It is obtained by finding the optimal short-term cutoff for the optimal transmission rate $R^t_l$. That is, finding the $s$ such that $\mathcal{P}_1(s) = \mathcal{P}_{av}$.

Condition (6.22) is also a sufficient condition that the optimal transmission rate $R^t_l$ and power cutoff $s^*_{R^t_l}$ satisfy. For transmission rate $R$ and cutoff $s_R$, we begin with

$$P_{out}(R, \mathcal{P}_{av}, 1) = 1 - e^{-\left( \frac{R - 1}{s_R} \right)}$$

which lets us define $T(R) = Re^{-\left( \frac{R - 1}{s_R} \right)}$. Taking the derivative $\frac{dT(R)}{dR}$ and setting to 0, we see that

$$s_R + R \left( \frac{e^R - 1}{s_R} \right) \frac{d(s_R)}{dR} - Re^R = 0. \quad (6.23)$$

By letting $g(s, R) = \mathcal{P}_1(s) - \mathcal{P}_{av}$ and performing implicit differentiation $\frac{ds_R}{dR} = -\frac{\frac{dg}{dR}}{\frac{dx}{ds_R}}$, we find that

$$\frac{ds_R}{dR} = -\frac{e^R[E_i(1, x) - xE_i(0, x)]}{x^2E_i(0, x)} \quad (6.24)$$
where $x = \frac{e^{r-1}}{s_R}$ and $E_i(0, x) = \int_1^\infty e^{-xt}dt$. Substituting this back into (6.23) and setting $s_R = s_{R^*}$, we arrive at (6.22). □

6.3 Examples and discussion

MZT\textsubscript{DT} quantifies the maximum throughput achievable with scheme DT. As is the case for constant power transmission, the benefit of allowing multiple (rather than a single) transmission attempts per codeword, with rate selection and power control, is an increased throughput for the same coding delay. In this section we illustrate this for the $\chi_2^2$ fading process.

6.3.1 Increased throughput with the multi-attempt approach

Within the single-attempt paradigm the need for a measure of zero-outage (error-free) communication performance for delay-limited systems led to the notion of delay-limited capacity, or $\epsilon$-capacity with $\epsilon = 0$. When only a single transmission attempt is allowed, the transmission rate $R$ must be supported on all possible $\alpha$. Thus, delay-limited capacity quantifies the error-free data rate that can be supported over all $\alpha$ in the support of the fading process.

With CSI-RT, delay-limited capacity is always 0 for $\chi_2^2$ fading when $K = 1$ [6]. However, when $K > 1$ non-zero delay-limited capacity is possible [24]. Figure 6.1 illustrates MZT\textsubscript{DT} and delay-limited capacity as a function of $P_{av}$ for $K = 2$, the smallest coding delay with non-zero delay-limited capacity. For the same coding de-
Figure 6.1: Delay-limited capacity and $\text{MZT}_{DT}$ for $K=2$ vs. SNR. Increased throughput is possible by allowing multiple transmission attempts per codeword rather than a single attempt. For example, for a throughput of 0.5 nats/sec/Hz, $\text{MZT}_{DT}$ provides a 2.4dB gain over delay-limited capacity.

lay, $\text{MZT}_{DT}$ is higher than delay-limited capacity for all $P_{av}$. The same phenomenon holds for all $K$ as is proven in Corollary 6.0.2 of Theorem 6.0.1. Clearly, the performance benefits of the multi-attempt approach over the single-attempt approach is an increased throughput for the same coding delay.

6.3.2 Importance of power control

The conventional view about optimal power control is that it yields "a negligible [ergodic] capacity gain" over constant power transmission [4]. This is quite evident when comparing $C_{\text{erg}}$ and $C_{\text{erg-pc}}$ as a function of $P_{av}$ in Figure 2.1. However, in Figure 6.2 we plot $\text{MZT}_{DT}$ as a function of $P_{av}$ with $K = 1$ for the constant, short-term
(equivalent to constant power allocation for $K = 1$) and long-term power allocation strategies. Comparing $MZT_{DT}^{\text{const}}$ and $MZT_{DT}^{\text{lt}}$, we see that the difference between the curves is large for all $P_{av}$; clearly power control is very important for delay-limited systems. Therefore the original statement about optimal power control should be qualified: *Power control provides negligible performance gains for delay-unconstrained systems, but for delay-limited systems the gains can be significant.*

The importance of power control is again shown in Figure 6.3 which plots $MZT_{DT}$ with $P_{av} = 10\text{dB}$. By observing $MZT_{DT}^{\text{lt}}$ as a function of $K$, we see that the throughput, under scheme DT, with optimal rate and power control converges very quickly to ergodic capacity. In fact, $MZT_{DT}^{\text{lt}} = 2.00 \text{ nats/sec/Hz}$ when $K = 10$, is just slightly lower than $C_{\text{erg-pc}} = 2.07 \text{ nats/sec/Hz}$, achievable only when $K = \infty$. Again, this illustrates that power control is more important than large coding delays for maximizing throughput. For example, a target throughput of $1 \text{ nat/sec/Hz}$ is achieved with $K = 1$ under the long-term average power constraint, but is not achievable even with $K = 100$ for constant power transmission. It is also worth noting that the more relaxed the power constraint the higher the throughput, i.e., $MZT_{DT}^{\text{const}} \leq MZT_{DT}^{\text{lt}} \leq MZT_{DT}^{\text{lt}}$. This relation holds for any coding delay $K$ since the constant power allocation is a special case of the short-term power allocation which in turn is a special case of the long-term power allocation strategy.

To reemphasize the importance of power control we again examine Figure 6.2. For $K = 1$ and a throughput of $1 \text{ nat/sec/Hz}$, $MZT_{DT}^{\text{lt}}$ is only about $0.5\text{dB}$ away from
Figure 6.2: $C_{\text{erg-const}}$, $C_{\text{erg-pc}}$, $\text{MZT}_{\text{DT}}^{\text{const}}$ and $\text{MZT}_{\text{DT}}^{\text{th}}$ for $K = 1$ as a function of SNR. We see that when $K = \infty$ power control provides negligible capacity gains, as the difference between $C_{\text{erg-const}}$ and $C_{\text{erg-pc}}$ shrinks as SNR increases. When $K < \infty$ power control clearly provides significant throughput gains as the difference between $\text{MZT}_{\text{DT}}^{\text{const}}$ and $\text{MZT}_{\text{DT}}^{\text{th}}$ remains large for all SNR.
ergodic capacity with power control, $C_{\text{erg}-\text{pc}}$ for which $K = \infty$. More surprisingly, for low SNR it is even greater than the ergodic capacity with constant power, $C_{\text{erg}-\text{const}}$. In this SNR region, we achieve a better average throughput, $\text{MZT}_{\text{DT}}$ for $K = 1$ with the delayed transmission scheme and power control than for $K = \infty$ with constant power allocation and the single-attempt approach, $C_{\text{erg}-\text{pc}}$. This implies that optimal power control is more important than the number of fading states affecting each codeword (ergodicity).
6.3.3 Importance of rate selection

As is the case for constant power transmission, our simulation results for variable power transmission show that the transmission rate must be selected carefully in order to maximize throughput. Selecting a suboptimal transmission rate can result in a throughput much smaller than $\text{MZT}_{\text{DT}}$. This can be seen in Figure 6.4 which plots the average throughput achieved with scheme DT as a function of transmission rate for the constant, short-term and long-term average power allocation strategies. The peak of each curve corresponds to $\text{MZT}_{\text{DT}}$. We also see that the larger the coding delay $K$ the larger the drop in throughput is if the optimal transmission rate is overshot. Therefore care must be taken to solve (6.3) and select the appropriate transmission rate for the power allocation policy at hand.

Figure 6.5 plots the optimal transmission rate, corresponding to $\text{MZT}_{\text{DT}}$, as a function of coding delay $K$. The optimal transmission rate, especially for small $K$, can fluctuate a great deal. In fact a very non-intuitive phenomenon is observed — in some cases, the optimal transmission rate can actually be higher than ergodic capacity. For example, when $K = 1$ the optimal transmission rate under the long-term average power constraint, $R^{\text{lt}} = 2.51 \, \text{nats/sec/Hz}$, is more than 21% higher than the ergodic capacity of the channel, $C_{\text{erg-pc}} = 2.07 \, \text{nats/sec/Hz}$. This is counter to common practice, where a transmission rate lower than capacity is normally used. We emphasize that this is not a violation of the ergodic capacity theorem [4], since
the resulting throughput is always less than ergodic capacity.

For a given power allocation policy, either the transmission rate or the outage probability, but not both, can be freely selected since they depend on one another. Figure 6.6 plots the outage probability associated with the optimal transmission rate. We see that the optimal outage probability can be quite high as was shown for constant power transmission. In fact, for $P_{av} = 10$dB the optimal outage probability when $K = 1$ is 0.37 and 0.27 for the short-term and long-term average power constraints. This is interesting because it is counter to conventional practice; in most communication literature $\epsilon$-capacity is normally measured for a small outage probability such as $\epsilon = 0.01$. However, we see that in order to maximize throughput the
outage probability should be much higher.

6.3.4 Effect of a peak power constraint

We previously saw in Section 5 that a peak power constraint can reduce the ability of a delay-limited communication system to prevent outage events, resulting in higher outage probabilities for the same transmission rate and average power. Clearly this will affect the $M_{ZT_{DT}}$ of the system as well. Here we illustrate the effect of the peak power constraint on the long-term average power scenario. Similar results, though to a lesser degree, can be observed for the short-term power scenario.

A peak power constraint limits the maximum throughput. Figure 6.7 show $M_{ZT_{DT}}$
Figure 6.6: Optimal outage probability vs. coding delay $K$ for $P_{av} = 10$. The optimal outage probability under the constant, short-term and long-term average power constraints are 0.4, 0.37 and 0.27, respectively.

when $K = 5$ as a function of $P_{av}$ both with and without a peak power constraint. Also plotted for reference is ergodic capacity (without a peak power constraint). $\text{MZT}_{DT}^h(P_{av}, K, P_p)$ is nearly identical to $\text{MZT}_{DT}^h(P_{av}, K)$ when $P_{av} \ll P_p$. However, for larger $P_{av}$ the outage probability, and therefore $\text{MZT}_{DT}^h(P_{av}, K, P_p)$, becomes limited by the peak power constraint. Further increasing the average power does not increase $\text{MZT}_{DT}^h(P_{av}, K, P_p)$ as the peak power constraint will not allow improvements in the minimum outage probability. Figure 6.7 also illustrates $\text{MZT}_{DT}^h$ for a fixed PAR. Here, $\text{MZT}_{DT}^h$ continues to increase with $P_{av}$, but the effect of the PAR restriction is obvious — $\text{MZT}_{DT}^h$ is less than that obtained without a peak power constraint. Figure 6.8 illustrates the analogous results for the short-term power con-
Figure 6.7: MZT\textsuperscript{lt} vs. $P_{av}$ with $K = 5$ and for both $P_p \in \{10, 13, 16, \infty\}$dB and for fixed PAR's of 1.5 and 2.0. Ergodic capacity is shown for reference. For a fixed $P_p$, MZT\textsubscript{DT}\textsuperscript{lt} becomes limited by the peak power constraint for $P_{av}$ near $P_p$. At a target spectral efficiency of 1 nat/sec/Hz MZT\textsubscript{DT}\textsuperscript{lt} is 0.9dB and 2.1dB away from ergodic capacity for PAR = 2.0 and PAR = 1.5, respectively.

The same effects are present but are not as pronounced due to the fact that the short-term average power constraint is less affected by an additional peak power constraint than the long-term average power constraint.

$\text{MZT}_{DT}^{lt}(P_{av}, K, P_p)$, is plotted against coding delay $K$ for $P_{av} = 10$dB and various values of $P_p$ in Figure 6.9(b). The smaller the $P_p$, the further $\text{MZT}_{DT}^{lt}(P_{av}, P_p, K)$ is from $\text{MZT}_{DT}^{lt}(P_{av}, K)$. An interesting phenomenon is observed as $K$ increases; we see that the peak power constraint affects the maximum throughput, and hence the outage probability, to a lesser degree. This is explained by the fact that the likelihood of a very poor channel $\alpha$ decreases for large $K$. Hence, the likelihood of
a power allocation vector which hits the peak power in several blocks also decreases and the effect of the peak power constraint diminishes. The same phenomenon can be seen with the short-term power constraint in Figure 6.10, though to a lesser degree.

Properly selecting the transmission rate remains important when a peak power constraint is imposed. Figure 6.11 plots the throughput against transmission rate with a long-term average power constraint for different values of $P_p$. It is critical to select the transmission rate that corresponds to $\text{MZT}^\text{LT}_{\text{DT}}(P_{av}, K, P_p)$, since a suboptimal selection can yield a large throughput drop. The effect of the peak power constraint is clearly seen on the throughput - small values of $P_p$ and/or for large values of $R$ the transmitted signal is peak limited. That is, the throughput is less that that if there is no peak power constraint. This same phenomenon is observed under the short-term
Figure 6.9: MZT_DT vs. $K$ for an long-term average power $P_{av} = 10$dB for peak power $P_p = \{10, 13, 16, \infty\}$dB. The effect of the peak power constraint diminishes for larger coding delays, $K$.

Figure 6.10: MZT_DT vs. $K$ under short-term average power $P_{av} = 10$dB and for peak power $P_p = \{10, 12, \infty\}$dB. The effect of the peak power constraint is not as pronounced as under the long-term power constraint.
Figure 6.11: Throughput vs. transmission rate under long-term and peak power constraints for $K = 5$, $P_{av} = 10\, \text{dB}$ and $P_{p} = \in (10, 13, 16, \infty)$ dB. The system is considered peak-limited for those $R$ and $P_{p}$ for which the throughput drops from the case when $P_{p} = \infty$. For $P_{p} = 10, 13\, \text{dB}$ the throughput is limited throughout the range of transmission rates. For $P_{p} = 16\, \text{dB}$ the throughput is peak-limited for $R > 3.5\, \text{nats/sec/Hz}$.

The power constraint in Figure 6.12 though not seen to the same degree as under the long-term power constraint.

Figure 6.13 and Figure 6.15 show the optimal transmission rate and the associated outage probability as a function of $K$ for various $P_{p}$ under the long-term average power constraint. We can make the same observations that were made without a peak power constraint: the optimal transmission rate can be higher than ergodic capacity and the optimal outage probability can be quite high. Both observations run counter to conventional practice. We notice that as $K$ grows that the difference in the optimal transmission rates, with and without a peak power constraint, decreases. Figure 6.14
Figure 6.12: Throughput vs. transmission rate under short-term and peak power constraints for $K = 5$, $P_{av} = 10\, \text{dB}$ and $P_p = \in\{10, 12, \infty\}\, \text{dB}$. The system is considered peak-limited for those $R$ and $P_p$ for which the throughput drops from the case when $P_p = \infty$.

and Figure 6.16 illustrate the analogous results under the short-term average power constraint.
Figure 6.13: The optimal transmission rate $R_{p}^{\text{st}}$ vs. $K$ under a long-term average power of $\mathcal{P}_{av} = 10$dB and peak powers $\mathcal{P}_{p} = \{10, 13, 16, \infty\}$dB. For larger values of $\mathcal{P}_{p}$ the optimal transmission rate can actually be higher than the erogdic capacity of the channel.

Figure 6.14: The optimal transmission rate $R_{p}^{\text{st}}$ vs. $K$ under a short-term average power of $\mathcal{P}_{av} = 10$dB and peak power $\mathcal{P}_{p} = \{10, 13, 16, \infty\}$dB. The difference in the transmission rates is not as pronounced as under the long-term power constraint.
Figure 6.15: The optimal outage probability vs. coding delay $K$ under the long-term average and peak power constraints with $P_{av} = 10$ dB. We can see that the optimal outage probability that maximizes the communications throughput can be quite high.

Figure 6.16: The optimal outage probability vs. coding delay $K$ under the short-term average and peak power constraints with $P_{av} = 10$ dB. We can see that the optimal outage probability that maximizes the communications throughput can be quite high.
Chapter 7

Throughput maximization with queuing delay constraints

The throughput maximization analysis of Chapters 4 and 6 measured communication performance under the multi-attempt paradigm. We maximized the average throughput for schemes RT, ID and DT. By allowing multiple transmission attempts per codeword, zero-outage communication is possible for finite coding delay $K$. This is not the case for the single-attempt approach, which often yields zero throughput for finite $K$. The improved throughput achieved with the multi-attempt approach does not occur without cost. A queueing delay, due to the random nature of the fading channel that is not present with the single attempt approach, occurs with the multi-attempt approach.

7.1 Mathematical formulation

In the following queueing analysis we assume a transmission system in which “time” is measured in multiples of the channel coherence time, or blocks of $N$ transmitted symbols in the BF-AWGN channel model. We assume that a codeword transmission attempt requires a single unit of time which corresponds to $K$ blocks if the coding delay is $K$. 


Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
The behavior of queuing systems depend on the nature of both the codeword arrival process and codeword service time. We denote \( a \) as the average number of arrivals per unit time and \( \frac{1}{a} \) as the average *interarrival time*, the amount of time between successive arrivals. Recall that the average service rate, \( \frac{1}{\mathbb{E}[S]} \), is the average number of codewords served per unit time. Using these quantities, the *queue utilization factor*

\[
\rho(a, R) := \frac{\text{average arrival rate}}{\text{average service rate}} = a\mathbb{E}[S(a, R)]
\]

(7.1)
is defined as the proportion of time that the transmitter is busy (re)transmitting codewords. Factoring the queue utilization yields

\[
T_{LT}(R, \mathcal{P}_w, K, a) = \rho(a, R) \frac{R}{\mathbb{E}[S]}
\]

(7.2)as the long-term average throughput for a particular transmission rate \( R \). The formulation is similar to (3.2) except for the scaling factor \( \rho(a, R) \) that accounts for the proportion of time the transmitter is busy. For example if the codeword arrival rate and transmission rate are such that the throughput is 2 nats/sec/Hz but \( \rho = 1/2 \), implying that the transmitter is busy only half of the time, then the long-term average throughput is 1 nat/sec/Hz.

We previously maximized the communications throughput without a constraint on the queueing delay. Implicit in the previous formulation for the maximum throughput is that \( \rho(a, R) = 1 \), that is the transmitter is always busy (re)transmitting data. This implies that average arrival rate of codewords into the queue is equal to the average
service rate of the codewords. This approach limits the coding delay to \( K \) blocks and provides the maximum throughput for a particular retransmission scheme without a constraint on the queueing delay.

In many applications, such as video or voice, excessive queueing delays cannot be tolerated. For these systems, operating at the maximum throughput \( T_{\text{max}}(P_{av}, K) \) (3.2) is not feasible as it would lead to excessive delay since \( \rho(a, R) = 1 \) [40]. For such applications the arrival rate and coding rate must be adjusted to ensure that the queueing delay is not excessive. The \textit{expected waiting-time}, or delay, is the amount of time that a codeword spends in the system (either in the queue or under service) [32, 33]. By constraining the waiting time we can constrain the average delay codewords experience. We illustrate for constant power transmission and scheme RT since it is the most analytically tractable; however, similar results can be derived for other multi-attempt schemes both with and without power control. Accounting for queuing delay, we can define

\[
MZT^{D}_{RT}(P_{av}, K) = \sup_{a, R} \{ \rho(a, R) \frac{R}{\mathbb{E}[S]} : \mathbb{E}[W(a, R)] \leq D \}
\]

\[
= \sup_{a, R} \{ \rho(a, R)R[1 - P_{\text{out}}(R, P_{av}, K)] : \mathbb{E}[W(a, R)] \leq D \}. \tag{7.3}
\]

as the maximum zero-outage throughput with average waiting time no greater than \( D \).

Reducing either \( a \) or \( R \) reduces \( \rho(a, R) \) and therefore the amount of information that the source transmits across the fading channel. Reducing \( a \), reduces the frequency of transmission of codewords containing \( RKN \) nats. On the other hand, reducing \( R \)
reduces the amount of information transmitted per codeword and since the arrival rate is fixed at \( a \), and ultimately the amount of information that the source transmits. By adjusting the queue utilization, a function of \( a \) and \( R \), the throughput can be maximized with respect to a waiting-time constraint.

### 7.2 Simulation Results

As previously mentioned the queuing behavior is dependent not only on the retransmission scheme, but on the nature of the arrival process. There has been much interest on the modelling of traffic patterns generated by a variety of different sources [34, 35, 36, 37, 38, 39]. In this paper we consider several representative models. To model non-bursty sources we consider the constant, Bernoulli and Poisson arrival processes. For the constant source, the interarrival period between codewords is deterministic; the time between codeword arrivals is always \( \frac{1}{a} \). For the Bernoulli source, the interarrival period is a random variable that is geometrically distributed on the positive integers with parameter \( a \) and mean \( \frac{1}{a} \). That is

\[
\text{Prob}(A = m) = a(1 - a)^{m-1}
\]  

(7.4)

is the probability that the interarrival period, \( A \), between successive codewords is \( m \).

For the Poisson arrival process with mean intensity \( a \), the interarrival period is an exponential random variable,

\[
\text{Prob}(A > \tau) = e^{-a\tau}.
\]  

(7.5)
Specifically (7.5) represents the probability that the interarrival period is greater than time \( \tau \). Recent work \([34, 35, 38]\) has shown that traffic patterns often have a bursty nature. In this thesis, we use a 2-state Markov modulated Poisson process (MMPP) \([34, 37]\), with average intensity \( a \), to model bursty traffic sources. This model is attractive due to its relative simplicity and the fact that it can approximate the correlated nature of many real-world traffic sources. In this model, a state variable toggles between a high traffic and low traffic state. In each state, codewords are generated according to a Poisson process with high and low intensities, respectively.

Let \( a_{\text{low}} \), \( a_{\text{high}} \) represent the arrival intensities in the low and high states, respectively, and \( p_{\text{low}} \) and \( p_{\text{high}} \) the probability of being in the low and high states, respectively. Since \( a_{\text{low}}p_{\text{low}} + a_{\text{high}}p_{\text{high}} = a \) and \( p_{\text{low}} + p_{\text{high}} = 1 \), a traffic source is more bursty for larger \( p_{\text{low}} \). That is, if the source is in the low traffic state for the majority of time, the bursts in the high traffic state must have very high intensity for the overall average intensity to be \( a \). For the purposes of our simulations we assume that the probability of being in the low and high intensity states are \( p_{\text{low}} = \in \{0.85, 0.95\} \) and \( p_{\text{high}} = \in \{0.15, 0.05\} \), respectively.

Figure 7.1 plots the optimal MZT\(_{\text{RT}}\)\(_{P}^{D}(P_{\text{av}}, K)\) as a function of the maximum average waiting time \( D \), for \( K = 1 \) and \( P_{\text{av}} = 10\)dB. For each value of \( R \) and \( a \) the expected waiting time was simulated and the throughput calculated. Then for a particular waiting time constraint \( D \), MZT\(_{\text{RT}}\)\(_{P}^{D}(P_{\text{av}}, K)\) was found by exhaustive search over the variables \( R \) and \( a \). Figure 7.2 also plots MZT\(_{\text{RT}}\)\(_{P}^{D}(P_{\text{av}}, K)\), obtained by the above
Figure 7.1: The maximum throughput $\text{MZT}_{\text{RT}}^K(P_{av}, K)$ is plotted as a function of $D$ for $K = 1$ and $P_{av} = 10$. The throughput for a finite coding delay of $K = 1$ and finite waiting-time $D$ can be predicted. MMPP1 represents the more bursty MMPP process with $p_{ow} = 0.95$, while MMPP2 represents the scenario with $p_{ow} = 0.85$.  

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
procedure, and the throughput corresponding to the upper and lower 95% confidence intervals of the expected waiting time. That is, for the upper and lower bounds, the confidence intervals of the mean waiting-time are used in the optimization procedure, rather than the mean waiting-time itself. The lack of symmetry occurs since the optimization is performed by exhaustive search and we have finite resolution of the variables $a$ and $R$. We see that the upper and lower bounds are rather tight, telling us that our simulation results are accurate.

From Figure 7.1, we see that $MZT_{RT}^D(P_{av}, K)$ is smaller than $MZT_{RT}(P_{av}, K)$. Restricting queue utilization allows the queueing delay to be constrained at the cost of reduced throughput. For each of the arrival processes we see that $MZT_{RT}^D(P_{av}, K)$ approaches $MZT_{RT}(K, P_{av})$ as the constraint on the waiting time is relaxed, i.e. $D \rightarrow \infty$. This figure is particularly useful is it allows us to predict the best case performance of a communication system using retransmission scheme RT with both a finite coding delay $K$ and a finite waiting-time $D$. It is also interesting to note that for small $D$, between 10 and 20 for the different arrival processes, $MZT_{RT}^D(P_{av}, K)$ approaches $MZT_{RT}(P_{av}, K)$. This tells us that the throughput with a relatively small delay constraint can approach the maximum throughput without any delay constraint. Additionally we see that for the bursty MMPP process that $MZT_{RT}^D(P_{av}, K)$ increases slower as a function of delay than the non-bursty sources. We also see that the convergence is the slowest for the more bursty MMPP source. This fits our intuition; the more bursty the source the more the average arrival rate needs to be constrained since
a burst of data can cause queues to build quickly. Therefore, in order to keep the average waiting-time below a threshold, $MZT^D_{RT}(P_{av}, K)$ must be limited to a greater degree than for less bursty sources.

For $K = 1$ and $P_{av} = 10$dB, the optimal coding rate $R_{MZT^D_{RT}}^*$ and codeword arrival rate $a_{MZT^D_{RT}}^*$ that maximize (7.3) are shown in Figures 7.3 and 7.4, respectively. For smaller $D$ we see that the optimal transmission rate is smaller than for larger $D$. Clearly, as the tolerable delay is decreased, the coding rate should be reduced. We also see that the optimal arrival rate for non-bursty sources is nearly constant regardless of the waiting-time constraint. This tells us that for non-bursty sources, in order to maximize the throughput while constraining the average waiting-time, the coding rate rather than the arrival rate should be reduced; the frequency of codeword arrivals should be left unchanged while the amount of information in each codeword should be decreased. This reduces the average waiting time since a smaller coding rate results in a smaller outage probability and a smaller $E[S]$. For bursty sources, reducing the coding rate alone is not sufficient to meet the waiting-time constraint. We see for smaller $D$, that the arrival rate must also be reduced. Again we see that both the coding rate and arrival rate must be reduced the most for MMPP scenario with the higher degree of burstiness.

For non-bursty sources, a reduction in the codeword arrival rate reduces the throughput to a greater extent than a reduction in the coding rate. This is non-intuitive as conventional flow-control algorithms, such as TCP [2], reduce the fre-
Figure 7.2: The maximum throughput $MZT_{RT}(P_{av}, K)$ is plotted as a function of $D$ for $K = 1$ and $P_{av} = 10$. We also plot the maximum throughput corresponding to the upper and lower 95% confidence intervals of the expected waiting time.
Figure 7.3: Optimal transmission rate $R_{MZTP}^*$ as function of the average waiting time $D$ for $K = 1$ and $P_{av} = 10$. For small $D$, $R_{MZTP}^*$ can be quite far from the asymptotic value, suggesting a reduction in the transmission rate to satisfy a waiting-time constraint. MMPP1 represents the less bursty MMPP process with $p_{low} = 0.85$, while MMPP2 represents the scenario with $p_{low} = 0.95$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
Figure 7.4: Optimal arrival rate $a_{\text{MZTP}}^*$ as a function of $D$ for $K = 1$ and $P_{av} = 10$. We see that for non-bursty sources, the optimal arrival rate remains relatively constraint. However, for the bursty MMPP source, the arrival rate drops for smaller $D$. Note that MMPP1 represents the more bursty MMPP process with $p_{low} = 0.95$, while MMPP2 represents the scenario with $p_{low} = 0.85$. 

Reproduced with permission of the copyright owner. Further reproduction prohibited without permission.
quency of packet generation when large queues build in communication networks [2].

The difference is reconciled by the fact that the underlying cause for the buildup of queues is different. The action that TCP takes is motivated by the assumption that queues build due to congestion in the network — that packets are being generated faster than the network can handle them. However, queues in fading channels grow due to the frequency of codeword generation and the fact that the medium itself is unreliable. For example, if the channel condition remains poor for 10 consecutive slots (resulting in outages on 10 consecutive transmission attempts) and zero new codewords arrive into the queue, then the queue size remains unchanged. However, if the link is assumed reliable and if zero codewords arrive into the queue for 10 consecutive slots, then the queue size shrinks by 10. This concept is intriguing as it allows for a novel method for waiting-time/delay (or queue-length) management in fading channels: If the average waiting-time is large then it can be reduced by using a smaller coding rate (codewords with a smaller amount of data) at the transmitter. Conversely, a larger coding rate (more information per codeword) can be used at the transmitter to increase communications throughput at the expense of a larger waiting-time.
7.3 Near-optimal throughput maximization with queuing delay constraints for Bernoulli sources

Since (7.3) is not necessarily a convex optimization problem it may have many local maxima. As such, numerical techniques to solve (7.3) may not converge to the globally optimal solution. In this situation a near-optimal optimization problem that is amenable to a numerical solution is highly desirable. We illustrate for the Bernoulli arrival process, due to its mathematical tractability, but the analysis could be extended to other arrival processes as well.

We begin with the observation that the optimal arrival rate, $a_{\text{MZT}_{\text{RT}}}^*$, in (7.3) does not fluctuate greatly a function of $D$. This phenomenon can clearly be seen in Figure 7.4. Similarly, we can see from Figure 7.3 that the optimal transmission rate drops significantly for small $D$. Clearly adjusting $R$ rather than $a$ is more important for controlling the waiting-time $D$ while maximizing the throughput. Therefore, it makes sense for the arrival rate to be fixed at

$$a = a_{\text{MZT}_{\text{RT}}}^*$$

$$= 1 - P_{\text{out}}(R_{\text{MZT}_{\text{RT}}}^*, P_{\text{av}}, K)$$

$$= e \left( \frac{R_{\text{MZT}_{\text{RT}}}^* - 1}{P_{\text{av}}} \right).$$

which is the arrival rate corresponding to 100% utilization using the optimal coding rate $R_{\text{MZT}_{\text{RT}}}^*$. For a fixed arrival rate $a = a_{\text{MZT}_{\text{RT}}}^*$ we define near-optimal $n\text{MZT}_{\text{RT}}^*$ for
coding delay $K$ and power $P_{av}$ with average waiting time $D$ as

$$n_{MZT}^D(R, P_{av}, K) = \sup_R \left\{ a^*_{MZT_{RT}} R : \frac{1 - a^*_{MZT_{RT}}}{1 - a^*_{MZT_{RT}} - P_{out}(R, P_{av}, K)} \leq D \right\}. \quad (7.7)$$

This is a convex optimization problem since both the objective function and set of feasible points are convex. Therefore, we can be assured that a globally optimal solution to (7.7) exists. The existence of a near-optimal convex optimization problem is also useful since both (7.3) and (7.7) must be solved numerically; if an optimization algorithm converges to a local maxima in both cases, then we can be assured that it is the globally optimal solution to (7.7) while it may not be for (7.3).

When $K = 1$ and the channel fading is $\chi^2$ a closed form solution to the near-optimal (7.7) can be found.

**Theorem 7.3.1.** If $K = 1$ and the channel fading process follows a $\chi^2$ fading distribution then

$$n_{MZT}^D(P_{av}, 1) = a^*_{MZT_{RT}} \log \left( 1 - P_{av} \log \left( \frac{1 - a^*_{MZT_{RT}}(1 - D)}{D} \right) \right). \quad (7.8)$$

**Proof.** We begin with

$$n_{MZT}^D(P_{av}, 1) = \sup_R \left\{ a^*_{MZT_{RT}} R : \frac{1 - a^*_{MZT_{RT}}}{1 - a^*_{MZT_{RT}} - P_{out}(R, P_{av}, 1)} \leq D \right\}. \quad (7.9)$$

For $K = 1$ and $\chi^2$ fading

$$P_{out}(R, P_{av}, 1) = 1 - e^{-\left( \frac{R-1}{P_{av}} \right)}. \quad (7.10)$$
Substituting this into the waiting-time constraint we obtain

\[
\frac{1 - a_{M\text{Z}T_{\text{RT}}}^*}{e^{\left(\frac{R}{P_{av}}\right)} - a_{M\text{Z}T_{\text{RT}}}^*} \leq D
\]

which after some algebraic manipulation yields

\[
R \leq \log \left(1 - P_{av} \log \left(\frac{1 - a_{M\text{Z}T_{\text{RT}}}^*(1 - D)}{D}\right)\right).
\]

Clearly the linear objective function in (7.9) is maximized by satisfying (7.12) with equality, resulting in (7.8). □

In Figure 7.5 we see that the near-optimal nM\text{Z}T_{\text{RT}}^D(P_{av}, K) achieved by only varying the coding rate performs nearly as well as M\text{Z}T_{\text{RT}}^D(P_{av}, K) achieved by optimizing the transmission rate and codeword arrival rate. A reduction in either a
and $R$ reduces both the expected waiting-time and maximum throughput. However, the maximum throughput suffers a great deal more if $a$ rather than $R$ is reduced, intuitively explaining the fact that $a^*_{MZTRT} \approx a^*_{MZTPR}$.

As with $R^*_{MZTPR}$, $R^*_{MZTRT}$ converges to $R^*_{MZTRT}$ as $D \to \infty$. For $K = 1$ this can be analytically seen by the fact that the throughput maximizing transmission rate

$$R^*_{MZTRT} = \log \left( 1 - P_{av} \log \left( \frac{1 - a^*_{MZTRT}(1-D)}{D} \right) \right) \quad D \to \infty \quad \log \left( 1 - P_{av} \log \left( a^*_{MZTRT} \right) \right) \quad = R^*_{MZTRT}.$$  

(7.13)

7.4 The queueing delay vs. coding delay tradeoff

In the previous section the waiting-time is constrained to be less than $D$ for a fixed coding delay of $K$. However, some applications using a communication system are affected by the total delay and it does not matter whether the delay is spent in coding or queueing. For small $K$ retransmissions are less costly in terms of delay but the instantaneous capacity, but the amount of information that can be reliably transmitted with each codeword, is small. For large $K$ the opposite is true, the instantaneous capacity is larger but retransmission is more costly in terms of delay. By optimizing over the coding delay $K$, an optimal balance can be struck.

Using this idea we can define

$$MZTPR_{RT}(P_{av}) = \sup_{K} \{MZTPR_{RT}(P_{av}, K) \}$$

(7.14)
Figure 7.6: $\text{MZT}_D^{RT}(P_{av}, K)$ as a function of $K$ for a waiting time of $D = 20$ and for $P_{av} = 10$dB. Clearly there is a tradeoff between coding delay and queueing delay. In this example a coding delay of $K = 16$ achieves maximum throughput for $D = 20$.

as the highest throughput for $P_{av}$ and average waiting-time $D$. This is achieved by solving (7.3) for each value of $K \in \{1, 2, \ldots, D\}$ and then taking the supremum, over $K$, of these values. To the end-user the average waiting time is $D$ for each coding delay $K$, however, the throughput is not. By optimizing over $K$, the throughput can be maximized without any effect on the average waiting-time of end users.

The tradeoff between coding delay and queueing delay is illustrated in Figure 7.6, which plots $\text{MZT}_D^{RT}(P_{av}, K)$ as a function of $K$ for $D = 20$ and $P_{av} = 10$dB. In this case we see that there is a unique coding delay, $K = 16$, that corresponds to $\text{MZT}_D^{RT}(P_{av})$. This indicates that for a total waiting-time of $D = 20$ that the
coding delay should be set to $K = 16$ and the codeword arrival and transmission rates found by solving (7.3). Also note that only zero throughput is achievable with $K = 20$. This is due to the fact that the minimum delay is $D = 20$ since $K = 20$ and a retransmission of any codeword would violate the average waiting-time constraint. Since retransmissions are not permitted in this case, and zero-outage communication is not possible with a single-transmission attempt (delay-limited capacity is zero), the throughput is zero.
Chapter 8
Conclusions and Future Work

8.1 Conclusions

Communication systems have traditionally been analyzed from the perspective that encoded data is transmitted only once. Performance metrics attempt to quantify the highest reliable data rate under this single- attempt paradigm, which is particularly well suited to delay-unconstrained systems. Practical communication systems are delay-limited and must use finite-length codewords. Here, transmitters need not restrict themselves to a single transmission attempt per codeword, but rather multiple transmission attempts can be permitted to ensure reliable communication. In this thesis we address the clear disconnect between the multi- attempt nature of practical communication systems and the conventional single- attempt measures of performance — the maximum average throughput with a multi- attempt approach provides a more accurate reflection of performance than the typical measures used today.

By matching the multi- attempt scheme to that of one used in a real- world system (ARQ, for example), the highest throughput for that system can be found. Maximum zero-outage throughput (MZT) and maximum e- throughput (MeT) represent the highest throughput achievable for a particular multi- attempt scheme if an unlimited and limited number of transmission attempts per codeword is permitted, respectively.
That is, MZT quantifies the highest error-free throughput, while McT quantifies the highest throughput with effective outage probability $\epsilon$.

Our analysis has resulted in several interesting conclusions. With CSI-R, throughput is maximized by optimally selecting the transmission rate. We saw that:

- While delay-limited capacity is zero for many common fading distributions, regardless of coding delay, it is possible to achieve non-zero MZT. This is due to the fact that we permit multiple transmission attempts per codeword.

- Ergodic capacity is a special case of the MZT definition, namely when $K = \infty$.

- Transmission rate and coding delay are intimately related. The transmission rate must be very carefully selected for the coding delay at hand to maximize the throughput of the system.

- It is difficult to assume ergodicity. If such an assumption were easy to make, then a throughput near capacity would be possible for some $K$ “large enough”. However, we see that making a false assumption can significantly degrade the throughput of a system. Clearly, the capacity promises are difficult to keep in practical delay-limited situations.

- Often the optimal operating point in terms of transmission rate and outage probability can be quite non-intuitive. For some systems the optimal transmission rate can actually be higher than the ergodic capacity of the channel.
Such high transmission rates are not typically used in practice. Additionally
the optimal outage probability can be quite high, in some cases greater than
0.5. That is, from a throughput point of view it is often preferable to lose ev­
er other codeword transmitted to outage. This is also rather non-intuitive as
normally small packet error rates such as $\epsilon = 0.01$ or $\epsilon = 0.05$ are targeted.

With CSI-RT, throughput is maximized by performing rate selection and power
control. Here we see that:

- $\text{MZT}_{\text{DT}}$ always exceeds delay-limited capacity. That is, for the same coding
delay $K$, a higher throughput is possible using the multi-attempt approach
than with the single-attempt approach.

- It is possible for $\text{MZT}_{\text{DT}}$ to approach ergodic capacity, achievable only with
infinite coding delay, with a small coding delay.

- The optimal transmission rate that maximizes communications throughput
must be selected carefully. As is the case with only CSI-R, the optimal trans­
mission rate with CSI-RT can actually be higher than the ergodic capacity of
the channel.

- Power control plays a significant role in the throughput of delay-limited sys­
tems. This is quite different to delay-unconstrained systems where it provides
"negligible capacity gains" [4]. In fact, we illustrate that power control is more
important than ergodicity (large coding delays).

The multi-attempt approach allows for higher error-free throughput for delay-limited systems than the single-attempt approach. The increased throughput comes with the cost of a queueing delay that is not present in single-attempt systems. We also quantify the maximum throughput with a constraint on the expected waiting-time, time spent by codewords either queued and/or under (re)transmission. Since end users only care about the total waiting time they are indifferent to whether the delay is mostly coding delay or queueing delay. We also examine the effect on throughput if queueing delay is traded for coding delay. We show that the nature of queueing delays in fading channels is quite different than that of wired communication links. As such flow control algorithms in fading channels must respond differently, by adjusting the coding rate and not the arrival rate of codewords.

8.2 Future Work

Our analysis focuses on communication systems that use a single transmission rate that is optimally selected. This makes it useful for the analysis of practical systems that typically use a single transmission rate or select from a small set of transmission rates. There is ample opportunity for further study based on the concepts provided in this thesis.

In this thesis we proposed several multi-attempt schemes. This by no means is an exhaustive list and others can be proposed. As an alternative to schemes RT and
ID, the transmitter need not send exact copies of the codeword in outage, but rather more complex functions of those codewords. A simple example of this may be to increase the power for retransmitted codewords to further reduce the likelihood of outage. Another alternative to scheme DT is to simply transmit the codeword at the average power even though an outage will occur and combine codewords at the receiver as in scheme ID. By substituting the form of the service time of any proposed retransmission scheme into our analysis framework we can analyze its performance in a straightforward manner.

Analysis can also be performed by relaxing some of the assumptions of this paper. In many practical systems feedback channel used to relay channel state information is limited. As such, the transmitter may only have a quantized estimate of the actual channel. Performing throughput maximization when only quantized CSI is available is also an important problem to examine. This analysis would be similar to that of Chapter 6, but would result in a finite set of possible power allocation levels since the number of possible channel realization is now finite.

We considered the situation in which the $K$ channel fades affecting a particular codeword are all seen simultaneously in a causal manner prior to transmission. This is the scenario in many multicarrier systems with $K$ parallel subchannels. However, having a code that spans $K$ blocks in time and performing causal power control in time is also a very relevant problem. This can be accomplished by modifying the results in Chapter 6 to use causal power allocation strategies along the lines of those
in [21].

Our analysis thus far has focused on single-user communications systems. However, the tools presented here can also be used to analyze other systems including multiuser, MIMO, relay and others. The main task for such analysis would be to determine the outage probability function, whose form would differ from the single user system we provide in this thesis. Along this same line, we can also analyze and design packet schedulers for communication systems over fading channels. This can be accomplished by adjusting the transmit power based not only on the condition of the fading channel but also on the number of packets that are enqueued at the transmitter.
Appendix A

Proof of $\text{MZT}_\text{DT}$ for different power constraints

A.1 Proof of Theorem 6.1.1

From the form of the outage probability (2.24) we see that

$$[1 - P_{\text{out}}(R, \gamma, K)] = E_\alpha \left[ I_P \left\{ R \leq C_K [\alpha, \gamma(\alpha)] \right\} \right]. \quad (A.1)$$

and thus

$$\text{MZT}_{\text{DT}}^\text{st}(P_{av}, K) = \sup_{R} \sup_{\gamma} \left\{ R E_\alpha \left[ I_F \left\{ R \leq C_K [\alpha, \gamma(\alpha)] \right\} \right] : \gamma \in \mathcal{O}_{\text{st}}^K(P_{av}) \right\}. \quad (A.2)$$

We realize that the outage minimizing power allocation strategy,

$$\gamma^\text{st}_k(\alpha) = \max \left\{ \lambda^\text{st}(\alpha) - \frac{1}{\alpha_k}, 0 \right\} \quad (A.3)$$

with $\lambda^\text{st}(\alpha)$ as the solution to

$$\sum_{k=0}^{K-1} \max \left( \lambda^\text{st}(\alpha) - \frac{1}{\alpha_k}, 0 \right) = KP_{av}, \quad (A.4)$$

are equivalent to (2.33) and (2.34), respectively. Substituting the form of the outage minimizing power allocation strategy into (A.2) yields

$$\text{MZT}_{\text{DT}}^\text{st}(P_{av}, K) = \sup_{R} \left( R E_\alpha \left[ I_P \left\{ R \leq \frac{1}{K} \log \left( 1 + \alpha_k \max \left\{ \lambda^\text{st}(\alpha) - \frac{1}{\alpha_k}, 0 \right\} \right) \right\} \right] \right) \quad (A.5)$$

as desired.
A.2 Proof of Theorem 6.1.2

From the form of the outage probability (2.24) we see that

\[ [1 - P_{\text{out}}(R, \gamma, K)] = \mathbb{E}_{\mathbf{x}} \left[ I_{F} \{ R \leq C_{K} [\alpha, \gamma(\alpha)] \} \right]. \]  

(A.6)

and that

\[ \text{MZT}^{\text{st}}_{\text{DT}}(P_{av}, K, P_{p}) = \sup_{R} \sup_{\gamma} \left\{ \mathbb{E}_{\mathbf{x}} \left[ I_{F} \{ R \leq C_{K} [\alpha, \gamma(\alpha)] \} \right] : \gamma \in \mathcal{C}_{K}^{\text{st}}(P_{av}, P_{p}) \right\}. \]

(A.7)

From Section 5.1 the optimal power allocation strategy under the short-term average and peak power constraints is

\[ \gamma^{\text{st}}_{k}(\alpha) = \begin{cases} \min \left( \max \left( \tilde{\lambda}^{\text{st}}(\alpha) - \frac{1}{\alpha_{k}}, 0 \right), P_{p} \right) & \text{if } P_{p} > P_{av} \\ P_{p} & \text{if } P_{p} \leq P_{av} \end{cases} \]

(A.8)

with \( \tilde{\lambda}^{\text{st}}(\alpha) \) as the solution to

\[ \sum_{k=0}^{K-1} \min \left( \max \left( \tilde{\lambda}^{\text{st}}(\alpha) - \frac{1}{\alpha_{k}}, 0 \right), P_{p} \right) = K P_{av}. \]

(A.9)

In the case \( P_{p} \leq P_{av} \) the optimal power allocation strategy is to always transmit at the peak power. The substituting \( \gamma_{k} = P_{p} \) into (A.7) yields

\[ \text{MZT}^{\text{st}}_{\text{DT}}(P_{av}, K, P_{p}) = \sup_{R} \left( \mathbb{E}_{\mathbf{x}} \left[ I_{F} \left( R \leq \frac{\log(1 + \alpha_{k} P_{p})}{K} \right) \right] \right). \]

(A.10)

On the other hand if \( P_{p} \leq P_{av} \) the optimal power allocation strategy is given by \( \gamma_{k} = \min \left( \max \left( \tilde{\lambda}^{\text{st}}(\alpha) - \frac{1}{\alpha_{k}}, 0 \right), P_{p} \right) \) which when substituted into (A.7) results in

\[ \text{MZT}^{\text{st}}_{\text{DT}}(P_{av}, K, P_{p}) = \sup_{R} \left( \mathbb{E}_{\mathbf{x}} \left[ I_{F} \left( R \leq \frac{\log(1 + \alpha_{k} \xi)}{K} \right) \right] \right) \]

with \( \xi = \min \left\{ \max \left( \tilde{\lambda}^{\text{st}}(\alpha) - \frac{1}{\alpha_{k}}, 0 \right), P_{p} \right\}, \) as desired.
A.3 Proof of Theorem 6.1.3

Under the long-term average power constraint the outage minimizing power allocation policy takes the form [24]

\[ \gamma_k^*(\alpha) = \begin{cases} 
\hat{\gamma}_k(\alpha), & \text{w/ prob 1} \quad \text{if } \langle \hat{\gamma}(\alpha) \rangle < s^* \\
\hat{\gamma}_k(\alpha), & \text{w/ prob } w^* \quad \text{if } \langle \hat{\gamma}(\alpha) \rangle = s^* \\
0, & \text{w/ prob } (1 - w^*) \quad \text{if } \langle \hat{\gamma}(\alpha) \rangle = s^* \\
0, & \text{w/ prob 1} \quad \text{if } \langle \hat{\gamma}(\alpha) \rangle > s^* 
\end{cases} \] (A.12)

for some $s^* \in \mathbb{R}_+$ and $w^* \in [0,1]$ and

\[ \hat{\gamma}_k(\alpha) = \max \left( \lambda_k^*(\alpha) - \frac{1}{\alpha_k}, 0 \right) \] (A.13)

with $\lambda_k^*(\alpha)$ as the solution to

\[ \frac{1}{K} \sum_{k=0}^{K-1} \log \left[ 1 + \alpha_k \max \left( \lambda_k^*(\alpha) - \frac{1}{\alpha_k}, 0 \right) \right] = R, \] (A.14)

which is equivalent to (2.44). From this and (2.24) we see that

\[ [1 - P_{\text{out}}(R, \gamma^*, K)] = \mathbb{E}_\alpha \left[ I_{\mathcal{P}_F} \left( \langle \hat{\gamma}(\alpha) \rangle < s^* \right) \right] + w^* \mathbb{E}_\alpha \left[ I_{\mathcal{P}_F} \left( \langle \hat{\gamma}(\alpha) \rangle = s^* \right) \right]. \] (A.15)

which when substituted along with the form of $\gamma^*$ into (6.3) yields

\[ \text{MZT}_{\text{DT}}^{\text{lt}}(\mathcal{P}_{\text{av}}, K) = \sup_{\mathcal{R}} \left\{ R \left[ \mathbb{E}_\alpha \left\{ I_{\mathcal{P}_F} \left( \frac{1}{K} \sum_{k=0}^{K-1} \max \left( \lambda_k^*(\alpha) - \frac{1}{\alpha_k}, 0 \right) < s^* \right) \right] \right] + w^* \mathbb{E}_\alpha \left\{ I_{\mathcal{P}_F} \left( \frac{1}{K} \sum_{k=0}^{K-1} \max \left( \lambda_k^*(\alpha) - \frac{1}{\alpha_k}, 0 \right) = s^* \right) \right] \right\}, \] (A.16)

as desired.
A.4 Proof of Theorem 6.1.4

Under the long-term average and peak power constraints the outage minimizing power allocation strategy from Section 5.2 is

\[ \tilde{\gamma}(\alpha) = \begin{cases} \tilde{\gamma}(\alpha), & \text{w/ prob } 1 \quad \text{if } \alpha \notin \mathcal{G}(\mathcal{P}_p) \text{ and } \langle \tilde{\gamma}(\alpha) \rangle < s^* \\ \tilde{\gamma}(\alpha), & \text{w/ prob } w^* \quad \text{if } \alpha \notin \mathcal{G}(\mathcal{P}_p) \text{ and } \langle \tilde{\gamma}(\alpha) \rangle = s^* \\ 0, & \text{w/ prob } (1 - w^*) \quad \text{if } \alpha \notin \mathcal{G}(\mathcal{P}_p) \text{ and } \langle \tilde{\gamma}(\alpha) \rangle = s^* \\ 0, & \text{w/ prob } 1 \quad \text{if } \alpha \notin \mathcal{G}(\mathcal{P}_p) \text{ and } \langle \tilde{\gamma}(\alpha) \rangle > s^* \\ 0, & \text{w/ prob } 1 \quad \text{if } \alpha \in \mathcal{G}(\mathcal{P}_p) \end{cases} \]  

(A.17)

for some \( s^* \in \mathbb{R}_+ \), \( w^* \in [0, 1] \), subset \( \mathcal{G}(\mathcal{P}_p) \subset \mathbb{R}^K_+ \) and

\[ \tilde{\gamma}_k(\alpha) = \min \left( \max \left( \lambda_k^*(\alpha) - \frac{1}{\alpha_k}, 0 \right), \mathcal{P}_p \right) \]  

(A.18)

with \( \lambda_k^*(\alpha) \) as the solution to

\[ \frac{1}{K} \sum_{k=0}^{K-1} \log \left[ 1 + \alpha_k \min \left( \max \left( \lambda_k^*(\alpha) - \frac{1}{\alpha_k}, 0 \right), \mathcal{P}_p \right) \right] = R. \]  

(A.19)

From this and (2.24) we see that

\[ [1 - P_{\text{out}}(R, \tilde{\gamma}_k^*, K)] = \]  

\[ \mathbb{E}_\alpha [\mathcal{I}_F (\alpha \notin \mathcal{G}_p)] \left\{ \mathbb{E}_\alpha \left[ \mathcal{I}_F (\langle \tilde{\gamma}(\alpha) \rangle < s^*) \right] + w^* \mathbb{E}_\alpha \left[ \mathcal{I}_F (\langle \tilde{\gamma}(\alpha) \rangle = s^*) \right] \right\}, \]  

(A.20)

which when substituted along with the form of \( \tilde{\gamma}_k^* \) into (6.3) yields

\[ \text{MZT}^\text{ht}_{\mathcal{D}_T}(\mathcal{P}_s, K, \mathcal{P}_p) = \sup_R \left\{ R \left[ \mathbb{E}_\alpha \left\{ \mathcal{I}_F (\alpha \notin \mathcal{G}_p) \right\} \right] \right\} \]  

\[ + w^* \mathbb{E}_\alpha \left[ \mathcal{I}_F (\kappa = s^*) \right] \]  

(A.21)

with \( \kappa = \frac{1}{K} \sum_{k=0}^{K-1} \min \left( \max \left( \lambda_k^*(\alpha) - \frac{1}{\alpha_k}, 0 \right), \mathcal{P}_p \right) \), as desired.
Bibliography


2002.


(Monticello, IL), 1999.


