Télécom Paris (ENST)

Institut EURECOM

THESE

présentée pour obtenir le grade de docteur

de l'école nationale supérieure

des télécommunications

Spécialité : Communications et Electronique

Daniela Tuninetti

On Multiple-Access Block-Fading Channels

Soutenue le 15 Mars 2002 devant le jury composé de:

Président	Emre Telatar, EPFL (Lausanne, CH)
Rapporteurs	Amos Lapidoth, ETHZ (Zurich, CH)
	Sergio Verdú, Princeton University (Princeton, NJ USA)
Examinateurs	Joseph Boutros, Telecom Paris (Paris, FR)
	Raymond Knopp, Institut Eurécom (Sophia Antipolis, FR)
Directeur de thèse	Giuseppe Caire, Institut Eurécom (Sophia Antipolis, FR)

Acknowledgments

First of all, I want to thank my advisor, Prof. Giuseppe Caire, for his guidance and initiative. Now, looking back at what those three years have meant for me, I would rather say I am most grateful for the chance he gave me. I am glad I have taken up the "challenge" and hope I did not disappoint him.

I want to thank Prof. Sergio Verdú for having spent some afternoons of his summer holidays to work with me, to read my papers and to give me suggestions. The last part of this thesis developed form one of his many fruitful ideas.

I am grateful to all the members of my defense committee, Prof. Amos Lapidoth, Prof. Sergio Verdú, Prof. Emre Telatar, Prof. Joseph Boutros and Prof. Raymond Knopp who accepted to be present at my defense despite of their over-booked agenda.

Second, *un tres grand merci* to the Eurecom stuff. The efficiency and competence of all these people make the success of our Institute.

Third, I wish to show my gratitude to the people I met here for having made my staying in France happy and interesting: Alex, Max, Chris, Houda, Jussi, Masato, Despina, David, Beppe, Neda, Souad, Kader, Albert, and all those that I do not list here for space limitation. I heartily hope the end of the thesis does not imply the end of our friendship.

Last but not the least, I would like to thank my parents, my sister, my family and all my "old" Italian friends for their unconditional support, encouragement and love.

January 2002

Abstract

In this work we analyze the performance of multiple-access fading channels with delay constraints, which arise since the fading dynamics is slow with respect to the tolerable decoding delay.

In the first part of the work, we consider a simple, decentralized, uncoordinated system where users access at random the channel whenever they have data to transmit. To cope with background noise, fading and interference from other users, packets that are negatively acknowledged by the receiver are retransmitted. At the receiver, packets related to the same information bits are combined together to increase decoding reliability. We study three different protocols in terms of total throughput (bit/s/Hz) as function of several different parameters. Closed form throughput formulas are derived by using random coding, typical set decoding and renewal-reward theory arguments. Then, we perform a comparison between systems that implement at network layer a repetition protocol and at physical layer different receiver structures. We compare the optimized throughput as a function of the average transmit energy per successfully received bit. In doing so we get insight into the optimal choice of the system parameters, like the transmission rate and the average channel load.

In the second part, we consider a completely centralized system, where users know the channel state and vary rate and power according to the channel conditions so that their rate is always inside the fading dependent achievable rate region. We define a variable coding scheme and study the corresponding long-term average capacity region. Since we assume that code-words can span a given number of consecutive time slots and that the channel is known only up to the current slot, the optimal solution is given in terms of Dynamic Programming algorithm. Then, we consider the performance in the low spectral efficiency regime, which is where the major benefits of transmitter feedback occur. We derive the long-term average capacity region per unit energy (infinite bandwidth regime) and the wideband slope region (wideband regime). We show that the simple one-shot policy that concentrates the available energy in only one of the fading states, chosen on the basis of its strength and of how likely it is that a more favorable fading state will appear before the end of the code-word, is wideband optimal.

Résumé

L'objet de ce travail est d'analyser l'effet des contraintes de retard sur la performance des canaux gaussiens multi-utilisaturs avec évanouissement. Dans notre modèle, dû à la dynamique très lente de l'évanouissement par rapport au retard de décodage tolérable, chaque mot de code est affecté par un nombre fini d'états d'évanouissement.

Dans la première partie, nous considérons un système simple, décentralisé et non coordonné qui est accédé au hasard par les utilisateurs toutes les fois qu'ils ont des données à transmettre. Pour faire face au bruit, à l'évanouissement et à l'interférence d'autres utilisateurs, les paquets de données qui ont été négativement reconnus par le récepteur sont retransmis. Au récepteur, les paquets liés au même bit d'information sont combinés pour augmenter la fiabilité de décodage. Nous étudions trois protocoles différents en termes de débit total (bit/s/Hz) en fonction du nombre d'utilisateurs, du retard, du débit du code, de la probabilité d'accéder au canal, du nombre de retransmissions et du rapport signal à bruit. Puis, nous exécutons une comparaison entre différents systèmes qui utilisent aux couches hautes un protocole à répétition et à la couche physique différentes stratégies de détection. Nous comparons le débit total en fonction de l'énergie moyenne par bit correctement reçu. En faisant une telle comparaison, nous obtenons la valeur optimale des paramètres de système, comme le débit de transmission et le chargement moyen de canal.

Dans la deuxième partie, nous considérons un système complètement centralisé, où les utilisateurs connaissent l'état du canal et peuvent changer le débit de transmission et la puissance selon l'état du canal de façon qu'ils soient toujours à l'intérieur de la région de capacité. Nous définissons un système de codage à débit variable et étudions sa région de capacité à long terme. Puisque nous supposons que les mots de code peuvent être étalés sur un nombre donné de slot consécutifs dans le temps et que le canal est connu seulement jusqu'au slot actuel, la solution optimale à notre problème est donnée en termes d'algorithme de programmation dynamique. Enfin, nous considérons la performance du système décrit quand il est utilisé sur un canal à bande très large. Nous caractérisons la performance quand la largeur de bande est infinie (valeur minimale de l'énergie moyenne par bit) ainsi que quand la largeur de bande est grande mais finie (région de pentes de l'efficacité spectrale). Nous prouvons que la simple politique d'allocation de puissances "one-shot", politique qui concentre toute l'énergie disponible dans un seul des états d'évanouissement, est optimale dans un système à très large bande.

Contents

List of Figures

Chapter 1

Multiple-access communication over fading channels

In this first chapter, we review briefly the model we have adopted for the fading channel and summarize the information theoretic results on multiple-access channels that are relevant to the rest of the dissertation. The chapter ends with the thesis outline and the list of our contributions.

1.1 Introduction

For many years after Shannon's landmark paper "A mathematical Theory of Communications" in 1948, information theory has seemed the play-field of mathematicians galvanized by the elegance of the newly born branch of mathematics.

Practical applicability of information theory results has been unclear for long time. On one hand, information theory models were thought to be "too simplistic" with respect to the complexity of the real-world systems and hence, the rigorous results derived from those models, could not be successfully implemented in any communication systems. On the other hand, capacity limits predicted by information theory were thought to be not applicable because of the complexity involved in the capacity achieving co-decoding schemes.

Nowadays, things have radically changed. First of all, complexity is less and less an issue and computation capability grows very fast. Moreover, modern coding techniques can approach the information theory limits

in at least some important single-user channel models. Finally, multi-user communication has been spured by the tremendous growth of wireless systems. The ever growing demand of wireless services every-where every-time "obliges" to access those limits that information theory predicts.

The result is that today information theory inspires the techniques at the basis of the design of compression, coding, signaling and detection of contemporary information systems. With the increasing demand of more capable systems, in terms of bandwidth and bit rate, information theory is bound to have an even stronger impact on future communication systems.

In this work, we shall focus on the multiple-access Gaussian channel with fading, commonly used to study the up-link of wireless systems, that consists of many senders transmitting to a common receiver by sharing the same communication channel impaired by additive noise and multiplicative fading. We start by describing the physical propagation channel and by characterizing its time-varying behavior from a statistical point of view. Then, we shall summarize background information theoretic results related to this multiple-access channel model and provide the motivations of our work. We conclude this introduction, outlining the main contributions of this thesis and by describing of the structure of the report.

1.2 Propagation channel model

A fading multipath channel is generally characterized as a linear, time-varying system with *impulse response* $c(t;\tau)$, or a time-varying frequency response $C(t;f)$, which is a wide-sense stationary random process in the variable t. Time variations of $c(t;\tau)$ result in frequency spreading of the transmitted signal, which is generally called Doppler spreading. Multipath propagation results in spreading the transmitted signal in time. By assuming that the multipath signals propagating through the channel at different delays are uncorrelated (widesense stationary uncorrelated scattering), a doubly spread channel can be characterized by the *scattering function* $S(\tau;\lambda)$, which measures the power spectrum of the channel at delay τ and frequency offset λ.

From the scattering function, we obtain the *delay power spectrum* of the channel by averaging over λ, i.e., $S_m(\tau) = \int S(\tau;\lambda)\mathrm{d}\lambda$ and the *Doppler power spectrum* by averaging over τ, i.e., $S_d(\lambda) = \int S(\tau;\lambda)\mathrm{d}\tau$. The range of values over which the delay power spectrum $S_m(\tau)$ is non-zero is defined as the *multipath spread* T_m of the channel. Similarly, the range of values over which the Doppler power spectrum $S_d(\lambda)$ is non-zero is defined as the *Doppler spread* B_d of the channel. The Doppler spread B_d provides a measure of how rapidly the channel impulse response varies in time: a slowly varying fading channel has large coherence time, where the *channel coherence time* is defined as $T_{\mathrm{coh}} = 1/B_d$, and a fast varying fading channel has

small coherence time. In a similar manner, the inverse of the multipath spread T_m is defined as *channel coherence bandwidth*, i.e., $B_{\text{coh}} = 1/T_m$, and measures the width of the interval of frequencies which are similarly affected by the channel response: the channel response at frequencies whose separation is smaller than B_{coh} is highly correlated.

The product $T_m B_d$ is called the *spread factor* of the channel. If $T_m B_d < 1$, the channel is said to be *underspread*, otherwise it is said to be *overspread*. Generally, if the spread factor $T_m B_d \ll 1$, the channel can be easily estimated by the receiver while, when the spread factor $T_m B_d > 1$, channel estimation is extremely difficult.

1.2.1 Flat fading channels vs. frequency selective channels

Let $x(t)$ be the transmitted signal, let $X(f)$ denote its Fourier transform and let W denote its bandwidth. The received signal, without additive noise, is given by

$$r(t) = \int c(t;\tau)x(t-\tau)\mathrm{d}\tau = \int C(t;f)X(f)\mathrm{e}^{\mathrm{j}2\pi ft}\mathrm{d}f \tag{1.1}$$

If the signal bandwidth is much smaller than the coherence bandwidth of the channel, i.e., $W \ll B_{\text{coh}}$, then all the frequency components in $X(f)$ undergo the same attenuation and phase shift. This implies that, the time-variant transfer function of the channel $C(t;f)$ is constant in within the bandwidth W, i.e., $C(t;f) = c(t)$. Such a channel is called frequency-nonselective or *flat fading*. In this case, the received signal $r(t)$ simplifies to

$$r(t) = c(t)\, x(t) \tag{1.2}$$

i.e., the multipath components of the channel are not resolvable because the signal bandwidth $W \ll B_{\text{coh}} = 1/T_m$ and the overall effect on the transmitted signal is a multiplicative attenuation.

A frequency-nonselective channel is said to be *slowly fading* if the time duration of a transmitted symbol, defined as T_s, is much smaller than the coherence time of the channel, i.e., $T_s \ll T_{\text{coh}}$. Since, in general, the signal bandwidth $W > 1/T_s$, it follows that a slowly fading frequency-nonselective channel is underspread. We define a *rapidly fading channel* as a channel that satisfies $T_s > T_{\text{coh}}$.

When the transmitted signal has a bandwidth W greater than the coherence bandwidth of the channel B_{coh}, the frequency components of $X(f)$ with frequency separation exceeding B_{coh} are subjected to different gains and phase shifts. In such a case, the channel is said to be *frequency selective*. The multipath components separated in delay by at least $1/W$ are

resolvable. The resulting channel model is a tapped-delay line with time-varying tap coefficients

$$c(t;\tau) = \sum_{n=1}^{N} c_n(t)\delta(\tau - n/W) \tag{1.3}$$

where $c_n(t)$ in the complex channel gain on the n-th path and $N = \lfloor T_m W \rfloor + 1$ represents the number of resolvable paths.

1.2.2 Ergodic channels vs. nonergodic channels.

Up to now, we distinguished between slowly varying and rapidly varying channels on the basis on the symbol duration T_s with respect to the channel coherence time. Other relevant parameters are the signaling bandwidth W and the transmission duration of the whole message (code-word) T. We distinguish between *ergodic* and *nonergodic* channels according to the whole code-word transmission duration T with respect to the variability of the fading process measured by $T_{\rm coh}$, assuming that $c(t;\tau)$ is a nondegenerate random process (otherwise $T_{\rm coh} \to \infty$). The fact that a specific channel is underspread ($T_{\rm coh}B_{\rm coh} > 1$) and that can be treated as a flat slow-fading process ($W \ll B_{\rm coh}$), does not imply that the total transmission duration may not span a large number of dimensions ($WT \gg 1$) so that the channel can be viewed as ergodic, hence giving rise to standard notions of the *ergodic capacity* [1]. Ergodic capacity is the classical Shannon type capacity whose operative definition is provided by the coding theorem. Note that the condition $WT \gg 1$ is required for Shannon type capacity to exists even for non-faded time-invariant channels as otherwise reliable communication is not possible.

When the product WT is not large, then the channel is nonergodic. This cases arises when the fading dynamic is slow with respect to the (tolerable) code-word duration. Nonergodic channels are of primary importance to study the effect of delay on the system performance. For non ergodic channels, capacity in the Shannon sense is not defined. The nonergodic case gives rise to interesting information-theoretic settings as capacity versus outage [2] and delay-limited capacities [3], all relying on notions of compound channels [4].

1.2.3 Statistical characterization of the fading

There are several probability distributions that have been used to model the statistical characteristics of the fading channel. Let $\alpha \stackrel{\Delta}{=} |c(t;\tau)|^2$ be the fading power, where for notation convenience we omit the variables t and τ. We indicate with $\Omega \stackrel{\Delta}{=} \mathrm{E}[\alpha]$ the average fading power and with $f_X(x)$ the probability density function (pdf) of the random variable X.

Rayleigh distribution. When there are a large number of scatterers in the channel that contribute to the signal at the receiver, an application of the central limit theorem leads to a Gaussian process model for $c(t;\tau)$. If the process is zero-mean, then the envelope of the channel impulse response has a Rayleigh probability distribution and the phase is uniformly distributed in the interval $[0, 2\pi]$, hence

$$f_{|c|}(x) = \frac{2x}{\Omega} \mathrm{e}^{-x^2/\Omega} \quad x \geq 0$$

and the fading power is exponentially distributed $f_\alpha(x) = \frac{1}{\Omega}\mathrm{e}^{-x/\Omega}$ for $x \geq 0$.

Nakagami-m distribution. An alternative statistical model for the envelope of the channel response is the Nakagami-m distribution. The pdf for this distribution is

$$f_{|c|}(x) = \frac{2m^m}{\Gamma(m)\Omega^m} x^{2m-1} \mathrm{e}^{-mx^2/\Omega} \quad x \geq 0 \tag{1.4}$$

and the parameter $m \geq 1/2$ is referred to as fading figure. By varying the parameters m and Ω, this distribution provides more accuracy in matching the observed signal statistics. The Nakagami-m distribution can be used, as an example, to model land mobile or indoor channels. It includes the Rayleigh distribution as a special case $m = 1$. The pdf of the fading power is $f_\alpha(x) = \frac{m^m}{\Gamma(m)\Omega^m} x^{m-1} \mathrm{e}^{-mx/\Omega}$ for $x \geq 0$.

Rice distribution. A distribution which is appropriate for modeling a Gaussian fading channel in which the impulse response has a nonzero mean component, usually called a specular component, in the Rice distribution. The pdf is

$$f_{|c|}(x) = \frac{2(1+K)x}{\Omega} \mathrm{e}^{-\frac{1+K}{\Omega}x^2 - K} I_0\left(2x\sqrt{\frac{K(1+K)}{\Omega}}\right) \quad x \geq 0 \tag{1.5}$$

where K is defined as the ratio of the nonfading (specular) signal component over the variance of the zero-mean Gaussian component and where $I_0(\cdot)$ is the zero order modified Bessel function of first kind. The two extreme cases of $K = 0$ and $K \to \infty$ give, respectively, the Rayleigh fading pdf and the degenerate case of constant fading. The Rice distribution is a particularly appropriate model for line-of-sight communication links, where there is a direct propagating signal component (the specular component) and multipath components due to secondary reflections. The pdf of the fading power is $f_\alpha(x) = \frac{(1+K)}{\Omega} \mathrm{e}^{-\frac{1+K}{\Omega}x - K} I_0\left(2\sqrt{\frac{xK(1+K)}{\Omega}}\right)$ for $x \geq 0$.

Two-state distribution. To model communication channels where either the signal is received undistorted or it is corrupted to result useless is the *two-state* model. The fading assumes only two possible values: $|c| = 1$ (good channel) with probability p and $|c| = o$ (bad channel) with probability $1-p$, i.e.,

$$f_{|c|}(x) = p\,\delta(x-1) + (1-p)\,\delta(x) \tag{1.6}$$

where $\delta(\cdot)$ is the Dirac delta function. The two-state channel model is commonly used in communication network to model the underlaying physical channel. It finds a natural application in modeling line-of-sight satellite communications.

1.2.4 Adopted fading model

Throughout the whole work, we shall model our communication channel as frequency non-selective ($W \ll B_{\text{coh}}$) and slowly varying ($T_s \ll T_{\text{coh}}$). In particular, we assume a *block-fading model*. The time axis is divided into slots of duration T_{coh} seconds; the fading is assumed to remain constant on the whole slot and to change on the subsequent slot in an i.i.d. (independent and identically distributed) fashion. The duration of a code-word is an integer multiple of the slot duration, i.e., $T/T_{\text{coh}} = N$,. On every slot the number of degree of freedom is large ($T_{\text{coh}}W \gg 1$) so that to guarantee reliable communication on each slot and to allow for reliable channel estimation. In our model the transmitted messages are delay-sensitive and the value of N is intended to measure the "tolerable" delay. If $N \to \infty$ then our channel is ergodic, otherwise it is not.

Most of the numerical examples shall assume a Rayleigh fading model for the fading coefficients. In Chapter 4, in the contest of capacity per unit energy, we shall also report results for the two-state channel and in Chapter 5, considering system performance in the wideband regime, we shall make use of the Rice distribution and the Nakagami-m distribution.

1.3 Multiple-user channel model

A general network model comprises N_{tx} transmitters and N_{rx} receivers, each of which equipped with a set of multiple antennae. The channel between a particular receiving antenna n_r and a particular transmitting antenna n_t is characterized by a time-varying linear filter with impulse $c_{n_t,n_r}(t;\tau)$. This communication system is characterized by specifying: a) to which degree the channels $c_{n_t,n_r}(t;\tau)$, for each (n_t, n_r) pair, are known at each transmitter and receiver, usually referred to as Channel State Information (CSI); b) the communication mode of each transmitter (in the general case, is a mix of multi-access broadcast interference and relay communication modes); c)

the network topology: the configuration and connectivity of the system as well as the mobility of senders and receivers; d) the power constraints that can be an average power applied to each of the transmitting antennas or an average over all the transmitting antennas, the average can be taken over the each code-word ("short-term" average) or over many transmitted code-words ("long-term" average); e) the bandwidth which usually is a critical design parameter; f) the delay constraint which poses a limitation on any practical system and determines the very existence of a Shannon type capacity region.

We shall consider a multi-access system with $N_{\rm tx} = K \geq 1$ senders and $N_{\rm rx} = 1$ receiver, each of which equipped with a single antenna. All senders and the receiver know the joint statistics of fading gains and the statistics of the noise. The receiver tracks perfectly the fading coefficients of all users, i.e., perfect receiver CSI, while for the transmitters we shall treat both the cases of no transmitter CSI and perfect transmitter CSI. About the other constraints, we shall discuss them in deeper details in the sequel.

For a superlative state-of-the-art tutorial on fading channels, the reader can refer to [5].

Before proceeding with an overview of the thesis content, we briefly summarize the information-theoretic results on multi-user fading channels relevant to our work.

1.3.1 What Information theory does study ...

A channel of bandwidth W is accessed by K users who send code-words of length T to a common receiver. The number of channel-symbols per code-word is $L = WT \gg 1$. The (discrete time) received complex signal y_s at time s is

$$y_s = \sum_{k=1}^{K} c_{k,s} x_{k,s} + n_s \tag{1.7}$$

where $x_{k,s}$ stands for the channel input of the k-th user at time s and $c_{k,s}$ designates the fading value at time s for user k. The ergodic assumption means that $\{c_{k,s}\}$ are jointly ergodic in the time index s and independent from user to user (in the index k). The additive-noise n_s is proper complex Gaussian random variable of zero mean and variance N_0. The k-th input is subjected to average-power constraints $\mathrm{E}[1/L \sum_{s=1}^{L} |x_{k,s}|^2] \leq P_k$. For the definitions of code, achievable rates and capacity region for channel (1.7) refer to [1].

Capacity region. Almost contemporarily, Ahlswede [6] and Liao [7] proved that the capacity region of a discrete memoryless multiaccess channel is the

set of vectors $(R_1, \cdots, R_K) \in \mathbb{R}_+^K$ that satisfy

$$\sum_{k \in S} R_k \leq I(\boldsymbol{X}(S); Y | \boldsymbol{X}(S^c), Q) \tag{1.8}$$

for all $S \subseteq \{1, 2, \cdots, K\}$. The input pdf is the form $\Pr(X_1, \cdots, X_K, Q) = \Pr(Q) \prod_{k=1}^K \Pr(X_k|Q)$ where Q is the auxiliary "time-sharing" random variable that makes the capacity region to be a convex set. The symbol $I(\boldsymbol{X}(S); Y|\boldsymbol{X}(S^c), Q)$ designates the mutual information between Y and $\boldsymbol{X}(S)$ (the input signals indexed by S) given Q and $\boldsymbol{X}(S^c)$ (the input signals indexed by the complementary set of S).

A general formula for the multiple-access capacity region was found by Te Sun Han in [8] extending the information-spectrum approach of [9]. Formally, the capacity region is determined by the inequalities (1.8) but where $I(\boldsymbol{X}(S); Y|\boldsymbol{X}(S^c), Q)$ has the meaning of the lim-inf in probability of the information density [9].

Capacity region of the constrained input Gaussian channels with receiver CSI only. We concentrate now on the additive Gaussian channel. In [10], Wyner showed that, if the channel gains c_k are perfectly known at the receiver but not at the transmitters, then the capacity region of (1.7) is determined by the following inequalities

$$\sum_{k \in S} R_k \leq \mathrm{E}\left[\log\left(1 + \frac{1}{N_0} \sum_{k \in S} |c_k|^2 P_k\right)\right] \tag{1.9}$$

for all $S \subseteq \{1, \cdots, K\}$ and where the average is with respect to the joint distribution of $(c_1, \cdots, c_K)$. Capacity (1.9) is achieved by independent and identically distributed (i.i.d.) proper complex Gaussian input $(X_1, \cdots, X_K)$ where X_k has zero mean and variance P_k. To achieve capacity (1.9), there is no need of variable-rate code. Long enough codebook ($T \gg T_{\mathrm{coh}}$), optimal for the unfaded case, are optimal also for the faded case provided that the whole statistics of the fading is revealed within the span of each code-word, i.e., to get the averaging effect of (1.9). For delay-sensitive messages transmitted over slowly-varying channels, the total delay need to achieve capacity can be intolerable.

Some insight into (1.9) can be gained by comparison the multiple-access capacity region with the TDMA achievable region. In a TDMA system, every user transmits for a fraction τ_k of the time with average power P_k/τ_k, where τ_k are non-negative "time-sharing" parameters such that $\sum_{k=1}^K \tau_k = 1$. Therefore, the k-th user achieves rate $R_k^{(\mathrm{tdma})} = \tau_k \, \mathrm{E}[\log(1+(P_k|c_k|^2)/(N_0\tau_k))]$.

Give a subset S, the corresponding rate-sum satisfies

$$\begin{aligned}
\sum_{k\in S} R_k^{(\text{tdma})} &= \sum_{k\in S} \tau_k \, \mathrm{E}\left[\log\left(1+\frac{1}{N_0}\frac{P_k}{\tau_k}|c_k|^2\right)\right] \\
&\overset{(a)}{\leq} \left(\sum_{k\in S}\tau_k\right) \mathrm{E}\left[\log\left(1+\frac{1}{N_0}\frac{\sum_{k\in S} P_k|c_k|^2}{\left(\sum_{k\in S}\tau_k\right)}\right)\right] \\
&\overset{(b)}{\leq} \mathrm{E}\left[\log\left(1+\frac{1}{N_0}\sum_{k\in S} P_k|c_k|^2\right)\right] \\
&\overset{(c)}{\leq} \log\left(1+\frac{1}{N_0}\sum_{k\in S} P_k \, \mathrm{E}[|c_k|^2]\right)
\end{aligned}$$

where (a) follows by log-sum inequality [1], (b) follows since $x\log(1+1/x)$ is increasing in $x \geq 0$ and $\sum_{k\in S}\tau_k \leq 1$ by definition of $\{\tau_k\}$ and (c) follows form Jensen's inequality. From this series of inequalities we can derive some of the most peculiar characteristics of multi-user fading channels.

First of all, the RHS of inequality (b) coincides with the rate-sum achieved by rate vectors on the closure of the capacity region (1.9). This proves that, for any choice of the time-sharing coefficients, the TDMA achievable region is strictly inside the multiple-access capacity region. Hence, any access scheme that orthogonalises the users, in the time domain, in the frequency domain or in the code space, is suboptimal for a multi-user channel.

Second, the inequality in (c) holds with equality only for unfaded channels, i.e., $|c_k|^2$ is a degenerate random variable. Hence, without transmitter CSI, the fading can only decrease capacity with respect to the unfaded Gaussian channel with gain equal to $\mathrm{E}[|c_k|^2]$. The common believe that "fading is bad", and hence has to be compensated for, might come form this simple application of Jensen's inequality. This idea is one the common misconception based on inaccurate information-theoretic analysis. In fact, in order to compensate for the "deleterious effect" of fading, some form of power control must be undertaken at the transmitters. Power control can be effective only if the transmitters know the fading values they are assumed to compensate for. However, in that case of transmitter CSI, the capacity region is no longer given by (1.9) since its derivation is based on the assumption that the transmitter cannot track the channel.

Third, the only case where the three inequalities in (a), (b) and (c) hold with equality is for $S = \{1, \cdots, K\}$, for unfaded channel and for $\tau_j = (\mathrm{E}[|c_j|^2]P_j)/(\sum_{k=1}^{K}\mathrm{E}[|c_k|^2]P_k)$. Therefor, TDMA is optimal, in the sense that the boundary of the TDMA achievable region touches the boundary of the capacity region, in one single point and only in the unfaded case.

Forth, the use of the central limit theorem for a symmetric system, i.e., same power constraint $P_k = P$ and same fading statistics $c_k \sim f_c(x)$ for

all the users, allows to write the maximum rate-sum (equation (1.9) for $S = \{1, \cdots, K\}$) in the limit for large K as

$$\mathrm{E}\left[\log\left(1 + \frac{KP}{N_0}\sum_{k=1}^{K}\frac{1}{K}|c_k|^2\right)\right] \overset{K\to\infty}{\longrightarrow} \log\left(1 + \frac{KP}{N_0}\mathrm{E}[|c_k|^2]\right) \qquad (1.10)$$

thus proving that, as the number of users grows, the effect of fading is mitigated by the average effect of many users.

Capacity region of the constrained input Gaussian channel with receiver and transmitter CSI. We assume now that every transmitters knows the whole set of fading coefficients $\{c_1, \cdots, c_K\}$. In this setting, the transmit power of each user can be varied according to the channel condition. We search for the best power allocation $p_k(c_1, \cdots, c_K)$ such that $\mathrm{E}[p_k(c_1, \cdots, c_K] = P_k$ for all k.

In the single-user case, Goldsmith and Varaiya showed that the optimal power allocation is "waterfilling" in time [11]. In the single-user case, the availability of CSI at the transmitter in addition to the receiver gives little advantage in terms of average reliable transmitted rate, and this small advantage is in particular pronounced for low signal-to-noise ratio (SNR) values, where the unfaded Gaussian capacity $\log(1 + P/N_0)$ may be surpassed.

In the multi-user setting channel state information at the transmitters has a tremendous impact. The optimal transmission strategy that maximizes the rate-sum is a form of channel-state driven TDMA while, for the case of CSI at the receiver only, TDMA is strictly suboptimal. Indeed, in [12], Knopp and Humblet showed that the rate-sum is maximized by letting only the user enjoying the best channel to be active and allocate power according to the waterfilling law. In contrast to the single-user case [11], where optimal power control marginally increases the average rate, in the multiple-user case, the optimal power control gives a substantial growth in capacity which increases with the number of users K. The reason for this result is that if K is large, then with high probability at least one of the users have a very high channel gain. Such a channel is in fact advantageous even over the unfaded Gaussian channel with the same average power gain. Fading creates a form of diversity that is often referred to as "multi-user diversity".

The whole capacity region for the fading multiple-access channel was found by Tse and Hanly in [13]. By exploiting the polymatroid structure of the multiaccess Gaussian capacity region, they provided the characterization of the optimal power allocation that achieves the boundary points of the capacity region. The optimal policy is such that in every fading state, only the subset of user enjoying "good enough" channel are allowed to transmit. Power and rate are allocated such that the active users can be decoded

sequentially: the first user is decoded by treating all the other active users as noise, then its code-word is re-encoded and subtracted from the overall received signal, at this point the the second strongest user is decoded treating the remaining users as interference. The process continues this way until all the users are decoded. Interestingly, the decoding order is fading independent and the joint decoding process is based on single-user decoding and stripping. Also in this case Gaussian codebooks are optimal. The delay incurred by applying optimal power policy can be very long, longer than in the case of no transmitter CSI since here users are active only when the corresponding channel is "good". With optimal policy, there are not only problems of delay, but also of fairness especially for users with "bad" channel statistics.

Capacity region per unit cost. An interesting problem is to assign a non-negative cost to each symbol of the channel input alphabet and to find the maximum number of bits that can be reliably transmitted on the channel per unit cost. For the Gaussian channel considered so far, the cost is the power and the system has an average cost constraint in the form $\mathrm{E}[|X_k|^2] \leq P_k$. Verdú in [14] showed that the capacity per unit cost of a memoryless stationary channel is

$$\mathcal{U} = \bigcup_{P_k>0} \{(r_1, \cdots, r_K) \in \mathbb{R}_+^K : (r_1 P_1, \cdots, r_K P_K) \in C(P_1, \cdots, P_K)\}$$

where $C(P_1, \cdots, P_K)$ is the standard Shannon capacity region with average constraints $\mathrm{E}[|X_k|^2] \leq P_k$ for all k. In particular, for additive channels the computation of $\mathcal{U}$ boils down to the computation of the single-user capacity per unit cost. The single-user capacity per unit cost is the inverse of the minimum transmit energy per reliable information bit. Minimum transmit energy per reliable information bit. is of particular importance in the case of system working in low-power regime, also called wideband regime.

Description of other interesting results on multiple access channels can be found in [5] and references therein. Among those the characterization of the capacity region for channels with ISI [15], for channels with feedback [16] and for arbitrarily varying channels [4]. For the non-ergodic case, relevant "rate measures" are the outage capacity [2] and delay limited capacity [3]. We do not mention these results only because we shall not refer to them in the sequel of the work.

1.3.2 .. and what Information theory does not consider

From the overview on ergodic capacity we gave in the previous section, it is clear that the implementation of capacity achieving strategies in practical

systems performs poorly when average delay and fairness among users is concerned. Moreover, the model adopted for multiple-access channels assumes a fix number of users who transmit continuously. In other words, information theory has always neglected the bursty nature of sources and the role of delay and concentrated on the tradeoff rate-power, while communication network has mainly focused on random arrival and collision resolution [17, 18].

In common models for communication network, a user accesses the channel when it has a message to transmit. An attempt to analyze the case where only a subset of the whole population of users may be active, is the so called "L-out-of-K multiple-access channel" model [19]. Here, at most L out of potential K users are simultaneously active, and the achievable reliable rate region, irrespective of the identity of the active users, is of interest. This model is motivated in a sense by random-access aspects, but it does not capture the fact that the number of transmitting users might itself be random and not fixed. Rather, it can be thought as an upper bound to the number of active users and the derived region as a worst case achievable rate region.

To maintain a fixed finite delay over fading channels, two information-theoretic approaches have been considered: outage capacity [2] and delay limited capacity [3]. The first consider the code-word rate as a random variable which depends on the fading current value. If the rate is smaller the actual transmission rate then an error (outage) occurs, otherwise the error probability behaves according to the error exponent. In the second case, power control is used to invert the channel so that the receiver sees an unfaded channel so that transmission rate is keep constant. The drawback of this approach is that certain channels, like a Rayleigh fading channel, cannot achieve any positive rate with finite average transmit power.

A remarkable attempt to combine information theory and communication network is due to Telatar and Gallager [20]. In [20] the multi-user system is assimilated to a processor-sharing system and model as a single-server queue. The service time required by the users is determined by considering the error exponent. The stability region is derived as well as the average delay for several different SNR. There are a certain number of recent works that bring information theory concepts into the communication network community, in particular Berry and Gallager [21], and Bettesh and Shamai [22, 23]. These works treat the single user-case and are concerned with the minimization of the average buffer length, and hence the average delay, with a given average power constraint.

Although, an ever-growing number of researchers in the information theoretic community looks at communication network problems, the "union between those two fields remain unconsumed" [18].

1.4 Thesis outline

From what summarized in the previous two sections, it emerges that incorporating "practical" system features into information theoretic models is a challenging open problem. It was precisely the idea of addressing this problem what we had in mind when three years ago this work started. We were interested in studying of the impact of delay constraints on the performance of multiple-access channels with fading. We analyzed the problem from two almost complementary point of view and those two views are reflected in the division in two parts of this report.

Part I. In Part I, inspired by the tutorial paper of Ephremides and Hajek [18] and the pioneer work of Telatar and Gallager [20], we study a simple multi-access system where users access at random the channel whenever they have data to transmit. Due to random activity, users cannot be coordinated and hence the systems operates in a completely decentralized way. The senders transmit at constant power and constant rate, since no channel state information is available at the transmitters. The decoder performs single-user decoding (interfering users are treated as noise). In such a system, users retransmit negatively acknowledged packets until successful decoding take place or a time-out expires. The introduction of a time-out models the fact that information is delay-sensitive and hence, if a packet is not received within a given maximum delay, the information becomes useless. To improve decoding at the receiver side, previously received packets are not discarded, but are combined in order to improve decoding at the next decoding attempt.

Many works are available in literature on repetition protocols in conjunction with packet combining techniques. All of them analyze a particular co-decoding scheme and the related results may not be easily extended to systems with different parameters. Here, we use information theoretic notions, like random coding and typical set decoding, for the analysis of the the throughput of three different protocols. Therefore, our results are independent of a particular co-decoding scheme and must be looked at as limiting performance in the usual information-theoretic sense.

In particular:

- In Chapter 2 we show that typical set decoding has very desirable properties for Hybrid-ARQ, in the limit for large product bandwidth codeword duration. From a renewal-reward theory approach, we obtain closed-form throughput formulas for three simple protocols: a generalization of slotted Aloha (ALO), a repetition time diversity scheme with maximal-ratio packet combining (RTD) and an incremental redundancy scheme based on progressively punctured codes (INR).

  Then, we analyze the effect of delay and rate constraints on the through-

put, as well as the limiting behavior with respect to the slot spectral efficiency, the channel load and the transmit SNR. Interestingly, all three protocols are not interference-limited, and achieve arbitrarily large throughput by simply increasing the transmit power of all users. Furthermore, for an optimal choice of the transmission rate the INR protocol achieves the ergodic rate of the underlaying block-fading channel.

Publications related to this chapter are:
[24] G.Caire and D.Tuninetti, *"ARQ Protocols for the Gaussian Collision Channel"*, in Proceedings 2000 IEEE International Symposium on Information Theory (ISIT2000), Sorrento (Italy), June 2000;
[25] G.Caire and D.Tuninetti, *"The throughput of Hybrid-ARQ protocols for the Gaussian collision channel"*, in IEEE Transactions on Information Theory, Volume n.47 Issue n.5, July 2001, Pages 1971-1988.

- In Chapter 3 we compare the performance of systems that implements, at MAC layer, one of the three different Hybrid-ARQ protocols analyzed in Chapter 2 and emploies, at physical layer, different receiver/decoding strategies. The chosen performance measure is maximum throughput versus average energy per successfully received information bit. In particular, we consider either single-user decoding based systems, as the system introduce in Chapter 1 and a random spread DS-CDMA system, as well as systems based on joint decoding, as successive cancelation and full joint decoding. In carrying out the optimization of the throughput with respect to the various systems parameters we get insight into optimal of the transmission rate (parameter depending on the users) and of the average channel load (value that must be kept close to its optimum by an appropriate admission control system).

  We show that the unspread system outperforms SUMF DS-CDMA, which is throughput-wise limited, but it is outperformed by MMSE DS-CDMA. All the systems have the same behavior in terms of throughput and of optimal system parameters. In the low E_b/N_0 regime, the optimized throughput is the same for all the systems and coincides with that of a SUMF DS-CDMA, achieved by an infinite number of users per degree of freedom transmitting at vanishing rate. In the high E_b/N_0 regime, while SUMF DS-CDMA is interference limited, the other systems are not. For this range of E_b/N_0 the optimized systems "self-orthogonalize", in the sense that optimal throughput is achieved by having on the average only one user per degree of freedom, i.e., one user per chip for the DS-CDMA and one active user per slot for the unspread system. All the SUD-based systems are outperformed

by MUD-based systems.

Our publications related to this chapter are:
[26] D.Tuninetti and G.Caire, *"The optimal throughput of some wireless multi-access systems"*, in Proceedings 2001 IEEE International Symposium on Information Theory (ISIT2001), Washington DC (USA), June 2001;
[27] D.Tuninetti and G.Caire, *"The optimal throughput of some wireless multi-access systems"*, to appear in IEEE Transactions on Information Theory.

Part II. In the second part, we take a somewhat complementary point of view with respect to Part I. We consider a completely centralized and coordinated system, where users are active all time, know the channel state and allocate rate and power according to the channel state in order to be always inside the fading dependent capacity region. We assume that the fading dynamics is slow with respect to the tolerable decoding delay, and hence code-words are affected by a finite number of different fading states. In contrast to capacity achieving coding strategies for ergodic channel, or "classical" approaches to deal with fixed code-word duration, like outage analysis or delay limited analysis, where the desired performance is achieved by constant rate transmission, here we consider a variable rate coding scheme. Since the transmission rate is a random variable, we define and characterize the "long-term average capacity region" as well as its asymptotic behavior for increasing (laxer) delay. Due to causal nature of feedback, the solution is given in terms of Dynamic Programming algorithm.

In particular:

- In Chapter 4 we formally define our variable rate coding scheme and give a characterization of the boundary of the corresponding long-time average capacity region.

  We prove that long-term average capacity is achieved, for delay equal to one slot, by constant power allocation, while, when the delay constrain is relaxed, the optimal causal policy tends to the optimal ergodic policy without delay constraint and non-causal channel state information [13]. Moreover, our setting gives the correct trade-off between peak-to-average constrained systems and complete freedom in the power allocation. Furthermore, it proves that past and future channel knowledge are immaterial when the delay constraint is not too severe.

  In a system characterized by energy limitation at the transmitter, a sensible design criteria is to look at the long-term average capacity per unit energy. We show that in this case the optimal power policy is "one-shot", i.e., the optimal policy concentrates all the energy in only

one of the fading states. That state is chosen on the basis of not only its strength, but also how likely it is that a more favorable fading state will appear before the end of the code-word.

Our publications related to this chapter are:
[28] D.Tuninetti and G.Caire. *"The long-term average capacity region per unit energy"*, in the Proceedings of the Thirty-fifth Annual Asilomar Conference on Signals Systems and Computers (Asilomar2001), Pacific Grove (USA), November 2001;
[29] D.Tuninetti and G.Caire. *"The long-term average capacity region per unit energy with application to protocols for wireless sensor networks"*, in Proceedings of the 2002 European Wireless Conference (EW2002), Firenze (Italy), February 2002. Best Student Paper Award.

- Since in the low-energy regime or wideband regime, it is not enough to look at capacity per unit cost, in Chapter 5 we study the wideband slope of the spectral efficiency curve at the point of minimum energy per bit (the inverse of the capacity per unit energy). We extend the single-user wideband analysis to the multi-user case by introducing the notion of wideband slope region. We show that the "one-shot" policy, that achieves capacity per unit energy, is also optimal in the sense of wideband slope.

  Our publications related to this chapter are:
  [30] S.Verdú and G.Caire and D.Tuninetti, *"Is TDMA optimal in the low power regime?"*, in Proceedings of the 2002 IEEE International Symposium on Information Theory (ISIT2002), Lausanne (CH), June 2002;
  [31] D.Tuninetti and G.Caire and S.Verdú, *"Fading multiaccess channels in the wideband regime: the impact of delay constraint"*, in Proceedings of the 2002 IEEE International Symposium on Information Theory (ISIT2002), Lausanne (CH), June 2002;
  [32] D.Tuninetti and G.Caire and S.Verdú, *"The impact of delay constraint and causal feedback on the wideband performance of block-fading multiple-access channels"*, submitted to IEEE Transactions on Information Theory, February 2002.

Chapter 6 concludes the report by briefly summarizing the main results our our work.

Part I

On the throughput analysis of ARQ protocols for completely uncoordinated decentralized systems with delay constraints

Chapter 2

Retransmission protocols for multi-user channels

In this first part of the thesis, we take an information-theoretic view of some simple protocols for reliable packet communication based on Hybrid-ARQ (Automatic Retransmission reQuest) over a slotted multi-user channel with noise and fading and study the system throughput and average delay. As an application of the Renewal-Reward theorem, we obtain closed-form throughput formula and then we consider its optimization with respect to the various system parameters. Since random coding and typical set decoding are assumed throughout the whole work, our results are independent of the particular coding/decoding technique and should be regarded as limit performance of the system in the information theoretic sense. We conclude the chapter with some considerations on practical implementation of Hybrid-ARQ strategies.

2.1 Introduction

In order to support new services (e.g., wireless mobile access to the Internet), next generation wireless communication systems will implement packet-oriented data transmission in addition to standard mobile telephony [33]. This implies bursty sporadic communication from a large population of users, that may require instantaneous large data rates and very small error probabilities for a short time. Motivated by the above consideration, we take an information-theoretic view of some simple protocols for reliable packet communication based on "Hybrid-ARQ", i.e., on combining *channel coding* and *Automatic Retransmission reQuest* (ARQ) [34, 35].

As remarkably well illustrated by Ephremides in [18], information theoretic techniques are not yet of widespread use in the domain of networking. Steps in this direction are represented by the work of Shamai and Wyner on cellular systems [36, 37] and of Telatar and Gallager [20]. In [20], the multiple access Gaussian channel is assimilated to a processor-sharing system and is analyzed as a queue with single server and an infinite buffer length. The required service time for each user is defined in terms of random coding bound on the error probability. A code-independent analysis of the mean transmission duration is obtained as an application of Little's Theorem [38].

Our work is mainly inspired by [20]. We study a system where users transmit their signal *bursts* in a completely uncoordinated way (user random activity) and where the transmission is governed by an Hybrid-ARQ protocol, designed to cope with background noise, fading and interference from other users (or "collisions", following the terminology introduced in [39]).

In packet data transmission two techniques are commonly used to control transmission errors: Forward Error Correction (FEC) and Automatic Retransmission reQuest (ARQ). With FEC, channel coding is used with the purpose of correcting the errors introduced by the channel before delivering, irrespectively of whether the errors have been successfully corrected, the packet to the end-user/application. With ARQ, a code is used to detect errors. When a packet is detected in error, the transmitter is informed via a feedback channel of the transmission failure and it is asked to retransmit the same packet. Retransmissions go on until the packet is positively ACKnowledged (ACK) by the receiver. The choice between these two strategies is dictated by the system constraints. ARQ is simple, provides high reliability (low probability of decoding error) but the throughput (number of information bits successfully delivered per unit of time) is not generally high and the latency (time interval between packet generation and its successful decoding) is large due to the repetition mechanism. On the other hand, FEC has a constant throughput, since a message is always delivered, but the reliability is quite low since it is possible that a wrong message is passed to the end-user [34]. For these reasons the two strategies are employed in different contest: FEC is used when a feedback channel is not available, delay requirements are strict and the probability of error needs not be very small, while ARQ is used when delay constraints are not so stringent but a small error rate is required.

FEC and ARQ can be combined together in what is called Hybrid-ARQ. In literature, several types of Hybrid-ARQ protocols have been proposed (see [34, 35] and references therein): Type I Hybrid-ARQ uses one code to detect and simultaneously correct errors, always the same packet is retransmitted; Type II Hybrid-ARQ uses two codes: one high rate code to detect errors and one low rate code to detect and correct errors, information and redundancy are alternatively retransmitted but only the last two received packets are taken into account for decoding; Type III Hybrid-ARQ

uses a code like Type II but only the redundancy part is retransmitted and the previously received packets are combined together in order to generate a code with lower rate. Packet combining can be based on hard decisions [40, 41, 42, 43, 44] or on soft channel outputs [45, 46, 47, 48], e.g., maximal-ratio, equal-gain or selection combining.

Type III Hybrid-ARQ can be generalized by generating several different redundancy packets from the same information bits and by sending them at each retransmission request. For example, soft decoding of maximal-ratio combined packets can be seen as an elementary form of generalized Type III Hybrid-ARQ, based on soft-decoding of a repetition code of variable length. Example of generalized Type III Hybrid-ARQ that employ different codes are [49, 47, 50]. In [49], a family of codes, called Rate Compatible Punctured Convolutional Codes, is designed so that all code bits of any code of the family are used by all lower rate codes. Transmission starts with the higher rate code and further coded bits are provided whenever necessary by using lower rate codes. Since each high rate code is part of a lower rate code, all codes can be decoded with the same decoder. In [47, 50] Compatible Punctured Convolutional Codes are introduced. All the codes of the family are derived from the same mother code, all have the same rate, the same distance property and give the original mother code when combined together. At each transmission the decoder tries to recover the information message from the last received packet and only in the case where a non-correctable error is detected, the last received packet is combined with the previous ones. More recently, Turbo-codes [51] have been suggested as candidates for packet combining, exploiting the fact that they are systematic and produce incremental redundancy by puncturing the parity bits [52] or by changing the interleaver [53].

Analysis of Hybrid-ARQ protocols in terms of throughput, error rates and delay can be found in [54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64]. Most works carry out a "separated analysis", i.e., consider a completely symmetric system with respect to any user, and study the behavior of the protocol for a particular reference user modeled as a Markov chain. In general, analysis depends on the type of codes and decoding/error-detection technique employed. Modeling the system as Markov chain might be complicated, since in each state one must convey all the information about the memory of the system. In [65], Zorzi proposes the use of *renewal theory* [66] in order to analyze ARQ protocols.

In this chapter we take an information theoretical view of some retransmission protocols in a scenario characterized by user random activity and time-sensitive information which imposes a maximum decoding delay. We assume that users transmit their signal in a completely uncoordinated way and access the channel at random, like in slotted Aloha systems [38]. Even though any point in the capacity region of multiple access channels can be implemented with low complexity by successive "stripping" [67], this re-

quires a good deal of coordination among the users [13, 3, 68] that may not be suited to random user activity, hence we assume that the receiver is formed by a bank of single-user decoders, and does not implement joint decoding, i.e., each decoder treats the signals from other users as noise. We study the system performance in terms of throughput and average delay for three simple idealized protocols: a coded version of slotted Aloha (Type I Hybrid-ARQ with user random activity), a repetition scheme with maximal-SINR (Signal to Interference plus Noise Ratio) packet combining (Type III Hybrid-ARQ with user random activity) and an incremental redundancy scheme with general coding (generalized Type III Hybrid-ARQ with user random activity). By applying the *renewal-reward* theorem [66], we obtain a closed-form throughput formula under a delay constraint and code rate constraint for a completely symmetric system. Since we consider random coding and typical set decoding, our results are independent of the particular coding/decoding technique and should be regarded as a limit in the information theoretic sense.

We derive closed form throughput formulas for all the three protocols, then we carry out their optimization with respect to the different system parameters. Interestingly, we show that the ARQ system is not interference-limited even if no multi-user detection or joint decoding is used, i.e., arbitrarily high throughput can be obtained simply by increasing the transmit power of all users, as opposed to conventional CDMA where the throughput tends to a finite limit as all users increase their transmit power [69, 70]. As a byproduct of this analysis, we provide a stronger operational meaning to the information outage probability of block-fading channels and we obtain the closed form probability distribution of SINR with Rayleigh fading and a Binomial- and Poisson-distributed number of interferers, extending the result of [37].

The rest of the chapter is organized as follows: in Section 2.2 we describe the system model; in Section 2.3 we deal with typical set decoding and error detection; in Section 2.4 we derive the system throughput and we prove that our Hybrid-ARQ strategy is not interference limited; in Section 2.5 we find the throughput for an unconstrained system; in Section 2.6 we present the optimization with respect to the transmission rate and in Section 2.7 we discuss some issues about the practical implementation of the proposed Hybrid-ARQ protocols. The proofs of the results are provided in the Appendices at the end of the chapter.

Our publications related to this chapter are:
[24] G.Caire and D.Tuninetti, *"ARQ Protocols for the Gaussian Collision Channel"*, in Proceedings 2000 IEEE International Symposium on Information Theory (ISIT2000), Sorrento (Italy), June 2000;
[25] G.Caire and D.Tuninetti, *"The throughput of Hybrid-ARQ protocols for*

the Gaussian collision channel", in IEEE Transactions on Information Theory, Volume n.47 Issue n.5, July 2001, Pages 1971-1988.

2.2 The slotted Gaussian channel with feedback

In the system under investigation, K users share a common radio channel of bandwidth W in order to transmit their information messages to a common receiver. Users are provided with a common time reference. The time axis is divided in *slots* of duration T and users transmit signal *bursts* of duration slightly less than T, aligned with the slots. Apart from the slotted transmission mode, users are completely uncoordinated. Each user can transmit about $L = \lfloor WT \rfloor$ independent complex symbols over one slot (assuming $WT \gg 1$ [1] [1]). The received signal over slot s can be written as

$$\boldsymbol{y}_s = \sum_{k \in \mathcal{K}(s)} c_{k,s} \boldsymbol{x}_{k,s} + \boldsymbol{\nu}_s \tag{2.1}$$

where $\mathcal{K}(s) \subseteq \{1, \ldots, K\}$ denotes the set of *active* users over slot s, $\boldsymbol{\nu}_s$ is a proper complex Gaussian random vector of dimension L with i.i.d. (independent and identically distributed) components of zero mean and variance N_0, $\boldsymbol{x}_{k,s}$ is the complex signal of user k transmitted in slot s with constant average energy $E_k \stackrel{\Delta}{=} \mathrm{E}[|\boldsymbol{x}_{k,s}|^2/L]$, $c_{k,s}$ is the complex fading coefficient for user k assumed constant (block-fading model [2]) over the whole slot with instantaneous power $\alpha_{k,s} \stackrel{\Delta}{=} |c_{k,s}|^2$, i.i.d. for all s and k, with cdf (cumulative distribution function) $F_\alpha(x)$. For finite L no positive rate is achievable. However, we can consider a sequence of channels indexed by the slot length L and study the achievable rates in the limit for $L \to \infty$. This is a standard mathematical abstraction in the study of the limit performance of block-fading channels [2] and it is motivated by the fact that, in many practical applications, the product WT is large and T is much smaller than the fading coherence time. [2]

User k encodes its information messages, of b bits each, independently of other users, by using a channel code with code book $\mathcal{C}_k \subset \mathbb{C}^{LM}$ of length LM over the complex numbers, where M is a given integer. Code-words are divided into M sub-blocks of length L, each of which is modulated into a

[1] For large WT, a complex symbol (or dimension) can be transmitted approximately in one second and one Hz. More precisely, the spectral efficiency expressed in bit/s/Hz can be obtained by multiplying the coding rate (bit/complex-symbol) by the modulation spectral efficiency (expressed in complex-symbols/s/Hz), that depends on the modulation excess bandwidth [71].

[2] For example, in the 3rd generation UMTS standard a packet-radio random access scheme is supported with variable slot duration $0.625 \leq T \leq 10$ ms, bandwidth $W = 5$ MHz [72] and modulation spectral efficiency up to 0.2, obtained by using direct-sequence spread-spectrum modulation with raised-cosine pulses with roll-off 0.22. This means that $L = \lfloor 0.2WT \rfloor$ is between 625 and 10000 complex symbols per slots.

signal burst and is transmitted over one slot. We let $\mathcal{C}_{k,m}$, for $m = 1, \ldots, M$, denote the punctured code of length mL obtained from $\mathcal{C}_k$ by deleting the last $M - m$ sub-blocks.

Each user selects the slots for transmission according to its own *time-hopping* random sequence, independently of the other users [73]. Time-hopping sequences can be seen as random "on-off" processes, where a user can transmit only when it is "on". We assume that the receiver knows *a priori* the time-hopping rule of all users in the system [73]. [3] Transmission is governed by the following retransmission protocol, run in a decentralized way by each user k. When a new code-word is ready for transmission, user k sends the first L symbols on the first allowed slot, say s_1, according to its time-hopping rule. The receiver decodes the code $\mathcal{C}_{k,1}$ by processing the received signal $\boldsymbol{y}_{s_1}$. If decoding is successful, a positive acknowledgment (ACK) is sent back to user k over an error-free and delay-free feedback channel and the transmission of the current code-word stops. On the contrary, if the receiver detects an error, a negative acknowledgment (NACK) is sent. In this case, user k sends the second block of L symbols of the same code-word on the next allowed slot, say s_2. Now, the receiver decodes the code $\mathcal{C}_{k,2}$ by processing the received signal blocks $\{\boldsymbol{y}_{s_1}, \boldsymbol{y}_{s_2}\}$. Again, if decoding is successful an ACK is sent and the transmission of the current code-word stops. On the contrary, if a decoding error is detected, a NACK is sent back and user k transmits the third block of L symbols of the same code-word on the next allowed slot. The process goes on this way: after the transmission of m bursts of the current code-word, code $\mathcal{C}_{k,m}$ is decoded by processing the received signal $\{\boldsymbol{y}_s : s \in \mathcal{S}_{k,m}\}$, where $\mathcal{S}_{k,m} \stackrel{\Delta}{=} \{s_1, \ldots, s_m\}$ denotes the sequence of slots where transmission of user k took place. If successful decoding occurs at the m-th transmission, the effective coding rate for the current code-word is R/m bits/s/Hz, where $R \stackrel{\Delta}{=} b/L$. In the sequel we shall refer to R as *information rate*.

In general, the slots $s \in \mathcal{S}_{k,m}$ are non-adjacent. We let n denote the delay (expressed in number of slots) between the instant where a code-word is generated and the current time (time ticks at the slot rate). Obviously, $m \leq n$ (see Fig. 2.1). In any practical application, an information message must be delivered to the receiver within a maximum delay of N slots, where for simplicity N is assumed common to all users and all messages. If successful decoding does not occur within delay N, the message becomes useless. Moreover, since the code-words of $\mathcal{C}_k$ have M sub-blocks, the same message can be transmitted in at most M signal bursts. If successful decoding does

[3] This assumption is not particularly restrictive, and is analogous to the standard assumption of CDMA with pseudo-random "long" spreading [74], where the receiver is assumed to know the spreading sequences of all users it wishes to decode. In practice, we might think of an access mechanism, run at a much slower time-scale than packet transmission, that assigns to new users entering the system a time-hopping sequence.

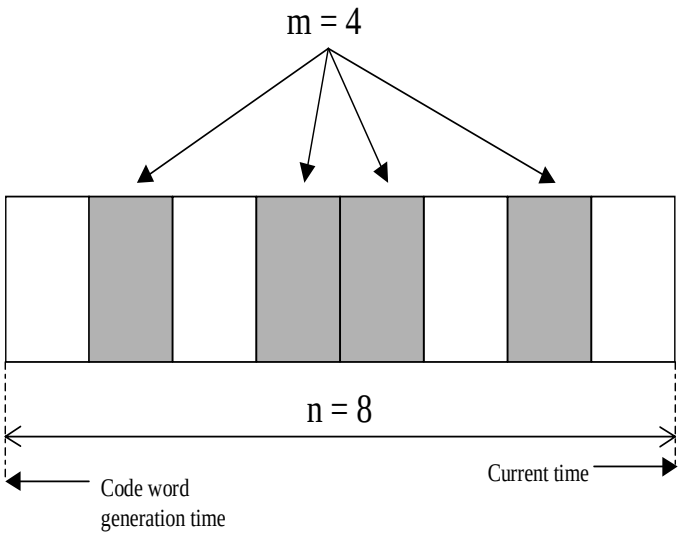

Figure 2.1: Example: $m = 4$ transmitted bursts (shadowed) over $n = 8$ slots since the current code-word generation.

not occur within M transmitted bursts, the message is lost. We shall refer to N and M as the "delay" and "rate" constraints, respectively. The transmission of a code-word can stop in three cases: i) Successful decoding occurs at the m-th transmitted burst and in n slots, with $m \leq M$ and $n \leq N$; ii) No successful decoding occurs after M transmitted bursts and $n \leq N$ slots; iii) No successful decoding occurs after N slots and $m \leq M$ transmitted bursts.

There are several ways to handle transmission failures (cases (ii) and (iii) above). For example, in the case of time-sensitive information, the current message is simply discarded. In other applications, delay is not a strict requirement (N is very large) and the current message may be kept in the transmission buffer for a later attempt. Several practical ARQ protocols have been proposed to handle transmission failures (see references in Section 2.1). The analysis carried out in the following considers the simplified scenario where an infinite sequence of messages is available to all users and, in the case of transmission failure, the current message is discarded and the next message is encoded and transmitted in exactly the same way. Since in our model there is not a packet arrival process, we are not concerned with instability and input buffer overflow, typical "problems" of classical analysis of Aloha protocols. The time-hopping sequence for slot selection is not modified by transmission failures (e.g., there is no idle state, waiting for better channel conditions). It is important to notice that each

user runs its own ARQ protocol independently of the other users. The only way in which users influence each others is through mutual interference that occurs when several users transmit their bursts over the same slot.

The single-user decoder for user k has perfect knowledge of the channel gain $\{\alpha_{k,s} : s \in \mathcal{S}_{k,m} \ , \ k \in \mathcal{K}(s)\}$

$$\beta_{k,s} \stackrel{\Delta}{=} \frac{\alpha_{k,s} E_k}{N_0 + \sum_{j \in \mathcal{K}(s): j \neq k} \alpha_{j,s} E_j} \tag{2.2}$$

Estimation of the channel gains can be accomplished in practice with high reliability by inserting training symbols into each signal burst, as currently done in most CDMA and TDMA cellular standards [75], at the price of a slight rate loss.

The ARQ protocol described above is a *general incremental redundancy scheme* (denoted by "INR" for brevity). We consider also the following particular cases.

Generalized Slotted Aloha. The slotted Aloha protocol [38] (denoted by "ALO" for brevity) is obtained by assuming for each user k a suboptimal decoder that considers only the last received signal block. In classical slotted Aloha it is assumed that a decoding failure occurs (and is detected) whenever a collision occurs. In mobile systems, users might be received at very different power levels because of fading, shadowing and different distances from the receiver. In this case, a packed can be decoded successfully even if a collision occurs (capture effect) [64]. Here, we consider a *generalized* ALO where channel coding is used and messages may be decoded correctly even in the presence of collisions, depending on the SINR.

Repetition time-diversity. A simple time-diversity scheme (denoted by "RTD" for brevity) is obtained by repeating the same burst of L symbols [45, 46] randomly interleaved at each retransmission. This is equivalent to construct the user code $\mathcal{C}_k$ as a concatenated code, where $\mathcal{C}_{k,1} \subset \mathbb{C}^L$ is the outer code and a simple repetition code of length M is the inner code. After the m-th transmission, the receiver performs maximal-ratio combining [45] of the de-interleaved signals and decodes the outer code $\mathcal{C}_{k,1}$ based on the combined signal.

2.3 Coding, decoding and error detection

We assume that all code books $\mathcal{C}_k$ are generated randomly and independently, with i.i.d. components, according to a given pdf $q(x)$ over $\mathbb{C}$ with mean zero and variance E_k. For each user, an encoding function $\psi_k : \{1, \ldots, \mathrm{e}^{RL}\} \to \mathcal{C}_k$ is defined and revealed to the receiver.

A key point of the ARQ schemes described in Section 2.2 is that decoding errors should be detected. Any *complete decoding* function, based on a parti-

tion of the channel output space into e^{RL} regions (e.g., MAP decoding or ML decoding), is not suited to this purpose, unless an explicit error-detection stage after channel decoding is introduced (e.g., in many cellular systems a CRC is inserted into the information message [35, 75]). This however is undesirable since it decreases the throughput by adding extra redundancy. An alternative is the use of possibly suboptimal decoders in terms of error probability, but featuring a built-in error detection capability. Moreover, it is desirable to decode all punctured codes $\mathcal{C}_{k,m}$, for $m = 1, \ldots, M$, by the same decoder.

In particular, we examine the following error correction/detection scheme. Consider decoding for user k after m received blocks, and let $\boldsymbol{x}_k^{(w)} = \psi_k(w)$ be the transmitted code-word corresponding to information message w. The decoder adds to the received signal $\{\boldsymbol{y}_s : s \in \mathcal{S}_{k,m}\}$ other $M - m$ dummy signal blocks $\boldsymbol{z}_i$, generated independently of the received signal, [4] to form the observation $\mathbf{Y} = (\boldsymbol{y}_{s_1}, \ldots, \boldsymbol{y}_{s_m}, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_{M-m})$ of length LM, and then decodes the "mother code" $\mathcal{C}_k$ according to the typical set rule $\phi_k : \mathbb{C}^{LM} \to \{1, \ldots, \mathrm{e}^{RL}, e\}$ (see [1] and Appendix 2.A for details) defined as follows:
Let $\mathcal{E}_w$ be the event that $\boldsymbol{x}_k^{(w)}$ is the unique code-word jointly typical with $\mathbf{Y}$. Then,

- $\phi_k(\mathbf{Y}) = \widehat{w}$ if, for some $\widehat{w} \in \{1, \ldots, \mathrm{e}^{RL}\}$, the event $\mathcal{E}_{\widehat{w}}$ occurs.
- $\phi_k(\mathbf{Y}) = e$ in any other case.

Since decoder k treats all other user signals as additive noise, it "sees" a virtual additive noise channel given by

$$\boldsymbol{y}_s = c_{k,s}\boldsymbol{x}_{k,s} + \boldsymbol{v}_{k,s} \tag{2.3}$$

where

$$\boldsymbol{v}_{k,s} = \boldsymbol{\nu}_s + \sum_{j \in \mathcal{K}(s): j \neq k} c_{j,s}\boldsymbol{x}_{j,s} \tag{2.4}$$

is the interference plus noise vector. We let $p_{k,s}(y|x)$ denote the single-letter transition pdf of the above channel (2.3), conditioned on the channel gains $\{c_{j,s} : j \in \mathcal{K}(s)\}$ and on the set of active users $\mathcal{K}(s)$, and we define $I(q(x), p_{k,s}(y|x))$ to be the mutual information (per letter) of channel (2.3), expressed as a functional of the pdfs $q(x)$ and $p_{k,s}(y|x)$. Obviously, $I(q(x), p_{k,s}(y|x))$ varies randomly from slot to slot, since it depends on the random set $\mathcal{K}(s)$ and on the random channel gains $\{c_{j,s} : j \in \mathcal{K}(s)\}$.

[4] In practice, in decoding of punctured convolutional codes dummy symbols are set to zero, but in the limiting case considered here it is sufficient that they are statistically independent of the channel input.

We examine the behavior of codes $\mathcal{C}_{k,m}$ with decoder ϕ_k defined above, for a given sequence of channel transition pdfs $\mathcal{P} \triangleq \{p_{k,s}(y|x) : s \in \mathcal{S}_{k,m}\}$. The average error probability is defined by

$$\Pr(\text{error}|\mathcal{P}, \mathcal{C}_k) \triangleq \mathrm{e}^{-RL} \sum_{w=1}^{\mathrm{e}^{RL}} \Pr(\overline{\mathcal{E}_w}|w, \mathcal{P}, \mathcal{C}_k)$$

A decoding error when message w is transmitted is not detected if, for some $\widehat{w} \neq w$, the event $\mathcal{E}_{\widehat{w}}$ occurs. Then, the average probability of undetected error is defined by

$$\Pr(\text{undetected error}|\mathcal{P}, \mathcal{C}_k) \triangleq \mathrm{e}^{-RL} \sum_{w=1}^{\mathrm{e}^{RL}} \Pr\left(\bigcup_{\widehat{w} \neq w} \mathcal{E}_{\widehat{w}} \middle| w, \mathcal{P}, \mathcal{C}_k \right)$$

The following results, proved in Appendix 2.A, show that the typical set decoder defined above is asymptotically optimal for both error and undetected error probabilities, for large burst length L:

Lemma 1 (achievability). For all $\epsilon > 0$ there exist L and codes $\mathcal{C}_k \in \mathbb{C}^{LM}$ of size e^{RL} with

$$\Pr(\text{error}|\mathcal{P}, \mathcal{C}_k) < \epsilon$$

for all $m = 1, \ldots, M$ and channel sequences $\mathcal{P}$ such that

$$\sum_{s \in \mathcal{S}_{k,m}} I(q(x), p_{k,s}(y|x)) > R \qquad \diamondsuit$$

Lemma 2 (converse). For all channel sequences $\mathcal{P}$ such that

$$\sum_{s \in \mathcal{S}_{k,m}} I(q(x), p_{k,s}(y|x)) < R$$

then

$$\Pr(\text{error}|\mathcal{P}, \mathcal{C}_{k,m}) \to 1$$

for any code $\mathcal{C}_{k,m} \in \mathbb{C}^{Lm}$ of size e^{RL} as $L \to \infty$. $\diamondsuit$

Lemma 3 (error detection). For all $\epsilon > 0$ and channel sequences $\mathcal{P}$ there exists L such that any code $\mathcal{C}_k \in \mathbb{C}^{LM}$ of size e^{RL} satisfies

$$\Pr(\text{undetected error}|\mathcal{P}, \mathcal{C}_k) < \epsilon \qquad \diamondsuit$$

The optimal input distribution $q(x)$ of the interference channel (2.3) is not known in general [1]. For the sake of mathematical tractability, we con-

sider (somewhat arbitrarily) proper complex Gaussian inputs for all users. [5] Then, the mutual information takes the form

$$I_{k,m} \triangleq \sum_{s \in \mathcal{S}_{k,m}} \log(1+\beta_{k,s}) \tag{2.5}$$

From the above results we have that, by using Gaussian codes and typical set decoding at each step m of the ARQ protocols of Section 2.2, the probability of decoding error is arbitrarily small if $R < I_{k,m}$, very large if $R > I_{k,m}$ and decoding errors are detected with arbitrarily large probability, for sufficiently large L. Practical future system for mobile data transmission will be characterized by a very large value of the product WT, in order to support large instantaneous bit rates. This motivates a system analysis under the assumption of very large L. In this regime, we shall assume that, for all k and m, $\Pr(\text{error}|R < I_{k,m}) = 0$, $\Pr(\text{error}|R \geq I_{k,m}) = 1$ and $\Pr(\text{undetected error}) = 0$.

In analogy to what done above for the INR scheme, we can define random coding and typical-set decoding for ALO and RTD. For the sake of brevity, we state without details that, as $L \to \infty$, also for these schemes there exist codes for which $\Pr(\text{error}|R < I_{k,m})$ and $\Pr(\text{undetected error})$ vanish and $\Pr(\text{error}|R \geq I_{k,m})$ goes to 1, provided that the correct expression for the mutual information is used. In RTD, the received signal (2.3) takes the form $\boldsymbol{y}_s = c_{k,s}\boldsymbol{\Pi}_{k,s}\boldsymbol{x}_k + \boldsymbol{v}_{k,s}$, since at each retransmission request the same codeword $\boldsymbol{x}_k$ is sent after being randomly interleaved with the random permutation matrix $\boldsymbol{\Pi}_{k,s}$. The deinterleaved received vector seen by the k-th decoder on slot s is then $\boldsymbol{\Pi}_{k,s}^{-1}\boldsymbol{y}_s = c_{k,s}\boldsymbol{x}_k + \boldsymbol{\Pi}_{k,s}^{-1}\boldsymbol{v}_{k,s}$, for $\boldsymbol{v}_{k,s}$ being the interference due to the other active users on slot s and the background noise (see definition in (2.4)). Assume now that user k was active on slots s and slots ℓ, i.e., $\{s,\ell\} \subseteq \mathcal{S}_{k,m}$ for some $m \geq 2$, as well as another user j, i.e., $\{k,j\} \subseteq \mathcal{K}(s) \cap \mathcal{K}(\ell)$. Now, the covariance matrix of the j-th user signals $\boldsymbol{x}_{j,s}$ and $\boldsymbol{x}_{j,\ell}$ is zero only if on slot ℓ user j was not attempting to retransmit codeword $\boldsymbol{x}_{j,s}$. In case of retransmission, the two signals are identical, i.e., $\boldsymbol{x}_{j,s} \equiv \boldsymbol{x}_{j,\ell}$. Hence, $\mathrm{E}[\boldsymbol{x}_{j,s}\boldsymbol{x}_{m,\ell}^H] = E_j \mathbf{I}_L \, 1\{\boldsymbol{x}_{j,s} \equiv \boldsymbol{x}_{j,\ell}\}$ where $\mathbf{I}_L$ is the $L \times L$ identity matrix. With the random permutation $\boldsymbol{\Pi}_{k,s}$, the vectors $\boldsymbol{\Pi}_{k,s}^{-1}\boldsymbol{v}_{k,s}$ for all $s \in \mathcal{S}_{k,m}$ are independent in the limit for large L. [6] This can be easily

[5] From the point of view of a practical implementation, it is interesting to access the limiting performance of Hybrid-ARQ schemes with input alphabets of finite cardinality. Lemmas 1, 2 and 3 remain valid provided that the correct expression for the mutual information is used. For example, in [76] the authors replicate the results of [24] in the case of codes mapped over BPSK by considering the mutual information of symmetric binary-inputs block-fading AWGN channel [77] instead of (2.5).

[6] The mutual information $I(\boldsymbol{y};x)$ between a scalar Gaussian input x of variance σ_x^2 and a vector of observations $\boldsymbol{y} = \boldsymbol{c}x + \boldsymbol{v}$, where the noise $\boldsymbol{v}$ is independent of x and Gaussian with covariance matrix $\Sigma_v = \mathrm{E}[\boldsymbol{v}\boldsymbol{v}^H]$, is given by $I(\boldsymbol{y};x) = \log(1+\sigma_x^2 \boldsymbol{c}^H \Sigma_v^{-1}\boldsymbol{c})$, assuming the vector $\boldsymbol{c}$ known. If Σ_v is diagonal then $I(\boldsymbol{y};x) = \log\left(1+\sigma_x^2 \sum_i \frac{|c_i|^2}{[\Sigma_v]_{i,i}}\right)$.

seen by

$$
\begin{aligned}
&\mathrm{E}[\boldsymbol{\Pi}_{k,s}^{-1}\boldsymbol{v}_s\boldsymbol{v}_\ell^H\boldsymbol{\Pi}_{k,\ell}^{-1}] \\
&= \mathrm{E}\left[\boldsymbol{\Pi}_{k,s}^{-1}\left[\sum_{j\neq k\,:\,j\in\mathcal{K}(s)} c_{j,s}\boldsymbol{\Pi}_{j,s}\boldsymbol{x}_{j,s}+\boldsymbol{\nu}_s\right]\left[\sum_{m\neq k\,:\,m\in\mathcal{K}(\ell)} c_{m,\ell}^*\boldsymbol{x}_{m,\ell}^H\boldsymbol{\Pi}_{m,\ell}^H+\boldsymbol{\nu}_\ell^H\right]\boldsymbol{\Pi}_{k,\ell}^{-H}\right] \\
&= N_0\,1\{s=\ell\}\,\mathbf{I}_L+\mathrm{E}\left[\boldsymbol{\Pi}_{k,s}^{-1}\sum_{j\neq k\,:\,j\in\mathcal{K}(s)\cap\mathcal{K}(\ell)} c_{j,s}c_{j,\ell}^*\left(\boldsymbol{\Pi}_{j,s}\boldsymbol{x}_{j,s}\boldsymbol{x}_{j,\ell}^H\boldsymbol{\Pi}_{j,\ell}^H\right)\boldsymbol{\Pi}_{k,\ell}^{-1}\right] \\
&= \begin{cases} N_0\,\mathbf{I}_L+\sum_{j\neq k\,:\,j\in\mathcal{K}(s)}|c_{j,s}|^2E_j\,\mathbf{I}_L & s=\ell \\ \sum_{j\neq k\,:\,j\in\mathcal{K}(s)} c_{j,s}c_{j,\ell}^*E_j\,1\{\boldsymbol{x}_{j,s}=\boldsymbol{x}_{j,\ell}\}\,\frac{1}{L}1_L & s\neq\ell \end{cases}
\end{aligned}
$$

since $\mathrm{E}[\boldsymbol{\Pi}]=1/L\,1_L$ where 1_L is the $L\times L$ matrix made of all 1. Therefore, the mutual information between $\boldsymbol{x}_k$ and the vector $\boldsymbol{y}_{\mathcal{S}_{k,m}}=[\boldsymbol{\Pi}_{k,s}^{-1}\boldsymbol{y}_s : s\in\mathcal{S}_{k,m}]$ in the limit for large L is given by

$$I_{k,m}=\log\left(1+\sum_{s\in\mathcal{S}_{k,m}}\beta_{k,s}\right) \tag{2.6}$$

ALO takes into account only the most recent received signal burst, therefore the corresponding $I_{k,m}$ is given by

$$I_{k,m}=\log(1+\beta_{k,s_m}) \tag{2.7}$$

Remark: Bounded distance and iterative decoding. Obviously, the typical set decoder considered above is not suited for practical implementations. However, it is interesting to notice that some non-ML practical decoding schemes show a behavior similar to the typical-set decoder. For example, bounded-distance decoding [78] outputs the message w if the received signal falls inside a *sphere* centered on the code-word corresponding to w, while if the received signal is not in any sphere, an error message e is declared. Another example is provided by the iterative decoding scheme [79] used to decode Turbo-codes. The component codes of the Turbo-code are individually decoded by symbol-by-symbol soft-in soft-out decoders sharing and updating some common information about the reliability of the symbol-wise decisions. Typically, if the code-word is correctly decoded all component decoders agree on the symbol-wise decisions, while in the presence of decoding errors the decoders keep on reversing the symbol decisions at each iteration [80]. This ill behavior, as well as the low reliability for some symbols, can be used as error indicators [81].

Remark: analogy with the block-fading channel. Under the assumption of Gaussian user code made here, the channel model(2.3) is totally analogous to the block-fading AWGN channel with perfect channel state information at the receiver introduced in [2]. In [2], decoding is always performed after M blocks and the probability of decoding failure for large L is

given by $\Pr(I_{k,M} \leq R)$, and is referred to as *information outage probability*. Outage probability finds a very natural interpretation as the limiting error probability for large block length averaged over the random coding ensemble and over the fading states [82]. A question left open in [2] and in many subsequent works is whether it exists a code sequence (for increasing values of the block length L) with error probability arbitrarily small *for all* fading states such that $I_{k,M} > R$. Notice that this is not a trivial question, since if the choice of the code sequence depends on the particular fading state, outage probability would not be achievable (it would require side information at the transmitter). The existence of codes *asymptotically good* for all fading states satisfying $I_{k,M} > R$ is given by Lemma 1 (see the details of the proof in Appendix 2.A). In this respect, information outage probability is not just an average probability of error over a code ensemble, but it can be approached by a given (deterministic) sequence of codes.

2.4 Throughput analysis

In this section we compute the throughput of the ARQ protocols of Section 2.2 with the coding and decoding scheme of Section 2.3, in the limit for large L. Our analysis is valid under the following idealized assumptions:

1. An infinite number of information messages is available for each users. As soon as a user stops the transmission of the current code-word, it encodes the next packet and starts its transmission in the next selected slot. As explained in Section 2.2, transmission of a code-word can stop either because successful decoding occurs, or because the delay or rate constraints N and M are violated (decoding failure).

2. The feedback channel is delay-free and error-free.

3. Users select slots for transmission so that the number of slots between two consecutive transmissions of the same user is i.i.d., geometrically distributed with identical parameter p_t for all users. In order words, on each slot s each user transmits a signal burst with probability p_t and does not transmit with probability $1 - p_t$. The expected number of users transmitting over a slot is $G = p_t K$ (average channel load).

4. The system is completely symmetric with respect to any user: all users have the same transmit SNR $\gamma \stackrel{\Delta}{=} E/N_0$, i.e., $E_k = E \ \forall k = 1, \ldots, K$ and the same transmission rate R.

Let t count the number of slots, $b_k(t)$ the number of information bits from user k successfully decoded up to slot t and $R_k(t) \stackrel{\Delta}{=} b_k(t)/L$ the correspond-

ing number of bit/s/Hz. The overall throughput $\eta_{N,M}$ is given by

$$\begin{aligned} \eta_{N,M} &= \lim_{t\to\infty} \frac{1}{tL} \sum_{k=1}^{K} b_k(t) \\ &= K \lim_{t\to\infty} \frac{1}{t} R_1(t) \end{aligned} \tag{2.8}$$

where the second line follows from the symmetry of the system with respect to any user.

Consider user 1 transmission. Under the above assumptions, the event that user 1 stops transmitting the current code-word is recognized to be a *recurrent event* [66]. A random *reward* $\mathcal{R}$ is associated to the occurrence of the recurrent event. In particular, $\mathcal{R} = R$ if transmission stops because successful decoding, and $\mathcal{R} = 0$ if transmission stops because delay/rate constraint violation. We can apply the renewal-reward theorem [66] and get

$$\eta_{N,M} = K \lim_{t\to\infty} \frac{1}{t} R_1(t) = K \frac{\mathrm{E}[\mathcal{R}]}{\mathrm{E}[\mathcal{T}]} \quad \text{with prob. 1} \tag{2.9}$$

where $\mathcal{T}$ is the random time between two consecutive occurrences of the recurrent event (inter-renewal time).

In order to evaluate $\mathrm{E}[\mathcal{R}]$, the mean reward, and $\mathrm{E}[\mathcal{T}]$, the mean inter-renewal time, we focus on the transmission of a given code-word of user 1 and we define the auxiliary random variable $\mathcal{M}$ to be the number of transmitted bursts between the instant when the code-word is generated and the instant when its transmission is stopped (i.e., between two consecutive occurrences of the recurrent event). We define the event $\mathcal{A}_m \stackrel{\Delta}{=} \{I_{1,m} > R\}$, and the probability $q(m)$ that the random sequence $I_{1,1}, I_{1,2}, \ldots, I_{1,m}, \ldots$ of mutual information at the user 1 decoder crosses level R at the m-th step (and not before), or, in other words, the probability of having successful decoding with m transmitted bursts. This is given by

$$\begin{aligned} q(m) &= \Pr(\overline{\mathcal{A}_1}, \ldots, \overline{\mathcal{A}_{m-1}}, \mathcal{A}_m) \\ &= \Pr(\overline{\mathcal{A}_1}, \ldots, \overline{\mathcal{A}_{m-1}}) - \Pr(\overline{\mathcal{A}_1}, \ldots, \overline{\mathcal{A}_m}) = p(m-1) - p(m) \end{aligned} \tag{2.10}$$

where

$$p(m) \stackrel{\Delta}{=} \Pr(\overline{\mathcal{A}_1}, \ldots, \overline{\mathcal{A}_m}) = 1 - \sum_{\ell=1}^{m} q(\ell) \tag{2.11}$$

The joint probability distribution of $\mathcal{T}$ and $\mathcal{M}$

$$f_{\mathcal{T},\mathcal{M}}(n,m) \stackrel{\Delta}{=} \Pr(\mathcal{T} = n, \mathcal{M} = m)$$

is obtained explicitly as follows (in the case $M \leq N$ otherwise the rate constraint is meaningless):

$$f_{\mathcal{T},\mathcal{M}}(n,m) = \begin{cases} (1-p_t)^N & n = N, m = 0 \\ v(N,m) + \binom{N}{m}(1-p_t)^{N-m} p_t^m p(m) & n = N, 1 \leq m \leq M-1 \\ v(n,M) + \binom{n-1}{M-1}(1-p_t)^{n-M} p_t^M p(M) & M \leq n \leq N, m = M \\ v(n,m) & m \leq n \leq N-1, 1 \leq m \leq M-1 \\ 0 & \text{elsewhere} \end{cases}$$

(we use the short-hand notation $v(n,m)$ for $\binom{n-1}{m-1}(1-p_t)^{n-m} p_t^m q(m)$). In Appendix 2.B we show that $f_{\mathcal{T},\mathcal{M}}(n,m)$ is a well-defined probability distribution for all $0 \leq p_t \leq 1$, $N \geq M > 0$ and non-negative non-increasing sequence $\{p(m)\}$ with $p(0) = 1$.

At this point, we are ready to compute $\mathrm{E}[\mathcal{R}]$ and $\mathrm{E}[\mathcal{T}]$. A reward R is obtained for $(\mathcal{T},\mathcal{M}) = (n,m)$ if successful decoding occurs in slot n after code-word generation and with m transmitted bursts. This corresponds to placing $m-1$ transmissions in the first $n-1$ slots without success, and the m-th transmission in the n-th slot with success, which occurs with probability $v(n,m)$. Therefore, the average reward is given by

$$\begin{aligned} \mathrm{E}[\mathcal{R}] &= R \sum_{m=1}^{M} \sum_{n=m}^{N} v(n,m) \qquad (2.12) \\ &= R\left[1 - \sum_{\ell=0}^{M-1} \binom{N}{\ell}(1-p_t)^{N-\ell} p_t^\ell p(\ell) - \sum_{\ell=M}^{N} \binom{N}{\ell}(1-p_t)^{N-\ell} p_t^\ell p(M)\right] \end{aligned}$$

and the average inter-renewal time is given by

$$\begin{aligned} \mathrm{E}[\mathcal{T}] &= \sum_{m=0}^{M} \sum_{n=1}^{N} n\, f_{\mathcal{T},\mathcal{M}}(n,m) \qquad (2.13) \\ &= \sum_{m=0}^{M-1} \frac{1}{p_t} p(m) \left[1 - \sum_{\ell=0}^{m} \binom{N+1}{\ell}(1-p_t)^{N+1-\ell} p_t^\ell - \binom{N}{m}(1-p_t)^{N-m} p_t^{m+1}\right] \end{aligned}$$

Finally, the desired closed-form expression for the system throughput is given by

$$\eta_{N,M} = RG \frac{\left[1 - \sum_{\ell=0}^{M-1} \binom{N}{\ell}(1-p_t)^{N-\ell} p_t^\ell p(\ell) - \sum_{\ell=M}^{N} \binom{N}{\ell}(1-p_t)^{N-\ell} p_t^\ell p(M)\right]}{\sum_{m=0}^{M-1} p(m) \left[1 - \sum_{\ell=0}^{m} \binom{N+1}{\ell}(1-p_t)^{N+1-\ell} p_t^\ell - \binom{N}{m}(1-p_t)^{N-m} p_t^{m+1}\right]} \qquad (2.14)$$

Protocols INR, RTD and ALO described before, for given parameters N, M, R, G, K and γ, differ in the probabilities $p(m)$. Consider first INR and RTD. These schemes have memory, since the receiver accumulates mutual information, for INR, or SINR, for RTD, over the sequence of slots $\mathcal{S}_{1,m}$. From (2.5) and (2.6), since $\beta_{1,s}$ is non-negative, it is apparent that the random sequence $\{I_{1,m}\}$ is non-decreasing with probability 1. Then, $\overline{\mathcal{A}_\ell} \subseteq \overline{\mathcal{A}_m}$ for all $\ell \leq m$ and we can write

$$p(m) = \Pr(\overline{\mathcal{A}_m})$$

For ALO, $I_{1,m}$ given by (2.7) has no particular monotone behavior. However, the receiver has no memory of past signal bursts and the events $\mathcal{A}_m$ are i.i.d., hence we can write

$$p(m) = \Pr(\overline{\mathcal{A}_1}, \ldots, \overline{\mathcal{A}_m}) = \prod_{i=1}^{m} \Pr(\overline{\mathcal{A}_i}) = \Pr(\overline{\mathcal{A}_1})^m$$

Finally, for all protocols examined we obtain a compact expression for $p(m)$ as

$$p(m) = \begin{cases} \Pr\left(\sum_{s\in\mathcal{S}_{1,m}} \log(1+\beta_{1,s}) \leq R\right) & \text{INR} \\ \Pr\left(\log(1+\sum_{s\in\mathcal{S}_{1,m}} \beta_{1,s}) \leq R\right) & \text{RTD} \\ \Pr\left(\log(1+\beta_{1,1}) \leq R\right)^m & \text{ALO} \end{cases} \tag{2.15}$$

From (2.14), it can be easily shown that $\eta_{N,M}$ is a decreasing function of the probabilities $p(m)$ and, from (2.15), that the probabilities $p(m)$ are related by

$$\begin{aligned} \Pr\left(\sum_{s\in\mathcal{S}_{1,m}} \log(1+\beta_{1,s}) \leq R\right) &\leq \Pr\left(\log(1+\sum_{s\in\mathcal{S}_{1,m}} \beta_{1,s}) \leq R\right) \\ &\leq \Pr\left(\log(1+\beta_{1,1}) \leq R\right)^m \end{aligned} \tag{2.16}$$

for every $m \geq 1$. Then, as expected, the three protocols are related by

$$\eta_{N,M}^{(\text{INR})} \geq \eta_{N,M}^{(\text{RTD})} \geq \eta_{N,M}^{(\text{ALO})} \tag{2.17}$$

The computation of $p(m)$ in (2.15) may not be done in closed form for every fading statistics and every protocol. For the INR and RTD it is not possible to find closed-form expressions for the probabilities $p(m)$. However, these can be calculated easily for any m as follows. Let $Z = \beta_{1,1}$ and $I = \log(1+\beta_{1,1})$. Then, from definitions (2.15), we see that for INR, $p(m)$ is the cdf of the sum of m i.i.d. RV's distributed as I, evaluated in R, and for RTD, $p(m)$ is the cdf of the sum of m i.i.d. RV's distributed as Z, evaluated in $2^R - 1$. For small m, $p(m)$ can be evaluated from the

distribution of $\beta_{1,1}$ (e.g., by using the characteristic function). Since this approach involves discrete Fourier transforms whose length increases with m, it cannot be applied for large m. In this case, from the central limit theorem [83] we have that $\frac{1}{\sqrt{m}}\sum_{s\in\mathcal{S}_{1,m}}\beta_{1,s}$ and $\frac{1}{\sqrt{m}}\sum_{s\in\mathcal{S}_{1,m}}\log(1+\beta_{1,s})$ are close to Gaussian RV's, for large m. Therefore, $p(m)$ can be easily evaluated from the Gaussian cdf. For the sake of brevity, we skip the details of numerical computations. However, it is interesting to notice that none of the results of this work are obtained by Monte Carlo simulation. For ALO, we have that

$$\eta_{N,M}^{(\mathrm{ALO})} = RG(1-p(1)) \tag{2.18}$$

independently of N and M. This result is expected, since ALO has no memory and both delay and rate constraints are irrelevant. Note that (2.18) is the throughput (2.14) for $M=1$, i.e., $\eta_{N,1}^{(\mathrm{INR})} = \eta_{N,1}^{(\mathrm{RTD})} = \eta_{N,1}^{(\mathrm{ALO})} = \eta_{N,M}^{(\mathrm{ALO})}$ independent of the delay constraint N.

Interestingly the throughput $\eta_{N,M}$ can be made arbitrarily large by increasing the user transmit SNR γ. Since the throughput for ALO is a lower bound for the other two protocols, it is sufficient to prove that ALO is not interference limited. Let J be the number of interfering users on a given slot and let $\Pr(J=k)$ the probability that J equals k, for $k\in\{0,\cdots,K-1\}$. We can write

$$\begin{aligned}
\lim_{\gamma\to\infty}\eta_{N,M} &= \lim_{\gamma\to\infty} GR\Pr(\log(1+\beta_{1,1})>R) \\
&= \lim_{\gamma\to\infty} GR\sum_{k=0}^{K-1}\Pr(J=k)\Pr\left(\beta_{1,1}>2^R-1|J=k\right) \\
&\geq \lim_{\gamma\to\infty} GR(1-p_t)^{K-1}\Pr\left(\alpha_{1,1}>\frac{2^R-1}{\gamma}\right)
\end{aligned}$$

where the last inequality follows from considering only the event $J=0$. Now, we choose an $\epsilon>0$ such that $F_\alpha(\epsilon)<1$, and we let $R=\log(1+\gamma\epsilon)$. Finally, we obtain

$$\lim_{\gamma\to\infty}\eta_{N,M} \geq \lim_{\gamma\to\infty} G(1-p_t)^{K-1}\left(1-F_\alpha(\epsilon)\right)\log(1+\gamma\epsilon) = \infty$$

as desired. This means that the ARQ system is not interference limited, even if no joint decoding is implemented at the receiver: arbitrarily high throughput can be obtained by simply increasing transmit SNR of all users, irrespectively of power control, fading, etc ... Intuitively, this is due to the fact that there is a non-zero probability that only one user is active on any given slot, and can transmit at very high instantaneous rate.

Fig. 2.2 shows $\eta_{N,M}$ vs. R for transmit SNR $\gamma=10$dB, $K=50$ users, load $G=1$, delay constraint $N=100$ and increasing values of rate constraint

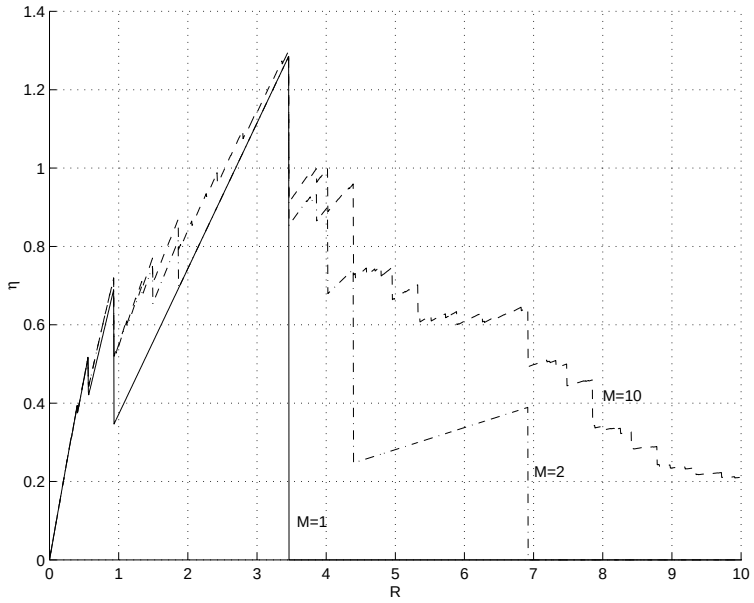

Figure 2.2: $\eta_{N,M}$ vs. R for γ=10dB, K=50, $G = 1$ and $N = 100$ for INR on AWGN channel.

M for the INR protocol on the AWGN channel (no fading). As already pointed out, the curve for $M = 1$ coincides with the throughput of the ALO protocol. The different curves overlap for small R since one transmitted burst is sufficient to decode. For $M > 1$ the throughput is non-zero also for R larger than $\log_2(1+\gamma) = 3.5$. For example, for $M = 2$ the maximum mutual information that can be accumulated is $2\log_2(1+\gamma) = 6.9$.

2.5 Unconstrained throughput

In this section, we study the throughput for an *unconstrained system*, i.e., for $N, M \to \infty$. In fact, form (2.16), we see that the sequence $p(m)$ for both INR and RTD is "sub-geometric", i.e., that $p(m) \leq p(1)^m$ for all $m \geq 1$, with equality only for $m = 1$. From this observation, it is possible to show that for, both INR and RTD, the throughput is increasing in N and M, i.e., that

$$\eta_{N+\ell,M+r} \geq \eta_{N,M}$$

for all $\ell, r \geq 0$, with equality for $\ell = 0, r = 0$ only. This result is intuitive, since it makes sense that the throughput is going to increase by relaxing the delay or the rate constraints. However, it is not completely trivial since both the numerator (average reward) and the denominator (average inter-renewal

time) of (2.14) are increasing functions of N and M. As a matter of fact, both the INR and the RTD protocols have the nice feature that "the longer we wait the more we gain".

In the following we indicate with η the unconstrained throughput, i.e., $\eta = \lim_{N,M\to\infty} \eta_{N,M}$. We notice here that all three protocols without constraints yield zero packet loss probability: the transmission of a code-word ends only when it is correctly decoded. Hence, the unconstrained throughput is easily obtained from (2.14) as

$$\eta = \frac{R\,G}{\sum_{m=0}^{\infty} p(m)} = \frac{R\,G}{\mathrm{E}[\mathcal{M}]} \tag{2.19}$$

where we used the fact that $\sum_{m=0}^{\infty} p(m) = \sum_{m=1}^{\infty} m\,q(m) = \mathrm{E}[\mathcal{M}]$, the average number of transmitted bursts needed for successful decoding. In passing, we notice that $\mathrm{E}[\mathcal{M}]/p_t$ is the mean delay (measured in slots) for the transmission of an information message, i.e., it is the average number of slots between the generation of a code-word and its successful decoding. It is worth pointing out that (2.19) holds for $p_t > 0$. In fact, for all finite N, we have

$$\lim_{p_t\to 0} \eta_{N,M} = RG(1 - p(1)) \tag{2.20}$$

In other words, in the limit for infinite population (for every finite G letting $p_t \to 0$ is equivalent to let $K \to \infty$), the INR and RTD protocols with finite delay constraint are equivalent to ALO. In fact, in this case a large number of users transmit with very small probability, and the probability that a user transmit more than once in any finite time N is negligible. Therefore, either the packet is successfully decoded at the first attempt, or it is discarded, like in ALO. On the contrary, for $N \to \infty$ the limit for infinite population is different for the three protocols and it is given by (2.19). In case we relax only the delay constraint, i.e., $N \to \infty$, then for any finite M the throughput is given by

$$\eta_{\infty,M} = RG\frac{1 - p(M)}{\sum_{m=0}^{M-1} p(m)} \tag{2.21}$$

Note that for ALO, $\eta = \eta_{N,M}$, given in (2.18), can be obtained explicitly for the AWGN channel and for the Rayleigh fading channel (see Appendix 2.E). For the channel without fading we have

$$\eta^{(\mathrm{ALO})} = R\,G \sum_{\ell=0}^{K(R,\gamma)-1} \binom{K-1}{\ell} \left(\frac{G}{K}\right)^{\ell} \left(1 - \frac{G}{K}\right)^{K-1-\ell} \tag{2.22}$$

where

$$K(R,\gamma) = \left\lfloor \frac{1}{2^R - 1} - \frac{1}{\gamma} \right\rfloor + 1 \tag{2.23}$$

is the maximum number of simultaneous users in a slot that can be correctly decoded (notice that, depending on R and γ, a collision does not correspond necessarily to an error, since $K(R,\gamma)$ might be larger than 1). For $K \to \infty$, (2.22) yields

$$\eta^{(\mathrm{ALO})} = R\,G \sum_{\ell=0}^{K(R,\gamma)-1} \mathrm{e}^{-G} \frac{G^\ell}{\ell!} \tag{2.24}$$

that for $K(R,\gamma) = 1$ reduces to the well-known result of classical slotted Aloha, $\eta^{(\mathrm{ALO})} = RG\mathrm{e}^{-G}$.
For the channel with Rayleigh fading we have

$$\eta^{(\mathrm{ALO})} = R\,G\,\mathrm{e}^{-(2^R-1)/\gamma} \left(1 - \frac{G}{K}(1 - 2^{-R})\right)^{K-1} \tag{2.25}$$

which for $K \to \infty$ yields

$$\eta^{(\mathrm{ALO})} = R\,G\mathrm{e}^{-(2^R-1)/\gamma - (1-2^{-R})G} \tag{2.26}$$

Up to our knowledge, the probability distribution function of the SINR with Rayleigh fading and Binomial- or Poisson-distributed number of interferers (see Appendix 2.E), used in (2.25) and (2.26), was not known in closed form prior to this work.

Figs. 2.3 and 2.4 show η vs. R, for the INR, RTD and ALO protocols, with $\gamma = 10$dB, $K = 50$ users, load $G = 1$ in AWGN and Rayleigh fading, respectively. For ALO on AWGN channel, η is zero for $R > \log_2(1+\gamma) = 3.5$ since for higher rates the SINR is not enough even in the absence of interferers (the system becomes power-limited rather than interference-limited). In the case of Rayleigh fading, η decreases with R but it is positive even for $R > \log_2(1+\gamma)$, since there is a non-zero probability that the fading gain is larger than one.

2.6 Optimal information rate

In Section 2.4 we expressed the throughput as a function of different system parameters: the delay constraint N, the rate constraint M, the transmit SNR γ, the information rate R, the number of users K, the channel load G and the different protocols $\{p(m)\}$. In Section 2.5 we showed that an ARQ system is not interference limited and that the throughput can be increased by relaxing the delay and rate constraint, i.e., the average reward

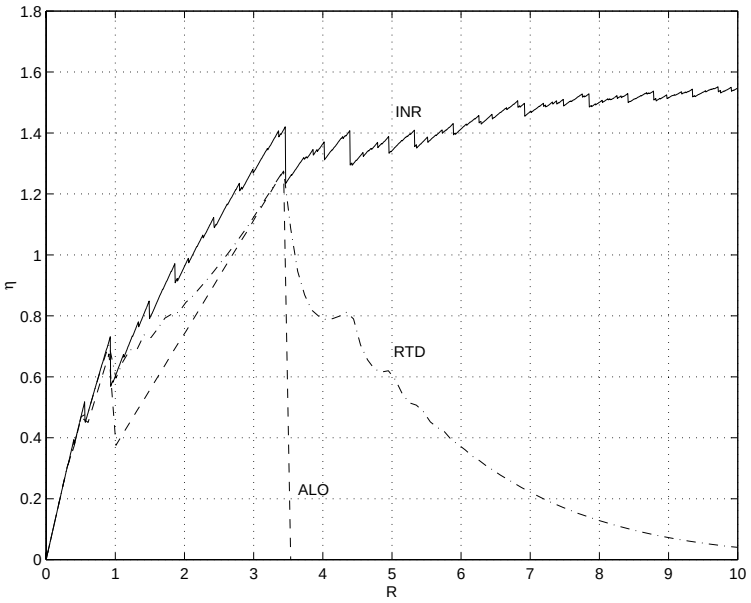

Figure 2.3: η vs. R for γ=10dB, K=50, $G = 1$ on AWGN channel.

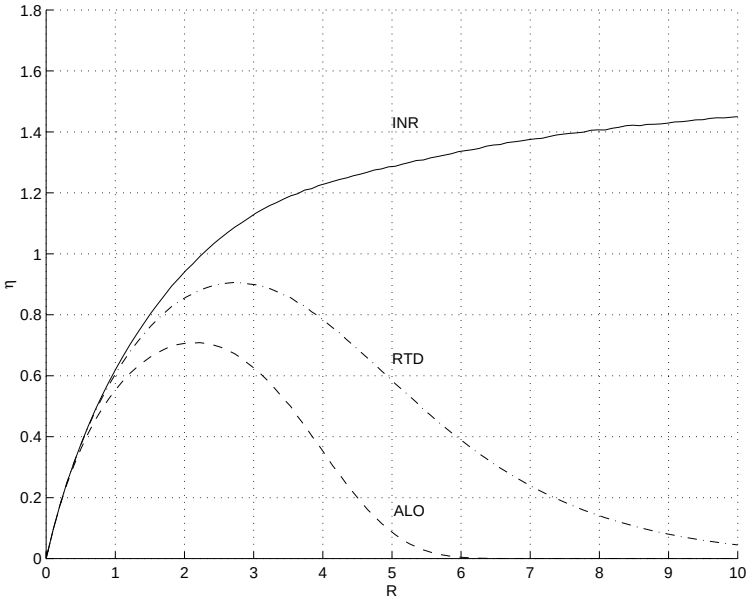

Figure 2.4: η vs. R for γ=10dB, K=50, $G = 1$ on Rayleigh fading channel.

$\mathrm{E}[\mathcal{R}]$ increases faster than the average delay $\mathrm{E}[\mathcal{T}] = \mathrm{E}[\mathcal{M}]/p_t$. From Figs. 2.3 and 2.4 it is clear that there exists optimal value of R for a given number of users K, channel load G and transmit SNR γ. The rest of the section is devoted to the determination of the optimal information rate. We indicate with $\bar{\eta}$ the R-optimized unconstrained throughput, i.e., $\bar{\eta} = \sup_{R \geq 0} \eta$.

We start by considering the limit for large R. In Appendix 2.C we show that

$$\lim_{R \to \infty} \eta = \begin{cases} G\,\mathrm{E}[\log(1+\beta_{1,1})] & \text{INR} \\ 0 & \text{RTD} \\ 0 & \text{ALO} \end{cases} \tag{2.27}$$

For INR, $\eta < G\,\mathrm{E}[\log(1+\beta_{1,1})]$ for all finite R. This fact is quite hard to show directly by using (2.19), since the probabilities $p(m)$ depend on R but a closed form is not available. However, we can provide a simple indirect proof of the statement as follows. The quantity $C \stackrel{\Delta}{=} \mathrm{E}[\log(1+\beta_{1,1})]$ is the capacity of the memoryless L-block interference channel [84] given in (2.3), where the interference signal $\boldsymbol{v}_{k,s}$ is proper complex Gaussian with i.i.d. components and where $\beta_{k,s}$ is the SINR for block s. A well-known result states that feedback does not increase the capacity of memoryless channels [1]. Then, even if the encoder has available the sequence of past received vectors $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_{s-1}$, the maximum transmissible rate for channel (2.3) is C.[7] Hence, we conclude that $\eta = GC$ is actually the maximum achievable throughput on this channel, irrespectively of the feedback and for any choice of R. From a practical system design point of view, in the absence of rate and delay constraints it is convenient to work with a very high information rate R, irrespectively of the channel load G and the transmit SNR γ.

For INR the maximum throughput is achieved for infinite delay. It is interesting to notice that, with infinite delay, the same maximum throughput (with zero packet loss probability) can be achieved by a system without feedback (just forward error correcting codes) [39]. It is natural to ask why the ACK/NACK feedback channel should be implemented at all. The answer is provided by closer examination of the average delay: the system without feedback needs a very large (infinite) delay in order to transmit with arbitrarily small packet loss probability for all values of η [39]. On the contrary, the INR protocol achieves zero transmission failure probability with finite average delay for all η strictly less than GC. Fig. 2.5 shows the average number of transmitted bursts $\mathrm{E}[\mathcal{M}]$ vs. η for the ALO and INR

[7] It is important to notice that (2.3) is memoryless at the block level, but not at the symbol level. Feedback does not provide any capacity increase if the feedback channel works at the slot rate, i.e., it sends back the whole received vector $\boldsymbol{y}_s$ at the end of each s-th slot. This is precisely the way the ACK/NACK feedback works. On the contrary, capacity would be clearly increased by a feedback working at faster rate, which sends back the components of $\boldsymbol{y}_s$ as soon as they are received, during each s-th slot.

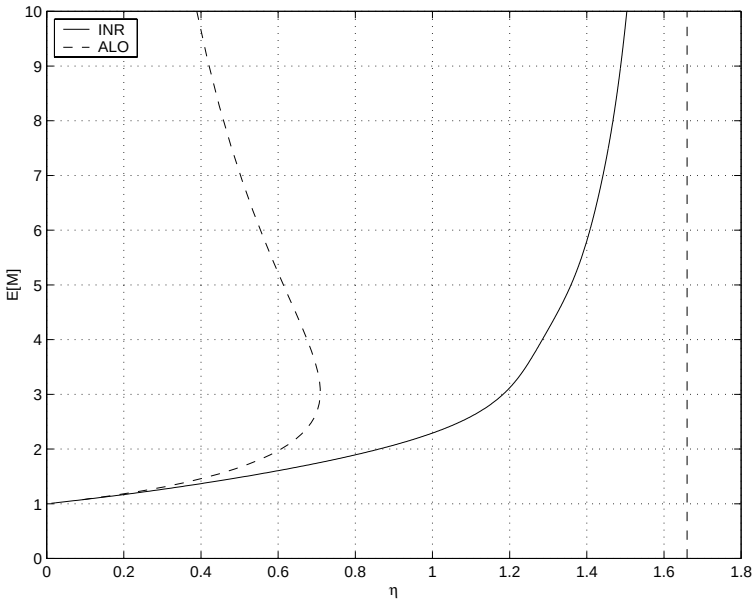

Figure 2.5: $\mathrm{E}[\mathcal{M}]$ vs. η for $\gamma = 10$dB, $K = 50$ and $G = 1$ for ALO and INR on Rayleigh fading channel.

protocols in the case of Rayleigh fading, for $\gamma = 10$dB, $K = 50$ users and load $G = 1$. The corresponding average delay is given by $\mathrm{E}[\mathcal{M}]/p_t$.

As opposed to INR, the throughput for very high R is zero for RTD and ALO. In fact, ALO and RTD involve strongly suboptimal coding schemes, for which $\mathrm{E}[\mathcal{M}]$ grows faster than R. Thus, the limiting η is zero. Since $\eta = 0$ for $R = 0$ and goes to 0 for large R, for both protocols there exist an optimal finite non-zero R. Unfortunately, the closed form expression of the optimal R seems infeasible, but, at least for the ALO protocol on AWGN, we can gain insight by taking a closer look at (2.22) and (2.24). It is easy to see that the supremum of η, for fixed G, K and γ, is always obtained when $R = \log(1 + \frac{\gamma}{1+J\gamma})$ for some integer J, where $J + 1$ is the maximum number of users that can collide on the same slots without causing a decoding error. Therefore, maximizing with respect to R is equivalent to searching for the maximum of the expression $G \log(1 + \frac{\gamma}{1+J\gamma}) \Pr(J \text{ interferes})$ over the non-negative integers $J \in \{0, \cdots, K-1\}$. In particular, for small G the maximum is obtained by $J = 0$. In this case, the throughput is maximized by choosing the largest possible R, i.e., $R = \log(1 + \gamma)$, and by letting the protocol alone to take care of collisions, like in conventional slotted Aloha. As G increases, the maximum is obtained by larger and larger J. In this case, the throughput is maximized by choosing R in order to tolerate up to J interferers, i.e., $R = \log(1 + \frac{\gamma}{1+J\gamma})$ (a decoding error occurs only when there are more than J interferers). In this way, the task of coping with

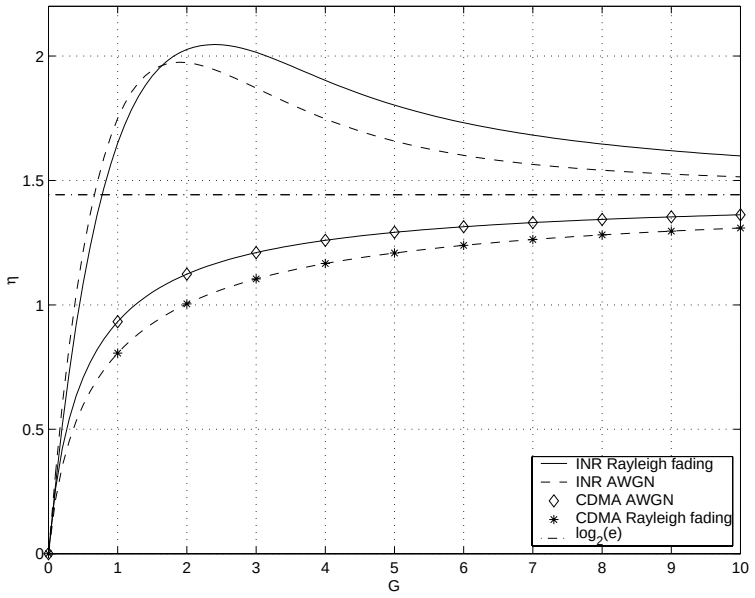

Figure 2.6: $\bar{\eta}$ vs. G for INR and CDMA, γ=10dB in AWGN and Rayleigh fading.

collisions is shared by channel coding and by the retransmission protocol: channel coding yields no errors for up to $J+1$ active users in the slot, while if the number of active users is larger than $J+1$ retransmission is needed. Figs. 2.6 and 2.7 show $\bar{\eta}$ vs. G for INR and ALO respectively, for γ=10dB on AWGN and Rayleigh fading channel.

From Figs. 2.6 and 2.7 it is apparent that there exist a limiting value of $\bar{\eta}$ as G grows towards infinity. In Appendix 2.D, we show that

$$\lim_{G\to\infty} \bar{\eta} = \begin{cases} \log(\mathrm{e}) & \text{INR} \\ \log(\mathrm{e}) & \text{RTD} \\ \log(\mathrm{e})\frac{\kappa}{\mu_\alpha} & \text{ALO} \end{cases} \tag{2.28}$$

where $\mu_\alpha = \mathrm{E}[\alpha_{1,1}]$ and $\kappa = \sup_{u\geq 0} u[1-F_\alpha(u)]$. For AWGN, $F_\alpha(u)$ is a step function with jump in $u=1$, therefore $\kappa = \mu_\alpha = 1$. For Rayleigh fading, $F_\alpha(u) = 1 - \mathrm{e}^{-u/\mu_\alpha}$, therefore $\kappa = \mu_\alpha/\mathrm{e}$ and $\lim_{G\to\infty} \bar{\eta} = \log(\mathrm{e})/\mathrm{e}$. This shows that for large channel load G all schemes are equivalent in AWGN, while ALO performs worse than INR and RTD in Rayleigh fading. In fact, ALO considers only the most recent received block for decoding. Hence, there is no "averaging effect" with respect to the fading affecting the useful signal over a long sequence of slots.

As G becomes large, a very large number of users transmit in every slot. In the limit, the system is equivalent to a CDMA system with an infinite

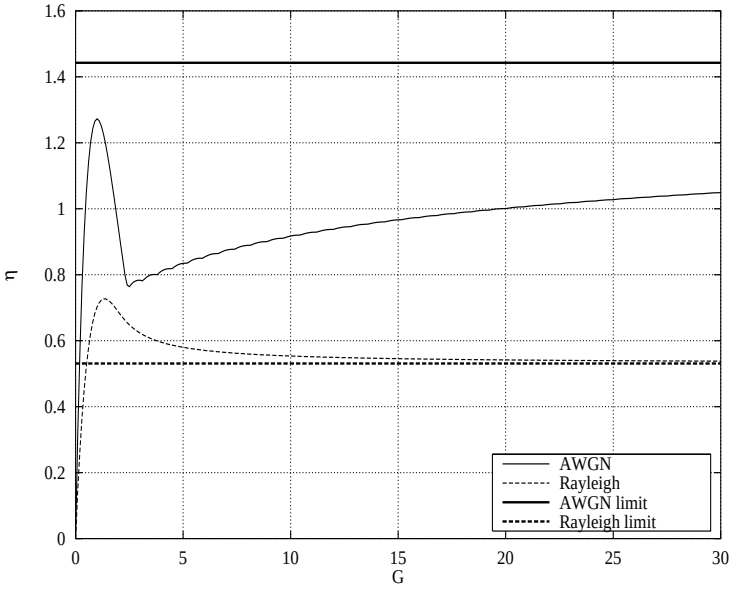

Figure 2.7: $\bar{\eta}$ vs. G for ALO, γ=10dB in AWGN and Rayleigh fading.

number of users K and infinite spreading gain N, such that the ratio K/N is equal to G [69, 70]. In fact, the channel load G is precisely the (average) number of users per dimension (per chip). The throughput of such CDMA system with single-user matched filter is given by [85, 86]

$$\eta^{(\mathrm{cdma})} = G\,\mathrm{E}\left[\log\left(1+\frac{\alpha_{1,1}\gamma}{1+G\mu_\alpha\gamma}\right)\right] \tag{2.29}$$

Its maximum is $\log_2(\mathrm{e})$ bit/s/Hz, obtained for $G\to\infty$. Interestingly, the throughput of CDMA is less than $\log_2(\mathrm{e})$ for all finite G, while for the INR, RTD and ALO schemes there might exist a range of G for which $\bar{\eta} > \log_2(\mathrm{e})$.

2.7 Concluding remarks

Combined channel coding and retransmission protocols appear to be a viable and simple solution for reliable packet-radio communication requiring high instantaneous rates and very low error probability and characterized by bursty sporadic transmission and by mild delay constraints.

In this chapter, we presented an information-theoretic throughput analysis of some ARQ protocols under idealized but fairly general conditions. We showed that typical set decoding has very desirable properties for Hybrid-ARQ, in the limit for large slot dimension. From a renewal-reward theory approach, we obtained closed-form throughput formulas for three simple

protocols: a generalization of slotted Aloha (ALO), a repetition time diversity scheme with maximal-ratio packet combining (RTD) and an incremental redundancy scheme based on progressively punctured codes (INR). We analyzed the effect of delay and rate constraints on the throughput, as well as the limiting behavior with respect to the information rate and the channel load. Interestingly, all three protocols are not interference-limited, and achieve arbitrarily large throughput by simply increasing the transmit power of all users.

The channel model described in Section 2.2 is, admittedly, quite simple and idealized. In order to get some insight into the effect on performance of the different system parameters, we made some hypothesis to make the model tractable. In the rest of the section, we discuss the way our model compares with practical systems and we also consider some practical implementation issue.

- We assumed that L, the number of complex symbols per slot, is large enough to guarantee reliable communication on each slot, in other words this assumption allows us to write $\Pr(\text{error}|R < I_{k,m}) = 0$ and $\Pr(\text{error}|R \geq I_{k,m}) = 1$. As we already pointed out in a footnote, this is quite realistic, in fact, in many practical applications the product WT is large.

- The assumption of block-fading is common in information theoretic studies of channels with fading. This is motivated by the fact the in relatively slowly moving environment and with reasonably system bandwidth, the Doppler spread is negligible. This assumption holds, for example, for indoor environments [72]. The i.i.d. assumption is realistic for slotted communication where slots are separated in time and/or frequency. In our model, due to user random activity, this hypothesis is not at all restrictive.

- The hypothesis of completely symmetric system with respect to any user was made for sake of simplicity. It may be restrictive since different class of users, characterized by different sets of six parameters $(M, N, p_t, R, \gamma, F_\alpha(x))$, are present in a practical system. The application of the renewal-reward theorem in still possible under the assumption of fading independent form slot to slot and from user to user with identical distribution with respect to the time index and single-user based decoding. The SINR is given by (2.2) and its cdf is easily

obtainable as

$$\begin{aligned} F_{\beta_k}(R_k) &\triangleq \Pr[\beta_{k,s} \le R_k] \\ &= \sum_{u_j \in \{0,1\}\,:\, j \in \{1,\cdots,K\},\, j \neq k} \prod_j [u_j p_j + (1-u_j)(1-p_j)] \cdot \\ &\quad \cdot \int_{\mathbb{R}_+^{K-1}} F_{\alpha_k}\left(\frac{R_k}{\gamma_k}\Big(1 + \sum_j u_j x_j \gamma_j\Big) \right) \prod_j \mathrm{d}F_{\alpha_j}(x_j) \end{aligned}$$

The probabilities $p(m)$ are still given by (2.15), where we need to compute the m-fold convolution of $F_{\beta_k}(R_k)$ (for RTD) or of $F_{I_k}(R_k) = F_{\beta_k}(2^{R_k}-1)$ (for INR). The joint distribution of $(\mathcal{T}, \mathcal{M})$ in (2.12) is still valid. For each user k, we need a $f_{\mathcal{T}_k,\mathcal{M}_k}(n_k, m_k)$ obtained by substitution of $(N, M, p(m), p_t)$ in (2.12) with the corresponding parameters for user k. Finally, the throughput is given by

$$\eta = \sum_{k=1}^{K} \frac{\mathrm{E}[\mathcal{R}_k]}{\mathrm{E}[\mathcal{T}_k]}$$

where $\mathrm{E}[\mathcal{R}_k]$ and $\mathrm{E}[\mathcal{T}_k]$ are given by (2.12) and (2.13) respectively. It is clear that in this way we have a not-manageable expression function of 6^K parameters. Alternatively, the system performance can be evaluated with simulation but the challenge of this work was to develop a closed form throughput analysis.

- The assumption of perfect receiver channel state information, in the sense of knowledge of fading gains for all the active users on every slot may appear unrealistic. Actually, the fading can be estimated very reliably by inserting a training sequence into each slot at a price of slight loss in rate. The assumption that the receiver knows exactly the time-hopping sequences of all users (the active set) might not be realistic. If user activity is random and not known to the receiver, our results can be seen as an upperbound on the achievable throughput obtained by a *genie-aided* receiver which knows *a priori* the active users in each slot. True random access, where the receiver must also detect which users are active, in order to make the appropriate packet combining, might be studied by inserting in our framework an *active user detection* scheme.

- Practical coding and decoding schemes based on incremental redundancy and featuring built-in error detection capability should be used with Hybrid-ARQ. As we already pointed out, a complete decoding scheme (es. Maximum Likelihood Viterbi decoding) is not suited for this propose; in this case a CRC must be inserted in each slot, as

currently done in all practical systems, at the price of rate loss. Another possibility to overcome the problem is the use of redundancy introduced by higher layer protocols or the use of the statistics of the metric computed at the receiver [76].

- In practice, binary codes are used which are mapped on bi-dimensional constellations before being sent on the channel, es. for high rate data transmission the UMTS standard proposes the use of Turbo Codes mapped on a QPSK constellation. In our analysis we assumed Gaussian coding, which would correspond to the use of a modulation with infinite points. Actually, in order to get theoretical inside into the performance of finite channel input alphabets, our analysis remains valid provided that the correct expressions for the mutual informations in (2.5), (2.6) and (2.7) are used. In [76], the authors computed a lower bound on the throughput performance of binary 1/M short constraint length convolutional codes mapped on BPSK based on worst case pair wise error probabilities. Their results shows that at low channel load performance predicted by theoretical analysis with binary input, very close to the theoretical performance with Gaussian input, are achieved by the considered convolutional codes. At high channel load, they show that the computed bound is too loose and that there is a great gap between simulations and theoretical performance with Gaussian input. The question left open is how to narrow that gap. It is shown by simulation that increasing the code complexity (larger constraint length) is not the way to go, the authors then suggested to use of more complex co/decoding schemes like Turbo Codes or the use of multi-point constellations. Turbo-codes (or other forms of concatenated coding) with iterative decoding appear to be a promising solution. However, the behavior of iterative decoders in the presence of decoding errors should be better characterized in order to exploit it for error detection.

- As in most of the available literature on Alhoa protocols, we assumed an error/delay free feedback channel. This is a bit unrealistic but it simplifies the analysis. As argued in [38], a delayed feedback causes no fundamental problems and doesn't change substantially the result of the simplified analysis. Feedback errors and delay were consider in [54, 55, 58, 59] by employing Markov chain since all these phenomena introduce memory in the system.

- Here, we concentrated on a very simple receiver that does not attempt to decode the users jointly. A natural direction for future research is to consider joint decoding at the receiver (e.g., implemented by stripping). A theoretical difficulty is represented by the user random activity [18]. In fact, because of random access, the capacity region varies from slot to slot and it is not known in advance, unless a com-

plicated reservation/allocation scheme is implemented. Also, the set of interfering users might be different from slot to slot, and it is not clear how to carry out joint decoding across the slots. First steps in this direction are taken in [87, 68].

We conclude with system considerations. In most practical applications, packet-radio networks must co-exist with other systems, as for example a connection-oriented CDMA system where a large number of low-power low-rate users transmit continuously. Quite a lot of work has been dedicated to the problem of power control for bursty transmission, where closed-loop schemes are not effective, in the fear that high-rate high-power bursty users might create too much interference to an underlying CDMA system. An appealing consequence of our study is the following: instead of trying to control bursty users, we can let them transmit at full-power. Thanks to the ARQ protocol, the signal from all bursty users can be eventually decoded correctly and subtracted from the received signal, so that the underlying CDMA system "sees" a clean channel, as if the bursty users were not there. In this way, the two quite different system could be layered one on top of the other. Obviously, in order to make this claim rigorous several issues must be addressed in the details: perhaps the most important of which is the delay. In fact, CDMA users can be decoded only after the signal from bursty users has been subtracted. Then, the variable decoding delay associated with the ARQ protocol imposes a variable decoding delay also on the CDMA system. If CDMA users have a strict delay constraint (e.g., due to real-time speech transmission, like in cellular telephony), outages due to the occurrence of large decoding delay events must be taken into account.

Appendix

2.A Proofs of Lemmas 1,2 and 3

Following standard continuity arguments [77], we consider a quantization of the input and a partition of the output of (2.3) and we work on the resulting discrete channel. The results for the continuous channel can be obtained by taking the supremum over all input quantizations and output partitions. Fix a sequence of channel transition probabilities $\mathcal{P} = \{p_{k,s}(y|x) : s \in \mathcal{S}_{k,M}\}$. Let $P(\{\boldsymbol{x}_{k,s}, \boldsymbol{y}_s : s \in \mathcal{S}_{k,m}\})$, $P(\{\boldsymbol{x}_{k,s} : s \in \mathcal{S}_{k,m}\})$ and $P(\{\boldsymbol{y}_s : s \in \mathcal{S}_{k,m}\})$ be the joint and the marginal probability distributions induced by $\mathcal{P}$ and by the input distribution $q(x)$. Since on every slot $s \in \mathcal{S}_{k,m}$ the quantized version of channel (2.3) is a time-invariant DMC, for the weak law of large

numbers [83] we have the following limits in probability

$$\begin{aligned}\lim_{L\to\infty}\frac{1}{L}\log P(\{\boldsymbol{x}_{k,s},\boldsymbol{y}_s : s\in\mathcal{S}_{k,m}\}) &= -\sum_{s\in\mathcal{S}_{k,m}} H_{k,s}(X,Y)\\ \lim_{L\to\infty}\frac{1}{L}\log P(\{\boldsymbol{x}_{k,s} : s\in\mathcal{S}_{k,m}\}) &= -mH_k(X)\\ \lim_{L\to\infty}\frac{1}{L}\log P(\{\boldsymbol{y}_s : s\in\mathcal{S}_{k,m}\}) &= -\sum_{s\in\mathcal{S}_{k,m}} H_{k,s}(Y)\end{aligned}$$

where

$$\begin{aligned}H_{k,s}(X,Y) &\triangleq -\sum_{x,y} q(x)p_{k,s}(y|x)\log q(x)p_{k,s}(y|x)\\ H_k(X) &\triangleq -\sum_{x} q(x)\log q(x)\\ H_{k,s}(Y) &\triangleq -\sum_{x,y} q(x)p_{k,s}(y|x)\log\sum_{x'} q(x')p_{k,s}(y|x')\end{aligned}$$

are the joint, input and output entropies per letter in slot s. The typical set $\mathcal{A}^{\epsilon}_{k,m}$ is defined as the set of all sequences $\{\boldsymbol{x}_{k,s},\boldsymbol{y}_s : s\in\mathcal{S}_{k,m}\}$ satisfying

$$\begin{aligned}\left|\frac{1}{L}\log P(\{\boldsymbol{x}_{k,s},\boldsymbol{y}_s : s\in\mathcal{S}_{k,m}\}) + \sum_{s\in\mathcal{S}_{k,m}} H_{k,s}(X,Y)\right| &\le \epsilon\\ \left|\frac{1}{L}\log P(\{\boldsymbol{x}_{k,s} : s\in\mathcal{S}_{k,m}\}) + mH_k(X)\right| &\le \epsilon\\ \left|\frac{1}{L}\log P(\{\boldsymbol{y}_s : s\in\mathcal{S}_{k,m}\}) + \sum_{s\in\mathcal{S}_{k,m}} H_{k,s}(Y)\right| &\le \epsilon\end{aligned}$$

By letting $I(q(x),p_{k,s}(y|x)) \triangleq H_k(X)+H_{k,s}(Y)-H_{k,s}(X,Y)$, and by following the same steps in [1, Th. 8.7.1], we get that any rate less than $\frac{1}{m}\sum_{s\in\mathcal{S}_{k,m}} I(q(x),p_{k,s}(y|x))$ is ϵ-achievable. In particular, for $m=M$ and given sequence of channels $\mathcal{P}$, for sufficiently large L there exists codes $\mathcal{C}_k$ of length LM and rate R/M with error probability (with typical set decoding) less than ϵ if

$$R < \sum_{s\in\mathcal{S}_{k,M}} I(q(x),p_{k,s}(y|x)) \tag{2.30}$$

In order to prove Lemma 1 we need to show that: i) there is a single code $\mathcal{C}_k$ having error probability uniformly less than ϵ over all sequences of channels $\mathcal{P}$ satisfying (2.30); ii) for all $1\le m\le M$, if $R<\sum_{s\in\mathcal{S}_{k,m}} I(q(x),p_{k,s}(y|x))$,

then the punctured code $\mathcal{C}_{k,m}$ obtained from $\mathcal{C}_k$ by taking the first m sub-blocks of length L has also error probability less than ϵ.

From the random coding achievability part and from the strong converse (it holds for every sequence of channels as shown by Lemma 2) we have that

$$\mathrm{E}_{\mathcal{C}}[\Pr(\text{error}|\mathcal{P},\mathcal{C})] \to 1\left\{\sum_s I(q(x),p_{k,s}(y|x)) \le R\right\}$$

as $L \to \infty$, where $1\{A\}$ denotes the indicator function of the event A and where $\mathrm{E}_{\mathcal{C}}$ denotes expectation over the ensemble of all codes of size e^{RL} and block length LM generated according to the input distribution $q(x)$. By averaging also with respect to the sequence of channels and exchanging expectations with respect to $\mathcal{C}$ and with respect to $\mathcal{P}$ (we can always do it, since the integrand is non-negative and bounded by 1) we obtain

$$\mathrm{E}_{\mathcal{C}}[\mathrm{E}_{\mathcal{P}}[\Pr(\text{error}|\mathcal{P},\mathcal{C})]] \to \Pr\left(\sum_{s\in\mathcal{S}_{k,M}} I(q(x),p_{k,s}(y|x)) \le R\right)$$

Then, there exists a family of codes $\mathcal{C}^*$ for increasing L such that

$$\mathrm{E}_{\mathcal{P}}[\Pr(\text{error}|\mathcal{P},\mathcal{C}^*)] \le \Pr\left(\sum_{s\in\mathcal{S}_{k,M}} I(q(x),p_{k,s}(y|x)) \le R\right) \tag{2.31}$$

for L sufficiently large. Because of the strong converse, $\Pr(\text{error}|\mathcal{P},\mathcal{C}^*) \to 1$ for all $\mathcal{P}$ such that $\sum_{s\in\mathcal{S}_{k,M}} I(q(x),p_{k,s}(y|x)) \le R$. Then, in order to satisfy (2.31) it must be $\Pr(\text{error}|\mathcal{P},\mathcal{C}^*) \to 0$ for all channel sequences $\mathcal{P}$ such that $\sum_{s\in\mathcal{S}_{k,M}} I(q(x),p_{k,s}(y|x)) > R$. This shows that, asymptotically, there exist codes $\mathcal{C}^*$ such that

$$\Pr(\text{error}|\mathcal{P},\mathcal{C}^*) \to 1\left\{\sum_s I(q(x),p_{k,s}(y|x)) \le R\right\}$$

for all channel sequences $\mathcal{P}$.

Now, let $\mathcal{C}_{k,M} = \mathcal{C}^*$ and assume that for a given sequence of channels

$$R < \sum_{s\in\mathcal{S}_{k,m}} I(q(x),p_{k,s}(y|x)) \tag{2.32}$$

for some $1 \le m \le M$. Then, we can extend the sequence of channels by adding to $\{p_{k,s}(y|x) : s \in \mathcal{S}_{k,m}\}$ other $M-m$ dummy useless memoryless channels whose output is independent of the input. Since the mutual information on the last $M-m$ blocks is zero and because of (2.32), the resulting sequence of M channels $\mathcal{P}'$ satisfies $\Pr(\text{error}|\mathcal{P}',\mathcal{C}_{k,M}) \to 0$. Notice that

extending the sequence of channels is equivalent to appending dummy output signal blocks $\boldsymbol{z}_i$ independent of the channel input to the received signal $\{\boldsymbol{y}_s : s \in \mathcal{S}_{k,m}\}$, as described in Section 2.3. This concludes the proof of Lemma 1.

In order to prove Lemma 2 we use the limit in probability

$$\lim_{L\to\infty} \frac{1}{L} \log \frac{P(\{\boldsymbol{x}_{k,s}, \boldsymbol{y}_s : s \in \mathcal{S}_{k,m}\})}{P(\{\boldsymbol{x}_{k,s} : s \in \mathcal{S}_{k,m}\})P(\{\boldsymbol{y}_s : s \in \mathcal{S}_{k,m}\})} = \sum_{s\in\mathcal{S}_{k,m}} I(q(x), p_{k,s}(y|x))$$

where the LHS is the limiting normalized *information density* over the m slots and where, for a fixed sequence of channels, the RHS is a constant. Therefore, the inf-information rate and the sup-information rate (see definitions in [9]) coincide and, from [9, Th. 7], the strong converse holds, conditionally on the sequence $\{p_{k,s}(y|x) : s \in \mathcal{S}_{k,m}\}$.

In order to prove Lemma 3 we use the simple relation

$$\bigcup_{\widehat{w}\neq w} \mathcal{E}_{\widehat{w}} \subseteq \left\{\{\boldsymbol{x}_{k,s}^{(w)}, \boldsymbol{y}_s : s \in \mathcal{S}_{k,m}\} \notin \mathcal{A}_{k,m}^{\epsilon}\right\}$$

$\forall\ w \in \{1, \ldots, \mathrm{e}^{RL}\}$ and for all $m = 1, \ldots, M$. This implies that

$$\begin{aligned} \Pr(\text{undetected error}|w, \mathcal{P}, \mathcal{C}_{k,m}) &\leq \Pr\left(\left\{\{\boldsymbol{x}_{k,s}^{(w)}, \boldsymbol{y}_s : s \in \mathcal{S}_{k,m}\} \notin \mathcal{A}_{k,m}^{\epsilon}\right\}\middle| w\right) \\ &< \epsilon \end{aligned} \tag{2.33}$$

where the second inequality holds for arbitrary $\epsilon > 0$ and sufficiently large L, since the probability that the channel input and output sequences are not jointly typical vanishes as $L \to \infty$ [1, Th. 8.6.1]. Then, Lemma 3 follows from averaging (2.33) over all transmitted messages.

2.B Probability distribution of the inter-renewal time

The joint pdf of $\mathcal{T}$ and $\mathcal{M}$ can be expressed by

$$f_{\mathcal{T},\mathcal{M}}(n,m) = \begin{cases} (1-p_t)^N & n = N, m = 0 \\ v(N,m) + g(N,m) & n = N, 1 \leq m \leq M-1 \\ v(n,M) + r(n,M) & M \leq n \leq N, m = M \\ v(n,m) & m \leq n \leq N-1, 1 \leq m \leq M-1 \\ 0 & \text{elsewhere} \end{cases} \tag{2.34}$$

where we define

$$\begin{aligned} v(n,m) &= \binom{n-1}{m-1}(1-p_t)^{n-m}p_t^m q(m) \\ r(n,M) &= \binom{n-1}{M-1}(1-p_t)^{n-M}p_t^M p(M) \\ g(N,m) &= \binom{N}{m}(1-p_t)^{N-m}p_t^m p(m) \end{aligned}$$

where $q(m)$ is defined in (2.10), $p(m)$ in (2.11) and they are related by $q(m) = p(m-1) - p(m)$.

We show that (2.34) is a well-defined probability distribution for any $N \geq M > 0$, $0 \leq p_t \leq 1$ and non-negative decreasing sequence $\{p(m)\}$ with $p(0) = 1$. Since all terms in (2.34) are non-negative, it is sufficient to show that their sum is 1. We use the identity

$$\sum_{n=k}^{N-1} \binom{n}{k} a^{n-k}(1-a)^{k+1} = 1 - \sum_{\ell=0}^{k} \binom{N}{\ell} a^{N-\ell}(1-a)^{\ell} \tag{2.35}$$

(for $0 \leq a \leq 1$) and write

$$\sum_{n,m} f_{\mathcal{T},\mathcal{M}}(n,m) = \sum_{m=1}^{M}\sum_{n=m}^{N} v(n,m) + \sum_{m=0}^{M-1} g(N,m) + \sum_{n=M}^{N} r(n,M) \tag{2.36}$$

For the sake of brevity, we let $s(\ell) \triangleq \binom{N}{\ell}(1-p_t)^{N-\ell}p_t^{\ell}$. The first, second and third terms in the RHS of (2.36) are given by

$$\sum_{m=1}^{M}\sum_{n=m}^{N} v(n,m) = 1 - \sum_{\ell=0}^{M-1} s(\ell)p(\ell) - p(M)\left(1 - \sum_{\ell=0}^{M-1} s(\ell)\right) \tag{2.37}$$

by

$$\sum_{m=0}^{M-1} g(N,m) = \sum_{m=0}^{M-1} s(m)p(m) \tag{2.38}$$

and by

$$\sum_{n=M}^{N} r(n,M) = p(M)\left(1 - \sum_{\ell=0}^{M-1} s(\ell)\right) \tag{2.39}$$

where we used the fact that $\sum_{k=\ell+1}^{N} q(k) = p(\ell) - p(N)$. The result follows by noting that the second and third term in the RHS of (2.37) are the opposite of the terms given in (2.38) and in (2.39).

2.C Limits for large R

We want to establish the limiting behavior of the unconstrained system throughput for large R. To this purpose we consider $\lim_{R\to\infty} 1/\eta$, where η is given in (2.19).

We need the following lemmas:

Lemma C.1. Let X be a RV with cdf $F_X(x)$. Then, $\forall\ y$,

$$1_{\{x\geq y\}}F_X(y) \leq F_X(x) \leq F_X(y) + 1_{\{x\geq y\}}(1 - F_X(y)) \tag{2.40}$$

♢

Lemma C.2. If $a_n \to a$ as $n \to \infty$, then for any non-negative finite integer k

$$b_n = \frac{1}{n+k}\sum_{i=1}^{n} a_i \to a \ \text{ for } \ n \to \infty \tag{2.41}$$

♢

INR protocol. We let $X_i \stackrel{\Delta}{=} \log(1+\beta_{1,s_i})$ for $s_i \in \mathcal{S}_{1,m}$ and $\mu_X \stackrel{\Delta}{=} \mathrm{E}[X_i]$. Then, for $\epsilon_1 > 0$ and $b = \mu_X + \epsilon_1$ we can write

$$\begin{aligned}
\lim_{R\to\infty}\frac{1}{\eta} &= \lim_{R\to\infty}\frac{1+\sum_{m=1}^{\infty}\Pr(\sum_{i=1}^{m}X_i < R)}{RG} \\
&\stackrel{(a)}{\geq} \lim_{R\to\infty}\frac{1}{RG}\sum_{m=1}^{\infty}1_{\{R\geq mb\}}\Pr\left(\frac{1}{m}\sum_{i=1}^{m}X_i < b\right) \\
&\stackrel{(b)}{\geq} \frac{1}{Gb}\lim_{R\to\infty}\frac{1}{\lfloor R/b\rfloor+1}\sum_{m=1}^{\lfloor R/b\rfloor}\Pr\left(\frac{1}{m}\sum_{i=1}^{m}X_i < b\right) \\
&\stackrel{(c)}{=} \frac{1}{G(\mu_X+\epsilon_1)}
\end{aligned} \tag{2.42}$$

where (a) follows by applying Lemma C.1 to the RV $\sum_{i=1}^{m}X_i$ with $x = R$ and $y = mb$; (b) follows by noting that $b/R \geq 1/(\lfloor R/b\rfloor + 1)$; (c) follows from Lemma C.2 with $k = 1$. In fact, for the Large Deviation Theorem [88, Sec.5.11], it exists a non-negative function $\phi(\epsilon)$, for every $\epsilon > 0$ such that $\Pr[X_i - \mu_X > \epsilon] > 0$, for which we can write

$$\Pr\left[\sum_{i=1}^{m}(X_i - \mu_X) > m\epsilon\right] \leq \mathrm{e}^{-m\phi(\epsilon)}$$

hence

$$1 - \mathrm{e}^{-m\phi(\epsilon)} \leq \Pr\left[\frac{1}{m}\sum_{i=1}^{m}X_i \leq \epsilon + \mu_X\right] \leq 1$$

which implies that

$$\lim_{m\to\infty} \Pr\left[\frac{1}{m}\sum_{i=1}^{m} X_i \le \mu_X + \epsilon\right] = 1$$

Similarly, for $\epsilon_2 > 0$ and $b = \mu_X - \epsilon_2$ we can write

$$\begin{aligned}
\lim_{R\to\infty} \frac{1}{\eta} &= \lim_{R\to\infty} \frac{1 + \sum_{m=1}^{+\infty} \Pr\left(\sum_{i=1}^{m} X_i < R\right)}{RG} \\
&\le \lim_{R\to\infty} \frac{1}{RG} \sum_{m=1}^{\infty} \left[\Pr\left(\frac{1}{m}\sum_{i=1}^{m} X_i < b\right) + 1_{\{R\ge mb\}}\right] \\
&\le \lim_{R\to\infty} \frac{1}{GR} \sum_{m=1}^{\infty} \Pr\left(\frac{1}{m}\sum_{i=1}^{m} X_i < b\right) + \frac{1}{G\,b} \lim_{R\to\infty} \frac{1}{\lfloor R/b\rfloor} \sum_{m=1}^{\lfloor R/b\rfloor} 1 \\
&\overset{(a)}{=} \frac{1}{G(\mu_X - \epsilon_2)}
\end{aligned} \tag{2.43}$$

In order to get (a), we use the fact that, for the Large Deviation Theorem [88, Sec.5.11] applied to the sum of m i.i.d. random variables $-X_i$, it exists a non-negative function $\psi(\epsilon)$, for every $\epsilon > 0$ such that $\Pr[-X_i + \mu_X > \epsilon] > 0$, for which we can write

$$\Pr\left[\sum_{i=1}^{m}(-X_i + \mu_X) > m\epsilon\right] \le \mathrm{e}^{-m\psi(\epsilon)}$$

which implies that

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m} X_i < \mu_X - \epsilon\right] \le \mathrm{e}^{-m\psi(\epsilon)}$$

hence, by summing over m, we get

$$\sum_{m=1}^{\infty} \Pr\left[\frac{1}{m}\sum_{i=1}^{m} X_i < \mu_X - \epsilon\right] \le \frac{\mathrm{e}^{-\psi(\epsilon)}}{1 - \mathrm{e}^{-\psi(\epsilon)}}$$

which gives a finite positive bound. Therefor, (a) follows. Eventually, we get $G\,(\mu_X - \epsilon_2) \le \lim_{R\to\infty} \eta \le G(\mu_X + \epsilon_1)$ and by letting $\epsilon_i \to 0$ for $i = 1, 2$ and recalling that, by definition, $\mu_X = \mathrm{E}[\log(1 + \beta_{1,1})]$ we obtain the desired result.

It is important to notice that the hypothesis $\Pr[X_i - \mu_X > \epsilon] > 0$ for some $\epsilon > 0$ does not hold for constant X_i (degenerate random variable). In our model, the SINR X_i is a deterministic constant only if $p_t = 1$ (no user random activity, i.e., $G = K$) and constant fading. It is immediate to show

that all the three protocols are maximized by $R = \left(1 + \frac{\gamma}{1+(K-1)\gamma}\right)$ and that

$$\begin{aligned} \sup_{R\geq 0} \eta^{(\mathrm{ALO})} &= \sup_{R\geq 0} \eta^{(\mathrm{RTD})} = \sup_{R\geq 0} \eta^{(\mathrm{INR})} \\ &= K \log\left(1 + \frac{\gamma}{1+(K-1)\gamma}\right) = K\,\mathrm{E}[\log(1+X_i)] \end{aligned}$$

therefor our statement $\sup_{R\geq 0} \eta^{(\mathrm{INR})} = G\,\mathrm{E}[\log(1+X_i)]$ holds in full generality.

ALO and RTD protocols. We let $X_i \triangleq \beta_{1,s_i}$ for all $s_i \in \mathcal{S}_{1,m}$ and $\mu_X \triangleq \mathrm{E}[X_i]$. Then, for $\epsilon > 0$, $b = \mu_X + \epsilon$ and by following the same steps that lead to (2.42), we can write

$$\begin{aligned} \lim_{R\to\infty} \frac{1}{\eta} &= \lim_{R\to\infty} \frac{1 + \sum_{m=1}^{\infty} \Pr\left(\sum_{i=1}^{m} X_i < 2^R - 1\right)}{RG} \\ &\geq \lim_{R\to\infty} \frac{1}{Gb} \frac{2^R - 1}{R} \frac{1}{\lfloor (2^R-1)/b \rfloor + 1} \sum_{m=1}^{\lfloor (2^R-1)/b \rfloor} \Pr\left(\frac{1}{m}\sum_{i=1}^{m} X_i < b\right) \\ &= \frac{1}{Gb} \lim_{R\to\infty} \frac{\mathrm{e}^R - 1}{R} \\ &= \infty \end{aligned} \tag{2.44}$$

This shows that $\lim_{R\to\infty} \eta^{(\mathrm{RTD})} = 0$ and since $\eta^{(\mathrm{ALO})} \leq \eta^{(\mathrm{RTD})}$, the same result holds for ALO.

2.D Limits for large G

We want to establish the limiting behavior of the unconstrained system throughput maximized with respect to the rate R, i.e. $\bar{\eta} = \sup_{R\geq 0} \eta$ for η given in (2.19), for large load G. To this end, we need to consider the limiting behavior of the RV $\sum_{j=1}^{J} \alpha_j$ where J is the number of interfering users in a given slot, binomially distributed and α_j is the channel gain of user j, assumed to be i.i.d. and independent of J, with finite mean μ_α and variance σ_α^2. Since $G = p_t K \leq K$, as $G \to \infty$ also $K \to \infty$. The mean and the variance of J, indicated with μ_J and σ_J^2 respectively, are given by

$$\begin{aligned} \mu_J &= p_t(K-1) = G\frac{K-1}{K} \leq G \\ \sigma_J^2 &= (1-p_t)p_t(K-1) \leq G/4 \end{aligned} \tag{2.45}$$

By iterating expectation, we obtain

$$\begin{aligned} \mathrm{E}\left[\sum_{j=1}^{J}\alpha_j\right] &= \mu_\alpha\mu_J \\ \mathrm{Var}\left[\sum_{j=1}^{J}\alpha_j\right] &= \sigma_J^2\mu_\alpha^2+\mu_J\sigma_\alpha^2 \end{aligned} \tag{2.46}$$

Putting together (2.45) and (2.46) we conclude that $\frac{1}{G}\sum_{j=1}^{J}\alpha_j$ converges in probability to μ_α as $G\to\infty$, in fact

$$\begin{aligned} \frac{\mathrm{Var}\left[\sum_{j=1}^{J}\alpha_j\right]}{G^2} &\le \frac{\mathrm{E}[\alpha^2]}{G}\to 0 \\ \frac{\mu_J}{G} &= G\frac{K-1}{K}\frac{1}{G}\to 1 \end{aligned} \tag{2.47}$$

From continuity of the functions $1/(1+x)$ and $\log(1+x)$ for $x>0$, the following limits for $G\to\infty$ hold in probability

$$\begin{aligned} G\beta_{1,1} &\to \frac{\alpha_{1,1}}{\mu_\alpha} \\ G\log(1+\beta_{1,1}) &\to \log(\mathrm{e})\,\frac{\alpha_{1,1}}{\mu_\alpha} \end{aligned} \tag{2.48}$$

INR and RTD protocols. By using (2.48), and the fact that, in the case of INR $\bar{\eta}=G\,\mathrm{E}[\log(1+\beta_{1,1})]$ (obtained for $R\to\infty$ as proved in Appendix 2.D), we have

$$\begin{aligned} \lim_{G\to\infty}\bar{\eta} &= \lim_{G\to\infty} G\,\mathrm{E}[\log(1+\beta_{1,1})] \\ &= \log(\mathrm{e})\,\mathrm{E}\left[\frac{\alpha_{1,1}}{\mu_\alpha}\right]=\log(\mathrm{e}) \end{aligned} \tag{2.49}$$

Notice that (2.48) implies $\sum_{s\in\mathcal{S}_{1,m}}\log(1+\beta_{1,s})\to\log(1+\sum_{s\in\mathcal{S}_{1,m}}\beta_{1,s})$ in probability, as $G\to\infty$. Then, the probabilities $p(m)$ given in (2.15) for INR and RTD are equal in the limit for large G. Since η depends on the particular protocol only through the probabilities $p(m)$, we conclude that limit 2.49) holds also for RTD.

ALO protocol. For ALO we have

$$\begin{aligned}
&\lim_{G\to\infty} \sup_{R\geq 0} RG\left[1-\Pr\left(\log(1+\beta_{1,1})\leq R\right)\right] \\
&= \lim_{G\to\infty} \sup_{R\geq 0} RG\left[1-\Pr\left(G\log(1+\beta_{1,1})\leq RG\right)\right] \\
&= \lim_{G\to\infty} \sup_{R\geq 0} RG\left[1-\Pr\left(\log(\mathrm{e})\frac{\alpha_{1,1}}{\mu_\alpha}\leq RG\right)\right] \\
&\overset{(a)}{=} \lim_{G\to\infty} \frac{\log(\mathrm{e})}{\mu_\alpha} \sup_{u\geq 0} u(1-F_\alpha(u))
\end{aligned} \tag{2.50}$$

where (a) follows by letting $u = RG\mu_\alpha/\log(\mathrm{e})$ and by noticing that the expression that must be maximized depends on the product RG and not on G alone, therefore maximization with respect to R or with respect to u yields the same result.

2.E Some useful cdf's

In order to simplify the notation of (2.2), we indicate the active users on slot s by $k = 0, 1, \ldots, |\mathcal{K}(s)| - 1$ (user 0 is the reference user) and we define the following RV's:

- User k instantaneous SNR, $X_k \triangleq \alpha_{k,s}\gamma$.
- The number of interfering users $J \triangleq |\mathcal{K}(s)| - 1$.
- The MAI instantaneous power-to-noise ratio $Y \triangleq \sum_{k=1}^{J} X_k$ with $Y = 0$ if $J = 0$.
- The instantaneous SINR $Z \triangleq \frac{X_0}{1+Y}$.
- The instantaneous mutual information (IMI) $I \triangleq \log(1+Z)$.

J is binomially distributed as

$$\Pr(J=u) = \binom{K-1}{u}\left(\frac{G}{K}\right)^u\left(1-\frac{G}{K}\right)^{K-u}$$

for $u = 0, \ldots, K-1$. For $K \to \infty$, this converges to the Poisson distribution

$$\Pr(J=u) = \mathrm{e}^{-G}\frac{G^u}{u!}$$

for $u \geq 0$.

Without fading, X_k is constant and equal to γ. Then, Y, Z and I takes on the values $u\gamma$, $\frac{\gamma}{1+u\gamma}$ and $\log(1+\frac{\gamma}{1+u\gamma})$ with probability $\Pr(J=u)$ given above.

In the case of normalized Rayleigh fading, X_k is exponentially distributed with mean γ,

$$F_X(x) = 1 - \mathrm{e}^{-x/\gamma} \tag{2.51}$$

The pdf of Y is readily obtained as a sum of u-fold convolutions of the pdf corresponding to (2.51), weighted by $\Pr(J = u)$. This yields the cdf

$$F_Y(x) = 1 - \sum_{u=1}^{K-1} \Pr(J = u) \sum_{k=0}^{u-1} \mathrm{e}^{-x/\gamma} \frac{(x/\gamma)^k}{k!} \tag{2.52}$$

The derivation of the cdf for the SINR Z is more involved (the details are postponed to the end of this Appendix). We obtain

$$\begin{aligned} F_Z(x) &= 1 - \sum_{u=0}^{K-1} \Pr(J = u) \frac{\mathrm{e}^{-x/\gamma}}{(1+x)^u} \\ &= 1 - \mathrm{e}^{-x/\gamma} \left(1 - p_t \frac{x}{1+x}\right)^{K-1} \end{aligned} \tag{2.53}$$

Finally, the cdf of the IMI I is obtained from (2.53) by a simple change of variable as

$$\begin{aligned} F_I(x) &= 1 - \sum_{u=0}^{K-1} \Pr(J = u) \mathrm{e}^{-(\mathrm{e}^x - 1)/\gamma} \, \mathrm{e}^{-xu} \\ &= 1 - \mathrm{e}^{-(2^x - 1)/\gamma} \left(1 - p_t(1 - 2^{-x})\right)^{K-1} \end{aligned} \tag{2.54}$$

obviously, all the above cdfs are defined for $x \geq 0$ and are zero for $x < 0$.

In the limiting case of $K \to \infty$, the pdf corresponding to (2.52) was found in [36], and it is given by

$$f_Y(x) = \mathrm{e}^{-G} \left[\delta(x) + \mathrm{e}^{-x/\gamma} \sqrt{\frac{G}{x\gamma}} I_1\left(\sqrt{\frac{4xG}{\gamma}}\right)\right] \tag{2.55}$$

where $\delta(x)$ is the Dirac delta function and $I_1(x)$ is the first-order modified Bessel function of the first kind. The SINR cdf for infinite users is given by

$$F_Z(x) = 1 - \exp\left(-\frac{x}{\gamma} - \frac{Gx}{1+x}\right) \tag{2.56}$$

and the corresponding IMI cdf is obtained from (2.56) by a change of variable as

$$F_I(x) = 1 - \exp\left(-\frac{2^x - 1}{\gamma} - (1 - 2^{-x})G\right) \tag{2.57}$$

Calculation of the SINR cdf conditioned on the number of interfering users. Let $\widetilde{X}$ and $\widetilde{Y}$ be two independent RV's obtained as the sum of A and B i.i.d. exponentially distributed RV's with mean $1/\lambda$, respectively. $\widetilde{X}$ and $\widetilde{Y}$ follow the *Gamma* cdf

$$F(x) = 1 - \sum_{k=0}^{N-1} \frac{(\lambda x)^k}{k!} \mathrm{e}^{-\lambda x} \tag{2.58}$$

for $N = A$ and $N = B$, respectively. For an arbitrary $b \geq 0$, consider the RV $\widetilde{Z} = \frac{\widetilde{X}}{b+\widetilde{Y}}$. Notice that, with $b = 1$, $\lambda = 1/\gamma$ and $A = 1$, $\widetilde{Y}$ is $Y|J = B$, $\widetilde{Z}$ is $Z|J = B$ and $\widetilde{X}$ is X. The following derivation generalizes the result obtained in [37]. The cdf of $\widetilde{Z}$ is given by

$$\begin{aligned}
F_{\widetilde{Z}}(z) &= \Pr\{\widetilde{Z} \leq z\} = \Pr\{\widetilde{X} \leq (b+\widetilde{Y})z\} \\
&= \int_{-\infty}^{+\infty} \mathrm{d}y \int_{-\infty}^{z(b+y)} \mathrm{d}x\, f_{\widetilde{X}}(x) f_{\widetilde{Y}}(y) \\
&= \int_{0}^{+\infty} f_{\widetilde{Y}}(y) F_{\widetilde{X}}(z(b+y))\mathrm{d}y \\
&= \int_{0}^{+\infty} \frac{\lambda}{(B-1)!}(\lambda y)^{B-1}\mathrm{e}^{-\lambda y} \left[1 - \sum_{k=0}^{A-1} \frac{(\lambda x)^k}{k!}\mathrm{e}^{-\lambda x}\right]_{x=0}^{z(b+y)} \mathrm{d}y \\
&= 1 - \sum_{k=0}^{A-1}\sum_{\ell=0}^{k} \int_{0}^{+\infty} \frac{\lambda(\lambda y)^{B-1}}{(B-1)!} \frac{(\lambda z b)^{k-\ell}}{(k-\ell)!} \frac{(\lambda z y)^{\ell}}{\ell!} \mathrm{e}^{-\lambda y - \lambda z b - \lambda z y}\mathrm{d}y \\
&= 1 - \frac{\mathrm{e}^{-\lambda z b}}{(1+z)^B} \left[\sum_{k=0}^{A-1}\sum_{\ell=0}^{k} \frac{(\lambda z b)^{k-\ell}}{(k-\ell)!} \left(\frac{z}{1+z}\right)^{\ell} \binom{B-1+\ell}{B-1}\right] \cdot \\
&\quad \cdot \left\{\int_{0}^{+\infty} \frac{\lambda(1+z)}{(B-1+\ell)!} \mathrm{e}^{-\lambda(1+z)y} [\lambda y(1+z)]^{B-1+\ell} \mathrm{d}y\right\}
\end{aligned} \tag{2.59}$$

where the integral in braces in the last line is equal to 1, since the integrand is a Gamma pdf.

For $A = 1$, the double summation in the last line of (2.59) reduces to one, and we obtain

$$F_{\widetilde{Z}}(z) = 1 - \frac{\mathrm{e}^{-\lambda z b}}{(1+z)^B} \tag{2.60}$$

Chapter 3

A system comparison

In this chapter, we compare the performance of systems that implements, at network layer, one of the three different Hybrid-ARQ protocols analyzed in the previous chapter and emploies, at physical layer, different receiver structures. The chosen performance measure is maximum throughput versus average energy per successfully received information bit. In particular, we consider the following single-user decoding based systems: the system introduce in chapter 2 and a random spread CDMA system. We conclude the chapter, and this first part of the thesis, by studying the throughput of retransmission protocols on top of decoders that performs joint decoding of the active users.

3.1 Introduction

In the previous chapter we introduced a simple (unspread) system that, in order to cope against multi-user interference and fading, retransmits erroneously received packets. We analyzed the throughput performance of three different packet combining techniques, referred to as ALO, RTD and INR protocol, as function of several system parameters: the number of users K, the system load G, the delay constraint N, the rate constraint M, the information rate R, the transmit SNR γ and the fading statistic $F_\alpha(x)$. Then, we showed that the throughput is increasing in both N and M and, for an unconstrained system ($M, N \to \infty$), we optimally designed the information rate R. In doing so, we showed that the optimized throughput for INR coincides with the ergodic capacity of the underlying single-user block fading channel, i.e., the ergodic capacity of a fading channel with fading statistics equal to that of the SINR of the considered multiple-user collision channel.

In this chapter, we present an information-theoretic comparison of some not-power-controlled multi-access wireless systems based on single-user decoding in an Hybrid-ARQ prospective. The systems we have chosen for our performance comparison have been considered in a number of previous works ([24, 85, 87] and references therein). These systems are intrinsically quite different: some assume an infinite number of users with vanishing coding rate but with non-zero probability of accessing the channel while others assume "bursty" users with instantaneously non-vanishing coding rate but with vanishing probability of accessing the channel. Here, we do not question the validity of these (quite idealized) models, on the contrary, the related basic results are our starting point for comparison. In order to provide a *fair* performance measure, we compare the systems in terms of their maximum unconstrained throughput versus E_b/N_0, the average transmit energy per correctly received information bit. The system throughput, as a function of E_b/N_0, is given in parametric form and its optimization is often involved. We present analytic closed-form results and very rarely we shall resort to numerical calculation (never to computer simulation). As a byproduct of this analysis, we obtain insight on the optimal choice of system parameters in order to obtain maximum throughput.

As pointed out in the concluding remarks of Chapter 2, a natural direction to extend the present analysis is to consider joint decoding at the receiver (e.g., implemented by stripping). A theoretical difficulty is represented by the user random activity [18]. In fact, because of random access, the capacity region varies from slot to slot and it is not known in advance, unless a complicated reservation/allocation scheme is implemented. Also, the set of interfering users might be different from slot to slot, and it is not clear how to carry out joint decoding across the slots. We conclude the chapter, and the first part of the thesis, by analyzing two ALO-based systems that jointly decode the largest possible subset of active users: the first is a Successive Interference Canceler (based on stripping) with Single-User Decoding at each decoding step (SIC-SUD) and the second is a Joint Multi User Decoder (JMUD) [87]. Finally, we analyze the throughput of an INR strategy on top of a joint decoder that either decode all the active users on none, thus providing a lower bound to any clever joint decoding INR strategy on top of random user activity.

We begin by considering the unspread system analyzed in Chapter 1, where a population of users access at random a channel and use retransmission as the only mean to combat fading and interference form other users. Then we turn to a random spread Direct Sequence Code Division Multiple Access (CDMA) system with linear detectors [85], namely, Single User Matched Filter (SUMF) and Minimum Mean Square Error (MMSE). In this channel model, users are assigned randomly and independently signature sequences that are assumed known at the receiver. In general, the SINR at the output of the linear filter depends on the correlation between the signa-

ture sequences of all the users. In a system with infinitely long spreading sequence and infinitely many users, the output SINR of any user is only function of a scaled version of the fading experienced by the user itself. In [85], the authors study the total capacity per second per Hz by characterizing the asymptotic performance in low E_b/N_0 regime and high E_b/N_0 regime. Here, we analyze the throughput performance in an Hybrid-ARQ prospective. After that, we concentrate on the joint decoding systems.

We show that the unspread system outperforms SUMF-CDMA, which is throughput-wise limited, but it is outperformed by MMSE-CDMA. All the systems have the same behavior in terms of throughput and of optimal system parameters. In the low E_b/N_0 regime, the optimized throughput is the same for all the systems and coincides with that of a SUMF-CDMA, achieved by an infinite number of users per degree of freedom transmitting at vanishing rate. In the high E_b/N_0 regime, while SUMF-CDMA is interference limited, the other systems are not. For this range of E_b/N_0 the optimized systems "self-orthogonalize", in the sense that optimal throughput is achieved by having on the average only one user per degree of freedom, i.e., one user per chip for the CDMA and one active user per slot for the unspread system. All the SUD-based systems are outperformed by MUD-based systems. In particular, we find that an INR protocol that either decodes all users or asks to all the users to retransmit is throughput-wise optimal, in the sense that it achieves the the ergodic rate-sum of the underlying block-fading multiple-access channel. This shows that an optimal INR strategy "forces" the user to transmit together, i.e., the system load is equal to the number of users, and does not attempt to decode a subset of the active users.

The rest of the chapter is organized as follows: in Section 3.2 we briefly revise the system model introduced in Chapter 1 to account for spreading; in Section 3.3 we derive the throughput versus E_b/N_0 curves for the unspread system and for the random spread CDMA system in conjunction with ALO, RTD and INR protocols; in Section 3.4 we analyze throughput versus E_b/N_0 for systems based on joint-decoding and in Section 3.5 we point out our conclusions.

Our publications related to this chapter are:
[26] D.Tuninetti and G.Caire, *"The optimal throughput of some wireless multi-access systems"*, in Proceedings 2001 IEEE International Symposium on Information Theory (ISIT2001), Washington DC (USA), June 2001;
[27] D.Tuninetti and G.Caire, *"The optimal throughput of some wireless multi-access systems"*, to appear in IEEE Transactions on Information Theory.

3.2 System model

In this section we generalize the system model introduced in the previous chapter in order to include systems with spreading. The system of Chapter 2 is then obtained as a special case of this more general model.

We assume a complex channel of bandwidth W Hz whose time axis is divided in slots of duration T seconds. Every slot can accommodate packets of $L \approx WT$ independent complex dimensions, in the limit of $WT \gg 1$. The channel is impaired by additive noise and by frequency-flat fading, assumed constant for the whole slot duration and independent for each slot and each user. The channel is accessed randomly and independently, with probability p_t on every slot, by a population of K users.

Each transmitter has an infinite sequence of information packets to encode, spread and transmit. Let S be the length of the spreading sequence and M be a given integer. User k encodes its packet of b data bits into a code word of length LM/S complex symbols and spreads it, so that each data packet corresponds to a channel packet of LM modulation symbols (chips) to send over the channel. Before actual transmission the channel packet is split into M "chunks" to fit the slot duration. The transmitters have no Channel State Information (CSI), hence users transmit at constant rate and constant power. Let E be the transmit energy per channel symbol, N_0 be the noise spectral density and $\gamma = E/N_0$ be the transmit SNR. Note that in the sequel, depending on the context, we shall use the terms "complex symbol", "degree of freedom", "dimension" or "chip" to indicate a basic channel use, i.e., one second per Hz. The quantity $R \stackrel{\Delta}{=} bS/L$ is referred to as the *information rate* and represents the number of data bit per coded symbol before spreading. The quantity $G \stackrel{\Delta}{=} K\, p_t/S$ is referred to as *channel load* and represents the average number of active users per dimensions. The system analyzed in Chapter 2 is obtained for $S = 1$;

Each time a transmitter is active, it sends on the current slot the not-yet transmitted chunk of L chips of the current channel packet. At the receiver, the sequence of slots where a user was active are collected, combined and used for decoding. On every slot, decoding of all active users is attempted. Then, via an error and delay free feedback link, an ACK is sent to all users for which decoding has been successful, while a NACK is sent to all users for which decoding has not been successful. When a user gets an ACK, it stops transmitting the current channel packet and the next time it is active it starts transmitting the first chunk of the next channel packet. On the contrary, when a user gets a NACK, the next time it is active it transmits the next chunk of the current channel packet. If successful decoding is not obtained after M transmitted chunks, or after N slots since the generation of the data packet, the packet is lost. M and N are referred to as the *rate constraint* and *delay constraint*, respectively. Three Hybrid-ARQ schemes are taken

into account: an ALOha-type scheme (ALO), where channel packets are made of the same basic code words of length L repeated M times, and where previously received chunks are discarded; a Repetition Time-Diversity scheme (RTD), where channel packets are also made of the same basic code words of length L repeated M times, but where previously received chunks are combined by *maximal ratio combining* before decoding; an INcremental Redundancy scheme (INR), where channel packets are effectively made of M different segments of L symbol each, and previously received chunks are all taken into account at each decoding attempt.

The receiver for each user is formed by a chip matched filter (the chip pulse-shaping waveform is assumed to be a Nyquist pulse), a sampler at chip rate, a linear filter, a packet combiner and a decoder. By stacking in a vector the S chips referring to the same coded symbol, the discrete-time model for the received signal (at symbol rate) is $\boldsymbol{y} = \sum_{k\in\mathcal{K}} c_k \boldsymbol{s}_k x_k + \boldsymbol{z}$ where $\boldsymbol{z} \in \mathbb{C}^S$ is a proper complex Gaussian noise vector of zero mean and per-component variance N_0, $\mathcal{K}$ is the set of active users during the current channel use, $\boldsymbol{s}_k \in \mathbb{C}^S$ is the spreading sequence of user k, $x_k \in \mathbb{C}$ is the k-th user transmit modulation symbol and $c_k \in \mathbb{C}$ is the k-th user fading amplitude. Let $\alpha_k \stackrel{\Delta}{=} |c_k|^2$ be the fading power, that we assume i.i.d. with cumulative distribution function (cdf) $F_\alpha(x)$. The receiver has perfect CSI, i.e., it knows the set of active users $\mathcal{K}$, the fading gains c_k and the spreading sequences $\boldsymbol{s}_k$ for all $k \in \mathcal{K}$. Without any claim of optimality, we assume that users employ Gaussian random codes and that the decoder is based on typical set decoding.

The systems under investigation belong to one of the following classes:
Unspread SUD-based systems. This systems are analyzed in [24]. In this case, $S = 1$ and $\boldsymbol{s}_k = 1$ for all k. For an infinite population of users ($K \to \infty$) and for all finite G, the probability p_t that a user transmit on any given slot is vanishing and the number of active users, i.e., the cardinality of the active set $\mathcal{K}$, is a Poisson distributed random variable.
Unspread MUD-based systems. Those systems have $S = 1$ and $\boldsymbol{s}_k = 1$ for all k. Following [87], we assume an infinite population of users that transmit with probability $p_t = 1$, i.e., $G = K \to \infty$, this results in a finite throughput (aggregate rate) and a vanishing per-user rate.
CDMA system with random spreading. In this case, following [69, 70, 85, 86], we assume $S, K \to \infty$ and $K/S \to d$, where d is the maximum number of active users per chips. The user spreading sequences $\boldsymbol{s}_k$ are random generated with i.i.d. components drawn according to an arbitrary probability assignment with zero mean, variance $1/S$ and bounded forth moment. Transmission follows again the same Hybrid-ARQ scheme described before. The linear filter is either a SUMF or a linear MMSE filters [89]. The channel load is given by $G = p_t d$, hence for all finite G and d, p_t is non-vanishing.

As we already pointed out in the introduction, those systems are intrinsi-

cally quite different. Here, we are interested in studying the performance of repetition protocols on top of different physical layer access techniques and not in questioning the system model validity or their practical applicability.

3.3 Optimized throughput

For a symmetric system with respect to any user, the unconstrained system throughput (expression (2.19) in Section 2.5) is given by

$$\eta = \frac{RG}{1+\sum_{m=1}^{\infty} \Pr\left(I(1) \leq R, \cdots, I(m) \leq R\right)} \tag{3.1}$$

where $I(m)$ is the accumulated mutual information between the receiver output and the transmitter after m received slots and it is given by

$$I(m) = \begin{cases} \sum_{s=1}^{m} \log(1+\beta_s) & \text{INR} \\ \log(1+\sum_{s=1}^{m} \beta_s) & \text{RTD} \\ \log(1+\beta_m) & \text{ALO} \end{cases} \tag{3.2}$$

for β_s being the SINR of the reference user in slot s. For the different systems we have

$$\beta_s = \begin{cases} \dfrac{\gamma\alpha_k}{1+\displaystyle\sum_{i\in\mathcal{K}:i\neq k} \gamma\alpha_i} & \text{Unspread system} \\ A\alpha & \text{SUMF} - \text{CDMA with} \quad A = \dfrac{\gamma}{1+G\gamma\,\mathrm{E}[\alpha]} \\ A\alpha & \text{MMSE} - \text{CDMA with} \quad A : \gamma = \dfrac{A}{1-G\,\mathrm{E}\left[\dfrac{A\alpha}{1+A\alpha}\right]} \end{cases} \tag{3.3}$$

The SINR expressions for the CDMA system are derived in [85].

As already pointed out, the throughput (3.1) optimized with respect to R coincides with the *ergodic single-user capacity* of the underlying block-fading channel for the INR protocol

$$\eta^{(\mathrm{INR})} = G\ \mathrm{E}\left[I_1\right] \tag{3.4}$$

and with the *outage single-user rate* for the ALO protocol

$$\eta^{(\mathrm{ALO})} = G \sup_{R\geq 0} R\ \Pr(I_1 > R) \tag{3.5}$$

with $\Pr(I_1 \leq R) = p(1)$ being the information outage probability. Note that $\eta^{(\mathrm{INR})} = \eta^{(\mathrm{ALO})}$ only if the mutual information I_1 is a deterministic constant. In general, $\eta^{(\mathrm{INR})} \geq \eta^{(\mathrm{RTD})} \geq \eta^{(\mathrm{ALO})}$.

Since the users transmit for a fraction p_t of the time and the average number of *received* information bits per modulation symbol per user is $S\eta/K$,

the average energy per received information bit is $E_b = (Ep_t)/(S\eta/K) = EG/\eta$. Hence, the user SNR γ is related to E_b/N_0 by

$$\gamma = \frac{E_b}{N_0}\frac{\eta}{G} \tag{3.6}$$

Notice the difference between this definition of E_b/N_0 and the common definition used by coding theorists in a single-user channel. There, E_b is the energy per *transmitted* information bit, irrespectively of error probability, i.e., on the fraction of erroneously received bits. Here, E_b denotes energy per successfully delivered information bit at the receiver, which in a multiuser channel prone to collisions and packet loss is a more sensible definition. For the sake of notation simplicity, in the rest of the chapter we use natural logarithms. Hence, η will be expressed in nat/s/Hz and we will use the notation E_n/N_0, instead of E_b/N_0, to indicate that E_n is the energy per received nat. All results can be readily translated in more usual bit/s/Hz vs. E_b/N_0 recalling that 1 nat=$\log_2$ (e) bits and that $\log(2) = -1.5917$ dB.

Next, we derive the throughput versus E_n/N_0 curves. After the optimization of (3.1) with respect to the information rate R (see Section (2.6)), the throughput is a function of G and γ, i.e., $\eta = h(G,\gamma)$. By substituting (3.6) into the throughput function we obtain the implicit equation $\eta = h(G,(E_n/N_0)(\eta/G))$ whose explicit solution, i.e., $\eta = g(G, E_n/N_0)$ that we shall refer to as *spectral efficiency*, is generally not available in closed form. The curve we are interest in is the optimized spectral efficiency, i.e., $\eta = \sup_{G\geq 0} g(G, E_n/N_0)$. Notice that, while the throughput $h(G,\gamma)$ is non-zero for every $\gamma > 0$, the spectral efficiency $g(G, E_n/N_0)$ is non-zero only for $E_n/N_0 > (E_n/N_0)_{\min}$ [85], where $(E_n/N_0)_{\min}$ is given by

$$\left(\frac{E_n}{N_0}\right)_{\min} = \lim_{\gamma\to 0^+}\frac{G\gamma}{h(G,\gamma)} = \left(\lim_{\gamma\to 0^+}\frac{1}{G}\frac{\partial h(G,\gamma)}{\partial\gamma}\right)^{-1} \tag{3.7}$$

We shall see that (3.7) is function of the protocol only and not of the system implemented at physical layer. In [14] it is shown that the inverse of $(E_n/N_0)_{\min}$ has the meaning of capacity per unit energy.

3.3.1 The single-user system

We start by analyzing a reference single-user system.

For the ALO protocol the throughput is given by

$$\eta = p_t \sup_{R\geq 0}\left\{R\left[1 - F_\alpha\left(\frac{\mathrm{e}^R - 1}{\gamma}\right)\right]\right\} \tag{3.8}$$

clearly maximized for $p_t = 1$, i.e., $G = 1$, since the optimal value of R only depends on γ. By noting that the maximization over R can be re-written as

$$\eta = \sup_{\theta\geq 0}\left\{\log(1+\gamma\theta)\left[1 - F_\alpha(\theta)\right]\right\} \tag{3.9}$$

by letting $R = \log(1+\gamma\theta)$, it follows after some algebra that

$$\lim_{\gamma\to 0}\frac{\mathrm{d}\eta}{\mathrm{d}\gamma} = \sup_{\theta\geq 0}\theta[1-F_\alpha(\theta)] \tag{3.10}$$

Then, the minimum value of E_n/N_0 is given by

$$\left(\frac{E_n}{N_0}\right)_{\min} = \frac{1}{\sup_{\theta\geq 0}\theta[1-F_\alpha(\theta)]} \tag{3.11}$$

For the RTD protocol the throughput is given by

$$\begin{aligned}\eta &= p_t \sup_{R\geq 0}\frac{R}{1+\sum_{m=1}^{\infty}\Pr\left[\sum_{s=1}^{m}\alpha_s \leq \frac{\mathrm{e}^R-1}{\gamma}\right]} \\ &= p_t \sup_{\theta\geq 0}\frac{\log(1+\gamma\theta)}{1+\sum_{m=1}^{\infty}\Pr\left[\sum_{s=1}^{m}\alpha_s \leq \theta\right]}\end{aligned} \tag{3.12}$$

again maximized for $p_t = 1$. The fist derivative with respect to γ at $\gamma = 0$ is given by

$$\begin{aligned}\lim_{\gamma\to 0}\frac{\mathrm{d}\eta}{\mathrm{d}\gamma} &= \sup_{\theta\geq 0}\frac{\theta}{1+\sum_{m=1}^{\infty}\Pr\left[\sum_{s=1}^{m}\alpha_s \leq \theta\right]} \\ &= \mathrm{E}[\alpha]\end{aligned} \tag{3.13}$$

where the last equality follows from Appendix 2.C. Hence, the minimum E_n/N_0 is

$$\left(\frac{E_n}{N_0}\right)_{\min} = \frac{1}{\mathrm{E}[\alpha]} \tag{3.14}$$

Note that $(E_n/N_0)_{\min}^{(\mathrm{ALO})} \geq (E_n/N_0)_{\min}^{(\mathrm{RTD})}$ with equality if and only if α is a deterministic constant, i.e., channel without fading. In this case, the throughput curves of all the three protocols, and hence the spectral efficiency curves, coincide.

For the INR protocol the maximum throughput with respect to R (see Appendix 2.C) is given by

$$\eta = p_t\,\mathrm{E}[\log(1+\gamma\alpha)] \tag{3.15}$$

which is maximized by $p_t = 1$ and gives the same $(E_n/N_0)_{\min}$ of RTD protocol. Interestingly, to achieve capacity per unit energy, i.e., $(E_n/N_0)_{\min}$, it is enough to adopt the simpler RTD strategy instead of the more complex INR strategy.

Examples. In the examples we consider two fading statistics: constant fading, i.e., $F_\alpha(x) = 1\{x \geq 0\}$, and Rayleigh normalized fading, i.e., $F_\alpha(x) = (1 - \mathrm{e}^{-x})1\{x \geq 0\}$. For all the protocols and for the two fading statistics considered here, except for ALO, we have $(E_n/N_0)_{\min} = 1 = 0\mathrm{dB}$. For ALO in Rayleigh fading we have $(E_n/N_0)_{\min} = \mathrm{e} = 4.3429\mathrm{dB}$.

The optimization of the throughput with respect to the information rate R (see Appendix 3.A) yields to the following parametric expressions of the spectral efficiency curve:

- ALO/RTD/INR without fading: for $\gamma \geq 0$

$$\begin{cases} \eta & = \log(1+\gamma) \\ \dfrac{E_n}{N_0} & = \dfrac{\gamma}{\eta} \end{cases} \tag{3.16}$$

- ALO with Rayleigh fading: for $R \geq 0$

$$\begin{cases} \eta & = R\mathrm{e}^{-\dfrac{\mathrm{e}^R - 1}{R\mathrm{e}^R}} \\ \dfrac{E_n}{N_0} & = \dfrac{R\mathrm{e}^R}{\eta} \end{cases} \tag{3.17}$$

- RTD with Rayleigh fading: for $R \geq 0$

$$\begin{cases} \eta & = \mathrm{e}^R + R - 1 \\ \dfrac{E_n}{N_0} & = \dfrac{1 + (R-1)\mathrm{e}^R}{\eta} \end{cases} \tag{3.18}$$

- INR with Rayleigh fading: for $\gamma \geq 0$

$$\begin{cases} \eta & = \mathrm{e}^{1/\gamma}\mathrm{Ei}(1/\gamma) \\ \dfrac{E_n}{N_0} & = \dfrac{\gamma}{\eta} \end{cases} \tag{3.19}$$

Fig. 3.1 shows the throughput η vs. E_n/N_0 for the single user system. Notice that, without fading, all the three protocols have the same spectral efficiency which is the upperbound to the throughput of any multi-access system without power control, as we shall see later. It is interesting to notice that the major benefits of RTD with respect to ALO occur at low E_n/N_0.

3.3.2 Unspread system

For the ALO protocol the throughput of the unspread system is given by

$$\begin{aligned} \eta &= G \sup_{R \geq 0} \left\{ R \Pr\left[\frac{\alpha_k}{1 + \sum_{i \in \mathcal{K}: i \neq k} \gamma \alpha_i} > \frac{\mathrm{e}^R - 1}{\gamma} \right] \right\} \\ &= G \sup_{\theta \geq 0} \left\{ \log(1+\gamma\theta)\, \mathrm{E}\left[1 - F_\alpha\left(\theta + \theta\gamma \sum_{i \in \mathcal{K} \neq k} \alpha_i \right) \right] \right\} \end{aligned} \tag{3.20}$$

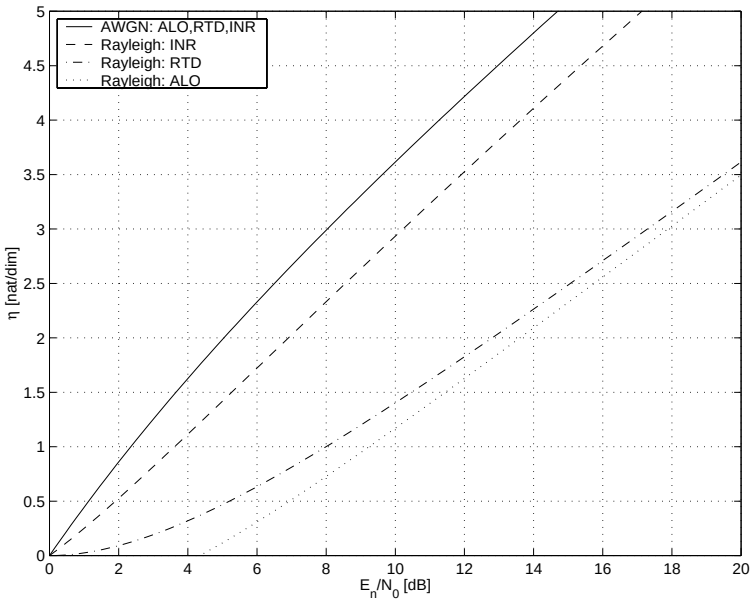

Figure 3.1: Throughput versus E_n/N_0 for a single user system.

and for vanishing γ we have

$$\eta = G \sup_{\theta \geq 0} \left\{ \gamma\theta \, \mathrm{E}\left[1 - F_\alpha(\theta)\right] + o(\gamma^2) \right\} \tag{3.21}$$

Hence, also for the unspread system with ALO we get

$$\left(\frac{E_n}{N_0}\right)_{\min} = \frac{1}{\sup_{\theta \geq 0} \theta[1 - F_\alpha(\theta)]} \tag{3.22}$$

In the same way, it is easy to see that the RTD and INR for the unspread system achieve

$$\left(\frac{E_n}{N_0}\right)_{\min} = \frac{1}{\mathrm{E}[\alpha]} \tag{3.23}$$

as in the single user case. Interestingly, $(E_n/N_0)_{\min}$ is function of the protocol and the fading statistics, but not of the system, and is the same for every G, i.e., the same minimum energy per bit can be achieved by systems with non optimized G.

Examples. Again, we give examples of the optimized spectral efficiency curves for the case of channels without fading and channels with Rayleigh fading. We assume that in the system there are infinite users, i.e., $K \to \infty$, with vanishing probability of transmitting, thus resulting in a finite channel

load G. The number of active users per slot follows a Poisson distribution, i.e., $\Pr(k \text{ active users}) = \mathrm{e}^{-G}\frac{G^k}{k!}$.

- ALO without fading

$$\eta = \sup_{J\geq 0}\left[\log\left(1+\frac{\gamma}{1+J\gamma}\right)\sum_{k=0}^{J} G\mathrm{e}^{-G}\frac{G^k}{k!}\right] \tag{3.24}$$

  It is easy to see that the supremum of η in (3.5), for fixed G and γ, is always obtained when $R = \log(1+\frac{\gamma}{1+J\gamma})$ for some integer J, where $J+1$ is the maximum number of users that can collide on the same slots without causing a decoding error. Therefore, maximizing with respect to R is equivalent to searching for the maximum of the expression $G\log(1+\frac{\gamma}{1+J\gamma})\Pr(J \text{ interferes})$ over the non-negative integers J.

- RTD without fading

$$\eta = G\sup_{R\geq 0}\frac{R}{1+\sum_{m=1}^{\infty}\Pr\left[\sum_{i=1}^{m}\beta_i \leq \mathrm{e}^R - 1\right]} \tag{3.25}$$

  where β_i are i.i.d. $\forall i$ with probability mass function $\Pr\left(\beta_i = \frac{\gamma}{1+k\gamma}\right) = \mathrm{e}^{-G}G^k/k!$ for $k \geq 0$. In this case we optimization over R and G has to be performed numerically.

- INR without fading

$$\eta = G\sum_{k\geq 0}\mathrm{e}^{-G}\frac{G^k}{k!}\log\left(1+\frac{\gamma}{1+k\gamma}\right) \tag{3.26}$$

  since in this case the optimization of the throughput over R yields (3.4).

- ALO with Rayleigh fading

$$\eta = G\sup_{R\geq 0} R\exp\left(-\frac{\mathrm{e}^R-1}{\gamma} - G(1-\mathrm{e}^{-R})\right) \tag{3.27}$$

  where the cumulative distribution function (cdf) of the SINR β with Rayleigh fading and a Poisson distributed number of interfering users was derived in Appendix 2.E.

- RTD with Rayleigh fading.
  In this case the throughput is again given by (3.25), but β_i are i.i.d. $\forall i$ with cdf $F_\beta(x) = [1-\exp(-x/\gamma - Gx/(1+x))]$ (see Appendix 2.E).

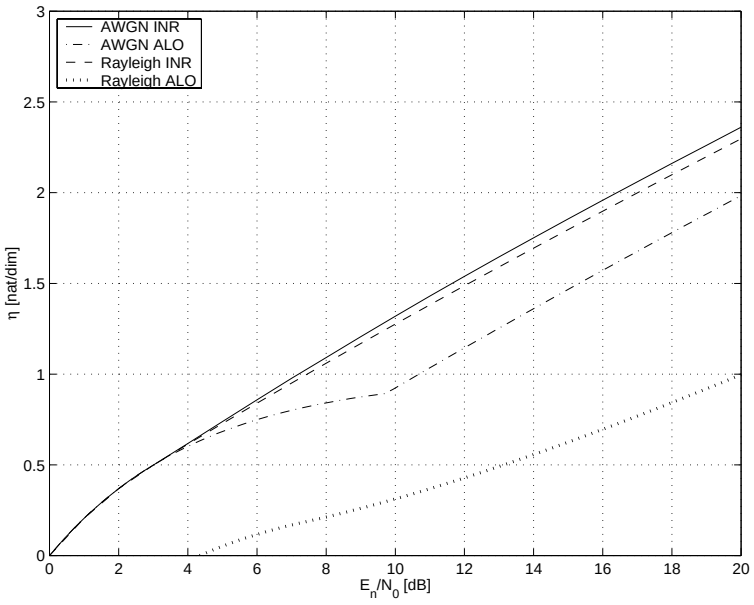

Figure 3.2: Throughput versus E_n/N_0 for an unspread system.

- INR with Rayleigh fading

$$\eta = G \int_0^\infty \exp\left(-\frac{\mathrm{e}^x - 1}{\gamma} - G(1 - \mathrm{e}^{-x})\right) \mathrm{d}x \tag{3.28}$$

here again, the maximization over R yields (3.4) and the distribution of the SINR β was derived in Appendix 2.E.

Fig. 3.2 shows the throughput η versus E_n/N_0 for the unspread system, with ALO and INR protocols only. The RTD case was not evaluated because of its complexity, due to the fact that for a given pair (G, γ) the cdf of $\sum_{i=1}^m \beta_i$ for all $m \geq 1$ needs to be computed. In Chapter 2 it is shown that RTD lies between ALO and INR. Fig. 3.3 shows the inverse of optimal G, i.e., the number of degree of freedom per user, versus E_n/N_0 for the unspread system with ALO and INR protocol.

A parametric closed form expression of the optimized throughput can be found for the ALO protocol with Rayleigh fading only. The derivation is reported in Appendix 3.B. By carrying out the optimization over G, it emerges that there exists an certain interval of E_n/N_0, the interval $E_n/N_0 \in [(E_n/N_0)_{\min}, (E_n/N_0)_{\text{th}}]$, for which η is maximized by $G \to \infty$ (see Appendix 3.B). In this interval, the maximum throughput is attained by infinite number of users per degree of freedom transmitting at vanishing

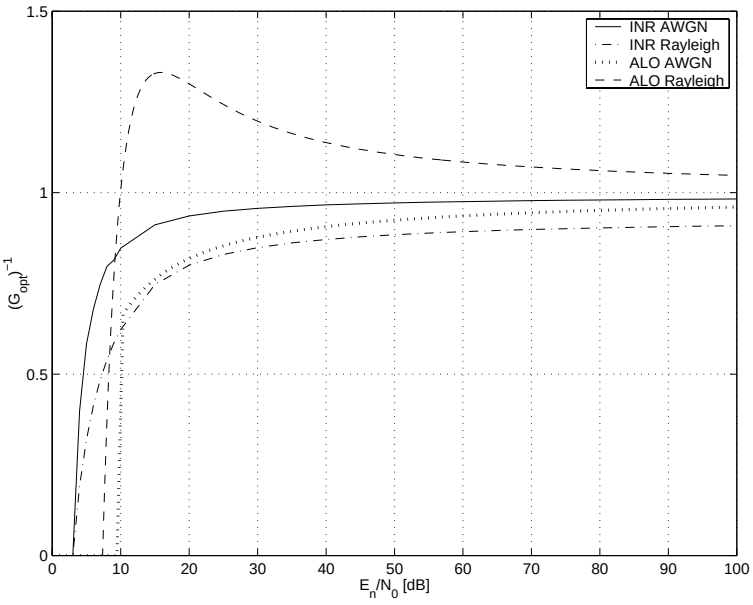

Figure 3.3: Inverse of optimal G versus E_n/N_0 for an unspread system with ALO and INR protocol.

rate ($G \to \infty$ implies $R \to 0$) and it is given by

$$\eta = \left(\frac{E_n}{N_0}\right)^{-1}_{\min} - \left(\frac{E_n}{N_0}\right)^{-1} \tag{3.29}$$

In the interval $E_n/N_0 > (E_n/N_0)_{\text{th}}$, the throughput is maximized by a finite G and non-zero R and furthermore, as $E_n/N_0 \to \infty$, optimal G tends to $G \to 1$. This means that maximum throughput is obtained by having on the average only one active user per degree of freedom transmitting at non-vanishing rate. The introduction of the parameter $(E_n/N_0)_{\text{th}}$, makes unambiguous the expressions "low E_n/N_0", that refers to the interval $E_n/N_0 \in [(E_n/N_0)_{\min}, (E_n/N_0)_{\text{th}}]$, and "large E_n/N_0", that refers to the interval $E_n/N_0 > (E_n/N_0)_{\text{th}}$. In the case of ALO with Rayleigh fading, we have $(E_n/N_0)_{\text{th}} = 2\text{e} = 7.3532$ dB.

The curve for ALO without fading was obtained via the numerical technique described in Appendix 3.C. It presents a change in slope at $(E_n/N_0)_{\text{th}} = 9.7305$ dB due to the fact that optimization over J in (3.24) gives either $J = 0$ for $E_n/N_0 > (E_n/N_0)_{\text{th}}$ or $J \to \infty$ for $E_n/N_0 \leq (E_n/N_0)_{\text{th}}$. This means that for low E_n/N_0 users must encode their messages at vanishing rate and transmit all the time, while for large E_n/N_0 users must encode their messages with non-vanishing rate and transmit (on the average) one at a time. This effect of self-orthogonalization of the optimized system is shown in Fig. 3.3.

The same numerical technique described in Appendix 3.C was used to obtain the curves for the INR protocol with and without fading. In both cases the E_n/N_0 threshold is $(E_n/N_0)_{\text{th}} = 2 = 3.0103$ dB and optimal G tends to $G \to 1$ as $E_n/N_0 \to \infty$. Again, this effect of self-orthogonalization is shown in Fig. 3.3.

3.3.3 CDMA system with random spreading

This system has been analyzed in [69, 70, 85]. The system is "single-user" like, in the sense that the SINR is function only of the fading experienced by the user itself, while the random nature of the system, i.e., random activity, fading and spreading sequences, is taken into account by the deterministic (in the limit for large K) constant A given in (3.3). For this reason, we can apply the results obtained in Appendix 3.A for the "single-user like" case for what concerns the maximization over the information rate R and for the determination of $(E_n/N_0)_{\min}$. The maximization over G is more complicated because A is also function of G.

From the expression of A in (3.3) we have that for SUMF

$$A = \gamma + o(\gamma) \quad \text{for} \quad \gamma \ll 1$$

while for MMSE

$$\gamma = A + o(A) \quad \text{for} \quad A \ll 1$$

Hence, for both systems $A \approx \gamma$ for $\gamma \to 0$, which implies $\beta = A\alpha \approx \gamma\alpha$, i.e., in the limit for small γ the random spread CDMA is equivalent to the single user system. We conclude that that $(E_n/N_0)_{\min}$ for ALO is given by (3.11) and for RTD and INR is given by (3.14).

In order to obtain the explicit expressions of the spectral efficiency curves, we first write the throughput of the three protocols as $\eta = GA\, g(A)$ where $g(A)$ depends on the protocol and is given by

$$g(A) = \begin{cases} \max\limits_{\theta \geq 0} \left\{ \dfrac{\log(1+A\theta)}{A\theta}\, \theta[1 - F_\alpha(\theta)] \right\} & \text{ALO} \\ \max\limits_{\theta \geq 0} \left\{ \dfrac{\log(1+A\theta)}{A\theta}\, \dfrac{\theta}{1 + \sum\limits_{m=1}^{\infty} \Pr\left[\sum\limits_{i=1}^{m} \alpha_i \leq \theta \right]} \right\} & \text{RTD} \\ \dfrac{\mathrm{E}[\log(1+A\alpha)]}{A} & \text{INR} \end{cases} \tag{3.30}$$

For all the three protocols, $g(A)$ is non-increasing in $A \geq 0$ and achieves it maximum for $A = 0$. Interestingly, $g(0) = (E_n/N_0)^{-1}_{\min}$. By substituting

$(\eta/G)(\mathrm{E}_n/N_0) = \gamma$ into the expression of A in (3.3), we obtain an expression of G parameterized by A

$$G = \frac{1}{A} \frac{\eta \dfrac{E_n}{N_0}}{1 + f(A)\, \eta \dfrac{E_n}{N_0}} \tag{3.31}$$

where $f(A)$ depends on the system and it is given by

$$f(A) = \begin{cases} \mathrm{E}[\alpha] & \text{SUMF} \\ \mathrm{E}\left[\dfrac{\alpha}{1 + A\alpha}\right] & \text{MMSE} \end{cases} \tag{3.32}$$

Also $f(A)$ is non-increasing in A and achieves it maximum for $A = 0$.

Seen the one-to-one relationship between G and A, the maximization of the spectral efficiency over $G \geq 0$ can be turned over the maximization over $A \geq 0$. Finally, with straightforward algebra, we get to

$$\eta = \sup_{A \geq 0} \frac{g(A) - \left(\dfrac{E_n}{N_0}\right)^{-1}}{f(A)} \tag{3.33}$$

that for SUMF, with any protocol, reduces to

$$\eta = \left[\left(\frac{E_n}{N_0}\right)^{-1}_{\min} - \left(\frac{E_n}{N_0}\right)^{-1}\right] \frac{1}{\mathrm{E}[\alpha]} \tag{3.34}$$

achieved for $A = 0$, which is equivalent to $G \to \infty$.

Examples. The functions $g(A)$ for the different protocols are

- ALO/RTD/INR without fading

$$g(A) = \frac{1}{A} \log(1 + A) \tag{3.35}$$

- ALO with Rayleigh fading

$$g(A) = \mathrm{e}^{-x - \frac{\mathrm{e}^x - 1}{x\mathrm{e}^x}}\Big|_{x\mathrm{e}^x = A} \tag{3.36}$$

- RTD with Rayleigh fading

$$g(A) = \mathrm{e}^{-x}\big|_{1 + (x-1)\mathrm{e}^x = A} \tag{3.37}$$

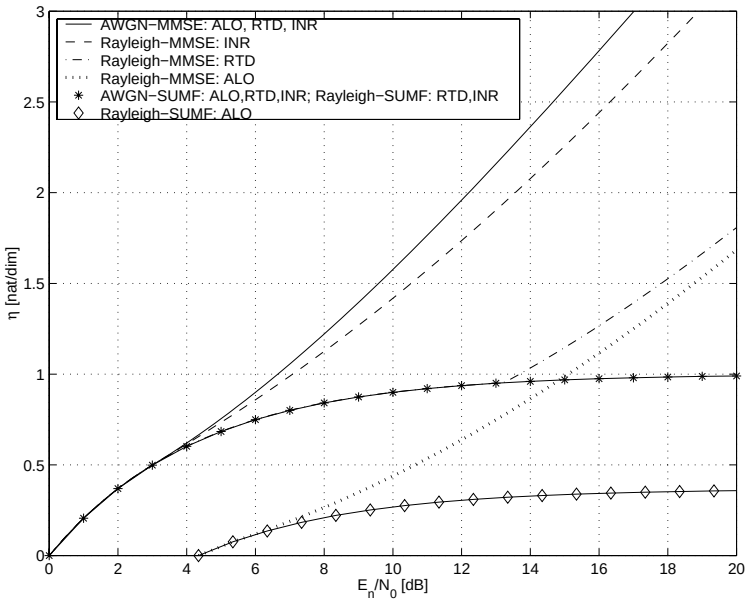

Figure 3.4: Throughput versus E_n/N_0 for a CDMA system with random spreading.

- INR with Rayleigh fading

$$g(A) = \frac{1}{A}\mathrm{e}^{1/A}\mathrm{Ei}\left(\frac{1}{A}\right) \tag{3.38}$$

and the functions $f(A)$ for the MMSE-CDMA system are

- Without fading

$$f(A) = \frac{1}{1+A} \tag{3.39}$$

- Without fading

$$f(A) = \frac{1}{A}\left[\frac{1}{A} - \mathrm{e}^{1/A}\mathrm{Ei}\left(\frac{1}{A}\right)\right] \tag{3.40}$$

and $f(A) = 1$ for SUMF-CDMA.

Fig. 3.4 shows the throughput η versus E_n/N_0 for CDMA with random spreading.

Fig. 3.5 shows the inverse of optimal G versus E_n/N_0 for the CDMA with MMSE. We did not report the curves for CDMA with SUMF because in this case optimal G is $G \to \infty$ $(A = 0)$ for all protocols, for every fading statistics

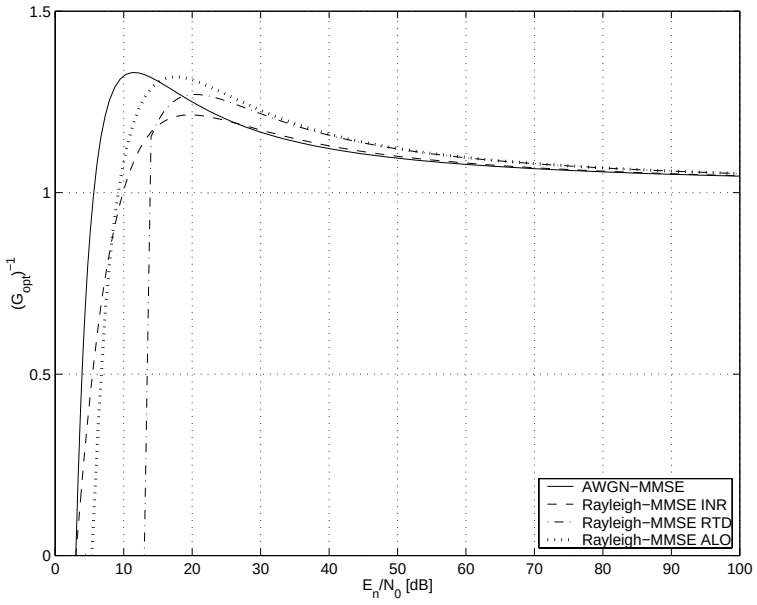

Figure 3.5: Inverse of optimal G versus E_n/N_0 for an MMSE-CDMA system with random spreading.

and for every $E_n/N_0 > (E_n/N_0)_{\min}$. Again, $G \to \infty$ for all E_n/N_0 means that η is maximized by infinite users per chip transmitting with vanishing coding rate.

For MMSE (see Appendix 3.D), we have $(E_n/N_0)_{\text{th}} = 2\text{e} = 7.3532$ dB for ALO with Rayleigh fading, $(E_n/N_0)_{\text{th}} = 13.18$ dB for RTD with Rayleigh fading, $(E_n/N_0)_{\text{th}} = 2 = 3.0105$ dB for INR with Rayleigh fading and all the protocols without fading. For $E_n/N_0 > (E_n/N_0)_{\text{th}}$ there exists a finite G optimizing the throughput with $G \to 1$ as $E_n/N_0 \to \infty$. This effect of self-orthogonalization of the optimized system is shown in Fig. 3.5. Notice that, for sufficiently high E_n/N_0, the optimal G for MMSE tends to one from below. In fact, as pointed out in [85], at high E_n/N_0 the MMSE system is interference limited if $G\Pr[\alpha > 0] > 1$.

3.4 Unspread system with joint decoding

In this last part of the chapter, we study the performance of three systems based on joint decoding.

We first consider the two systems described and analyzed in [87]. The setting is analogous to the unspread system with continuous user activity, i.e., $p_t = 1$, but users transmit and are decoded on a strict slot-by-slot basis (in our terminology, it is an ALO protocol). The decoder is either a

Successive Interference Canceler (stripping) with Single-User Decoding at each decoding step (SIC-SUD) or a Joint Multi User Decoder (JMUD). In both cases, users are ranked in decreasing received SNR order. The SIC-SUD strips users once at a time starting from the strongest and the stripping process continues until it is possible to decode user reliably. The JMUD attempts to decode all K users. If decoding is not successful, it treats the weakest user as noise and attempts the decoding of the remaining $K-1$ users. It proceeds in this way until successful decoding of a subset of users occurs. While in an ergodic setting SIC-SUD and JMUD are equivalent, in the sense they both achieve maximum throughput, in the realm of outage (our setting) they perform differently.

Here, we set $p_t = 1$ because it is not clear how to define packet combining techniques in conjunction with joint decoding. In fact, when user access at random the channel, the set of active user differs from slot to slot.

While in the previous sections we have optimized the spectral efficiency of SUD-based system with respect to the information rate R and to the channel load $G = K\,p_t$, here the channel load is fixed to be to the equal the number of users in the system. Hence, for these MUD-based systems, the information rate R is optimized but the channel load is not. For this reason, we shall refer to those systems as "G-non-optimized". The spectral efficiency curves we shall obtain are then to be intended as lower bound to the performance of systems optimized with respect to G.

Furthermore, a closed form analysis of the SIC-SUD and JMUD is possible only in the limit for $K \to \infty$. As we shall see, having imposed $G = K \to \infty$ turns to be heavily suboptimal for SIC-SUD.

We conclude the section, with the analysis of a system where users transmit with probability $p_t = 1$ and adopt an INR protocol. Here, we assume that the decoder does not attempt to decode the largest possible subset of users, as in the case of SIC-SUD and JMUD, instead if decoding of all the active users was unsuccessful, then all the users retransmit new redundancy. Therefore, this INR-MUD scheme is a lower-bound to any "clever" joint decoding INR strategy on top of random user activity. As we shall prove, this INR-MUD scheme achieves the ergodic rate-sum of the underlaying multiple-access channel, hence it is throughput-wise optimal.

As for the SUD-based systems, we assume that users have the same SNR γ and transmit at the same rate R. The receiver has perfect CSI while the transmitters have no CSI.

3.4.1 Results for G-non-optimized ALO with SIC-SUD

First we develop the analysis of ALO with SIC-SUD for a given number of users $K \geq 1$ and a given probability of accessing the channel $p_t \in [0,1]$. Then, following [87], we show that closed form throughput expressions can be found for $p_t = 1$ and $K \to \infty$.

With SIC-SUD, users are sorted in decreasing received SNR order and the strongest is decoded first, by considering the others as noise. Its decoded message is re-modulated and subtracted from the overall received signal. Then, the second strongest users is decoded, re-modulated and subtracted, and so on until it is possible to decode users reliably. Let J be the number of active users on a given slot. J is a binomial distributed random variable with probability mass function (pmf) $\Pr[J=j] = \binom{K}{j} p_t^j (1-p_t)^{K-j}$ for $j = 0, \cdots, K$.

Let $(\tau_1, \tau_2, \cdots, \tau_J)$ the permuted version of the fading power vector $(\alpha_1, \alpha_2, \cdots, \alpha_J)$ such that $\tau_1 \geq \tau_2 \geq \cdots \geq \tau_J$. The cumulative distribution function (cdf) of the k-th ranked fading power τ_k is given by [87]

$$F_{\tau_k}(x) = \int_0^x \frac{K!}{(K-k)!(k-1)!} \left(1 - F_\alpha(u)\right)^{k-1} F_\alpha(u)^{K-k} \mathrm{d}F_\alpha(u) \quad (3.41)$$

where $F_\alpha(x)$ is the cdf of the fading power α. In the limit for large K, the cdf (3.41) becomes

$$F_{\tau_k}(x) \overset{K \gg 1}{\longrightarrow} 1\left\{1 - F_\alpha(u) \leq \frac{k-1}{K-1}\right\} \quad (3.42)$$

where $1\{A\}$ is the indicator function of the event A. Relation (3.42) means that, for $K \to \infty$, the fading powers of each user are a still i.i.d. random variable but the τ's are deterministic. This phenomenon is referred to as *hardening effect* in [87]. From (3.42), in the limit for large K, we have

$$\tau_k \overset{K \gg 1}{\longrightarrow} \mathrm{E}[\tau_k] = F_\alpha^{-1}\left(1 - \frac{k-1}{K-1}\right) \quad (3.43)$$

and

$$\frac{1}{K}\sum_{k=\ell}^{n} \tau_k \overset{K \gg 1}{\longrightarrow} \int_{\ell/K}^{n/K} F_\alpha^{-1}(1-u)\,\mathrm{d}u \triangleq \beta\left(\frac{\ell}{K}, \frac{n}{K}\right) \quad (3.44)$$

for all $k, \ell \leq K$.

Given J active users, define for all $m = 1, \cdots, J$ the event

$$\mathcal{B}_{m,J}(R) = \Bigg\{ \quad \log\left(1 + \frac{\gamma\tau_j}{1 + \sum_{\ell=j+1}^{J} \gamma\tau_\ell}\right) > R \quad \forall j = 1, \cdots, m\,,$$
$$\log\left(1 + \frac{\gamma\tau_j}{1 + \sum_{\ell=j+1}^{J} \gamma\tau_\ell}\right) \leq R \quad \forall j = m+1, \cdots, J \Bigg\}$$

$\mathcal{B}_{m,J}(R)$ is the event the SIC-SUD receiver can decode at most m users each transmitting at rate R, out of the J that were active. The throughput of ALO with SIC-SUD is

$$\eta = \sum_{j=1}^{K} \Pr[J=j] \sum_{m=1}^{j} m\, R\, \Pr[\mathcal{B}_{m,j}(R)] \quad (3.45)$$

For any finite K and p_t, the probabilities $\Pr[\mathcal{B}_{m,j}(R)]$ are not easy to compute. On the contrary, for $p_t = 1$ and infinite population of users, i.e., $G = K \to \infty$, the fraction of users that can be decoded at rate R becomes deterministic, what is random is the identity of the decoded users. Let $m_0 \in \{1, \cdots, K\}$ be the number of user that can be reliably decode, then we can write

$$\eta \overset{K \gg 1}{\longrightarrow} m_0 \inf_{0<j\leq m_0} \log \left(1 + \frac{\frac{1}{K}\tau_j}{\frac{1}{K\gamma} + \frac{1}{K}\sum_{\ell=m_0+1}^{K} \tau_\ell} \right) \tag{3.46}$$

where the logarithm to be minimized in (3.46) represents the most stringent rate constraint of the event $\mathcal{B}_{m_0,K}(R)$. Since the fraction inside the logarithm in (3.46) vanishes as K increases, throughput (3.46) becomes

$$\eta \overset{K \gg 1}{\longrightarrow} \inf_{0<j\leq m_0} \frac{\frac{m_0}{K}\tau_j}{\frac{1}{K\gamma} + \frac{1}{K}\sum_{\ell=m_0+1}^{K} \tau_\ell} \tag{3.47}$$

Therefore, by recalling that $(E_n/N_0)(\eta/K) = \gamma$, by defining $y = m_0/K$ (the fraction of decoded users) and $x = j/K$, and by using (3.43) and (3.44), in the limit for large K we have

$$\eta = y \inf_{x\in]0,y]} \frac{F_\alpha^{-1}(1-x)}{\left(\frac{E_n}{N_0}\right)^{-1}\eta^{-1} + \beta(x,1)} \tag{3.48}$$

Notice that y is the probability that a randomly chosen user is in the set of users that can be decoded reliably.

Fig. 3.6 shows the throughput η (optimized with respect to $y \in [0,1]$) versus E_n/N_0 for the system with joint decoding, both SIC-SUD and JMUD.

Notice that, without fading, SIC-SUD is equivalent to CDMA with SUMF, i.e.,

$$\eta = 1 - \left(\frac{E_n}{N_0}\right)^{-1} \tag{3.49}$$

because the condition for successful decoding for the first user coincides with that of CDMA with SUMF and, since all the users are received with identical power, either all or none of them can be decoded.

For increasing E_n/N_0, since η in non-decreasing in E_n/N_0, either η converges to a finite value (in case of throughput-wise limited system) or it grows unbounded. In both cases $(E_n/N_0)\eta \to \infty$. Hence, in the limit for

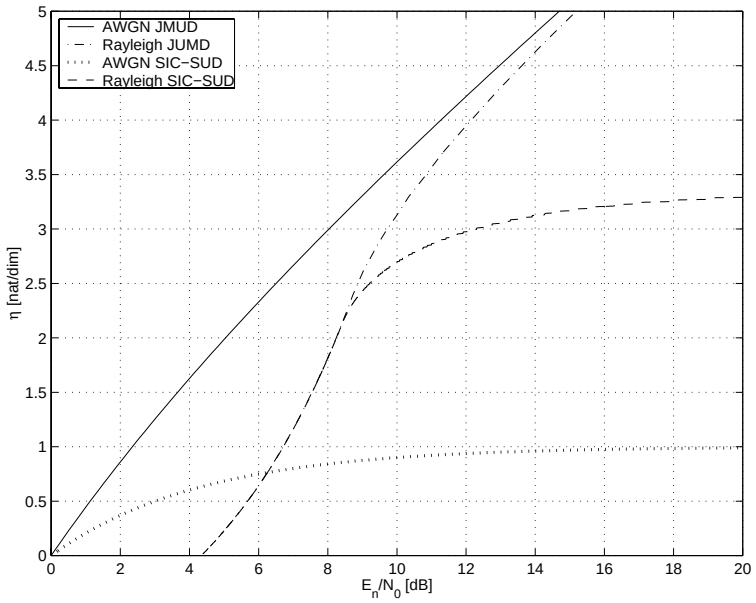

Figure 3.6: Throughput versus E_n/N_0 for a system with joint decoding.

$E_n/N_0 \to \infty$ the throughput (3.46) optimized with respect to y can be written as

$$\eta = \max_{y\in[0,1]} y \inf_{x\in]0,y]} \frac{F_\alpha^{-1}(1-x)}{\beta(x,1)} \tag{3.50}$$

that for the Rayleigh fading case gives

$$\eta = \sup_{u\geq 0} \mathrm{e}^{-u} \inf_{\theta\geq u} \frac{\theta}{1-\mathrm{e}^{\theta}(1+\theta)} = 3.3509 \tag{3.51}$$

From (3.51) we are tempted to conclude that SIC-SUD is throughput-wise limited. Actually, this is due to the fact the the channel load G is not optimized but kept fix to $G = K \to \infty$. In fact, a SIC-SUD system cannot perform worth than unspread ALO (which is not throughput-wise limited) since the condition for decoding the strongest user is the same in both systems but the all the other users in SIC-SUD are decoded at higher SINR than in the unspread system, we conclude that $G = K \to \infty$ is not optimal in all range of E_n/N_0. In particular, it is strongly suboptimal for high E_n/N_0. It is not clear how to define a SIC-SUD receiver in presence of packet combining and random user activity, hence we cannot optimize ALO with SIC-SUD with respect to G.

3.4.2 Results for a G-non-optimized ALO with JMUD

As for the SIC-SUD system, we first develop the analysis of ALO with JMUD for a given number of users $K \geq 1$ and a given probability of accessing the channel $p_t \in [0,1]$. Then, following [87], we show that closed form throughput expressions can be found for $p_t = 1$ and $K \to \infty$.

With JMUD, users are again sorted in decreasing received SNR order. Let J be the number of active users on a given slot. The receiver attempts to joint decode all J users. If the equal-rate point falls inside the capacity region for the J users, then decoding is successful. If not, the weakest user is treated as noise and the receiver considers the capacity region for the remaining $J-1$ users. If the equal-rate point is not inside it, the two weakest users are treated as noise, and so on until it is possible to decode a subset of the J users. For all $m = 1, \cdots, J$ define the event

$$\begin{aligned} \mathcal{C}_{m,J}(R) &= \left\{ \forall S \subseteq \{1,2,\cdots,m\} \quad \sum_{i \in S} R < \log\left(1 + \frac{\sum_{i\in S}\tau_i\gamma}{1+\sum_{\ell=m+1}^{J}\tau_\ell\gamma}\right)\right\} \\ &= \left\{ \forall \ell \in \{1,2,\cdots,m\} \quad R < \frac{1}{\ell}\log\left(1 + \frac{\sum_{i=m-\ell+1}^{m}\tau_i\gamma}{1+\sum_{i=m+1}^{J}\tau_i\gamma}\right)\right\} \end{aligned}$$

$\mathcal{C}_{m,J}(R)$ is the event the JMUD receiver decodes m users, out of the J that were active, each transmitting at rate R, by treating the other $J-m$ user signals as noise. The receiver can successfully decode all the users if the event $\mathcal{C}_{J,J}(R)$. If not (the event $\overline{\mathcal{C}_{J,J}(R)}$ is true), it can decode $J-1$ users if $\mathcal{C}_{J-1,J}(R)$ is true. If not (both the evnts $\overline{\mathcal{C}_{J,J}(R)}$ and $\overline{\mathcal{C}_{J-1,J}(R)}$ are true), it can decode $J-2$ users if $\mathcal{C}_{J-2,J}(R)$ is true, and so on and so forth. Then, the throughput of ALO with JMUD is

$$\eta = \sum_{j=1}^{K} \Pr[J=j] \sum_{m=1}^{j} m\, R\, \Pr[\overline{\mathcal{C}_{j,j}(R)}, \cdots, \overline{\mathcal{C}_{m+1,j}(R)}, \mathcal{C}_{m,j}(R)] \quad (3.52)$$

In general the throughput can be upper bounded as

$$\eta \geq \sum_{j=1}^{K} \Pr[J=j]\, j\, R\, \Pr[\mathcal{C}_{j,j}(R)] \quad (3.53)$$

achieved by a decoder that decodes either all the active users on none. We shall use a similar "lower-bound" in the analysis of INR with JMUD.

Again, for any finite K and p_t, the joint probability of the events $\mathcal{C}_{m,j}(R)$ are not easy to compute. On the contrary, for $p_t = 1$ and infinite population of users, i.e., $G = K \to \infty$, thanks to the hardening of the ordered fading power vector [87], the number of user that can be decoded reliably at rate R becomes a deterministic constant. Let $m_0 \in \{1, \cdots, K\}$ be the number

of user that can be reliably decoded, then we can write

$$\eta \overset{K\gg 1}{\longrightarrow} m_0 \inf_{0<\ell\le m_0} \frac{1}{\ell}\log\left(1+\frac{\frac{1}{K}\sum_{i=m_0-\ell+1}^{m_0}\tau_i}{\frac{1}{K\gamma}+\frac{1}{K}\sum_{i=m_0+1}^{K}\tau_i}\right) \tag{3.54}$$

By recalling that $(E_n/N_0)(\eta/K)=\gamma$, by using (3.43) and (3.44), and by defining $x=\ell/m_0$ and $y=m_0/K$, in the limit for large K we can write (3.54) as

$$\eta = \inf_{x\in]0,1]} \frac{1}{x}\log\left(1+\frac{\beta(y(1-x),y)}{\left(\frac{E_n}{N_0}\right)^{-1}\eta^{-1}+\beta(y,1)}\right) \tag{3.55}$$

Intuitively the logarithm to be minimized in (3.54) represents the most stringent constraint, in terms of aggregated rate, for which the equal rate point falls inside the capacity region defined by the set of the best "Ky" users, when treating the rest as noise.

Fig. 3.6 shows the throughput η (optimized with respect to $y\in[0,1]$) versus E_n/N_0 for the system with joint decoding, both SIC-SUD and JMUD.

Without fading JMUD is equivalent to the unfaded single user system. To show this, we consider first the case of finite $G=K$ and then we take the limit for $G=K\to\infty$. For a given K let $R=\frac{1}{K}\log(1+K\gamma)$ be the information rate of the users. Since all the users transmit (and are received) with identical power, the equal rate point is on the dominant face of the capacity region for these K users, hence the throughput is

$$\eta = K\,R = \log(1+K\gamma) = \log\left(1+\frac{E_n}{N_0}\eta\right) \tag{3.56}$$

Being (3.56) valid for all finite K, by continuity, it is valid for $K\to\infty$. Note that, without fading, all the users are always decoded, i.e., optimal y is $y=1$ for all $E_n/N_0\ge(E_n/N_0)_{\min}$. Therefore, the optimal G is indeed $G\to\infty$. Throughput (3.56) is the maximum possible throughput for a multiple-access system without power control.

With Rayleigh fading, the optimization of (3.54) cannot be carried out in closed form and hence we used numerical evaluation. As pointed out in [87], for high E_n/N_0 the throughput (3.54) tends to

$$\eta \to \log\left(1+\frac{\frac{E_n}{N_0}\eta}{1+\delta}\right) \tag{3.57}$$

for some $\delta>0$ and outage probability vanishes as $\sqrt{\delta/(\eta\,E_n/N_0)}$. This means that at high E_n/N_0 ALO-JMUD approaches unfaded single user performance, i.e., asymptotically there is no loss in performances with respect

to the optimal system. An open question is whether $G \to \infty$ is indeed optimal for every finite E_n/N_0. It is interesting to note that, in Rayleigh fading channel, also with JMUD $(E_n/N_0)_{\min} = \mathrm{e} = 4.3429$ dB as it is for SUD-based systems.

3.4.3 Results for INR with JMUD

As a third, and last, case of MUD-based system, we analyze an INR with JMUD where the decoder either decodes all the users or asks to all the users to retransmit new redundancy. Since in principle this might not be the optimal INR strategy with INR, our result is a lower bound to "clever" INR systems with MUD. We shall prove that this scheme actually achieves the ergodic rate-sum of the underlying block-fading channel, thus proving its optimality.

In a system characterized by user random activity, a given user is active on a sequence of slots that, in general, are not consecutive in time and have different set of active users. In this scenario, it is not clear how to carry out joint decoding across the slots. In the two cases discussed above, joint decoding across the slots is not needed since an ALO protocol was consider. In general, for any protocol and any receiver structure the throughput cannot exceed the ergodic rate-sum of the underlaying MAC block-fading channel. For a system symmetric with respect to all the K users, the ergodic throughput is given by

$$\begin{aligned} \eta^{(\text{ergodic})} &= \mathrm{E}\left[\log\left(1+\sum_{j=1}^{K}\gamma\alpha_j\right)\right] \\ &\overset{(a)}{=} \mathrm{E}\left[\log\left(1+\gamma K\alpha_{\text{eq}}\right)\right]\Big|_{\alpha_{\text{eq}}=\frac{1}{K}\sum_{j=1}^{K}\alpha_j} \\ &\overset{(b)}{\leq} \log\left(1+\gamma K\,\mathrm{E}[\alpha]\right) \end{aligned} \tag{3.58}$$

where (a) shows that the K-user ergodic throughput is the single-user INR throughput with equivalent fading given by the arithmetic mean of the fading powers of the different users; and (b) follows from Jensen inequality and shows that the single-user AWGN capacity is an upperbound to all system without power control at the transmitters. Since $\eta^{(\text{ergodic})}$ is increasing in K, the single-user unfaded performance is achieved by letting $K \to \infty$. The convergence to the RHS of (3.58) follows from the central limit theorem.

Any system with user random activity has throughput that satisfies $\eta \leq \eta^{(\text{ergodic})}$, hence also INR with JMUD (assuming we were able to define what INR with JMUD is!). A lower bound to this INR with JMUD can be obtained by imposing the (possibly) sub-optimal decoding scheme "all or none", i.e. either the decoder is able to decode all the users or asks to all

the users to transmit a new "chunk" of their codeword. We can write

$$\frac{K\,R}{1+\sum_{m=1}^{\infty} p(m)} \leq \eta \leq \mathrm{E}\left[\log\left(1+\sum_{k=1}^{K}\gamma\alpha_k\right)\right] \tag{3.59}$$

where $p(m)$ is the probability that successful joint decoding does not occur after m received blocks. From the definition of our decoding strategy we have that, for every $m \geq 1$, $p(m)$ is given by

$$p(m) = 1 - \Pr\left[\bigcap_{S\subseteq\{1,2,\cdots,K\}}\left\{|S|R \leq \sum_{i=1}^{m}\log\left(1+\sum_{k\in S}\gamma\alpha_{k,i}\right)\right\}\right] \tag{3.60}$$

Since for any given set A and B we have: (a) $\Pr[\overline{A\cap B}] = \Pr[\overline{A}\cup\overline{B}]$ and (b) $\Pr[\overline{A}\cup\overline{B}] = \Pr[\overline{B}] + \Pr[\overline{A}\cap B]$, we have that (3.60) is equivalent to

$$\begin{aligned} p(m) &= \Pr\left[\bigcup_{S\subseteq\{1,2,\cdots,K\}}\left\{|S|R > \sum_{i=1}^{m}\log\left(1+\sum_{k\in S}\gamma\alpha_{k,i}\right)\right\}\right] \\ &= \Pr\left[KR > \sum_{i=1}^{m}\log\left(1+\sum_{k=1}^{K}\gamma\alpha_{k,i}\right)\right] \\ &+ \Pr\left[\left\{\bigcup_{S\subset\{1,2,\cdots,K\}}|S|R > \sum_{i=1}^{m}\log\left(1+\sum_{k\in S}\gamma\alpha_{k,i}\right)\right\}\bigcap\right. \\ &\quad \left.\bigcap\left\{KR \leq \sum_{i=1}^{m}\log\left(1+\sum_{k=1}^{K}\gamma\alpha_{k,i}\right)\right\}\right] \end{aligned} \tag{3.61}$$

By using $\Pr[A\cup B] \leq \Pr[A] + \Pr[B]$ and the assumption of i.i.d. fading for all the users, i.e., the probabilities in (3.61) depend only on the cardinality of S and not on S itself, it follows

$$\begin{aligned} p(m) &\leq \Pr\left[KR > \sum_{i=1}^{m}\log\left(1+\sum_{k=1}^{K}\gamma\alpha_{k,i}\right)\right] \\ &+ \sum_{s=1}^{K-1}\binom{K}{s}\Pr\left[\left\{sR > \sum_{i=1}^{m}\log\left(1+\sum_{k=1}^{s}\gamma\alpha_{k,i}\right)\right\}\right. \\ &\quad \left.\bigcap\left\{KR \leq \sum_{i=1}^{m}\log\left(1+\sum_{k=1}^{K}\gamma\alpha_{k,i}\right)\right\}\right] \end{aligned} \tag{3.62}$$

Denote with $\mu(K) \stackrel{\Delta}{=} \mathrm{E}\left[\log\left(1+\sum_{k=1}^{K}\gamma\alpha_k\right)\right]$ the ergodic throughput of a fading multiple-access channel with K symmetric users. It easy to see that

$\mu(K)$ is increasing in K while $\mu(K)/K$ is decreasing in K. At this point, we can bound the probabilities in (3.62) as follows

$$\Pr\left[\frac{1}{m}\sum_{i=1}^{m}\log\left(1+\sum_{k=1}^{K}\gamma\alpha_{k,i}\right)<\frac{KR}{m}\right]\leq\begin{cases}\mathrm{e}^{-m\phi_1} & \frac{R}{m}<\frac{\mu(K)}{K}\\ 1 & \frac{R}{m}\geq\frac{\mu(K)}{K}\end{cases} \tag{3.63}$$

and

$$\begin{aligned}&\Pr\left[\left\{\sum_{i=1}^{m}\log\left(1+\sum_{k=1}^{s}\gamma\alpha_{k,i}\right)<sR\right\}\bigcap\left\{\sum_{i=1}^{m}\log\left(1+\sum_{k=1}^{K}\gamma\alpha_{k,i}\right)\geq KR\right\}\right]\\ &\leq\begin{cases}\Pr\left[\frac{1}{m}\sum_{i=1}^{m}\log\left(1+\sum_{k=1}^{s}\gamma\alpha_{k,i}\right)<\frac{sR}{m}\right] & \frac{R}{m}<\frac{\mu(s)}{s}\\ \Pr\left[\frac{1}{m}\sum_{i=1}^{m}\log\left(1+\sum_{k=1}^{K}\gamma\alpha_{k,i}\right)\geq\frac{KR}{m}\right] & \frac{R}{m}\geq\frac{\mu(s)}{s}\end{cases}\\ &\leq\begin{cases}\mathrm{e}^{-m\phi_2} & \frac{R}{m}<\frac{\mu(s)}{s}\\ \mathrm{e}^{-m\phi_3} & \frac{R}{m}\geq\frac{\mu(s)}{s}\end{cases}\end{aligned} \tag{3.64}$$

where ϕ_i for $i=1,2,3$ are positive constants guaranteed to exist by the *Large deviation Theorem* [88, Sec.5.11]. With the bounds in (3.64), the sum of the $p(m)$ over $m\geq 1$ is bounded by

$$\sum_{m=1}^{\infty}p(m)\leq\sum_{m\leq\frac{KR}{\mu(K)}}1+\sum_{m>\frac{KR}{\mu(K)}}\mathrm{e}^{-m\phi_1}+\sum_{s=1}^{K-1}\binom{K}{s}\left(\sum_{m\leq\frac{sR}{\mu(s)}}\mathrm{e}^{-m\phi_3}+\sum_{m>\frac{sR}{\mu(s)}}\mathrm{e}^{-m\phi_2}\right)$$

Now, by defining $n_1=\left\lceil\frac{KR}{\mu(K)}\right\rceil+1$ and $n_2(s)=\left\lceil\frac{sR}{\mu(s)}\right\rceil+1$, we get

$$\begin{aligned}\frac{1}{\eta}=\frac{1+\sum_{m=1}^{\infty}p(m)}{K\,R} &\leq \frac{1+\left\lceil\frac{KR}{\mu(K)}\right\rceil+\frac{\mathrm{e}^{-n_1\phi_1}}{1-\mathrm{e}^{-\phi_1}}}{K\,R}\\ &+\frac{\sum_{s=1}^{K-1}\binom{K}{s}\left[\frac{1-\mathrm{e}^{-n_2(s)\phi_2}}{1-\mathrm{e}^{-\phi_2}}+\frac{\mathrm{e}^{-n_2(s)\phi_3}}{1-\mathrm{e}^{-\phi_3}}\right]}{K\,R}\end{aligned} \tag{3.65}$$

and by taking the limit for $R\to\infty$, and by recalling (3.59), we arrive at the desired results

$$\eta=\mu(K)=\mathrm{E}\left[\log\left(1+\sum_{k=1}^{K}\gamma\alpha_k\right)\right] \tag{3.66}$$

This proves that an optimal system "forces" the users to transmit all the time, i.e. $G=K$ and for every K, and decodes either all of them or no one. While for SUD-based systems with infinite population, optimal $G\to\infty$ in $E_n/N_0\in[(E_n/N_0)_{\min},(E_n/N_0)_{\mathrm{th}}]$ and $G\to 1$ in $E_n/N_0\gg 1$, for this scheme $G=K\to\infty$ for all E_n/N_0.

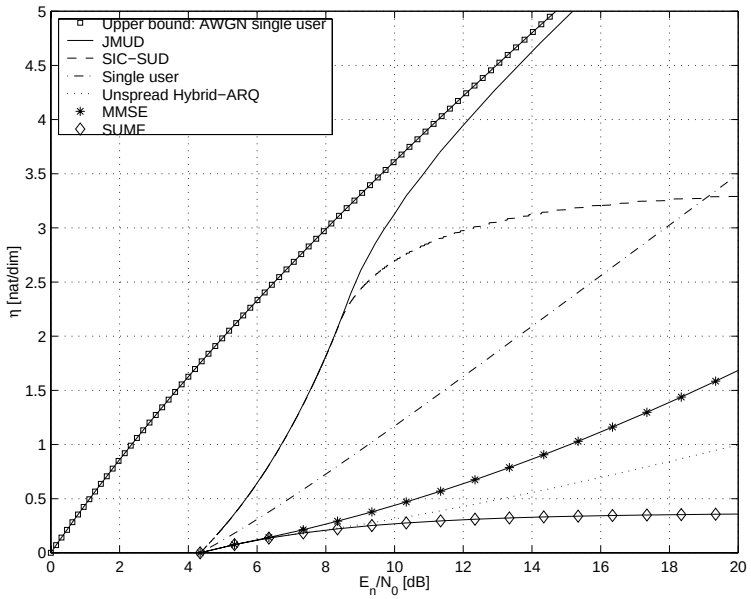

Figure 3.7: Throughput versus E_n/N_0 for ALO with Rayleigh fading.

3.5 Conclusions

To conclude the chapter, we put on the same chart the throughput of all the systems analyzed so far, in order to compare them and draw some conclusions. We claim that our comparison based on equal received energy per bit is fair. However, it is worth reminding that from a practical point of view there are other quantities of great interest, average delay and probability of packet loss for example, that were not considered in this work. Furthermore, comparison is made on total throughput versus E_n/N_0, the choice of other performance measures, like per-user rate or delay, etc, may lead to different conclusions from the one we are going to present here.

In the following we refer to Fig. 3.7,which shows the throughput curves versus E_n/N_0 of all the analyzed systems with ALO protocol in Rayleigh fading (we did not reported the cases without fading to make the picture more readable) and to Fig. 3.8, which shows the throughput curves versus E_n/N_0 of all the analyzed systems with INR without fading (again, the curves referring to the Rayleigh fading case were not added to make the picture more readable).

Minimum E_n/N_0. We have identified two values of $(E_n/N_0)_{\min}$ under which the throughput is zero. One is $(E_n/N_0)_{\min} = \mathrm{e} = 4.3429$ dB for ALO with Rayleigh fading, the other is $(E_n/N_0)_{\min} = 1 = 0$ dB for ALO, RTD, INR without fading and RTD and INR with Rayleigh fading. This shows

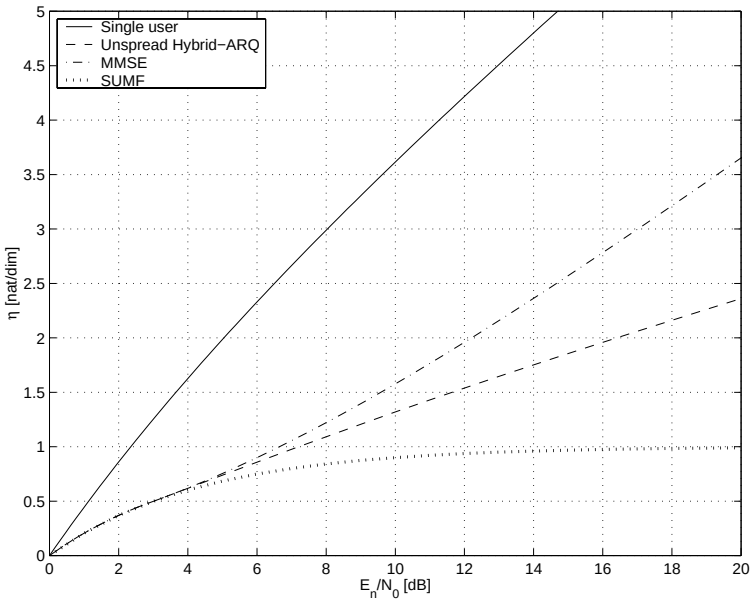

Figure 3.8: Throughput versus E_n/N_0 for INR without fading.

the benefit of packet combining over discarding previous transmission in a faded environment.

Low E_n/N_0 regime. For low E_n/N_0 all the protocols, regardless the fading statistic, have the same throughput which corresponds to vanishing information rate R and infinite channel load G. This means that users must be active all the time, in the unspread case, or be infinitely more than the spreading gain, in the CDMA system. Fading does not impair the system performance, compared to the unfaded case, because the available power is so little that fading fluctuations do not matter. In this regime the simplest system can be implemented without loosing in throughput, in fact CDMA with SUMF does not loose in performance with respect to more complicated schemes as CDMA with MMSE. For the ALO with Rayleigh fading, the curves of Fig. 3.7 show that JMUD is superior to CDMA and unspread system. In this regime SIC-SUD is equivalent to JMUD and again the simpler systems have the same performance of the more complex ones.

High E_n/N_0 regime. The differences in performance among systems and protocols are visible at high E_n/N_0. For all SUD-based systems but CDMA with SUMF, the optimal information rate R grows with E_n/N_0, the channel load $G \to 1$ and the throughput grows linearly with $\log(E_n/N_0)$. This means that when the available power is high, the best strategy consists of encoding

at non vanishing rate and having, on average, only one user per slot, in the unspread system, or one user per chip, in the CDMA system. This does not hold for CDMA with SUMF for which optimal G is always $G \to \infty$ and the throughput is limited to 1 in the case without fading and to e^{-1} in the case with Rayleigh fading. The unspread system behaves always better than CDMA with SUMF but is outperformed by CDMA with MMSE. At $E_n/N_0 = 18$ dB, for example, the difference between unspread system and CDMA with MMSE is about 1 nat/dim.

Without fading JMUD is optimal, in the sense that it has the same performance as unfaded single user case, while with Rayleigh fading it approaches the unfaded single user performance for $E_n/N_0 \to \infty$. SIC-SUD with $G \to \infty$ is highly suboptimal.

Performance in a faded environment. Fading does not improve the throughput performance of the analyzed systems since CSI is not available at the transmitter, i.e., no power control is possible. The single-user unfaded case upperbounds all the systems. This bound is reached by ALO-JMUD without fading, the INR-JMUD technique in Section 3.4.3 and approached by ALO-JMUD with fading in an high E_n/N_0 regime.

Performance-complexity trade-off. The charts in Fig. 3.8 and Fig. 3.7 basically show that performance can be increased by augmenting the complexity of the receiver. First of all, it is clear that SUMF-CDMA is not worth the implementation since it is throughput-wise limited and it is outperformed by the simple unspread system. Among the remaining systems, the unspread system is the simplest of all and its implementation requires the knowledge of the fading gains only; the simplicity is paid by a slower growth of the throughput with E_n/N_0 when compared to the other access techniques. Just above the unspread system, we position MMSE-CDMA: the system is still based on SUD decoding but the improved performance are due to the joint detection step implemented with the MMSE filter. The construction of an MMSE filter requires the knowledge of the spreading sequences, in addition to the fading power levels, and a matrix inversion that, especially for large systems, can be computationally heavy. The best system throughput-wise is obtained by INR protocols and JMUD decoder for load equal to the number of users. The top performance are obtained by mean of a complex decoding process whose practical realization in not clear. In fact, in our setting characterized by outage, joint decoding cannot be implemented with stripping as we showed that SIC-SUD (at least with ALO) is not equivalent to JMUD.

Appendix

3.A Throughput optimization for the single-user like system

In a single-user like case, the SINR can be expressed as

$$\beta = A\alpha \tag{3.67}$$

where α is the (random) fading power and A is a system (deterministic) constant. For an actual single user system, A coincides with the SNR, but in general it depends on other system parameters, as for the case of CDMA with random spreading. The probabilities of decoding failure $p(m)$ at the denominator of (3.1) are

$$p(m) = \begin{cases} F_1\left(\frac{\mathrm{e}^R-1}{A}\right)^m & \text{ALO} \\ F_m\left(\frac{\mathrm{e}^R-1}{A}\right) & \text{RTD} \\ \Pr[\sum_{s=1}^m \log(1+\beta_s) \le R] & \text{INR} \end{cases} \tag{3.68}$$

where we define the cdf

$$F_m(x) = \Pr\left[\sum_{s=1}^m \alpha_s \le x\right] = \begin{cases} 1_{\{x>m\}} & \text{without fading} \\ 1 - \sum_{j=0}^{m-1} \frac{x^j}{j!}\mathrm{e}^{-x} & \text{Rayleigh fading} \end{cases} \tag{3.69}$$

Note that $F_1(x) \triangleq F_\alpha(x)$.

3.A.1 Result for ALO

In this case, the $p(m)$'s form a geometric series, hence

$$\eta = RG\,[1-p(1)] = \left\{ \begin{array}{ll} RG\,1_{\{x\le 1\}} & \text{without fading} \\ RG\,\mathrm{e}^{-x} & \text{Rayleigh fading} \end{array} \right|_{x=(\mathrm{e}^R-1)/A} \tag{3.70}$$

The maximization over R gives

$$\eta = \left\{ \begin{array}{ll} G\log(1+A) & \text{without fading} \\ G\,A\,\mathrm{e}^{-x-\frac{\mathrm{e}^x-1}{x\mathrm{e}^x}} & \text{Rayleigh fading} \end{array} \right|_{x\mathrm{e}^x=A} \tag{3.71}$$

3.A.2 Result for RTD

Without fading the denominator of (3.1) yields

$$1 + \sum_{m\ge 1} p(m) \;=\; \sum_{m\ge 0} 1_{\{m<x\}} = 1 + \lfloor x \rfloor \tag{3.72}$$

while with Rayleigh fading we get

$$1+\sum_{m\geq 1} p(m) = 1+\sum_{m\geq 1}\left[1-\sum_{j=0}^{m-1}\frac{x^j}{j!}\mathrm{e}^{-x}\right] = 1+\sum_{m\geq 1}\sum_{j=m}^{\infty}\frac{x^j}{j!}\mathrm{e}^{-x}$$
$$= 1+\sum_{j=1}^{\infty} j\frac{x^j}{j!}\mathrm{e}^{-x} = 1+x \tag{3.73}$$

By substitution of (3.72) and (3.73) in (3.1) we obtain

$$\eta = \left\{ \begin{array}{ll} \frac{RG}{1+\lfloor x \rfloor} & \text{without fading} \\ \frac{RG}{1+x} & \text{Rayleigh fading} \end{array} \right. \Bigg|_{x=(\mathrm{e}^R-1)/A} \tag{3.74}$$

The maximization over R gives

$$\eta = \left\{ \begin{array}{ll} G\log(1+A) & \text{without fading} \\ G\,A\mathrm{e}^{-x} & \text{Rayleigh fading} \end{array} \right. \Bigg|_{1+(x-1)\mathrm{e}^x=A} \tag{3.75}$$

3.A.3 Result for INR

The optimization over R yields $R\to\infty$ (see [24]) and the throughput is

$$\eta = \mathrm{E}[\log(1+\beta)] \tag{3.76}$$

hence

$$\eta = \left\{ \begin{array}{ll} G\log(1+A) & \text{without fading} \\ G\,\mathrm{e}^{1/A}\mathrm{Ei}(1/A) & \text{Rayleigh fading} \end{array} \right. \tag{3.77}$$

where $\mathrm{Ei}(x) \triangleq \int_x^\infty \mathrm{e}^{-t}/t\,\mathrm{d}t$ is the exponential integral function.

Remark on the optimization over G. For an actual single user system $A=\gamma$ and $G\in[0,1]$, hence the optimization over G is trivial and gives for all the protocols with every fading statistics $G=1$. In general, G can be a function of the other system parameters, hence the maximization in more involved and must be carried out case by case.

3.B The unspread system: ALO with Rayleigh fading

Suppose the following implicit equation is given

$$H(x,y)=0 \tag{3.78}$$

that in a neighborhood of the solution can be put in explicit form as

$$y = f(x) \tag{3.79}$$

The derivative of $f(\cdot)$ can be obtained by solving the following system, which involves the differential of (3.78) and (3.79),

$$\begin{cases} \mathrm{d}x H_x + \mathrm{d}y H_y = 0 \\ \mathrm{d}y = f_x \mathrm{d}x \end{cases} \tag{3.80}$$

where $H_x \triangleq \frac{\partial H}{\partial x}$ and $f_x \triangleq \frac{\mathrm{d}f}{\mathrm{d}x}$, and get

$$f_x = \frac{\mathrm{d}y}{\mathrm{d}x} = -\frac{H_x}{H_y} \tag{3.81}$$

(for more details see [90], Sections 2.10 and 2.11).

In the case at hand, from Appendix 2.E, we have

$$H\left(G, R, \eta, \frac{E_n}{N_0}\right) = -\eta + RG \exp\left(-G\frac{\mathrm{e}^R - 1}{\frac{E_n}{N_0}\eta} - G(1 - \mathrm{e}^{-R})\right) \tag{3.82}$$

and in a neighborhood of the solution

$$\eta = f\left(G, R, \frac{E_n}{N_0}\right) \tag{3.83}$$

Applying (3.81) we obtain that the derivative of (3.83) with respect to G is

$$\frac{\partial f}{\partial G} = -\frac{H_G}{H_\eta} = \frac{\eta}{1 - G\frac{\mathrm{e}^R - 1}{\frac{E_n}{N_0}\eta}} \left[\frac{1}{G} - \frac{\mathrm{e}^R - 1}{\frac{E_n}{N_0}\eta} - (1 - \mathrm{e}^{-R})\right] \tag{3.84}$$

By equating (3.84) to zero and solving with respect to G, we get

$$G = \frac{1}{\frac{\mathrm{e}^R - 1}{\frac{E_n}{N_0}\eta} + (1 - \mathrm{e}^{-R})} \tag{3.85}$$

Equation (3.85) can be view as a parametric definition of G. By substitution of (3.85) in (3.82) and writing explicitly η with respect to E_n/N_0, we get

$$\eta = \frac{\mathrm{e}^{-1} R}{1 - \mathrm{e}^{-R}} - \frac{\mathrm{e}^R}{\frac{E_n}{N_0}} \tag{3.86}$$

Note that (3.86) is positive $\forall R \geq 0$ iff $E_n/N_0 > \mathrm{e}$. The optimization of (3.86) with respect to R gives

$$\frac{E_n}{N_0} = \mathrm{e}\frac{\mathrm{e}^R(1 - \mathrm{e}^{-R})^2}{1 - \mathrm{e}^{-R}(R + 1)} \tag{3.87}$$

The limit of (3.87) for $R \to 0$ is $E_n/N_0 = 2\mathrm{e}$, in fact for $E_n/N_0 \in [\mathrm{e}, 2\mathrm{e}]$ the function (3.86) in monotonic decreasing and has it maximum for $R = 0$. Hence the final expression is

$$\begin{cases} \frac{E_n}{N_0} \in [\mathrm{e}, 2\mathrm{e}] \\ \eta = \mathrm{e}^{-1} - \left(\frac{E_n}{N_0}\right)^{-1} \end{cases}$$
$$\begin{cases} R \geq 0 \\ \frac{E_n}{N_0} = \mathrm{e}\dfrac{\mathrm{e}^R(1-\mathrm{e}^{-R})^2}{1-\mathrm{e}^{-R}(R+1)} \\ \eta = \frac{\mathrm{e}^{-1}R}{1-\mathrm{e}^{-R}} - \frac{\mathrm{e}^R}{\frac{E_n}{N_0}} \end{cases} \tag{3.88}$$

Note that $R \to 0$ in (3.85) means $G \to \infty$.

3.C A numerical technique for throughput optimization

In general, after the maximization of η with respect to R, we have

$$\begin{cases} \eta & = f(G,\gamma) \\ \dfrac{E_n}{N_0} & = \dfrac{G\gamma}{f(G,\gamma)} \end{cases} \tag{3.89}$$

where $f(G,\gamma)$ is a function that depends on the protocol and on the cdf of the fading. Equations in (3.89) define η as a function of E_n/N_0 in a parametric form that depends on two parameters: the channel load G and the average SNR γ. In principle, we could let G and γ vary in all R_+^2, for every pair (G,γ) plot the corresponding point $(\eta, E_n/N_0)$ in a Cartesian plane and then take the upper-envelope of the obtained set of points. Numerically this in not a well defined problem, unless we are reasonably sure of taking enough "good" pairs (G,γ) that are on the upper-envelope of our set of points.

The procedure just discussed is equivalent to fixing a value of E_n/N_0, searching for the pairs (G,γ) solving $\frac{G\gamma}{f(G,\gamma)} = E_n/N_0$ and among all these pairs take the point yielding the maximum η. In formulas, we define

$$\mathcal{A}_x = \left\{(G,\gamma) \in R_+^2 : \frac{G\gamma}{f(G,\gamma)} = x\right\} \tag{3.90}$$

then

$$\eta = \frac{1}{x} \sup_{(G,\gamma)\in\mathcal{A}_x} G\gamma \tag{3.91}$$

Assuming we are able to compute $\mathcal{A}_x$ for every x, the formulation (3.91) is much more appealing.

Let apply (3.91) to the case of ALO without fading. The the throughput formula is

$$\eta = \log\left(1+\frac{\gamma}{1+K\gamma}\right)\sum_{k=0}^{K} G\mathrm{e}^{-G}\frac{G^k}{k!} \tag{3.92}$$

(see (3.24)), i.e., every user encodes its messages with information rate $R = \log(1+\gamma/(1+K\gamma))$, that allows for successful decoding with all sets of less than K simultaneous interfering users. The optimization of (3.92) over K, i.e. choosing the best information rate for a given pair (G,γ), cannot be carried out in closed form, but assuming known the best value of K we can apply the method in (3.91) and write

$$\eta = \frac{1}{x}\sup_{(G,\gamma)\in\mathcal{A}_x^{(K)}} G\gamma \tag{3.93}$$

$$\mathcal{A}_x^{(K)} = \left\{(G,\gamma)\in R_+^2 : \sum_{k=0}^{K}\mathrm{e}^{-G}\frac{G^k}{k!}\frac{1}{\gamma}\log\left(1+\frac{\gamma}{1+K\gamma}\right)=\frac{1}{x}\right\} \tag{3.94}$$

We can solve (3.94) for $K \geq 0$ and plot curves of η indexed by K: for each value of E_n/N_0 we choose the point of the curve with attains maximum η and that gives us also the optimum K. Fig. 3.9 shows the curves obtained for different values of K. It turns out that the best value of K is either zero or infinity. In the regime of low E_n/N_0 the optimum K tends to $K\to\infty$ and the optimum G tends to $G\to\infty$ as $\gamma\to 0$. This means that users must encode their messages at vanishing rate and transmit all the time. The resulting optimized throughput in this region is $\eta = 1-(E_n/N_0)^{-1}$. For large E_n/N_0, the throughput is maximum for $K=0$ and $G\to 1$ as $\gamma\to\infty$, This means that users must encode their messages with non-vanishing rate, which allows correct decoding only if there are no collisions, and transmit very rarely so that on average there is one active user per slot. The resulting parametric throughput expression in this region is $\eta = G\mathrm{e}^{-G}\log(1+\gamma)$. Note that $\eta = G\mathrm{e}^{-G}\log(1+\gamma)$ is the standard slotted-Aloha throughput.

3.D The random spread CDMA system

In general, in the CDMA with random spreading, the throughput can be written as

$$\eta = \frac{g(A)-\left(\frac{E_n}{N_0}\right)^{-1}}{f(A)} \tag{3.95}$$

where $f(A)$ is given in (3.32) and $g(A)$ in (3.35) for the case without fading, in (3.36), in (3.37) and in (3.38) for ALO, RTD and INR, respectively, for the case with Rayleigh fading

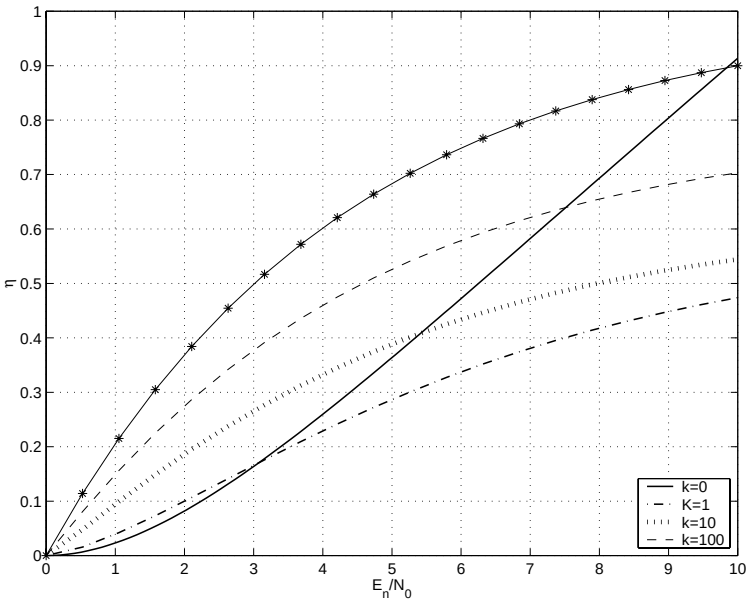

Figure 3.9: η as a function of E_n/N_0 for different values of K

For all the case above, but RTD with Rayleigh fading, the optimization with respect to A is quite simple since η is either a decreasing function of A or it has just one global maximum. In the former case, the optimal value is $A = 0$, in the latter is some positive A. We can find the range of E_n/N_0 for which the optimum is achieved by $A = 0$ as the solution of

$$\left.\frac{\partial \eta}{\partial A}\right|_{A=0} \leq 0 \tag{3.96}$$

In the case without fading and INR with Rayleigh fading, this gives $(E_n/N_0) \leq (E_n/N_0)_{\text{th}} = 2$. In the case of ALO with Rayleigh fading, this gives $(E_n/N_0)_{\text{th}} = 2\text{e}$. In the case of RTD with Rayleigh fading, η is either decreasing function of A or it has a maximum and a minimum. In the former case optimal A is again $A = 0$, in the latter case the optimal A is $A > 0$ if and only if the local maximum of η is actually the absolute maximum. In this case, numerically we found $(E_n/N_0)_{\text{th}} = 13.18$ dB.

In all cases, in the range $(E_n/N_0) > (E_n/N_0)_{\text{th}}$, the optimization over A gives the following parametric equations

$$\begin{cases} \eta & = \left(\dfrac{\mathrm{d}g(A)}{\mathrm{d}A}\right) / \left(\dfrac{\mathrm{d}f(A)}{\mathrm{d}A}\right) \\ \left(\dfrac{E_n}{N_0}\right)^{-1} & = g(A) - f(A)\eta \end{cases} \tag{3.97}$$

for $A \geq 0$, while in the range $(E_n/N_0)_{\min} \leq (E_n/N_0) < (E_n/N_0)_{\text{th}}$, the throughput is given by $\eta = 1 - (E_n/N_0)^{-1}$.

Part II

On the effect of delay constraints on the wideband performance of coordinated centralized systems

Chapter 4

Causal feedback and delay constraint

We consider a wireless multi-access system where users must deliver a message within a given maximum delay by spending a given finite energy. If the message is not transmitted within the required delay then it becomes useless and the residual energy is wasted. The channel is block-fading, with independent fades for each user and each slot. Users know the fading levels up to the current slot but do not know the future fading levels. The receiver collects the signal on all the slots of the frame and performs joint decoding of all the messages. We characterize the region of long-term average achievable rates and the long-term average capacity region per unit energy by finding the optimal power/rate allocation.

4.1 Introduction

The literature on the capacity of fading channels has followed two distinct approaches to characterize power constraints: A) Power constraint on a per-symbol basis (averaged over the codebook); B) Power constraint on a per-codeword basis (averaged over the length of the codeword and the codebook).

Basic information theory results [1] have shown that the laxer constraint B offers no advantage in unfaded channels or in fading channels where the transmitter does not know the channel. However, when the transmitter has instantaneous knowledge of the channel fading coefficients, constraint B leads to strictly larger capacity than A because it enables the use of "power control" which avoids wasting power at symbols where the channel undergoes

deep fades. Under B, the optimum strategy as shown in [11] is water-filling in time. In this setting, the fading process is assumed to be stationary and ergodic and the codewords are long enough for the fading distribution to be revealed within the span of one codeword. If the fading dynamics are slow, this leads to intolerably long blocklength, and consequently delay. Furthermore, waterfilling power control leads to very large peak-to-average ratio of the transmitted waveform, in the low power (or "wideband") regime.

In the high power regime, constraints A and B, although leading to different optimum transmission strategies, achieve very similar single-user capacity. Only in conjunction with multi-access and multiuser detection do optimum power control strategies lead to noticeable advantages in the high power regime [85]. On the other hand, in the low spectral efficiency regime, constraint B enables (for fading distributions with infinite support) reliable communication with energy per bit as small as desired, in stark contrast to constraint A which requires a minimum transmitted energy per bit that is bounded away from zero. Therefore, it is natural to focus on the wideband regime when analyzing the impact of delay constraints on capacity.

Incorporating delay constraints in Shannon theoretic settings is a perennial challenge. In fading channels, it is essential to specify the duration of a codeword with respect to the fading process coherence time and the time interval on which the average input power constraint is enforced. Although vanishing error probability is unattainable unless the number of degrees of freedom grows without bound, that number grows with the product of time duration and bandwidth. Thus, in the wideband regime, an asymptotic analysis is feasible even in a setting of fixed duration codebooks.

In [3] the concept of "delay-limited" capacity region for a multi-access fading channel is introduced. In this setting, each codeword spans a single fading state (i.e., the fading coherence time is much longer than the codeword duration) but the input power constraint is even laxer than B given above: it is imposed over an arbitrarily long sequence of codewords (we shall refer to such constraint as *long-term*). The delay-limited capacity region is the set of rates which can be achieved *for all* fading states (up to a set of measure zero), subject to the long-term input constraint. In other words, the coding rates are fixed while the transmit power fluctuates.

In [2], the concept of "capacity versus outage" is introduced. In this model, each codeword spans a finite number of fading states and consequently the accumulated mutual information at the end of transmission is a random variable. An outage is defined as the event that the mutual information is below the transmission rate. In [91], the authors derived the optimal power allocation policy that minimizes the outage for a given target transmission rate. The optimal policy turns off transmission if, with the current fading realization, the transmission rate cannot be sustained without violating the power constraint. Moreover, in the very low-power regime, it has the characteristic to concentrate the available power on the slot with

highest fading gain.

The derivation of the optimal power control law in [91] is based on the assumption that the fading gains are known prior to transmission. This assumption may not be realistic for transmission taking place over consequent time slots, which the scheme we have considered so far in our thesis. In [92, 93] the authors incorporated in their model the causal knowledge of the channel together with a constraint on the maximum number of slots a codeword is allowed to span. In this scenario, the accumulated mutual information at the end of transmission is a random variable whose value is known only when of whole codeword is received. Both the problem of maximizing the expected rate and of minimization of the outage probability are considered. Notice that the problem is not trivial. The transmitter, in fact, must trade-off between waiting for good channel state and running the risk of violating the delay constraint. Moreover, if the transmitter waits too long, it may be forced to send data on poor channel or, in case a maximum power constraint at each transmission attempt is also incorporated (as in [94]), it may happen that the available power is not used up completely.

In this work we assume a block fading model where a codeword spans a *finite* number of slots N, with fading constant over each slot and varying independently from slot to slot, and the power constraint is enforced on a per-codeword basis (constraint B above). However, we allow *variable rate coding* so that users can coordinate their rates in order to be always inside the fading-dependent capacity region. Here, the transmit power is fixed while the coding rates fluctuate. Consequently, we define the *long-term average* capacity region as the set of all achievable rates averaged over an arbitrarily long sequence of codewords. In this energy-limited setting, we characterize also the "long-term average capacity region per unit energy". Finally, we assume that the fading coefficient affecting each slot, the so-called fading state, is revealed causally to the transmitters [92, 93, 94], i.e., precisely at the beginning of the slot. This is an idealized model of practical schemes that use training symbols. The receiver performs coherent joint decoding after having collected the user signals on all the slots of the frame.

We prove that long-term average capacity is achieved for $N = 1$ by constant power allocation [10], while, as N increases, the optimal causal policy tends to the optimal ergodic policy without delay constraint and non-causal channel state information [13]. Our setting gives the correct trade-off between peak-to-average constrained systems ($N = 1$) and complete freedom in the power allocation ($N \to \infty$) and proves that past and future channel knowledge are immaterial when the delay constraint is not too severe. On the other hand, the optimal policy achieving long-term average capacity per unit energy is "one-shot", i.e., transmission occurs in only one slot of the frame whose selection depends on the fading on the channel. Furthermore, with the "one-shot" policy, transmission occurs at minimum energy per bit needed for reliable communication, which implies not only that the energy

is used in the most efficient way but also that interference to other users is reduced to the minimum.

Our work is mainly inspired by Negi and Cioffi [92, 93] who investigated the optimal causal power control law and its implications on average capacity and outage performance in a single user system. They identified the "one-shot" law as an approximation of the long-term average capacity achieving policy in the low Signal-to-Noise-Ratio (SNR) regime. As a matter of fact, their argument can be made rigorous by using the framework of capacity per unit-cost as introduced by Verdú in [14], which is the approach taken in this work. Furthermore, in our work we do not just state a "capacity formula", we give a coding theorem (achievable and converse part) to prove that the quantity maximized in [92, 93] is the long-term average capacity of the channel. We also give a limiting analysis as the delay constraint N is relaxed and we prove the convergence of our long-term average quantities to the corresponding ergodic capacities. Finally, we quantify the loss of the optimal causal strategy with the optimal strategy with non-causal channel knowledge of the channel [91].

The paper is organized as follows: in Section 4.2 we briefly describe the system model and define our variable rate coding scheme; in Section 4.3 and Section 4.4 we characterize the long-term average capacity region and the long-term average capacity region per unit energy, respectively, for every finite delay N as well as their behavior for infinitely large N; in Section 4.5 we give some numerical examples and in Section 4.6 we conclude with some practical implications of the optimal power allocation policy achieving capacity per unit energy and discuss its application to protocols for wireless sensor networks [95, 96, 97]. All proofs are reported in the Appendices.

Our publications related to this chapter are:
[28] D.Tuninetti and G.Caire. *"The long-term average capacity region per unit energy"*, in the Proceedings of the Thirty-fifth Annual Asilomar Conference on Signals Systems and Computers (Asilomar2001), Pacific Grove (USA), November 2001;
[29] D.Tuninetti and G.Caire. *"The long-term average capacity region per unit energy with application to protocols for wireless sensor networks"*, in Proceedings of the 2002 European Wireless Conference (EW2002), Firenze (Italy), February 2002. Winner of the Best Student paper Award.

4.2 System model and basic definitions

We consider a block-fading Gaussian Multi-Access Channel (MAC) where K transmitters must deliver their message within N slots to the receiver by spending a fixed maximum energy. The number of complex dimensions per slot is $L = \lfloor WT \rfloor$, where T is the slot duration and W is the channel

bandwidth. The baseband complex received vector in slot n is

$$\boldsymbol{y}_n = \sum_{k=1}^{K} c_{k,n}\boldsymbol{x}_{k,n} + \boldsymbol{z}_n \tag{4.1}$$

where $\boldsymbol{z}_n$ is a proper complex Gaussian random vector of dimension L with i.i.d. (independent and identically distributed) components of zero mean and normalized unit variance, $\boldsymbol{x}_{k,n}$ is the complex signal of user k transmitted in slot n, $c_{k,n}$ is the complex fading coefficient for user k with instantaneous power $\alpha_{k,n} \triangleq |c_{k,n}|^2$ with *continuous* cdf (cumulative distribution function) $F_{\alpha}^{(k)}(x)$ i.i.d. for all $n = 1, \cdots, N$ and mutually independent for $k = 1, \cdots, K$.

The receiver has perfect Channel State Information (CSI) while the transmitters have perfect *causal* CSI [92, 93], i.e., in slot n the transmitters know the channel state up to time n, defined by

$$\mathcal{S}_n \triangleq \{c_{k,i} : k = 1, \cdots, K,\ i = 1, \cdots, n\} \tag{4.2}$$

Each transmitter k is subject to the per-codeword input constraint B given above

$$\frac{1}{NL}\sum_{n=1}^{N} |\boldsymbol{x}_{k,n}|^2 \leq \gamma_k \tag{4.3}$$

where γ_k is the transmitted energy per symbol, and because of the noise variance normalization adopted here it has the meaning of *transmit* SNR. In the following we will use the notation

$$\beta_{k,n} \triangleq \frac{1}{L}|\boldsymbol{x}_{k,n}|^2 \tag{4.4}$$

for the *instantaneous* SNR of transmitter k in slot n.

For finite N and L no positive rate is achievable. However, we can consider a sequence of channels indexed by the slot length L and study the achievable rates in the limit for $L \to \infty$ and fixed N. This is a standard mathematical abstraction in the study of the limit performance of block-fading channels [2] and it is motivated by the fact that, in many practical applications, the product WT is large and T is much smaller than the fading coherence time. Even in the limit of large L, the rate K-tuple at which reliable communication is possible over a frame of N slots is a random vector, because only a fixed number N of fading coefficients affect each user codeword. We allow *variable rate coding* so that users can coordinate their rates in order to be always inside the fading-dependent capacity region.

Variable-rate coding in our setting is essentially different from variable-rate coding in an ergodic setting, such as in [13, 11], where actually capacity can be achieved with constant transmission rate and constant energy

per codeword. Here, we assume that each transmitter has an infinite "bit-reservoir" and, depending on the fading instantaneous realization, transmits a variable number of bits per frame. We model this setting by letting the message set size depend on the fading state. Consider user k, let $\mathcal{W}_{k,n} = \{W_{k,n}(\mathcal{S}_n) : \mathcal{S}_n \in \mathbb{C}^{nK}\}$ be a collection of message sets indexed by the channel state $\mathcal{S}_n$ and $|W_{k,n}(\mathcal{S}_n)| = M_{k,n}(\mathcal{S}_n)$ denote the cardinality of the message set $W_{k,n}(\mathcal{S}_n)$.

Definition 1. A variable-rate coding system is defined by:
a) An assignment of message sets to the fading states defined by $\mathcal{W}_{k,n}$ given above;
b) A sequence of encoding functions $\phi_{k,n} : W_{k,n}(\mathcal{S}_n) \times \mathbb{C}^{nK} \to \mathbb{C}^L$ such that $\phi_{k,n} : (w, \mathcal{S}_n) \mapsto \boldsymbol{x}_{k,n}$, where $w \in W_{k,n}(\mathcal{S}_n)$, and such that the resulting codeword satisfies (4.3);
c) A decoding function $\psi : \mathbb{C}^{NL} \times \mathbb{C}^{NK} \to \bigotimes_{k=1}^{K} \bigotimes_{n=1}^{N} \{\cup \mathcal{W}_{k,n}\}$ such that $\psi : (\{\boldsymbol{y}_n : n = 1, \cdots, N\}, \mathcal{S}_N) \mapsto (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_K)$, where $\{\cup \mathcal{W}_{k,n}\}$ is a shorthand notation to indicate the union of all message sets $W_{k,n}(\mathcal{S}_n) \in \mathcal{W}_{k,n}$, and where $\boldsymbol{w}_k = (w_{k,1}, \cdots, w_{k,N})$ is a sequence of messages such that $w_{k,n} \in W_{k,n}(\mathcal{S}_n)$. $\Diamond$

For given $\mathcal{S}_N$, the coding rate for user k of the above scheme is given by

$$R_k(\mathcal{S}_N) = \frac{1}{NL} \sum_{n=1}^{N} \log\left(M_{k,n}(\mathcal{S}_n)\right) \tag{4.5}$$

for all $k = 1, \cdots, K$ and the error probability is given by

$$\begin{aligned} \mathrm{P_e}(\mathcal{S}_N) &= \frac{1}{\prod_{k=1}^{K} \prod_{n=1}^{N} M_{k,n}(\mathcal{S}_n)} \cdot \\ &\sum_{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_K} \Pr\left(\psi(\{\boldsymbol{y}_n\}, \mathcal{S}_N) \neq (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_K) | (\boldsymbol{w}_1, \cdots, \boldsymbol{w}_K)\right) \end{aligned} \tag{4.6}$$

Consider a sequence of frames, where coding and decoding are performed frame-by-frame according to a variable-rate coding scheme defined above, and where the channel state sequence $\mathcal{S}_N$ over each frame is generated according to some ergodic and stationary process. By the law of large numbers, the long-term average coding rate and error probability are given by $R_k = \mathrm{E}[R_k(\mathcal{S}_N)]$ for all $k = 1, \cdots, K$ and by $\mathrm{P_e} = \mathrm{E}[\mathrm{P_e}(\mathcal{S}_N)]$, where expectation is with respect to the joint statistics of the channel state $\mathcal{S}_N$.

A variable-rate coding scheme for frame length N, slot length L, with average rate k-tuple $\boldsymbol{R} = (R_1, \cdots, R_K)$, with power constraint defined by the K-tuple $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_K)$ and average probability of error $\mathrm{P_e} \leq \epsilon$ is said to be a $(N, L, \boldsymbol{R}, \boldsymbol{\gamma}, \epsilon)$-code. The operative definitions of long-term average capacity region and of long-term average capacity region per unit-energy mimic, respectively, the standard capacity region definition for input

constrained channels [1] and definition of capacity region per unit cost given in [14].

Definition 2. A rate K-tuple $\boldsymbol{R}^\star \in \mathbb{R}_+^K$ is long-term average ϵ-achievable if for all $\lambda > 0$ there exist $\overline{L}$ such that for $L \geq \overline{L}$ variable-rate $(N, L, \boldsymbol{R}, \boldsymbol{\gamma}, \epsilon)$-codes can be found with $R_k > R_k^\star - \lambda$ for $k = 1, \cdots, K$. A rate K-tuple is achievable if it is ϵ-achievable for all $0 < \epsilon < 1$. The long-term average capacity region $C_{K,N}(\boldsymbol{\gamma})$ is the closure of the convex hull of all long-term achievable rate K-tuples. $\Diamond$

Definition 3. A K-tuple $\boldsymbol{r}^\star \in \mathbb{R}_+^K$ is a long-term average ϵ-achievable rate per unit energy if for all $\lambda > 0$ there exist an energy vector $\bar{\boldsymbol{\nu}} = (\bar{\nu}_1, \cdots, \bar{\nu}_K)$ such that for $\boldsymbol{\nu} \geq \bar{\boldsymbol{\nu}}$ [1] variable-rate $(N, L, \boldsymbol{R}, \boldsymbol{\nu}/(NL), \epsilon)$-codes can be found with $(LNR_k)/\nu_k > r_k^\star - \lambda$ for $k = 1, \cdots, K$. A rate K-tuple is achievable if it is ϵ-achievable for all $0 < \epsilon < 1$. The long-term average capacity region per unit-energy $U_{K,N}$ is the set of all long-term achievable rate K-tuples per unit-energy. $\Diamond$

In this setting, it is meaningful to study the largest achievable *long-term average rate region*, subject to the short-term power constraint (4.3). Moreover, in the energy-limited case investigated here, a meaningful system design criterion is to look for the largest achievable *long-term average capacity per unit energy* (bit/joule). Next, in analogy with [13, 14], we characterize the long-term average capacity region and the long-term average capacity per unit energy for our system. We also give limiting theorems for large delay N.

4.3 Long-term average capacity region

We have the following result:

Theorem 1. The long-term average capacity region is given by

$$C_{K,N}(\boldsymbol{\gamma}) = \bigcup_{\boldsymbol{\beta} \in \Gamma_{K,N}(\boldsymbol{\gamma})} \left\{ \boldsymbol{R} \in \mathbb{R}_+^K \; : \; \forall \mathcal{A} \subseteq \{1, \cdots, K\} \right. \\ \left. \sum_{k \in \mathcal{A}} R_k \leq \mathrm{E}\left[\frac{1}{N} \sum_{n=1}^{N} \log\left(1 + \sum_{k \in \mathcal{A}} \alpha_{k,n} \beta_{k,n}(\mathcal{S}_n) \right) \right] \right\} \quad (4.7)$$

where expectation is with respect to the channel state $\mathcal{S}_N$ and where $\Gamma_{K,N}(\boldsymbol{\gamma})$ is the set of *feasible* causal short-term power allocation policies $\boldsymbol{\beta} = \{\beta_{k,n} :$

[1] For two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$, the notation $\boldsymbol{a} \geq \boldsymbol{b}$ means that the difference $\boldsymbol{a} - \boldsymbol{b}$ has nonnegative components.

$k = 1, \cdots, K, n = 1, \cdots, N\}$ defined as

$$\Gamma_{K,N}(\boldsymbol{\gamma}) \triangleq \left\{ \boldsymbol{\beta} \in \mathbb{R}_+^{KN} : \frac{1}{N} \sum_{n=1}^{N} \beta_{k,n}(\mathcal{S}_n) \leq \gamma_k \right\} \tag{4.8}$$

where $\beta_{k,n}(\mathcal{S}_n)$ indicates the causality constraint.

Proof. The achievable part easily follows by constructing random codes with entries $\boldsymbol{x}_{k,n} \sim \mathcal{N}(\mathbf{0}, \beta_{k,n}\mathbf{I})$ such that the variances satisfy (4.8) and the rates satisfy all the inequalities in (4.7). The converse part follows by showing that the capacity of the N-slot extension channel is achieved by Gaussian input distributions in the form of those used in the achievable part of the theorem. For details see Appendix 4.A. □

Remark. The region $C_{K,N}(\boldsymbol{\gamma})$ is convex in $\boldsymbol{\gamma}$, in fact, by applying Jensen's inequality it is straightforward to see that if $\boldsymbol{R}^{(a)} \in C_{K,N}(\boldsymbol{\gamma})$ and $\boldsymbol{R}^{(b)} \in C_{K,N}(\boldsymbol{\gamma})$ then, for every $\lambda \in [0,1]$ we have $\lambda \boldsymbol{R}^{(a)} + (1-\lambda)\boldsymbol{R}^{(b)} \in C_{K,N}(\boldsymbol{\gamma})$. For this reason in Theorem 1 the convex-hull operation is not needed.

For a given power policy $\boldsymbol{\beta}$ in $\Gamma_{K,N}(\boldsymbol{\gamma})$, let $\mathcal{P}(\boldsymbol{\beta})$ be the set of long-term average rates achievable by applying $\boldsymbol{\beta}$. Theorem 1 states that the long-term average capacity region $C_{K,N}(\boldsymbol{\gamma})$ is the union of all the polymatroids $\mathcal{P}(\boldsymbol{\beta})$

$$C_{K,N}(\boldsymbol{\gamma}) = \bigcup_{\boldsymbol{\beta} \in \Gamma_{K,N}(\boldsymbol{\gamma})} \mathcal{P}(\boldsymbol{\beta}) \tag{4.9}$$

Such formulation of $C_{K,N}(\boldsymbol{\gamma})$ is not useful unless we can determine its closure set. We explicitly characterize the boundary surface of the $C_{K,N}(\boldsymbol{\gamma})$, following the approach of [13], as the *closure of the convex-hull* all K-tuples $\boldsymbol{R} \in \mathbb{R}_+^K$ that solve

$$\max_{\boldsymbol{R} \in C_{K,N}(\boldsymbol{\gamma})} \sum_{k=1}^{K} \mu_k R_k \tag{4.10}$$

for some $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_K) \in \mathbb{R}_+^K$. As in [13], the optimization in (4.10) can be written as

$$\max_{\boldsymbol{\beta} \in \Gamma_{K,N}(\boldsymbol{\gamma})} \max_{\boldsymbol{R} \in \mathcal{P}(\boldsymbol{\beta})} \sum_{k=1}^{K} \mu_k R_k \tag{4.11}$$

Since $\mathcal{P}(\boldsymbol{\beta})$ is a polymatroid, the solution of the inner maximization in (4.11) is attained by one of the (at most) $K!$ vertices of $\mathcal{P}(\boldsymbol{\beta})$. Such vertex is univocally determined by the entries of the vector $\boldsymbol{\mu}$: it is the one corresponding to the decoding order $\pi_K, \pi_{K-1}, \cdots, \pi_1$ where $\boldsymbol{\pi}$ is the permutation

of $\{1,2,\cdots,K\}$ that orders $\boldsymbol{\mu}$ in decreasing order, i.e., $\mu_{\pi_1} > \cdots > \mu_{\pi_K}$. Hence, for any policy $\boldsymbol{\beta}$ we have

$$\max_{\boldsymbol{R}\in\mathcal{P}(\boldsymbol{\beta})} \sum_{k=1}^{K} \mu_k R_k = \sum_{k=1}^{K} \mu_{\pi_k} \mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N} \log\left(1+\frac{\alpha_{\pi_k,n}\,\beta_{\pi_k,n}}{1+\sum_{j<k}\alpha_{\pi_j,n}\,\beta_{\pi_j,n}}\right)\right] \quad (4.12)$$

where $\boldsymbol{\pi}$ only depends on $\boldsymbol{\mu}$ and not on $\boldsymbol{\beta}$. We have thus turned the maximization (4.10) into the maximization of the right hand side (RHS) of (4.12) over the power policies $\boldsymbol{\beta} \in \Gamma_{K,N}(\boldsymbol{\gamma})$. Due to the causal nature of the channel state information, the maximization of (4.12) with respect to $\boldsymbol{\beta}$, and hence the solution of (4.10), is obtained by Dynamic Programming. We have the following:

Theorem 2. Define for $n = 1,\cdots,N$ the Dynamic Programming recursion

$$\begin{aligned} S_n(P_1,\cdots,P_K;\boldsymbol{\mu}) &= \mathrm{E}\left[\max_{\forall k: p_k\in[0,P_k]} \sum_{k=1}^{K} \mu_{\pi_k} \log\left(1+\frac{\alpha_{\pi_k}\,p_{\pi_k}}{1+\sum_{j<k}\alpha_{\pi_j}\,p_{\pi_j}}\right)\right. \\ &\quad \left. + S_{n-1}(P_1-p_1,\cdots,P_K-p_K;\boldsymbol{\mu})\right] \end{aligned} \quad (4.13)$$

with initial condition $S_0(P_1,\cdots,P_K;\boldsymbol{\mu}) = 0$, where the expectation is with respect to $\boldsymbol{\alpha} = (\alpha_1,\cdots,\alpha_K)$ and where π is the permutation that orders $\boldsymbol{\mu}$ in decreasing order, i.e., $\mu_{\pi_1} > \cdots > \mu_{\pi_K}$. Let $(\widehat{p}_{1,n}(\boldsymbol{\alpha};\boldsymbol{\mu},\boldsymbol{P}),\cdots,\widehat{p}_{K,n}(\boldsymbol{\alpha};\boldsymbol{\mu},\boldsymbol{P}))$ the value of $(p_1,\cdots,p_N)$ that achieves the maximum in (4.13) at step n. Then, the boundary surface of $C_{K,N}(\boldsymbol{\gamma})$ is the closure of

$$\text{convex–hull}\left\{\widehat{\boldsymbol{R}}_N(\boldsymbol{\mu},\boldsymbol{\gamma}) \,:\, \boldsymbol{\mu}\in\mathbb{R}_+^K, \sum_{k=1}^{K}\mu_k = 1\right\} \quad (4.14)$$

where the rates $\widehat{\boldsymbol{R}}_N = [\widehat{R}_{1,N},\cdots,\widehat{R}_{K,N}]$ are given by

$$\widehat{R}_{k,N}(\boldsymbol{\mu},\boldsymbol{\gamma}) = \mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N}\log\left(1+\frac{\alpha_{k,n}\,\widehat{\beta}_{k,n}(\mathcal{S}_n;\boldsymbol{\mu},\boldsymbol{\gamma})}{1+\sum_{j<\pi^{-1}(k)}\alpha_{\pi_j,n}\,\widehat{\beta}_{\pi_j,n}(\mathcal{S}_n;\boldsymbol{\mu},\boldsymbol{\gamma})}\right)\right] \quad (4.15)$$

($\pi^{-1}(k)$ gives the position of index k in the permuted vector $\boldsymbol{\pi}$) and where the optimal power policy $\widehat{\boldsymbol{\beta}} = \left\{\widehat{\beta}_{k,n}(\mathcal{S}_n;\boldsymbol{\mu},\boldsymbol{\gamma})\ ,\ \forall k\ ,\ \forall n\right\}$ is given by

$$\widehat{\beta}_{k,n}(\mathcal{S}_n;\boldsymbol{\mu},\boldsymbol{\gamma}) = \widehat{p}_{k,N-n+1}\left(\boldsymbol{\alpha}_n, N\boldsymbol{\gamma} - \sum_{j=1}^{n-1}\widehat{\boldsymbol{\beta}}_j(\mathcal{S}_j;\boldsymbol{\mu},\boldsymbol{\gamma});\boldsymbol{\mu}\right) \quad (4.16)$$

for all n and k.

Proof. Recursion (4.13) and the optimal power policy (4.16) follow easily from the general theory of Dynamic Programming [98] when the cost function to be maximized is given by the RHS of (4.12) and the system evolves,

from slot n to slot $n+1$, according to $(\boldsymbol{\alpha}_n, \boldsymbol{P}) \to (\boldsymbol{\alpha}_{n+1}, \boldsymbol{P} - \widehat{\boldsymbol{p}}_n)$. Details can be found in Appendix 4.B. □

Note that

$$\sum_{k=1}^{K} \mu_k \, \widehat{R}_{k,N}(\boldsymbol{\mu}, \boldsymbol{\gamma}) = \frac{1}{N} S_N(N\gamma_1, \cdots, N\gamma_K; \boldsymbol{\mu}) \tag{4.17}$$

In [92], the authors computed numerically the recursion (4.13) for $K = 1$ in the Rayleigh fading case.

Remark. In contrast with [13], the convex-hull operation in (4.14) is needed since the rates $\widehat{\boldsymbol{R}}_N(\boldsymbol{\mu}, \boldsymbol{\gamma})$ might not be continuous functions of $\boldsymbol{\mu}$. Consider, as an example, the case of $N = 1$. The region $C_{K,1}(\boldsymbol{\gamma})$ coincides with the ergodic capacity region of a fading channel without CSI at the transmitters, the dominant face of which is an hyper-plane in K dimensions. Due to the polymatroid structure of $C_{K,1}(\boldsymbol{\gamma})$, the solution (4.15) is one of the (at most) $K!$ vertices of the dominant face. Hence, as $\boldsymbol{\mu}$ varies in $\mathbb{R}_+^K$, the set of $\widehat{\boldsymbol{R}}_N(\boldsymbol{\mu}, \boldsymbol{\gamma})$ contains at most $K!$ points. It is clear that the convex hull operation is needed here.

From Theorem 2, by solving recursion (4.13) for $n = 1$, we see that the optimal solution is $\widehat{p}_{k,1} = P_k$ for every $\boldsymbol{\mu}$ and for every $\boldsymbol{\alpha}$. Hence, from (4.16) with $N = 1$, we have that $\widehat{\beta}_{k,1} = \gamma_k$ for all k, i.e., the optimal solution for $N = 1$ is constant power allocation. From (4.16), we also see that $\widehat{\beta}_{k,N} = N\gamma_k - \sum_{j=1}^{N-1} \widehat{\beta}_{k,j}$ which means that, on the last available slot, all the remaining energy is used regardless of the fading value, which is sensible since the remaining energy cannot be used on the next frame.

Due to the heavy notation of Theorem 2, it might not be so straightforward to understand how to construct the optimal policy $\widehat{\boldsymbol{\beta}}$ and the role of the recursion (4.13). Hence, we give some explanation. To characterize the boundary surface of $C_{K,N}(\boldsymbol{\gamma})$, first we need to solve the recursion (4.13) in order to determine the optimal values $\{\widehat{p}_{k,n}(\boldsymbol{\alpha}; \boldsymbol{\mu}, \boldsymbol{P}) \ \forall k = 1 \cdots, K\}$ for $n \geq 1$. Then, for a given delay N, we built up the optimal power policy (4.16) by considering the "length-N window" of optimal values $\{\widehat{p}_{k,n}(\boldsymbol{\alpha}; \boldsymbol{\mu}, \boldsymbol{P}) \ \forall k = 1 \cdots, K\}$ for $n = N, N-1, \cdots, 1$. On the first slot ($n = 1$), the optimal policy $\widehat{\boldsymbol{\beta}}_1 = [\widehat{\beta}_{1,1}, \cdots, \widehat{\beta}_{K,1}]$ is derived from $\{\widehat{p}_{k,N}(\boldsymbol{\alpha}; \boldsymbol{\mu}, \boldsymbol{P}) \ \forall k = 1 \cdots, K\}$ computed for $\boldsymbol{\alpha}$ equal to the actual fading values $\boldsymbol{\alpha}_1 = [\alpha_{1,1}, \cdots, \alpha_{K,1}]$ and for power $\boldsymbol{P}$ equal to the total available energies $N\boldsymbol{\gamma} = [N\gamma_1, \cdots, N\gamma_K]$, i.e., for each user $k = 1, \cdots, K$

$$\widehat{\beta}_{k,1}(\mathcal{S}_1; \boldsymbol{\mu}, \boldsymbol{\gamma}) = \widehat{p}_{k,N}(\boldsymbol{\alpha}_1, N\boldsymbol{\gamma}; \boldsymbol{\mu})$$

On the second slot ($n = 2$), the optimal policy $\widehat{\boldsymbol{\beta}}_2 = [\widehat{\beta}_{1,2}, \cdots, \widehat{\beta}_{K,2}]$ is derived from $\{\widehat{p}_{k,N-1}(\boldsymbol{\alpha}; \boldsymbol{\mu}, \boldsymbol{P}) \ \forall k = 1 \cdots, K\}$. Given the fading realization $\boldsymbol{\alpha}_2 = [\alpha_{1,2}, \cdots, \alpha_{K,2}]$ and the remaining energies $N\boldsymbol{\gamma} - \widehat{\boldsymbol{\beta}}_1(\mathcal{S}_1; \boldsymbol{\mu}, \boldsymbol{\gamma})$, the optimal power allocation is

$$\widehat{\beta}_{k,2}(\mathcal{S}_2; \boldsymbol{\mu}, \boldsymbol{\gamma}) = \widehat{p}_{k,N-1}(\boldsymbol{\alpha}_2, N\boldsymbol{\gamma} - \widehat{\boldsymbol{\beta}}_1(\mathcal{S}_1; \boldsymbol{\mu}, \boldsymbol{\gamma}); \boldsymbol{\mu})$$

On the third slot ($n = 3$), the optimal policy $\widehat{\boldsymbol{\beta}}_3$ is derived from $\{\widehat{p}_{k,N-2}(\boldsymbol{\alpha};\boldsymbol{\mu},\boldsymbol{P}) \;\; \forall \; k \;\; = \;\; 1 \;\cdots,\; K\}$. With fading realization $\boldsymbol{\alpha}_3$ and remaining available energies $N\boldsymbol{\gamma} - \widehat{\boldsymbol{\beta}}_1(\mathcal{S}_1;\boldsymbol{\mu},\boldsymbol{\gamma}) - \widehat{\boldsymbol{\beta}}_2(\mathcal{S}_2;\boldsymbol{\mu},\boldsymbol{\gamma})$, the optimal policy is

$$\widehat{\beta}_{k,3}(\mathcal{S}_2;\boldsymbol{\mu},\boldsymbol{\gamma}) = \widehat{p}_{k,N-2}(\boldsymbol{\alpha}_3, N\boldsymbol{\gamma} - \widehat{\boldsymbol{\beta}}_1(\mathcal{S}_1;\boldsymbol{\mu},\boldsymbol{\gamma}) - \widehat{\boldsymbol{\beta}}_2(\mathcal{S}_2;\boldsymbol{\mu},\boldsymbol{\gamma});\boldsymbol{\mu})$$

for each user $k = 1,\cdots,K$. The procedure continues until $n = N$ With the optimal policy built in such a way, we can compute the rates (4.15).

Notice that, having determined the optimal values $\{\widehat{p}_{k,n}(\boldsymbol{\alpha};\boldsymbol{\mu},\boldsymbol{P})\}$ for $n = 1,\cdots,N$, allows us to compute $C_{K,n}(\boldsymbol{\gamma})$ for all $n = 1,\cdots,N$. In other words, varying the delay from N_1 to $N_2 > N_1$ only requires the computation of other $N_2 - N_1$ steps of the recursion (4.13), those for $n = N_1 + 1,\cdots,N_2$.

Although for finite N a closed form solution of (4.13) seems infeasible, for large N we can prove:

Theorem 3. In the limit for large N, the long-term average capacity region $C_{K,N}(\boldsymbol{\gamma})$ tends to the ergodic capacity region [13]

$$C_K^{(\mathrm{erg})}(\boldsymbol{\gamma}) \;\; = \bigcup_{\boldsymbol{\beta}\in\Gamma_K^{(\mathrm{erg})}(\boldsymbol{\gamma})} \Big\{\boldsymbol{R}\in\mathbb{R}_+^K \;:\; \forall\mathcal{A}\subseteq\{1,\cdots,K\} \sum_{k\in\mathcal{A}} R_k \le \mathrm{E}\left[\log\left(1+\sum_{k\in\mathcal{A}}\alpha_k\beta_k(\boldsymbol{\alpha})\right)\right]\Big\} \tag{4.18}$$

where expectation is with respect to the instantaneous channel state $\boldsymbol{\alpha} = (\alpha_1,\cdots,\alpha_K)$ and $\Gamma_K^{(\mathrm{erg})}(\boldsymbol{\gamma})$ is the set of *feasible* power allocation functions $\boldsymbol{\beta} = \{\beta_k \;:\; k = 1,\cdots,K\}$ defined by

$$\Gamma_K^{(\mathrm{erg})}(\boldsymbol{\gamma}) \stackrel{\Delta}{=} \{\boldsymbol{\beta}\in\mathbb{R}_+^K \;:\; \mathrm{E}[\beta_k(\boldsymbol{\alpha})] \le \gamma_k\} \tag{4.19}$$

Proof. The proof follows three steps: 1) we show that the ergodic capacity of the underlying block-fading channel contains $C_{K,N}(\boldsymbol{\gamma})$ for every $N \ge 1$; 2) for every $N \ge 1$ we construct an inner bound region by choosing the following (possibly sub-optimal) feasible power policy: on every slot allot power according to the optimal ergodic policy, if this does not violate the energy constraint, otherwise waits for the next slot; 3) finally, since in the limit for large N the probability that the above policy in unable to allocate power according to the ergodic law on all the N slots of the frame is vanishing, we show that the inner-bound region coincides with the ergodic capacity, hence proving the statement. See the details in Appendix 4.C. □

Remark. As the delay constraint is relaxed, i.e., N increases, the penalty incurred by the use of a short-term causal power allocation policy with respect to the ergodic power allocation policy decreases. This means

that, when the delay is not too strict, the past information becomes irrelevant and the power policy tends to become memoryless in the sense that the same law is applied on every slot in an "i.i.d. fashion".

Theorem 2 and Theorem 3 show that in out setting the parameter N can be consider as a measure of the peak-to-average power ratio. In fact, for $N = 1$, optimal power policy is constant power allocation (peak-to-average ratio equal to one) while as N increases optimal power policy tends to ergodic power allocation which allots power only on the "most" favorable fading gains. In the low power regime, the ergodic policy concentrates power only on those slots whose fading is close to the maximum possible fading value.

4.4 Long-term average capacity region per unit energy

A byproduct of the proof of Theorem 1 is that the long-term average capacity region coincides with the standard "ergodic" capacity region of the N-slot extension channel, which is frame-wise memoryless (since the power control "correlates" only symbols inside the same frame). The following theorem is an immediate consequence of this fact and of the general theory of capacity per unit cost [14]:

Theorem 4. The long-term average capacity region per unit energy is

$$U_{K,N} = \bigcup_{\boldsymbol{\gamma} \in \mathbb{R}_+^K} \left\{ \boldsymbol{r} \in \mathbb{R}_+^K : (\gamma_1 r_1, \cdots, \gamma_K r_K) \in C_{K,N}(\boldsymbol{\gamma}) \right\} \tag{4.20}$$

Proof. The proof follows immediately from [14, Theorem 5]. □

In analogy with [14], it is easy to show:

Theorem 5. The long-term average capacity region per unit energy is the hyper-rectangle

$$U_{K,N} = \left\{ \boldsymbol{r} \in \mathbb{R}_+^K : r_k \leq s_N^{(k)} \right\} \tag{4.21}$$

where $s_N^{(k)}$, given by

$$s_N^{(k)} = \lim_{\gamma_k \to 0} \frac{1}{\gamma_k} \sup_{\boldsymbol{\beta} \in \Gamma_{1,N}(\gamma_k)} \mathrm{E}\left[\frac{1}{N} \sum_{n=1}^{N} \alpha_{k,n} \beta_{k,n}(\mathcal{S}_n) \right] \tag{4.22}$$

is the k-th user single-user long-term average capacity per unit energy.

Proof. See Appendix 4.D. □

The explicit solution of (4.22) was found in [92]. We report it here in our notation for later use:

Theorem 6. The k-th user single-user long-term average capacity per unit energy $s_N^{(k)}$ is given by the Dynamic Programming recursion

$$s_n^{(k)} = \mathrm{E}[\max\{s_{n-1}^{(k)}, \alpha_k\}] \tag{4.23}$$

for $n = 1, \ldots, N$ with initial condition $s_0^{(k)} = 0$ and where expectation is with respect to $\alpha_k \sim F_\alpha^{(k)}(x)$. Furthermore, $s_N^{(k)}$ is achieved by the "one-shot" power allocation policy defined by

$$\beta_{k,n}^\star = \begin{cases} N\gamma_k & \text{if } n = n_k^\star \\ 0 & \text{otherwise} \end{cases} \tag{4.24}$$

where the random variable $n_k^\star$, function of $(\alpha_{k,1}, \cdots, \alpha_{k,N})$, is defined as

$$n_k^\star = \min\left\{ n \in \{1, \ldots, N\} : \alpha_{k,n} \geq s_{N-n}^{(k)} \right\} \tag{4.25}$$

Proof. Expression (4.23) is the solution of the Dynamic Programming algorithm when the cost function to be maximized is (4.22). For more details, see the proof given in [92]. □

We have nicknamed the optimal policy $\beta^\star$ "one-shot" because the whole available energy $N\gamma_k$ is spent all at once in a single slot. In fact, in each slot $n \in \{1, \cdots, N\}$, the transmitter compares the instantaneous fading gain $\alpha_{k,n}$ with the time varying threshold $s_{N-n}^{(k)}$, if the fading is above the threshold then it transmits on the current slot by using all the available energy otherwise it waits for the next slot. Since the threshold to be used on the last available slot is $s_0^{(k)} = 0$, the available energy is used within the required delay of N slots with probability one. This feature of optimal policy was already found out in [92], but in this work the authors did not realize that what is the restricted context of an "approximation for low SNR" is actually the general solution to long-term average capacity per unit energy for every finite delay N. Fig. 4.1 shows a fading realization over a window of $N = 10$ slots. We can this that in this case transmission takes place in slot $n = 6$.

From a practical implementation point of view, the "one-shot" policy is appealing. It requires virtually no computation, just a comparison of the instantaneous fading amplitude with a threshold. The threshold sequence $\{s_n^{(k)}\}_{n=0}^\infty$ can be pre-computed and stored in memory since it only depends on the fading statistic and not on the instantaneous fading values. Policy (4.24) is "memoryless" in the sense that the only information needed about the past slots of the frame is whether transmission has already took place and it is decentralized, i.e., $n_k^\star$ only depends on $(\alpha_{k,1}, \cdots, \alpha_{k,N})$. Moreover, when varying the delay requirements from N_1 to N_2, the threshold

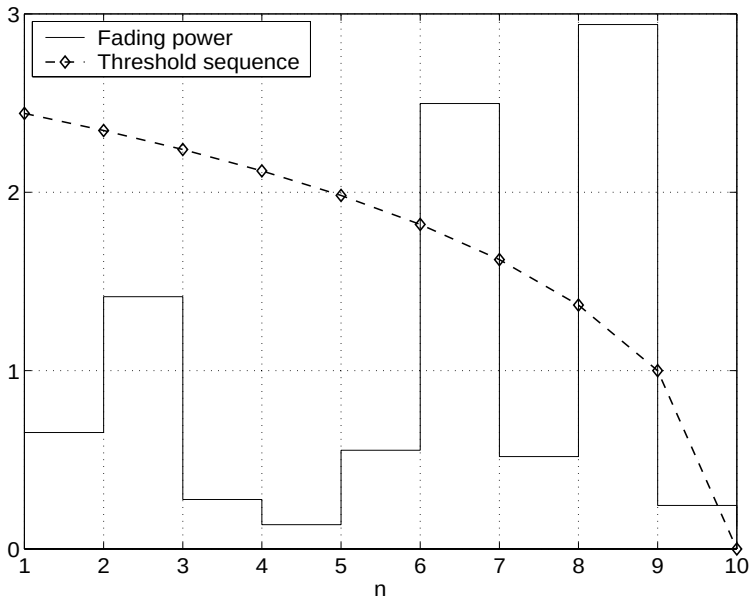

Figure 4.1: Fading realization over a frame of $N = 10$ slots.

sequence need not to be re-computed, only a different "chunk" $\{s_n^{(k)}\}_{n=0}^{N_2-1}$, instead of $\{s_n^{(k)}\}_{n=0}^{N_1-1}$, has to be used. Notice also that the number of active users K does not affect the value of the thresholds.

The behavior of $s_N^{(k)}$ when N grows to infinity is given by the following:
Theorem 7. For large N, the k-th user single-user long-term average capacity per unit energy $s_N^{(k)}$ tends to the k-th user single-user ergodic capacity per unit energy given explicitly by

$$\lim_{N\to\infty} s_N^{(k)} = \sup\{\alpha_k\} \tag{4.26}$$

where $\sup\{\alpha_k\} \stackrel{\Delta}{=} \inf\{x \geq 0 : F_\alpha^{(k)}(x) = 1\}$.
Proof. See Appendix 4.E. □

Remark. Let $C(\gamma)$ be the capacity expressed in nat/dimension as a function of γ, and let $\mathsf{C}(E_b/N_0)$ denote the corresponding spectral efficiency in bit/s/Hz as a function of the energy per bit vs. noise power spectral density, E_b/N_0, given implicitly by

$$\begin{cases} \frac{E_b}{N_0} = \frac{\gamma \log 2}{C(\gamma)} \\ \mathsf{C}\left(\frac{E_b}{N_0}\right) = \frac{C(\gamma)}{\log 2} \end{cases} \tag{4.27}$$

The value $(E_b/N_0)_{\min}$ for which $\mathsf{C}(E_b/N_0) > 0 \Leftrightarrow E_b/N_0 > (E_b/N_0)_{\min}$ is given by [85] $\left(\frac{E_b}{N_0}\right)_{\min} = \frac{\log 2}{\dot{C}(0)}$ where $\dot{C}(0)$ is the first derivative of the

capacity function $C(\gamma)$ at $\gamma = 0$. From the proof of Theorem 5, we see immediately that the reciprocal of $(E_b/N_0)_{\min}$ for the k-th user is its capacity per unit energy (expressed in bit/joule), of the channel, i.e.,

$$\left(\frac{E_b}{N_0}\right)_{\min} = \frac{\log 2}{s_N^{(k)}} \tag{4.28}$$

The "one-shot" policy not only makes the most efficient use of the energy, by maximizing the number of expected correctly received number of bits per joule, but also reduces to the minimum the interference to other users since all the users transmit at minimum E_b/N_0. Notice that as the delay constraint is relaxed, i.e., N grows, the minimum required E_b/N_0 lowers down. Of course nothing is for free: the fact that the system works at the minimum E_b/N_0 is because it uses of a large number of degree of freedom (L) per information bit, i.e., the system works in the so-called "infinite bandwidth regime". As shown recently in [99], information theoretic performance in "wideband regime" is not only characterized by the minimum energy per bit since minimum E_b/N_0 alone is unable to give the correct tradeoff bandwidth-power. Transmitting at minimum E_b/N_0 implies using of an infinitely large bandwidth (infinite bandwidth regime) and hence having zero spectral efficiency while, by increasing a bit E_b/N_0 from its minimum value, the required bandwidth for reliable communication is large but finite (wideband regime) as well as the spectral efficiency. The analysis on the wideband performance of our system, characterized by causal feedback and delay constraint, will be the topic of next chapter.

The non-causal policy achieving long-term average capacity per unit-energy. At this point is interesting to compare the optimal "one-shot" (causal) policy with the optimal non-causal policy achieving long-term average capacity per unit energy. We consider only the single user case, since we saw that in the multi-user case the long-term average capacity region is the Cartesian product of the single-user long-term average capacities. [2] If we allow the input to depend on the whole CSI $\mathcal{S}_N$ in a non-causal way, it is immediate to show that the optimal policy is "maximum selection"

$$\beta_{k,n}^{\star\,(\mathrm{nc})} = \begin{cases} \frac{N\gamma_k}{|M_k|} & \text{if } n \in M_k \\ 0 & \text{otherwise} \end{cases} \tag{4.29}$$

where

$$M_k = \{n \,:\, \alpha_{k,n} = \max\{\alpha_{k,1}, \cdots, \alpha_{k,N}\}\} \tag{4.30}$$

[2] The K-user capacity region per unit cost is equal to the Cartesian product of the K single-user capacities per unit cost only if a) every user has an alphabet that contains a symbol of zero cost and b) for a given user, the use of the zero-cost symbol by all the other users corresponds to the most favorable single-user channel seen by the considered user. Those two hypothesis are always satisfy by additive channels.

and $|M_k| \in \{1, \cdots, N\}$ denotes the cardinality of the non-empty set M_k. Power policy (4.29) equally divides the available energy among the slots whose fading is equal to the maximum. Hence, the non-causal long-term average capacity per unit energy is

$$s_N^{(k,\mathrm{nc})} = \mathrm{E}[\max\{\alpha_{k,1}, \cdots, \alpha_{k,N}\}] \tag{4.31}$$

Notice that, with continuous fading distribution, $\Pr[|M_k| > 1] = 0$ and interestingly, also in the non-causal setting, the optimal policy achieving capacity per unit energy is "one-shot".

4.5 Numerical results

In this section we give numerical values of the long-term average capacity per unit energy for two types of channel: *discrete two states fading channel* and *continuous Rayleigh fading channel*. We also compare the long-term average rates per unit energy achievable with causal feedback with those achievable with non-causal feedback.

The two states fading channel. This fading statistics models a communication system with a line of sight, as low orbit satellite communication systems. The fading can be either $\alpha = 0$ (bad channel, i.e., no line of sight) or $\alpha = 1$ (good channel). The probability of the good state is $\delta = \Pr[\alpha = 1] = 1 - \Pr[\alpha = 0]$ with $\delta \in [0, 1]$. For this channel $\mathrm{E}[\alpha] = \delta$ and $\sup\{\alpha\} = 1$. We have that long-term average capacity per unit energy (4.23) is

$$s_N = 1 - (1 - \delta)^N \tag{4.32}$$

Note that for this special channel $s_N = s_N^{(\mathrm{nc})}$ i.e. the non-causal knowledge of the channels gains does not improve the performance of the system. Fig. 4.2 shows the value of s_N as a function of N for the two states fading channel. With delay $N = 6$ the performance are almost that of the ergodic system (attainable for $N \to \infty$).

The Rayleigh fading channel. The channels gain are i.i.d with cdf $F_\alpha(x) = 1 - \mathrm{e}^{-x}$ for $x \geq 0$. For this channel $\mathrm{E}[\alpha] = 1$ and $\sup\{\alpha\} = \infty$. The long-term average capacity per unit energy (4.23) can be computed from the recursion

$$s_N = s_{N-1} + \mathrm{e}^{-s_N - 1} \tag{4.33}$$

with initial condition $s_0 = 0$. With non-causal knowledge of the channels gains the long-term average capacity per unit energy (4.31) is

$$s_N^{(\mathrm{nc})} = \sum_{n=1}^{N} \binom{N}{n} \frac{(-1)^{n+1}}{n} \tag{4.34}$$

strictly larger than s_N for all $N > 1$. Fig. 4.3 shows the value of s_N and $s_N^{(\mathrm{nc})}$ as a function of N for the Rayleigh fading channel. For example, with delay $N = 5$ the system doubles its capacity per unit energy, i.e., $s_5 \approx 2s_1$, which means that the required minimum transmit energy per bit (4.28) is 3dB lower than in the case $N = 1$.

The "one-shot" policy is only optimal in the low SNR regime. Consider the single-user long-term average rate that can be achieved by applying policy $\boldsymbol{\beta}^\star$. In order to be always inside the instantaneous fading dependent capacity region, the user must encode on the slot n at rate

$$r_n^\star = \log(1 + \alpha_n N\gamma\, 1\{n^\star = n\}) \tag{4.35}$$

where $n^\star$ is given in (4.25). Hence, on a long-term average, the user acchieves rate $C_{1,N}^\star(\gamma)$ given by

$$\begin{aligned} C_{1,N}^\star(\gamma) &= \frac{1}{N}\,\mathrm{E}\left[\sum_{n=1}^{N} \log(1 + \alpha_n N\gamma\, 1\{n^\star = n\})\right] \\ &= \frac{1}{N} S_N^\star(N\gamma) \end{aligned} \tag{4.36}$$

where $S_N^\star(P)$ is given by the Dynamic Programming recursion

$$\begin{aligned} S_N^\star(P) &= \Pr[\alpha_j < s_{N-1}] S_{N-1}^\star(P) \\ &\quad + \int_{s_{N-1}}^{\infty} \log\,(1 + Px)\,\mathrm{d}F_\alpha(x) \end{aligned} \tag{4.37}$$

for $n = 1, \cdots, N$ and with initial condition $S_0^\star(P) = 0$. Fig. 4.4 shows $C_{1,N}^\star(\gamma)$ for the Rayleigh fading case for different value of N. Note that for small γ, $C_{1,N}^\star(\gamma)$ increases with N but for higher γ it decreases, proving that $\boldsymbol{\beta}^\star$ is optimal only in the energy limited (low SNR or wideband) regime. As N increases the rate $C_{1,N}^\star(\gamma)$ drops to zero. Fig. 4.5 shows the spectral efficiency for the Rayleigh fading case for different value of N as function of E_b/N_0.

4.6 Conclusions

In this chapter we have analyzed an idealized fading model where each codeword sees N independently drawn fading states, known to the transmitter causally. The power control algorithm at the transmitter must decide what portion of the available energy to allocate to each fading state based only on the knowledge of current and past fading states. We have solved for the optimal power control policy and capacity for fixed arbitrary N and for arbitrary number of users. The optimal policy is to concentrate all the energy

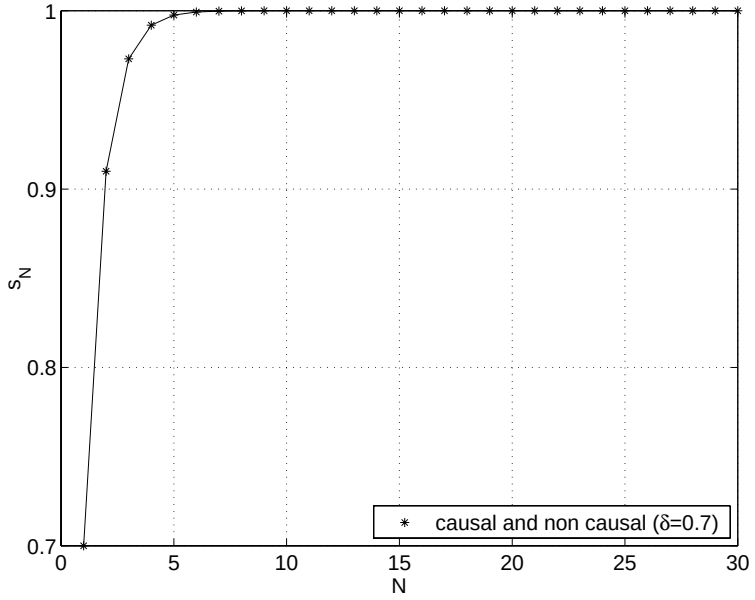

Figure 4.2: s_N vs. N for the two states channel.

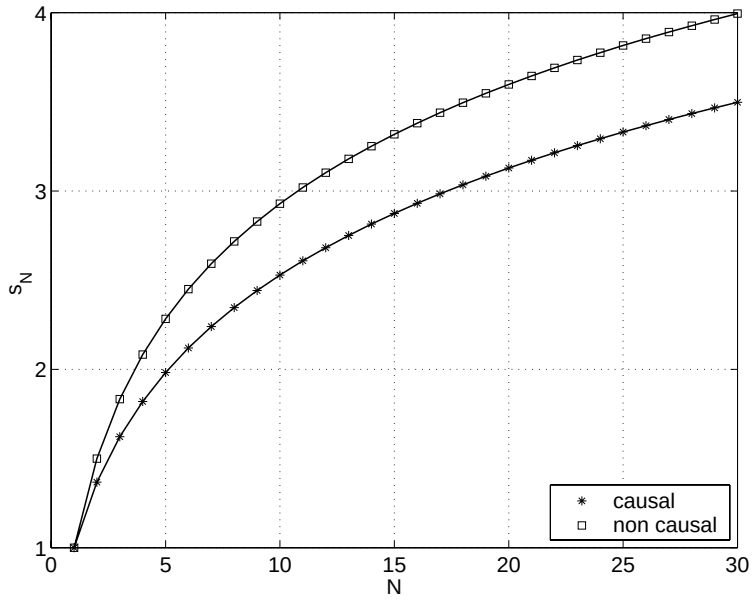

Figure 4.3: s_N vs. N for the Rayleigh channel.

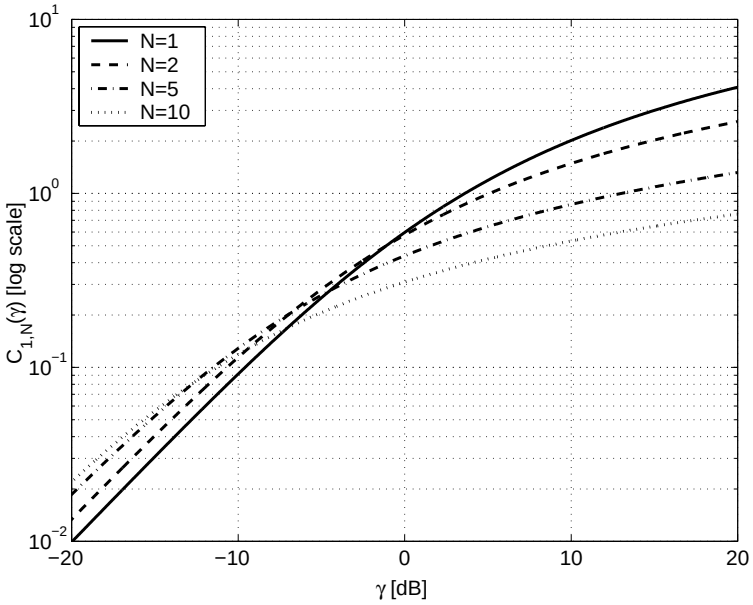

Figure 4.4: $C^{\star}_{1,N}$ vs. γ for the Rayleigh channel.

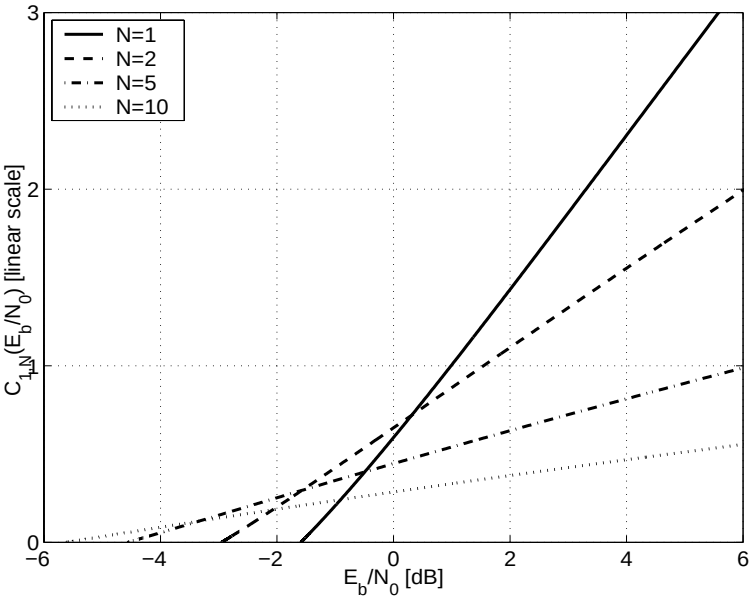

Figure 4.5: $C^{\star}_{1,N}$ vs. E_b/N_0 for the Rayleigh channel.

in only one of the fading states. That state is chosen on the basis of not only its strength, but also how likely it is that a more favorable fading state will appear before the end of the codeword.

We insisted that the "one-shot" power policy is simple, decentralized and we listed a number of "practical" advantages of this fact. Actually, in order to achieve rate points on the closure of the long-term average capacity $C_{K,N}(\boldsymbol{\gamma})$ users not only must use their power on the most favorable channel conditions but they also need to coordinate their transmission rate in order to be always inside the instantaneous fading dependent capacity region. To be more clear, let assume that there is only one user in the system and that allocates power according to $\boldsymbol{\beta}^\star$ and rate according to (4.35), then it achieves rate $C^\star_{1,N}(\gamma)$ given in (4.37). Add to the system another user that transmits with the same rate/power allocation policy. If the two users happen to transmit on the same slot ($n_1^\star = n_2^\star$) then the receiver can not jointly decode them and the system is in outage. The probability of outage, i.e., $\mathrm{P}_{\mathrm{out}}(N) \stackrel{\Delta}{=} \Pr[n_1^\star = n_2^\star]$, can be computed with the following recursion

$$\begin{aligned} \mathrm{P}_{\mathrm{out}}(N) &= \left(1 - F_\alpha^{(1)}(s_{N-1}^{(1)})\right)\left(1 - F_\alpha^{(2)}(s_{N-1}^{(2)})\right) \\ &\quad + F_\alpha^{(1)}(s_{N-1}^{(1)}) F_\alpha^{(2)}(s_{N-1}^{(2)}) \cdot \mathrm{P}_{\mathrm{out}}(N-1) \end{aligned} \tag{4.38}$$

with initial condition $\mathrm{P}_{\mathrm{out}}(0) = 0$. Fig. 4.6 shows the probability of outage in the Rayleigh fading case as a function of N. It can be can seen that at $N = 10$ the two users are going to collide on 1 frame out of 10 ($\mathrm{P}_{\mathrm{out}}(10) = 0.1$). To avoid outage users must coordinate their rates.

Could rate coordination be avoided by using other strategies, like TDMA? The question is legitimate since, from the proof of Theorem 5, the long-term capacity per unit energy can be achieved either with superposition coding and optimum joint decoding but also with TDMA inside each slot. Actually, the analysis that lead to the derivation of the long-term capacity per unit energy deals with the "infinite bandwidth regime" and not with the "wideband regime". Recent works [30, 32] have shown that actually TDMA can be heavily sub-optimal, in term of achievable rates, especially in a multiuser faded environment which is intrinsically rich in diversity. In fact, when many users are active in a faded environment, with high probability the best user enjoys channel gain that is larger than its average, hence the performance is dominated not by the average but by the maximum. We shall go deeply into the subject in the next chapter.

What do we conclude? In order to fully exploit diversity we need joint processing in the form of rate coordination and joint decoding at the receiver, but this is "expensive" in complexity.

From this theoretical analysis we can draw some guidelines for the design of practical systems. In wideband regime, in order to optimize the average number of received bits per joule, sequential polling of the active users by

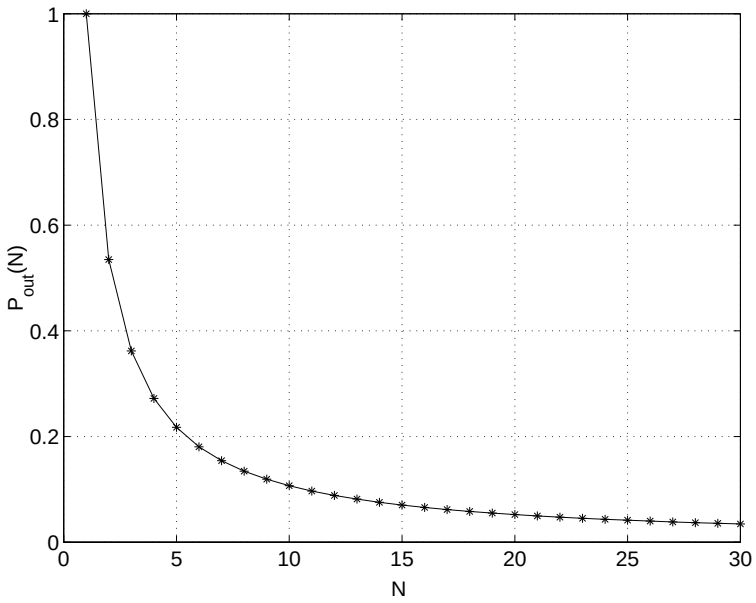

Figure 4.6: $\mathrm{P_{out}}(N)$ vs. N for the Rayleigh channel.

the master is not needed. This analysis suggests that the master station should send periodically a "probe" signal; if a user has a message to send, then it starts a timeout and measures the attenuation of its own channel on every slot of the frame: on the first slot where the channel gain is higher than the time varying threshold s it sends its packet with all the available energy, then it resets the timeout and waits for the next packet to send. Note that an optimal system actually does not require the time windows of the active users to be synchronous and the users can have different delay requirements, this allowing for extra flexibility.

We close the section, and the chapter, by describing a practical system where our results could be of importance. Although our model is admittedly quite simplified with respect to the reality, it fits the characteristics of a wireless sensor network. Briefly, a WSN comprises many stationary nodes and a very small number of mobile nodes. Unlike terminals in conventional wireless networks, i.e., ad-hoc networks or cellular systems, the sensors nodes in a WSN operate under very dynamic/different conditions (take measurements, elaborate the acquired data, discover other nodes to establish links, act as relay for other nodes in case of lack of connectivity or unreachable target, etc.) and work unattended. In order to guarantee network connection and long operational lifetime, energy must managed carefully, especially because sensors run on batteries whose frequent substitution might be impossible and/or impractical. In terms of energy consumption, transmitting

data on the air is of much higher cost than non-real time processing, hence source and channel coding are of primary importance in order to lower down the transmission rate. Mobile nodes periodically collect data from the sensors. They send a reference signal that sensors use to detect their presence, to synchronize and to measure their channels. Since the mobile moves, the channel from the mobile to the sensors changes over time, hence sensors must transmit at variable rate/quality in order to deliver in any case a useful message. Transmission must take place within the time the mobile node is reachable otherwise the data will be lost.

As an example, imagine sensors for tele-surveillance located over a large geographical area and a non-geostationary satellite that periodically flies above them so that the sensors are in the coverage of its spot beam antenna. The sensors have solar cells to charge their batteries. At every passage of the satellite, and within the time the satellite is reachable, the sensors send their data, an image or a measurement, by using the energy that they have stored in the batteries. Because of phenomena like tropospheric scattering, rain or physical obstacles, the channel between each sensor and the satellite is slowly time-varying and can be considered frequency flat. Due to time variation of atmospheric conditions, the rate at which reliable communication is possible is a random variable. Since in this kind of application it is important to deliver some measurement, even if not at the best quality, the sensors encode the data by layered source coding and, depending on the instantaneous channel conditions, transmit the fundamental coarse information and more or less refinement. Notice that source coding need not be in real time, so it costs (virtually) no energy. On the contrary, transmission must be done with the accumulated energy in the battery. A sensible criterion for this setting is to maximize the expected number of transmitted bit per joule.

Appendix

4.A Proof of Theorem 1

Achievability is easily obtained by considering a particular variable-rate coding system that encodes and decodes independently over the N slots. For each channel state [3] $\mathcal{S}_N$, the users construct a sequence of Gaussian code books of length L with i.i.d. entries of zero mean and unit variance and sizes $M_{k,n}(\mathcal{S}_n)$, satisfying the set of inequalities

$$\frac{1}{L}\sum_{k\in\mathcal{A}}\log\left(M_{k,n}(\mathcal{S}_n)\right) < \log\left(1+\sum_{k\in\mathcal{A}}\alpha_{k,n}\beta_{k,n}(\mathcal{S}_n)\right) \tag{4.39}$$

[3] For a rigorous treatment in the case where the fading is a continuous random vector we should use the argument of [13] based on the discretization of the fading state. For the sake of brevity, we cut short and we assume that we can define a code book *for each* channel state.

for all $\mathcal{A} \subseteq \{1, \cdots, K\}$, where $\boldsymbol{\beta} \in \Gamma_{K,N}(\boldsymbol{\gamma})$. Each transmitter k, during slot n, after observing $\mathcal{S}_n$, selects a message uniformly on $W_{k,n}(\mathcal{S}_n) = \{1, \ldots, M_{k,n}(\mathcal{S}_n)\}$ and independently of the other transmitters, and sends the corresponding codeword amplified by the transmit power level $\beta_{k,n}(\mathcal{S}_n)$. The receiver perform decoding on a slot-by-slot basis (even though it is allowed to wait until the end of the frame of N slots). From the standard Gaussian MAC [1], any rate K-tuple satisfying the set of inequalities (4.39) is achievable, i.e., the conditional decoding error probability given the channel state $\mathcal{S}_n$ vanishes as $L \to \infty$. By summing over N slots we get

$$\frac{1}{NL} \sum_{k \in \mathcal{A}} \log \left(\prod_{n=1}^{N} M_{k,n}(\mathcal{S}_n) \right) < \frac{1}{N} \sum_{n=1}^{N} \log \left(1 + \sum_{k \in \mathcal{A}} \alpha_{k,n} \beta_{k,n}(\mathcal{S}_n) \right) \quad (4.40)$$

with conditional (w.r.t. $\mathcal{S}_N$) error probability not larger than N times the maximum over the N slots of the conditional error probability in the n-th slot. Finally, by taking expectation with respect to the channel state of both sides in (4.40) and of the error probability and by applying Definition 2, we find that the set of rates defined in (4.7) are long-term average achievable.

For the converse part, we consider the N-slot extension of our channel, with input "blocks" $\mathbf{X}_k = \{\boldsymbol{x}_{k,n} : n = 1, \cdots, N\}$ and output "block" $\mathbf{Y} = \{\boldsymbol{y}_n : n = 1, \cdots, N\}$, where the input constraint is given "block-wise" by (4.3). [4] One frame of the original channel corresponds to a channel use of the new channel.

The new channel is (block-wise) memoryless. We consider a sequence of such channels indexed by increasing L, and define the capacity region of the N-slot extension channel as the closure of the union of all regions for $L = 1, 2, \cdots$. Any error probability (averaged over an arbitrarily large number of frames) achievable by a coding system constrained to perform coding and decoding frame-by-frame can be also achieved by performing coding and decoding over an arbitrarily long sequence of frames. Hence, any long-term achievable rate K-tuple of the original channel is achievable (in the usual ergodic sense) by the N-slot extension channel (rates are always expressed in nat per dimension of the original channel). We conclude that the capacity region of the N-slot extension channel is an outer bound to the long-term average capacity region of the original channel.

Let $\mathbf{X} = \{\mathbf{X}_k : k = 1, \ldots, K\}$ and, for any $\mathcal{A} \subseteq \{1, \ldots, K\}$, let $\mathbf{X}(\mathcal{A}) \stackrel{\Delta}{=} \{\mathbf{X}_k : k \in \mathcal{A}\}$ and $R(\mathcal{A}) \stackrel{\Delta}{=} \sum_{k \in \mathcal{A}} R_k$. From standard results on memoryless MAC [6, 7, 36, 13, 1], the capacity region of the N-slot extension channel is given by

$$\bigcup_{\Pr(\mathbf{X}, V, \mathcal{S}_N)} \left\{ \boldsymbol{R} \in \mathbb{R}_+^K \ : \ R(\mathcal{A}) \leq \frac{1}{LN} I\left(\mathbf{X}(\mathcal{A}); \mathbf{Y} | \mathbf{X}(\overline{\mathcal{A}}), \mathcal{S}_N, V\right) \ \forall \mathcal{A} \subseteq \{1, \cdots, K\} \right\} \quad (4.41)$$

[4] Similar "blocking" techniques have been used to prove coding theorems for channels with ISI [15, 100].

where the joint probability of $(\mathbf{X}, V, \mathcal{S}_N)$ satisfies

$$\Pr(\mathbf{X}, V, \mathcal{S}_N) = \left(\prod_{k=1}^{K} \prod_{n=1}^{N} \Pr(\boldsymbol{x}_{k,n} | \mathcal{S}_n, V, \boldsymbol{x}_{k,1}, \cdots, \boldsymbol{x}_{k,n-1}) \right) \Pr(V | \mathcal{S}_N) \Pr(\mathcal{S}_N) \quad (4.42)$$

and each factor $\prod_{n=1}^{N} \Pr(\boldsymbol{x}_{k,n} | \mathcal{S}_n, V, \boldsymbol{x}_{k,1}, \cdots, \boldsymbol{x}_{k,n-1})$ puts zero probability outside the sphere $\frac{1}{NL} \sum_{n=1}^{N} |\boldsymbol{x}_{k,n}|^2 \leq \gamma_k$. The input probability in the form (4.42) expresses the fact that encoding is independent for all transmitters, when conditioned with respect to the common CSI $\mathcal{S}_N$ and the time-sharing variable V, and that the common CSI is causal, i.e., that $\boldsymbol{x}_{k,n}$ depends only on $\mathcal{S}_n$ and not on the whole $\mathcal{S}_N$. Notice that we allow the time-sharing variable V to depend on the whole CSI $\mathcal{S}_N$, even if the CSI is only revealed causally to the transmitters (again, this can only increase the capacity region).

Fix an input probability distribution $P(\mathbf{X}, V, \mathcal{S}_N)$ in the form (4.42) with conditional component-wise second-order moments

$$\beta_{k,n}^{(\ell)}(\mathcal{S}_n, V) = \mathrm{E}[|x_{k,n}^{(\ell)}|^2 | \mathcal{S}_n, V] \quad (4.43)$$

where $x_{k,n}^{(\ell)}$ denotes the ℓ-th component of $\boldsymbol{x}_{k,n}$. Since the channel is additive and the input second-order moment is constrained, the boundary of the region (4.41) is clearly achieved only if $P(\mathbf{X}, V, \mathcal{S}_N)$ satisfies $\mathrm{E}[\mathbf{X} | \mathcal{S}_N, V] = \mathbf{0}$. Then, we shall restrict to this case. Let $P(\mathbf{Y}, \mathbf{X}, \mathcal{S}_N, V)$ be the joint input-output probability corresponding to $P(\mathbf{X}, V, \mathcal{S}_N)$ and to the transition probability of the channel. Let $\Phi(\mathbf{Y}, \mathbf{X}, \mathcal{S}_N, V)$ be the joint input-output probability for input $\mathbf{X}$ conditionally Gaussian with independent components of zero conditional mean and conditional variance as in (4.43). Notice that such input distribution is valid, in the sense that it is in the form (4.42).

For every subset $\mathcal{A}$ we have

$$\begin{aligned} & I(\mathbf{X}(\mathcal{A}); \mathbf{Y} | \mathbf{X}(\overline{\mathcal{A}}), \mathcal{S}_N, V) \\ &= D\left(\Pr(\mathbf{Y} | \mathbf{X}, \mathcal{S}_N, V) \| \Phi(\mathbf{Y} | \mathbf{X}(\overline{\mathcal{A}}), \mathcal{S}_N, V) | \Pr(\mathbf{X}, \mathcal{S}_N, V)\right) \\ &\quad - D\left(\Pr(\mathbf{Y} | \mathbf{X}(\overline{\mathcal{A}}), \mathcal{S}_N, V) \| \Phi(\mathbf{Y} | \mathbf{X}(\overline{\mathcal{A}}), \mathcal{S}_N, V) | \Pr(\mathbf{X}, \mathcal{S}_N, V)\right) \\ &\overset{(a)}{\leq} D\left(\Pr(\mathbf{Y} | \mathbf{X}, \mathcal{S}_N, V) \| \Phi(\mathbf{Y} | \mathbf{X}(\overline{\mathcal{A}}), \mathcal{S}_N, V) | \Pr(\mathbf{X}, \mathcal{S}_N, V)\right) \\ &= D\left(\mathcal{N}_{\mathbb{C}}(\boldsymbol{\mu}, \mathbf{I}) \| \mathcal{N}_{\mathbb{C}}(\boldsymbol{\nu}, \boldsymbol{\Lambda}) | \Pr(\mathbf{X}, \mathcal{S}_N, V)\right) \end{aligned} \quad (4.44)$$

where (a) follows from the non-negativity of divergence [1] and where we defined the conditional mean vectors of dimension $NL \times 1$ as

$$\boldsymbol{\mu} = \begin{bmatrix} \sum_{k=1}^{K} c_{k,1} \boldsymbol{x}_{k,1} \\ \vdots \\ \sum_{k=1}^{K} c_{k,N} \boldsymbol{x}_{k,N} \end{bmatrix}, \quad \boldsymbol{\nu} = \begin{bmatrix} \sum_{k \notin \mathcal{A}} c_{k,1} \boldsymbol{x}_{k,1} \\ \vdots \\ \sum_{k \notin \mathcal{A}} c_{k,N} \boldsymbol{x}_{k,N} \end{bmatrix} \quad (4.45)$$

and the conditional covariance matrix of dimension $NL \times NL$ as

$$\mathbf{\Lambda} = \mathrm{diag}\left(1 + \sum_{k\in\mathcal{A}} \alpha_{k,1}\beta_{k,1}^{(1)}(\mathcal{S}_1, V), \ldots, 1 + \sum_{k\in\mathcal{A}} \alpha_{k,N}\beta_{k,N}^{(L)}(\mathcal{S}_N, V)\right) \quad (4.46)$$

By applying the general formula for the divergence of two Gaussian complex circularly symmetric distributions [99] we obtain

$$\begin{aligned}
& D\left(\mathcal{N}_{\mathbb{C}}(\boldsymbol{\mu}, \mathbf{I}) \| \mathcal{N}_{\mathbb{C}}(\boldsymbol{\nu}, \mathbf{\Lambda}) \,|\, \Pr(\mathbf{X}, \mathcal{S}_N, V)\right) \\
&\quad = \mathrm{E}\left[\log \prod_{n=1}^{N}\prod_{\ell=1}^{L}\left(1 + \sum_{k\in\mathcal{A}} \alpha_{k,1}\beta_{k,n}^{(\ell)}(\mathcal{S}_n, V)\right)\right. \\
&\qquad \left. + \sum_{n=1}^{N}\sum_{\ell=1}^{L} \frac{\left|\sum_{k\in\mathcal{A}} c_{k,n} x_{k,n}^{(\ell)}\right|^2 - \sum_{k\in\mathcal{A}} \alpha_{k,n}\beta_{k,n}^{(\ell)}(\mathcal{S}_n, V)}{1 + \sum_{k\in\mathcal{A}} \alpha_{k,1}\beta_{k,n}^{(\ell)}(\mathcal{S}_n, V)}\right] \\
&\quad \overset{(a)}{=} \mathrm{E}\left[\mathrm{E}\left[\sum_{n=1}^{N}\sum_{\ell=1}^{L} \log\left(1 + \sum_{k\in\mathcal{A}} \alpha_{k,1}\beta_{k,n}^{(\ell)}(\mathcal{S}_n, V)\right)\middle|\, \mathcal{S}_N\right]\right] \\
&\quad \overset{(b)}{\leq} \mathrm{E}\left[\mathrm{E}\left[\sum_{n=1}^{N} L \log\left(1 + \sum_{k\in\mathcal{A}} \alpha_{k,1}\beta_{k,n}(\mathcal{S}_n, V)\right)\middle|\, \mathcal{S}_N\right]\right] \\
&\quad \overset{(c)}{\leq} \mathrm{E}\left[\sum_{n=1}^{N} L \log\left(1 + \sum_{k\in\mathcal{A}} \alpha_{k,1}\beta_{k,n}(\mathcal{S}_n)\right)\right] \qquad (4.47)
\end{aligned}$$

where (a) follows by taking conditional expectation with respect to $\mathbf{X}$, given $\mathcal{S}_N$ and V, and by using the fact that, from (4.42) the $\mathbf{X}_k$ are mutually independent given $\mathcal{S}_N$ and V, (b) follows by defining $\beta_{k,n}(\mathcal{S}_n, V) \overset{\Delta}{=} \frac{1}{L}\sum_{\ell=1}^{L} \beta_{k,n}^{(\ell)}(\mathcal{S}_n, V)$ and from Jensen's inequality applied to the concave function $\log(1+x)$, and (c) follows by defining $\beta_{k,n}(\mathcal{S}_n) \overset{\Delta}{=} \mathrm{E}[\beta_{k,n}(\mathcal{S}_n, V)|\mathcal{S}_n]$ and again from Jensen's inequality.

From (4.44) and (4.47) we have that

$$\frac{1}{NL} I(\mathbf{X}(\mathcal{A}); \mathbf{Y}|\mathbf{X}(\overline{\mathcal{A}}), \mathcal{S}_N, V) \leq \mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N} \log\left(1 + \sum_{k\in\mathcal{A}} \alpha_{k,1}\beta_{k,n}(\mathcal{S}_n)\right)\right] \quad (4.48)$$

and that the LHS of (4.48) is achieved by degenerate V (i.e., a constant) and $P(\mathbf{X}|\mathcal{S}_N, V)$ Gaussian with conditionally independent elements $x_{k,n}^{(\ell)} \sim \mathcal{N}_{\mathbb{C}}(0, \beta_{k,n}(\mathcal{S}_n))$. Since this holds for arbitrary $\mathcal{A}$ and input distribution $P(\mathbf{X}, \mathcal{S}_N, V)$, we conclude that (4.41) coincides with (4.7), thus proving the converse.

4.B Dynamic Programming

Notation. A discrete time stochastic dynamic system is defined by the following equations (following the notation of [98]):

$$\left\{\begin{array}{lcll} y_n & = & h_n(x_n, z_n) & \text{measurement} \\ u_n & = & f_n(y_0, y_1, \cdots, y_n, u_0, u_1, \cdots, u_{n-1}) & \text{control command} \\ x_{n+1} & = & g_n(x_n, w_n, u_n) & \text{state evolution} \\ c_n & = & c_n(y_n, u_n) & \text{cost/reward} \end{array}\right. \tag{4.49}$$

where

- x_n is the state of the system at time n;
- y_n is the noisy measurement of the system state x_n;
- u_n is the control command defined on a set $\mathcal{U}_n$;
- w_n is an i.i.d. noise on the state;
- z_n is an i.i.d. noise on the measurement;
- c_n is the immediate cost/reward for being in "state" y_n and having applied a command u_n;

Indicate with x_1 the initial state and with $c_{N+1} = c_{N+1}(y_{N+1})$ the final cost/reward.

The Dynamic Programming Algorithm gives a recursive solution to the problem of finding the control policy $\{u_n\}_{n=0}^N$ that maximizes the sum of the rewards over a finite horizon [98]

$$J^\star = \max_{\{u_n \in \mathcal{U}_n\}} \sum_{n=1}^{N} c_n(y_n, u_n) + c_{N+1}(y_{N+1}) \tag{4.50}$$

Assuming perfect knowledge of the state, i.e., $y_n = x_n$, the Dynamic Programming Algorithm is: with initial condition $V_{N+1}(x) = c_{N+1}(x)$ compute for $n = N, \cdots, 1$ the functions

$$V_n(x) = \sup_{u \in \mathcal{U}_n} \{c_n(x, u) + \mathrm{E}_w [V_{n+1}(g_n(x, w, u))]\} \tag{4.51}$$

$$u_n^\star(x) = \operatorname*{argsup}_{u \in \mathcal{U}_n} \{c_n(x, u) + \mathrm{E}_w [V_{n+1}(g_n(x, w, u))]\} \tag{4.52}$$

then, the optimal value of the average cost is

$$J^\star = \mathrm{E}_{x_1}[V_1(x_1)] \tag{4.53}$$

and the optimal policy to be applied at step n with observed state (a random variable) x_n is

$$u_n^\star = u_n^\star(x_n) \tag{4.54}$$

that depends only on the current system state x_n and not on the whole sequence of states $\{x_j\}_{j=0}^{n}$ and control commands $\{u_j\}_{j=0}^{n-1}$, i.e., the optimal optimal policy is said to be "Markovian".

Notice that the "backward" recursion (4.51) can be written in form of a "forward" recursion by defining $\mathrm{E}_w\left[V_{n+1}(g_n(x,w,u))\right] = S_{N-n}(x)$.

The single-user case. Without loss of generality and for the sake of simplicity, we drop the user index k. Maximizing $\mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N}\alpha_n\beta_n(\mathcal{S}_n)\right]$ or $\mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N}\log(1+\alpha_n\beta_n(\mathcal{S}_n))\right]$, subject to $\frac{1}{N}\beta_n(\mathcal{S}_n) \leq \gamma$ falls in the class of optimal control of stochastic dynamical systems with an additive cost function over a finite horizon where the system state is measured without error [98]. In fact, identify

- The state of the system at time n

$$x_n = \begin{bmatrix} \alpha_n \\ P_n \end{bmatrix} \tag{4.55}$$

  where P_n is the energy still available at time n and where the sequence $\alpha_1, \cdots, \alpha_N$ is independent with $\alpha_n \sim F_\alpha(z)$, a given probability distribution. Note that α plays the role of w, the noise on the state.

- The initial state

$$x_1 = \begin{bmatrix} \alpha_1 \\ N\,\gamma \end{bmatrix} \tag{4.56}$$

- The energy allocation function at time n

$$p_n = f_n(x_1, \cdots, x_n, p_1, \cdots, p_{n-1}) \in [0, P_n] \equiv \mathcal{U}_n \tag{4.57}$$

  that can depend on all the observed channels states $\{x_j\}_{j=1}^{n}$ and on the already allotted energies $\{p_j\}_{j=1}^{n-1}$. Note that the power to allocate p plays the role of u, the control command.

- The system dynamics

$$x_{n+1} = g(x_n, \alpha_{n+1}, p_n) = \begin{bmatrix} \alpha_{n+1} \\ P_n - p_n \end{bmatrix} \tag{4.58}$$

  Note that the equation that defines the state evolution does not depend on the time index n, i.e., $g_n(\cdot) = g(\cdot)$.

- The immediate reward/cost

$$c_n(x_n) = \log(1 + \alpha_n p_n) \tag{4.59}$$

  or $c_n(x_n) = \alpha_n p_n$. Note that the equation that defines the reward does not depend on the time index n, i.e., $c_n(\cdot) = c(\cdot)$.

- The final reward/cost $c_{N+1}(x_{N+1}) = 0$.

With the above definitions, the problem at hands is equivalent to (4.50), a part a factor $1/N$, and is solved by (4.51) and (4.52). The optimal value is given by (4.53) and the optimal policy by (4.54).

The multi-user case. The generalization to the multi-user case follows straightforwardly when the additive total cost to be maximized is given by (4.12) and the system evolves, from slot n to slot $n+1$ according to $(\boldsymbol{\alpha}_n, \boldsymbol{P}) \to (\boldsymbol{\alpha}_{n+1}, \boldsymbol{P} - \widehat{\boldsymbol{p}}_n)$.

4.C Proof of Theorem 3

In order to fix ideas, we treat first the single-user case ($K = 1$). The proof of Theorem 3 follows by applying the same technique in the slightly more involved multiuser case.

For simplicity, we drop the user index k. With a slight abuse of notation, we indicate the single-user ergodic capacity with

$$\begin{aligned} & C_1^{(\mathrm{erg})}(\gamma) = \max_{\boldsymbol{\beta}} \mathrm{E}\left[\log\left(1+\alpha\beta(\alpha)\right)\right] \\ & \text{subject to} \quad \beta(\alpha) \geq 0 \quad \text{and} \quad \mathrm{E}[\beta(\alpha)] \leq \gamma \end{aligned} \tag{4.60}$$

the single-user long-term average capacity with causal CSI, delay N and "short-term" power constraint with

$$\begin{aligned} & C_{1,N}(\gamma) = \max_{\boldsymbol{\beta}} \mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N}\log\left(1+\alpha_n\beta_n(\mathcal{S}_n)\right)\right] \\ & \text{subject to} \quad \beta_n(\mathcal{S}_n) \geq 0 \quad \text{and} \quad \tfrac{1}{N}\textstyle\sum_{n=1}^{N}\beta_n(\mathcal{S}_n) \leq \gamma \end{aligned} \tag{4.61}$$

and the single-user long-term average capacity with non-causal CSI, delay N and "long-term" power constraint with

$$\begin{aligned} & C_{1,N}^{(\mathrm{LT-nc})}(\gamma) = \max_{\boldsymbol{\beta}} \mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N}\log\left(1+\alpha_n\beta_n(\mathcal{S}_N)\right)\right] \\ & \text{subject to} \quad \beta_n(\mathcal{S}_N) \geq 0 \quad \text{and} \quad \mathrm{E}\left[\tfrac{1}{N}\textstyle\sum_{n=1}^{N}\beta_n(\mathcal{S}_N)\right] \leq \gamma \end{aligned} \tag{4.62}$$

When user k is considered, the mean values in (4.60), (4.61) and (4.62) are computed with respect to α_n i.i.d. $\sim F_\alpha^{(k)}(x)$ and for $\gamma = \gamma_k$.

Problem (4.60) has solution [11]

$$C_1^{(\mathrm{erg})}(\gamma) = \mathrm{E}\left[\log\left(1+\alpha\beta^{(\mathrm{erg})}(\alpha;\gamma)\right)\right] = \mathrm{E}\left[\log\left(\frac{\alpha}{\lambda}\right)^{+}\right] \tag{4.63}$$

where $\beta^{(\mathrm{erg})}(\alpha;\gamma)$ is the ergodic waterfilling power allocation

$$\beta^{(\mathrm{erg})}(\alpha;\gamma) = \left[\frac{1}{\lambda} - \frac{1}{\alpha}\right]^+ \tag{4.64}$$

and the Lagrangian multiplier λ satisfies

$$\mathrm{E}\left[\beta^{(\mathrm{erg})}(\alpha;\gamma)\right] = \gamma \tag{4.65}$$

It is immediate to see that for every N

$$C_{1,N}(\gamma) \leq C_{1,N}^{(\mathrm{LT-nc})}(\gamma) = C_1^{(\mathrm{erg})}(\gamma) \tag{4.66}$$

where the inequality in (4.66) follows since the set of feasible "short-term" causal power allocations is a subset of the set of feasible "long-term" non-causal power allocations, and the equality in (4.66) follows straightforwardly. It is also easy to see that, since $C_1^{(\mathrm{erg})}(\gamma)$ is a non-decreasing continuous function of γ, for every $\epsilon > 0$ it exist $\delta > 0$ such that

$$C_1^{(\mathrm{erg})}(\gamma) + \epsilon = C_1^{(\mathrm{erg})}(\gamma + \delta) \tag{4.67}$$

Next, we find a lower bound on $C_{1,N}(\gamma)$ by choosing a particular "short-term" causal power allocation policy, and we show that, in the limit for $N \to \infty$, the lower bound can be made arbitrarily close to the upper bound $C_1^{(\mathrm{erg})}(\gamma)$. For every N and for $\delta \in [0,\gamma]$, consider the (sub-optimal) power allocation $\widetilde{\boldsymbol{\beta}} \in \Gamma_{1,N}(\gamma)$ defined by

$$\widetilde{\beta}_n(\mathcal{S}_n) = \begin{cases} \beta^{(\mathrm{erg})}(\alpha_n;\gamma-\delta) & \text{if } \sum_{i=1}^n \widetilde{\beta}_i(\mathcal{S}_i) \leq N\gamma \\ 0 & \text{otherwise} \end{cases} \tag{4.68}$$

Hence, the desired lower bound is given by

$$\mathrm{E}\left[\left(\frac{1}{N}\sum_{n=1}^N \log\left(1+\alpha_n\beta^{(\mathrm{erg})}(\alpha_n;\gamma-\delta)\right)\right) 1\left\{\frac{1}{N}\sum_{n=1}^N \beta^{(\mathrm{erg})}(\alpha_n;\gamma-\delta) \leq \gamma\right\}\right] \tag{4.69}$$

Note that both $\{\log\left(1+\alpha_n\beta^{(\mathrm{erg})}(\alpha_n;\gamma-\delta)\right)\}$ and $\{\beta^{(\mathrm{erg})}(\alpha_n;\gamma-\delta)\}$ are i.i.d. random variables for all n. Since $\mathrm{E}[\beta^{(\mathrm{erg})}(\alpha_n;\gamma-\delta)] = \gamma - \delta$ by definition (4.65) and because of the law of large numbers, the indicator function $1\left\{\frac{1}{N}\sum_{n=1}^N \beta^{(\mathrm{erg})}(\alpha_n;\gamma-\delta) \leq \gamma\right\}$ tends to the constant value 1 almost surely. For the same reasons, $\frac{1}{N}\sum_{n=1}^N \log\left(1+\alpha_n\beta^{(\mathrm{erg})}(\alpha_n;\gamma-\delta)\right)$ tends to $\mathrm{E}[\log\left(1+\alpha_n\beta^{(\mathrm{erg})}(\alpha_n;\gamma-\delta)\right)] = C_1^{(\mathrm{erg})}(\gamma-\delta)$ almost surely. Hence, because of (4.67), we have that the RHS of (4.69) converges almost surely to $C_1^{(\mathrm{erg})}(\gamma) - \epsilon$ for some $\epsilon > 0$. Finally, since

$$C_1^{(\mathrm{erg})}(\gamma) - \epsilon \leq \lim_{N\to\infty} C_{1,N}(\gamma) \leq C_1^{(\mathrm{erg})}(\gamma) \tag{4.70}$$

holds for every $\epsilon > 0$, we have that

$$\lim_{N\to\infty} C_{1,N}(\gamma) = C_1^{(\mathrm{erg})}(\gamma) \tag{4.71}$$

In order to extend this result to the multiuser case and prove the statement of Theorem 3, we consider the explicit characterization of the boundary of $C_K^{(\mathrm{erg})}(\boldsymbol{\gamma})$ given in [13]. A rate K-tuple $\boldsymbol{R} = (R_1, \ldots, R_K)$ is on the boundary surface of $C_K^{(\mathrm{erg})}(\boldsymbol{\gamma})$ if it is the solution of

$$\max_{\boldsymbol{R} \in C_K^{(\mathrm{erg})}(\boldsymbol{\gamma})} \sum_{k=1}^{K} \mu_k R_k \tag{4.72}$$

for some $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_K) \in \mathbb{R}_+^K$. A point $(R_1^{(\mathrm{erg})}(\boldsymbol{\mu}, \boldsymbol{\gamma}), \cdots, R_K^{(\mathrm{erg})}(\boldsymbol{\mu}, \boldsymbol{\gamma}))$ is solution of the above problem if it exists a vector of Lagrangian multipliers $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_K) \in \mathbb{R}_+^K$ such that

$$\mathrm{E}[\beta_k^{(\mathrm{erg})}(\boldsymbol{\alpha}; \boldsymbol{\mu}, \boldsymbol{\gamma})] = \gamma_k \tag{4.73}$$

$$\mathrm{E}[r_k^{(\mathrm{erg})}(\boldsymbol{\alpha}; \boldsymbol{\mu}, \boldsymbol{\gamma})] = R_k^{(\mathrm{erg})}(\boldsymbol{\mu}, \boldsymbol{\gamma}) \tag{4.74}$$

where the average is with respect to $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_K)$ and

$$\left\{ \begin{array}{l} u_k(z) \stackrel{\Delta}{=} \frac{\mu_k}{1+z} - \frac{\lambda_k}{\alpha_k} \quad \text{for} \quad z \in \mathbb{R}_+ \\ u^\star(z) \stackrel{\Delta}{=} \max\limits_{k=1,\cdots,K} \{[u_k(z)]^+\} \\ \beta_k^{(\mathrm{erg})}(\boldsymbol{\alpha}; \boldsymbol{\mu}, \boldsymbol{\gamma}) \stackrel{\Delta}{=} \frac{1}{\alpha_k} \int 1\{u_k(z) = u^\star(z)\} \mathrm{d}z \\ r_k^{(\mathrm{erg})}(\boldsymbol{\alpha}; \boldsymbol{\mu}, \boldsymbol{\gamma}) \stackrel{\Delta}{=} \int \frac{1}{1+z} 1\{u_k(z) = u^\star(z)\} \mathrm{d}z \end{array} \right. \tag{4.75}$$

Note that $r_k^{(\mathrm{erg})}(\boldsymbol{\alpha}; \boldsymbol{\mu}, \boldsymbol{\gamma})$ and $\beta_k^{(\mathrm{erg})}(\boldsymbol{\alpha}; \boldsymbol{\mu}, \boldsymbol{\gamma})$ are, respectively, the instantaneous rate and instantaneous power allocated to user k in fading state $\boldsymbol{\alpha}$. It is clear that if $\boldsymbol{\gamma}_1 \leq \boldsymbol{\gamma}_2$ then $C_K^{(\mathrm{erg})}(\boldsymbol{\gamma}_1) \subseteq C_K^{(\mathrm{erg})}(\boldsymbol{\gamma}_2)$ and for any $\boldsymbol{\mu} \in \mathbb{R}_+^K$

$$\max_{\boldsymbol{R} \in C_{K,N}(\boldsymbol{\gamma}_1)} \sum_{k=1}^{K} \mu_k R_k \leq \max_{\boldsymbol{R} \in C_{K,N}(\boldsymbol{\gamma}_2)} \sum_{k=1}^{K} \mu_k R_k \tag{4.76}$$

Conversely, if (4.76) holds for any direction vector $\boldsymbol{\mu}$, then $C_K^{(\mathrm{erg})}(\boldsymbol{\gamma}_1) \subseteq C_K^{(\mathrm{erg})}(\boldsymbol{\gamma}_2)$ and $\boldsymbol{\gamma}_1 \leq \boldsymbol{\gamma}_2$.

With arguments analogous to the single-user case, we can show that the upper bound $C_{K,N}(\boldsymbol{\gamma}) \subseteq C_{K,N}^{(\mathrm{LT-nc})}(\boldsymbol{\gamma}) \equiv C_K^{(\mathrm{erg})}(\boldsymbol{\gamma})$ holds for every delay N. For an arbitrary direction $\boldsymbol{\mu} \in \mathbb{R}_+^K$, an inner bound to $C_{K,N}(\boldsymbol{\gamma})$ is obtained by fixing the allocation policy $\widetilde{\boldsymbol{\beta}}$ as follows: for given $\boldsymbol{\delta} \in \mathbb{R}_+^K$ such that $\boldsymbol{\gamma} - \boldsymbol{\delta} \geq 0$, we define

$$\widetilde{\beta}_{k,n}(\mathcal{S}_n) = \left\{ \begin{array}{ll} \beta_k^{(\mathrm{erg})}(\boldsymbol{\alpha}_n; \boldsymbol{\mu}, \boldsymbol{\gamma} - \boldsymbol{\delta}) & \text{if } \sum_{j=1}^{n} \widetilde{\beta}_{k,j}(\mathcal{S}_j) \leq N \gamma_k \\ 0 & \text{otherwise} \end{array} \right. \tag{4.77}$$

The inner bound implies that

$$\mathrm{E}\left[\left(\sum_{k=1}^{K}\mu_k\frac{1}{N}\sum_{n=1}^{N}r_k^{(\mathrm{erg})}(\boldsymbol{\alpha}_n;\boldsymbol{\mu},\boldsymbol{\gamma}-\boldsymbol{\delta})\right)\prod_{k=1}^{K}1\left\{\frac{1}{N}\sum_{n=1}^{N}\beta_k^{(\mathrm{erg})}(\boldsymbol{\alpha}_n;\boldsymbol{\mu},\boldsymbol{\gamma}-\boldsymbol{\delta})\le\gamma_k\right\}\right]$$
$$\le\sum_{k=1}^{K}\mu_k\widehat{R}_{k,N}(\boldsymbol{\mu},\boldsymbol{\gamma}) \tag{4.78}$$

where $\widehat{R}_{k,N}(\boldsymbol{\mu},\boldsymbol{\gamma})$ are the rates on the boundary surface of $C_{K,N}(\boldsymbol{\gamma})$, given in (4.15). Now, since both $\{\sum_{k=1}^{K}\mu_k\, r_k^{(\mathrm{erg})}(\boldsymbol{\alpha}_n;\boldsymbol{\mu},\boldsymbol{\gamma}-\boldsymbol{\delta})\}$ and $\{\beta_k^{(\mathrm{erg})}(\boldsymbol{\alpha}_n;\boldsymbol{\mu},\boldsymbol{\gamma}-\boldsymbol{\delta})\}$ are i.i.d. random variables for all n, the indicator functions in the RHS of (4.78) tend to the constant value 1 almost surely and the sum of instantaneous rates tends to $\sum_{k=1}^{K}\mu_k\, R_k^{(\mathrm{erg})}(\boldsymbol{\mu},\boldsymbol{\gamma}-\boldsymbol{\delta})$ almost surely. Again, the RHS of (4.78) converges almost surely to $\sum_{k=1}^{K}\mu_k R_k^{(\mathrm{erg})}(\boldsymbol{\mu},\boldsymbol{\gamma}-\boldsymbol{\delta})$ and hence

$$\sum_{k=1}^{K}\mu_k R_k^{(\mathrm{erg})}(\boldsymbol{\mu},\boldsymbol{\gamma}-\boldsymbol{\delta})\le\lim_{N\to\infty}\sum_{k=1}^{K}\mu_k\widehat{R}_{k,N}(\boldsymbol{\mu},\boldsymbol{\gamma})\le\sum_{k=1}^{K}\mu_k R_k^{(\mathrm{erg})}(\boldsymbol{\mu},\boldsymbol{\gamma}) \tag{4.79}$$

Since $\boldsymbol{\delta}$ is arbitrary and (4.79) holds for any $\boldsymbol{\mu}$, we conclude that

$$\lim_{N\to\infty}C_{K,N}(\boldsymbol{\gamma})\equiv C_K^{(\mathrm{erg})}(\boldsymbol{\gamma}) \tag{4.80}$$

4.D Proof of Theorem 5

In the following we indicate with $C_{1,N}^{(k)}(\gamma)$ the single-user long-term average capacity for user k as defined in (4.61), where the extra superscript "(k)" stresses the fact that the mean value is computed using cdf $F_\alpha^{(k)}(x)$. Note that $C_{1,N}^{(k)}(\gamma)=\widehat{R}_{k,N}(\mathbf{1}_k,\boldsymbol{\gamma})$ for $\widehat{R}_{k,N}(\boldsymbol{\mu},\boldsymbol{\gamma})$ defined in (4.15) and where $\mathbf{1}_k$ is the vector of length K of all zeros but a "1" in position k.

Consider the following inner and outer bounds for $C_{K,N}(\boldsymbol{\gamma})$

$$\left\{\boldsymbol{R}\in\mathbb{R}_+^K:R_k\le\frac{1}{K}C_{1,N}^{(k)}(K\gamma_k)\right\}\subseteq C_{K,N}(\boldsymbol{\gamma})\subseteq\left\{\boldsymbol{R}\in\mathbb{R}_+^K:R_k\le C_{1,N}^{(k)}(\gamma_k)\right\} \tag{4.81}$$

where the inner bound is clearly achievable by TDMA, i.e., by letting each user transmit for a fraction $1/K$ of the slot time, and the outer bound is the Cartesian product of the single user long-term average capacity regions. Theorem 4 implies the following inner and outer bounds for $U_{K,N}$

$$\left\{\boldsymbol{r}\in\mathbb{R}_+^K:r_k\le\sup_{\gamma_k>0}\frac{1}{K\gamma_k}C_{1,N}^{(k)}(K\gamma_k)\right\}\subseteq U_{K,N}\subseteq\left\{\boldsymbol{r}\in\mathbb{R}_+^K:r_k\le\sup_{\gamma_k>0}\frac{1}{\gamma_k}C_{1,N}^{(k)}(\gamma_k)\right\} \tag{4.82}$$

Define the feasible power allocation policy

$$(\beta_{k,1}^\star,\cdots,\beta_{k,N}^\star)=\arg\sup_{\boldsymbol{\beta}\in\Gamma_{1,N}(\gamma_k)}\mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N}\alpha_{k,n}\beta_{k,n}(\mathcal{S}_n)\right] \tag{4.83}$$

and indicate with $(\widehat{\beta}_{k,1}, \cdots, \widehat{\beta}_{k,N})$ the k-th user single-user long-term average capacity achieving policy. The boundary surface of the outer region in (4.82) is given by

$$\begin{aligned}
\sup_{\gamma_k>0} \frac{1}{\gamma_k} C_{1,N}^{(k)}(\gamma_k) &= \sup_{\gamma_k>0} \frac{\mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N} \log\left(1+\alpha_{k,n}\widehat{\beta}_{k,n}\right)\right]}{\gamma_k} \\
&\overset{(a)}{=} \lim_{\gamma_k\to 0} \frac{\mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N} \log\left(1+\alpha_{k,n}\widehat{\beta}_{k,n}\right)\right]}{\gamma_k} \\
&\overset{(b)}{=} \lim_{\gamma_k\to 0} \frac{\mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N} \alpha_{k,n}\beta^{\star}_{k,n}\right]}{\gamma_k} \triangleq s_N^{(k)} \qquad (4.84)
\end{aligned}$$

where: (a) follows since $C_{1,N}^{(k)}(\gamma_k)$ is concave in γ_k (see the Corollary to Lemma 1 at the end of this section) and (b) follows for Lemma 2 at the end of this section.

With similar steps, we find that the boundary surface of the inner region in (4.82) is also given by (4.84). We conclude that the K-user long-term average capacity region per unit energy is the hyper-rectangle

$$U_{K,N} = \left\{ \boldsymbol{r} \in \mathbb{R}_+^K : r_k \le s_N^{(k)} \right\}. \qquad (4.85)$$

for $s_N^{(k)}$ given in (4.84) and that $\boldsymbol{\beta}^{\star} = \{\boldsymbol{\beta}^{\star}_{k,n} \ : \ k = 1, \cdots, K, n = 1, \cdots, N\}$ is the optimal K-user long-term average capacity region per unit energy achieving policy. □

In the following we will drop the superscript "(k)" since no confusion can arise.

Lemma 1. $C_{1,N}(\gamma)$ given in (4.61) is a concave function of γ.

Proof. Consider the single-user long-term average capacity achieving power allocation, that for notation convenience we re-write as follow

$$\left(\widehat{\beta}_1(\mathcal{S}_1;\gamma), \cdots, \widehat{\beta}_N(\mathcal{S}_N;\gamma)\right) = \arg \sup_{\boldsymbol{\beta}\in\Gamma_{1,N}(\gamma)} \mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N} \log\left(1+\alpha_n\beta_n(\mathcal{S}_n)\right)\right] \qquad (4.86)$$

to explicitly denote the dependency on the constraint γ. For every $\lambda \in [0,1]$ and for every $\gamma_a, \gamma_b \geq 0$ consider the convex combination

$$\begin{aligned}
&\lambda C_{1,N}(\gamma_a) + (1-\lambda) C_{1,N}(\gamma_b) \\
&\quad = \lambda \, \mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N} \log\left(1+\alpha_n \widehat{\beta}_n(\mathcal{S}_n;\gamma_a)\right)\right] \\
&\qquad +(1-\lambda)\,\mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N} \log\left(1+\alpha_n \widehat{\beta}_n(\mathcal{S}_n;\gamma_b)\right)\right] \\
&\quad \overset{(a)}{\leq} \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}\left[\log\left(1+\alpha_n \lambda\widehat{\beta}_n(\mathcal{S}_n;\gamma_a) + \alpha_n(1-\lambda)\widehat{\beta}_n(\mathcal{S}_n;\gamma_b)\right)\right] \\
&\quad \overset{(b)}{\leq} \frac{1}{N}\sum_{n=1}^{N} \mathrm{E}\left[\log\left(1+\alpha_n \widehat{\beta}_n(\mathcal{S}_n;\lambda\gamma_a + (1-\lambda)\gamma_b)\right)\right] \\
&\quad = C_{1,N}(\lambda\gamma_a + (1-\lambda)\gamma_b) \qquad (4.87)
\end{aligned}$$

where: (a) follows from Jensen's inequality and (b) because the feasible power policy $\lambda\widehat{\boldsymbol{\beta}}(\cdot;\gamma_a) + (1-\lambda)\widehat{\boldsymbol{\beta}}(\cdot;\gamma_b)$ does not coincide in general with the optimal power allocation (4.86) for $\gamma = \lambda\gamma_a + (1-\lambda)\gamma_b$. □

Corollary. Since $C_{1,N}(\gamma)$ is nonnegative and concave we have

$$\sup_{\gamma>0} \frac{C_{1,N}(\gamma)}{\gamma} = \dot{C}_{1,N}(0) \qquad (4.88)$$

where $\dot{C}_{1,N}(0)$ denotes the first derivative of $C_{1,N}(\gamma)$ at $\gamma = 0$.
In fact, since $C_{1,N}(\gamma)$ is concave, its second derivative is non-positive, i.e., $\ddot{C}_{1,N}(\gamma) \leq 0$, and hence its first derivative is non-increasing, i.e. $\dot{C}_{1,N}(\gamma) \leq \dot{C}_{1,N}(0)$. Since $C_{1,N}(\gamma)$ is nonnegative, by integrating both sides of the inequality $\dot{C}_{1,N}(\gamma) \leq \dot{C}_{1,N}(0)$ and imposing the initial condition $C_{1,N}(0) = 0$ we get

$$0 \leq C_{1,N}(\gamma) \leq \gamma \dot{C}_{1,N}(0) \qquad (4.89)$$

hence (4.88) follows. □

Lemma 2. Let $\widehat{\boldsymbol{\beta}} = \arg\max \mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N} \log\left(1+\alpha_n\beta_n(\mathcal{S}_n)\right)\right]$ and $\boldsymbol{\beta}^\star = \arg\max \mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N} \alpha_n\beta_n(\mathcal{S}_n)\right]$, where in both case $\boldsymbol{\beta} \in \Gamma_{1,N}(\gamma)$, then the following relation holds

$$\begin{aligned}
\mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N} \log\left(1+\alpha_n\beta_n^\star\right)\right] &\overset{(a)}{\leq} \mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N} \log\left(1+\alpha_n\widehat{\beta}_n\right)\right] \\
&\overset{(b)}{\leq} \mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N} \alpha_n\widehat{\beta}_n\right] \overset{(c)}{\leq} \mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N} \alpha_n\beta_n^\star\right] \qquad (4.90)
\end{aligned}$$

where (a) and (c) follow by definition and (b) follows since $\log(1+x) \leq x$ for $x \geq 0$. By recalling the definition of $\boldsymbol{\beta}^\star$, relation (4.90) implies

$$\begin{aligned} \mathrm{E}\left[\frac{1}{N\gamma}\sum_{n=1}^{N}\log\left(1+N\gamma\alpha_n 1\{n^\star=n\}\right)\right] &\leq \mathrm{E}\left[\frac{1}{N\gamma}\sum_{n=1}^{N}\log\left(1+\alpha_n\widehat{\beta}_n\right)\right] \\ &\leq \mathrm{E}\left[\sum_{n=1}^{N}\alpha_n 1\{n^\star=n\}\right] \end{aligned} \tag{4.91}$$

and by letting $\gamma \to 0$, equality (4.84) follows.

4.E Proof of Theorem 7

Relation (4.88) and definition (4.84) imply $\dot{C}_{1,N}^{(k)}(0) = s_N^{(k)}$ for every delay N, while in the limit for large N equation (4.71) holds for every $\gamma \geq 0$. By putting together those facts, we have

$$\lim_{N\to\infty} s_N^{(k)} = \lim_{N\to\infty} \dot{C}_{1,N}^{(k)}(0) = \dot{C}_1^{(\mathrm{erg})}(0) \tag{4.92}$$

The single-user ergodic capacity is given by the waterfilling formula (4.63) parameterized by the Langangian multiplier λ satisfying (4.65). Hence, we have

$$\begin{aligned} \lim_{\gamma\to 0}\frac{\mathrm{d}C_1^{(\mathrm{erg})}(\gamma)}{\mathrm{d}\gamma} &= \lim_{\gamma\to 0}\left.\frac{\frac{\mathrm{dE}\left[\left(\log\frac{\alpha}{\lambda}\right)^+\right]}{\mathrm{d}\lambda}}{\frac{\mathrm{dE}\left[\left(\frac{1}{\lambda}-\frac{1}{\alpha}\right)^+\right]}{\mathrm{d}\lambda}}\right|_{\gamma=\mathrm{E}[(1/\lambda-1/\alpha)^+]} \\ &= \lim_{\lambda\to\sup\{\alpha\}} \lambda\, 1\{\lambda\leq\sup\{\alpha\}\} = \sup\{\alpha\} \end{aligned} \tag{4.93}$$

which concludes the proof. □

Chapter 5

Wideband performance

We analyze the performance of the system introduced in the previous chapter (a multi-user block-fading channel with causal feedback and delay constraint) in a regime where the number of transmitted data bits per received dimension is small, commonly referred to as "wideband" or "low-power" regime. We use two analysis tools: the minimum transmit energy per bit required for reliable communication and the wideband slope of the spectral efficiency curve vs. energy per bit. We show that the "one-shot" policy, derived in the previous chapter in the contest of capacity per unit energy, is wideband optimal for every number of users and every delay, i.e., it achieves both the minimum energy per bit and the wideband slope.

5.1 Introduction

In the previous chapter we characterized the long-term average capacity region and long-term average capacity region per unit energy of a Gaussian block-fading multi-user channel where users know the channel causally and are constraint to send a codeword within a fame of N slots. We showed that for $N = 1$ the optimal policy coincides with constant power allocation [10], while, as N increases, it tends to the ergodic policy found by Tse and Hanly [13]. In particular, the policy achieving long-term average capacity per unit energy is "one-shot": all the energy is concentrated on one slot of the available N whose selection is fading-dependent. Since such slot must be chosen on the basis of causal feedback, the transmitter cannot simply choose the most favorable slot in the frame (optimal non-causal policy). Rather, the solution is obtained through Dynamic Programming and has the structure of a comparison of the instantaneous fading amplitude with a

decreasing threshold [92]. Moreover, since capacity per unit energy is the inverse of the minimum transmit energy required for reliable communication, we have concluded that the "one-shot" policy is optimal in the infinite bandwidth regime [14, 85, 99].

As shown recently in [85, 99], the minimum energy per bit, on which traditionally information theoretic analysis of wideband regime have focused, fails to capture the fundamental power-bandwidth tradeoff. To study that tradeoff is necessary to analyze not only the minimum energy per bit but also the "wideband slope" of the spectral efficiency curve vs. energy per bit at the point of minimum energy per bit. In the multi-user setting, every user rate must be characterized in terms of its wideband slope, hence the notion of wideband slope region emerges to replace the simple wideband slope notion used in the single-user case. The analysis of the wideband slope region can be done by extending of the single-user approach of [85, 99], as we showed in [30]. As an application of the methodology developed in [30] to case of fading multiuser channel with delay and energy constraint, we show that a "one-shot" power policy is optimal in terms of minimum energy per bit (see Theorem 5 in Chapter 4) as well as of the wideband slope. Moreover, we show that TDMA is not as good as superposition with optimum joint decoding in the wideband regime, thus disproving the common believe that "the penalty TDMA suffers with respect to an optimal system vanishes in the low-power/wideband regime".

The paper is organized as follows: Section 5.2 briefly revises the notions of optimality in the infinity bandwidth regime and in the wideband regime for the single-user case [99] and extends those results to the multi-user case; Section 5.3 deals with the wideband analysis of the causal multi-user system and Section 5.4 of the multi-user non-causal system; Section 5.5 gives numerical examples and Section 5.6 concludes the chapter with some final practical remarks. All proofs are collected in the Appendices.

Our publications related to this chapter are:
[30] S.Verdú and G.Caire and D.Tuninetti, *"Is TDMA optimal in the low power regime?"*, in Proceedings of the 2002 IEEE International Symposium on Information Theory (ISIT2002), Lausanne (CH), June 2002;
[31] D.Tuninetti and G.Caire and S.Verdú, *"Fading multi-access channels in the wideband regime: the impact of delay constraint"*, in Proceedings of the 2002 IEEE International Symposium on Information Theory (ISIT2002), Lausanne (CH), June 2002;
[32] D.Tuninetti and G.Caire and S.Verdú, *"The impact of delay constraint and causal feedback on the wideband performance of block-fading multiple-access channels"*, submitted to IEEE Transactions on Information Theory, February 2002.

5.2 Wideband analysis

The single-user case. In [14], Verdú formulates the problem of finding the capacity per unit cost for the class of memoryless stationary channels. Given a nonnegative cost function $b(\cdot)$ defined on the channel input alphabet X and given a maximum cost Γ, then the Shannon capacity is given by

$$C(\Gamma) = \sup_{p_X : \mathrm{E}[b(X)] \leq \Gamma} I(X;Y) \tag{5.1}$$

where the supremum of the mutual information $I(X;Y)$ is over all the probability distribution functions p_X that satisfy the average cost constraint $\mathrm{E}[b(X)] \leq \Gamma$. The capacity $C(\Gamma)$ represents the number of bits per channel use that can be reliably transmitted through the channel with average cost Γ. Hence, the minimum of $\Gamma/C(\Gamma)$ is the minimum cost incurred for the reliable transmission of one bit and its reciprocal has the meaning of capacity per unit cost. Therefore, the capacity per unit cost U is given by

$$U = \sup_{\Gamma > 0} \frac{C(\Gamma)}{\Gamma} = \frac{\mathrm{d}}{\mathrm{d}\Gamma}\{C(\Gamma)\}\Big|_{\Gamma=0} \tag{5.2}$$

where the last equality follows since $C(\Gamma)$ is a non-decreasing concave function for $\Gamma \geq 0$. [1] In order to achieve (5.2) the use of optimal input distribution p_X achieving the supremum in (5.1) is not mandatory. In [14], it is proved that it is enough to restrict attention to "binary" codes, i.e., codes that use only one non-zero-cost symbol in addition to the zero-cost symbol. Based on this observation, we say that an input distribution is optimal in terms of capacity per unit cost if the first derivative of the corresponding mutual information at $\Gamma = 0$ achieves the first derivative of capacity at $\Gamma = 0$.

From a practical point of view, it is sensible to consider as cost the energy at the transmitter. In this case the capacity per unit cost is the maximum number of bits transmitted per unit joule and its inverse is the minimum transmit energy per bit required for reliable communication.

To fix ideas, consider a single-user AWGN channel of bandwidth W and noise power spectral density N_0. This channel is used to transmit codewords of duration T made up of L complex symbols, in the assumption $L = WT \gg 1$. Given an average transmit energy per channel symbol E_s, the capacity, measured in bits per channel use, is $C = \log(1+\gamma)$ where $\gamma = E_s/N_0$ is the SNR. The minimum transmit energy per bit, normalized with respect to the noise power spectral density N_0, is easily obtainable as the inverse of the

[1] Actually, the main contribution of [14] resides in an alternate way to compute U for channels whose input alphabets contain a symbol of zero cost. The alternate expression involves the maximization over the input alphabet of the Kullback-Leibler divergence between conditional probabilities. That formulation has the advantage that does not require the computation/knowledge of the Shannon capacity function $C(\Gamma)$.

first derivative of the capacity at $\gamma = 0$ and is given by $(E_b/N_0)_{\min} = \log(2)$. In general, the average transmit energy per bit E_b is related to the average transmit energy per channel symbol E_s by $E_s = C\,E_b$, assuming that the transmitter uses a capacity achieving code of rate C. Then, the system of equations

$$\begin{cases} \frac{E_s}{N_0} = \mathsf{C}\,\frac{E_b}{N_0} \\ \mathsf{C} = \log_2\left(1 + \frac{E_s}{N_0}\right) \end{cases} \tag{5.3}$$

define implicitly the so-called *spectral efficiency* function $\mathsf{C}(E_b/N_0)$. Note that the function $\mathsf{C}(E_b/N_0)$ is positive only for $E_b/N_0 > (E_b/N_0)_{\min}$, while $C(\gamma)$ is defined for all $\gamma \geq 0$. In a right interval of $(E_b/N_0)_{\min}$ the spectral efficiency function can be approximated as follows

$$\mathsf{C}\left(\frac{E_b}{N_0}\right) = \mathcal{S}_0 \cdot \left(\frac{\frac{E_b}{N_0}}{\left(\frac{E_b}{N_0}\right)_{\min}} - 1\right) + o\left(\frac{E_b}{N_0}\right) \tag{5.4}$$

as $E_b/N_0 \to (E_b/N_0)_{\min}$. In our example, $\mathcal{S}_0 = 2$. The constant $\mathcal{S}_0$ is called in [99] the *wideband slope* of the spectral efficiency curve. Notice that capacity $C(\gamma)$ can also be achieved by binary antipodal signaling if $\gamma \ll 1$. The equivalence between the two input distribution is proved by showing that the ratio of mutual information to capacity approaches one for $\gamma \to 0$ or, in other words, that mutual information with binary input attains the first derivative of capacity at $\gamma = 0$, i.e., the binary input distribution achieves $(E_b/N_0)_{\min}$. Fig 5.1 reports the capacity curve $C(\gamma)$ vs. γ in linear scale, while Fig. 5.2 reports the spectral efficiency curve $\mathsf{C}(E_b/N_0)$ vs. E_b/N_0 in dB scale, for Gaussian input and binary input. From Fig. 5.2 we see that the spectral efficiency curve with binary input achieves $(E_b/N_0)_{\min} = -1.59$dB but has a wideband slope of just $\mathcal{S}_0 = 1$. This very simple example shows that the comparison of systems on the basis of their $(E_b/N_0)_{\min}$ may lead to conclusions that are not longer valid when $\mathcal{S}_0$ is considered.

In [99], the optimality of a coding scheme in the wideband regime (in terms of $\mathcal{S}_0$) is defined and studied for several input-constrained additive noise channels. Following the terminology introduced in [99], we say that $(E_b/N_0)_{\min}$ is the performance measure of the channel in the *infinite bandwidth regime* while $\mathcal{S}_0$ is the performance measure in the *wideband regime*. The terminology "infinite bandwidth regime" refers to a system operating with infinite bandwidth, with zero power, and hence achieving zero spectral efficiency. "Wideband regime" refers to the case with low but finite power, large but not infinite bandwidth and hence small but non-zero spectral efficiency. It is precisely the wideband regime that is of practical importance since the 3G and 4G wireless systems, as well as ad-hoc networks and sensor networks, will be operating in that regime.

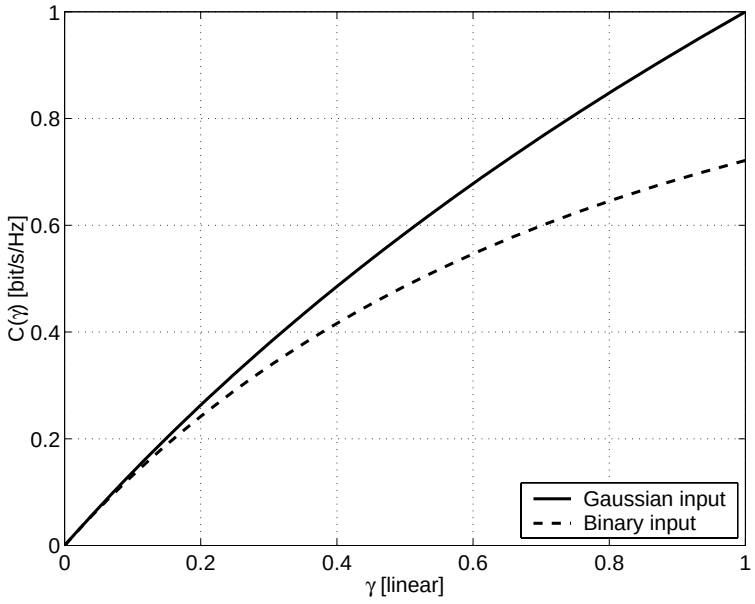

Figure 5.1: Capacity vs. γ (linear) for the single-user AWGN channel: Gaussain and binary input.

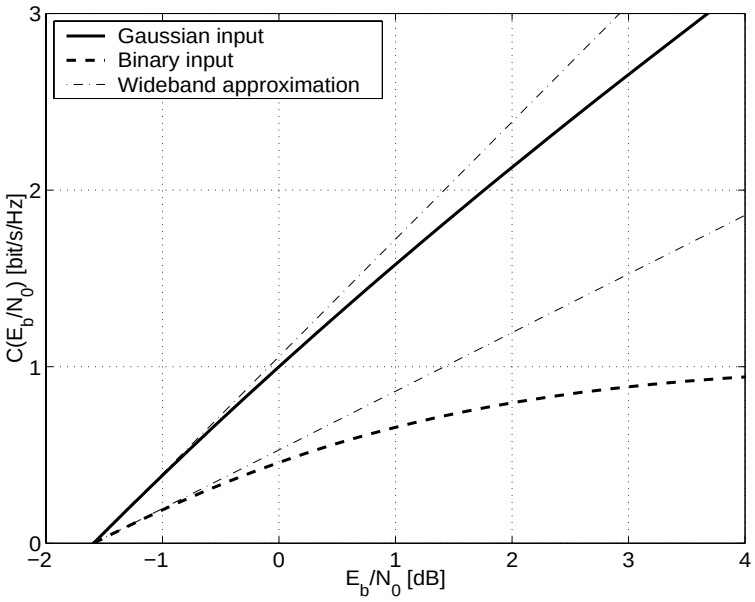

Figure 5.2: Spectral efficiency vs. E_b/N_0 (dB) for the single-user AWGN channel: Gaussain and binary input.

In general, let $C(\gamma)$ be the capacity expressed in nat/dimension (or nat/s/Hz) as a function of the (transmit) SNR γ and let E_b/N_0 be the transmit energy per bit, then the spectral efficiency C expressed as a function of E_b/N_0 is given implicitly by the parametric equation

$$\begin{cases} \frac{E_b}{N_0} &= \frac{\gamma}{C(\gamma)} \log 2 \\ \mathsf{C} &= C(\gamma) \frac{1}{\log 2} \end{cases} \tag{5.5}$$

The value $(E_b/N_0)_{\min}$, for which $C(E_b/N_0) > 0 \Leftrightarrow E_b/N_0 > (E_b/N_0)_{\min}$, is given by [99]

$$\left(\frac{E_b}{N_0}\right)_{\min} = \frac{\log 2}{\dot{C}(0)} \tag{5.6}$$

where $\dot{C}(0)$ indicates the first derivative of $C(\gamma)$ at $\gamma = 0$. In [99] it is show that $\dot{C}(0)$ essentially is the channel gain and depends on the transmitter channel state information only. In the wideband regime, the behavior of spectral efficiency in a (right) neighborhood of $(E_b/N_0)_{\min}$ is captured by the slope of spectral efficiency at $(E_b/N_0)_{\min}$, given by (see [99, Theorem 6])

$$\mathcal{S}_0 = \frac{2\left(\dot{C}(0)\right)^2}{-\ddot{C}(0)} \tag{5.7}$$

where $\ddot{C}(0)$ indicates the second derivative of $C(\gamma)$ at $\gamma = 0$. The definition of $\mathcal{S}_0$ as in (5.4) has two advantages: first it is invariant to the channel gain (see the detailed discussion in [99]) and second it can be given the meaning of "slope per 3dB" [99, 85], in fact

$$10\log_{10}\left(\frac{E_b}{N_0}\right) = 10\log_{10}\left(\frac{E_b}{N_0}\right)_{\min} + \frac{C}{\mathcal{S}_0} 10\log_{10}(2) + o(C)$$

as $C \to 0$.

The wideband slope $\mathcal{S}_0$ quantifies the bandwidth requirement for a given desired data rate. In fact, the data rate R_b (bit/s), the channel bandwidth W (Hz) and the energy per bit (E_b/N_0) are related via the spectral efficiency $\mathsf{C}(E_b/N_0)$ (bit/s/Hz) as follow

$$R_b = W\,\mathsf{C}\left(\frac{E_b}{N_0}\right) \tag{5.8}$$

For any $\epsilon > 0$ and $(E_b/N_0) = (E_b/N_0)_{\min}(1+\epsilon)$ we have

$$R_b \simeq W\,\mathcal{S}_0\,\epsilon \tag{5.9}$$

for small ϵ, which means that the system with higher wideband slope $\mathcal{S}_0$ requires less bandwidth for fix ϵ and R_b.

Following [99] we have:

Definition 1. A signaling strategy in a single-user system is said to be *first-order optimal* if it achieves $(E_b/N_0)_{\min}$ (fulfill the criterion of the first derivative at $\gamma = 0$) and *second-order optimal* if it achieves both $(E_b/N_0)_{\min}$ and $\mathcal{S}_0$ (fulfill the criterion of both the first and second derivative at $\gamma = 0$). $\Diamond$

The multi-user case. Also for the multi-user case, the analysis of infinite-bandwidth regime may lead to conclusion that are not valid in the wideband regime. In a multiple-access channel, the individual user energy per bit over the noise power N_0 are defined by $E_k/N_0 \stackrel{\Delta}{=} \log(2)\,\gamma_k/R_k$, where γ_k is the transmit SNR and R_k is the rate in nat/s/Hz of user k. In general, $R_k = R_k(\gamma_1, \cdots, \gamma_K)$ such that the point $(R_1, \cdots, R_K)$ is achievable. Without loss of generality we can consider only rate K-tuples on the boundary surface of the K-user capacity region. We indicate by $\mathcal{S}_0^{(k)}$ the k-th user single-user slope and by $\mathcal{S}_k$ the slope of user k in the multi-user case. Since the presence of an interferer can not increase the rate, we have that $\mathcal{S}_k \in [0, \mathcal{S}_0^{(k)}]$.

In order to fix ideas, consider the 2-user AWGN system. The capacity region is the polymatroid [6, 7]

$$\begin{cases} R_1 & \leq & \log(1+\gamma_1) \\ R_2 & \leq & \log(1+\gamma_2) \\ R_1+R_2 & \leq & \log(1+\gamma_1+\gamma_2) \end{cases} \tag{5.10}$$

As well known, the points of the boundary surface (often called "dominant face" [67]) of (5.10) are achieved by superposition and optimum joint decoding at the receiver. In contrast, TDMA achieves

$$\bigcup_{\tau\in[0,1]} \begin{cases} R_1 & \leq & \tau \log\left(1+\frac{\gamma_1}{\tau}\right) \\ R_2 & \leq & (1-\tau)\log\left(1+\frac{\gamma_2}{1-\tau}\right) \end{cases} \tag{5.11}$$

Figs. 5.3 and 5.4 show the capacity region and the TDMA region for $\gamma_1 =$ 10dB and $\gamma_2 =$ 13dB and $\gamma_1 = -10$dB and $\gamma_2 = -7$dB respectively. We see that the boundary of the TDMA region touches the boundary of the capacity region for $\tau = 0$ (only user 2 active), $\tau = \frac{\gamma_1}{\gamma_1+\gamma_2}$ and $\tau = 1$ (only user 1 active). If we lower down the values of the SNR's we see that both regions shrink and that the TDMA region occupies an increasing larger fraction of the capacity region. This can be formalized by showing that the TDMA region converges to the Cartesian product of the single-user capacity regions in the following sense

$$\lim_{\gamma_i\to 0} \frac{\tau\log\left(1+\frac{\gamma_1}{\tau}\right)}{\log(1+\gamma_1)} + \frac{(1-\tau)\log\left(1+\frac{\gamma_2}{1-\tau}\right)}{\log(1+\gamma_2)} = 2 \tag{5.12}$$

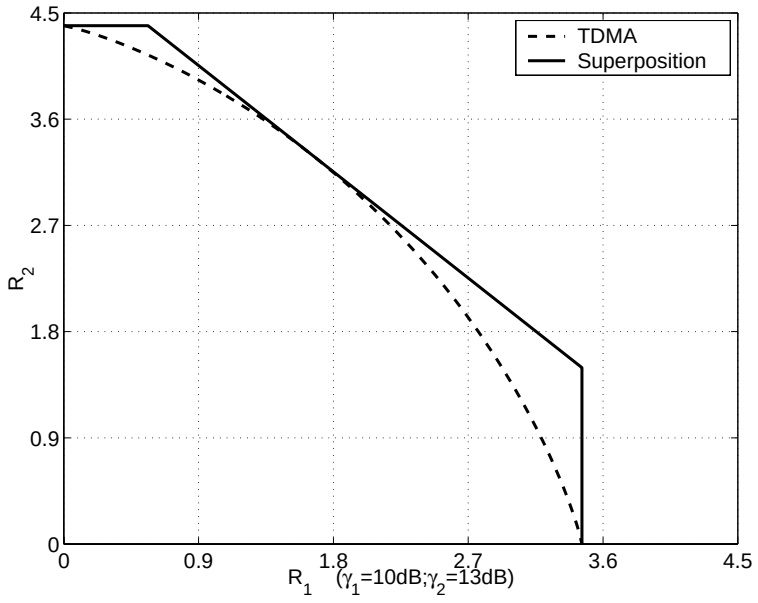

Figure 5.3: Capacity region for the 2-user AWGN channel (high SNR regime)

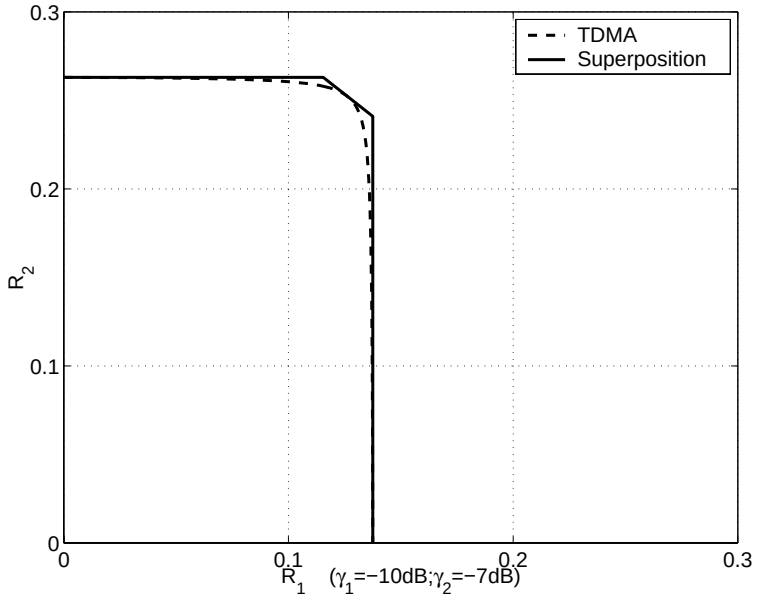

Figure 5.4: Capacity region for the 2-user AWGN channel (low SNR regime)

We are now interested in evaluating the minimum energy per bit and the slope region for the 2-user channel. By applying (5.6) and (5.7) to the per-user-rate of the TDMA system (5.11) and by considering $\tau \in [0,1]$ a fix parameter (non dependent on (γ_1, γ_2)) we get

$$\left(\frac{E_1}{N_0}\right)_{\min} = \left(\frac{E_2}{N_0}\right)_{\min} = \log(2) \tag{5.13}$$

since the first derivatives at zero SNR are

$$\frac{\mathrm{d}}{\mathrm{d}\gamma_1}\left\{\tau \log\left(1+\frac{\gamma_1}{\tau}\right)\right\}\Big|_{\gamma_1=0} = \frac{\mathrm{d}}{\mathrm{d}\gamma_2}\left\{(1-\tau)\log\left(1+\frac{\gamma_2}{1-\tau}\right)\right\}\Big|_{\gamma_2=0} = 1 \tag{5.14}$$

Since the convergence of the limits (5.14) is uniform over τ, we conclude that the result holds even for τ dependent on (γ_1, γ_2). Since TDMA achieves a subset of the capacity region and in the low-power regime achieves the same minimum energy per bit of the interference-free channel, we are tempted to assert that the advantages of superposition over TDMA vanish in the low-power regime. In that case, the increase in receiver complexity required to achieve rate in the capacity region outside the TDMA region would be hardly justifiable unless other factors come into the play.

As for the single-user case, the minimum energy per bit on its own is not "sufficient" to compare different signaling strategies in the wideband regime. Whereas the capacity region supplies the tradeoff of rates for fixed powers, we can define a corresponding "slope region" that gives the tradeoff of individual user slopes for a fixed ratio at which the individual rates vanish. Although formula (5.7) applies to the single-user channel, it turns out to be sufficient for the analysis of the multi-user case.

For the TDMA system we have $\mathcal{S}_1 = 2\tau$ and $\mathcal{S}_1 = 2(1-\tau)$ since the second-order derivatives, to be used in (5.7), of the per-user-rates are

$$\frac{\mathrm{d}^2}{\mathrm{d}\gamma_1^2}\left\{\tau \log\left(1+\frac{\gamma_1}{\tau}\right)\right\}\Big|_{\gamma_1=0} = \frac{1}{\tau}$$
$$\frac{\mathrm{d}^2}{\mathrm{d}\gamma_2^2}\left\{(1-\tau)\log\left(1+\frac{\gamma_2}{1-\tau}\right)\right\}\Big|_{\gamma_2=0} = \frac{1}{1-\tau}$$

By letting τ varying in $[0,1]$ and by recalling that $\mathcal{S}_0 = 2$, we get that the slope region for TDMA is

$$\{(\mathcal{S}_1, \mathcal{S}_2) : 0 \le \mathcal{S}_1 + \mathcal{S}_2 \le \mathcal{S}_0\} \tag{5.15}$$

The analysis of the system with superposition is more involved. Without loss of generality we consider rate-couples on the dominant face of the capacity region (since those are the only points in the capacity region for which is not possible to further increase the rate of one user without having to decrease

the rate of the other users in order to be inside the capacity region). Let $\lambda \in [0,1]$ and $\theta = \gamma_1/\gamma_2$. Every point of the dominant face of (5.10) can be written as

$$R_1 = R_1(\gamma_1) = \lambda \log(1+\gamma_1) + (1-\lambda)\log\left(1+\frac{\gamma_1}{1+\gamma_1/\theta}\right) \quad (5.16)$$

$$R_2 = R_1(\gamma_2) = \lambda \log\left(1+\frac{\gamma_2}{1+\gamma_2\,\theta}\right) + (1-\lambda)\log(1+\gamma_2) \quad (5.17)$$

By taking the first and second derivative at zero SNR and by applying (5.7), we get

$$\begin{cases} \mathcal{S}_1 = 2\frac{1}{1+2(1-\lambda)/\theta} \\ \mathcal{S}_2 = 2\frac{1}{1+2\lambda\theta} \end{cases} \quad (5.18)$$

By solving (5.18) for λ by recalling that $\mathcal{S}_0 = 2$, we obtain

$$\frac{1}{\theta}\left(\frac{1}{\mathcal{S}_1}-\frac{1}{\mathcal{S}_0}\right)+\theta\left(\frac{1}{\mathcal{S}_2}-\frac{1}{\mathcal{S}_0}\right)=1 \quad (5.19)$$

this curve, for a given θ, is what we call the *slope trade-off boundary*. The slope region is

$$\bigcup_{\theta\geq 0}\left\{(\mathcal{S}_1,\mathcal{S}_2) : 0 \leq \mathcal{S}_k \leq \mathcal{S}_0, \quad \frac{1}{\theta}\left(\frac{1}{\mathcal{S}_1}-\frac{1}{\mathcal{S}_0}\right)+\theta\left(\frac{1}{\mathcal{S}_2}-\frac{1}{\mathcal{S}_0}\right)=1\right\} \quad (5.20)$$

were the condition $\mathcal{S}_k \leq \mathcal{S}_0$ follows from the fact that the existence of an interferer cannot improve the rate. Fig. 5.5 shows the slope region, for $\theta = 1$ and $\theta = 10$, and the achievable slope region with TDMA. It is clear that TDMA is suboptimal with respect to superposition when slopes are considered, thus dispelling the longstanding misconception that TDMA is optimal in the wideband regime. Note that both users can achieve slope arbitrarily close to the single-user slope provided that they use superposition, optimum decoding and their powers are sufficiently unbalanced (either $\theta \ll 1$ or $\theta \gg 1$). In our paper [30] we did the same kind of analysis for the broadcast channel and we got to the same conclusions: TDMA is not wideband optimal.

In general for the K-user channel, we fix a vector $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K) \in \mathbb{R}_+^K$ and we let the user SNR's vanish with fixed ratio $\gamma_k/\gamma_j = \theta_k/\theta_j$, for all $i,j \in \{1,\cdots,K\}$. The fact that, from the general theory of capacity per unit cost of additive channels, the capacity region per unit energy is an hyper-rectangle implies that for vanishing SNR the user rates are directly proportional, through the capacity per unit energy, to their SNR's. Hence,

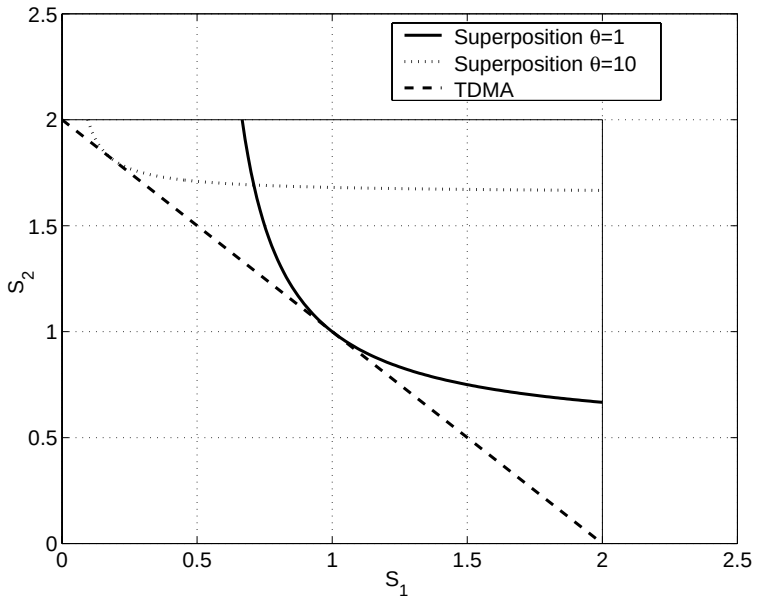

Figure 5.5: Slope region for the 2-user AWGN channel ($\mathcal{S}_0 = 2$).

imposing SNR ratios is equivalent to fix rate ratios. For the channel model introduced in Chapter 4, since $R_k \approx s_N^{(k)} \gamma_k$, we have

$$\frac{\gamma_k}{\gamma_j} = \frac{\theta_k}{\theta_j} \quad \Rightarrow \quad \frac{R_k}{R_j} = \frac{s_N^{(k)}\,\theta_k}{s_N^{(j)}\,\theta_j} \tag{5.21}$$

The rate for user k as function of γ_k, and of the parameter $\boldsymbol{\theta}$, is given by

$$R_k = R_k(\gamma_k) = R_k\left(\frac{\theta_1}{\theta_k}\gamma_k, \cdots, \frac{\theta_K}{\theta_k}\gamma_k\right) \tag{5.22}$$

Hence by taking the first and second derivative of (5.22) with respect to γ_k and by applying (5.7), we express the wideband slope of the k-th user as

$$\mathcal{S}_k = \frac{2\left(\sum_{j=1}^{K} \theta_j \cdot \partial_j R_k(0,\cdots,0)\right)^2}{-\sum_{j=1}^{K}\sum_{m=1}^{K} \theta_j\theta_m \cdot \partial_{j,m} R_k(0,\cdots,0)} \tag{5.23}$$

where $\partial_j R_k(0,\cdots,0)$ is the shorthand notation for

$$\partial_j R_k(0,\cdots,0) = \lim_{\boldsymbol{\gamma}\to\mathbf{0}} \frac{\partial R_k(\gamma_1,\cdots,\gamma_K)}{\partial \gamma_j} \tag{5.24}$$

and $\partial_{j,m} R_k(0,\cdots,0)$ for

$$\partial_{j,m} R_k(0,\cdots,0) = \lim_{\boldsymbol{\gamma}\to\mathbf{0}} \frac{\partial^2 R_k(\gamma_1,\cdots,\gamma_K)}{\partial \gamma_j\,\partial \gamma_m} \tag{5.25}$$

Notice that the wideband slope of user k is completely characterized by the gradient and the Hessian matrix of the rate function $R_k = R_k(\gamma_1, \cdots, \gamma_K)$ computed for $\boldsymbol{\gamma} = \mathbf{0}$. By letting the rate vector $(R_1, \cdots, R_K)$ span the whole boundary of the capacity region for a fix $\boldsymbol{\theta}$, the corresponding slopes (5.23) define a curve in the K-dimensional space we shall refer to as *slope tradeoff boundary*.

Definition 2. We say that a signaling strategy in a K-user system is *first-order optimal* if it achieves $(E_k/N_0)_{\min}$ for all the users and *second-order optimal* if it achieves both $(E_k/N_0)_{\min}$ and the slope tradeoff boundary for every $\boldsymbol{\theta}$. $\Diamond$

In the following, we characterize the slope region corresponding to the long term average capacity region $C_{K,N}(\boldsymbol{\gamma})$ and determine whether the policy $\boldsymbol{\beta}^\star$ (which is first order optimal) is also second order optimal, i.e., for every $\boldsymbol{\theta}$ it achieves the slope tradeoff boundary.

5.3 Second-order optimality of $\beta^\star$ in the causal system

The single-user case. We deal first with the single user case, i.e., $K = 1$. For simplicity we drop the user index.

We indicate the single-user long-term average capacity given in Theorem 2, with a slight abuse of notation, as

$$C_{1,N}(\gamma) = \frac{S_N(N\gamma)}{N} \tag{5.26}$$

where, for simplicity we re-write recursion (4.13) omitting the irrelevant parameter μ as follow

$$S_n(P) \;=\; \mathrm{E}\left[\max_{p\in[0,P]} \log\left(1+\alpha\, p\right) + S_{n-1}(P-p)\right] \tag{5.27}$$

for $n = 1, \cdots, N$ with initial condition $S_0(P) = 0$. When user k is considered, the mean value in (5.27) is computed with respect to $\alpha \sim F_\alpha^{(k)}(x)$ and the SNR in (5.26) is $\gamma = \gamma_k$.

Even if we cannot give a closed form expression for $S_N(P)$ and $\widehat{\boldsymbol{\beta}}$, the wideband characterization of the single-user long-term average capacity and the second-order optimality of the one-shot policy $\beta^\star$ are given by the following:

Theorem 8. $(E_b/N_0)_{\min}$ and $\mathcal{S}_0$ for the single-user block fading channel

with causal transmitter CSI and delay N are given by

$$\left(\frac{E_b}{N_0}\right)_{\min} = \frac{\log 2}{\dot{S}_N(0)} \tag{5.28}$$

$$\mathcal{S}_0 = \frac{2\left(\dot{S}_N(0)\right)^2}{-N\,\ddot{S}_N(0)} \tag{5.29}$$

where $\dot{S}_N(0)$ and $\ddot{S}_N(0)$ are, respectively, the first and the second derivative of $S_N(P)$ in (5.27) at $P=0$. The first derivative is given by

$$\dot{S}_N(0) = s_N \tag{5.30}$$

where s_N is given (for the k-th user) by the recursion (4.23), i.e., $s_n = \mathrm{E}[\max\{s_{n-1}, \alpha\}]$, and the second derivative is given by the recursion

$$\begin{aligned} -\ddot{S}_n(0) &= \mathrm{E}\left[\alpha^2 | \alpha \geq s_{n-1}\right] \Pr(\alpha \geq s_{n-1}) \\ &\quad -\ddot{S}_{n-1}(0) \Pr(\alpha < s_{n-1}) \end{aligned} \tag{5.31}$$

for $n = 1, \ldots, N$, with $\ddot{S}_n(0) = 0$.

Furthermore, the one-shot power allocation policy $\boldsymbol{\beta}^\star$ achieves $(E_b/N_0)_{\min}$ and slope $\mathcal{S}_0$, i.e., it is first and second-order optimal.

Proof. Expressions (5.28) and (5.29) follow by using (5.26) in (5.6) and (5.7). Statement (5.30) follows immediately for Theorem 5 and Theorem 6, i.e., $\dot{C}_{1,N}(0) = \dot{S}_N(0)$ from (5.26) and $s_N = \dot{C}_{1,N}(0)$ from the proof of Theorem 5. The proof of statement (5.31) and of the second-order optimality of $\boldsymbol{\beta}^\star$ can be found in Appendix 5.A. □

TDMA achievable slope region. Before carrying on the characterization of the slope region for the multi-user case, we investigate the achievable slope region of TDMA in conjunction with power policy $\boldsymbol{\beta}^\star$. In Section 4.4 we have shown that the one-shot power allocation $\boldsymbol{\beta}^\star$ (in conjunction with Gaussian variable-rate coding) achieves the capacity region per unit energy, i.e., achieves $(E_k/N_0)_{\min}$ for all users. Then, we conclude that the one-shot policy is first-order optimal for any number of users K. From the proof of Theorem 5 it follows that first-order optimality can be obtained either by using superposition coding or by using TDMA inside each slot. Because of the second-order optimality of $\boldsymbol{\beta}^\star$ in the single user we have:

Theorem 9. For any arbitrary ratios γ_k/γ_j, as the rates vanish, the largest achievable slope region under TDMA is

$$\left\{ \mathcal{S}_{k,\mathrm{tdma}} \geq 0 \quad \forall k = 1, \cdots, K \;:\; \sum_{k=1}^{K} \frac{\mathcal{S}_{k,\mathrm{tdma}}}{\mathcal{S}_0^{(k)}} \leq 1 \right\} \tag{5.32}$$

and this region is achieved by $\boldsymbol{\beta}^\star$.

Proof. For $\boldsymbol{\tau} = (\tau_1, \cdots, \tau_K) \in \mathbb{R}_+^K$ such that $\sum_{k=1}^K \tau_k = 1$ the maximum

achievable rates under TDMA are $R_k = (\tau_k/N) \cdot S_N(\gamma_k N/\tau_k)$. By straightforward application of (5.23), we have $\mathcal{S}_{k,\text{tdma}} = \tau_k \, \mathcal{S}_0^{(k)}$ hence, by considering the union over all possible choice of $\boldsymbol{\tau}$, we get (5.32). □

The multi-user case. The optimal slope region under the causal power constraint is given by the following:

Theorem 10. Fix a vector $\boldsymbol{\theta} \in \mathbb{R}_+^K$. For vanishing user rates while keeping fixed the ratios $\gamma_k/\gamma_j = \theta_k/\theta_j$, the optimal slope region is given by the parametric form

$$\bigcup_{\boldsymbol{\lambda}} \left\{ \mathcal{S}_k \geq 0 \quad \forall k = 1, \cdots, K \; : \; \mathcal{S}_k \leq \frac{\mathcal{S}_0^{(k)}}{1 + \sum_{\boldsymbol{\pi}} \lambda_{\boldsymbol{\pi}} \sum_{j < \pi^{-1}(k)} \frac{\theta_{\pi_j}}{\theta_k} \mathcal{K}_{k,\pi_j}} \right\} \tag{5.33}$$

for

$$\mathcal{K}_{k,j} = \frac{2 \sum_{n=1}^{N} \mathrm{E}[\alpha_{k,n} 1\{n_k^\star = n\}] \, \mathrm{E}[\alpha_{j,n} 1\{n_j^\star = n\}]}{\sum_{n=1}^{N} \mathrm{E}\left[\alpha_{k,n}^2 1\{n_k^\star = n\}\right]} \tag{5.34}$$

where $\sum_{\boldsymbol{\pi}}$ denotes the sum over all permutations of $\{1, \ldots, K\}$, where $\boldsymbol{\lambda} = \{\lambda_{\boldsymbol{\pi}}\}$ are $K!$ nonnegative "time-sharing" coefficients (indexed by the permutations $\boldsymbol{\pi}$) such that $\sum_{\boldsymbol{\pi}} \lambda_{\boldsymbol{\pi}} = 1$ and where $n_k^\star$ is given in (4.25).

Furthermore, the one-shot policy $\boldsymbol{\beta}^\star$ achieves the slope tradeoff boundary, i.e., it is second-order optimal in the multi-user case.

Proof. See Appendix 5.B. □

5.4 Second-order optimality of $\beta^\star$ in the non-causal system

The single-user case. If we allow the input to depend on the whole CSI $\mathcal{S}_N$ in a non-causal way the optimal allocation policy is given in (4.29) and the corresponding capacity per unit energy in (4.31). It follow easily that:

Theorem 11. $(E_b/N_0)_{\min}^{(\text{nc})}$ and $\mathcal{S}_0^{(\text{nc})}$ for the single-user block fading channel with non-causal transmitter CSI, delay N and continuous fading distribution [2] are given by

$$\begin{aligned} \left(\frac{E_b}{N_0}\right)_{\min}^{(\text{nc})} &= \frac{\log 2}{\mathrm{E}[\max\{\alpha_1, \cdots, \alpha_N\}]} \\ \mathcal{S}_0^{(\text{nc})} &= \frac{2 \left(\mathrm{E}[\max\{\alpha_1, \cdots, \alpha_N\}]\right)^2}{N \, \mathrm{E}[(\max\{\alpha_1, \cdots, \alpha_N\})^2]} \end{aligned} \tag{5.35}$$

[2] In case the fading distribution is not-continuous, the second derivative of capacity at $\gamma = 0$ is given by

$$-\frac{1}{N} \ddot{C}_{1,N}^{(\text{nc})}(0) = \mathrm{E}\left[\frac{(\max\{\alpha_1, \cdots, \alpha_N\})^2}{\sum_{\ell=1}^{M} 1\{\alpha_\ell = \max\{\alpha_1, \cdots, \alpha_N\}\}}\right]$$

Furthermore, the one-shot power allocation policy $\boldsymbol{\beta}^{\star\,(\mathrm{nc})}$ achieves both $(E_b/N_0)_{\min}^{(\mathrm{nc})}$ and $\mathcal{S}_0^{(\mathrm{nc})}$.
Proof. See Appendix 5.C. □

TDMA achievable slope region. In a multiuser scenario, as TDMA is concerned, because of second-order optimality of $\boldsymbol{\beta}^{\star\,(\mathrm{nc})}$ in the single-user case, we have:
Theorem 12. For any arbitrary SNR ratios γ_k/γ_j, as the rates vanish, the largest achievable slope region under TDMA is given by

$$\left\{ \mathcal{S}_{k,\mathrm{tdma}}^{(\mathrm{nc})} \geq 0 \quad \forall k = 1, \cdots, K \;:\; \sum_{k=1}^{K} \frac{\mathcal{S}_{k,\mathrm{tdma}}^{(\mathrm{nc})}}{\mathcal{S}_0^{(k)\,(\mathrm{nc})}} \leq 1 \right\} \tag{5.36}$$

and this region is achieved by $\boldsymbol{\beta}^{\star\,(\mathrm{nc})}$.
Proof. As for Theorem 9. □

The multi-user case. The optimal slope region is given by the following:
Theorem 13. The non-causal one-shot policy $\boldsymbol{\beta}^{\star\,(\mathrm{nc})}$ in conjunction with joint decoding achieves the slope tradeoff boundary given by the expression in (5.33) but where

$$\mathcal{K}_{k,j} = \frac{2}{N} \frac{\mathrm{E}[\max\{\alpha_{k,1}, \cdots, \alpha_{k,N}\}]\ \mathrm{E}[\max\{\alpha_{j,1}, \cdots, \alpha_{j,N}\}]}{\mathrm{E}[(\max\{\alpha_{k,1}, \cdots, \alpha_{k,N}\})^2]} \tag{5.37}$$

i.e., $\boldsymbol{\beta}^{\star\,(\mathrm{nc})}$ is first and second-order optimal for any number of users K and any delay N.
Proof. See Appendix 5.D. □

5.5 Numerical examples

In order to illustrate the results of previous sections we consider the case of i.i.d. Rayleigh fading, i.e., the channel gain law is $F_\alpha(x) = 1 - \mathrm{e}^{-x}$ for $x \geq 0$ for all the users.

Comparison between causal and non-causal power policy. The one-shot policy $\boldsymbol{\beta}^\star$ is completely determined by the thresholds given by the recursion

$$s_n = s_{n-1} + \mathrm{e}^{-s_{n-1}} \tag{5.38}$$

for $n = 1, 2, \cdots$ with $s_0 = 0$. The first-order derivative of the long-term average capacity region $C_{1,N}(\gamma)$ satisfies $\dot{C}_{1,N}(0) = s_N$, while the recursion

for the second-order derivative satisfies $\ddot{C}_{1,N}(0) = N\ddot{S}_N(0)$ with $\ddot{S}_n(0)$ given by the recursion

$$-\ddot{S}_n(0) = \mathrm{e}^{-s_{n-1}}(2 + 2s_{n-1} + s_{n-1}^2) - \ddot{S}_{n-1}(0)(1 - \mathrm{e}^{-s_{n-1}}); \qquad (5.39)$$

for $n = 1, 2, \cdots$ with $\ddot{S}_0(0) = 0$. When the input is allowed to depend on the whole CSI $\mathcal{S}_N$, we have

$$\mathrm{E}[\max\{\alpha_1, \cdots, \alpha_N\}] = \sum_{n=1}^{N} \binom{N}{n} (-1)^{n+1} \frac{1}{n} \qquad (5.40)$$

$$\mathrm{E}[\max\{\alpha_1, \cdots, \alpha_N\}^2] = \sum_{n=1}^{N} \binom{N}{n} (-1)^{n+1} \frac{2!}{n^2} \qquad (5.41)$$

The first-order and the second-order derivative at $\gamma = 0$ determine the value of $(E_b/N_0)_{\min}$ and $\mathcal{S}_0$ according to (5.6) and (5.7). Figs. 5.6 and 5.7 show $(E_b/N_0)_{\min}$ and $\mathcal{S}_0$ versus the delay N and for both the causal and non-causal knowledge of the channel.

For a given delay N, the spectral efficiency curves as function of E_b/N_0 of the causal system and of the non-causal system start at different $(E_b/N_0)_{\min}$, smaller for the non-causal system, with an almost equal slope. The gain due to the causal vs. non-causal transmit channel state information is large and increasing with N, as far as $(E_b/N_0)_{\min}$ is concerned while it is almost negligible in terms of wideband slope.

Comparison between TDMA and superposition coding. For a desired user rate R_b (in bit/s) common to all users, and assuming that all users transmit with equal power, i.e., they have the same E_b/N_0 such that $(E_b/N_0)_{\mathrm{dB}} - ((E_b/N_0)_{\min})_{\mathrm{dB}} = \epsilon$, the system bandwidth is given approximately by [99]

$$W \approx \frac{R_b}{\min_k \mathcal{S}_0^{(k)} \epsilon} \qquad (5.42)$$

We quantize the bandwidth expansion required by TDMA w.r.t. superposition coding for a given delay N.

Since (5.42) is determined by the minimum slope, in order to minimize the system bandwidth we maximize the minimum slope. By the use Theorems 9 and 10 we determine the max-min slope of an equal-rate system. For equal rates $\theta_j/\theta_k = 1$ for all k, j, the denominator of (5.33) becomes

$$\begin{aligned} 1 + \mathcal{K}_0 \sum_{\boldsymbol{\pi}} \lambda_{\boldsymbol{\pi}} \sum_{j < \pi^{-1}(k)} 1 &= 1 + \mathcal{K}_0 \sum_{\boldsymbol{\pi}} \left(\lambda_{\boldsymbol{\pi}} \pi^{-1}(k) - 1\right) \\ &= 1 - \mathcal{K}_0 + \mathcal{K}_0 \sum_{\boldsymbol{\pi}} \lambda_{\boldsymbol{\pi}} \pi^{-1}(k) \qquad (5.43) \end{aligned}$$

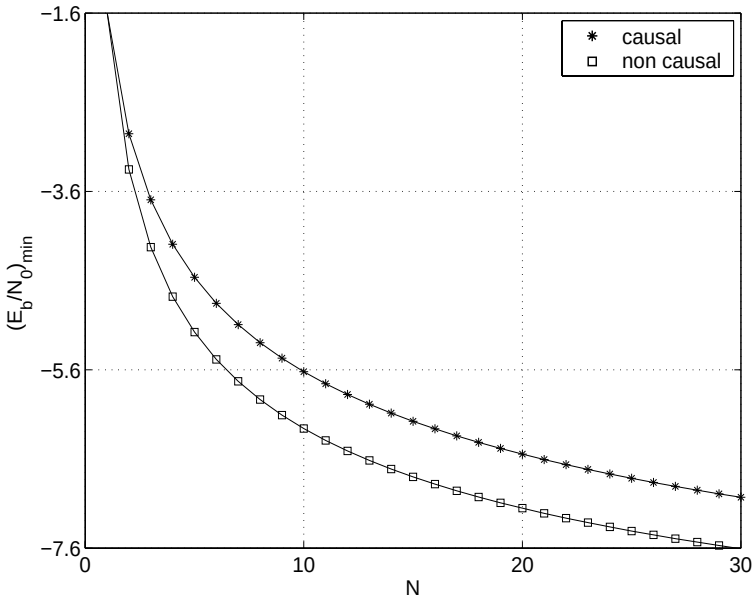

Figure 5.6: $(E_b/N_0)_{\min}$ in dB vs. N for the Rayleigh fading channel.

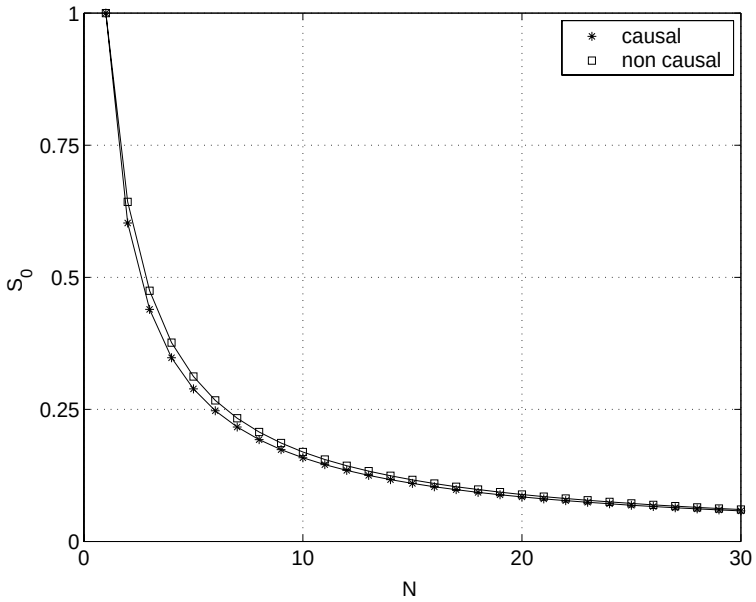

Figure 5.7: S_0 vs. N for the Rayleigh fading channel.

where, for i.i.d. fading, $\mathcal{K}_{k,j}$ in (5.34) are all equal to $\mathcal{K}_0$ for every $k, j = 1, \cdots, K$ and given by

$$\mathcal{K}_0 = 2\frac{\sum_{n=1}^{N} \left(\mathrm{E}\left[\alpha_n 1\{n^\star = n\}\right]\right)^2}{\sum_{n=1}^{N} \mathrm{E}\left[\alpha_n^2 1\{n^\star = n\}\right]} \tag{5.44}$$

As $\boldsymbol{\pi}$ varies over all $K!$ permutations, $\pi^{-1}(k)$ takes on each value $1, \ldots, K$ exactly $(K-1)!$ times. Because of symmetry, maximizing the minimum slope is achieved by letting $\mathcal{S}_0^{(k)} = \text{const.}$, i.e., $\lambda_{\boldsymbol{\pi}} = 1/K!$ for all $\boldsymbol{\pi}$. This yields to the max-min slope

$$\max_k \min \mathcal{S}_k = \frac{\mathcal{S}_0}{1 + \mathcal{K}_0(K-1)/2} \tag{5.45}$$

For TDMA, the max-min slope is obtained by letting $\tau_k = 1/K$, we have

$$\max_k \min \mathcal{S}_{k,\mathrm{tdma}} = \mathcal{S}_0/K \tag{5.46}$$

Therefore, the bandwidth expansion factor of TDMA with respect to superposition coding is given by

$$\delta = \frac{K}{1 + \mathcal{K}_0(K-1)/2} < \frac{2}{\mathcal{K}_0} \tag{5.47}$$

From (5.44) we have immediately that $\mathcal{K}_0 < 2$, i.e., TDMA is strictly wideband-suboptimal, for any non-degenerate fading distribution. Notice also that the case of equal E_b/N_0 for all users is the most favorable for TDMA [30]. As already noticed, for a very imbalanced system the bandwidth expansion factor can be much larger than (5.47).

Fig. 5.8 shows the asymptotic expansion factor $2/\mathcal{K}_0$ versus the delay N for different fading statistics and Fig. 5.9 shows the bandwidth expansion factor δ versus the number of users K and different delays for the Rayleigh fading case. For example, at delay $N = 2$ and $K = 4$ users in the Rayleigh fading case, the TDMA requires more than twice the bandwidth necessary for reliable communications by a system with superposition coding (Fig. 5.9) and asymptotically for a large population of users the TDMA requires more than three times the bandwidth (Fig. 5.8).

Notice, form Fig. 5.8 and Fig. 5.9, that by increasing either the delay N and/or the population size K the TDMA gets more and more suboptimal, since δ is increasing in both N and K. In the wideband regime, we can tradeoff system complexity versus bandwidth: if the complexity is the key issue, then TDMA is better that superposition coding but more bandwidth than the strictly necessary has to be used; on the contrary, if bandwidth is the critical parameter then superposition coding is better than TDMA, but the system complexity increases due to joint decoding and rate coordination among users.

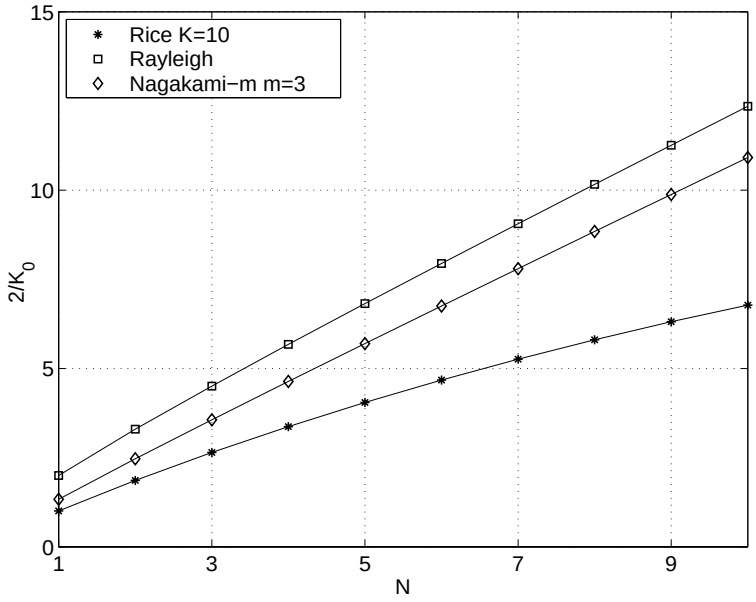

Figure 5.8: Limiting bandwidth expansion factor of TDMA over superposition coding vs. N for different fading distributions.

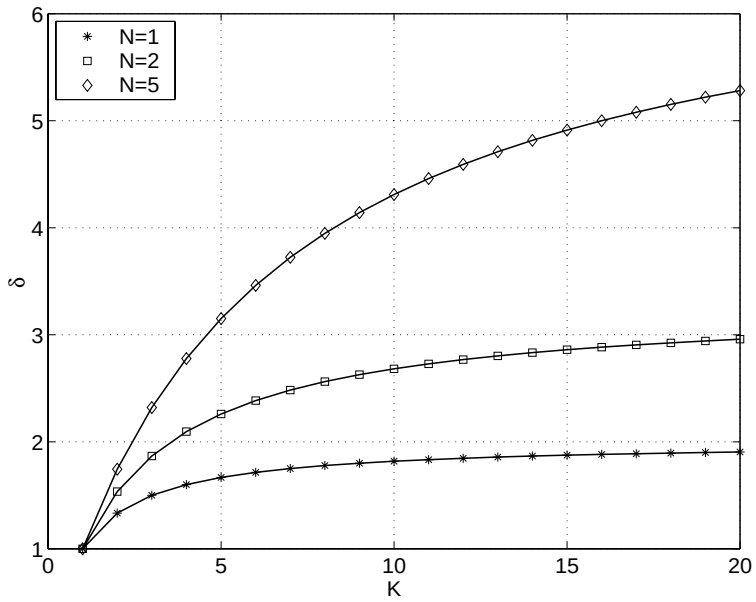

Figure 5.9: Bandwidth expansion factor of TDMA over superposition coding vs. the number of users K for the Rayleigh fading channel.

Slope region for the two user case. Next, we study in more detail the case $K = 2$. For superposition coding, by letting $\theta = \theta_1/\theta_2$, we have

$$\begin{cases} \mathcal{S}_1 & \leq & \frac{\mathcal{S}_0}{1+\mathcal{K}_0(1-\lambda)\frac{1}{\theta}} \\ \mathcal{S}_2 & \leq & \frac{\mathcal{S}_0}{1+\mathcal{K}_0\lambda\theta} \end{cases} \tag{5.48}$$

By eliminating the time-sharing parameter λ we obtain the slope region boundary as

$$\bigcup_{\theta\geq 0}\left\{\left(\frac{\mathcal{S}_0}{\mathcal{S}_1}-1\right)\theta+\left(\frac{\mathcal{S}_0}{\mathcal{S}_2}-1\right)\frac{1}{\theta}\leq\mathcal{K}_0,\quad 0\leq\mathcal{S}_k\leq\mathcal{S}_0\right\} \tag{5.49}$$

With TDMA we obtain the boundary $\mathcal{S}_{1,\text{tdma}}+\mathcal{S}_{2,\text{tdma}}=\mathcal{S}_0$.

We might wonder if for some θ TDMA achieves the same slope trade-off of superposition coding, i.e., if the two boundaries of the slope regions intersects at some point $(\mathcal{S}_1, \mathcal{S}_2)$. By substituting in (5.49) $\mathcal{S}_1 = \tau\mathcal{S}_0$ and $\mathcal{S}_2 = (1-\tau)\mathcal{S}_0$ for $\tau \in [0,1]$, we find

$$\left(\frac{1}{\tau}-1\right)\theta+\left(\frac{1}{1-\tau}-1\right)\frac{1}{\theta}=\mathcal{K}_0 \tag{5.50}$$

which yields

$$\tau=\frac{\theta}{2}\frac{2\theta+\mathcal{K}_0\pm\sqrt{\mathcal{K}_0^2-4}}{\theta^2+\mathcal{K}_0\theta+1} \tag{5.51}$$

Again, for $\mathcal{K}_0 < 2$ (non-constant fading), TDMA is strictly suboptimal, for any choice of the rate ratio θ.

Fig. 5.10 shows the 2-user optimal slope region for different rate ratios. The optimal region achievable by TDMA is shown for comparison. This figure clearly illustrates that even though TDMA achieves the capacity per unit energy, it is actually very suboptimal in the wideband regime, especially in a fading scenario.

For example, for $\theta = 1$ and $\mathcal{S}_1 = \mathcal{S}_2$, form Fig. 5.10 we see that TDMA achieves only $\frac{\mathcal{S}_k}{\mathcal{S}_0} = \frac{1}{2} = 50\%$ of the single user slope, while the optimal system with superposition coding achieves $\frac{\mathcal{S}_k}{\mathcal{S}_0} = (1+\mathcal{K}_0/2)^{-1} = 87\%$ of the single user slope.

5.6 Conclusions

In this last part of our thesis work, we have focused the analysis of the delay constraint system with causal feedback, introduced in Chapter 4, in the low spectral efficiency regime, which is where the major benefits of transmitter feedback occur. We have analyzed not only the rates achievable in the absence of bandwidth constraints (minimum energy per bit), but also the

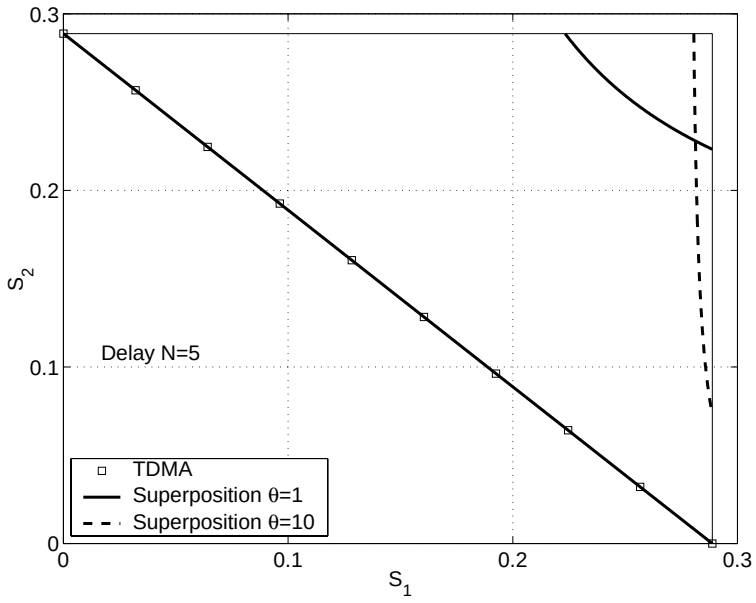

Figure 5.10: Slope region for the 2-user Rayleigh fading channel with delay $N = 5$.

bandwidth required to provide a given rate with given power in the low power regime (wideband slope). As a result of this analysis we have quantified the bandwidth expansion required by using the suboptimal TDMA strategy. The TDMA bandwidth penalty, which can be rather substantial, depends on the fading distribution, and grows with both the number of fading states N and the number of users K.

Appendix

5.A Proof of Theorem 8

Let

$$S_n(P) \quad = \quad \mathrm{E}\left[\max_{u\in[0,P]} \log\left(1+\alpha\, u\right) + S_{n-1}(P-u)\right] \tag{5.52}$$

and

$$\widehat{u}_n(\alpha,P) = \arg\max_{u\in[0,P]} \log\left(1+\alpha\, u\right) + S_{n-1}(P-u) \tag{5.53}$$

for $n = 1, \cdots, N$ and initial condition $S_0(P) = 0$. In Theorem 2 we showed that $C_{1,N}(\gamma) = S_N(N\gamma)/N$ and in Theorem 5 that $\dot{C}_{1,N}(0) = s_N$, these together imply $\dot{S}_N(0) = s_N$ which proves statement (5.30).

In order to prove statement (5.31) we need to analyze in detail expression (5.52). Because of the concavity of $S_n(P)$ (from Lemma 1 in Appendix 4.D since $S_n(P) = nC_{1,n}(P/n)$ from Theorem 2), $\widehat{u}_n(\alpha, P)$ in (5.53) can be written as

$$\widehat{u}_n(\alpha, P) = \begin{cases} 0 & \text{if } \alpha < \dot{S}_{n-1}(P) \\ P & \text{if } \dfrac{\alpha}{1+\alpha P} \geq \dot{S}_{n-1}(0) \\ u_n^* & \text{elsewhere} \end{cases} \tag{5.54}$$

with u_n^* the unique solution of

$$\frac{\alpha}{1+\alpha u_n^*} = \dot{S}_{n-1}(P - u_n^*) \tag{5.55}$$

The first and second derivative of $S_n(P)$ are given by

$$\begin{aligned} \dot{S}_n(P) &= \mathrm{E}\left[\dot{S}_{n-1}(P)\, 1\{\widehat{u}_n(\alpha, P) = 0\}\right] \\ &\quad + \mathrm{E}\left[\frac{\alpha}{1+\alpha P}\, 1\{\widehat{u}_n(\alpha, P) = P\}\right] \\ &\quad + \mathrm{E}\left[\frac{\alpha}{1+\alpha u_n^*}\, 1\{\widehat{u}_n(\alpha, P) = u_n^*\}\right] \end{aligned} \tag{5.56}$$

and by

$$\begin{aligned} -\ddot{S}_n(P) &= \mathrm{E}\left[\left(-\ddot{S}_{n-1}(P)\right)\, 1\{\widehat{u}_n(\alpha, P) = 0\}\right] \\ &\quad + \mathrm{E}\left[\left(\frac{\alpha}{1+\alpha P}\right)^2\, 1\{\widehat{u}_n(\alpha, P) = P\}\right] \\ &\quad + \mathrm{E}\left[\left(\frac{\alpha}{1+\alpha u_n^*}\right)^2 \frac{\partial u_n^*}{\partial P}\, 1\{\widehat{u}_n(\alpha, P) = u_n^*\}\right] \end{aligned} \tag{5.57}$$

Now, as $P \to 0$ we have

$$\widehat{u}_n(\alpha, P) = \begin{cases} 0 & \text{if } \alpha < \dot{S}_{n-1}(0) - (-\ddot{S}_{n-1}(0))\, P + o(P) \\ P & \text{if } \alpha \geq \dot{S}_{n-1}(0) + (\dot{S}_{n-1}(0))^2\, P + o(P) \\ P\, \dfrac{\partial u_n^*}{\partial P}\Big|_{P=0} & \text{elsewhere} \end{cases} \tag{5.58}$$

Hence, by substituting (5.58) in (5.57) and by letting $P \to 0$ we obtain

$$-\ddot{S}_n(0) = \mathrm{E}\left[\alpha^2\, 1\{\alpha \geq \dot{S}_{n-1}(0)\}\right] - \ddot{S}_{n-1}(0)\, \mathrm{E}\left[1\{\alpha < \dot{S}_{n-1}(0)\}\right] \tag{5.59}$$

which coincides with (5.31).

Next, in order to prove the second-order optimality of the policy $\boldsymbol{\beta}^\star$, we show that the rate function $C^\star_{1,N}(\gamma)$, defined as

$$C^\star_{1,N}(\gamma) = \mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N}\log\left(1+N\gamma\alpha_n\,1\{n^\star=n\}\right)\right] \tag{5.60}$$

obtained by applying $\boldsymbol{\beta}^\star$, has the first and second derivative at $\gamma=0$ equal to those of $C_{1,N}(\gamma)$.

It follows immediately that the first and second derivative of (5.60) w.r.t. γ computed for $\gamma=0$ are

$$\dot{C}^\star_{1,N}(0) = \mathrm{E}\left[\sum_{n=1}^{N}\alpha_n 1\{n^\star=n\}\right] \tag{5.61}$$

$$-\frac{1}{N}\ddot{C}^\star_{1,N}(0) = \mathrm{E}\left[\sum_{n=1}^{N}\alpha_n^2 1\{n^\star=n\}\right] \tag{5.62}$$

From the proof of Theorem 5 it follows that $\dot{C}^\star_{1,N}(0)=\dot{C}_{1,N}(0)$, i.e., $\boldsymbol{\beta}^\star$ achieves $(E_b/N_0)_{\min}$. Next we show that (5.62) is equal to $-\ddot{S}_N(0)$ which implies $\ddot{C}^\star_{1,N}(0)=\ddot{C}_{1,N}(0)$. To actually show the identity of the second order derivatives we show that the recursion to compute (5.62) is (5.59).

The probability that transmission occurs in slot n is

$$\Pr(n^\star=n) = \Pr\left(\alpha_n \ge s_{N-n}\right)\prod_{j=1}^{n-1}\Pr\left(\alpha_j < s_{N-j}\right) \tag{5.63}$$

Obviously, $\sum_{n=1}^{N}\Pr(n^\star=n)=1$. For every $n=1,\cdots,N$, the cdf of $\alpha_n 1\{n^\star=n\}$ is given by

$$\begin{aligned}
&\Pr(\alpha_n 1\{n^\star=n\}\le x)\\
&\quad= \Pr(\alpha_n 1\{n^\star=n\}\le x|n^\star=n)\Pr(n^\star=n)\\
&\qquad+\Pr(\alpha_n 1\{n^\star=n\}\le x|n^\star\ne n)\Pr(n^\star\ne n)\\
&\quad= \Pr(\alpha_n\le x|n^\star=n)\Pr(n^\star=n)+\Pr(0\le x|n^\star\ne n)\Pr(n^\star\ne n)\\
&\quad= \Pr(\alpha_n\le x|n^\star=n)\Pr(n^\star=n)+\Pr(n^\star\ne n)\quad\text{for}\quad x\ge 0
\end{aligned} \tag{5.64}$$

By recalling the expression of $\Pr(n^\star=n)$ in (5.63) we finally get

$$\begin{aligned}
\Pr(\alpha_n 1\{n^\star=n\}\le x) &= \prod_{j=1}^{n-1}\Pr(\alpha_j<s_{N-j})\Pr(\alpha_n\le x,\alpha_n\ge s_{N-n})\\
&\quad+\left(1-\Pr(\alpha_n\ge s_{N-n})\prod_{j=1}^{n-1}\Pr(\alpha_j<s_{N-j})\right)
\end{aligned} \tag{5.65}$$

and hence, for every value r, we have

$$\mathrm{E}\left[\alpha_n^r 1\{n^\star = n\}\right] = \prod_{j=1}^{n-1} F_\alpha(s_{N-j}) \int_{s_{N-n}}^{\infty} x^r \mathrm{d}F_\alpha(x) \tag{5.66}$$

By summing the terms in (5.66) over $n = 1, \cdots, N$ for $r = 1$ and $r = 2$ we get respectively (5.61) and (5.62). Let $\mu_N(r) \stackrel{\Delta}{=} \sum_{n=1}^{N} \mathrm{E}\left[\alpha_n^r 1\{n^\star = n\}\right]$, then by using (5.66) we have

$$\begin{aligned}
\mu_N(r) &= \sum_{n=1}^{N} \prod_{j=1}^{n-1} F_\alpha(s_{N-j}) \int_{s_{N-n}}^{\infty} x^r \, \mathrm{d}F_\alpha(x) \\
&= \int_{s_{N-1}}^{\infty} x^r \, \mathrm{d}F_\alpha(x) + \sum_{n=2}^{N} \prod_{j=1}^{n-1} F_\alpha(s_{N-j}) \int_{s_{N-n}}^{\infty} x^r \, \mathrm{d}F_\alpha(x) \\
&= \int_{s_{N-1}}^{\infty} x^r \, \mathrm{d}F_\alpha(x) + F_\alpha(s_{N-1}) \, \mu_{N-1}(r) \\
&= \mathrm{E}\left[\alpha_N^r \, 1\{\alpha_N \geq s_{N-1}\}\right] + \mathrm{E}\left[1\{\alpha < s_{N-1}\}\right] \mu_{N-1}(r) \qquad (5.67)
\end{aligned}$$

Since $\dot{S}_n(0) = s_n$ for all n and that $-\ddot{S}_N(0)$ and $\mu_N(2)$ satisfy the same recursion and have the same initial condition for $N = 0$, they coincide for all N. This concludes the proof.

Remark. The cdf (5.64) can be used to compute $C_{1,N}^\star(\gamma)$ as defined in (5.60) for all γ. In fact, with initial condition $S_0^\star(P) = 0$, we have

$$\begin{aligned}
S_N^\star(P) &\stackrel{\Delta}{=} \mathrm{E}\left[\sum_{n=1}^{N} \log\left(1 + P\alpha_n \, 1\{n^\star = n\}\right)\right] \\
&= \sum_{n=1}^{N} \prod_{j=1}^{n-1} \Pr[\alpha_j < s_{N-j}] \int_{s_{N-n}}^{\infty} \log\left(1 + Px\right) \mathrm{d}F_\alpha(x) \\
&= \int_{s_{N-1}}^{\infty} \log\left(1 + Px\right) \mathrm{d}F_\alpha(x) + \Pr[\alpha_j < s_{N-1}] S_{N-1}^\star(P) \qquad (5.68)
\end{aligned}$$

and $C_{1,N}^\star(\gamma) = \frac{1}{N} S_N^\star(N\gamma)$.

5.B Proof of Theorem 10

Consider the following inner and outer bound to the long-term average capacity region

$$\left\{ \boldsymbol{R} \in \mathbb{R}_+^K : \forall \mathcal{A} \quad \sum_{j \in \mathcal{A}} R_j \leq g^{(\mathcal{A})} \right\} \subseteq C_{K,N}(\boldsymbol{\gamma}) \subseteq \left\{ \boldsymbol{R} \in \mathbb{R}_+^K : \forall \mathcal{A} \quad \sum_{j \in \mathcal{A}} R_j \leq f^{(\mathcal{A})} \right\} \tag{5.69}$$

where we define the set functions

$$f^{(\mathcal{A})} \triangleq \mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N}\log\left(1+\sum_{j\in\mathcal{A}}\alpha_{j,n}\beta_{j,n}^{(\mathcal{A})}\right)\right] \tag{5.70}$$

where

$$\{\beta_{j,n}^{(\mathcal{A})} : j\in\mathcal{A}, n=1,\cdots,N\} \triangleq \arg\max_{\boldsymbol{\beta}\in\Gamma_{K,N}(\boldsymbol{\gamma})} \mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N}\log\left(1+\sum_{j\in\mathcal{A}}\alpha_{j,n}\beta_{j,n}\right)\right] \tag{5.71}$$

and

$$g^{(\mathcal{A})} \triangleq \mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N}\log\left(1+\sum_{j\in\mathcal{A}}\alpha_{j,n}\beta_{j,n}^{\star}\right)\right] \tag{5.72}$$

for all $\mathcal{A}\subseteq\{1,\cdots,K\}$. The inner bound in (5.69) is the long-term average capacity region when users apply the one-shot policy $\boldsymbol{\beta}^{\star}$ and the outer bound in (5.69) is obtained by applying the the "max-flow-min-cut" theorem for multi-terminal networks [1, Theorem 14.10.1] to our system.

Before proceeding, we point out some characteristics of the set functions $g^{(\mathcal{A})}$ and $f^{(\mathcal{A})}$. First, they do not depend on the whole SNR vector $\boldsymbol{\gamma}=(\gamma_1,\cdots,\gamma_K)$ but only on $\{\gamma_j\}_{j\in\mathcal{A}}$. Second, by recalling that $\beta_{j,n}^{\star}=N\,\gamma_j\,1\{n_j^{\star}=n\}$ for $n_j^{\star}$ defined in (4.25), it easy to see that, in the limit for $\boldsymbol{\gamma}\to\mathbf{0}$, the first-order partial derivative of $g^{(\mathcal{A})}(\{\gamma_j\}_{j\in\mathcal{A}})$ w.r.t. γ_ℓ for all $\ell\in\mathcal{A}$ is given by

$$\partial_\ell g^{(\mathcal{A})}(\mathbf{0}) = \sum_{n=1}^{N}\mathrm{E}\left[\alpha_{\ell,n}1\{n_\ell^{\star}=n\}\right] \tag{5.73}$$

and that the second-order partial derivative of $g^{(\mathcal{A})}(\{\gamma_j\}_{j\in\mathcal{A}})$ w.r.t. γ_ℓ and γ_m for all $\ell,m\in\mathcal{A}$ is given by

$$\partial_{\ell,m} g^{(\mathcal{A})}(\mathbf{0}) = -N\sum_{n=1}^{N}\mathrm{E}\left[\alpha_{\ell,n}1\{n_\ell^{\star}=n\}\alpha_{m,n}1\{n_m^{\star}=n\}\right] \tag{5.74}$$

Notice that, since $n_\ell^{\star}$ only depends on the fading sequence of user ℓ, in equation (5.74) the mean value factorizes when $\ell\neq m$. From Theorem 8 we have

$$\partial_\ell g^{(\mathcal{A})}(\mathbf{0}) = s_N^{(\ell)} = \dot{C}_{1,N}^{(\ell)}(0) \tag{5.75}$$

$$\partial_{\ell,\ell} g^{(\mathcal{A})}(\mathbf{0}) = \ddot{C}_{1,N}^{(\ell)}(0) \tag{5.76}$$

where $C_{1,N}^{(\ell)}(\gamma_\ell)$ is the ℓ-th user single-user long-term average capacity. Hence, we can write the single user wideband slope as $\mathcal{S}_0^{(\ell)} = -2(\partial_\ell g^{(\mathcal{A})}(\mathbf{0}))^2/\partial_{\ell,\ell} g^{(\mathcal{A})}(\mathbf{0})$.

Now we derive an achievable slope region based on the inner bound in (5.69). For a given permutation $\boldsymbol{\pi} = (\pi_1 \cdots, \pi_K)$ of $\{1, \ldots, K\}$, corresponding to the decoding order $\pi_K, \pi_{K-1}, \cdots, \pi_1$, we have the following vertex of the inner bound region

$$R_{\pi_k}(\boldsymbol{\pi}) = g^{\{\pi_1 \cdots, \pi_k\}} - g^{\{\pi_1 \cdots, \pi_{k-1}\}} \tag{5.77}$$

Every point on the dominant face of the inner bound region can be expressed as a convex combination of the $K!$ vertices, of coordinates (5.77), as follow

$$R_k = \sum_{\boldsymbol{\pi}} \lambda_{\boldsymbol{\pi}} R_{\pi_{\pi^{-1}(k)}}(\boldsymbol{\pi}) \tag{5.78}$$

where $\pi^{-1}(k)$ gives the position of the integer k in the permuted vector $\boldsymbol{\pi}$, where $\sum_{\boldsymbol{\pi}}$ denotes the sum over the $K!$ permutations of $\{1, \ldots, K\}$ and where $\boldsymbol{\lambda} = \{\lambda_{\boldsymbol{\pi}}\}$ are nonnegative "time-sharing" coefficients (indexed by the permutations $\boldsymbol{\pi}$) such that $\sum_{\boldsymbol{\pi}} \lambda_{\boldsymbol{\pi}} = 1$.

For fixed $(\theta_1, \cdots, \theta_K) \in \mathbb{R}_+^K$ we let $\gamma_k/\gamma_j = \theta_k/\theta_j$ for all $i, j \in \{1, \cdots, K\}$ and we compute the derivatives of $R_{\pi_k}(\boldsymbol{\pi})$ in (5.77), expressed as a function of γ_{π_k}, that for simplicity we indicate with x. The rate is given by

$$R_{\pi_k}(\boldsymbol{\pi}) = g^{\{\pi_1 \cdots, \pi_k\}}\left(\frac{\theta_{\pi_1}}{\theta_{\pi_k}}x, \cdots, \frac{\theta_{\pi_{k-1}}}{\theta_{\pi_k}}x, x\right) - g^{\{\pi_1 \cdots, \pi_{k-1}\}}\left(\frac{\theta_{\pi_1}}{\theta_{\pi_k}}x, \cdots, \frac{\theta_{\pi_{k-1}}}{\theta_{\pi_k}}x\right) \tag{5.79}$$

Its first derivative is

$$\dot{R}_{\pi_k}(\boldsymbol{\pi}) = \sum_{j=1}^{k-1} \frac{\theta_{\pi_j}}{\theta_{\pi_k}} \partial_{\pi_j}\left[g^{\{\pi_1 \cdots, \pi_k\}} - g^{\{\pi_1 \cdots, \pi_{k-1}\}}\right] + \partial_{\pi_k} g^{\{\pi_1 \cdots, \pi_k\}} \tag{5.80}$$

and its second derivative is

$$\begin{aligned} \ddot{R}_{\pi_k}(\boldsymbol{\pi}) &= \sum_{j=1}^{k-1}\sum_{\ell=1}^{k-1} \frac{\theta_{\pi_j}}{\theta_{\pi_k}} \frac{\theta_{\pi_\ell}}{\theta_{\pi_k}} \partial_{\pi_j, \pi_\ell}\left[g^{\{\pi_1 \cdots, \pi_k\}} - g^{\{\pi_1 \cdots, \pi_{k-1}\}}\right] \\ &\quad + 2\sum_{j=1}^{k-1} \frac{\theta_{\pi_j}}{\theta_{\pi_k}} \partial_{\pi_j, \pi_k} g^{\{\pi_1 \cdots, \pi_k\}} + \partial_{\pi_k, \pi_k} g^{\{\pi_1 \cdots, \pi_k\}} \end{aligned} \tag{5.81}$$

In the limit for $x \to 0$ we get

$$\lim_{x \to 0} \dot{R}_{\pi_k}(\boldsymbol{\pi}) = \partial_{\pi_k} g^{\{\pi_1 \cdots, \pi_k\}}(\mathbf{0}) \tag{5.82}$$

$$\lim_{x \to 0} \ddot{R}_{\pi_k}(\boldsymbol{\pi}) = \partial_{\pi_k, \pi_k} g^{\{\pi_1 \cdots, \pi_k\}}(\mathbf{0}) + 2\sum_{j=1}^{k-1} \frac{\theta_{\pi_j}}{\theta_{\pi_k}} \partial_{\pi_j, \pi_k} g^{\{\pi_1 \cdots, \pi_k\}}(\mathbf{0}) \tag{5.83}$$

Note that the summation in (5.83) accounts for the users not decoded yet according to the decoding order $\pi_K, \cdots, \pi_1$. Finally, by substituting (5.82)

and (5.83) in (5.78) we get

$$\begin{aligned}\lim_{x\to 0}\dot{R}_k &= \sum_{\boldsymbol{\pi}}\lambda_{\boldsymbol{\pi}}\partial_{\pi_k}g^{\{\pi_1\cdots,\pi_k\}}(\mathbf{0})\\ &= \partial_k g^{\{\pi_1\cdots,\pi_k\}}(\mathbf{0}) \qquad (5.84)\\ \lim_{x\to 0}\ddot{R}_k &= \sum_{\boldsymbol{\pi}}\lambda_{\boldsymbol{\pi}}\left(\partial_{\pi_k,\pi_k}g^{\{\pi_1\cdots,\pi_k\}}(\mathbf{0})+2\sum_{j=1}^{k-1}\frac{\theta_{\pi_j}}{\theta_{\pi_k}}\partial_{\pi_j,\pi_k}g^{\{\pi_1\cdots,\pi_k\}}(\mathbf{0})\right)\\ &= \partial_{k,k}g^{\{\pi_1\cdots,\pi_k\}}(\mathbf{0})+2\sum_{\boldsymbol{\pi}}\lambda_{\boldsymbol{\pi}}\sum_{j<\pi^{-1}(k)}\frac{\theta_{\pi_j}}{\theta_k}\partial_{\pi_j,k}g^{\{\pi_1\cdots,\pi_k\}}(\mathbf{0}) \qquad (5.85)\end{aligned}$$

By recalling (5.75) and (5.76), and from expression (5.74), we get

$$\mathcal{S}_k=\frac{\mathcal{S}_0^{(k)}}{1+2\sum_{\boldsymbol{\pi}}\lambda_{\boldsymbol{\pi}}\sum_{j<\pi^{-1}(k)}\frac{\theta_{\pi_j}}{\theta_k}\frac{\sum_{n=1}^{N}\mathrm{E}[\alpha_{\pi_j,n}1\{n^\star_{\pi_j}=n\}]\,\mathrm{E}[\alpha_{k,n}1\{n^\star_k=n\}]}{\sum_{n=1}^{N}\mathrm{E}[\alpha^2_{k,n}1\{n^\star_k=n\}]}} \qquad (5.86)$$

Since $\boldsymbol{\beta}^\star$ is a suboptimal policy, the slope region obtained as union over all $\boldsymbol{\lambda}$ of (5.86) for all k is in general an inner bound to the optimal slope region. Similarly, the slope region obtained considering the outer bound (5.69) is in general an outer bound to the optimal slope region. Next we prove that those two bounds coincide, thus proving that policy $\boldsymbol{\beta}^\star$ in conjunction with superposition-coding is second-order optimal for any number of users K and any delay N.

In order to express a general point on the dominant face of the outer bound in (5.69) we follow the same steps that led to (5.86). In particular we need the gradient and Hessian matrix of $f^{(\mathcal{A})}$, computed in $\boldsymbol{\gamma}=\mathbf{0}$, for all subsets $\mathcal{A}$. The proof that the the outer bound yields the same slope region of the inner bound is hence complete if we show that $\partial_{\ell,m}g^{(\mathcal{A})}(\mathbf{0})=\partial_{\ell,m}f^{(\mathcal{A})}(\mathbf{0})$ for all $\ell,m\in\mathcal{A}$ and $\ell\neq m$ and for all subsets $\mathcal{A}$. In fact it is obvious that $\partial_\ell g^{(\mathcal{A})}(\mathbf{0})=\partial_\ell f^{(\mathcal{A})}(\mathbf{0})$, otherwise the points on the outer-bound region would achieve higher minimum energy per bit than the points on the inner-bound region, and that $\partial_{\ell,\ell}g^{(\mathcal{A})}(\mathbf{0})=\partial_{\ell,\ell}f^{(\mathcal{A})}(\mathbf{0})$, otherwise the numerator of the equivalent of (5.86) for the outer bound region would be different from the optimal ℓ-th single-user wideband slope $\mathcal{S}_0^{(\ell)}$.

For every subset $\mathcal{A}$, for all $n=1,\cdots,N$ let

$$S_n\left(\{P_j\}_{j\in\mathcal{A}};\mathcal{A}\right)=\mathrm{E}\left[\max_{\forall j\in\mathcal{A}:u_j\in[0,P_j]}\log\left(1+\sum_{j\in\mathcal{A}}\alpha_j\,u_j\right)+S_{n-1}\left(\{P_j-u_j\}_{j\in\mathcal{A}};\mathcal{A}\right)\right] \qquad (5.87)$$

with initial condition $S_0\left(\mathbf{0};\mathcal{A}\right)=0$, then

$$f^{(\mathcal{A})}(\{\gamma_j\}_{j\in\mathcal{A}})=\frac{1}{N}S_N\left(\{N\,\gamma_j\}_{j\in\mathcal{A}};\mathcal{A}\right) \qquad (5.88)$$

Let $\boldsymbol{b} \in \{0,1\}^{|\mathcal{A}|}$, then a necessary condition for $\{u_j = P_j\, b_j\}_{j\in\mathcal{A}}$ to be solution of (5.87) is

$$\frac{\alpha_\ell}{1+\sum_{j\in\mathcal{A}}\alpha_j\, P_j\, b_j} - \partial_\ell S_{n-1}\left(\{P_j(1-b_j)\}_{j\in\mathcal{A}};\mathcal{A}\right) \begin{cases} <0 & \text{if} \quad b_\ell = 0 \\ \geq 0 & \text{if} \quad b_\ell = 1 \end{cases} \tag{5.89}$$

Then it follows easily that in the limit for small $\{P_j\}_{j\in\mathcal{A}}$ we have $u_\ell = 0$ if $\alpha_\ell < \partial_\ell S_{n-1}(\mathbf{0};\mathcal{A})$ and $u_\ell = P_\ell$ if $\alpha_\ell \geq \partial_\ell S_{n-1}(\mathbf{0};\mathcal{A})$. Then we can write

$$\begin{aligned} S_n\left(\{P_j\}_{j\in\mathcal{A}};\mathcal{A}\right) &= \sum_{\boldsymbol{b}} \mathrm{E}\left[\left(\log\left(1+\sum_{j\in\mathcal{A}}\alpha_j\, P_j\, b_j\right) + S_{n-1}\Big(\{P_j(1-b_j)\}_{j\in\mathcal{A}};\mathcal{A}\Big)\right)\cdot\right. \\ &\qquad \left.\cdot\prod_{j\in\mathcal{A}} 1\{u_j = P_j\, b_j\}\right] + \text{vanishing terms with} \quad \{P_j\}_{j\in\mathcal{A}} \end{aligned} \tag{5.90}$$

Finally, in the limit for vanishing $\{P_j\}_{j\in\mathcal{A}}$ the second-order partial derivative of $S_n\left(\{P_j\}_{j\in\mathcal{A}};\mathcal{A}\right)$ w.r.t. P_ℓ and P_m is

$$\begin{aligned} \partial_{\ell,m} S_n(\mathbf{0};\mathcal{A}) &= \sum_{\boldsymbol{b}} \mathrm{E}\Big[\left(-b_\ell b_m \alpha_\ell \alpha_m + (1-b_\ell)(1-b_m)\partial_{\ell,m} S_{n-1}(\mathbf{0};\mathcal{A})\right)\cdot \\ &\qquad \cdot \prod_{j\in\mathcal{A}:b_j=0} 1\{\alpha_j < \partial_\ell S_{n-1}(\mathbf{0};\mathcal{A})\} \prod_{j\in\mathcal{A}:b_j=1} 1\{\alpha_j \geq \partial_\ell S_{n-1}(\mathbf{0};\mathcal{A})\}\Big] \\ &= \mathrm{E}\Big[-\alpha_\ell\alpha_m 1\{\alpha_\ell \geq \partial_\ell S_{n-1}(\mathbf{0};\mathcal{A})\}1\{\alpha_m \geq \partial_m S_{n-1}(\mathbf{0};\mathcal{A})\} \\ &\qquad + \partial_{\ell,m} S_{n-1}(\mathbf{0};\mathcal{A}) 1\{\alpha_\ell < \partial_\ell S_{n-1}(\mathbf{0};\mathcal{A})\}1\{\alpha_m < \partial_m S_{n-1}(\mathbf{0};\mathcal{A})\}\Big] \\ &= \frac{1}{N}\partial_{\ell,m} f^{(\mathcal{A})}(\mathbf{0}) \end{aligned} \tag{5.91}$$

In order to prove that $N\partial_{\ell,m} S_N(\mathbf{0};\mathcal{A})$ indeed coincides with (5.74) we must show that (5.91) is the recursion to compute (5.74). In fact, by recalling (5.66), we can write

$$\mu_N(\ell,m) \stackrel{\Delta}{=} \sum_{n=1}^{N} \mathrm{E}\left[\alpha_{\ell,n} 1\{n_\ell^\star = n\}\right] \mathrm{E}\left[\alpha_{m,n} 1\{n_m^\star = n\}\right] \tag{5.92}$$

$$= \sum_{n=1}^{N}\prod_{j=1}^{n-1} F_\alpha^{(\ell)}(s_{N-j}^{(\ell)}) \int_{s_{N-n}^{(\ell)}}^{\infty} x\,\mathrm{d}F_\alpha^{(\ell)}(x) \cdot \prod_{j=1}^{n-1} F_\alpha^{(m)}(s_{N-j}^{(m)}) \int_{s_{N-n}^{(m)}}^{\infty} x\,\mathrm{d}F_\alpha^{(m)}(x)$$

now, by separating the term for $n = 1$ in the summation, we can write

$$\begin{aligned}\mu_N(\ell,m) &= \int_{s_{N-1}^{(\ell)}}^{\infty} x\mathrm{d}F_\alpha^{(\ell)}(x) \cdot \int_{s_{N-1}^{(m)}}^{\infty} x\mathrm{d}F_\alpha^{(m)}(x) \\ &\quad + \sum_{n=2}^{N} \prod_{j=1}^{n-1} F_\alpha^{(\ell)}(s_{N-j}^{(\ell)}) \int_{s_{N-n}^{(\ell)}}^{\infty} x\mathrm{d}F_\alpha^{(\ell)}(x) \cdot \prod_{j=1}^{n-1} F_\alpha^{(m)}(s_{N-j}^{(m)}) \int_{s_{N-n}^{(m)}}^{\infty} x\mathrm{d}F_\alpha^{(m)}(x) \\ &= \mathrm{E}\left[\alpha_\ell 1\{\alpha_\ell \geq s_{N-1}^{(\ell)}\}\right] \cdot \mathrm{E}\left[\alpha_m 1\{\alpha_m \geq s_{N-1}^{(m)}\}\right] \\ &\quad + \mathrm{E}\left[1\{\alpha_\ell < s_{N-1}^{(\ell)}\}\right] \cdot \mathrm{E}\left[1\{\alpha_m < s_{N-1}^{(m)}\}\right] \cdot \mu_{N-1}(\ell,m) \end{aligned} \tag{5.93}$$

which, by recalling $\partial_\ell S_n(\mathbf{0};\mathcal{A}) = s_n^{(\ell)}$ for all $n = 1, 2, \cdots$ and all $\ell \in \{1, \cdots, K\}$, coincides with (5.91) for $n = N$. This concludes the proof that $\partial_{\ell,m} g^{(\mathcal{A})}(\mathbf{0}) = \partial_{\ell,m} f^{(\mathcal{A})}(\mathbf{0})$ for all $\ell, m \in \mathcal{A}$ and for all subsets $\mathcal{A}$, thus proving that the optimal slope region, parameterized by $\boldsymbol{\theta}$ can be written as in (5.33)

$$\bigcup_{\boldsymbol{\lambda}} \left\{ \mathcal{S}_k \geq 0 \quad \forall k = 1, \cdots, K \right. \tag{5.94}$$

$$\left. \mathcal{S}_k \leq \frac{\mathcal{S}_0^{(k)}}{1 + 2\sum_{\boldsymbol{\pi}} \lambda_{\boldsymbol{\pi}} \sum_{j<\pi^{-1}(k)} \frac{\theta_{\pi_j}}{\theta_k} \frac{\sum_{n=1}^{N} \mathrm{E}[\alpha_{\pi_j,n} 1\{n_{\pi_j}^\star = n\}]\, \mathrm{E}[\alpha_{k,n} 1\{n_k^\star = n\}]}{\sum_{n=1}^{N} \mathrm{E}[\alpha_{k,n}^2 1\{n_k^\star = n\}]}} \right\}$$

and that the one-shot policy $\boldsymbol{\beta}^\star$ is second-order optimal.

Remark. With a technique similar to that that lead to (5.93) it can be shown, for all $\mathcal{A}$, that

$$g^{(\mathcal{A})}(\{\gamma_j\}_{j\in\mathcal{A}}) = \frac{1}{N} S_N^\star\left(\{N\,\gamma_j\}_{j\in\mathcal{A}};\mathcal{A}\right) \tag{5.95}$$

where

$$\begin{aligned} S_N^\star\left(\{P_j\}_{j\in\mathcal{A}};\mathcal{A}\right) &= \mathrm{E}\left[\log\left(1 + \sum_{j\in\mathcal{A}} \alpha_j P_j\right) \prod_{j\in\mathcal{A}} 1\{\alpha_j \geq s_{N-1}^{(j)}\}\right] \\ &\quad + \prod_{j\in\mathcal{A}} \Pr[\alpha_j < s_{N-1}^{(j)}] S_{N-1}^\star\left(\{P_j\}_{j\in\mathcal{A}};\mathcal{A}\right) \end{aligned}$$

with initial condition $S_0^\star\left(\{P_j\}_{j\in\mathcal{A}};\mathcal{A}\right) = 0$.

5.C Proof of Theorem 11

The proof follows the same steps of the proof of Theorem 8 in Appendix 5.A. First compute the first and second derivative at $\gamma = 0$ of the long-term average rate achieved by applying the suboptimal "maximum selection" policy

$\boldsymbol{\beta}^{\star\,(\mathrm{nc})}$. Then, compute the first and second derivative at $\gamma = 0$ of the long-term average capacity obtained by applying the optimal "waterfilling" policy $\widehat{\boldsymbol{\beta}}^{(\mathrm{nc})}$. It is easy to show that the derivatives coincide by writing the long-term average capacity as

$$\begin{aligned} C_{1,N}^{(\mathrm{nc})}(\gamma) &= \mathrm{E}\left[\frac{1}{N}\log\left(1+N\gamma\alpha_{\max}\right)\cdot 1\left\{\frac{\alpha_{\max}}{1+\alpha_{\max}N\gamma}\geq\alpha_n\ \forall\alpha_n\neq\alpha_{\max}\right\}\right] \\ &\quad +o(\gamma)\quad \text{as}\quad \gamma\to 0 \end{aligned} \tag{5.96}$$

where $\alpha_{\max} = \max\{\alpha_1, \cdots, \alpha_N\}$ and where the only event that "counts" in the limit for $\gamma \to 0$ is the event in the indicator function in (5.96) since

$$\lim_{\gamma\to 0}\Pr\left(\frac{\alpha_{\max}}{1+\alpha_{\max}N\gamma}\geq\alpha_n\ \forall\alpha_n\neq\alpha_{\max}\right)=1 \tag{5.97}$$

5.D Proof of Theorem 13

The proof follows the same steps of the proof of Theorem 10 in Appendix 5.B. As in (5.69), the long-term average capacity region $C_{K,N}^{(\mathrm{nc})}(\boldsymbol{\gamma})$ contains the the inner bound region obtained by applying $\boldsymbol{\beta}^{\star\,(\mathrm{nc})}$ and is contained in the outer bound region obtained by mean of the "max-flow-min-cut" theorem for multi-terminal networks [1, Theorem 14.10.1]. We must prove that, for any subset $\mathcal{A} \subseteq \{1, \cdots, K\}$, the gradient and the Hessian matrix at $\boldsymbol{\gamma} = \mathbf{0}$ of the partial rate-sums that define the inner and outer bound regions coincide.

By construction $\boldsymbol{\beta}^{\star\,(\mathrm{nc})}$ is first-order optimal, hence the gradients of the partial rate-sums of inner and outer bound regions coincide for any subset $\mathcal{A} \subseteq \{1, \cdots, K\}$.

In order to prove that the Hessian matrix of the partial sum-rates of the inner and outer bound coincide, it is enough to write the partial rate-sums of the outer bound region as follow

$$\begin{aligned} &\max_{\boldsymbol{\beta}^{(\mathrm{nc})}}\mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N}\log\left(1+\sum_{j\in\mathcal{A}}\alpha_{j,n}\beta_{j,n}\right)\right] \\ &=\sum_{\boldsymbol{b}\in\{1,\cdots,N\}^{|A|}}\mathrm{E}\left[\frac{1}{N}\sum_{n=1}^{N}\log\left(1+\sum_{j\in\mathcal{A}}\alpha_{j,n}N\gamma_j\,1\{n=b_j\}\right)\cdot\right. \\ &\left.\cdot\prod_{k\in\mathcal{A}}1\left\{\frac{\alpha_{k,b_k}}{1+\sum_{j\in\mathcal{A}}\alpha_{j,b_k}N\gamma_j\,1\{b_k=b_j\}}\geq\frac{\alpha_{k,\ell}}{1+\sum_{j\in\mathcal{A}}\alpha_{j,\ell}N\gamma_j\,1\{\ell=b_j\}}\quad\forall\ell\neq b_k\right\}\right] \\ &+\text{vanishing terms with }\ \boldsymbol{\gamma} \end{aligned} \tag{5.98}$$

where we have obtained (5.98) by imposing the (necessary) Khun-Taker conditions for $\{\beta_{j,n} = N\gamma_j\,1\{n = b_j\}\ \ n = 1, \cdots, N\ \ j \in \mathcal{A}$ to be the optimal

solution. Only those type of solutions "count" in the limit for $\boldsymbol{\gamma} \to \mathbf{0}$ since the argument of indicator functions in (5.98) tends to $\{\alpha_{k,b_k} \geq \alpha_{k,\ell} \ \ \forall \ell \neq b_k\}$ which is clearly a partition of the whole fading space.

The expression for $\mathcal{K}_{k,j}$ in (5.37) comes form (5.64) with $n_k^\star = \max\{\alpha_{k,1}, \cdots, \alpha_{k,N}\}$, i.e.,

$$\Pr[\alpha_{k,n} 1\{n_k^\star = n\} \leq x] = \frac{N-1}{N} + \frac{1}{N}\Big(F_\alpha^{(k)}(x)\Big)^N \tag{5.99}$$

Chapter 6

Conclusions

In this report we have addressed the problem of multi-access for wireless fading channels subject to delay constraint, where the maximum delay is assumed to be a fixed design parameter. We have taken two "complementary" point of view.

Part I. In one case, we have considered *user random activity* in a system without channel state information at the transmitters, i.e., constant-rate transmission with constant per-symbol energy, and where erroneously received packets are *retransmitted and combined* to improve decoding reliability.

- In Chapter 2 we presented an information-theoretic throughput analysis of some Hybrid-ARQ protocols under idealized but fairly general conditions. We showed that typical set decoding has very desirable properties for Hybrid-ARQ, in the limit for large slot dimension. From a renewal-reward theory approach, we obtained closed-form throughput formulas for three simple protocols: a generalization of slotted Aloha (ALO), a repetition time diversity scheme with maximal-ratio packet combining (RTD) and an incremental redundancy scheme based on progressively punctured codes (INR). We analyzed the effect of delay and rate constraints on the throughput, as well as the limiting behavior with respect to the slot spectral efficiency, the channel load and the transmit SNR. Interestingly, all three protocols are not interference-limited, and achieve arbitrarily large throughput by simply increasing the transmit power of all users. Publications related to this chapter are [24, 25].

- Chapter 3 presented a comparison between three different multi-access strategies in a scenario characterized by random activity of an infinite population of uncoordinated users. Different retransmission protocols and combing techniques are considered in presence of block fading and additive noise. To make the comparison fair, the system throughput is optimized with respect to all the system parameters and expressed as function of the individual E_b/N_0. The best performance is obtained by joint decoding even without packet combing. Among SUD-based systems, MMSE-CDMA outperforms the system without spreading, while the SUMF-CDMA is heavily suboptimal and interference limited. We showed that at low E_b/N_0 all SUD-based system are equivalent to CDMA with SUMF, suggesting that practical system operating in this region do not need to be complex and that users must transmit continuously with vanishing rate. At high E_b/N_0 the best strategy, for SUD-based system, is having on the average one active user per degree of freedom transmitting at non-vanishing rate, this makes the system not interference limited. In practice, a call admission control scheme should keep the channel load G close to its optimum value, depending on the operating E_b/N_0. Publications related to this chapter are [26, 27].

We have concluded that repetition together with code combining is a simple and viable strategy to overcome fading and multiuser interference in a system where user coordination (rate/power allocation) seems problematic due to user random activity. The simplicity of the system is paid in term of total throughput with respect to more complex systems that implement joint detection, like MMSE-CDMA, or joint decoding. Our repetition strategy is not interference limited and outperforms "naive" SUMF-CDMA.

Part II. In the second case, we have assumed that users transmit continuously, have *causal* channel state information, hence allowing for rate and power to vary according to channel conditions, and are subject to a per-codeword power constraint due to transmitters *energy limitation*.

- In Chapter 4 we analyzed an idealized fading model where each codeword sees N independently drawn fading states, known to the transmitter causally. The power control algorithm at the transmitter must decide what portion of the available energy to allocate to each fading state based only on the knowledge of current and past fading states. We have solved for the optimal power control policy and capacity for fixed arbitrary N and for arbitrary number of users. The optimal policy is to concentrate all the energy in only one of the fading states. That state is chosen on the basis of not only its strength, but also how likely it is that a more favorable fading state will appear before the end of the codeword. Our publications related to this chapter are [28, 29].

- Then, in Chapter 5, we focused on the low spectral efficiency regime, which is where the major benefits of transmitter feedback occur. We have analyzed not only the rates achievable in the absence of bandwidth constraints (minimum energy per bit), but also the bandwidth required to provide a given rate with given power in the low power regime (wideband slope). As a result of this analysis we have quantified the bandwidth expansion required by using the suboptimal TDMA strategy. The TDMA bandwidth penalty, which can be rather substantial, depends on the fading distribution, and grows with both the number of fading states N and the number of users K. Our publications related to this chapter are [30, 31, 32].

We have concluded that the penalty incurred by causal feedback in an energy limited system becomes negligible, with respect to the optimal ergodic system, when the delay constraint is not too strict. We have identified a rate/power allocation strategy the is optimal in the wideband regime with respect to both the minimum energy per bit and the wideband slope. In performing the wideband analysis, we have also shown that TDMA (with optimal power and rate allocation within each sub-slot) is heavily suboptimal in a fading multi-user system thus disproving the common belief that "TDMA is optimal in the low-power/wideband regime".

Bibliography

[1] T.Cover and J.Thomas, *Elements of information theory*. New York: Wiley, 1991.

[2] L.Ozarow, S.Shamai(Shitz), and A.D.Wyner, "Information theoretic considerations for cellular mobile radio," *IEEE Trans. on Vehic. Tech.*, vol. 43, pp. 359–378, May 1994.

[3] S.Hanly and D.Tse, "Multiaccess fading channels-part II: Delay limited capacities," *IEEE Trans. on Inform. Theory*, vol. 44, pp. 2816–2831, November 1998.

[4] A. Lapidoth and P. Narayan, "Reliable communication under channel uncertainty," *IEEE Trans. on Inform. Theory*, vol. 44, pp. 2148 –2177, October 1998.

[5] E.Biglieri, J.Proakis, and S.Shamai(Shitz), "Fading channels: information-theoretic and communications aspects," *IEEE Trans. on Inform. Theory*, vol. 44, pp. 2619 –2692, October 1998.

[6] R.Ahlswede, "Multi-way communication channels," in *Proceedings 1971 IEEE International Symposium on Information Theory*, (Armenian S.S.R), 1973.

[7] H.Liao, *Multiaccess channels*. PhD thesis, Dep.Elec.Eng.Univ.Hawaii, Honlulu, 1972.

[8] T. Han, "An information-spectrum approach to capacity theorems for the general multiple-access channel," *IEEE Trans. on Inform. Theory*, vol. 44, pp. 2773 –2795, Nov 1998.

[9] S.Verdú and T.S.Han, "A general formula for channel capacity," *IEEE Trans. on Inform. Theory*, vol. 40, pp. 1147–1157, July 1994.

[10] A.D.Wyner, "Shannon-theoretic approach to the gaussian multiple-access channel," *IEEE Trans. on Inform. Theory*, vol. 40, pp. 1713–1726, November 1994.

[11] A.Goldsmith and P.P.Varaiya, "Capacity of fading channels with channel state information," *IEEE Trans. on Inform. Theory*, vol. 43, pp. 1986–1992, November 1997.

[12] R.Knopp and P.A.Humblet, "Information capacity and power control in single-cell multiuser communications," in *IEEE International Conference on Communications, 1995 (ICC '95), 'Gateway to Globalization'*, vol. 1, (Seattle), pp. 331–335, July 1999.

[13] D.Tse and S.Hanly, "Multiaccess fading channels-part I: Polymatroid structure, optimal resource allocation and throughput capacities," *IEEE Trans. on Inform. Theory*, vol. 44, pp. 2796–2815, November 1998.

[14] S.Verdú, "On channel capacity per unit cost," *IEEE Trans. on Inform. Theory*, vol. 36, pp. 1019–1030, September 1990.

[15] R.S.Cheng and S.Verdú, "Gaussian multiaccess channels with ISI: capacity region and multiuser water-filling," *IEEE Trans. on Inform. Theory*, vol. 39, pp. 773 –785, May 1993.

[16] Gaarder and Woolf, "The capacity region of a multiple access discrete memory less channel can increase with feedback," *IEEE Trans. on Inform. Theory*, vol. 21, pp. 100–102, 1975.

[17] R.G.Gallager, "A prospective on multiaccess channels," *IEEE Trans. on Inform. Theory*, vol. 31, pp. 124–142, March 1985.

[18] A.Ephremides and B.Hajek, "Information theory and communication networks: An unconsummed union," *IEEE Trans. on Inform. Theory*, vol. 44, pp. 2416–2434, October 1998.

[19] R. Alsugair, A.A.and Cheng, "Symmetric capacity and signal design for l-out-of-k symbol-synchronous cdma gaussian channels," *IEEE Trans. on Inform. Theory*, vol. 41, pp. 1072 –1082, July 1995.

[20] E.Telatar and R.G.Gallager, "Combining queuing theory with information theory for multiaccess," *IEEE J. Select. Areas Commun.*, vol. 13, pp. 963–969, August 1995.

[21] R.Berry and R.Gallager, "Communication over fading channels with finite buffer constraints - single user and multiple access cases," in *Proceedings 2001 IEEE International Symposium on Information Theory*, (Washington DC, USA), June 2001.

[22] I.Bettesh and S.Shamai(Shitz), "Optimal power and rate control for fading channels," in *Proceedings of VTC 2001 Spring*, (Greece), 2001.

[23] I.Bettesh and S.Shamai(Shitz), "Queuing analysis of the single user fading channel," in *Proceedings of the 21st IEEE Convention of the Electrical and Electronic Engineers*, (Tel Aviv, Israel), April 11-12, 2000.

[24] G.Caire and D.Tuninetti, "ARQ protocols for the Gaussian collision channel," in *Proceedings 2000 IEEE International Symposium on Information Theory*, (Sorrento, Italy), June 2000.

[25] G.Caire and D.Tuninetti, "The throughput of Hybrid-ARQ protocols for the Gaussian collision channel," *IEEE Trans. on Inform. Theory*, vol. 47, pp. 1971–1988, July 2001.

[26] D.Tuninetti and G.Caire, "On the optimal throughput of some wireless systems," in *Proceedings 2001 IEEE International Symposium on Information Theory*, (Washington DC, USA), June 2001.

[27] D.Tuninetti and G.Caire, "On the optimal throughput of some wireless systems," *submitted to Trans. on Inform. Theory*, October 2000.

[28] D. Tuninetti and G. Caire, "The effect of delay constraint and causal feedback on the wideband performance of multiaccess block-fading channels," in *Proceedings Asilomar conference 2001*, (Pacific Grove (Ca, USA)), 2001.

[29] D. Tuninetti and G. Caire, "The long-term average capacity per unit energy with application to protocols for sensor networks," in *Proceedings of 2002 European Wireless Conference (EW2002)*, (Firenze, IT), 2001.

[30] S.Verdú, G.Caire, and D.Tuninetti, "s TDMA optimal in the low power regime?," in *Proceedings 2001 IEEE International Symposium on Information Theory*, (Lausanne, CH), June 2002.

[31] D.Tuninetti, G.Caire, and S.Verdú, "Fading multiaccess channels in the wideband regime: the impact of delay constraints," in *Proceedings 2001 IEEE International Symposium on Information Theory*, (Lausanne, CH), June 2002.

[32] D.Tuninetti, G.Caire, and S.Verdú, "On fading multiple-access channels in the wideband regime: the impact of delay constraints and causal feedback." submitted to, February 2002.

[33] T.Ojanpera, "Overview of research activities for third generation mobile communication," *Wireless Communications TDMA vs. CDMA (S.G. Glisic and P.A.Leppanen, Eds.)*, pp. 415–446, 1997.

[34] S.Lin and D.Costello, *Error control coding : fundamentals and applications*. New York: Prentice-Hall, 1983.

[35] D.Costello, J.Hagenauer, H.Imai, and S.Wicker, "Applications of error-control coding," *IEEE Trans. on Inform. Theory*, vol. 44, pp. 2531–2560, October 1998.

[36] S.Shamai(Shitz) and A.D.Wyner, "Information-theoretic considerations for symmetric, cellular, multiple-access fading channels - Part I," *IEEE Trans. on Inform. Theory*, vol. 43, pp. 1877–1894, November 1997.

[37] S.Shamai(Shitz) and A.D.Wyner, "Information-theoretic considerations for symmetric, cellular, multiple-access fading channels - Part II," *IEEE Trans. on Inform. Theory*, vol. 43, pp. 1895–1911, November 1997.

[38] D.Bertsekas and R.Gallager, *Data Networks (2nd ed.)*. New York: Prentice-Hall, 1987.

[39] G.Caire, E.Leonardi, and E.Viterbo, "Modulation and coding for the Gaussian collision channel," *IEEE Trans. on Inform. Theory*, vol. 46, pp. 2007 –2026, September 2000.

[40] P.S.Sindhu, "Retransmission error control with memory," *IEEE Trans. on Commun.*, vol. 25, pp. 473–479, May 1977.

[41] G.Benelli, "An ARQ scheme with memory and soft error detection," *IEEE Trans. on Commun.*, vol. 33, pp. 285–288, March 1985.

[42] S.B.Wicker, "Adaptive error control through the use of diversity combining majority logic decoding in Hybrid ARQ protocol," *IEEE Trans. on Commun.*, vol. 39, pp. 380–385, March 1991.

[43] S.S.Chakraborty, M.Liinaharja, and E.Yli-Juuti, "An adaptive ARQ scheme with packet combining for time varying channels," *IEEE Commun. Letters*, vol. 3, pp. 52–54, February 1999.

[44] S.S.Chakraborty, E.Yli-Juuti, and M.Liinaharja, "An adaptive ARQ scheme with packet combining," *IEEE Commun. Letters*, vol. 2, pp. 200–202, July 1998.

[45] D.Chase, "Code combining - A maximum-likelihood decoding approch for combining an arbitrary number of noisy packets," *IEEE Trans. on Commun.*, vol. 33, pp. 385–393, May 1985.

[46] B.A.Harvey and S.B.Wicker, "Packet combining system based on the Viterbi decoder," *IEEE Trans. on Commun.*, vol. 42, pp. 1544–1557, February/March/April 1994.

[47] S.Kallel, "Analysis of a Type-II Hybrid ARQ scheme with code combining," *IEEE Trans. on Commun.*, vol. 38, pp. 1133–1137, August 1990.

[48] M.Schwartz, W.Bennet, and S.Stein, *Communication Systems and Techniques*. New York: McGraw-Hill, 1966.

[49] J.Hagenauer, "Rate-compatible punctured convolutional codes (RCPC codes) and their applications," *IEEE Trans. on Commun.*, vol. 36, pp. 389–400, April 1988.

[50] S.Kallel, "Complementary punctured convolutional (CPC) codes and their applications," *IEEE Trans. on Commun.*, vol. 43, pp. 2005–2009, June 1995.

[51] C.Berrou and A.Glavieux, "Near optimum error correcting coding and decoding: Turbo-codes," *IEEE Trans. on Commun.*, vol. 44, pp. 1261–1271, October 1996.

[52] W.C.Chan, E.Geraniotis, and V.D.Nguyen, "An adaptive Hybrid FEC/ARQ protocol using turbo codes for multi-media traffic," in *Conference Record of IEEE 6th International Conference on Universal Personal Communications*, vol. 2, pp. 541 –545, 1997.

[53] K.R.Narayanan and G.L.Stuber, "A novel ARQ techique using the turbo coding principle," *IEEE Commun. Letters*, vol. 1, pp. 49–51, March 1997.

[54] M.Zorzi and R.R.Rao, "Performance of ARQ Go-Back-N protocol in Markov channels with unreliable feedback," *Mobile Networks and Applications*, vol. 2, pp. 183–193, 1997.

[55] M.Zorzi and R.R.Rao, "Throughput performance of ARQ selective-repeat with time diversity in Markov channels with unreliable feedback," *WIRELESS NETWORK*, vol. 2, pp. 63–75, March 1996.

[56] M.Zorzi and F.Borgonovo, "Performance of capture-division paket access with slow shadowing and power control," *IEEE Trans. on Vehic. Tech.*, vol. 46, pp. 687 –696, August 1997.

[57] M.Zorzi, "Mobile radio slotted ALOHA with capture, diversity and retransmission control in the presence of shadowing," *Wireless Networks*, vol. 4, pp. 379–388, August 1998.

[58] R. S.Kallel and S.Bakhtiyari, "Throughput performance of memory ARQ scheme," *IEEE Trans. on Vehic. Tech.*, vol. 48, pp. 891–899, May 1999.

[59] R.Cam and C.Leung, "Throughput analysis of some ARQ protocols in the presence of feedback errors," *IEEE Trans. on Commun.*, vol. 45, pp. 35–44, January 1997.

[60] Q.Zhang, T.F.Wong, and J.S.Lehnert, "Performance of Type-II Hybrid ARQ protocol in slotted DS-SSMA packet radio systems," *IEEE Trans. on Commun.*, vol. 47, pp. 281–290, February 1999.

[61] A.M.Y.Bigloo, T.A.Gulliver, and V.K.Bhargava, "Maximum-likelihood decoding and code combining for DS-SSMA slotted Aloha," *IEEE Trans. on Commun.*, vol. 45, pp. 1602–1612, December 1997.

[62] S.Souissi and S.B.Wicker, "A diversity combining DS/SSMA system with convolutional encoding and Viterbi decoding," *IEEE Trans. on Vehic. Tech.*, vol. 44, pp. 304–312, May 1995.

[63] R.K.Morrow and J.S.Lehnert, "Packet throughput in ALOHA DS/SSMA radio system with random signature sequences," *IEEE Trans. on Commun.*, vol. 40, pp. 1223–1230, July 1992.

[64] I.Habbab, M.Kaverhad, and C.E.Sundberg, "ALOHA with capture over slow and fast fading radio channels with coding and diversity," *IEEE J. Select. Areas Commun.*, vol. 7, pp. 79–88, January 1989.

[65] M.Zorzi and R.R.Rao, "On the use of renewal theory in the analysis of ARQ protocols," *IEEE Trans. on Commun.*, vol. 44, pp. 1077–1081, September 1996.

[66] R.Wolff, *Stochastic modeling and the theory of queues*. Upper Saddle River, New York: Prentice-Hall, 1989.

[67] B.Rimoldi and R.Urbanke, "A rate splitting approach to the Gaussian multiple-access channel," *IEEE Trans. on Inform. Theory*, vol. 42, pp. 364–375, March 1996.

[68] M.Mecking, "Expected rate and power allocation for Gaussian random multiple-access channels," in *Proceedings 2000 IEEE International Symposium on Information Theory*, (Sorrento, Italy), June 2000.

[69] D.Tse and S.Hanly, "Linear multiuser receivers: effective interference, effective bandwidth and capacity," *IEEE Trans. on Inform. Theory*, vol. 45, pp. 641–657, March 1999.

[70] S.Verdú and S.Shamai(Shitz), "Spectral efficiency of CDMA with random spreading," *IEEE Trans. on Inform. Theory*, vol. 45, pp. 622–640, March 1999.

[71] J.G.Proakis, *Digital Communications (3nd ed.)*. New York: McGraw Hill, 1995.

[72] E. SMG2, "The ETSI UMTS terrestrial radio access (UTRA) ITU-R RTT candidate submission." Tdoc SMG2 260/98, June 1998.

[73] J.Massey and P.Mathys, "The collision channel without feedback," *IEEE Trans. on Inform. Theory*, vol. 31, pp. 192–204, February 1985.

[74] A.J.Viterbi, *CDMA principles of Spread Spectrum Communication.* New York: Addison-Wesley, 1995.

[75] T.Rappaport, *Wireless Communications.* Englewood Cliffs, New York: Prentice-Hall, 1996.

[76] G.Caire and C.F.Leanderson, "Throughput performance of an incremental redundancy ARQ scheme in the block -fading Gaussian collision channel," in *Proceedings 2001 IEEE International Symposium on Information Theory*, (Whasington, USA), June 2001.

[77] R.Gallager, *Information theory and reliable communication.* New York: Wiley, 1968.

[78] R.E.Blahut, *Principles and practice of information theory.* New York: Addison-Wesley, 1987.

[79] J.Hagenauer, E.Offer, and L.Papke, "Iterative decoding of binary block and convolutional codes," *IEEE Trans. on Inform. Theory*, vol. 42, pp. 429–445, March 1996.

[80] E.Biglieri, G.Caire, and G.Taricco, "On the convergence of the iterated decoding algorithm," in *Proceedings of the* SBT/IEEE International Telecommunications Symposium, *Rio de Janeiro, Brasil*, (Rio de J.), August 22-25 1994.

[81] F.Burkert, G.Caire, J.Hagenauer, T.Hindelang, and G.Lechner, "Turbo decoding with unequal error protection applied to GSM speech coding," in *IEEE GLOBECOM '96*, (London, UK), November, 18-22 1996.

[82] G.Caire, R.Knopp, and P.Humblet, "System capacity of F-TDMA cellular systems," *IEEE Trans. on Commun.*, vol. 46, pp. 1649–1661, December 1998.

[83] P.Billingsley, *Probability and measure.* New York: Wiley, 1986.

[84] R.J.McEliece and W.E.Stark, "Channels with block interference," *IEEE Trans. on Inform. Theory*, vol. 30, pp. 44–53, January 1984.

[85] S.Shamai(Shitz) and S.Verdú, "The effect of flat fading on the spectral efficiency of CDMA with random spreading," *IEEE Trans. on Inform. Theory*, vol. 47, pp. 1302–1327, May 2001.

[86] E.Biglieri, G.Caire, G.Taricco, and E.Viterbo, "CDMA with fading: effective bandwidth and spreading-coding tradeoff," in *Proceedings 2000 IEEE International Symposium on Information Theory*, (Sorrento, Italy), June 2000.

[87] I.Bettesh and S.Shamai(Shitz), "Outages, expected rates and delays in multiple-users fading channels," in *Proceedings of the 2000 Conference on Information Science and Systems - Vol I*, (Princeton, New Jersey), March 2000.

[88] G.R.Grimmet and D.R.Strizaker, *Probability and Random Processes*. New York: Oxford University Press, 2nd ed., 1992.

[89] S.Verdú, *Multiuser Detection*. Cambridge University Press, 1998.

[90] W.Kaplan, *Advanced Calculus*. New York: Addison-Wesley, 1991.

[91] G.Caire, G.Taricco, and E.Biglieri, "Optimum power control over fading channel," *IEEE Trans. on Inform. Theory*, vol. 45, pp. 1468–1489, July 1999.

[92] R.Negi, M.Charikar, and J.Cioffi, "Transmission over fading channels with channel state information and delay constraint," in *Global Telecommunication Conference (GLOBECOM'99)*, (???), pp. 2550–2554, 1999.

[93] R.Negi, M.Charikar, and J.Cioffi, "Minimum outage transmission over fading channels with delay constraint," in *IEEE International Conference on Communications 2000 (ICC 2000)*, (???), pp. 282–286, 2000.

[94] A.Fu, E.Modiano, and J.Tsitsiklis, "Transmission scheduling over a fading channel with energy and deadline constraints." submitted to the 2002 Conference on Information Science and Systems, 2002.

[95] K.Sohrabi, J.Gao, V.Ailawadhi, and G.Pottie, "Protocols foe self-organizing of wireless sensor networks," *IEEE Personal Communications*, pp. 16–27, October 2000.

[96] G.Pottie, "Wireless sensor networks," in *ITW 1998*, (Killamey, Ireland, June 22-26), pp. 139–140, 1998.

[97] M.Ahmed and G.Pottie, "Information theory of wireless sensor networks: the n-helpers Gaussian case," in *Proceedings 2000 IEEE International Symposium on Information Theory (ISIT 1998)*, (Sorrento, Italy, June 25-30), p. 436, 2000.

[98] D.P.Bertsekas, *Dynamic programming and optimal control, Vol.1 and Vol.2*. Athena Scientific, 1995.

[99] S.Verdú, "Spectral efficiency in the wideband regime," *submitted to Trans. on Inform. Theory: special issue on Shannon Theory: perspective, trends and applications*, 2001.

[100] A.J.Goldsmith and M.Effros, "The capacity region of broadcast channels with intersymbol interference and colored gaussian noise," *IEEE Trans. on Inform. Theory*, vol. 47, pp. 219 –240, Jannuary 2001.