A Robust and Tractable Analytical Foundation for Radio Resource Management: Centralized and Decentralized Optimizations Involving Data and Media Communication

D I S S E R T A T I O N

Submitted in Partial Fulfillment
of the Requirements for the

Degree of

DOCTOR OF PHILOSOPHY (Electrical Engineering)

at the

POLYTECHNIC UNIVERSITY

by

Virgilio Rodríguez

May 2004

Approved:

_____
Department Head

Copy No. _____

_____
Date

Replace this page with signature page done separately

Acknowledgements

I am indebted to my advisor, Prof. David Goodman, in many respects. I was fortunate that, early on in my studies, he assigned to me a well-defined research project perfectly suited to my skills and interests. That project, which led to chapters 8 and 9 of this document, introduced me to wireless resource management, and allowed me to identify the analytical structure which permeates this entire dissertation. Dr. Goodman graciously allowed me to pursue certain research questions independently, even when it appeared they had only marginal or no relevance to his interests. These pursuits led to the writing of important parts of this documents. He also generously funded my participation in numerous academic conferences, which had a positive impact on this work, and on my overall experience as a doctoral student.

I am also thankful to the other members of my PhD committee, Profs. ZhongPing Jiang, Philip Sarachik, Andrej Stefanov, and Roy Yates, for having graciously agreed to serve on my committee, and provided me with helpful comments and questions. I am especially grateful to Dr. Jiang, for having selflessly contributed crucial funding during the last semester of my studies.

The contributions of my coauthors in some of the publications derived from this work have improved and enriched this dissertation. For instance, the programming wizardry of Zory Marantz was instrumental in the numerical illustrations of chapter 8. Likewise, the video streaming analysis in chapters 1 and 6 benefited greatly from Prof. Yao Wang's input. She also provided helpful comments on other material related to "media" in chapters 5 and 7. Although Dr. Wang is not listed as a member of my PhD committee, in practice she was.

Moreover, I am grateful to those at the Polytechnic University and elsewhere whose fine teaching helped me understand many concepts and methods relevant to this work. I also appreciate many valuable comments and questions I received during my conference presentations, and during my seminars at Bell Labs, and at the universities Stanford, Lehigh, Rutgers, Georgia Tech, and Princeton. Likewise, I appreciate very much the helpfulness and professionalism of our support staff, especially Ms. Valerie Davis.

Last, but by no means least, I am thankful to the many individuals around the world who have altruistically contributed to the excellent "open source" software tools I have used in preparing this document, and its associated papers and slide presentations. These tools include the Linux operating system and its debian "distribution", the associated LaTeX system, the PDFslide LaTeX add-on, and especially, LyX, an elegant LaTeX front-end. LyX gave me access to the power and flexibility of LaTeX, while demanding only a modicum of effort from me.

Vita

From the Fall of 2000, through the Spring of 2004, Virgilio Rodríguez was a Research Fellow and a PhD student with the Electrical and Computer Engineering Department of the Polytechnic University. He worked at the Integrated Information Systems Laboratory, under the supervision of Prof. David J. Goodman, on resource management in the context of wireless communications. Many results from his work have been published at such IEEE conferences as VTC, WCNC, ICASSP, ICC and Globecom, as well as at the CISS and Allerton conferences. Virgilio has received master's degrees in engineering from both Purdue University and the Virginia Polytechnic Institute. At one time or another, he has also benefited from instruction at New York University, the New Jersey Institute of Technology, Northwestern University, and the Instituto Tecnológico de Santo Domingo.

Abstract

Many radio resource optimizations of practical interest share a common analytical core. A framework focused on this core enables robust and tractable analysis, and provides clear answers that apply to a wide variety of physical layer configurations. This framework has three key elements: (i) a tractable abstraction of the human sensory system, (ii) a tractable abstraction of the physical layer of a wireless communication link, and (iii) a fundamental technical result. In these 3 elements, a function about which *all that is known* is that it is sigmoidal, that is, that its graph is an "S-curve", plays a central role. The fundamental result involves the maximization of the ratio f(x)/x.

Fractional programming is the study of the optimization of a ratio of two functions. Current fractional programming literature involves ratios of concave and convex functions. But a sigmoidal function is neither concave nor convex. This work characterizes the maximization of the ratio f(x)/x for any function f having sigmoidal shape. Without imposing any particular algebraic functional form ("equation") on the considered function, this work shows that the maximizer always exists, is unique, and can be graphically described and determined. Additionally, the ratio f(x)/x is shown to be quasi-concave. The maximization of the ratio f(x)/g(x), with g a monotonic function, can be approached by writing this ratio as h(t)/t with t=g(x). If h(t) retains the sigmoidal shape, the preceding analysis can be applied.

This analytical framework is applied to various issues of current interest, involving resource optimization in the context of wireless communications, with emphasis on third-generation cellular systems. The applications include (i) decentralized power control (ii) power and data rate assignment for maximal data throughput when data and media terminals share a CDMA cell, (iii) power and coding rate optimization for the wireless transfer of image or video files, and (iv) choosing an optimal level of media distortion when fidelity is expensive.

Contents

List of Figures

Chapter 1

Introduction and Overview

Many radio resource optimizations of practical interest share a common analytical core. A framework focused on this core enables robust and tractable analysis, and provides clear answers that apply to a wide variety of physical layer configurations. This framework has three key elements: (i) a tractable abstraction of the human sensory system, (ii) a tractable abstraction of the physical layer of a wireless communication link, and (iii) a fundamental technical result. In these 3 elements, a function about which all that is known is that it is a sigmoidal; i.e., that its graph is an "S-curve", plays a central role. The fundamental result involves the maximization of the ratio $f(x)/x$ with f an S-curve. Examples of radio resource optimizations of practical interest to which this framework can be applied include decentralized power control, power and data rate assignment for maximal network throughput, power and coding rate selection for the transfer of media files which have been scalably encoded, and choosing the “right amount” of tolerable media distortion when less distortion means higher cost.

1.1 Overview

The three elements of this framework arise naturally in the context of one of the mentioned applications: power and coding rate selection for video streaming, which is the topic of chapter 6.

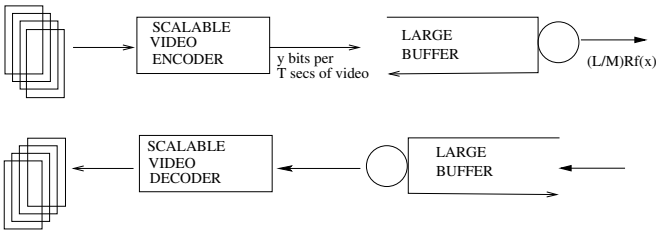

Figure 1.1: Schematic diagram of a system for the streaming of scalably encoded video over a wireless link

Figure 1.1 shows schematically the video streaming system of interest. An energy-limited termi-

nal needs to transfer over a wireless link a “long” video sequence. Each T secs of video is encoded as a fully embedded bit stream, as supported by MPEG standards. This stream is “scalable” in the sense that it can be truncated at an arbitrary point, y, and decoded, leading to various levels of reproduced media quality. The file corresponding to a given segment must be transferred in a deadline of Δ seconds. Files will be split into small packets for transmission purposes and error-control coding bits will be added. Packets received in error which cannot be corrected result in ideal re-transmissions until correctly received and confirmed. Correctly received packets are placed in a large buffer. Transferring each file complete will result in maximal quality per segment, but also in a greater expenditure of energy per file, a hence a shorter battery life. Conversely, transferring few bits per segment will lengthen the battery life, at the expense of possibly unacceptable segment viewing quality. The terminal must jointly optimize both the truncation point of the embedded bit stream (coding rate), and its transmission power.

Performing this joint optimization necessitates three crucial elements: (i) a function $U(y)$ giving the end-user “perceptual quality” or “utility” of a decoded video segment when there are y bits in the corresponding *truncated* file (coding rate); (ii) a function $f_s(x)$ giving the probability of successful reception of a data packet when the signal-to-interference ratio (SIR) at the receiver is x; (iii) a criterion leading to an index to be optimized as function of the quality of individual video segments, and the energy spent per segment.

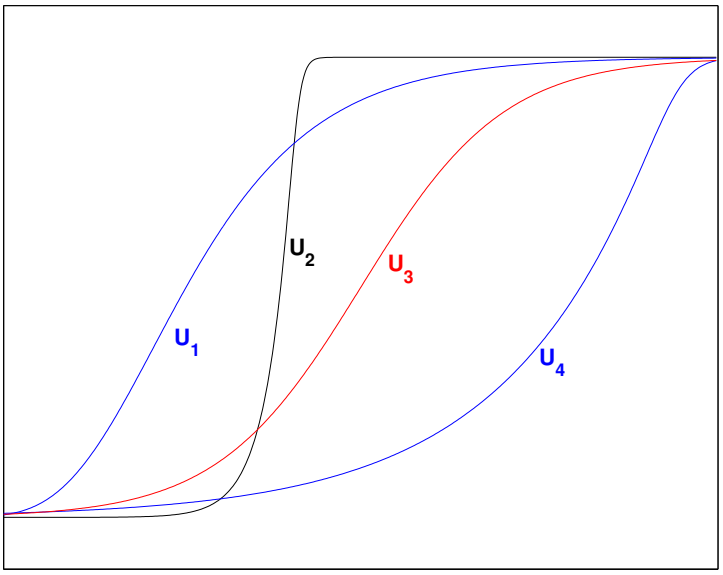

Figure 1.2: Representative S-curves. U_1 is "mostly" concave. U_4 is "mostly" convex. U_2 approximates a "step" function. U_3 includes a "ramp" that follows a straight line over a limited range.

The frame-success function (FSF), which gives the probability that a data packet is received successfully as a function of the terminal’s signal-to-interference ratio (SIR) at the receiver, is determined by physical attributes of the system, including the modulation technique, the forward error

detection scheme, the nature of the channel, and properties of the receiver, including its demodulator, decoder, and antenna diversity, if any. Obtaining an exact expression for this function for a realistic model of a wireless communication setting may be prohibitively difficult or impossible. And even when this function is available, it may be intractable or very inconvenient, and highly dependent on the chosen physical layer configuration. However, one can safely assume that, whatever this function is, its graph is S-shaped, as shown in fig. 1.2. Consequently, the analysis will apply to many physical layer configurations of practical interest, as long as they give rise to an FSF that has an S-shaped graph.

There are additional practical reasons why the S shape may be chosen for modeling a monotonic function of interest. An arbitrary S-curve starts out convex and smoothly transitions to concave. But the inflexion (transition) point is arbitrarily placed. Therefore, as fig. 1.2 shows, this family of curves in fact contains as special cases curves that are "mostly" concave (inflexion point is "very close" to the origin) and others that are "mostly" convex (inflexion point is "very far" from the origin). Furthermore, the "ramp" of an S-curve may be nearly vertical, in which case the curve behaves like a "step" (threshold) function. Or this "ramp" can approximate a straight line, in which case the S-curve expresses a near linear relation over certain range. These shapes should accommodate many situations of interest.

A quality-rate theory is not readily available, but can be arrived at through the concept of distortion. Distortion is typically defined as a relatively simple mean square measure of the difference between a signal and its copy. The properties of any function $D(R)$ giving distortion as a function of coding rate for a given information source are well known. It is generally accepted that the $D(R)$ function is decreasing and convex. The *perceptual* quality of an "imperfect" copy of a signal is determined by the human sensory system (visual, auditory, etc). Common distortion measures behave poorly when distortion is large. However, within certain range it seems reasonable to assume that the perceptual quality is somehow determined by distortion; i.e., that a function $Q(D)$ that translates distortion into perceptual quality can be found. The quality-distortion function cannot be derived, and should not be imposed. It should be obtained by psychophysical experimentation. However, one can make some reasonable assumptions about the properties that any such function should possess. Figure 1.2 shows some plausible basic relations, which are explained in the corresponding caption.

Further reflection indicates that it is reasonable to assume that the graph of the $Q(D)$ function is a "reversed" S-curve, as shown by fig. 1.4. This graph strictly generalizes the step function often assumed in the literature. And, like the family of regular S-curves, this family also includes as special cases curves that are "mostly" convex, others that are "mostly" concave, and some whose "ramps" follow closely a straight line over a given interval. Thus, if the analyst assumes that *all that is known* about the $Q(D)$ curve is that it is a reverse S-curve, and conducts the analysis on the basis of properties derived from this shape, the solution procedure and conclusions will be valid for a wide variety of plausible $Q(D)$ relations.

With $Q(D)$ denoting the reversed S-curve giving *perceptual* quality as function of distortion, it

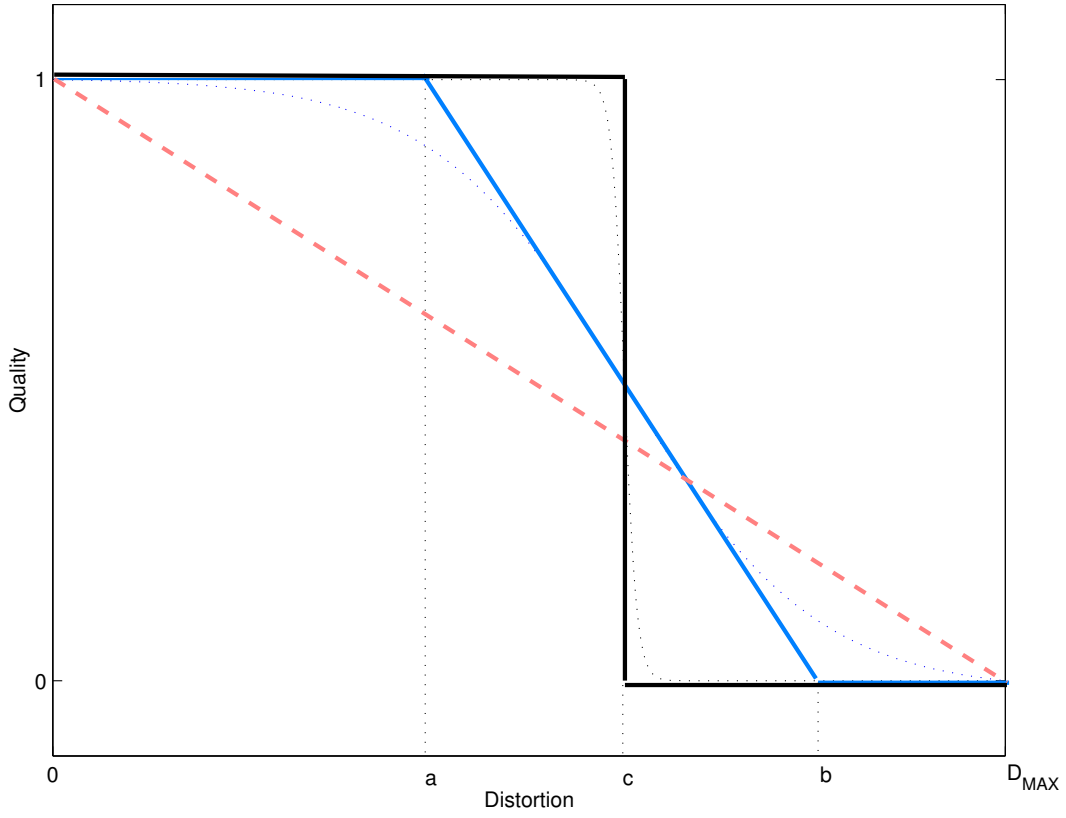

Figure 1.3: Quality vs. distortion: Some plausible simple relations are: (i) fidelity equals quality (red dashed line); (ii) hard threshold (step); (iii) ramp (blue broken line). The ramp includes as special case the threshold ($a = b = c$) and the linear relation ($a = 0$, $b = D_{\mathrm{MAX}}$). But as shown by the next figure, the reverse S-curve includes all of these cases and more.

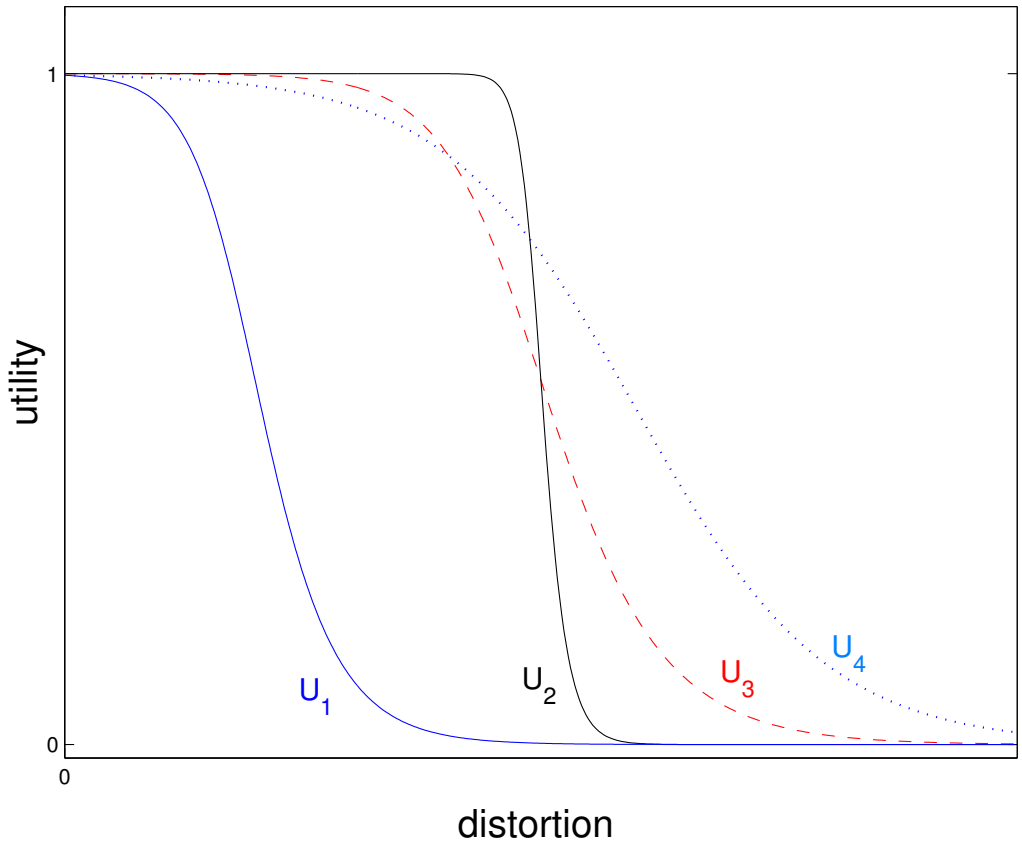

Figure 1.4: Perceptual quality ("utility") as a function of distortion: The family of (reversed) S-curves generalizes the step function often assumed in the literature, and includes many interesting special cases.

is clear that the composite function $Q(D(R)) := U(R)$ yields perceptual quality directly as a function of the coding rate. It is then of interest to characterize the composite function $Q(D(R))$ when *all that is known* about $D(R)$ is that it is decreasing and convex, and *all that is known* about $Q(D)$ is that it is a "reversed" S-curve. The caption of fig. 1.5 contains an approximate analysis that suggests that the graph of $U(R) = Q(D(R))$ is a (non-reversed) S-curve, as displayed in fig. 1.2. Figure 1.6 confirms this conjecture for specific $Q(D)$ curves, and the $D(R)$ function of the memoryless Gaussian source.

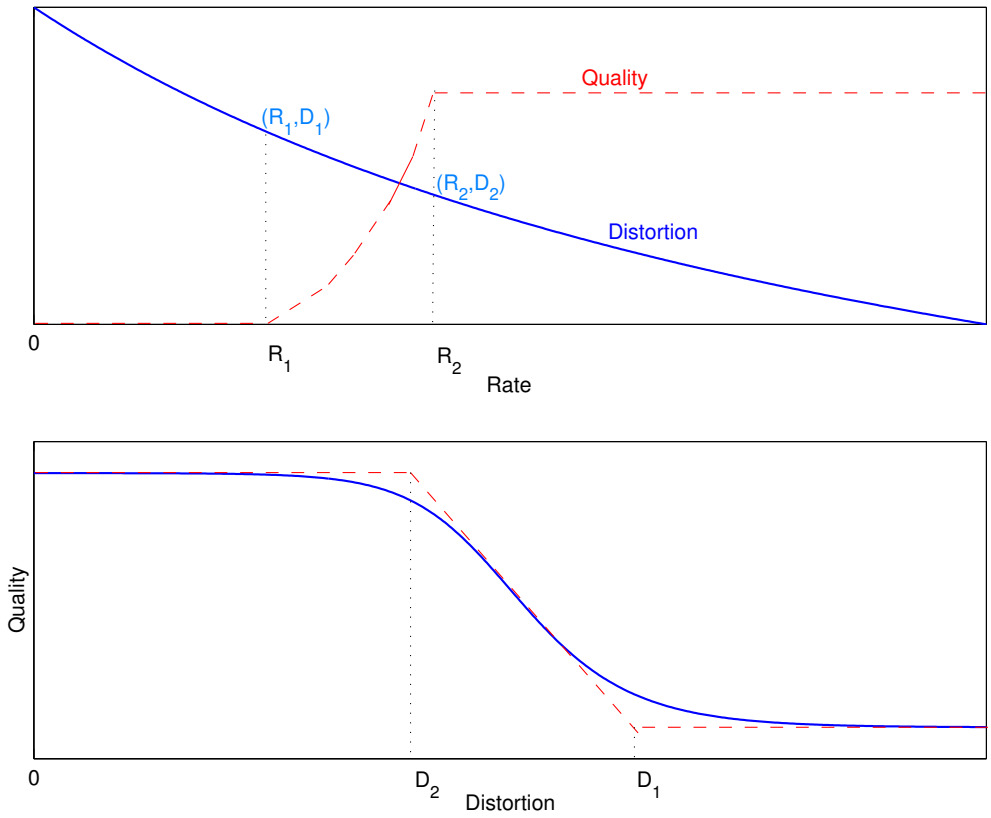

Figure 1.5: The convex curve at the top corresponds to $D(R)$, which relates coding rate to distortion. The S-curve at the bottom, $Q(D)$, relates perceptual quality to distortion. The composite function $Q(D(R)) := U(R)$ yields perceptual quality directly as a function of the coding rate. Some reflection indicates that if $Q(D)$ is approximated by the broken red line at the bottom, the resulting $U(R)$ is the broken red line at the top. For Q a reversed S-curve, we expect $U(R) = Q(D(R))$ to yield an (increasing) S-curve.

At this point, of the three items identified previously as key to the analysis, two has been found: both the FSF and the quality-rate function can be taken to be S-curves. The index to be optimized needs to be defined. A reasonable objective is to maximize the total perceptual quality (utility) that gets transferred by the time energy runs out; that is, to maximize $n \times u(y)$, where $n = E/c(y)$, with E the available energy, and $c(y)$ the energy cost of successfully transmitting a y-long file in Δ secs. This is equivalent to maximizing $u(y)/c(y)$ (subject to an appropriate constraint), or perceptual quality per Joule.

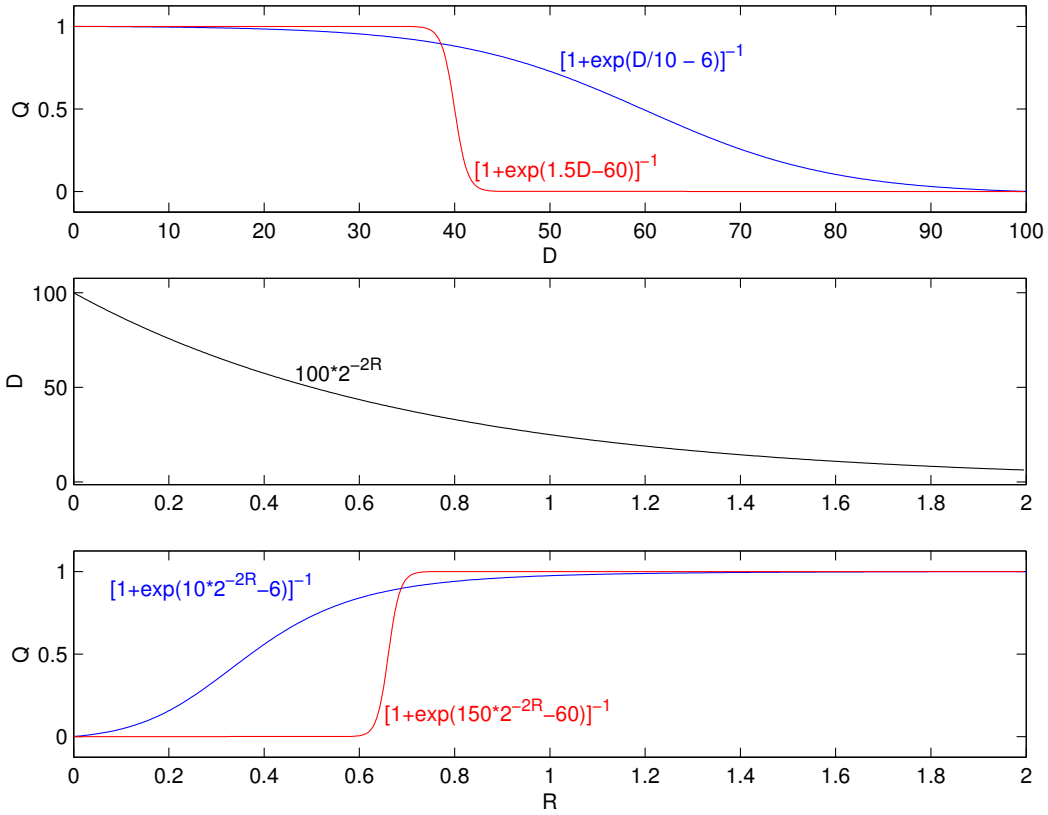

Figure 1.6: The preceding analysis can be applied to the specific case of a memoryless Gaussian source, whose $D(R) \propto 2^{-2R}$ (2nd subplot). At the top, there are two plausible quality-distortion curves. At the bottom are the graphs of the composite functions $Q(D(R)) := U(R)$ corresponding to the Q functions at the top, and the Gaussian rate-distortion curve. As the preceding analysis suggests for the general case, in these examples the graphs of $Q(D(R))$ are (increasing) S-curves.

The detailed analysis of the maximization of $u(y)/c(y)$ is found in chapter 6; an outline is presented now. First, it is easily established that, on the average, $(L/M)Rf(x)\Delta$ information bits are correctly received in Δ secs. L/M is the ratio of information bits to the packet length, R is the raw data rate of the terminal, and f is a slight modification of the FSF. For given y and Δ, there is a specific SIR $x(y)$ that satisfies $(L/M)Rf(x)\Delta = y$, and there is a specific transmitted power, $P(y)$, that yields $x(y)$. Thus, the total number of T-sec video segments of quality $u(y)$ that can be transferred with an energy budget of E is $E/(P(y)\Delta)$. The total quality viewed is $(E/\Delta)u(y)/P(y)$. For fixed E and Δ, it is sufficient to maximize $u(y)/P(y)$.

Thus, the terminal must solve:

$$\begin{array}{rcl} \max\limits_{x,y} \dfrac{u(y)}{x} & & \max\limits_{x} \dfrac{u(Bf(x))}{x} \\ \text{s.t. } y = Bf(x) & \text{OR} & \text{s.t. } 0 \le x \le \bar{x} \\ 0 \le x \le \bar{x} & & \end{array}$$

with $B = (L/M)R\Delta$ interpreted as the maximum amount of information bits ("best case scenario") that can be transferred in the deadline Δ, and x the SIR.

u and f are both S-curves. The composite function $h(x) := u(Bf(x))$ is expected to retain the S-shape. Hence, in order to solve this problem, the solution to maximizing $h(x)/x$ when *all that is known* about h is that it is an S-curve needs to be found.

1.2 Content and Organization

The research reported herein has several "branches" that were pursued quasi-independently. Most chapters started as self-contained papers. While an effort has been made to integrate the various papers into a coherent report, the document still has the "flavor" of an edited collection of papers. While redundancy and multiplication of information do exist, intentionally or otherwise, an advantage of this fact is that chapters are largely self-contained. Pertinent literature is reviewed in each chapter.

This work continues by investigating the maximization of the ratio $f(x)/x$ when *all that is known* about f is that its graph "starts out" convex at the origin, and "smoothly" transitions to concave as it approaches a horizontal asymptote. Problems involving the optimization of ratios of functions have been intensively studied in the last few decades, and are commonly called "fractional programming". These problems arise naturally in many contexts, including macroeconomics, finance, inventory control, and numerical analysis, among others. Reference [5] is a very recent survey of this literature. However, the most general formulations studied in this literature involve ratios of concave and convex functions. In a few cases, the definitions of concavity and/or convexity are relaxed to include a somewhat larger class of functions. But, the sigmoidal functions studied herein are, by definition, neither concave nor convex (very loosely speaking they are "half and half"),

and are, therefore, excluded from the current fractional programming literature. Without imposing any particular algebraic functional form ("formula") on the considered functions, chapter 2 shows that the solution to this maximization problem always exists, is unique, and can be graphically described and determined. A tangent line drawn from the origin to the graph of f specifies the optimal solution. Additionally, the ratio $f(x)/x$ is shown to be quasi-concave.

The remainder of this work applies the basic analytical core afforded by sigmoidal fractional programming to various issues of current interest, involving the optimization of power, data rate, and coding rate in wireless communications, with emphasis on third-generation cellular communication systems.

Decentralized power control in a multiple-transmission-rate scenario relevant to third-generation wireless networks is studied first. Chapter 3 addresses the critical aspect of specifying a well-behaved, sensible quality-of-service (QoS) index ("utility function") for a data terminal to maximize. An index that exhibits solid technical behavior, is physically significant, intuitively appealing, and applicable to a wide variety of physical layer configurations is proposed. Subsequently, in chapter 4, decentralized power control is set up as a "game" in which each data transmitting terminal maximizes its QoS. Closed-form Nash equilibrium conditions and power levels are derived "from first principles". It has been known for some time that Nash-equilibria are generally "inefficient". In fact, when each data-transmitting wireless terminal chooses a transmission power level to maximize a sensible QoS index, they settle on equilibrium power levels that are "too high". The challenge is to induce the terminals to move toward a more efficient operating point in a decentralized fashion. This chapter proposes a relatively simple mechanism, available in the economics literature, to achieve an efficient decentralized allocation of power.

The next chapters focus on media transmission. Chapter 5 analyzes resource management involving scalably encoded information. Scalable encoders, as that of the JPEG 2000 standard, produce files which can be truncated at an arbitrary point and decoded. An energy-efficient policy for the transmission over a wireless network of scalably-encoded images is found. At the core of the analysis is an "S-curve" yielding a measure of "quality" of the decoded information as a function of the "truncation point" (coding rate). Transmission power, and the coding rate are jointly optimized. The single-user case is fully analyzed, and a closed-form solution given, which can be clearly identified, graphically. The analysis leads to the maximization, over an appropriate region, of the product $kf(x)/x \times u(y)/y$, where x is the received SIR, f is the frame success function, y is the chosen number of decoded bits, and u is the "quality" function. $kf(x)/x$ has the unit bits/Joule, while quality/bit is the unit of $u(y)/y$. Thus, the maximized product is an intuitively appealing index in quality/Joule.

Chapter 6 extends the analysis of the preceding chapter to the more interesting case of scalable video streaming (this is the sample application discussed in the present chapter). The analysis leads to the maximization of the quality-to-power ratio, which is equivalent to maximizing quality per Joule. Although the problem is set up as a joint optimization of power and coding rate, the analysis indicates that any one of these two variables fully determines the other, when the underlying

streaming application constrains the transmission time. With $u(y)$ the quality of a video segment as a function of the coding rate, f the frame success function, and B certain constant, the terminal should choose its transmission power so that the received SIR x maximizes the ratio $g(x)/x$, with g a composite function of both S-curves, $u(Bf(x))$.

The quality-distortion curve introduced in the present chapter can also be interpreted as a "utility function" giving the "usefulness" to an observer of an "imperfect" signal. A key difference between perceptual quality and "utility" is that utility is application-dependent. For instance, for a given observer, a level of distortion deemed "unbearable" for a "serious" application, may be perfectly acceptable (to the same observer) in a less "serious" situation. Chapter 7 takes the "utility" point of view. A "utility function" on distortion, assumed to be a reversed S-curve, mathematically captures the idea that media signals can be useful to humans at various degrees of noticeable distortion. When less distortion means a higher cost, an end-user may prefer more distortion, in exchange for energy, money or other savings. In chapter 7, two problems related to this issue are analyzed. First, a consumer with a limited budget can acquire more media files, by accepting more distortion per file. The amount of distortion that maximizes *the sum* of the utility of each purchased file is found and clearly identified in the graph of the utility function. Second, an energy-limited transmitter with many media files to transfer can, statistically, reduce distortion per file, at the expense of fewer transferred files. A solution that maximizes his *total expected* utility is given through the graph of the *expected* utility as a function of the received SIR. Because the proposed family of utility functions contains as a special case the step function typically assumed by the literature, this formulation adds to the literature, and takes nothing away.

All three analyses involving media files can be extended to consider multiple terminals, through the application of game theory, as done in chapter 4. In fact, this is the reason why CDMA quantities are used in defining the signal-to-interference ratio (SIR). For a more general analysis, one can replace (in the single-terminal situation) the SIR with the familiar ratio E_b/N_0, with the numerator denoting energy per bit, and the denominator denoting "noise energy".

The situations discussed so far focus on the terminal/user. That is, the analysis seeks the best allocation from the standpoint of the terminal, as opposed to the network's administrator or owner. By contrast, chapter 8 seeks centralized power and data rate allocations in order to maximize the cell *weighted* throughput. This setting is relevant to the uplink of a variable spreading gain (VSG) CDMA cell, a technology capable of accommodating multi-rate traffic, which is supported by third-generation standards. A weight is associated with the throughput of each terminal. The weights admits various practical interpretations, including per-bit utility, priority, or unit price paid to the network by the user. The traffic is assumed delay tolerant, and the cell is assumed interference-limited (out-of-cell interference and random noise are deemed negligible). First, a two-terminal-only scenario is fully solved. This special case establishes the terminology and the solution procedure, and provides a great deal of intuition. Subsequently, the analysis is extended to an arbitrary number of terminals. A main conclusion of the analysis is that at least one terminal should operate at the highest available data rate, and that terminals *not* operating at this rate should operate at the same

signal-to-interference ratio (SIR), a value that maximizes the ratio $f(x)/x$, with f a slight modification of the frame-success function. The development in this chapter describes a solution procedure leading to the global optimizer, for the special case in which only two weights are considered.

Chapter 9 discusses how to extend the preceding model to consider three additional items: (i) transmission power limits, (ii) non-negligible out-of-cell interference, and (iii) the presence of media-transmitting terminals with fixed bit rates and inflexible SIR requirements. Power limitations are important for obvious reasons. However, when out-of-cell interference is negligible, the power allocation question reduces to finding a vector of carrier-to-interference ratios involving the received powers of the terminals. The specific power levels are, in theory, arbitrary. However, when the noise term includes strong out-of-cell interference, the values in absolute terms of the power levels are important, and the power limitations of the terminals need to be taken explicitly into account. Additionally, there may be media-transmitting terminals operating at fixed bit rates and SIR. These media terminals can be thought of as additional sources of "noise", which decrease the total throughput of the data terminals. Chapter 9 focuses on the interaction of a power-limited media terminal, with two data terminals, one of which is more "important" than the other. The aim is to show that much of the analysis of the preceding chapter can still be applied, with relatively minor modifications, to the more complicated and realistic situation of this chapter.

Chapter 10 discusses some of the general limitations of this work, suggests extensions and related topics for future research, and highlights some of the main contributions.

The appendices provide various technical results. Appendices C and D are of special note. In appendix C, the procedure used to find the equilibrium allocation of the game of chapter 4 is extended to address a somewhat more general issue: power allocation when terminals SIR requirements are "elastic". That is, each terminal has a preferred or optimal SIR value, but is willing and able to operate at lower values. The proposed procedure maximizes the number of terminals operating at their preferred SIRs, subject to the constraint that no terminal be sacrificed to help another. Closed-form analytical expressions are provided through the development.

Appendix D focuses on macro-diversity, a scheme in which the cellular structure of a CDMA system is removed and each transmitter is jointly decoded by all "receivers". This scheme has been shown to increase the capacity of CDMA wireless networks. The available macrodiversity capacity results rely on a "self-interference" approximation, which may *not* be appropriate for 3rd generation cellular systems. Explicitly considering power constraints, and without resorting to this approximation, this appendix applies well established mathematical results to derive capacity results that are less conservative than those previously available.

Chapter 2

Sigmoidal Fractional Programming

2.1 Introduction

Sigmoidal functions are particularly useful, having played a fundamental role in the modeling of a wide variety of interesting phenomena in the physical, biological and social sciences. One reason for their ubiquity is that the graph of the solution to the differential equation $x'(t) = rx(t)(1-x(t)/k)$ has the sigmoidal shape ("logistic growth"). This equation, which arises naturally in many dynamical systems, was introduced in [44], in the context of population growth. In this context, $x(t)$ denotes the size, at time t, of certain population, whose instantaneous growth rate is directly proportional to both its current size, and the difference between this size and the environment's "carrying capacity" (maximal sustainable population size), k. Reference [28] introduces the generalization $x'(t) = rx(t)[1-(x(t)/k)^{\beta}]$, and argues its usefulness; and [25] describes the statistical fitting of the four-parameter family of curves introduced by [28]. A recent survey, [39], discusses other generalizations, and introduces its own. Reference [22] argues that sigmoidal functions may be even more useful than traditionally thought, because in many interesting situations, a complex process whose growth behavior may not seem sigmoidal, can be fruitfully modeled via the superposition of various sigmoidal functions in a single model. This reference provides examples or suggests applications of this approach in many domains, including ecology, psychology, and socio-technological inquiries. Likewise, in computing, sigmoidal functions have played the important role of "activation functions" of processing elements in artificial neural networks.

In the previously mentioned studies, the sigmoidal curves are tied to specific algebraic functional forms ("equations") arising as a solution to certain differential equations. The analysis in this chapter significantly differs from the literature in that the S-curves studied herein are *not* described in algebraic terms. The curves are described geometrically, and the analysis follows from properties derived from their shape.

This chapter focuses on the maximization of the ratio $f(x)/x$, for any real-valued, univariate function f having the specified sigmoidal shape. This ratio may admit different interpretations depending on the context. For example, if $x(t)$ is associated with the "logistic growth" of certain process, the ratio $[x(t)-x(0)]/t$, the average growth rate at time t, has the form of the ratio be-

ing studied here. More concretely, many radio resource optimizations of practical interest depend critically on the maximization of an expression of the form $f(x)/x$. The specific function and its argument depend on the problem being analyzed; but f is, typically, monotonic, and can be assumed to be a member of the family of "S-curves", for reasons given in chapter 1. Some specific applications are mentioned in chapter 1, and are discussed in detail in subsequent chapters.

It has also been mentioned in chapter 1 that problems involving the optimization of ratios of functions arise naturally in many contexts, and have been intensively studied in the last few decades, under the name "fractional programming"[5]. But, the sigmoidal functions studied here are excluded from the current fractional programming literature, because, by definition, they are neither concave nor convex.

This chapter analyzes the "context-free" maximization of the ratio $f(x)/x$ for any function f having the specified sigmoidal shape, and characterizes the optimal solution strictly in terms of geometrical properties derived from this shape. Specifically, without imposing any particular algebraic functional form on the considered functions, this chapter shows that, under the assumptions herein, the solution to this maximization problem exists, is unique, and can be graphically described and determined. Additionally, the ratio $f(x)/x$ is shown to be quasi-concave.

Below, the considered class of functions is formally characterized. Then, the solution to the maximization problem of interest is derived. Subsequently, the quasi-concavity of the ratio is established. Finally, some closing comments are given. Appendix A reproduces or fully develops certain key technical results.

2.2 Formalization of the functions of interest

2.2.1 Basic Assumptions

Figure 2.1 provides a graphical illustration of a function representative of the class of functions to be considered. Any such function, f, has the following characteristics:

1. Its domain is the non-negative part of the real line; that is, the interval $[0,\infty)$
2. Its range is the interval $[0,B)$, where, for convenience, and without loss of generality, we take $B=1$.
3. It is increasing.
4. ("Initial convexity") It is strictly convex over the interval $[0,x_f]$, with x_f a positive number.
5. ("Eventual concavity") It is strictly concave over any interval of the form $[x_f,L]$, where L is a positive number greater than x_f
6. It has a continuous derivative.

Notice that *no* assumptions about the second derivative of the function f are explicitly made.

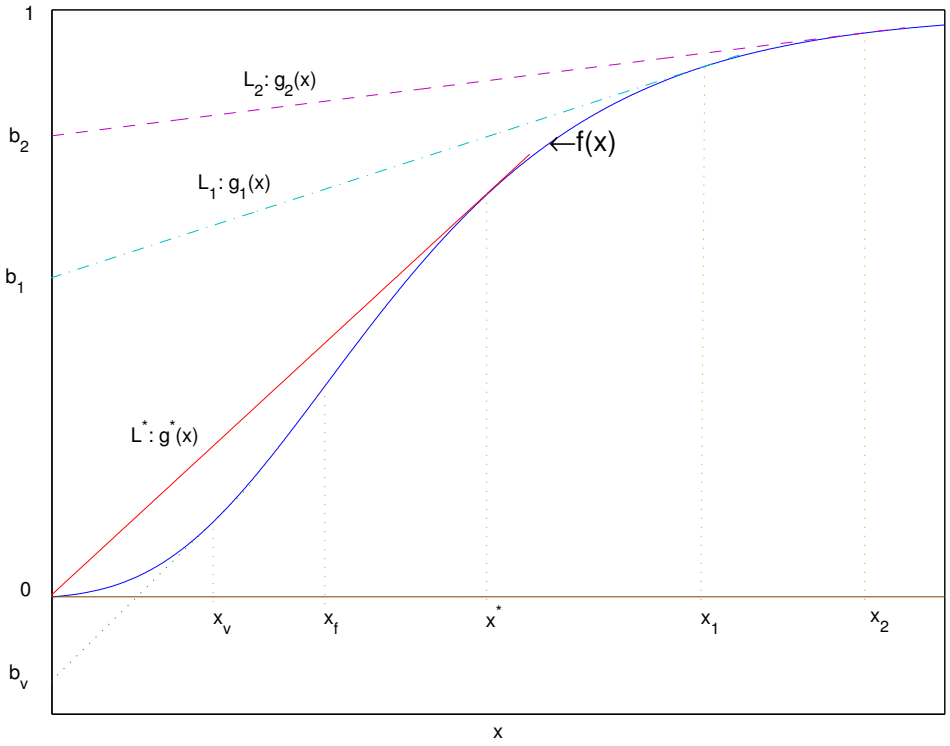

Figure 2.1: A representative function and some of its tangents

2.2.2 Immediately Implied Characteristics

1. Assumptions (1), (2) and (3) imply that $f(0) = 0$.

2. Assumptions 4 ("initial convexity") and 5 ("eventual concavity") imply that the function is continuous for any $x > 0$. (See Theorem 1.3, Chapter III, in reference [2]). And this implication, together with the preceding one further imply that f is continuous overall.

3. The "initial convexity" assumption 4 and the continuous derivative assumption 6 together imply that $f'(0) < \infty$ (See subsections A.2.1 and 2.3.2.1). This ensures that $\lim_{x\to 0} f(x)/x$ is finite, by L'Hopital rule

4. Assumption 6 also implies the continuity of f .

2.3 Maximization

Below, the following optimization problem is solved:

$$\text{Max: } f(x)/x \text{ subject to } 0 \leq x \leq M$$

2.3.1 An interior solution

First, it is presumed that a "stationary" point exists within the allowable range of x.

2.3.1.1 First-order conditions for a maximum

The first-order necessary conditions are:

$$f(x) - xf'(x) = 0 \tag{2.1}$$

It will prove useful to observe that the equation of a straight line tangent, at the point $(x_1, f(x_1))$, to the curve described by the graph of the function f can be written as

$$g_1(x) = f(x_1) + f'(x_1)(x - x_1) \text{ or } g_1(x) = b(x_1) + f'(x_1)x \tag{2.2}$$

where $b(t) := f(t) - tf'(t)$ represents the ordinate at the origin (y-intercept) of the straight line tangent at the point $(t, f(t))$ to the curve described by the graph of f (see fig. 2.1). Therefore, equation (2.1) can be stated as $b(x) = 0$, which is discussed further in section 2.3.1.5.

2.3.1.2 Existence of a Solution

A solution to equation (2.1) always exists. This follows from these facts:

i) $b(x) = f(x) - xf'(x)$ is a continuous function.

ii) For sufficiently large x_L, $b(x_L) > 0$

iii) For any x_v in $(0, x_f]$, $b(x_v) < 0$.

Statement (i) follows directly from the fact that both $f(x)$ and $f'(x)$ have been assumed to be continuous.

Statement (ii) is a direct consequence of the fact that, by assumption, $\lim_{x\to\infty} f(x) = 1$. Hence, in the limit, the tangent line to the graph of f is the line $y = 1$. The y-intercept of this line is, of course, 1. So, $\lim_{x\to\infty} b(x) = 1$, for which $b(x)$ is bound to take on positive values "sooner or later".

Statement (iii) follows from the essential property of tangent lines of continuously differentiable strictly convex functions (see section A.2.1). Over the interval $[0, x_f]$, f is assumed to be strictly convex. Taking $x_2 = 0$ and x_1 equal to an arbitrary number in $(0, x_f]$, denoted as x_v, inequality (A.3) yields $f(0) > f(x_v) + f'(x_v) \cdot (0 - x_v)$ or, equivalently, $b(x_v) = f(x_v) - x_v f'(x_v) < 0$.

Statements (i), (ii), and (iii) above have been shown to be valid. These three facts imply the existence of an x^*satisfying $b(x^*) = 0$, because a continuous function cannot go from a negative to a positive value without taking on the value zero.

Furthermore, notice that the validity of statement (iii) immediately implies that any such x^*must be greater than x_f (that is, any such x^*must be in the interval over which f is concave), since, $x < x_f \to b(x) < 0$.

2.3.1.3 Uniqueness of the solution

In subsection 2.3.1.2 it was established that any solution to $b(x) = f(x) - xf'(x) = 0$ must must lie inside the interval where f is strictly concave. The uniqueness of this solution follows directly from the "monotone intercepts" corollary, presented in subsection A.2.2. This results indicates that if x_1 and x_2 are points in an interval of the real line over which the function f is strictly concave, then $x_2 > x_1$implies that $b(x_2) > b(x_1)$. Hence, if $x^*$is such that $b(x^*) = 0$, any $x \neq x^*$must be such that $b(x) \neq 0$

2.3.1.4 Optimality of the solution

The derivative of the ratio $f(x)/x$ can be expressed as

$$\frac{xf'(x) - f(x)}{x^2} = -\frac{b(x)}{x^2} \tag{2.3}$$

with $b(x)$ as previously defined. The derivative is well-defined with the possible exception of the boundary value $x = 0$. The case $x = 0$ is discussed in the subsection 2.3.2. For the purposes of this section, x is assumed to be positive.

The monotone intercepts corollary of subsection A.2.2 specifies that for any $x > x^*$, $b(x) > b(x^*) = 0$. Therefore, the ratio $f(x)/x$ is strictly *decreasing* for any $x > x^*$.

The same argument leads to the conclusion that the ratio $f(x)/x$ is strictly *increasing* for any $x_f < x < x^*$.

In subsection 2.3.1.2 it was established that $b(x) < 0$ for any x in $(0, x_f]$. Therefore, the derivative of the ratio $f(x)/x$ is positive for any such x, (see equation 2.3 above), which means this ratio is increasing over $(0, x_f]$.

In conclusion, the ratio $f(x)/x$ is less than $f(x^*)/x^*$ for any positive $x \neq x^*$.

2.3.1.5 Description of the solution: The characteristic tangent

The solution to the first-order necessary optimizing conditions given by equation (2.1) can be directly identified in the graph of the function f. Only one positive value, x^*, satisfies equation (2.1). $(x^*, f(x^*))$ is the only point at which a line tangent to the curve describing the function passes through the origin. Thus, the equation of any such tangent line is $g^*(x) = f'(x^*)x$. (See the tangent line drawn at x^* in fig. 2.1). This tangent line is termed "the *characteristic tangent*" of a given sigmoidal function. Of course, different sigmoids may have the same characteristic tangent.

The value of the objective function at the solution, x^*, can be obtained graphically as the slope of the characteristic tangent, which is $f(x^*)/x^*$. This observation can be useful for conceptual "sensitivity analyses". The effect on the optimal solution of changing one sigmoid for another (for example via a change in certain parameter) immediately manifests itself, visually, through the new characteristic tangent, and its slope.

2.3.2 "Boundary" solution

The development so far has ignored the constraint that $x \leq M$ for some M. Below, this issue is addressed. Before that, the possibility that the optimal value be zero is formally discarded.

2.3.2.1 The non-optimality of x=0

By construction, and the application of L'Hopital rule, $\lim_{x\to 0} f(x)/x = f'(0) < \infty$. In sub-sections 2.3.1.2 and 2.3.1.4 it was discussed why the ratio $f(x)/x$ is increasing over the interval $(0, x_f]$. Hence, $x = 0$ is *not* the maximizer.

2.3.2.2 The global optimality of the smallest of M and x^*

Given the discussion in subsections 2.3.1.4 and 2.3.2.1, it is clear that the ratio $f(x)/x$ is increasing over the interval $[0, x^*]$, where x^* is the only value of x satisfying the first-order necessary optimizing conditions given by equation (2.1). Hence, if the maximum allowable value for x, denoted as M, is less than x^*, $f(M)/M$ is the highest achievable value for the ratio $f(x)/x$. But if x^* is less than M, $x = x^*$ is clearly the optimizing choice. Therefore, the smallest of the numbers M and x^* is the global maximizer.

2.4 The Quasi-concavity of $f(x)/x$

In the preceding development, it has been determined that, for the class of functions under consideration, the ratio $f(x)/x$ is "single-peaked"; that is, there is a number x^* such that this ratio is strictly increasing for all $x \in [0, x^*)$ and strictly decreasing for all $x \in (x^*, \infty)$. This implies the quasi-concavity of this ratio. For a general discussion about quasi-concavity and various related concepts and results, see [27].

Below, the definition of quasi-concavity is given, and the compliance of $f(x)/x$ with this definition is formally established.

2.4.1 Definition of Quasi-concavity

Definition: The function $h : I \to R$, defined on an interval $I \subset \Re$, is said to be quasi-concave if its upper contour sets, $\{x \in I : h(x) \geq t\}$, are convex sets; that is, for any $t \in \Re$, any $\alpha \in [0, 1]$, and any $x_1, x_2 \in I$, $h(x_1) \geq t$ and $h(x_2) \geq t$ imply that

$$h(\alpha x_1 + (1-\alpha)x_2) \geq t \tag{2.4}$$

The function h is said to be *strictly* quasi-concave if the implied inequality in (2.4) holds strictly whenever $x_1 \neq x_2$ and $\alpha \in (0, 1)$.

2.4.2 Verification of Quasi-concavity

The function $f(x)/x$ is strictly quasi-concave.

Proof:

For notational convenience, let $h(x) \doteq f(x)/x$ and let $h(x^*) \doteq P^*$.

Let $t \in (0, P^*)$. Notice that verifying (2.4) is trivial for t outside this interval.

Suppose $0 \leq x_1 < x_2$, $h(x_1) \geq t$ and $h(x_2) \geq t$

Because $h(x)$ is continuous and strictly *increasing* in the interval $[0, x^*)$, there is an x_t' such that $h(x) \geq t$ for all x between x_t' and x^*, and $h(x) < t$ for $x < x_t'$. Likewise, since $h(x)$ is continuous and strictly *decreasing* in the interval (x^*, ∞), there is an x_t'' such that $h(x) \geq t$ for all x between $x^*$and x_t'', and $h(x) < t$ for $x > x_t''$.

Then, clearly, any x for which $h(x) \geq t$ must be between x_t' and x_t'', and any x between x_t' and x_t'' is such that $h(x) \geq t$. That is, $x_t' \leq x \leq x_t'' \Leftrightarrow h(x) \geq t$.

Therefore, $h(x_1) \geq t$ and $h(x_2) \geq t$ implies $x_t' \leq x_1 < x_2 \leq x_t''$

And for $\alpha \in (0,1)$, $x_1 < \alpha x_1 + (1-\alpha)x_2 < x_2$. This implies $x_t' < \alpha x_1 + (1-\alpha)x_2 < x_t''$, which further implies $h(\alpha x_1 + (1-\alpha)x_2) \geq t$

Q.E.D.

2.5 Discussion

The maximization of the ratio $f(x)/x$ for any function f having a “sigmoidal” shape has been studied, and its optimal solution been characterized without imposing any particular algebraic functional form (“equation”) on the considered functions. “Sigmoidness” has been captured in a strictly geometric manner, by assuming that the considered functions “start out” convex at the origin, and “smoothly” transition to concave as they approach a horizontal asymptote. This *geometric* construction had not been found in the scientific literature, although sigmoidal functions have been studied in numerous contexts, including in technological, biological and socio-economic environments. On the basis of geometrical properties derived from this shape, this note shows that the solution to the maximization problem of interest always exists, is unique, and can be graphically described and determined.

The graphical identification of the solution could be valuable as a conceptual tool to understand the meaning of the solution, as well as a “sensitivity analysis" tool, to visualize how a change in the considered function can impact the optimal solution.

Central to the development and fully developed herein, the observation that the “y-intercepts” of concave and convex functions are monotonic may be useful beyond the particular aims of this work.

Along the way, the ratio $f(x)/x$ has been shown to be quasi-concave, which is by no means obvious given the arbitrary sigmoidal shape of the function in the numerator. This fact can be beneficial in situations in which this maximization is embedded into a larger problem, as in the

"game" discussed in chapter 3 and in references [6, 20, 33], where certain important results (such as Debreu's "general equilibrium" theorem) can be invoked because of the quasi-concavity of this ratio.

Through the remainder of this work it will be made clear that the maximization of a ratio of the form $f(x)/x$, with f some "S-curve", is particularly relevant to several important problems involving resource management for data communication over a wireless medium. This includes decentralized power control, power and data rate assignment for maximal network throughput in a 3G-CDMA context, and resource management for scalably-encoded visual information, as with the JPEG-2000 and MPEG-4 standards.

Chapter 3

Robust Modeling for Wireless Data

3.1 Introduction

Several recent scholarly publications, following an approach suggested by Ji, [12], recognize that algorithms useful for engineering applications can be obtained via the formulation of radio resource management issues, in particular power control in wireless data applications, on the foundations of microeconomic theory (References [6] and [20] are recent surveys of this literature). This approach is centered around the notion of a quality-of-service (QoS) index, often referred to, by analogy with economics literature, as a "utility function", defined as a real-valued function of certain physically-significant quantities. Algorithms are designed seeking the maximization, under appropriate rules and constraints, of the utility of each transmitter.

Utility maximization in a practical setting need *not* involve a human user instantaneously choosing utility-maximizing levels of resources during transmission. Rather, it may be implemented by software inside transmitting terminals. Depending upon the service agreement, a human "customer" may or may not have control of the embedded program. This observation is important because an inappropriate QoS index may lead the terminal to behave in a manner inconsistent with human intelligence.

Utility maximization, like other radio resource optimizations of practical interest, depends critically on a function giving the probability of the correct reception of a data packet in terms of the signal-to-interference ratio (SIR) at the receiver. This "frame-success" function (FSF) is determined by physical attributes of the system, including the modulation technique, the forward error detection scheme, the nature of the channel, and properties of the receiver, including its demodulator, decoder, and antenna diversity, if any. It may be prohibitively difficult or impractical to obtain and/or work with an exact expression of this function. Therefore, functions corresponding to highly simplified situations are often utilized in analytical studies. Regrettably, the obtained results may only be valid for the rare situations for which the assumed functional form is appropriate.

In view of the above, it is highly desirable that analytical studies be based on generalized frame-success/utility functions, whose assumed characteristics match most realistic situations. Results obtained on the basis of such "generic" functions would then be robust, in the sense that they would

apply to a wide variety of physical layer configurations and practical situations.

Perhaps the only non-trivial feature which can be assumed to match most, if not all, frame-success functions of practical interest is “sigmoidness”; that is, the graph of any such function is S-shaped. Below it is assumed that the frame-success function f_s of interests obeys the technical properties of the generalized S-curve discussed in greater detail in chapter 2.

Another critical issue is specifying an appropriate utility function, which is the QoS index whose maximization is assumed to be sought by each user. Below, after discussing other such indices available in the literature, the *earned*-throughput-to-power ratio (ETPR) is discussed. As a QoS index, the ETPR is shown to exhibit good mathematical behavior, be physically significant, attain or surpass the intuitive appeal of related measures already accepted by the scientific literature, and, perhaps more significantly, be defined for arbitrary frame-success functions of practical interest. In chapter 4, a game in which terminals with dissimilar data rates choose transmission power seeking to independently maximize their respective ETPR is analyzed.

3.2 A Generalized “frame-success” function (FSF)

It is assumed that the function f_s, which gives the probability of the successful reception of a transmitted data packet in terms of the signal-to-interference ratio (SIR) at the base station, is such that the related function f defined by $f(x) = f_s(x) - f_s(0)$ obeys the general properties of the generalized sigmoidal function discussed in 2.2. It is further assumed that f_s (and hence f) has a continuous second derivative.

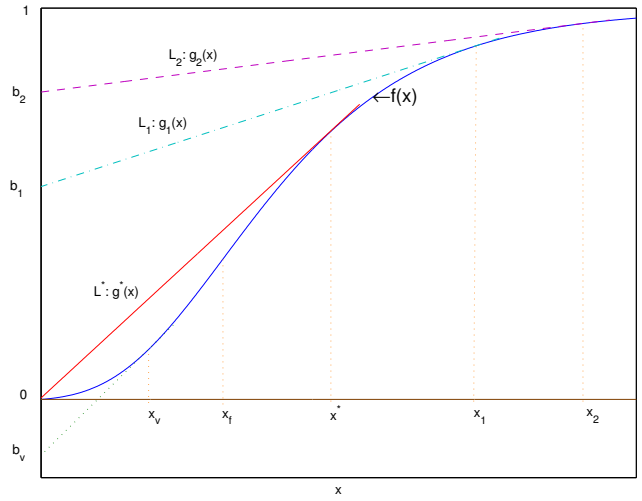

Figure 3.1: A typical “corrected” frame-success function and its “critical tangent”

3.3 Early QoS indices for wireless data

3.3.1 The Intuitive Index and Its Problem.

The ratio of a terminal's throughput to the power employed by it was introduced in [48] in an analysis of re-transmission schemes of data packets. Specifically, let $f_s(\gamma)$ denote the frame success function, and γ the received SIR. The TPR is proportional to the quantity $Rf_s(\gamma)/P$, where P is the transmission power of the concerned transmitter, and R its transmission rate. This yields a physically significant measure in bits per Joule of considerable appeal as a user's quality-of-service index. Below, a development leading to this measure "from first principles" is given. But, generally, $f_s(0) > 0$, which implies that the TPR grows without bound as the transmission power approaches zero.

The zero-power issue can become a practical problem. The implementation of utility maximization in a practical setting may take the form of an algorithm, not necessarily controllable by a human operator, possibly embedded into a transmitting terminal. Thus, the misbehavior of the TPR near zero could drive the algorithm toward arbitrarily small transmission power levels, or no transmission at all, in situations where such behavior would be inappropriate. To counter this, the algorithm would need to be endowed with additional "intelligence", which would increase its computational complexity.

3.3.2 The Efficiency Function remedy and its problems.

Reference [35] and the literature that followed it replaced the frame success function in the numerator of the TPR with an "efficiency function", $f_e(\gamma)$, which gives (as a function of the SIR in the received signal) a "measure of the efficiency of the transmission protocol" [35]. Then, they defined the utility function as proportional to the ratio $Rf_e(\gamma_i)/P_i$.

But f_e was only specified, as $(1-2\mathrm{BER}(\gamma_i))^M$, for frame-success functions of the simple form $(1-\mathrm{BER}(\gamma_i))^M$ (BER denotes the bit error rate). Moreover, there is no clear physical or probability interpretation for this function, nor for the utility function obtained from it. Furthermore, power control algorithms designed with this efficiency function can be highly suboptimal (of the order of 18 to 1 in a specific example) [31].

3.4 The ETPR: An Improved QoS Index

3.4.1 A QoS Metric from First Principles

The development of a QoS index from first principles may provide some additional valuable insights into this issue, and is done below.

3.4.1.1 Decision Scenario

It is assumed that the underlying communication technology is CDMA, although the general approach could be extended to other technologies (one can set up the analysis in terms of the familiar ratio E_b/N_0; that is, energy-per-bit over "noise"). Specifically:

Given:

- A certain amount of energy, E_i, available for transmission
- A fixed transmission rate of R_i bits per second
- A long sequence of blocks of bits ("frames") of length M_i containing $L_i < M_i$ data bits (plus "overhead").
- A certain fixed level of interference (noise), I_i
- A frame-success function f_s as described in section 3.2.

the transmitter wants to choose its transmission power in order to satisfy a reasonable optimality criterion. The transmission power will be set at the start of the transmission, and held constant until energy runs out.

3.4.1.2 Performance for a Fixed Power Level

Since only one terminal is being considered in this development, the subscript i is dropped. Let $Q = P \cdot h$ be the power at the receiver when a certain data packet is transmitted with power P; and let I be the interference (noise) power. Then, $f_s(GQ/I)$ is the probability that said packet is correctly received. G is the spreading (processing) gain, defined as the ratio of the channel's "chip rate", R_c, to the transmission bit rate, R.

Assuming that, once a packet is received in error, re-transmission is ideal, then the total number of times a given packet needs to be transmitted, including re-transmissions, is a geometric random variable, whose probability distribution is of the form $\pi(1-\pi)^{k-1}$ with $\pi = f_s(GQ/I)$. The expected value of this random variable is $1/\pi$, interpreted as the average number of times the same packet needs to be transmitted to ensure correct reception.

The average amount of energy that needs to be spent in order to achieve the successful reception of a data packet when transmission power is set to P can be obtained as follows. Each packet requires an amount of energy equal to the product of P times the length in time of a packet (given the transmission rate R) times the average number of times the same packet needs to be transmitted to ensure correct reception. Each bit lasts $1/R$ secs, so each M-bit frame lasts M/R secs. Therefore, the average amount of energy required by a packet is $P \cdot (M/R) \cdot (1/\pi) = PM/(\pi R)$. Thus, with transmission power fixed at P, the average number of information bits which can be successfully transmitted with an energy budget E is then (assuming all variables are continuous)

$$L\left(E \div \frac{PM}{\pi R}\right) = ER\frac{L}{M}\frac{\pi}{P} \equiv E\frac{L}{M}\frac{Rf_s(GhP/I)}{P} \tag{3.1}$$

3.4.2 A Refined energy-expenditure criterion

The preceding analysis has led naturally to the throughput-to-power ratio, TPR. It is tempting to assume that the terminal should choose its power in order to maximize this index, which would result in the maximal average number of bits transmitted before energy runs out. But it has already been discussed that doing so leads to technical difficulties of both theoretical and practical importance.

3.4.2.1 Throughput: Earned vs. Serendipitous

In order to prevent the technicalities in question, while preserving the physical meaning and probability interpretation of the relevant quantities, one must distinguish between two additive components of the throughput: the earned throughput, and the serendipitous (trivial) throughput. The earned throughput is the result of the expenditure of transmission power. On the other hand, the serendipitous throughput is that obtained without power expenditure, due to serendipity (a detector's wild guesses), which yields a correct detection of a packet with a probability of 2^{-M}.

3.4.2.2 An appropriate criterion

An appropriate objective for the terminal is to choose its transmission power in order to maximize the ratio of the *earned* throughput derived by a transmitter to the transmission power, or the earned-throughput-to-power ratio (ETPR). This results in the maximal average number of *earned* successfully transmitted bits before the available energy is exhausted.

Specifically, if $f_s(\gamma)$ gives the probability that a packet sent by terminal i is correctly detected, when its SIR at the base station is $\gamma = GhP/I$, then the ETPR ("utility") of terminal i is defined as:

$$\begin{aligned} \mathrm{ETPR}(\gamma) &= \frac{f_s(\gamma) - f_s(0)}{P} \quad \text{for } \gamma > 0 \\ \mathrm{ETPR}(0) &= \lim_{\gamma \downarrow 0} \mathrm{ETPR}(\gamma) \end{aligned} \tag{3.2}$$

If one wishes to make the range of the numerator equal to the interval $[0,1]$, one can divide the ETPR by $(1 - f_s(0))$. Likewise, by multiplying, as in the original index, by the data rate R, one obtains a physically meaningful QoS index in bits per Joule. However, in the subsequent development, the scaling constants are ignored.

3.4.3 Technical behavior of the ETPR

As long as G , h, and I are fixed, $k := Gh/I$ is a constant. Thus, maximizing $(f_s(GhP/I) - f_s(0))/P$ is equivalent to maximizing $k(f_s(kP) - f_s(0))/kP$, or simply maximizing $(f_s(x) - f_s(0))/x$ with $x := kP \equiv GhP/I$. Much relevant information about the technical behavior of the ETPR can be found in chapter 2, which discusses the "context-free" maximization of the ratio $f(x)/x$, with f an arbitrary function with an S-shaped graph, as discussed in section 3.2. Chapter 2 shows that this ratio is quasi-concave, and admits a unique global maximizer. The maximizer can be graphically

identified in figure (3.1) as x^*, the abscissa of the only point at which a line tangent to f passes through the origin. Below, the behavior of the ETPR around 0 is of special interest.

The generalized frame-success function being considered is strictly convex over the interval $[0, x_f]$, with x_f a positive number. It is well-known that the continuously differentiable function $f_s : I \to R$ defined on an interval $I \subset \Re$ is strictly convex if and only if, $\forall x_1, x_2 \in I$,

$$f_s(x_2) > f_s(x_1) + f_s'(x_1) \cdot (x_2 - x_1)$$

Taking $x_1 = 0$ and x_2 equal to an arbitrary $x \in (0, x_f]$, the preceding inequality yields

$$f_s(x) > f_s(0) + f_s'(0) \cdot (x) \tag{3.3}$$

This inequality, (3.3), immediately implies that if f_s "starts out" convex, as assumed, $f_s'(0)$ must be finite. And a simple application of L'Hospital rule shows that the ETPR (see equation (3.2)) goes to $k f_s'(0)$ when its argument goes to zero. Therefore, as long as f_s has the assumed shape, $\text{ETPR}(0) = \lim_{x \downarrow 0} \text{ETPR}(x)$ is finite. Furthermore, inequality (3.3) can be re-written as $k(f_s(x) - f_s(0))/x > k f_s'(0)$. The left hand side of this inequality is the ETPR evaluated at x (equation (3.2)), and the right-hand side is ETPR(0). Thus the ETPR is actually *minimized* at 0, and, for the assumed family of frame-success functions, an ETPR-maximizing algorithm will *never* choose 0 as the maximizer.

It is interesting to note that if f_s were a function which "starts out" strictly *concave,* inequality (3.3) would be reversed, and could be written as $(f_s(x) - f_s(0))/x < f_s'(0)$ for any $x \leq x_f$. In that case, zero would be, indeed, a (local) *maximizer* for $(f_s(x) - f_s(0))/x$.

3.4.4 Discussion

In most, if not all, practical systems, the serendipitous throughput is negligible, and so is the difference between the earned-throughput-to-power ratio (ETPR) and the (total) throughput-to-power ratio (TPR). However, the mis-behavior of the TPR for low transmission power is of theoretical and practical importance, as has already been explained.

By contrast, the ETPR is well-behaved throughout its entire domain. Not only does this facilitate mathematical analysis. It also means that an ETPR-maximizing algorithm will not choose an unreasonably small transmission power because of the technical misbehavior of the objective function. This additional reliability comes without any significant complexity cost.

Intellectual curiosity may lead one to consider which one of these two ratios, regardless of the technical issue at the origin, come closer to an 'ideal' QoS index. The TPR divides the average amount of data successfully transmitted (per time unit) by the energy spent (in each time unit). This yields a sensible measure in bits per Joule which is appealing as a guide for energy-expenditure decisions. On the other hand, the ETPR compares the amount of energy spent (in each time unit), to the average amount of data (per time unit) the transmitter *could not have delivered* without energy expenditure. Hence, the ETPR, in fact, reflects a refinement of the intuition leading originally to

the TPR. As it turns out, this refinement solves a problem of practical and theoretical importance, without exacting any significant cost.

Finally, one may wonder why the transformation leading to the ETPR, which may superficially seem 'obvious', was not made in earlier works. A plausible answer is that, if some increasing function g is such that $g(0) > 0$ which makes $g(x)/x$ go to infinity as $x \downarrow 0$, the transformed ratio $(g(x) - g(0))/x$ may *also* go to infinity as $x \downarrow 0$. An example of this is $g(x) = \sqrt{x} + 1$, for which $(g(x) - g(0))/x = 1/\sqrt{x}$ which clearly goes to infinity as $x \downarrow 0$. In fact, it was shown in section 3.4.3 that for any function g which "starts out" *concave*, $(g(x) - g(0))/x$ indeed reaches a (local) maximum at zero! One has to invoke the "initial convexity" of the frame-success function in order to show that an ETPR-maximizing algorithm will not converge to zero. Interestingly, all this implies that if a communication channel is ever found with a *concave* frame-success function, then an ETPR-maximizing terminal should decline to use this channel.

Chapter 4

Efficient Decentralized Power Allocation via Mechanism Design

4.1 Introduction

Several recent scholarly publications, following an approach suggested by Ji, [12], recognize that algorithms useful for engineering applications can be obtained via the formulation of radio resource management issues, in particular power control in wireless data applications, on the foundations of microeconomic theory. This approach is centered around the *decentralized* maximization, under appropriate rules and constraints, of a quality-of-service (QoS) index, referred to as a "utility function". This maximization *may* or may *not* involve a human user choosing resources during transmission. The choices may be made by "software agents" inside transmitting terminals. These agents may be entirely programmed by the network administrator, so that they behave in the best interests of the network. Or these agents may be controlled and/or tuned or trained by the actual human operator.

In either case, decentralized QoS maximization can be modeled as a "game": a situation in which each of several "selfish" agents choose a "strategy" in order to maximize its own "payoff". Generally, the payoff to a given player depends on the chosen strategies by all players. For instance, in a wireless network, the transmission power chosen by a terminal becomes interference for others. And this interference affects the payoff/utility (QoS) of all terminals.

Below, a game in which each data transmitting terminal maximizes the QoS index introduced in chapter 3 is studied. This index is the *earned*-throughput-to-power ratio (ETPR), which was shown to exhibit good mathematical behavior, be physically significant, and attain or surpass the intuitive appeal of related measures already accepted by the scientific literature. The data rates are fixed but may be different among the terminals.

A key solution concept is a Nash Equilibrium (NE); i.e., an allocation (a strategy per player) such that no player would be better off by *unilaterally* "deviating" (changing strategy). In this game, a NE specifies a power level per terminal, such that no terminal would increase its QoS index

by unilaterally adjusting its power. It is made clear below, that, if transmission power is limited, a Nash equilibrium does exist. And even if power is unlimited, a Nash equilibrium may exist under certain circumstances. However, NE tend to be "inefficient", which is verified in this case. The challenge is to get selfish terminals to move toward a more efficient operating point "on their own".

An approach to guide competing selfish entities toward a "socially optimal" outcome is to design an appropriate set of "rules of interaction"; i.e., a set of procedures, penalties and rewards designed to guide the entities toward a desired operating point. In order to achieve an efficient decentralized allocation of power among mutually interfering terminals, this chapter proposes the application of a relatively simple mechanism introduced in [43]. In order for this mechanism to work, there must exist one "transferable good" with which terminals can compensate each other. This good could be money, or some form of service credits, such as time of usage ("minutes").

Below, first, the system model is built. Then, the game in which each data transmitting terminals chooses its transmission power in order to maximize its QoS index, without any mechanism present, is analyzed. Subsequently, it is shown by two methods that the NE of this game is "inefficient". Finally, the compensation mechanism is introduced and discussed.

4.2 System Model

This work discusses the application of a mechanism (a set of rules for the interaction of some "players") to guide some mutually-interfering data-transmitting terminals to an "efficient" allocation of power. The mechanism could be applied under many physical layer configurations, and multi-access schemes. For simplicity, this work focuses on a single CDMA cell.

In this simple model, the following quantities and/or concepts are of interest:

i) N is the number of terminals transmitting data simultaneously to the base station. For most of the development, $N = 2$ is assumed. Extensions are discussed at the end.

ii) R_i bits per second is the source data rate of terminal i

iii) R_C chips per second is the chip rate ("bandwidth") of the channel

iv) $G_i = R_C/R_i$ is the processing gain of terminal i.

vii) $f_S(G_i\alpha_i)$ is the probability of correct reception of a data packet, where α_i is the carrier-to-interference ratio (CIR) of the receiver tuned to transmitter i, and is defined by:

$$\alpha_i = \frac{P_i h_i}{\sum_{\substack{j=1 \\ j\neq i}}^{N} P_j h_j + \sigma^2} = \frac{Q_i}{\sum_{\substack{j=1 \\ j\neq i}}^{N} Q_j + \sigma^2} \doteq \frac{Q_i}{I_i} \tag{4.1}$$

In this expression, $P_i \leq P_{\max}$ is the transmitted power of terminal i, h_i is the path gain from terminal i to the base station, and σ^2 is the noise power in the base station receiver. $Q_i = P_i h_i$ is the received power at the base station in the signal transmitted by terminal i. For notational convenience, I_i denotes the total level of interference experience by terminal i.

viii) Absent of other incentives, the earned-throughput-to-power ratio (ETPR) discussed in chapter 3 is the quality-of-service (QoS) index whose maximization is desired by each terminal.

It is obtained as :

$$\frac{R_i\,(L/M)\,(f_S(G_i\alpha_i)-f_S(0))}{P_i} \doteq R_c\left(\frac{L}{M}\right)h_i\frac{f(G_i\alpha_i)}{G_iQ_i}$$

L/M is the ratio of the number of information bits in a data packet to the total number of bits in the packet. $f(x) \doteq f_S(x)-f_S(0)$ has been set. $f_S(0)$ is typically very small, but, as discussed in chapter 3, this correction is necessary to avoid technical and practical problems.

It is assumed that *all that is known* about f is that its graphs has an "S-shape", as shown in fig. 3.1. This should accommodate most physical layer configurations of practical interest. The technical characterization of a function with an "S-shaped" graph is discussed in chapter 2.

4.3 Decentralized ETPR Maximization: No Mechanism

4.3.1 Objective Function and constraints

For a given level of interference, I_i , terminal i wants to choose its transmission power, P_i , to maximize:

$$\frac{G_i}{I_i}\frac{f\,(G_iQ_i/I_i)}{G_iQ_i/I_i} \text{ or simply } \frac{f(x)}{x} \text{ with } x \doteq G_i\frac{Q_i}{I_i} \tag{4.2}$$

subject to:

$$0 \le x \le x_{M_i} \text{with } x_{M_i} = \frac{G_i}{I_i}Q_{M_i} = \frac{G_i}{I_i}h_iP_{max}$$

4.3.2 Best Response Function

As discussed in section 3.4.3, the maximization of the ratio $f(x)/x$ for function f as described in section 3.2 is well understood. Chapter 2 shows that $f(x)/x$ is quasi-concave, and admits as unique global maximizer x^*, which is the only positive number satisfying $xf'(x) = f(x)$ (see figure (3.1)).

This implies that the maximizer sought in the problem of section 4.3.1 is the smallest of x_{M_i} and x^*. Let $x_i^* = \min(x^*, x_{M_i})$. It follows that, for a given interference level I_i, transmitter i will respond with a P_i^* such that

$$Q_i^* = \frac{I_i}{G_i}x_i^* = \min\left(\frac{I_i}{G_i}x^*, h_iP_{max}\right) \tag{4.3}$$

4.3.3 Nash-equilibria

In this context, a Nash equilibrium is a power vector, specifying a power level for each active terminal, such that no terminal can increase its quality of service by *unilaterally* changing its power level.

The preceding development indicates that the "best response" of each terminal is such that, for a given interference level, each would like to set its transmission power to achieve a "received" signal-to-interference ratio of x^*, a constant determined by the physical layer through the frame-success function. When a terminal cannot reach the power level leading to x^*, it transmits at the highest possible power level. However, the interference level is not a fixed constant, but rather, a variable

determined by the transmission power levels of all active terminals. Thus, it is, in principle, unclear whether an equilibrium power vector will exist.

It can be shown on the basis of a well-known result by Gerard Debreu that, if transmission power is limited and utility functions are quasi-concave (which the ETPR is), a Nash-equilibrium does exist (see [33] for further details). Nevertheless, below the conditions for the existence of Nash-equilibria of various forms (with and without transmission power limits) are explored "from first principles", without explicitly invoking Debreu's or similar results.

4.3.3.1 Equal-received-SIR Nash equilibrium (ERSNE)

This section seeks conditions under which a solution exists for a set of N equations of the form:

$$\frac{Q_i}{I_i} \equiv \frac{Q_i}{\sum_{\substack{j=1 \\ j\neq i}}^{N} Q_j + \sigma^2} = \frac{x^*}{G_i} := \alpha_i \tag{4.4}$$

This problem is fully discussed in appendix B. The equations defining the α_i's (equation (4.4)) lead to a system of equations:

$$\begin{pmatrix} 1 & -\alpha_1 & \cdots & -\alpha_1 \\ -\alpha_2 & 1 & \cdots & -\alpha_2 \\ \vdots & \vdots & \vdots & \vdots \\ -\alpha_N & -\alpha_N & \cdots & 1 \end{pmatrix} \begin{pmatrix} Q_1 \\ Q_2 \\ \vdots \\ Q_N \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{pmatrix} \sigma^2 \tag{4.5}$$

Then, one can show that if the condition

$$s := \sum_{k=1}^{N} \frac{\alpha_k}{1+\alpha_k} \equiv \sum_{k=1}^{N} \frac{x^*}{x^* + G_k} < 1 \tag{4.6}$$

is satisfied, the system (4.5) has a unique solution, in which each component of the received power vector is given by:

$$Q_k^* = \frac{\sigma^2}{1-s}\frac{\alpha_k}{1+\alpha_k} = \frac{\sigma^2}{1-s}\frac{x^*}{x^*+G_k} \tag{4.7}$$

Evidently, if all G_i's are identical, then $\alpha_i = \alpha = x^*/G := 1/\hat{G}$, and the feasibility condition given by (4.6) reduces to:

$$s = \frac{N\alpha}{1+\alpha} := \frac{N}{\hat{G}+1} < 1 \tag{4.8}$$

Likewise, equation (4.7) becomes:

$$Q_k^* = \frac{\sigma^2}{\hat{G}-N+1} := Q_{sym}(N,\sigma^2) \tag{4.9}$$

This development leads to the following conclusion about the feasibility of the ERSNE. In order

for the ERSNE to be feasible, condition (4.6) must be satisfied. When this condition is satisfied, equation (4.7) gives the levels of received power which would lead the terminals to the desired SIR, x^*. Therefore, the ERSNE may fail for either of two reasons: failure of condition (4.6), or inability by any terminal to reach the required power level. In either case, the possibility that a non-ERS equilibrium exists needs to be explored.

4.3.3.2 ERSoMP-1 Nash Equilibrium

This section explores conditions under which a Nash equilibrium exists in which one terminal operates at maximal power, while all others operate at whichever power level is necessary to achieve the optimal SIR of x^*. This case will be identified as an ERSoMP-NE-1 for an equal-received-SIR or maximal-power Nash equilibrium of order 1.

For expositional convenience, it is assumed that terminal N is the one operating at maximal power. In this scenario, the received power from terminal N, Q_N, is presumed fixed at $h_N P_{max} := \bar{Q}_N$, while others need to be found to satisfy :

$$\frac{Q_i}{\sum_{\substack{j=1 \\ j\neq i}}^{N-1} Q_j + \Sigma^2} = \frac{x^*}{G_i} := \alpha_i \tag{4.10}$$

where $\Sigma^2 := \bar{Q}_N + \sigma^2$. For $i = 1 \ldots (N-1)$ the value of each α_i is the same as in the original equation (4.4).

Evidently, the equations of the form (4.10) lead to a system analogous to (4.5), except that it is of order $N-1$, and Σ^2 replaces σ^2. From the development leading to condition (4.6), the feasibility condition for the solution of this new system is:

$$s_1 := \sum_{k=1}^{N-1} \frac{\alpha_k}{1+\alpha_k} < 1 \tag{4.11}$$

Likewise, if inequality (4.11) is satisfied, a unique solution exists, in which the first $N-1$ components of the received power vector satisfy:

$$Q_k^* = \frac{\Sigma^2}{1-s_1} \frac{\alpha_k}{1+\alpha_k} \tag{4.12}$$

Notice that if inequality (4.6) is satisfied, so is inequality (4.11). But the converse is obviously not true.

Again, if $\forall i$, $\alpha_i = x^*/G = \alpha$, the feasibility condition given by (4.11) reduces to:

$$s_1 = \frac{\alpha(N-1)}{1+\alpha} = \frac{N-1}{\hat{G}+1} < 1 \tag{4.13}$$

and equation (4.12) becomes:

$$Q_k^* = \frac{\Sigma^2}{1-s_1} \frac{\alpha}{1+\alpha} = \frac{\bar{Q}_N + \sigma^2}{\hat{G} - N + 2} \tag{4.14}$$

Even if the new feasibility condition (4.11) is satisfied, and each of the terminals from 1 to $N-1$ can reach the required power level (4.12), the possibility that this allocation may *not* be a Nash equilibrium needs to be explored. According to the development in section 4.3.2, the best-response function of terminal N is given by equation (4.3) as $Q_N^* = \min\left(\frac{I_N}{\hat{G}}, \bar{Q}_N\right)$. This means that if $\bar{Q}_N > I_N/\hat{G}$, terminal N would be better off by lowering its power, and the allocation being considered would *fail* to be a Nash equilibrium. This possibility is explored below for the identical-rates case. The extension of this procedure to consider nonidentical rates is straightforward.

On the basis of equation (4.14), I_N can be obtained as

$$I_N = (N-1)\frac{\bar{Q}_N + \sigma^2}{\hat{G} - N + 2} + \sigma^2 \tag{4.15}$$

In order to ascertain whether $\hat{G}\bar{Q}_N < I_N$, this inequality can be expressed as

$$\bar{Q}_N\left(\hat{G} - \frac{N-1}{\hat{G}-N+2}\right) \overset{?}{<} \sigma^2\left(\frac{N-1}{\hat{G}-N+2} + 1\right)$$

or, since $\hat{G} - N + 2 > 0$ by condition (4.13), as

$$\bar{Q}_N[\hat{G}(\hat{G}-N+2) - N + 1] \overset{?}{<} (\hat{G}+1)\sigma^2$$

Notice that $\hat{G}(\hat{G}-N+1+1) - N + 1$ can be written as $\hat{G}(\hat{G}-N+1) + (\hat{G}-N+1)$ which can be factored as $(\hat{G}+1)(\hat{G}-N+1)$. This leads to checking whether,

$$\bar{Q}_N(\hat{G}-N+1) \overset{?}{<} \sigma^2 \tag{4.16}$$

If condition (4.8) is satisfied, which means that, without a power limitation, the original ERSNE would have been feasible, then inequality (4.16) can be written as $\bar{Q}_N < \sigma^2/(\hat{G}-N+1)$. But the right-hand side of this inequality is precisely the received power level required for the equal-received-SIR Nash Equilibrium (ERSNE) (equation(4.9)). Thus, inequality (4.16) is satisfied if condition (4.8), which determines the feasibility of the power-unlimited ERSNE, was satisfied, but terminal N could not, because of its power constraint, reach the power level necessitated by ERSNE.

On the other hand, if condition (4.8) failed, which means that the original ERSNE would have been *impossible* even without a power limitation, then the left-hand-side of inequality (4.16) is *negative*, which directly implies that this inequality is necessarily satisfied.

In conclusion, the ERSoMP-NE-1 exists whenever three requirements are met: (i) condition (4.11) is satisfied, (ii) each of the terminals from 1 to $N-1$ can reach the required power level (4.12), and (iii) the ERSNE failed to exist. (For this purpose, it does *not* matter whether the ERSNE failed because condition (4.8) failed, or because terminal N could not reach the required power level).

4.3.3.3 ERSoMP Nash-Equilibrium of order M

The preceding development suggests the following extension to the more general equilibrium in which M terminals operate at maximal power, with the remaining ones operating with received SIR equal to x^*; i.e., an equal-received-SIR or maximal-power Nash equilibrium of order M (ERSoMP-NE-M). As discussed in the introduction to section 4.3.3, given the quasi-concavity of our utility function, such equilibrium exists, whenever transmission power is limited [33].

For expositional convenience, it is assumed that the terminals have been labeled so that if M terminals cannot reach the required power level, they are terminals $N-M+1$ through N. For instance, this will happen if both the transmission bit rates, and the maximal transmission power levels are constant across terminals, but the path gains satisfy $h_1 > \cdots > h_N$.

First, check whether condition (4.8), which determines the feasibility of the power-*un*limited ERSNE, is satisfied, and all terminals can reach the appropriate power level given by equation (4.7). If this is the case, then the ERSNE is the only available NE. If condition (4.8) fails and transmission power is unlimited, then *no* NE is available. If an ERSNE is *not* possible (for whatever reason), and transmission power is limited, then set terminal N at maximal power and determine whether an ERSoMP-NE-1 is possible. If condition (4.13) fails, or if this condition is satisfied but one or more of the first $N-1$ terminals cannot reach the required power level, (equation (4.12)), then an ERSoMP-NE-1 is *not* possible. Hence, set both terminal N and terminal $N-1$ at maximal power, and proceed to verify whether an ERSoMP-NE-2 is possible. Continue this recursion, until an ERSoMP Nash equilibrium of order M is reached.

4.3.4 Discussion

A game in which terminals carrying multi-rate traffic choose transmission power in order to maximize the ETPR index has been analyzed. The key solution concept is a Nash equilibrium; i.e., an allocation such that no terminal would be better off by *unilaterally* "deviating". Closed-form equilibrium conditions and power levels has been derived from first principles. It has been shown that all terminals want the same signal-to-interference ratio (SIR), but, because of power limitations, some terminals cannot reach the necessary power level. At equilibrium, a number of terminals transmit at maximal power, and the others achieve the same optimal SIR. This SIR value can be easily identified in the graph of f as the abscissa of the only point where a ray emanating from the origin is tangent to the graph (see x^* in fig. 3.1). A basic rationale to search for these equilibria has been given.

4.4 Efficiency Analysis of the Equilibria

4.4.1 Overview

With limited transmission power, an equilibrium always exists. And even with unlimited power, an equilibrium *may* exist if certain sum of simple terms is less than 1. However, the equilibrium

allocation can be shown to be inefficient (not Pareto optimal): there are other feasible allocations under which the utility of some terminals could be increased without decreasing the utility of any other. To show the inefficiency of the equilibrium, two separate arguments are given. The first argument follows that given in [33], and concludes that if the terminals operating at the optimal SIR lower their equilibrium power levels by certain fraction, the utility of *each* terminal increases. Thus, equilibrium power levels are "too high". Subsequently, an alternative way of showing that the equilibrium power levels are inefficient based on economics theory is provided, for a 2-terminal situation.

4.4.2 Description of the equilibrium allocation

As discussed in section 4.3.2, each terminal wants to operate with SIR x^*(see x^* in fig. 3.1). If the condition $s := \sum_{k=1}^{N} x^*/(x^* + G_k) < 1$ is satisfied, and power limits are sufficiently high, each terminal can reach the received power level leading to the SIR x^*. When all terminals have the same transmission rate, this power level is, as a multiple of the noise average power,:

$$q^* = \frac{1}{G/x^* - 1} \rightarrow u_i^* \propto h_i R_i (G - x^*) \frac{f(x^*)}{x^*}$$

u_i^* is the utility derived at equilibrium by terminal i.

4.4.3 Inefficiency of equilibrium allocation-I

If all terminals simultaneously change their received power to εq^*, with $0 < \varepsilon \leq 1$, then the new SIR is

$$\gamma_\varepsilon = \frac{G q^*}{(N-1)q^* + \varepsilon^{-1}}$$

and the new utility level is $u_i^\varepsilon \propto f(\gamma_\varepsilon)/\varepsilon q^*$.

$$\frac{\partial u_i^\varepsilon}{\partial \varepsilon} \propto \left(f'(\gamma_\varepsilon) \frac{\partial \gamma_\varepsilon}{\partial \varepsilon} - \varepsilon^{-1} f(\gamma_\varepsilon) \right)$$

$$\frac{\partial \gamma_\varepsilon}{\partial \varepsilon} = \frac{G q^* \varepsilon^{-2}}{\left((N-1)q^* + \varepsilon^{-1}\right)^2} = \frac{1}{G q^*} \left(\frac{\gamma_\varepsilon}{\varepsilon} \right)^2$$

$$\therefore \quad \frac{\partial u_i^\varepsilon}{\partial \varepsilon} \propto \frac{\gamma_\varepsilon}{\varepsilon} \frac{\gamma_\varepsilon f'(\gamma_\varepsilon)}{G q^*} - f(\gamma_\varepsilon)$$

Recall that for $\varepsilon = 1$, $\gamma_\varepsilon = x^*$ and $f(x^*) = x^* f'(x^*)$. Thus,

$$\left. \frac{\partial u_i^\varepsilon}{\partial \varepsilon} \right|_{\varepsilon=1} \propto f(x^*) \left(\frac{x^*}{G q^*} - 1 \right) \propto \frac{1}{(N-1)q^* + 1} - 1 < 0$$

The fact that this derivative is negative for $\varepsilon = 1$ implies that for some scaling factor ε such that $0 < \varepsilon < 1$, if *all* terminals simultaneously scale down their equilibrium power levels by ε, each terminal would increase its utility over its equilibrium value.

This argument only considers a situation in which all terminals can reach the universally desired SIR. However, [33] extends this argument to show that, even when some terminals must operate at maximal power, if all the terminals operating at the optimal SIR scale down their equilibrium power levels by an appropriate factor, the utility of *each* terminal, including those operating at maximal power, increases.

4.4.4 Inefficiency of equilibrium allocation-II

The first-order necessary conditions that must be met by an allocation in order to be Pareto-efficient are the same as those of an allocation that maximizes a weighted sum of the utilities of each terminal [42, p. 332]. Therefore, an allocation that fails to meet these conditions is *not* Pareto-efficient. Below, it is verified for the two terminal situation that the Nash equilibrium of the game played by two data-transmitting terminals fails the necessary conditions for Pareto-efficiency, with the possible exception of the situation in which both terminals operate at maximal power.

For two terminals, the centralized utility-maximization problem is:

$$\begin{array}{rl} \max_{q_1,q_2} & \beta_1 h_1 \frac{f(\gamma_1)}{G_1 q_1} + \beta_2 h_2 \frac{f(\gamma_2)}{G_2 q_2} \\ \text{subject to} & q_1 \leq \bar{q_1}\ ;\ q_2 \leq \bar{q_2} \end{array}$$

β_i denotes an arbitrary weight, $q_i := Q_i/\sigma^2$ and

$$\gamma_i = \frac{G_i q_i}{q_j + 1} \rightarrow \frac{\partial \gamma_i}{\partial q_i} = \frac{\gamma_i}{q_i}\ ;\ \frac{\partial \gamma_i}{\partial q_j} = -\frac{\gamma_i^2}{G_i q_i}$$

The augmented objective function (Lagrangian) is:

$$\frac{\beta_1 h_1}{G_1}\frac{f(\gamma_1)}{q_1} + \frac{\beta_2 h_2}{G_2}\frac{f(\gamma_2)}{q_2} + \mu_1(\bar{q_1} - q_1) + \mu_2(\bar{q_2} - q_2)$$

The corresponding first-order necessary conditions for a maximum are:

$$\begin{bmatrix} \frac{\beta_1 h_1}{G_1}\frac{\gamma_1 f'(\gamma_1) - f(\gamma_1)}{q_1^2} - \frac{\beta_2 h_2}{G_2}\frac{\gamma_2^2 f'(\gamma_2)}{G_2 q_2^2} + \mu_1 \\ -\frac{\beta_1 h_1}{G_1}\frac{\gamma_1^2 f'(\gamma_1)}{G_1 q_1^2} + \frac{\beta_2 h_2}{G_2}\frac{\gamma_2 f'(\gamma_2) - f(\gamma_2)}{q_2^2} + \mu_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{4.17}$$

$$\mu_i(\bar{q_i} - q_i) = 0 \text{ and } \mu_i \leq 0 \tag{4.18}$$

For an interior solution, $\mu_i = 0$, and equation (4.17) implies,

$$\frac{\beta_1 h_1}{G_1}\frac{\gamma_1 f'(\gamma_1) - f(\gamma_1)}{q_1^2} = \frac{\beta_2 h_2}{G_2}\frac{\gamma_2^2 f'(\gamma_2)}{G_2 q_2^2} \tag{4.19}$$

$$\frac{\beta_2 h_2}{G_2}\frac{\gamma_2 f'(\gamma_2) - f(\gamma_2)}{q_2^2} = \frac{\beta_1 h_1}{G_1}\frac{\gamma_1^2 f'(\gamma_1)}{G_1 q_1^2} \tag{4.20}$$

Quite clearly, a finite power vector in which $\gamma_1 = \gamma_2 = x^*$ cannot possibly satisfy equations (4.19

and 4.20), because $x^* f'(x^*) = f(x^*)$, which would make the left-hand side of these equations equal to zero, while the right-hand side is greater than zero ($f'(x) > 0$).

If terminal 2 was operating at the optimal SIR x^*, with terminal 1 operating at maximal power, $\mu_2 = 0$, and equation (4.20) would still apply. But again, the left-hand side of equation (4.20) would continue to equal zero, while its right-hand side would be greater than zero. A similar situation happens if the terminals switched roles.

Finally, with both terminals operating at maximal power, $q_i = \bar{q_i}$, and the complementary-slackness condition, equation (4.18), would require $\mu_i \leq 0$. First, observe that :

$$\bar{\gamma_i} := \frac{G_i \bar{q_i}}{\bar{q_j} + 1} \longrightarrow \frac{G_i \bar{q_i}}{\bar{\gamma_i}} \equiv \bar{q_j} + 1$$

From equations (4.19),

$$\begin{aligned} \mu_1 &= \frac{\beta_2 h_2}{G_2} \frac{\bar{\gamma_2}^2 f'(\bar{\gamma_2})}{G_2 \bar{q_2}^2} - \frac{\beta_1 h_1}{G_1} \frac{\bar{\gamma_1} f'(\bar{\gamma_1}) - f(\bar{\gamma_1})}{\bar{q_1}^2} \\ &\equiv \beta_2 h_2 \frac{f'(\bar{\gamma_2})}{(\bar{q_1}+1)^2} - \frac{\beta_1 h_1}{G_1} \frac{\bar{\gamma_1} f'(\bar{\gamma_1}) - f(\bar{\gamma_1})}{\bar{q_1}^2} \end{aligned} \tag{4.21}$$

Notice that $\bar{\gamma_1} f'(\bar{\gamma_1}) - f(\bar{\gamma_1})$ has the sign of the derivative of $f(t)/t$ evaluated at $\bar{\gamma_1}$. Chapter 2 shows that this ratio is "single-peaked", and reaches its maximum at γ_0. Thus, its derivative is negative for any $t > \gamma_0$. Therefore, $\bar{\gamma_1} > \gamma_0$ makes μ_1 positive. Thus, operating at maximal power when the preferred SIR γ_0 is achievable is not Pareto-efficient, as intuition suggests. On the other hand, with $\bar{\gamma_1} < \gamma_0$, which is really the interesting case, the second term in the right-hand side of equation (4.21) becomes negative, but the first one remains positive. A similar analysis applies to μ_2. So, this test is inconclusive when both terminals operate at maximal power. This is intuitively appealing. If both terminals are very poorly situated with respect to the base station, it could be perfectly reasonable ("efficient") for both of them to operate at maximal power.

$$\mu_2 = \frac{\beta_1 h_1}{G_1} \frac{\bar{\gamma_1}^2 f'(\bar{\gamma_1})}{G_1 q_1^2} - \frac{\beta_2 h_2}{G_2} \frac{\bar{\gamma_2} f'(\bar{\gamma_2}) - f(\bar{\gamma_2})}{q_2^2} \equiv \beta_1 h_1 \frac{f'(\bar{\gamma_1})}{(\bar{q_2}+1)^2} - \frac{\beta_2 h_2}{G_2} \frac{\bar{\gamma_2} f'(\bar{\gamma_2}) - f(\bar{\gamma_2})}{q_2^2}$$

4.5 A Simple Efficiency-Inducing Mechanism

Section 4.4 shows that the allocation arising as a Nash equilibrium of the game in which each data terminal chooses its transmission power to maximize the ETPR is not efficient. The challenge is to guide the "selfish" terminals toward an efficient operating point "on their own".

An approach employed in [33] to induce the terminals toward a lower-power equilibrium is to introduce a "tax" on transmission power. That is, terminals are programmed to maximize an expression of the form $u(p;I) - cp$, where $u(p;I)$ denotes the utility of the terminal when its transmission power is p, and its interfering power (caused by noise and the other terminals) equals I; and c is a "tax" on power. This leads to lower power levels at equilibrium, and an increase in the utility to the terminals. However, there are several problems with this approach: (i) while the new

equilibrium allocation is an improvement, it is still inefficient; (ii) there is no clear and convenient expression giving the optimal tax, and (iii) the approach may require certain additional impositions and technical assumptions which are best avoided.

This section proposes and discusses the application of a "mechanism" introduced in [43] to guide the terminals toward an efficient allocation of transmission power. In much of the development below, only two terminals are considered. However, the discussion section addresses some of the issues involved in a multi-terminal extension of this approach.

4.5.1 What is a mechanism?

A "mechanism" is a set of procedures, penalties and rewards intended to guide selfish entities toward a desired outcome. An example of a simple, well-known mechanism is Vickery's Second Price Auction. In this situation, a valuable object if offered for sale to several interested parties. Each submits a sealed bid, and the object is awarded to the highest bidder. However, the winner only pays the second-highest bid! The key advantage of this auction is that it has been proved that each player's best response is to bid its exact true valuation of the object, which is private information only known to him/her. That is, in this arrangement, "truth telling" is optimal [45].

4.5.2 Economic Model

In order to achieve an efficient decentralized allocation of power among mutually interfering terminals through the proposed mechanism, there must exist one transferable good (say money) with which terminals can compensate each other.

The basic economic model is that of partial-equilibrium analysis and a quasi-linear utility function, as discussed, for instance in [42, Ch. 10]. Each terminal is assumed to have both an energy budget, E_i, and a monetary budget, D_i. The terminal's payoff is $\beta_i B_i + y_i$ where (i) β_i is the monetary value to the terminal of one information bit successfully transferred, (ii) B_i is the (average) number of ("earned") information bits the terminal gets to successfully transfer by the time its energy runs out, and (iii) y_i is the amount of money the terminal has left after compensation, and penalties are computed.

Without penalties and rewards, the terminal keeps its complete monetary budget, D_i, intact. Thus, the terminal's payoff is $\beta_i B_i + D_i$. But when a mechanism is introduced, the second term becomes D_i plus any reward/compensation received, minus any penalty/compensation paid. This will be further clarified below.

4.5.3 The compensation mechanism

The mechanism is implemented in two stages: (i) announcement: the terminals announce the prices c_{12}^1, c_{21}^1, c_{21}^2, c_{12}^2, where the superscript indicates which terminal sets the price, and the subscript ij denotes that money flows from i to j. (ii) choice: each terminal chooses its power level to maximize its payoff, given the announced prices. If the compensation offered by terminal i does not match

what terminal j wants, terminal i must pay a penalty $\left(c_{ij}^{i}-c_{ij}^{j}\right)^2$ to a third party. Thus, once all choices have been made, the payoff to terminal 1 is :

$$\underbrace{\beta_1}_{\$/\text{bits}}\underbrace{B_1(P_1;P_2)}_{\text{"earned" bits}}+\underbrace{D_1}_{\text{budget}}+\underbrace{c_{21}^2P_2}_{\text{from 2}}-\underbrace{c_{12}^2P_1}_{\text{to 2}}-\underbrace{\left(c_{12}^1-c_{12}^2\right)^2}_{\text{Penalty}} \tag{4.22}$$

4.5.4 Describing the equilibrium for the asymmetric case

Reference [43] shows that the allocation arising from this game is efficient. Nevertheless, it is interesting to describe the powers and prices arising at equilibrium. This is done below for the special case in which terminal 1 interferes with terminal 2 but *not* vice-versa (successive interference cancellation (SIC) decoding). In this case, c_2 denotes the unit compensation terminal 2 ("injured" terminal) demands, and c_1 the compensation offered by terminal 1 (interferer). Since the injured terminal makes no payments, it is convenient to set its monetary budget $D_2 = 0$.

4.5.4.1 Second-stage payoffs

After all choices have been made, the payoffs for the asymmetric game are :

$$\underbrace{\beta_1}_{\$/\text{bits}}\underbrace{B_1(P_1;I_1)}_{\text{bits}}+\underbrace{D_1}_{\text{budget}}-\underbrace{c_2P_1}_{to2}-\underbrace{(c_1-c_2)^2}_{\text{penalty}} \tag{4.23}$$

$$\underbrace{\beta_2}_{\$/\text{bits}}\underbrace{B_2(P_2;I_2)}_{\text{bits}}+\underbrace{c_1P_1}_{\text{from 1}}-\underbrace{(c_2-c_1)^2}_{\text{penalty}} \tag{4.24}$$

As discussed in chapter 3, for a given level of interference, I_i

$$B_i(P_i,I_i)=E_i\frac{L}{M}R_i\frac{f(G_ih_iP_i/I_i)}{P_i}\equiv R_cE_i\frac{L}{M}\frac{h_i}{I_i}\frac{f(x_i)}{x_i} \tag{4.25}$$

with x_i the signal to interference ratio (SIR) at the receiver (L/M is the ratio of information bits to total bits in a packet) . For a given I_i, choosing transmission power is equivalent to choosing x_i. With a slight abuse of notation, $B_i(x_i,I_i)$ can replace $B_i(P_i,I_i)$.

4.5.4.2 Characterizing the equilibrium

4.5.4.2.1 General approach This is a 2 stage game. To solve it, one first looks at the second stage (choosing the power levels), as if the first-stage choices had been pre-determined. This gives power levels that are a function of c_1 and c_2. With this information, the first-stage of the game can be solved. But notice from eq. (4.23), that the choice of c_1 only impacts the interferer's payoff through the penalty term, $(c_2-c_1)^2$. c_1 *does* influence the power chosen by terminal 2, but this power has no effect on the interferer's payoff because, by assumption, terminal 1's only impairment is random noise. Thus, at equilibrium $c_1 = c_2$ because it is always optimal for the interferer to

avoid the penalty. Hence, in characterizing the equilibrium allocation, one can focus on the case $c_1 = c_2 = c$.

4.5.4.2.2 Interferer's power choice By assumption, terminal 1 interferes with terminal 2 but not vice-versa. That is, $I_1 = \sigma^2$, while $I_2 = P_1 + \sigma^2$.

Terminal 1 will choose P_1 so that x_1 maximizes

$$R_c \frac{L}{M} \frac{E_1}{\sigma^2} h_1 \beta_1 \frac{f(x_1)}{x_1} - c \frac{\sigma^2}{G_1 h_1} x_1 \equiv \hat{\beta}_1 u(x_1) - \hat{c}_1 x_1 \tag{4.26}$$

where for notational convenience

$$u(x_1) := \frac{f(x_1)}{x_1} \; ; \hat{\beta}_1 := \frac{L}{M} \frac{R_c E_1 h_1}{\sigma^2} \beta_1 \; ; \hat{c}_1 := \frac{c\sigma^2}{G_1 h_1}$$

Equation (4.26) has the general form $u(x) - cx$.

4.5.4.2.3 Maximizing $u(x) - cx$ Chapter 2 shows that for f an S-curve, $f(x)/x$ is "single peaked", as shown in fig. 4.1. Thus, the maximization of $u(x) - cx$, where *all that is known* about u is that it is "single peaked", needs to be understood.

As fig. 4.1 clearly shows, if c exceeds certain critical value, c_L, the line cx lies entirely over the curve $u(x)$ except at the origin. Thus, $u(x) - cx < 0$ for any positive x, which implies that $u(x) - cx$ has its maximum at $x = 0$. At the other extreme, if $c \approx 0$, the maximum occurs at x^*, which, as discussed in chapter 2, is shown in fig. 3.1 at the tangency point between f and a line from the origin. For $0 < c \leq c_L$there is an interval (a,b) on which $u(x) > cx$ (when $c = c_L$, a and b "merge" into x_L). The function $u(x) - cx$ is continuous; therefore it must have a maximum over the closed and bounded interval $[a,b]$. The maximum occurs at the point x^{**} where $u'(x) = c$ (that is, a point at which a tangent to the curve is parallel to the line).

With power limitations, the terminal may not be able to exceed a certain maximal SIR $\bar{x}$. In this case, if $a < \bar{x} \leq x^{**}$ it is optimal for this terminal to operate at $\bar{x}$. However, if $\bar{x} < a$ the optimal choice for this terminal is 0, since $u(x) - cx < 0$ for $0 < x < a$. If $\bar{x} = a$ the terminal is indifferent between choosing 0 or a. In the interest of simplicity, it is assumed that when operating and not operating give an identical utility, the terminal will choose to operate.

In conclusion, the problem of maximizing $u(x) - cx$ is well defined, and has a solution. Depending upon the value of c and $\bar{x}$, the maximizer could be 0, $\bar{x}$, or x^{**}. Thus, for fixed $\bar{x}$, the function $x(c)$ giving the maximizer of $u(x) - cx$ is well-defined.

4.5.4.2.4 Injured terminal's power choice From the preceding analysis, it is clear that for a given c one can properly refer to a function $x_1(c)$ giving the optimal SIR for the interferer as a function of the unit compensation paid. $x_1(c)$ directly yields $P_1(c)$, the corresponding transmission power level.

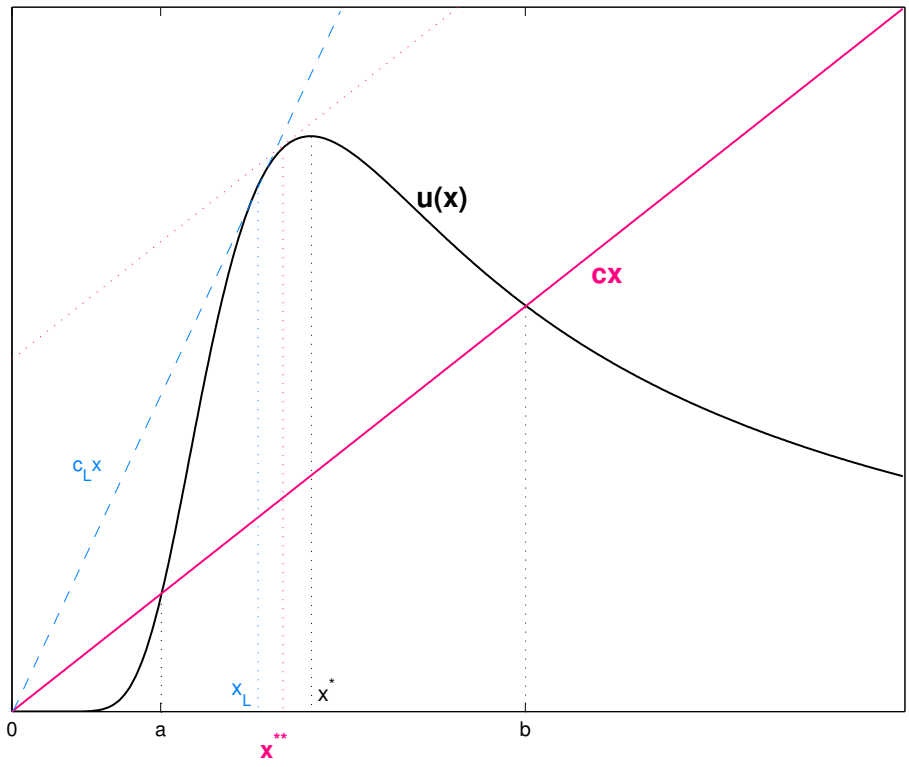

Figure 4.1: x^{**} uniquely maximizes $u(x) - cx$, unless $c > c_L$, in which case, 0 is the maximizer.

Terminal 2 will chooses its power level to maximize

$$R_c \frac{L}{M} \frac{E_2}{I_2} h_2 \beta_2 \frac{f(x_2)}{x_2} + c \frac{\sigma^2}{G_1 h_1} x_1 \equiv \hat{\beta}_2 u(x_2) + \hat{c}_1 x_1 \tag{4.27}$$

Presumably, at this stage, c has already been chosen, and so has the interferer's power as a function of c. Clearly, terminal 2 must choose its power so that its received SIR maximizes $u(x_2) := f(x_2)/x_2$. As discussed in chapter 2, the maximizer occurs at x^*, which is shown in fig. 3.1 at the tangency point between f and a line from the origin.

4.5.4.2.5 Injured terminal's price choice Through the development in sections 4.5.4.2.2 through 4.5.4.2.4, the second stage ("choice") of the asymmetric compensation game has been characterized. The same must be done for the stage of the game in which terminals announce their prices. As remarked in section 4.5.4.2.1, at equilibrium the interferer's compensation will match that demanded by the injured terminal. Thus, all that remains is to specify the price that terminal 2 will demand. This is done with the understanding that for any chosen compensation (c), the interferer will choose its power so that its received SIR is $x_1(c)$ (section 4.5.4.2.3), and the injured terminal will choose to operate at the SIR x^* (section 4.5.4.2.4).

The injured terminal will choose c to maximize its overall utility (taking into account what will happen in the next stage of the game). That is, it will maximize

$$v(c) := \frac{A}{h_1 P_1(c) + \sigma^2} + c P_1(c) \tag{4.28}$$

with

$$A := R_c \frac{L}{M} h_2 E_2 \beta_2 \frac{f(x^*)}{x^*}$$

$v(c)$ is a single variable function which can be readily maximized, whether analytically or numerically. $P_1(c)$ gives the optimally chosen power of the interferer for any given c, which follows from the analysis in section 4.5.4.2.3. Disregarding power constraints, the function $x_1(c)$ (optimal SIR or the interferer) can be assumed to vary smoothly with c (i.e., to be continuously differentiable), over the interval $[0, c_L]$ (see fig. 4.1), and the same can be said about $P_1(c)$. Over this range, $v(c)$ is then a composite function of continuous functions, which is therefore continuous, and must have a maximum over a closed and bounded set. Therefore, $c^* := \arg\max v(c)$ is well defined. Furthermore, with $P_1(c)$ differentiable over $[0, c_L]$, the derivative of $v(c)$ is well defined, and can be obtained and set to zero.

$v(c^*)$ is the best the injured terminal can do, with a price low enough to induce the interferer to pay and operate. If $c > c_L$ (say $c = c_L + \varepsilon$), the interferer will choose *not* to operate, deriving a total utility of D_1, its original monetary budget. In this case, the injured terminal will have the channel to itself, will receive no compensating money, and will get the the bits/Joule performance of a random noise channel (say B_0 total bits, given its energy budget). Thus, if terminal 2 sets $c = c_L + \varepsilon$, its utility will be $\beta_2 B_0$. This must be compared against $v(c^*)$, for a final choice of the price c (either c^*

or $c_L + \varepsilon$).

4.5.5 Discussion

A compensation mechanism has been applied to the achieve an efficient allocation of power among two mutually-interfering, data-transmitting terminals. The mechanism is efficient because it induces the terminals to "fairly" compensate each other, by way of money or some transferable good. With 2 mutually interfering terminals, each terminal must quote two prices: one to *be paid to* the other as compensation; the second to *be charged* as compensation. But each terminal faces a penalty if its offered price differs from what the other wants as compensation.

The intuition of this mechanism can best be captured by considering a 2-terminal situation in which only terminal 1 interferes with terminal 2 (but *not* vice-versa), which can actually happen with successive interference cancellation (SIC). Terminal 2 must declare the amount of the transferable good it wishes to *charge* terminal 1 as compensation for each unit of interference. Likewise, terminal 1 must quote the price it offers *to pay* terminal 2 as compensation. But terminal 1 faces a penalty increasing with any difference between its offered price and what terminal 2 demands. At equilibrium, the interfering terminal will pay the true cost caused on the other terminal by its interference, which is precisely the "fair thing to do". This is so because if the amount paid by terminal 1 exceeded the cost its interference causes on terminal 2, then terminal 2 would in fact "make a profit" per unit of interference. But then, it would be optimal for this terminal to induce terminal 1 to *increase* its interference, and to do so, terminal 2 would *decrease* what it charges.

The development provides further insights into the equilibrium allocation, for the special case in which terminal 1 interferes with terminal 2, but *not* vice-versa (SIC decoding). The injured terminal will operate at the optimal SIR, x^* (fig. 3.1), which is the bits-per-Joule-maximizing value a terminal would choose with random noise as its only impairment. The interferer will either stay out of the channel, or pay the exact compensation price c demanded by terminal 2, and operate at the SIR $x_1(c)$ illustrated in fig. 4.1. $x_1(c)$ is always less than x^*. The optimal c maximizes $v(c)$, a relatively simple function (eq. (4.28)) which has a continuous derivative for values of c that are low enough to entice the interferer into operating. A complete characterization and understanding of the power and money allocations arising from this mechanism necessitates additional analytical and numerical work.

This framework can be extended to accommodate many mutually-interfering terminals, and can be applied outside the cellular architecture. With many terminals, the exchange of pricing signals between terminals becomes an issue. However, the fact that terminals only care about the total interference helps, because a terminal's charge per unit of interference should be independent of the source of the interference. But each terminal may, in principle, quote a different value. The rate of convergence toward the equilibrium prices and power levels is also a concern. But [43] shows that a simple updating algorithm exists that leads to the equilibrium, even when terminals don't know "everything" about each other. In an ad-hoc network, the main challenge may be to set up a practical accounting system to track down the compensating payments among terminals.

The impact of this mechanism on several issues involving communication networks should be explored. For instance, it is known that mobile terminals using a cellular system from “bad locations” can stress the system, and reduce its capacity. This can be more severe if a poorly-situated terminal transmits media content (e.g., video) that demands a high data rate, and an inflexible signal-to-interference target. These terminals should, ideally, defer transmission pending a better location, unless their information is “urgent”, which is only known to the transmitter. Implementing a mechanism such as this should induce a more judicious use of the network by these terminals.

Chapter 5

Power and Coding Rate Allocation for Scalably Encoded Information

5.1 Introduction

At the foundations of the JPEG 2000 image compression standard there are ideas found in the embedded zero-tree wavelet coding (EZW) algorithm introduced by [36], a technique which produces a fully "embedded" bit stream[41]. An embedded bit stream is "scalable", in the sense that it can be truncated at an arbitrary point, and decoded. If bits are decoded as they are received, at any instant the "quality" of the decoded information is the best available for the number of bits received up to that moment. Thus, an image compression ratio can be varied simply by truncating the coded bit stream. Similar ideas can be applied to video coding. In fact, fine-granular scalability (FGS) is at the core of the MPEG-4 video-compression standard.

Scalable coding can be fruitfully exploited in many practical applications, including: (i) image database browsing (ii) progressive image transmission (where the consumer can examine the improving decoded image as bits are received, and can abort the transfer when the image quality reaches a satisfactory level), and (iii) multimedia web serving (a single file can serve a variety of consumer requirements and capabilities, and also various congestion/channel conditions).

These files introduce interesting resource management issues, because their special structure can be exploited to allocate scarce resources efficiently. Such analysis necessitates a relatively simple model combining the properties of analytical tractability, with flexibility to accommodate a wide variety of situations. This work proposes such model.

In the situation under study, a terminal with a limited supply of energy and a long sequence of scalably encoded images to transfer over a wireless link seeks to manage its energy efficiently. At the center of this inquiry is a function yielding the "quality" of the resulting information (image) in terms of the fraction of the encoded file which is chosen for decoding. Below, it is postulated that all that is known about this function is that its graph is an S-curve, as introduced in [29] and discussed further in [30] (see fig. 5.2). In chapter 1, this relation was arrived at via rate-distortion theory.

As discussed in previous chapters, there are practical reasons why the S shape is chosen. An arbitrary S-curve starts out convex and smoothly transitions to concave. But, as shown by fig. 5.2, the inflexion (transition) point is arbitrarily placed. Therefore, this curve in fact contains as special cases a ("mostly") concave curve (inflexion point is "very close" to the origin, e.g. U_1) and a ("mostly") convex curve (inflexion point is "very far" from the origin, e.g. U_4). Likewise, some S-curves behave like smoothed out "step" functions (e.g. U_2). And the "ramp" displayed by S-curves such as U_3, can express a near linear relation, over a range of interest. Thus, by assuming an S shape for the function giving the "quality" of the image recovered from the truncated file in terms of the number of decoded bits, this work allows not only the S-shape proper, but also the concave and the convex shape, as well as steps and ramps. These shapes should accommodate most, if not all situations of interest.

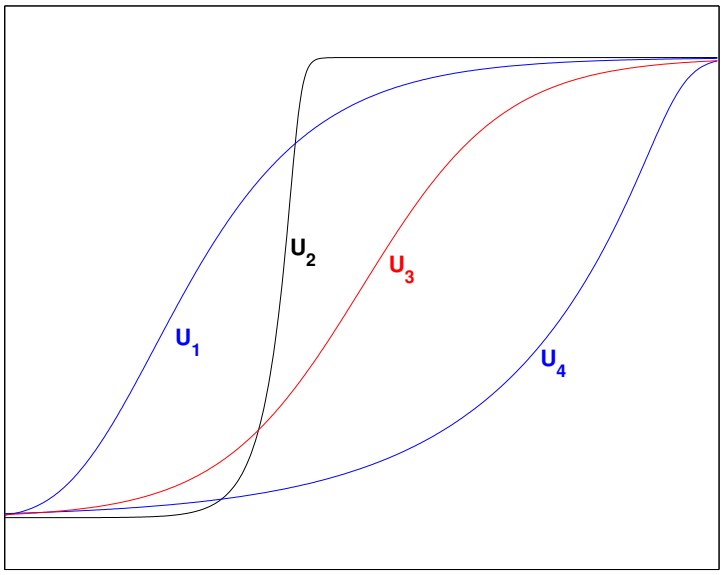

Figure 5.1: Some representative S-curves

Additionally, as discussed in fundamental psychology texts (see, for instance, [8, Chapter 7]), the S-curve naturally arises in psychophysical experiments involving human perception. In these experiments, a graph is made in which the horizontal axis denotes the "intensity" of a stimulus applied to a subject. The vertical axis denotes the probability that the subject correctly identifies or detects the presence of the stimulus. These graphs have often the shape of fig. 5.2.

The peak signal-to-noise ratio (PSNR) is the image quality metric most commonly found in the literature. This is a simple to calculate index, which can be sensible and useful in many situations. However, as an indicator of image quality as perceived by a human observer, the PSNR is at best a very crude measure. Dansereau and Kinsner [4] argues this further, while proposing a metric specifically aimed at progressive image transmission: the Renyi dimension spectrum. But this measure is much too complicated for resource management studies.

This analysis also depends critically on a function giving the probability of success of the transmission of a data packet in terms of a signal to interference measure at the receiver. This "frame-success" function (FSF) is determined by the physical layer of the system. It can be safely assumed that for any physical layer, any such function has an S-shaped graph. Thus, two different S-curves are at the core of this analysis.

The single-terminal case is fully analyzed, and the foundation is laid for a multi-terminal analysis. The problem is set up as a joint optimization in which two key variables are jointly optimized: transmission power, and the number of bits of each file to be decoded. A closed-form solution is given.

The scientific literature contains various works involving power allocation and the transmission of scalably encoded information. The most relevant may be [14], which considers files which have been "layer coded" (a form of discrete scalable coding) and seeks a power allocation policy across the various layers, minimizing the overall end-to-end distortion. However, previous works seeking a joint power, and coding rate selection in order to maximize an image quality metric appear unavailable.

5.2 Conceptual framework

5.2.1 Quality as a function of the number of decoded bits

At the center of this inquiry is a function yielding the quality of the decoded image as a function of the number of bits in the fraction of the encoded file which is decoded. In chapter 1, an argument grounded on rate-distortion theory led to a characterization of a quality-rate curve. Below, an "axiomatic" approach is undertaken. Image quality is a subjective matter. Nevertheless, certain basic assumptions can be made about the properties of a function giving quality as a function of received bits. About this function, it is postulated that:

1) Its domain is the interval $[0, M]$, where M is the length in bits of the entire encoded file.

2) Its range is the interval $[0, 1]$. This is just a normalization. A 1 denotes the best possible quality of the decoded image (say the quality of the original).

3) It must be strictly increasing (more decoded bits yield a better image quality, by design)

4) Its graph is S-shaped, as in figure (5.2). In practical terms, sigmoidness further implies that: (a) If the number of decoded bits is sufficiently large, the quality of the decoded image will be sufficiently close to "perfect". (b) After a sufficiently large number of bits have been decoded, the marginal contribution to image quality of an additional bit becomes "very small" and is decreasing. (c) If the number of decoded bits is sufficiently small, the quality of the decoded image will be sufficiently close to zero. (d) Bits at the beginning of the encoded file contribute to the perceived "quality" of the image at an increasing rate ("initial convexity"). One plausible interpretation is that even a highly distorted image may provide enough information to identify its "meaning" (what is it? a bird?, a person's face?, etc.). This essential semantic information is provided by the bits at the beginning of the encoded file ("base layer") .

Chapter 2 and references [30, 29] discuss the technical characterization of a generic S-shaped function. A fixed function could work for different images, in particular if the images are sufficiently "similar" (e.g., each corresponds to a passport picture of a respective adult).

5.2.2 A Generalized frame-success function

The frame-success function (FSF) yields the probability that a data packet is received successfully as a function of the received signal to interference ratio. This function is determined by physical attributes of the system, including the modulation technique, the forward error detection scheme, the nature of the channel, and properties of the receiver. It is assumed that all that is known about the FSF, f_s, is that its graph exhibits a sigmoidal shape as in figure (5.2). More specifically, it is assumed that the function defined by $f(x) = f_s(x) - f_s(0)$ obeys the properties of the generalized sigmoidal function introduced in [29] and discussed further in [30].

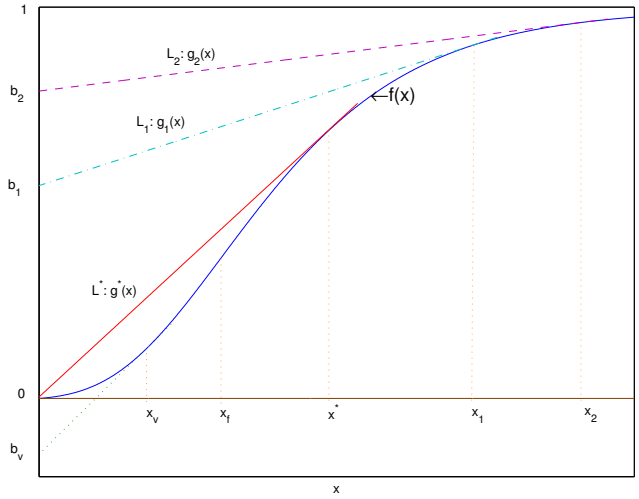

Figure 5.2: An S-curve and some of its tangents

5.3 Single-user analysis

5.3.1 Problem statement

The problem faced by a single transmitter in a wireless (in particular CDMA) network can be formulated as follows.

It is taken as given a (1) certain amount of energy, E, available for transmission, (2) fixed transmission rate of R bits per second, (3) long sequence of files each of length M, each divided into blocks of bits (packets/frames) of length $L \ll M$ and each corresponding to equally important *similar* images encoded scalably, (4) function g as defined in section 5.2.1 giving the quality of an image obtained by decoding a truncated encoded file as a function of the numbers of bits decoded,

(5) certain level of interference (noise), (6) function f_s as described in section 5.2.2 giving the probability that a data frame is received successfully as a function of the signal to interference ratio at the receiver.

The transmitter wants to choose optimally (i) the number of successfully received bits at which point a given file can be considered successful, so that the transmission of the next file is started (that is, the "optimal" level of image quality at which point it is considered "good enough"), and (ii) the transmission power. The objective is to maximize the weighted number of images transferred by the time the available energy runs out, where the weight is the quality of each image. This criterion can also be stated as maximizing the "total quality" transferred.

Because packets have been assumed much smaller than a file, the fact that the number of bits in a frame and file is an integer is ignored. Because the images are similar enough (e.g. each image corresponds to a (respective) human face), the same function works for all images.

5.3.2 Objective function

Suppose that at a certain instant of time, $y < M$ bits of the current file have been received. Then, $g(y) \leq 1$ gives the quality of the image that would result if the file containing the received bits is decoded.

Let $Q = P \cdot h$ be the power at the receiver when a certain data packet is transmitted with power P; and let I be the interference (noise) power. Then, $f_s(GQ/I)$ is the probability that said packet is correctly received (G is a CDMA constant, the spreading/processing gain).

Assuming that, once a packet is received in error, re-transmission is ideal, then the total number of times a given packet needs to be (re)-transmitted is a geometric random variable, whose probability distribution is of the form $\pi(1-\pi)^{k-1}$, with $\pi = f_s(GQ/I)$. The expected value of this random variable is $1/\pi$, interpreted as the average number of times the same packet needs to be transmitted to ensure correct reception.

The average amount of energy that needs to be spent in order to achieve the successful reception of an image of quality $g(y)$ when transmission power is set to P can be obtained as follows. Each packet requires an amount of energy equal to the product of 3 factors: the power P, the length in time of a packet (given the transmission rate R), and the average number of times the same packet needs to be transmitted to ensure correct reception. Each L-bit packet lasts L/R secs. Therefore, the average amount of energy required by a packet is $P \cdot (L/R) \cdot (1/\pi)$. Since y bits of data contain y/L packets, the average amount of energy necessitated by the successful reception of an image of quality $g(y)$ is given by

$$\frac{PL}{\pi R} \cdot \frac{y}{L} = \frac{Py}{\pi R} \tag{5.1}$$

To obtain the average number of images of quality $g(y)$ which can be successfully transmitted with an energy budget E, we divide E by the preceding expression (eq. (5.1)), and obtain:

$$\frac{\pi RE}{Py} \equiv RE\,\frac{f_s(GhP/I)}{Py} \tag{5.2}$$

To obtain the total received "image quality", the preceding expression needs to be multiplied by $g(y)$, the quality of each image. Therefore, the user wants to maximize

$$RE\,\frac{f_s(GhP/I)g(y)}{Py} = RE\,\frac{f_s(GhP/I)}{P}\,\frac{g(y)}{y} \tag{5.3}$$

For technical reasons discussed in [30], $f_s(x)$ is replaced with $f(x) = f_s(x) - f_s(0)$. Then we can re-write equation () as

$$\frac{REGh}{I}\,\frac{f(GhP/I)}{GhP/I}\,\frac{g(y)}{y} \propto \frac{f(x)}{x}\,\frac{g(y)}{y} \tag{5.4}$$

with $x := GhP/I$.

5.3.3 Optimization Model and Solution

In view of the preceding analysis, the objective of the single user can be summarized as

$$\max\ \frac{f(x)}{x}\,\frac{g(y)}{y} \tag{5.5}$$

$$\text{s.t. } 0 \le y \le M \tag{5.6}$$

$$0 \le x \le \bar{x} \tag{5.7}$$

where $\bar{x} := Gh\bar{P}/I$ with $\bar{P}$ the largest available transmission power.

Notice that the ratios in the objective function (5.5) are mutually independent; i.e., one does not influence or constrain the other. Therefore, the ratios $f(x)/x$ and $g(y)/y$ can be maximized independently, and the maximum of the product of the ratios can be obtained as the product of the individual maxima. This problem can be easily solved by invoking the results provided by chapter 2 and references [30, 29]. These works discuss finding the maximum of the ratio $f(x)/x$ s.t. $0 \le x \le \bar{x}$ where all that is known about f is that its graph is S-shaped. The maximizer is the lesser of $\bar{x}$ and x^*. $x^*$is the abscissa of the unique point where the graph of f is tangent to a ray emanating from the origin (See figure (5.2)).

From the preceding paragraph, the maximum of $f(x)/x$ s.t. $0 \le x \le \bar{x}$ is obtained at $x^{**} = \min\{x^*, \bar{x}\}$. Likewise, the maximum of $g(y)/y$ s.t. $0 \le y \le M$ is obtained at $y^{**} = \min\{y^*, M\}$. The single-user problem is solved.

5.4 Discussion

The problem faced by an energy-limited terminal with a long list of scalably encoded similar images to transfer over a wireless link has been solved. A tractable model, based on two "S-curves", has been discussed. A closed-form solution is given in terms of a point which can be easily identified in the graph of the pertinent S-curve. The analysis leads to the maximization, over an appropriate region, of the product $Rf(x)/P \times g(y)/y$, where x is the received SIR, P is the transmission power, R the data transmission rate, f is the "frame success" function, y is the chosen number of decoded

bits, and g is the "quality" function. $Rf(x)/P$ has the unit bits/Joule, well known in the power control literature (see chapter 4 and reference [30]), while quality/bit is the unit of $g(y)/y$. Hence, the maximized product is an intuitively appealing index in quality/Joule.

Although the problem is set up as a joint optimization of power and coding rate, the development indicates that both variables can be "decoupled". In retrospect, this seems reasonable. The files are transmitted in small segments (data packets) which are assumed much smaller than the files, constant, and independent of y, the number of bits chosen for decoding. Power is needed to increase the probability that a data packet is received successfully. But the physical layer treats each packet in the same way, irrespective of the file to which it belongs, or its position within its file. Thus, the point, y, at which a given file is truncated to start the transmission of the next file has no effect on the probability of success of the intervening packets. Future research could consider the possibility that packet length be a variable dependent on y (a shorter packet length for a smaller y).

The S-curve practically contains as special cases a strictly convex and a strictly concave curve. However, it is shown in chapters 2 and 3, that, if f (respectively, g) were strictly concave, the ratio $f(x)/x$ (respectively, $g(y)/y$) would be maximized at zero. In this case, the power level, $\propto x$, (respectively, the "truncation point", y) should be set as small as possible. Likewise, if f (respectively, g) were strictly convex, then the power level, (respectively, the "truncation point") should be set as large as possible.

This analysis can be extended to include many terminals sharing a CDMA channel. In this case, each terminal's "noise" must include the interference caused by others. The problem can be set up as a "game" in which each terminal seeks to maximize its quality/Joule index. In this formulation, the key question is the existence and characterization of a "Nash equilibrium" (NE); i.e., a feasible allocation (of power and file size here) to each terminal, such that no terminal would be better off by *unilaterally* changing its allocation. Both of the ratios ($f(x)/x$ and $g(y)/y$) making up the quality/Joule index are quasi-concave [29]. It is well-known that a game in which "payoff" functions are quasi-concave, and each player's "strategy space" (power and file size here) is closed and bounded does have a NE. Game theory has been fruitfully applied to the transmission of conventional data over a wireless channel in chapter 4, and in other works, such as [20, 33].

Chapter 6

Coding Rate and Power Allocation for Scalably Encoded Video Streaming

6.1 Introduction

Modern media encoders, such as those in the the JPEG 2000 (still images) and MPEG-4 (video) compression standards, support scalability. Fine granular scalability produces an "embedded" bit stream, which can be truncated at an arbitrary point, and decoded, leading to various levels of reproduced media quality. Video scalability can be achieved along various dimensions, including SNR, spatial (size), temporal (frame rate), and frequency; and these scalability modes may be combined [46, Ch.11].

In the present chapter, the model discussed in chapter 5 for still images is extended to consider the transfer over a wireless link of scalably encoded video. This chapter partially overlaps the material in chapter 1. Each T secs of video leads to a Y-bit embedded bit stream, which is independent of the other segments. For example, T may correspond to one group of pictures (GOP), or several GOPs, in video coded according to MPEG standards. An energy-limited terminal seeks to jointly optimize both the truncation point of the embedded bit stream (coding rate), and its transmission power.

We postulate that *all that is known* about the function yielding the "utility" or "quality" of the resulting video segment in terms of the number of bits in the truncated file (coding rate) is that its graph is an S-curve. As shown in fig. 5.1, this family of curves contains as special cases ("mostly") concave curves (e.g, U_1), ("mostly") convex curves (e.g., U_4), and smoothed out "step" functions (e.g. U_2). And the "ramp" displayed by S-curves such as U_3, can express a near linear relation, over a range of interest. These shapes should accommodate most, if not all situations of interest. Other reasons for adopting this family are given in chapters 1 and 5.

Another critical function is that giving the probability of success of the transmission of a data packet in terms of a signal to interference measure at the receiver. It can be safely assumed that for any physical layer, any such function has an S-shaped graph. Thus, two different S-curves are at the

core of this analysis.

The scientific literature registers at least one previous use of the idea of maximizing end-user utility in video streaming in [18], later extended to [19]. But that work focuses on a wired network with renegotiable CBR services, does not consider scalability, and only considers a logarithmic utility function. There are also various works involving power allocation and the wireless transmission of video. Typically, power is minimized, and possibly other parameters are adjusted, while holding "end-to-end" distortion to an acceptable level. For instance, [47] specifically targets scalably-encoded video, while seeking an optimal power allocation, with joint source-channel coding. However, previous works seeking a joint power, and coding rate selection in order to maximize a video quality metric within an analytical model appear unavailable.

Below, we describe the system model, and discuss more formally the key functions. Then, after formally stating the problem, we build and analytically solve an optimization model, and provide a numerical example. We conclude by discussing our results, and commenting on possible extensions.

6.2 Conceptual framework

6.2.1 System model

Fig. 6.1 shows schematically the system engaged in the wireless transmission of scalably encoded live video. Each T secs of video is encoded as a fully embedded bit stream of length Y, which may be truncated to length y. For $y \leq Y$, the reproduced video is imperfect. Its quality or utility is $u(y)$, with u an increasing function discussed below. The bit stream is broken up into packets. Each packet may have added error-control bits (error-control system *not* shown). These packets enter a large buffer prior to transmission. Packets are wirelessly transmitted at the rate of R bps. To ensure continuous video play out at the receiver, the actual transmission time alloted to the y bits corresponding to a given T-sec segment is $\Delta \leq T$ secs. (i.e., the coding rate cannot exceed $R\Delta/T$). A $\Delta < T$ may account for processing and propagation time not being modeled, and a certain "guard time". The probability that a packet is successfully received is $f_s(x)$, with x the signal-to-interference ratio (SIR) at the receiver, which is determined by the chosen transmission power, any path loss, and the interference (noise) present at the receiver. The function f_s is discussed further below. Packets received in error which cannot be corrected result in ideal re-transmissions until correctly received and confirmed. Correctly received packets are placed in a large buffer. Other symbols shown in fig. 6.1 are discussed as introduced below.

6.2.2 Quality as a function of the coding rate

At the core of this inquiry is a function yielding the quality or utility of the decoded video as a function of the number of bits in the truncated encoded file. This function cannot be derived; it is fully determined by the end-user, in the same way in which the "utility function" at the core of economic studies resides within the consumer. $u(y)$ could be obtained by psychophysical experimentation.

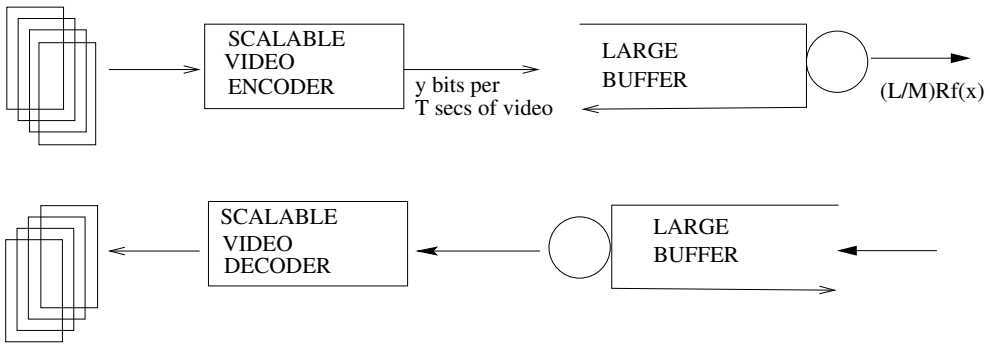

Figure 6.1: Schematic diagram of the wireless streaming of scalably encoded video.

We postulate that this function is such that its graph is an S-curve. Some of the implications of this assumption are discussed further in section 5.2.1. A fixed function could work for different video segments, in particular if the segments are sufficiently "similar" (e.g., each corresponds to different parts of the same sporting event). Further justification is found in chapters 1 and 5. Chapter 2 discusses the technical characterization of a generic S-curve.

6.2.3 A Generalized frame-success function

The frame-success function (FSF) yields the probability that a data packet is received successfully as a function of the signal to interference ratio at the receiver. This function is determined by physical attributes of the system, including the modulation technique, the forward error detection scheme, the nature of the channel, and properties of the receiver. We assume that *all that is known* about the FSF, f_s, is that its graph exhibits a sigmoidal shape as in figure (6.2). For good technical reasons similar to those discussed in chapter 3, $f(x) := f_s(x) - f_s(0)$ replaces f_s in the analysis below ($f_s(0)$ is generally very small, but not zero).

6.3 Analysis

For our purposes, it is convenient to regard the wireless channel as if it was a deterministic channel producing the throughput that the actual channel produces on the average. Thus, we assume that, when the SIR at the receiver is x, $(L/M)Rf(x)$ information bits are received each second at the decoder buffer. The intuition is as follows. With a perfect channel, each packet would be filled with information bits (no ECC), and would be received successfully at first try. Thus, R information bits would be received each sec. However, with an imperfect channel, $M-L$ ECC bits are introduced in each packet, and still, on the average, only $nf(x)$ out of every n packets are received successfully. Thus, an average of $(L/M)Rf(x)$ information bits are successfully transferred each second.

6.3.1 Problem statement

It is taken as given a (1) certain amount of energy, $\bar{E}$, available for transmission, (2) fixed transmission rate of R bits per second, (3) long sequence of files, each of length $y \leq Y$, each divided into packets of length $L \ll y$ and each corresponding to a video segment of length T secs which has been encoded scalably ($L-M$ error-control bits are added to each packet), (4) maximal time $\Delta \leq T$ secs. to complete the transmission of the y bits corresponding to a given T-sec segment (i.e., the coding rate cannot exceed $R\Delta/T$) (5) utility/quality function u as defined in section 6.2.2, (6) certain level of interference (noise), I, (7) frame-success function f_s as described in section 6.2.3.

The transmitter wants to choose optimally (i) the truncation point (coding rate) and (ii) the transmission power, in order to maximize the sum of the quality or utility of each one of the video segments that can be viewed at the receiver before energy runs out.

6.3.2 Objective Function

For a given level of desired quality, $\bar{u}$, there is a corresponding number of information bits, y, that produces this quality ($u(y) = \bar{u}$). Thus, the total number of information bits received successfully after Δ secs. must be not less than this y. And spending energy to exceed this level would be unwise, because it would decrease the total number of segments of quality $\bar{u}$ that are delivered before energy runs out. Thus, for given y and Δ, the terminal must choose its transmission power so that

$$\frac{L}{M} R f(x) \Delta = y \tag{6.1}$$

There is one specific SIR value, $x(y)$, that satisfies eq. (6.1), and a specific transmitted power, $P(y)$, that yields the SIR $x(y)$ at the receiver. Thus, for a given Δ, y determines the transmission power.

The total amount of energy spent on the transmission of a video segment of quality $u(y)$ is $P(y)\Delta$. Thus, the total number of T-sec video segments of quality $u(y)$ that can be transferred with an energy budget of $\bar{E}$ is $\bar{E}/(P(y)\Delta)$. Then, the total quality viewed, which the terminal wishes to maximize, is

$$\frac{\bar{E}}{\Delta} \frac{u(y)}{P(y)} \tag{6.2}$$

For a fixed level of energy, $\bar{E}$, the terminal only needs to maximize $u(y)/(\Delta P(y))$ (quality per Joule), and if Δ is also fixed, just maximize $u(y)/P(y)$, the quality-to-power ratio (QPR).

6.3.3 Optimization Model and Solution

In view of the preceding analysis, the objective of the single user can be expressed as maximizing $u(y)/P(y)$. Assuming a CDMA technology, with a spreading gain of $G := R_c/R$ (chip rate over bit rate), channel gain of h, and interfering power I, the received SIR and the transmitted power are

related as $x = GPh/I$. Thus, the terminal objective is equivalent to :

$$\begin{array}{lll} \max\limits_{x,y} \dfrac{u(y)}{x} & & \max\limits_{x} \dfrac{u(Bf(x))}{x} \\ \text{s.t. } y = Bf(x) & \text{OR} & \text{s.t. } 0 \le x \le \bar{x} \\ 0 \le x \le \bar{x} & & \end{array}$$

where $B := (L/M)R\Delta$ and $\bar{x} := Gh\bar{P}/I$ with $\bar{P}$ the largest available transmission power.

With $u(Bf(x)) := s(x)$, the terminal should maximize the ratio $s(x)/x$. It can be shown that, as shown in fig. 6.2, the composite function $u(Bf(x))$ retains the S-shape of both u and f. As discussed in [30, 29], for *any* S-curve S, $S(x)/x$ is always maximized at x^*, the abscissa of the tangency point between the S-curve and a straight line that passes through the origin.

6.3.4 Numerical example

Fig. 6.2 summarizes a numerical example. We assume that it has been experimentally determined that, for this end-user, the utility or quality function$u(y) = [1+\exp((60-y)/10)]^{-1}$(plotted at the top of fig. 6.2). With x denoting SIR at the receiver, the frame-success function is assumed to be $f_s(x) = [1-\frac{1}{2}\exp(x/2)]^{80}$(whose graph is second from the top), which corresponds to non-coherent FSK modulation, no FEC and 80-bit packet size. Suppose that T secs of video can be scalably encoded, at full rate, to $Y = 100$ (in some multiple of bits). The parameters R, L, M, and Δ are such that $B = (L/M)R\Delta = 110$ (in the same unit as Y). The third subplot corresponds to the composite function $u(Bf(x)) := s(x)$, which clearly retains the S-shape of both u and f. The terminal must choose its transmission power so that the ratio $s(x)/x$ (plotted at the bottom) is maximized. The maximizer is $x^* \approx 10.5$, and its matching truncation point is $y^* \approx 110 * f(10.5) = 88$. Thus, for this user, under this physical layer, the scalable file should be truncated to about 88% its size, leading to a per-segment video quality of about 94% that of the original.

6.4 Discussion

We have investigated the problem faced by an energy-limited terminal transferring over a wireless link a long sequence of files, each corresponding to a segment of video which has been scalably encoded, as supported by the MPEG-4 standard. We have discussed a tractable analytical model, based on two key functions: $u(y)$ which gives the perceptual quality or utility of a video segment as function of the coding rate, and $f(x)$, the packet success probability as function of the signal-to-interference ratio (SIR) at the receiver. By assuming that *all that is known* about these 2 functions is that they are S-curves, we are de facto allowing the possibility that ("mostly") concave, convex, "step", and linear functions play those roles (fig. 5.1). We have postulated that the terminal wishes to maximize the "cumulative utility" (or quality) from all the segments that reach the receiver before energy runs out. Our analysis has led us to maximize the quality-to-power ratio, which is equivalent to maximizing quality per Joule. Although we have set up the problem as a joint optimization of

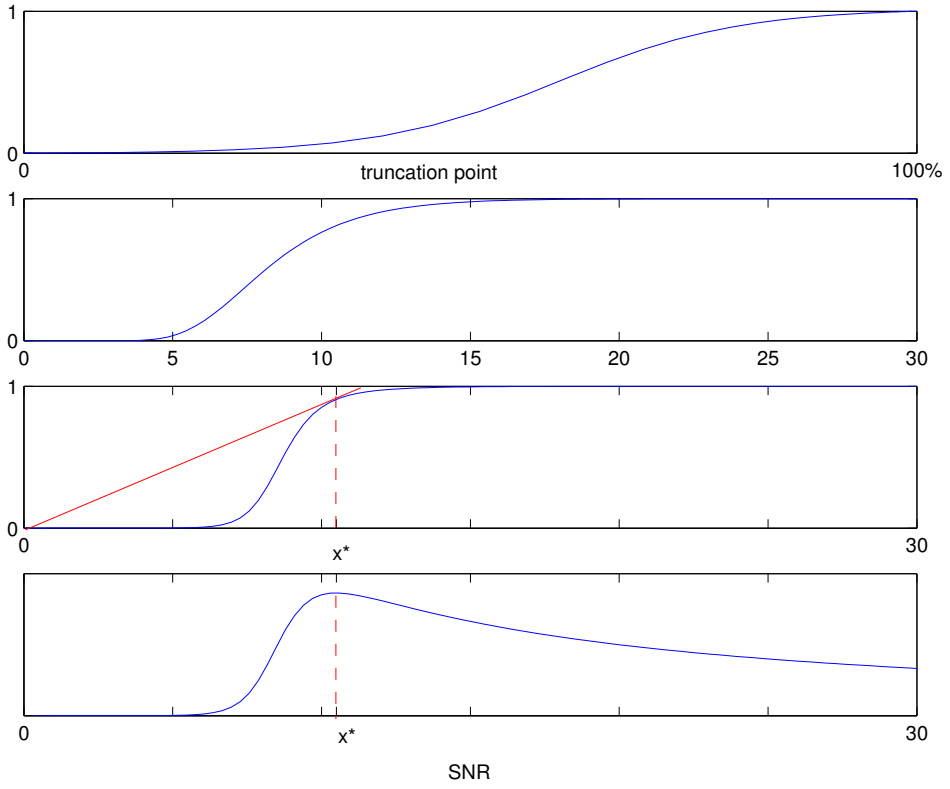

Figure 6.2:
From the top, (i) the S-curve $u(y)$ giving the perceptual quality of a video segment, as a function of the coding rate, (ii) $f(x)$, the probability of successful reception of a packet as a function of the SIR, (iii) the composite function $u(Bf(x)) := s(x)$, (iv) the ratio $s(x)/x$ which the terminal should maximize. For *any* S-curve S, $S(x)/x$ is always maximized at x^*, found at the tangency point between the S-curve and a straight line from the origin.

power and coding rate, our analysis indicates that, when the transmission time is constrained by the underlying streaming application, any one of these variables fully determines the other. The terminal should choose its transmission power so that the received SIR x maximizes the ratio $u(Bf(x))/x$, which occurs at the tangency point between a straight line from the origin, and the graph of the composite function $u(Bf(x))$ (also an S-curve). If the terminal lacks sufficient power to reach that SIR, it should operate at maximal power, unless the resulting video quality is unacceptably low.

Direct implications of our analysis include: (i) if $u(y) \approx ky$ so that the quality-coding-rate relation is nearly linear, the optimal SIR is determined by the physical layer, as the maximizer of $f(x)/x$ (which is *the same* SIR that a data-transmitting terminal would choose, as discussed in chapter 3); (ii) if u behaves like a step function, the terminal should truncate just past the point where the step occurs; and (iii) if f behaves like a step function, then the optimal SIR is just past the point where f jumps.

Even with a fixed physical layer (f function), the optimal operating point could change due to a variation in the perception of quality (u function) at the receiver, or movement that may force the transmitter to operate at an SIR below the optimal level due to power limitations. If the streamed video has been encoded prior to transmission, scalability is essential to achieve such adaptation, via a change in the truncation point of the embedded bit stream. But if coding is being performed concurrent with transmission, a non-scalable encoder that can adapt its rate in real-time could provide a more efficient solution, at a possibly higher computational cost. We can also apply our analysis to optimally choose the coding rate of the non-scalable encoder.

A situation in which several video transmitters share a CDMA channel can be set up as a "game" in which each terminal seeks to maximize its quality-to-power index, with each terminal's "noise" including the interference caused by others. Game theory has been fruitfully applied to the wireless transmission of data, in chapter 4 and in other works, such as [20, 33].

Chapter 7

Quality-Distortion Theory: Distortion Management when Fidelity is Expensive

7.1 Introduction

Distortion measures the difference between a signal and its copy. It is an important QoS measure in the processing and transmission of error-tolerant information, such as media signals intended for human consumption. Typically, when dealing with distortion, the resource-management literature assumes that up to a level, distortion is of no consequence, but beyond that level, it makes the signal totally useless. Such “hard threshold” seems at odds with the way humans process media signals. These signals can be useful at various degrees of noticeable distortion. And when a reduction of distortion is costly, the consumer can prefer more distortion, in exchange for energy, money, or other savings. Furthermore, scientific work has shown that judiciously relaxing the distortion constraint by a small amount can lead, under certain conditions, to a disproportionately larger increase in the capacity of a CDMA network[15].

Hence, a tractable model is needed for the way humans perceive the quality of “imperfect” signals. Below, a model that establishes a quality-distortion relation is (re)introduced. The model is sufficiently flexible to capture a wide variety of plausible quality-distortion relationships, and includes as special cases some of the simpler cases, such as the step function often assumed by the literature. It is postulated that the perceptual quality of an imperfect copy of a signal is determined by a sensible decreasing function of its distortion. No specific algebraic functional form (“equation”) is imposed. Rather, a general family of Q-D functions is assumed. Any such function has the general shape shown in fig. 7.2. This shape can accommodate a wide variety of quality-distortion relations (“step”, “ramp”, convex, concave, etc). Further discussion on this matter is found in chapter 1.

As remarked above, the literature generally assumes that distortion has no noticeable effect up to a certain level, and completely spoils the signal after that level. Reference [18] takes a somewhat more general approach by postulating that the end-user wishes to maximize the “utility” of an im-

perfect media signal. But this reference focuses on video over a wired network, and only considers the special case of a logarithmic utility function.

The quality-distortion curve (first introduced in chapter 1) can also be interpreted as a "utility function" giving the "usefulness" to an observer of an "imperfect" signal. A key difference between perceptual quality and "utility" is that utility is application-dependent. For instance, for a given observer, a level of distortion deemed unacceptable for a "serious" application, may be perfectly acceptable (to the same observer) in a less demanding situation. Because the family of Q-D curves ("utility functions") assumed in the present chapter includes as special cases both the logarithmic and the step function, the present approach is a strict generalization of the literature.

Under this approach, the "right amount" of distortion is a variable to be chosen optimally, whether directly, or, by choosing other resources, indirectly. Below, a situation in which distortion is directly chosen is considered first. A consumer is offered media files at various degrees of distortion. Both his "utility" and the cost of acquiring a file are decreasing in the amount of distortion in the file. With a limited budget, which could be in money, energy, time or any other valuable resource, the consumer faces a classical quantity vs. quality trade-off. He can obtain relatively few high-quality media files, or relatively many low-quality ones. What is the optimal choice? It turns out that with linear pricing the optimal amount of distortion can be quite clearly described. It is obtained by drawing a tangent line from the point $(0,\bar{D})$ to the graph of the utility function ($\bar{D}$ is the largest available distortion level). With non-linear pricing, a similar but somewhat more involved procedure can be applied.

A more specific communication scenario is also considered. An energy-limited transmitter with many media files (images) to transfer over a wireless link wants to choose optimally its transmission power. At low transmission power, many bit errors occur, which produce a highly distorted image at the receiver. High transmission power produces less distortion, at the expense of higher energy consumption per file. Again, a quality vs. quantity trade-off arises. The transmitter opts to maximize the total *weighted* number of files transferred before energy runs out. The weight of each file is its expected "utility" (perceptual quality), which is a function of its distortion. This distortion is a random variable determined by the number of bit errors during the transmission of the file, which is itself determined by the signal-to-interference ratio (SIR), γ, at the receiver. With $\bar{U}(\gamma)$ denoting the expected utility of a media file, the analysis leads to choosing an SIR γ^* to maximize an index in utility/Joule, which is proportional to $\bar{U}(\gamma)/\gamma$. For bit-error functions of practical interest, $\bar{U}(\gamma)$ has the familiar S-shape, and γ^* can be obtained by drawing a tangent line from the origin to the graph of $\bar{U}(\gamma)$ (see x^* in fig. 5.2).

Below, the general properties of the proposed family of Q-D curves are formally given and discussed. Then, the situation in which the degree of distortion of media files can be directly chosen optimally given a cost function is analyzed. Subsequently, the more specific telecommunication problem is solved. Finally, some general summarizing comments are given. (Below, the phrase "perceptual quality" and the word "utility" are used exchangeably. Strictly speaking, a difference could be established between the two, as discussed above)

7.2 Quality/distortion Theory

Distortion is typically defined as a relatively simple mean square measure of the difference between a signal and its copy. As an indicator of media quality as perceived by a human observer, this index is, at best, a very crude measure. The *perceptual* quality of an "imperfect" copy of a signal is determined by the human sensory system (visual, auditory, etc). It seems reasonable to assume that the perceptual quality is somehow determined by distortion; i.e., that a function $Q(D)$ that translates distortion into perceptual quality can be found. The quality-distortion function cannot be derived, and should not be imposed. It should be obtained by psychophysical experimentation. However, one can make some reasonable assumptions about the properties that any such function should possess. Then one can analyze a problem of interest and (optimistically) describe its solution by employing the general properties of the curve.

7.2.1 Intuitive specification

Figure 7.1 illustrates some plausible, simple $Q(D)$ curves. First is, of course, the supposition that perceptual quality falls linearly as distortion increases from zero to its highest value ("quality equals fidelity"). This assumption would greatly simplify the analysis. But it essentially means that the human visual system (HVS) (or auditory, etc) is perfectly "tuned" to a very simple mean squared measure, ..., in all cases, ..., for all people. Such a strong assumption would be adventurous, and likely to be refuted by experimentation. Another highly simplifying assumption often employed in the literature is that distortion is unnoticeable up to a level (c in fig. 7.1) but it totally spoils the signal beyond that point ($Q(D)$ is a "step function"). But our own experience tells us that media signals can be useful at various degrees of noticeable distortion. Furthermore, when a reduction of distortion is costly, a human may choose to tolerate more distortion, in exchange for energy, money or other savings. But the step function assumption precludes the study of such trade-offs. A third possibility illustrated in fig. 7.1 is the "ramp" $Q(D)$, implying that distortion has no noticeable effect up to a level (a), and completely spoils the signal beyond another level (b), while varying linearly between these two points. Presumably, a and b would be determined by the specific user/application combination. The ramp includes as special case the threshold ($a = b = c$) and the linear relation ($a = 0$, $b = D_{\mathrm{MAX}}$); but still its "piecewise linearity" is a big imposition which may not be supported by experimentation.

Further reflection indicates that it is reasonable to assume that the graph of the $Q(D)$ function is a "reversed" S-curve, as shown by fig. 7.2. This graph strictly generalizes the step function often assumed in the literature. And the family of S-curves includes as special cases curves that are "mostly" convex, others that are "mostly" concave, and some whose "ramps" follow closely a straight line over a given interval. Thus, if the analyst assumes that *all that is known* about the $Q(D)$ curve is that it is a reverse S-curve, and conducts the analysis on the basis of properties derived from this shape, the solution procedure and conclusions will be valid for a wide variety of plausible $Q(D)$ relations.

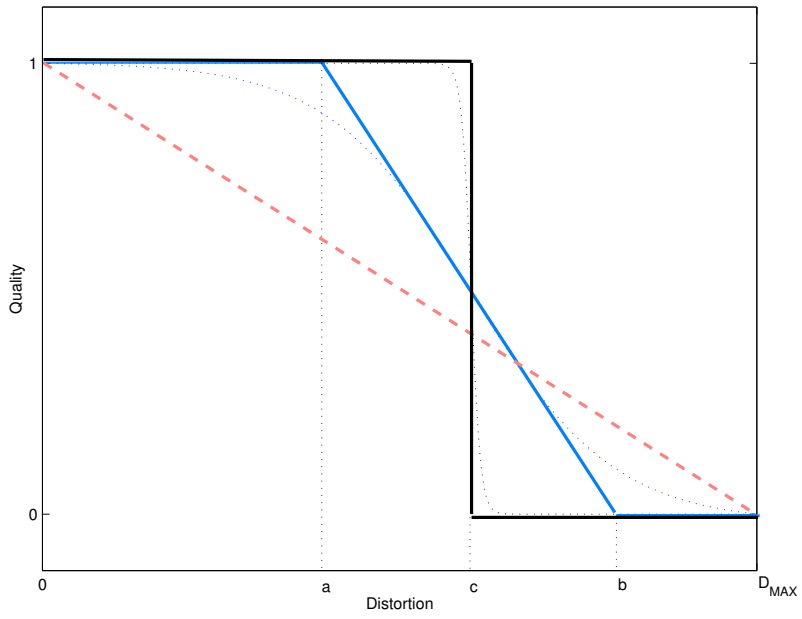

Figure 7.1:
Quality vs. distortion: Some plausible simple relations are: (i) fidelity equals quality (red dashed line); (ii) hard threshold (step); (iii) ramp (blue broken line). The ramp includes as special case the threshold ($a = b = c$) and the linear relation ($a = 0$, $b = D_{\mathrm{MAX}}$). But the reverse S-curve includes all these cases and more (see next figure).

7.2.2 Formal definition

The Q-D curve (“utility function”) has the following properties:

1) Its domain is the interval $[0, \bar{D}]$, where $\bar{D}$ is the largest available level of distortion.

2) Its range is the interval $[0, 1]$. This is just a normalization. A 1 denotes the best possible quality of the decoded file (say the quality of the original) and a zero is the ’quality’ of a maximally distorted file .

3) It is strictly decreasing (distortion worsens quality)

4) Its graph is “reversed” S-shaped, as in fig. 7.2. In practical terms, reversed-sigmoidness further implies that: (a) If the distortion is sufficiently small, the quality of the decoded file will be sufficiently close to “perfect”. (b) After distortion has been sufficiently reduced, the marginal contribution to media quality of further reductions of distortion becomes “very small” and is decreasing. (c) If the distortion is sufficiently large, the quality of the decoded image will be sufficiently close to zero. (d) The function becomes convex as distortion increases (“eventual convexity”). One plausible interpretation is that even a highly distorted image may provide enough information to identify its “meaning” (what is it? a bird?, a person’s face?, etc.). This essential semantic information is provided at high levels of distortion. Thus, the utility of the distorted image *increases at a fast rate* as distortion is *reduced from its highest level* (right to left in the graph).

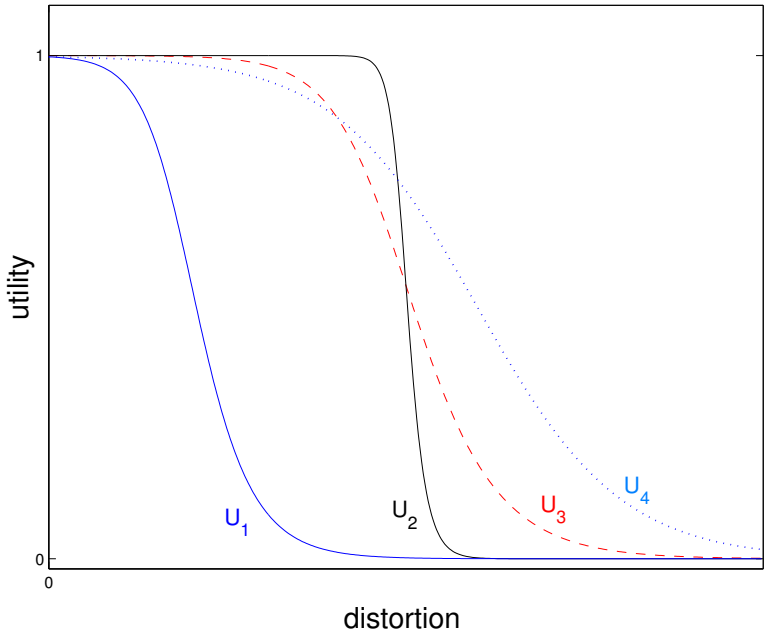

Figure 7.2:
In the eyes of the beholder: Media signals can be useful to end users at various degrees of noticeable distortion. This is captured by a “utility function” indicating the “usefulness” of the distorted signal.

7.2.3 An alternate view: fidelity vs. distortion

Rather than basing the argument on the distortion, y, of the recovered signal, one can focus on the variable $x = \bar{D} - y$, interpreted as the "fidelity", or the amount of distortion which has been "avoided" or "removed". When $x = 0$ the resulting signal is "fully" distorted ($y = \bar{D}$). We can think of this as a signal obtained by guessing all the bits in the concerned file, which yields the "cheapest" possible image. To get an image with any less distortion necessitates some kind of expenditure. The larger x (the difference between $\bar{D}$ and y), the higher the quality of the image, and the greater its cost. Thus, this analysis can be based on the derived function $s(x) := u(\bar{D} - y)$. The graph of $u(-y)$ is the "mirror image" of that of u ("time reversal"). And the graph $u(\bar{D} - y)$ is the same as that of $u(-y)$ but shifted to the right $\bar{D}$ units. Thus, the graph $s(x)$ yields a "standard" S-curve, as displayed in fig. 5.2. This observation will prove useful in the technical development.

7.3 Acquiring Variably Distorted Information

Pedagogically, it may be useful to set up the problem of interest in a general scenario, before introducing communication issues.

7.3.1 Problem statement

A consumer can acquire files corresponding to perceivable media (say images), each available at varied degrees of distortion, $y \in [0, \bar{D}]$. The cost of any one image (in terms of money, energy, or any other scarce resource that the consumer has and values) is $c(y)$, which is always positive and decreasing in the level of distortion, y. For convenience, let $c(\bar{D}) = 0$ and $c(0) = c_0$. Images are equally valuable to the consumer, in the sense that he is indifferent between any one of two images, if they both have the same level of distortion. The usefulness, quality, or "utility" to the consumer of a distorted image is determined as a function of its distortion, y, by a function $u(y)$, whose properties are discussed in section 7.2.2.

The consumer wants to spend his budget B optimally. That is, he wants to determine, given u, c and B, what is the "right" amount of distortion he should choose. If he chooses to acquire images with very small distortion ($y \approx 0$), the cost of each image, $c(y)$, will be "high", and the number of images he will get to view, $B \div c(y)$, will be small. On the other hand, choosing a large y will result in a large number of highly distorted images.

Notice that, as discussed in section 7.2.3, the problem can be stated in terms of $x = \bar{D} - y$, which is interpreted as the amount of distortion which has been "avoided" or "removed from" the image, or simply its "fidelity". In this case, the pertinent cost function is denoted as $c_x(x)$.

7.3.2 Objective Function and Constraints

Some reflection indicates that the consumer should maximize his total utility, which is obtained as the product of the quality (or utility) of each image by the total number of images he gets acquire.

Hence, the consumer should solve

$$\max_{0\le y\le D} \quad \frac{u(y)}{c(y)} \tag{7.1}$$

$$\text{or } \max_{0\le x\le D} \quad \frac{s(x)}{c_x(x)} \tag{7.2}$$

(multiplying by the constant B would make no difference to the solution).

Now, the index being maximized, $u(y)/c(y)$ or $s(x)/c_x(x)$, has the unit quality/dollar, quality per Joule, or quality per second, depending upon the customer's scarce resource.

7.3.3 First-order optimizing conditions

The first-order necessary conditions (FONOC) for an interior solution to this problem is

$$c(y)u'(y) = c'(y)u(y) \tag{7.3}$$

$$\text{or } c_x(x)s'(x) = c'_x(x)s(x) \tag{7.4}$$

Inspection of this equation immediately indicates that if $c(y) \propto u(y)$ then any value of y (or x) would satisfy it.

7.3.4 Solutions

7.3.4.1 Linear cost function

If c is such that $c(y) = (\bar{D} - y)\bar{c}$, ($c_x(x) = \bar{c}x$) then the objective function (eq. (7.3)) can be written, as

$$\max_{0\le x\le D} \frac{u(D-x)}{x} = \frac{s(x)}{x} \tag{7.5}$$

As discussed in section 7.2.3, the graph of $s(x)$ has the form shown in fig. 5.2; that is, $s(x)$ is a standard "S-curve". The solution to maximizing $s(x)/x$, with s an S-curve, is well understood. It is the unique positive number obtained as the abscissa of the point at which a tangent line emanating from the origin meets the graph of s. (see x^* in fig. 5.2). The optimal distortion level is $y^* = \bar{D} - x^*$. Equivalently, the desired solution can be obtained by drawing a tangent from the point $(\bar{D}, 0)$ to the graph of the original $u(y)$.

7.3.4.2 General cost functions

The preceding development can be extended, with due attention to certain technical details, to a more general cost function. The key step is to make the non-linear coordinate transformation. For instance, suppose that $c(y) = (D-y)^2$. Let $t := (D-y)^2$, so that $y = D - \sqrt{t}$. The objective function can then be written as:

$$\max_{0\le t\le D^2} \frac{u(D-\sqrt{t})}{t} := \frac{s_t(t)}{t} \tag{7.6}$$

It can be argued that the graph of $s_t(t)$ is still a "stretched" S-curve. Hence, the value that maximizes $s_t(t)/t$ can be obtained, under appropriate technical assumptions, as before, by drawing a tangent line from the origin to the graph of $s_t(t)$.

7.4 Distortion and power management

Below, the analysis focuses on the more specific scenario of transmission of error-tolerant files ("media") over a wireless link. For simplicity, each information bit in a file is viewed as corresponding to a pixel of an *uncoded* black and white image.

7.4.1 Problem statement

It is taken as given: (1) a certain amount of energy, E, available for transmission; (2) a fixed transmission rate of R bits per second; (3) a long sequence of files, each corresponding to an equally important image, and each divided into N blocks of bits (packets) with a total of M bits, of which L are information bits; (4) a certain level of interference (noise), I. The transmission proceeds one packet at a time, *without* retransmissions. An error-control system is *assumed* to operate as follows. Up to m bit errors per packet can be corrected; and if $m+1 \leq k \leq L$ bit errors occur in a given packet, each will ultimately contribute one error in the decoded file. These errors creates distortion. Thus, there is also a function u as defined in section 7.2 giving the utility (quality) of a received file as a function of its distortion.

The signal-to-interference ratio (SIR) at the receiver determines the bit error probability. Thus, a larger transmission power leads to fewer errors, statistically lower values of distortion, and greater *expected* utility. But, with limited energy, more transmission power means fewer total images transferred. The transmitter wishes to utilize its energy efficiently.

7.4.2 Distortion analysis

The error-control system is viewed as a "black box" whose net effect is that a packet with $m+1 \leq k \leq L$ bit errors contribute k errors to the decoded file. Distortion is, generally, defined as of sum of squares of differences between the reconstructed signal and the original. This sum equals the total number of bit errors in the reconstructed image, in this scenario.

For example, suppose that the number of packets per file is 2, and that the code being used can correct up to 3 bit errors per packet. Suppose that 2 and 5 bit errors occur during the transmission of the first and second packet, respectively. Then, the first packet is corrected, so that all its information bits coincide with the original. But the 5 errors in the second and final packet are not corrected, and contribute 5 errors in the decoded file. Thus, the total distortion of this image will be 5. The utility function of the user will determine how good or bad a distortion of 5 is.

It is worth noting that it is not obvious, at least in this problem, what is the worst case scenario for distortion. In principle, it would seem that having each and every bit in error should be the worst

that can happen. However, given the idiosyncrasies of the human visual system, if a black and white image were to have each and every bit reversed, the result would be a perfectly intelligible image, in which black and white simply switch roles! However, this fact is not considered in the analysis below.

7.4.3 Expected utility of distorted image

In this scenario, distortion is a discrete random variable. The transmission power determines the bit-error rate (BER), and frame-error rate (FER), which indirectly determines the probability distribution of distortion. When the number of packets per file , N, is large, expressing this probability distribution in terms of the BER is quite cumbersome and tedious. This task is, however, relatively straightforward when each image fits into a single packet. Let this be the case. Under the assumptions that have been made about the error-control system, distortion is zero, if m or less bit errors have occurred during the transmission of the packet. When the number of bit errors exceed the number that can be corrected by the code, what happens depends on more specific details of the error control system. Let us assume, pessimistically, that if $m+1$ to L errors occur, each will cause an error among information bits in the decoded file.

Assuming independent bit errors, the probability of k bit errors in an L bit packet is given by $\binom{M}{k}\varepsilon^k(1-\varepsilon)^{M-k}$, with ε the bit-error rate (BER) which is determined byγ, the signal-to-interference ratio (SIR) at the receiver.

For the single-packet file, the ***expected utility*** of a file $U_E(\gamma)$ is

$$u(0)\underbrace{\left(\sum_{k=0}^{m}\binom{M}{k}\varepsilon^k(1-\varepsilon)^{M-k}\right)}_{\text{Prob of 0 to m bit errors}} + \sum_{k=m+1}^{L}\binom{M}{k}\varepsilon^k(1-\varepsilon)^{M-k}u(k) \tag{7.7}$$

7.4.4 Solution

The expected utility function $U_E(\gamma)$ is a representative measure of the expected quality of each image, given a transmission power level, P, which determines the received SIR, γ. Notice, however, that the BER is 1/2 when $\gamma = 0$, which means that $U_E(0) > 0$. To avoid technical problems involving "transmissions" with 0 power, $\bar{U}(\gamma)$:=$\bar{U}(\gamma)$-$\bar{U}(0)$, the "earned" expected utility of an image, is chosen as the representative quality figure of merit (see chapter 3 for a relevant discussion involving error-intolerant data transmissions).

Since each bit lasts $1/R$ secs., (R is the transmission bit rate), the total energy consumed by the transmission of the single-packet image is PM/R. Thus, $ER \div MP$ images can be transferred with E Joules. The transmitter wishes to maximize its total (earned) expected utility, and must solve:

$$\max_{0\le P\le \bar{P}} \frac{R}{M}\frac{\bar{U}(\gamma)}{P} \equiv \max_{0\le\gamma\le\bar{\gamma}} \frac{R_c}{M}\frac{h}{I}\frac{\bar{U}(\gamma)}{\gamma}$$

with h the path loss, I the interference, R_c the "chip rate" (a CDMA constant closely related to the bandwidth), $\bar{P}$ the highest available transmission power, and $\bar{\gamma} = (R_c/R)h\bar{P}/I$, the highest achievable SIR.

It can be argued that for BER functions of practical interest, the graph of $\bar{U}(\gamma)$ has the S-shape displayed in fig. 5.2. Then, by the argument given in section 7.3.4.1, the value γ^* which maximizes $\bar{U}(\gamma)/\gamma$ can be obtained by drawing a tangent from the origin to the graph of $\bar{U}(\gamma)$. This value determines the transmission power, and solves the single user problem. In any case, it is discussed in chapter 3 that if $\bar{U}(\gamma)$ was convex, the optimal would occur at the highest available power level, and that if $\bar{U}(\gamma)$ was concave it would be optimal to "operate" at zero power.

The preceding development also applies when each media file is divided into many packets. Extending the preceding analysis to consider a multi-packet image file is conceptually simple, but very tedious. The procedure to find the probability distribution of distortion is more cumbersome. But once done, it is straightforward to find the "earned" expected utility of a file as a function of the received SIR, $\bar{U}(\gamma)$. The shape of the graph of this function should not be affected by the number of packets per file.

7.5 Discussion

Media signals can be useful at various degrees of distortion. A proposed model captures this fact mathematically, and enables its exploitation, when avoiding/reducing distortion requires the expenditure of limited resources. Two interesting problems involving a quality versus quantity trade-off are formulated and solved. In one case, media files are offered at various degrees of distortion, at a price that is *decreasing* in distortion. A consumer willing to accept a higher degree of distortion, can acquire more files. A more specific version of this problem involves an energy-limited transmitter wishing to transfer many images over a wireless link. Spending more energy per packet reduces bit errors, and hence distortion, but also leads to fewer images transferred.

At the core is a function relating the perceptual quality ("utility") of an "imperfect" media signal to its distortion; i.e., a quality-distortion (Q-D) curve. In the development, no specific "equation" (logarithmic, logistic, etc) is imposed as a Q-D function. Rather, it is assumed that *all that is known* about this curve is that it belongs to certain family characterized by a "reversed" S-shaped graph. The analysis follows from the general properties of this family; so that it applies to *any* Q-D curve, as long as its graph has the assumed shape. This shape contains as special case the "sharp threshold" (step) often assumed in the literature, as well as many plausible Q-D relations (convex, concave, "ramps", etc). This level of generality is important, because the "true" Q-D curve can only be obtained by psychophysical experimentation with human subjects. The actual curve will, generally, depend on the specific targeted human user, and quite possibly on the specific application. Because of its generality, this analysis and its conclusions are robust, and should hold for many user/application combinations.

Chapter 8

Data Rate and Power Allocation for Throughput Maximization

8.1 Introduction

Modern wireless networks will accommodate simultaneous transceivers operating at very different bit rates. Some of the transceivers may be transferring data, while others transfer media content, such as voice, images, or video. Several technologies have been proposed to accommodate multi-rate traffic in such networks. Reference [26] discuss several multi-rate schemes based on Direct Sequence Code-Division Multiple Access (DS-CDMA). One such scheme is variable spreading gain (VSG) CDMA, as described, for example, in [11]. In a VSG-CDMA system, each terminal's spreading gain is determined as the ratio of the common chip rate to the terminal's bit rate.

The model discussed in this chapter is relevant to an interference-limited single-cell VSG-CDMA system in which each data terminal can operate within a range of bit rates, which is assumed continuous for tractability. An allocation specifying, for each active terminal, a choice of data rate and power level is sought that will maximize the network weighted throughput. The weights admit various interpretations, including levels of importance or priority, "utilities", or monetary prices (contribution to the network's revenues). The traffic is assumed to be delay-tolerant ("best-effort").

Similar situations have been considered by the literature. This formulation has much in common with that of [40]. Major differences between this reference and the present work include (a) the weights (b) the "generalized" frame-success function adopted herein, and (c) the simplifying linearization involved in the solution procedure given in the reference. Reference [16] seeks data rates and power allocation, and consider a "sigmoidal-like" frame-success function, but focuses on the downlink, does not consider weights, and provides a sub-optimal algorithmic solution based on pricing. The present work has also many similarities with [37], which maximizes a fairly general "capacity function". But [37] does not consider weights, and assumes that the terminal's data rates are fixed exogenous parameters, as opposed to variables to be chosen optimally.

At the core of this analysis is the frame-success function (FSF), which gives the probability

that a data packet is received successfully as a function of the terminal's signal-to-interference ratio (SIR) at the receiver. This function is determined by physical attributes of the system, including the modulation technique, the forward error detection scheme, the nature of the channel, and properties of the receiver, including its demodulator, decoder, and antenna diversity, if any. No particular algebraic functional form ("equation") is imposed as FSF. Rather, it is assumed that all that is known about this function is that its graph is a smooth S-shaped curve, as displayed in fig. (8.1) (see chapter 3 for further discussion of this approach). The development exploits properties derived from this shape. Hence, the present analysis should apply to many physical layer configurations of practical interest, as long as they give rise to an FSF that has an S-shaped graph, and satisfies certain mild technical assumptions.

Below, a relatively simple optimization model relevant to uplink data transmission in one VSG-CDMA cell is built. Afterward, an outline of the general solution procedure is provided. Then, the two-terminal special case is completely solved analytically, including the verification of the second-order optimality conditions. This case is thoroughly discussed, as it provides insights useful for the general analysis. Subsequently, the analysis focuses on a specific N-terminal scenario. The scenario studied is one in which a few equally "important" terminals share a cell with many "ordinary" terminals. It is presumed that the system can accommodate all the important terminals at the highest available data rate. But it is not clear how many, if any, of the ordinary terminals should be set to operate at this high rate, in order to maximize the cell's weighted throughput. A general solution procedure for this scenario is given. Finally, the results given in this chapter are discussed, emphasizing the technical limitations of this analysis.

8.2 General Formulation

8.2.1 Problem Statement

We seek to solve:

$$\max_{G_i,\alpha_i} \sum_{i=1}^{N} \beta_i T_i(G_i,\alpha_i) \tag{8.1}$$

subject to

$$\sum_{i=1}^{N} \frac{\alpha_i}{1+\alpha_i} = 1 \tag{8.2}$$

$$G_i \geq G_0 \quad i \in \{1,\cdots,N\} \tag{8.3}$$

In this simple model,

1. N is the number of terminals.

2. The throughput of terminal i is defined as $R_C T_i(G_i, \alpha_i)$, with

$$T_i(G_i, \alpha_i) := \frac{f(G_i \alpha_i)}{G_i} \tag{8.4}$$

3. $G_i = R_C/R_i$, $i \in \{1, \ldots, N\}$ is the spreading gain of terminal i; i.e., the ratio of the channel's chip rate , R_C to the terminal's data transmission rate R_i (bits per second). $G_0 \geq 1$ is the lowest available spreading gain (determined by the highest available data rate).

4. α_i is the carrier-to-interference ratio (CIR) of the signal from terminal i received at the base station. α_i is defined as,

$$\alpha_i := \frac{P_i h_i}{\sum_{\substack{j=1 \\ j \neq i}}^{N} P_j h_j + \sigma^2} = \frac{Q_i}{\sum_{\substack{j=1 \\ j \neq i}}^{N} Q_j + \sigma^2} \tag{8.5}$$

   with P_i the transmission power of terminal i, h_i its "gain" (path loss) coefficient, $h_i P_i := Q_i$ its received power, and σ^2 a representative of the average noise power and, possibly, out-of-cell interference. It is shown in appendix B, that with $\sigma^2 = 0$, the CIR's must be such that $\sum \alpha_i/(1+\alpha_i) = 1$ (constraint (8.2)) to ensure feasibility.

5. The product $G_i \alpha_i$, denoted as γ_i , is terminal i's signal to interference (SIR) ratio.

6. $\beta_i \geq 1$ is a weight, which admits various practical interpretations. Without loss of generality, we set $1 = \beta_1 \leq \cdots \leq \beta_N$. If only 2 classes of terminals are considered, say N_1 "light weight" terminals and N_2 "important" ones, then $1 = \beta_1 = \cdots = \beta_{N_1}$ and $\beta = \beta_{N_1+1} = \cdots = \beta_{N_1+N_2}$ with $N_1 + N_2 = N$.

7. We assume that there is a real-valued frame-success function (FSF) which gives the probability of the correct reception of a data packet in terms of the received SIR. We assume that this function is such that $f(x) := f_S(x) - f_S(0)$ has the general properties of the generalized "S-curve" discussed in chapter 2 (see fig. (8.1)), and that it has a continuous second derivative. Because $f_S(0)$ is very small, the difference between f_S and f is generally negligible. Nevertheless, this correction is made for technical reasons. It is stressed that no actual function is used, except to provide numerical examples. Our analysis should apply to a wide variety of physical layer configurations, as long as they give rise to an FSF with an S-shaped graph. To provide numerical examples, we use the FSF corresponding, under suitable assumptions, to non-coherent FSK modulation, with no FEC, and packet size 80, which is,

$$f(x) = \left[1 - \frac{1}{2}\exp\left(-\frac{x}{2}\right)\right]^{80} \tag{8.6}$$

8. Certain technical results require a few additional assumptions that are stated when needed, and discussed at the end of the chapter.

It is sometimes useful to observe that constraint (8.2) can be expressed as

$$\sum_{i=1}^{N} \frac{1}{1+\alpha_i} = N-1 \tag{8.7}$$

In the development below, an asterisk used as a superscript on a variable denotes a specific value of the variable which satisfies certain optimality condition. Terminals operating at maximal data rate are referred to as "favored" or "favorite", and terminals in the high-weight class are called "important", as opposed to "ordinary". Some ordinary terminals may be "favored" in the sense that they may be allowed to operate at the highest available data rate.

8.2.2 General solution procedure

The general procedure is as follows:

- Create an "augmented" objective function, combining the original objective function with Lagrange multipliers and the constraint equations
- Set up the first-order necessary optimizing conditions (FONOC). This involves setting the partial derivative of the *augmented* objective function with respect to each variable equal to zero. Moreover, inequalities of the form $G_0 - G_i \leq 0$ contribute equations of the form $\mu_i(G_0 - G_i) = 0$, (complementary slackness condition), where μ_i is a Lagrange multiplier.
- Solve FONOC. Evidently, each equation of the form $\mu_i(G_0 - G_i) = 0$ requires that if $G_i > G_0$, then μ_i must equal zero; and that if $\mu_i \neq 0$, G_i must equal G_0. Both possibilities must be considered separately while finding various solutions to FONOC. It is necessary that each μ_i be non-positive, for a maximizer.
- A solution to FONOC provides a candidate for a maximizer. The second-order sufficient conditions (SOSC) *may* confirm the candidate as a maximizer. This maximizer may *not* be global. If the SOSC are not verified, then the solution is obtained by directly verifying which of all the points satisfying FONOC yields the highest weighted throughput.

8.3 Special Case: N=2

For pedagogical reasons, a two-terminal-only situation is considered first.

To be solved:

$$\text{Maximize } \frac{f(G_1\alpha_1)}{G_1} + \frac{\beta f(G_2\alpha_2)}{G_2} \tag{8.8}$$
$$\text{subject to } \alpha_1\alpha_2 = 1 \text{ ; } G_1 \geq G_0 \text{ ; } G_2 \geq G_0$$

It can be easily verified that for N=2, the constraint (8.2) reduces to $\alpha_1\alpha_2 = 1$. This also follows from the fact that, with negligible noise, $\alpha_1 := Q_1/Q_2 := 1/\alpha_2$.

8.3.1 Augmented objective function

The augmented objective function is

$$\phi(G_1, G_2, \alpha_1, \alpha_2) = \frac{f(G_1\alpha_1)}{G_1} + \frac{\beta f(G_2\alpha_2)}{G_2} + \lambda(1 - \alpha_1\alpha_2) + \sum_{i=1}^{2} \mu_i(G_0 - G_i) \tag{8.9}$$

8.3.2 First-Order Necessary Optimizing Conditions (FONOC)

The FONOC can be expressed in vector form, with $\gamma_i = G_i\alpha_i$, as:

$$\begin{bmatrix} (\gamma_1 f'(\gamma_1) - f(\gamma_1))/G_1^2 - \mu_1 \\ \beta(\gamma_2 f'(\gamma_2) - f(\gamma_2))/G_2^2 - \mu_2 \\ f'(\gamma_1) - \lambda\alpha_2 \\ \beta f'(\gamma_2) - \lambda\alpha_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{8.10}$$

$$\text{with} \begin{cases} \alpha_1\alpha_2 = 1 \\ \mu_1(G_0 - G_1) = 0 \quad \mu_1 \leq 0 \\ \mu_2(G_0 - G_2) = 0 \quad \mu_2 \leq 0 \end{cases} \tag{8.11}$$

8.3.3 Hessian Matrix

In order to check the sufficient second-order conditions the Hessian matrix of second partial derivatives of the augmented objective function , denoted as $\phi\mathbf{2_{xx}}$, is needed. This matrix is given by:

$$\phi\mathbf{2}_{xx} = \begin{bmatrix} \psi(G_1, \alpha_1) & 0 & \alpha_1 f''(\gamma_1) & 0 \\ 0 & \beta\psi(G_2, \alpha_2) & 0 & \beta\alpha_2 f''(\gamma_2) \\ \alpha_1 f''(\gamma_1) & 0 & G_1 f''(\gamma_1) & -\lambda \\ 0 & \beta\alpha_2 f''(\gamma_2) & -\lambda & \beta G_2 f''(\gamma_2) \end{bmatrix} \tag{8.12}$$

In equation (8.12), strictly for notational convenience, the function ψ is defined, with $\gamma_i = G_i\alpha_i$, as:

$$\psi(G_i, \alpha_i) = \frac{2}{G_1^3}\left[f(\gamma_i) - \gamma_i f'(\gamma_i) + \frac{1}{2}\gamma_i^2 f''(\gamma_i)\right] \tag{8.13}$$

8.3.4 Finding the optimizer

8.3.4.1 Looking inside the feasible region

It is natural to start looking for a solution to FONOC that lies in the interior of the feasible region. That is, $\mu_1 = \mu_2 = 0$ is set, which allows both G_1 and G_2 to be greater than G_0 (see equations (8.11)).

8.3.4.1.1 An Interior solution to FONOC Working with the top 2 rows of the matrix equation (8.10), $\gamma_i f'(\gamma_i) = f(\gamma_i)$ is obtained, which is an equation of the general form:

$$x f'(x) = f(x) \tag{8.14}$$

Chapter 2 shows that for the class of generalized sigmoidal functions, such as f, there is a unique positive value γ_0 which satisfies equation (8.14). This value can be graphically identified in figure (8.1) as the abscissa of the point where the graph of f is tangent to a ray emanating from the origin; that is, tangent to the straight line $y = f'(\gamma_0)x$.

Therefore, if any values of the variables of interest satisfy, under the stated hypotheses, equations (8.10) and (8.11) , they must be such that:

$$G_1^* \alpha_1^* = G_2^* \alpha_2^* = \gamma_0 \tag{8.15}$$

By working with the bottom half of the matrix equation (8.10), it is established that:

$$\lambda = \frac{f'(G_1^* \alpha_1^*)}{\alpha_2^*} = \frac{\beta f'(G_2^* \alpha_2^*)}{\alpha_1^*} \tag{8.16}$$

Now, substituting equation (8.15) into equation (8.16), $\alpha_1^*/\alpha_2^* = \beta$ results, which leads to a complete "interior" solution to FONOC:

$$\begin{bmatrix} G_1 \\ G_2 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} \gamma_0/\sqrt{\beta} \\ \sqrt{\beta}\gamma_0 \\ \sqrt{\beta} \\ 1/\sqrt{\beta} \end{bmatrix} \tag{8.17}$$

Notice that, in order for these values to be feasible, $G_i^* \geq G_0$; i.e., $G_0\sqrt{\beta} \leq \gamma_0$. Replacing these values into the objective function yields

$$T_B = \frac{f(\gamma_0)}{G_1^*} + \frac{\beta f(\gamma_0)}{G_2^*} = \frac{f(\gamma_0)\sqrt{\beta}}{\gamma_0} + \frac{\beta f(\gamma_0)}{\gamma_0\sqrt{\beta}} \tag{8.18}$$

This is a closed form solution. If the function f is known, γ_0 can be easily obtained graphically (see figure (8.1)) or equation(8.14) can be solved numerically. For instance, for the FSF given by equation (8.6), $\gamma_0 = 10.75$, $f(\gamma_0) = 0.83$.

This allocation has an interesting property: it is 'balanced' in the sense that both users experience the same weighted throughput: $f(\gamma_0)\sqrt{\beta}/\gamma_0$.

8.3.4.1.2 Verifying the Second-order sufficient conditions To characterize the interior "stationary point" that was just found, the second order conditions, which depend upon $\phi \mathbf{2}_{xx}$, the matrix of second partial derivatives (Hessian matrix) of the augmented objective function ϕ.

Essentially, at a point satisfying the FOC, i.e., a "stationary" point, for any vector $\vec{h}$ along a feasible direction, the triple product $\vec{h}^{\mathbf{T}} \phi \mathbf{2}_{xx} \vec{h}$ is positive if the "stationary" point corresponds to a local minimum, and this product is negative if the stationary point corresponds to a local maximum. If neither of these conditions hold, then the point is a "saddle point".

A feasible direction is one that is tangent to the curve representing the constraint relationship. Hence, denoting the constraint curve as $b(G_1, G_2, \alpha_1, \alpha_2) = \alpha_1\alpha_2 - 1 = 0$, only vectors $\vec{h}$ satisfying $\nabla\mathbf{b} \bullet \vec{h} = 0$ needs to be considered, that is, vectors normal to the gradient of the constraint curve. At the interior stationary point,

$$\nabla\mathbf{b} = \begin{bmatrix} 0 \\ 0 \\ \alpha_2^* \\ \alpha_1^* \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1/\sqrt{\beta} \\ \sqrt{\beta} \end{bmatrix} \propto \begin{bmatrix} 0 \\ 0 \\ 1 \\ \beta \end{bmatrix}$$

Then, it is easily verified that any vector $\vec{h}$ of the form $\begin{bmatrix} a_1 & a_2 & \beta a_3 & -a_3 \end{bmatrix}^{\mathrm{T}}$, where the a_i's are arbitrary real numbers, satisfies $\nabla\mathbf{b} \bullet \mathbf{h} = 0$.

It will prove convenient to express such vector as the product of a "transformation" matrix, $\mathbf{M}$ times an arbitrary vector $\vec{a} = \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix}^{\mathbf{T}}$. It is trivial to verify that

$$\mathbf{M} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \beta \\ 0 & 0 & -1 \end{bmatrix} \quad \text{is such that} \quad \tilde{\mathbf{h}} \triangleq \mathbf{M} \times \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix}$$

satisfies the desired condition.

In terms of $\mathbf{M}$ and $\vec{a}$, the second-order conditions for the stationary point under consideration can be re-stated as follows. At such point, for any vector $\vec{a}$, the product $\vec{a}^{\mathbf{T}}\mathbf{M}^{\mathbf{T}}\phi_{xx}\mathbf{M}\vec{a}$ is positive if the stationary point corresponds to a local minimum, and this product is negative if the stationary point corresponds to a local maximum. If neither of these conditions hold, then the stationary point is a "saddle point".

Because the components of $\vec{a}$ are arbitrary, the above conditions can be expressed in terms of the matrix $\mathbf{M}^{\mathbf{T}}\phi_{xx}\mathbf{M}$. This matrix is positive definite if the stationary point corresponds to a local minimum, and it is negative definite if the stationary point corresponds to a local maximum. If this matrix is indefinite this point is a "saddle point".

The matrix of second-partial derivatives is given by equation (8.12), which must be evaluated at the point of interest, given by equation (8.17). For these values, $\phi\mathbf{2}_{xx}$ becomes, with $\rho_0 = f'(\gamma_0)/f''(\gamma_0)$:

$$\phi\mathbf{2}_{xx} = \begin{bmatrix} \frac{\beta}{\gamma_0} & 0 & 1 & 0 \\ 0 & \frac{1}{\gamma_0\beta} & 0 & 1 \\ 1 & 0 & \frac{\gamma_0}{\beta} & -\rho_0 \\ 0 & 1 & -\rho_0 & \beta\gamma_0 \end{bmatrix} \sqrt{\beta} f''(\gamma_0)$$

Some algebra yields:

$$\frac{\mathbf{M}^{\mathrm{T}} \times \phi\mathbf{2}_{xx}}{\sqrt{\beta} f''(\gamma_0)} = \begin{bmatrix} \frac{\beta}{\gamma_0} & 0 & 1 & 0 \\ 0 & \frac{1}{\gamma_0 \beta} & 0 & 1 \\ \beta & -1 & \gamma_0 + \rho_0 & -\beta(\gamma_0 + \rho_0) \end{bmatrix}$$

And some more algebra yields:

$$\frac{\mathbf{M}^{\mathrm{T}} \times \phi\mathbf{2}_{xx} \times \mathbf{M}}{\sqrt{\beta} f''(\gamma_0)} = \begin{bmatrix} \frac{\beta}{\gamma_0} & 0 & \beta \\ 0 & \frac{1}{\gamma_0 \beta} & -1 \\ \beta & -1 & 2\beta(\gamma_0 + \rho_0) \end{bmatrix} \tag{8.19}$$

Given the development in chapter 2, $\mathbf{f}''(\gamma_0)$ **is negative**. Thus, if the matrix $\left(\mathbf{M^T}\phi\mathbf{2}_{xx}\mathbf{M}\right)/f''(\gamma_0)$ is positive definite the point being tested is a local maximum. If all three principal minor determinants of a matrix are positive, the matrix is positive definite.

The first determinant is simply the first element of the matrix, which clearly is a positive number. The second determinant is $1/\gamma_0^2$, which is also positive. However, after some algebra the determinant of the whole matrix is obtained as:

$$\frac{2\beta f'(\gamma_0)}{\gamma_0^2 f''(\gamma_0)}$$

But, this expression is negative, because the first derivative of f is positive everywhere, and its second derivative is negative at γ_0.

Hence, the first two principal minor determinants are positive, while the third one negative. The concerned matrix is indefinite. Therefore, the interior stationary point is neither a local minimizer nor a local maximizer. It is a "saddle point".

8.3.4.2 A Single Favorite Boundary Solution (SFBS)

In the preceding section, an interior solution to FONOC was identified. But that allocation is a non-maximizer, which suggests that a maximizer be sought over the "boundary" of the feasible region; i.e., when $G_i = G_0$ for one or both i. Below, single favorite boundary solution (SFBS) solution, in which the important terminal is the only one transmitting at the highest allowable data rate, is found. That is, $G_2 = G_0$, and $\mu_1 = 0$ (which allows $G_1 \geq G_0$) are set. (With only two terminals, the phrase "single favorite" is redundant, since there can be at most one favorite. But the phrase is kept because it has a similar usage in the poly-terminal scenario)

8.3.4.2.1 Finding the SFBS For the reader's convenience, equation (8.10) is reproduced below:

$$\begin{bmatrix} (\gamma_1 f'(\gamma_1) - f(\gamma_1))/G_1^2 - \mu_1 \\ \beta(\gamma_2 f'(\gamma_2) - f(\gamma_2))/G_2^2 - \mu_2 \\ f'(\gamma_1) - \lambda\alpha_2 \\ \beta f'(\gamma_2) - \lambda\alpha_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Working with the first row of equation (8.10), and keeping in mind that$\mu_1 = 0$ has been set,

$$G_1\alpha_1 = \gamma_0 \tag{8.20}$$

is obtained, with γ_0 as defined by equation (8.14).

Working with the bottom half of equation (8.10), and using the preceding result, it is established that:

$$\frac{f'(\gamma_1)}{\alpha_2} = \lambda = \frac{\beta f'(\gamma_2)}{\alpha_1} \tag{8.21}$$

Combining equations (8.20) and (8.21), one obtains

$$\frac{G_0^2 f'(\gamma_0)}{G_0\alpha_2} = \beta f'(\gamma_2) G_0\alpha_2 \quad \Rightarrow \quad \frac{x^2 f'(x)}{f'(\gamma_0)} = \frac{G_0^2}{\beta} \tag{8.22}$$

with $x := G_0\alpha_2 = \gamma_2$. Hence, α_2^* is obtained by solving equation (8.22).

It is observed that, for the class of functions being considered, $x^2 f'(x)$ is a "bell-shaped" function, as shown by figure 8.1.

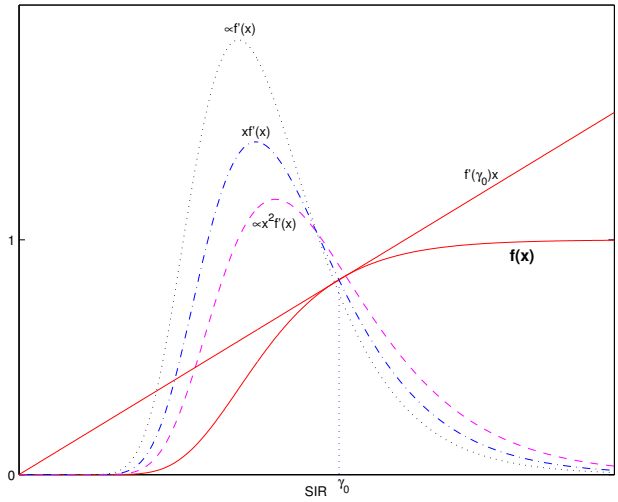

Figure 8.1: A particular $f(x)$, $xf'(x)$, and *scaled* versions of $f'(x)$, and $x^2 f'(x)$. γ_0 satisfies $xf'(x) = f(x)$

This implies that, if G_0^2/β surpasses the "peak" of the function on the left hand side of equation (8.22), then, this equation has *no* solutions. If G_0^2/β is sufficiently small, two values of x will satisfy equation (8.22). Denote the chosen value as δ_0.

Now the second row of equation (8.10) yields the multiplier associated with the constraint $G_0 - G_2 \leq 0$ as

$$\mu_2 = \frac{\delta_0 f'(\delta_0) - f(\delta_0)}{G_0^2/\beta} \tag{8.23}$$

It is necessary for a maximizer that $\mu_2 \leq 0$. This condition is best interpreted by writing it as

$$\frac{G_0^2}{\beta}\mu_2 = t^2 \frac{d}{dt}\left(f(t)/t\right)\Big|_{t=\delta_0} \leq 0$$

The development in chapter 2 shows that, for the class of functions f being considered, the derivative of $f(t)/t$ is positive for $t < \gamma_0$, is zero at γ_0, and is negative for $t > \gamma_0$(with γ_0 defined by equation (8.14)). Thus, in order for δ_0 to lead to a maximizer, it is necessary that

$$\delta_0 \geq \gamma_0 \tag{8.24}$$

As displayed in figure 8.1, it is possible that even if G_0^2/β falls below the maximal value of the function $x^2 f'(x)/f'(\gamma_0)$, it may still be too high, because the resulting intersection points may both be less than γ_0, which would violate a necessary condition for a maximizer.

In view of the preceding development, of the two values satisfying equation (8.22), the larger value, to the right of the peak, is chosen as a prospective maximizer. That is, δ_0 is the largest value satisfying:

$$\frac{\delta_0^2 f'(\delta_0)}{f'(\gamma_0)} = \frac{G_0^2}{\beta} \tag{8.25}$$

In terms of δ_0, a complete solution to FONOC is identified. By definition, $\delta_0 = G_0\alpha_2$, which implies that $\alpha_2^* = \delta_0/G_0$ satisfies FONOC, and obviously so does $\alpha_1^* = 1/\alpha_2^* = G_0/\delta_0$. And since FONOC requires that $G_1^*\alpha_1^* = \gamma_0$, then G_1^* can be obtained as $\gamma_0/\alpha_1^* = \gamma_0\delta_0/G_0$.

Hence, the following single-favorite solution has been found:

$$\begin{bmatrix} G_1^* \\ G_2^* \\ \alpha_1^* \\ \alpha_2^* \end{bmatrix} = \begin{bmatrix} \gamma_0\delta_0/G_0 \\ G_0 \\ G_0/\delta_0 \\ \delta_0/G_0 \end{bmatrix} \tag{8.26}$$

But feasibility requires that $G_1^* \geq G_0$, which imposes that $G_0^2 \leq \gamma_0\delta_0$, in addition to the requirements discussed in the preceding paragraphs. But by definition δ_0 must satisfy equation (8.25). Thus, $G_0^2 \leq \gamma_0\delta_0$ implies that

$$\frac{\beta\delta_0^2 f'(\delta_0)}{f'(\gamma_0)} \leq \gamma_0\delta_0 \rightarrow \beta\delta_0 f'(\delta_0) \leq \gamma_0 f'(\gamma_0) = f(\gamma_0) \tag{8.27}$$

It is observed in figure 8.1, that the function $xf'(x)$ has a bell shaped graph. Thus, in order for condition (8.27) to be satisfied, δ_0 must be significantly larger than γ_0. This further limits the highest value of the ratio G_0^2/β for which the SFBS exists.

For the frame-success function introduced previously as equation (8.6), $\gamma_0 = 10.75$, and $f(\gamma_0) = 0.83$. When $G_0 = 2$ and $\beta = 2$, both $x = 22.1$ and $x = 3.97$ satisfy equation (8.22) . Hence, δ_0 = 22.1. This gives $T_{SFBS} = 1.01$. By comparison, the 'balanced' solution only yields $T_B = 0.15\sqrt{2} = 0.21$,

which is much less.

8.3.4.2.2 Second-order sufficient conditions As discussed in section 8.3.4.1.2, the optimality of the SFBS depends upon the matrix $\phi\mathbf{2}_{xx}$ given by equation (8.12). Essentially, at a point satisfying the FONOC, i.e., a stationary point, if for *any* vector $\vec{h}$ along a feasible direction the triple product $\vec{h}^{\mathrm{T}} \times \phi\mathbf{2}_{xx} \times \vec{h}$ is positive, then the stationary point corresponds to a local minimum, and if this product is negative then the stationary point corresponds to a local maximum.

A feasible direction is one that is tangent to the curve representing the equality constraint, as well as to any curve corresponding to an "active" inequality constraint. An "active" inequality constraint is one satisfied as equality. In the case under discussion, exactly one inequality is presumed to be active: $G_0 - G_2 \leq 0$, since $G_2 = G_0$. Hence, denoting the equality constraint curve as $b(G_1, G_2, \alpha_1, \alpha_2) = \alpha_1\alpha_2 - 1 = 0$, and the active inequality as $d(G_1, G_2, \alpha_1, \alpha_2) = G_0 - G_2 = 0$, only vectors $\vec{h}$ satisfying $\nabla\mathbf{b} \bullet \vec{h} = 0$ AND $\nabla\mathbf{d} \bullet \vec{h} = 0$ need to be considered. It is immediate that:

$$\nabla\mathbf{b}^{\mathrm{T}} = \begin{bmatrix} 0 & 0 & \alpha_2 & \alpha_1 \end{bmatrix} = \begin{bmatrix} 0 & 0 & \frac{1}{\alpha_1} & \alpha_1 \end{bmatrix} \tag{8.28}$$

$$-\nabla\mathbf{d}^{\mathrm{T}} = -\begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix} \tag{8.29}$$

Thus, only vectors $\vec{h}$ of the form $\begin{bmatrix} a_1 & 0 & -\alpha_1 a_2 & \frac{1}{\alpha_1} a_2 \end{bmatrix}^{\mathrm{T}}$ with a_1 and a_2 arbitrary, need to be considered.

The matrix of second-partial derivatives is given by equation (8.12), which must be evaluated at the point of interest, given by equation (8.26). For these values, $\phi\mathbf{2}_{xx}$ becomes:

$$\phi\mathbf{2}_{xx} = \begin{bmatrix} \frac{G_0^3}{\gamma_0\delta_0^3} f''(\gamma_0) & 0 & \frac{G_0}{\delta_0} f''(\gamma_0) & 0 \\ 0 & \beta\psi_{00} & 0 & \frac{\beta\delta_0}{G_0} f''(\delta_0) \\ \frac{G_0}{\delta_0} f''(\gamma_0) & 0 & \frac{\gamma_0\delta_0}{G_0} f''(\gamma_0) & -\frac{\beta\delta_0}{G_0} f'(\delta_0) \\ 0 & \frac{\beta\delta_0}{G_0} f''(\delta_0) & -\frac{\beta\delta_0}{G_0} f'(\delta_0) & \beta G_0 f''(\delta_0) \end{bmatrix}$$

with

$$\psi_{00} = \frac{2}{G_0^3}\left[f(\gamma_{00}) - \gamma_{00} f'(\gamma_{00}) + \frac{1}{2}\gamma_{00}^2 f''(\gamma_{00}) \right]$$

For $\vec{h}^{\mathrm{T}} = \begin{bmatrix} a_1 & 0 & -\alpha_1 a_2 & \frac{1}{\alpha_1} a_2 \end{bmatrix}$, $\vec{h}^{\mathrm{T}} \times \phi\mathbf{2}_{xx}$ is obtained as

$$\begin{bmatrix} a_1 & 0 & -\frac{G_0}{\delta_0} a_2 & \frac{\delta_0}{G_0} a_2 \end{bmatrix} \begin{bmatrix} \frac{G_0^3}{\gamma_0\delta_0^3} f''(\gamma_0) & 0 & \frac{G_0}{\delta_0} f''(\gamma_0) & 0 \\ 0 & \beta\psi(G_2, \alpha_2) & 0 & \beta\alpha_2 f''(\delta_0) \\ \frac{G_0}{\delta_0} f''(\gamma_0) & 0 & \frac{\gamma_0\delta_0}{G_0} f''(\gamma_0) & -\frac{\beta\delta_0}{G_0} f'(\delta_0) \\ 0 & \frac{\beta\delta_0}{G_0} f''(\delta_0) & -\frac{\beta\delta_0}{G_0} f'(\delta_0) & \beta G_0 f''(\delta_0) \end{bmatrix} =$$

$$\begin{bmatrix} \frac{G_0^3 f''(\gamma_0)}{\gamma_0 \delta_0^3} a_1 - \frac{G_0^2 f''(\gamma_0)}{\delta_0^2} a_2 \\ \frac{\beta \delta_0^2 f''(\delta_0)}{G_0^2} a_2 \\ \frac{G_0 f''(\gamma_0)}{\delta_0} a_1 - \gamma_0 f''(\gamma_0) a_2 - \frac{\beta \delta_0^2 f'(\delta_0)}{G_0^2} a_2 \\ \beta \left(f'(\delta_0) + \delta_0 f''(\delta_0) \right) a_2 \end{bmatrix}^{\mathrm{T}}$$

Now, $\vec{h}^{\mathrm{T}} \times \phi \mathbf{2}_{xx} \times \vec{h}$ is obtained by scalarly multiplying the vector just obtained by the vector $\begin{bmatrix} a_1 & 0 & -\alpha_1 a_2 & \frac{1}{\alpha_1} a_2 \end{bmatrix}^{\mathrm{T}}$, which yields :

$$\frac{G_0^3}{\gamma_0 \delta_0^3} f''(\gamma_0) a_1^2 - \frac{G_0^2}{\delta_0^2} f''(\gamma_0) a_1 a_2 - \frac{G_0^2}{\delta_0^2} f''(\gamma_0) a_1 a_2 \quad + \\ \frac{G_0 \gamma_0}{\delta_0} f''(\gamma_0) a_2^2 + \frac{\beta \delta_0}{G_0} f'(\delta_0) a_2^2 + \frac{\beta \delta_0}{G_0} f'(\delta_0) a_2^2 + \frac{\beta \delta_0^2}{G_0} f''(\delta_0) a_2^2$$

This can be simplified as,

$$\begin{aligned} \frac{G_0^3}{\gamma_0 \delta_0^3} f''(\gamma_0) a_1^2 - 2 \frac{G_0^2}{\delta_0^2} f''(\gamma_0) a_1 a_2 + \frac{G_0 \gamma_0}{\delta_0} f''(\gamma_0) a_2^2 \quad & + \\ 2 \frac{\beta \delta_0}{G_0} f'(\delta_0) a_2^2 + \frac{\beta \delta_0^2}{G_0} f''(\delta_0) a_2^2 \quad & = \\ \frac{G_0 \gamma_0}{\delta_0} f''(\gamma_0) \left(\frac{G_0^2}{\gamma_0^2 \delta_0^2} a_1^2 - 2 \frac{G_0}{\gamma_0 \delta_0} a_1 a_2 + a_2^2 \right) + \frac{\beta}{G_0} a_2^2 \left(2 \delta_0 f'(\delta_0) + \delta_0^2 f''(\delta_0) \right) \quad & = \\ \frac{G_0 \gamma_0}{\delta_0} f''(\gamma_0) \left(\frac{G_0}{\gamma_0 \delta_0} a_1 - a_2 \right)^2 + \frac{\beta \delta_0}{G_0} a_2^2 \left(2 \delta_0 f'(\delta_0) + \delta_0^2 f''(\delta_0) \right) \quad & \end{aligned}$$

The first term is clearly negative, since the development in chapter 2 shows that γ_0 occurs to the right of the inflexion point of f, where this function is concave. The second term is also negative, because $2\delta_0 f'(\delta_0) + \delta_0^2 f''(\delta_0)$ equals the derivative of the function $x^2 f'(x)$ evaluated at δ_0. It is observed in figure 8.1 that $x^2 f'(x)$ is a bell-curve. Thus, if δ_0 is to the right of the "peak" of this function (see discussion immediately preceding equation (8.25)), $x^2 f'(x)$ is *decreasing* at δ_0, which means its derivative is *negative* at δ_0.

8.3.4.3 "Greedy" allocation

The preceding section considered the SFBS, in which *only* the "important" terminal operates at the lowest available spreading gain (highest data rate). It was observed that the SFBS fails to exist or is infeasible if G_0^2/β is "too large". This section seeks a "greedy" (favoriteless) solution to FONOC, in which both terminals operate at the highest available data rate. Specifically, $G_1 = G_2 = G_0$ is set.

8.3.4.3.1 Describing the greedy allocation For the reader's convenience, equation (8.10) is reproduced once again:

$$\begin{bmatrix} (\gamma_1 f'(\gamma_1) - f(\gamma_1))/G_1^2 - \mu_1 \\ \beta(\gamma_2 f'(\gamma_2) - f(\gamma_2))/G_2^2 - \mu_2 \\ f'(\gamma_1) - \lambda\alpha_2 \\ \beta f'(\gamma_2) - \lambda\alpha_1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Working with the last two rows of equation (8.10) it is established that:

$$\lambda = \frac{f'(\gamma_1)}{\gamma_2/G_0} = \frac{\beta f'(\gamma_2)}{\gamma_1/G_0} \Rightarrow \gamma_1 f'(\gamma_1) = \beta\gamma_2 f'(\gamma_2) \tag{8.30}$$

with the constraint

$$\gamma_1\gamma_2 = G_0^2 \tag{8.31}$$

In order to satisfy the general first-order necessary optimizing conditions (FONOC), the SIR's of the important and the ordinary terminals, denoted respectively as x and y, must satisfy equation (8.30), as well as the constraint equation (8.31). Equations (8.30) and (8.31) form a system of two non-linear equations in two unknown which is in principle solvable, and may even be reduced to a single-unknown equation. The solution is best described graphically, through figure 8.2.

It is observed that, for the class of functions being considered, the graph of $t f'(t)$ is "bell-shaped", as displayed at the top of figure (8.2). That is, there is exactly one point t^* at which this function has a global maximum, and for every $t_1 \leq t^*$ there is a $t_2 \geq t^*$ such that $t_1 f'(t_1) = t_2 f'(t_2)$. Thus, for every pair (x_2, y_2) which satisfies equation (8.30), with $x_2 \geq t^*$ and $y_2 \geq t^*$, there is a corresponding pair (x_1, y_1), with $x_1 \leq t^*$ and $y_1 \leq t^*$, which also satisfies this equation, and so do (x_1, y_2), and (x_2, y_1).

For a value of x (the SIR of the important terminal), there are two values of y (the SIR of the ordinary terminal) which satisfies $\beta x f'(x) = y f'(y)$. When all points (x, y) satisfying this equation are plotted, an "X-shaped" graph arises, as shown at the bottom sub-figure of figure 8.2. For a fixed β, this graph has four distinct branches. The "North-East" branch corresponds to points like (x_2, y_2) (top sub-figure), which satisfy $\beta x_2 f'(x_2) = y_2 f'(y_2)$, and are both to the right of the peak of $x f'(t)$. The "South-East" branch corresponds to points like (x_2, y_1), which also satisfies $\beta x_2 f'(x_2) = y_1 f'(y_1)$, with y_1 to the left of the peak. Analogously, the "North-West" and "South-West" branches corresponds to points like (x_1, y_2) and (x_1, y_1), respectively, in the top sub-figure. When $\beta = 1$, all four branches have exactly one common point. In that case, $y = x$ always satisfies equation (8.30), but another possibility exists for any x in the SE branch.

But in order to satisfy FONOC, x and y must also satisfy the constraint equation (8.31). Plotting on the same axes this constraint, gives rise to the hyperbolic (L-shaped) curves. The intersection points between the L-shaped and X-shaped graphs for the given (G_0, β) pair lead to feasible solutions to FONOC.

8.3.4.3.2 Eliminating some candidates It is necessary that $\mu_i = (\gamma_i f'(\gamma_i) - f(\gamma_i))/G_0^2$ (obtained from the top two rows of equation (8.10)) be non-positive, for a maximizer. This condition is best

interpreted by writing it as

$$G_0^2\mu_i = t^2\frac{d}{dt}\left(f(t)/t\right)\Big|_{t=\gamma_i} \leq 0 \tag{8.32}$$

That is, in order for a considered point, say $(G_0, G_0, y/G_0, x/G_0)$, to be a maximizer, it is necessary that the derivative of $f(t)/t$ be non-positive when evaluated at x, and also when evaluated at y. The development in chapter 2 shows that, for the class of functions f being considered, the derivative of $f(t)/t$ is positive for $t < \gamma_0$, is zero at γ_0, and is negative for $t > \gamma_0$(with γ_0 defined by equation (8.14)). Thus, in order for the point $(G_0, G_0, y/G_0, x/G_0)$ to be a maximizer, it is necessary that

$$\min\{x, y\} \geq \gamma_0 \tag{8.33}$$

Figure 8.1 shows that the value γ_0 satisfying $t\,f'(t) = f(t)$ occurs to the right of the peak of the graph of $x\,f'(x)$; that is, $\gamma_0 > t^*$, where t^* is the value that maximizes $t\,f'(t)$. When $\gamma_0 > t^*$, only points in the NE "leg" of the "X" can be maximizers, since a pair (x, y) in any one of the other branches would have at least one the coordinates less than γ_0. However, the possibility that $\gamma_0 \leq t^*$ has *not* been ruled out, theoretically.

8.3.4.3.3 Verifying the Second-order Sufficient Conditions Suppose that the L and X graphs intercept at $(x, y) \equiv (x, G_0^2/x)$.

As discussed in sections 8.3.4.1.2 and 8.3.4.3.3, a point satisfying FONOC leads to a (local) maximum, if at such point, for *any* vector $\vec{h}$ along a feasible direction, the triple product $\vec{h}^{\mathrm{T}} \times \phi\mathbf{2}_{xx} \times \vec{h}$ is negative.

A feasible direction is one that is tangent to the curve representing the equality constraint, as well as to any curve corresponding to an "active" inequality constraint. In the case under discussion, both inequalities are presumed active. Hence, denoting the equality constraint as $b(G_1, G_2, \alpha_1, \alpha_2) = 1 - \alpha_1\alpha_2 = 0$, and the inequality constraints as $d_1(G_1, G_2, \alpha_1, \alpha_2) = G_0 - G_1 = 0$ and $d_2(G_1, G_2, \alpha_1, \alpha_2) = G_0 - G_2 = 0$, only vectors $\vec{h}$ satisfying $\nabla\mathbf{b} \bullet \vec{h} = 0$ AND $\nabla\mathbf{d}_i \bullet \vec{h} = 0$ need to be considered. It is immediate that:

$$\nabla\mathbf{b}^{\mathrm{T}} = -\begin{bmatrix} 0 & 0 & \alpha_2 & \alpha_1 \end{bmatrix} = -\frac{1}{G_0}\begin{bmatrix} 0 & 0 & x & y \end{bmatrix} \tag{8.34}$$

$$-\nabla\mathbf{d}_1^{\mathrm{T}} = \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} \tag{8.35}$$

$$-\nabla\mathbf{d}_2^{\mathrm{T}} = \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix} \tag{8.36}$$

Thus, only vectors $\vec{h}$ of the form $a\begin{bmatrix} 0 & 0 & -y & x \end{bmatrix}^{\mathrm{T}}$ with a arbitrary, need to be considered.

The matrix of second-partial derivatives (reproduced below) is given by equation (8.12), which must be evaluated at the point of interest.

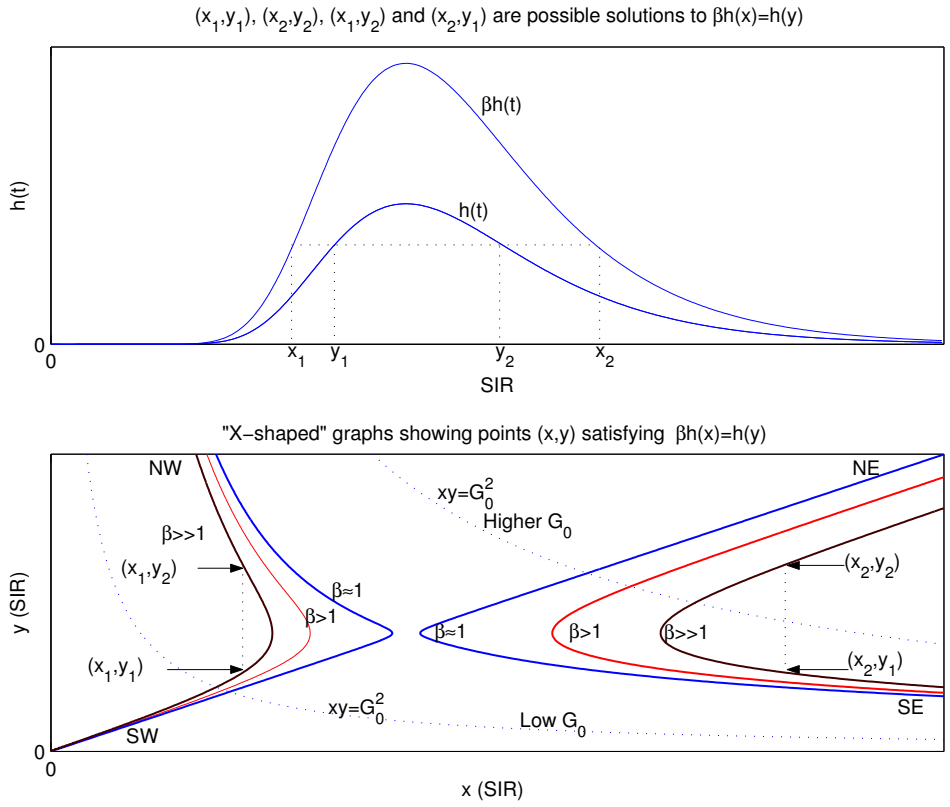

Figure 8.2:
With the SIR of the important and ordinary terminal denoted, respectively, as x and y, FONOC requires that $\beta h(x) = h(y)$ (eq. (8.30)), with $h(t) := t f'(t)$. Any of the pairs (x_1,y_1), (x_2,y_2), (x_1,y_2), or (x_2,y_1) (top) satisfies this equation, but may not be feasible. Plotting all such points reveals an "X-shaped" graph (NE, NW, SW and SE are directional labels). Plotting on the same axes the constraint equation (8.31) gives rise to the hyperbolic (L-shaped) curves. The intersection points between the L-shaped and X-shaped graphs for the given (G_0,β) pair lead to feasible solutions to FONOC. When G_0 is "large", the "L" intersects the NE leg of the "X", which yields a maximizer. If G_0 is low enough, the hyperbola "L" *only* intersects the SW leg of the X-curve, which leads to a minimum.

$$\phi\mathbf{2}_{xx} = \begin{bmatrix} \psi(G_1,\alpha_1) & 0 & \alpha_1 f''(\gamma_1) & 0 \\ 0 & \beta\psi(G_2,\alpha_2) & 0 & \beta\alpha_2 f''(\gamma_2) \\ \alpha_1 f''(\gamma_1) & 0 & G_1 f''(\gamma_1) & -\lambda \\ 0 & \beta\alpha_2 f''(\gamma_2) & -\lambda & \beta G_2 f''(\gamma_2) \end{bmatrix}$$

At the point $(G_0, G_0, y/G_0, x/G_0) = (G_0, G_0, G_0/x, x/G_0)$, taking into account that, from equation (8.30),

$$\frac{G_0 f'(y)}{x} \equiv \frac{y}{G_0} f'(y) = \lambda = \frac{\beta G_0 f'(x)}{y} = \frac{\beta x}{G_0} f'(x)$$

$\phi\mathbf{2}_{xx}$ becomes:

$$\phi\mathbf{2}_{xx} = \begin{bmatrix} * & 0 & * & 0 \\ 0 & * & 0 & * \\ * & 0 & G_0 f''(y) & -\frac{1}{G_0} y f'(y) \\ 0 & * & -\frac{\beta}{G_0} x f'(x) & \beta G_0 f''(x) \end{bmatrix}$$

(An asterisk denotes a non-zero element of this matrix that will not be needed, given the form of the vectors $\vec{h}$ being considered.)

For $\vec{h}^{\mathrm{T}} = \begin{bmatrix} 0 & 0 & -y & x \end{bmatrix}$, $\vec{h}^{\mathrm{T}} \times \phi\mathbf{2}_{xx}$ is obtained as

$$\begin{bmatrix} * & * & -G_0 y f''(y) - \frac{\beta}{G_0} x^2 f'(x) & \frac{1}{G_0} y^2 f'(y) + \beta G_0 x f''(x) \end{bmatrix}$$

$$\frac{1}{G_0} \vec{h}^{\mathrm{T}} \times \phi\mathbf{2}_{xx} \times \vec{h} = y^2 f''(y) + \frac{\beta}{G_0^2} x^2 y f'(x) + \frac{1}{G_0^2} x y^2 f'(y) + \beta x^2 f''(x)$$

Observing that $xy = G_0^2$, the preceding sum can be written as:

$$\beta x f'(x) + \beta x^2 f''(x) + y f'(y) + y^2 f''(y) = \beta x \left(f'(x) + x f''(x) \right) + y \left(f'(y) + y f''(y) \right)$$

Thus, if the sum $\beta x (f'(x) + x f''(x)) + y (f'(y) + y f''(y))$ is negative, the tested point is a (local) maximizer, and if the sum is positive, the point is a minimizer. A deeper understanding of this condition is gained by re-writing the sum as:

$$\beta x \frac{d}{dt} \left(t f'(t) \right) \Big|_{t=x} + y \frac{d}{dt} \left(t f'(t) \right) \Big|_{t=y} \tag{8.37}$$

As displayed at the top of figure 8.2, the graph of the function $t f'(t)$ is "bell-shaped". Thus, $t f'(t)$ has a global maximum at a specific value, say t^*, and its derivative is positive for $t < t^*$, is zero at t^*, and is negative otherwise.

If the point being tested lies on the NE leg of the X, both x and y are to the right of t^*. In this case, both terms in the sum (8.37) are negative, and the tested point is a *maximizer*. Likewise, if the tested point lies on the SW leg of the X, both x and y are to the *left* of t^*. In this case, both terms in the sum (8.37) are positive, and the tested point is a *minimizer*. If the tested point lies on any one of

the other two legs of the X, it is not clear, a priori, whether the sum will be positive or negative.

8.3.4.3.4 The symmetric special case ($\beta = 1$) Considering the special case in which β=1 provides further insights into the general greedy allocation. In this special case, it is evident that $x = y = G_0$ ($\alpha_1 = \alpha_2 = 1$) (equal received powers) satisfies equation (8.30) and the constraint equation (8.31), and hence FONOC. In this case, the sum (8.37) can be written as

$$2G_0 \frac{d}{dt}\left(t f'(t)\right)\Big|_{t=G_0}$$

whose sign is determined by the position of G_0 with respect to the value t^*that maximizes $t f'(t)$. The equal-received power allocation is a maximizer if $G_0 > t^*$, and a minimizer if $G_0 < t^*$.

In particular, for the function given as equation (8.6), $x f'(x)$ reaches its maximum at $t^* = 7.95$. Thus, in this example, with $\beta = 1$, $\gamma_1 = \gamma_2 = G_0$ is a maximizer for $G_0 > 7.95$, but a minimizer for $G_0 < 7.95$.

8.3.5 Discussion of the special case

The optimum power levels and data rates for two terminals transmitting to one base station, in a scenario relevant to variable spreading gain CDMA, have been derived. The objective function is the weighted network throughput, where the weights admit various practical interpretations, including monetary prices paid by the terminals. The analysis identifies three allocations satisfying the first-order necessary optimality conditions (FONOC): (i) a "balanced" allocation, in which both terminals operate at the "preferred" SIR, γ_0, and achieve equal weighted throughput; (ii) an "unfair" assignment in which the important terminal operates at the highest available data rate, with the other terminal achieving the SIR, γ_0; and (iii) a "greedy" assignment in which both terminals operate at the highest available data rate.

The balanced assignment is always suboptimal, implying that "fairness" (in the sense of equal weighted throughput) comes at the expense of performance. The important terminal should always operate at maximal data rate. Only when the ratio $G_0/\sqrt{\beta}$ is larger than certain threshold determined by the physical layer through the FSF should both terminals operate at maximal data rate (G_0 is the smallest available spreading gain and β is the weight of the favorite terminal). This makes intuitive sense, because when G_0 is "large", the highest available data rate is relatively small, and keeping only one terminal operating at maximal data rate is not appealing, unless that terminal has "a lot of weight". However, when the highest available data rate is very high, an allocation in which *only one* terminal operates at this rate is more appealing.

The "greedy" allocation is particularly treacherous, which is particularly clear when both terminals are equally weighted. In this case, an equal-received-power assignment satisfies FONOC. But this assignment can lead to either a maximum or a minimum, depending upon whether G_0 exceeds a specific value determined by the physical layer.

It is significant that the greedy and the unfair allocations are complementary in this sense: a

low G_0 (highest available data rate is large) may turn the greedy allocation into a minimizer, but the unfair allocation, which is a maximizer, needs a low G_0 in order to be feasible.

8.4 Throughput Optimization with N terminals

In the preceding section, the 2-terminal weighted throughput maximization problem is completely solved analytically, including the verification of the second-order optimality conditions. This special case illustrate the general solution procedure, and provides insights useful for the general analysis. When N terminals are present, the analysis is more complicated. In particular, verifying the second-order conditions symbolically does not appear practical. However, identifying the set of points that satisfy the first-order optimality conditions (FONOC) is quite useful, because these points are relatively few. Thus, for a given physical layer and system parameters, the optimizer can be found by directly verifying which of these points yields the highest weighted throughput.

The present section focuses on a specific N-terminal scenario. The scenario studied is one in which a few equally "important" terminals share a cell with many "ordinary" terminals. It is presumed that the system can accommodate all the important terminals at the highest available data rate. But it is not clear how many, if any, of the ordinary terminals should be set to operate at this high rate, in order to maximize the cell's weighted throughput. A general solution procedure for this scenario is given. The cell-throughput maximizing data rates (through the corresponding spreading gains) and the transmission power levels (through the corresponding carrier-to-interference ratios) for all terminals are specified.

8.4.1 Augmented objective function

The pertinent augmented objective function is $\phi(G_1,\ldots,G_N,\alpha_1,\ldots,\alpha_N) =$

$$\sum_{i=1}^{N}\beta_i T_i(G_i,\alpha_i)+\lambda\left(\sum_{i=1}^{N}\frac{\alpha_i}{1+\alpha_i}-1\right)+\sum_{i=1}^{N}\mu_i(G_0-G_i) \tag{8.38}$$

8.4.2 General First-Order Necessary Optimizing Conditions (FONOC)

The general FONOC can be expressed in vector form, with $\gamma_i = G_i\alpha_i$, as:

$$\begin{bmatrix} \beta_1\partial T_1(G_1,\alpha_1)/\partial G_1-\mu_1 \\ \vdots \\ \beta_N\partial T_N(G_N,\alpha_N)/\partial G_N-\mu_N \\ \beta_1 f'(\gamma_1)+\lambda(1+\alpha_1)^{-2} \\ \vdots \\ \beta_N f'(\gamma_N)+\lambda(1+\alpha_N)^{-2} \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \tag{8.39}$$

$$\text{with} \begin{cases} \sum_{i=1}^{N}(1+\alpha_i)^{-1} = N-1 \\ \mu_1(G_0 - G_1) = 0 \\ \vdots \\ \mu_N(G_0 - G_N) = 0 \end{cases} \tag{8.40}$$

Notice that

$$\frac{\partial T_i(G_i,\alpha_i)}{\partial G_i} = \frac{\gamma_i f'(\gamma_i) - f(\gamma_i)}{G_i^2} \tag{8.41}$$

In order to verify the second-order sufficient conditions one needs the Hessian matrix of second partial derivatives of our augmented objective function, denoted as ϕ_{xx}. For $N = 3$, this matrix can be written as:

$$\phi 3_{xx} = \begin{bmatrix} \psi_1 & 0 & 0 & \omega_1 & 0 & 0 \\ 0 & \psi_2 & 0 & 0 & \omega_2 & 0 \\ 0 & 0 & \psi_3 & 0 & 0 & \omega_3 \\ \omega_1 & 0 & 0 & \chi_1 & 0 & 0 \\ 0 & \omega_2 & 0 & 0 & \chi_2 & 0 \\ 0 & 0 & \omega_3 & 0 & 0 & \chi_3 \end{bmatrix} \tag{8.42}$$

Where, for notational convenience, :

$$\psi_i := \beta_i \frac{\partial^2 T_i(G_i,\alpha_i)}{\partial G_i^2} = \beta_i \frac{\gamma_i^2 f''(\gamma_i) + 2\left(f(\gamma_i) - \gamma_i f'(\gamma_i)\right)}{G_i^3} \tag{8.43}$$

$$\chi_i := \beta_i G_i f''(\gamma_i) - 2\lambda(1+\alpha_i)^{-3} \tag{8.44}$$

and

$$\omega_i := \alpha_i \beta_i f''(\gamma_i) \tag{8.45}$$

The general ϕ_{xx} can best be expressed as :

$$\phi_{xx} = \begin{bmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{12} & \mathbf{D}_{22} \end{bmatrix} \tag{8.46}$$

In equation (8.46), $\mathbf{D}_{11}$, $\mathbf{D}_{22}$ and $\mathbf{D}_{12}$ are $N \times N$ diagonal sub-matrices, defined as:

$$\mathbf{D}_{11} = \text{diag}\,(\psi_1, \ldots, \psi_N) \tag{8.47}$$

$$\mathbf{D}_{22} = \text{diag}\,(\chi_1, \ldots, \chi_N) \tag{8.48}$$

$$\mathbf{D}_{12} = \text{diag}\,(\omega_1, \ldots, \omega_N) \tag{8.49}$$

with ψ_i, χ_i and ω_i defined by equations (8.43,8.44,8.45) respectively.

8.4.3 Solving FONOC

8.4.3.1 Looking inside the feasible region

It is natural to start looking for a solution to FONOC that lies in the interior of the feasible region. That is, $\mu_i = 0$ is set, which allows G_i to be greater than G_0 for each i (see equation (8.40)).

8.4.3.1.1 Identifying an interior solution to FONOC From the top half of the vector equation (8.39), $\gamma_i f'(\gamma_i) = f(\gamma_i)$ is obtained. Therefore, from the discussion following equation (8.14), it follows that

$$G_i^* \alpha_i^* = \gamma_0 \tag{8.50}$$

From the bottom half of the vector equation (8.39), it is established that:

$$-\lambda = \beta_i f'(G_i^* \alpha_i^*)(1+\alpha_i^*)^2 = \beta_j f'(G_j^* \alpha_j^*)\left(1+\alpha_j^*\right)^2 \tag{8.51}$$

Now, replacing equation (8.50) into equation (8.51) yields :

$$\frac{1}{1+\alpha_j^*} = \frac{\sqrt{\beta_j/\beta_i}}{1+\alpha_i^*} \tag{8.52}$$

α_j^* ($j > 1$) can be expressed in terms of α_1 through equation (8.52). This way, the constraint relation (8.7) can be turned into an equation which can solved for α_1^*:

$$\frac{\sum_{j=1}^N \sqrt{\beta_j/\beta_1}}{1+\alpha_1^*} = N-1 \Rightarrow \alpha_1^* = \frac{B}{(N-1)} - 1 \tag{8.53}$$

with $\beta_1 = 1$, and

$$B := \sum_{j=1}^N \sqrt{\beta_j} \tag{8.54}$$

Once the value of α_1^* is known, equation (8.52) gives the value of each α_j^*. And once each $\alpha_i^*$is known, equation (8.50) yields immediately the corresponding G_i^* as γ_0/α_i^*. Therefore, a complete "interior" solution to FONOC has been found in closed-form solution :

$$\alpha_i^* + 1 = \frac{B}{(N-1)\sqrt{\beta_i}} \tag{8.55}$$

$$G_i^* = \gamma_0/\alpha_i^* \tag{8.56}$$

Notice that, in order for these values to be feasible, $G_i^* = \gamma_0/\alpha_i^* \geq G_0$ or $\alpha_i^* \leq \gamma_0/G_0$. Under the construction $1 = \beta_1 \leq \cdots \leq \beta_N$, the largest α_i^* is actually α_1^* (see equation (8.55)). Thus, this condition requires that $B = \sum_{j=1}^N \sqrt{\beta_j} \leq (N-1)\gamma_0/G_0$.

It is stressed that this is a closed form solution. γ_0 can be easily obtained from the graph of function f (see fig. (8.1)), or equation (8.14) can be solved numerically.

It is noteworthy that, if $\beta_i = 1$ for all i (terminals are equally "important"), $B = N$ and equation (8.55) reduces to $\alpha_i^* = 1/(N-1)$. Thus, all terminals enjoy the same throughput.

8.4.3.1.2 Is the interior solution to FONOC **a maximizer?** A procedure similar to that applied in section 8.3.4.1.2 can show that the previously found allocation (equations (8.55 and8.56)) is neither a maximizer nor a minimizer, but a saddle point. But it is straightforward to argue that this allocation is *not* the global maximizer.

The non-optimality of the interior solution Each terminal, regardless of its weight, operates with an SIR of γ_0. The spreading gain G_i is obtained as γ_0/α_i with α_i given by equation (8.55). This equation yields an α_i that is inversely proportional to $\sqrt{\beta_i}$. Thus, the largest α_i is that of terminal 1, which has the lowest weight, and the smallest α_i is that of terminal N. Thus, the terminal with the least weight operates with the smaller spreading gain G_i, which means the highest data rate of those assigned, whereas the terminal with the most weight operates with the lowest data rate, of those assigned!

This allocation does not maximize the weighted throughput. Simply re-assigning the α_i in such a way that α_1^* is assigned to terminal N , and α_N^* is assigned to terminal 1 produces a still feasible allocation that yields a higher weighted throughput. Specifically,

$$\sum_{i=1}^{N} \beta_i \frac{f(G_i\alpha_i)}{G_i} = \sum_{i=1}^{N} \beta_i \frac{f(\gamma_0)}{\gamma_0/\alpha_i^*} \propto \sum_{i=1}^{N} \beta_i\alpha_i^* = \alpha_1^* + \sum_{i=2}^{N-1} \beta_i\alpha_i^* + \beta_N\alpha_N^* \tag{8.57}$$

By assigning α_1^* to terminal N and α_N^* to terminal 1, the preceding sum is replaced by

$$\alpha_N^* + \sum_{i=2}^{N-1} \beta_i\alpha_i^* + \beta_N\alpha_1^* \tag{8.58}$$

Subtracting (8.57) from (8.58) yields

$$\beta_N(\alpha_1^* - \alpha_N^*) - (\alpha_1^* - \alpha_N^*) = (\alpha_1^* - \alpha_N^*)(\beta_N - 1) > 0$$

Thus, (8.58) is an improvement over (8.57). The interior solution to FONOC is a non-maximizer.

Second-order sufficient conditions The optimality of this stationary point depends upon the matrix of second partial derivatives (Hessian matrix) of ϕ, our augmented objective function, denoted as ϕ_{xx}. Essentially, if at a point satisfying the FONOC, i.e., a stationary point, for any vector $\vec{h}$ along a feasible direction, the triple product $\vec{h}^{\mathrm{T}}[\phi_{xx}]\vec{h}$ is positive, then the stationary point corresponds to a local minimum, and if this product is negative then the stationary point corresponds to a local maximum. If neither of these conditions hold, then the point is a "saddle point".

A feasible direction is one that is tangent to the curve representing the constraint relationship. Hence, if we denote our constraint curve as $b(G_1, G_2, \alpha_1, \alpha_2) = 0$, (i.e., $b(G_1, \ldots, G_N, \alpha_1, \ldots, \alpha_N) \doteq N - 1 - \sum_{i=1}^{N}(1+\alpha_i)^{-1}$), we only need to consider vectors $\vec{h}$ satisfying $\nabla\mathbf{b} \bullet \vec{h} = 0$, that is, vectors

normal to the gradient of the constraint curve. For us,

$$\nabla \mathbf{b} = \begin{bmatrix} \partial b/\partial G_1 \\ \vdots \\ \partial b/\partial G_N \\ \partial b/\partial \alpha_1 \\ \vdots \\ \partial b/\partial \alpha_N \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ (1+\alpha_1^*)^{-2} \\ \vdots \\ (1+\alpha_N^*)^{-2} \end{bmatrix} \tag{8.59}$$

but at our interior stationary point, α_i^* is given by equation (8.55). Therefore, $\nabla \mathbf{b}$ becomes

$$\nabla \mathbf{b} = \frac{(N-1)^2}{B^2} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \beta_1 \\ \vdots \\ \beta_N \end{bmatrix} \propto \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \beta_1 \\ \vdots \\ \beta_N \end{bmatrix} \tag{8.60}$$

Then, it is easily verified that any vector $\vec{h}$ of the form $\begin{bmatrix} a_1 & \cdots & a_N & b_1 & \cdots & b_N \end{bmatrix}^{\mathrm{T}}$, where the a_i's and b_i's are real numbers, with $\sum_{i=1}^{N} \beta_i b_i = 0$, satisfies $\nabla \mathbf{b} \bullet \vec{h} = 0$. That is, the vector $\vec{b} := \begin{bmatrix} b_1 & \cdots & b_N \end{bmatrix}^{\mathrm{T}}$ must be orthogonal to the vector $\vec{\beta} := \begin{bmatrix} \beta_1 & \cdots & \beta_N \end{bmatrix}^{\mathrm{T}}$.This will happen, for instance, if $b_N = -\left(\sum_{i=1}^{N-1} \beta_i b_i\right)/\beta_N$.

It will prove convenient to express such vector as the product of a "projection" matrix, M, by an arbitrary vector $\vec{a} := \begin{bmatrix} a_1 & \cdots & a_N & b_1 & \cdots & b_{N-1} \end{bmatrix}^{\mathrm{T}}$ of length $2N-1$. Let us describe this process for the special case in which $N = 3$. Subsequently, we will generalize it.

With N=3, the matrix M takes the form

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & -\frac{\beta_1}{\beta_3} & -\frac{\beta_2}{\beta_3} \end{bmatrix} \tag{8.61}$$

so that with $\vec{a} := \begin{bmatrix} a_1 & a_2 & a_3 & b_1 & b_2 \end{bmatrix}^{\mathrm{T}}$ an arbitrary 5-dimensional vector,

$$\mathbf{M} \times \vec{a} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ b_1 \\ b_2 \\ (-\beta_1 b_1 - \beta_2 b_2)/\beta_3 \end{bmatrix} \tag{8.62}$$

It is easily verified that the scalar product of the vector $\mathbf{M} \times \vec{a}$ by $\begin{bmatrix} 0 & 0 & 0 & \beta_1 & \beta_2 & \beta_3 \end{bmatrix}^{\mathrm{T}}$ $\propto \nabla \mathbf{b}$ always equals zero.

For a general N, the desired matrix has the form:

$$\mathbf{M} = \begin{bmatrix} \mathbf{I}_N & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{N-1} \\ 0_{1\times N} & \mathbf{r} \end{bmatrix} \tag{8.63}$$

(compare equations (8.63) and (8.61)), where $\mathbf{I}_N$ denotes the identity matrix of size N, $\mathbf{0}$ denotes the zero matrix of appropriate dimension, $0_{1\times N}$ denotes an all-zero row of length N, and r is a row vector of length $N-1$ of the form:

$$\mathbf{r} = -\frac{1}{\beta_N} \begin{bmatrix} \beta_1 & \cdots & \beta_{N-1} \end{bmatrix}$$

The second-order conditions for the stationary point under consideration can be expressed in terms of the matrix $\mathbf{M}^{\mathrm{T}} \times \phi_{xx} \times \mathbf{M}$. If this matrix is positive definite, then the stationary point corresponds to a local minimum, and if it is negative definite then the stationary point corresponds to a local maximum. If this matrix is indefinite, then this point is a "saddle point". A square matrix is positive definite if all its principal minor determinants are positive.

When $N = 3$, $\phi 3_{xx}$ is given by equation (8.42), in terms of ψ_i , χ_i , and ω_i. After some algebra, we obtain, $\mathbf{M}^{\mathrm{T}} \times \phi 3_{xx} =$

$$\begin{bmatrix} \psi_1 & 0 & 0 & \omega_1 & 0 & 0 \\ 0 & \psi_2 & 0 & 0 & \omega_2 & 0 \\ 0 & 0 & \psi_3 & 0 & 0 & \omega_3 \\ \omega_1 & 0 & -\frac{\beta_1}{\beta_3}\omega_3 & \chi_1 & 0 & -\frac{\beta_1}{\beta_3}\chi_3 \\ 0 & \omega_2 & -\frac{\beta_2}{\beta_3}\omega_3 & 0 & \chi_2 & -\frac{\beta_2}{\beta_3}\chi_3 \end{bmatrix}$$

And some more algebra yields $\mathbf{M}^{\mathrm{T}} \times \phi 3_{xx} \times \mathbf{M} =$

$$\begin{bmatrix} \psi_1 & 0 & 0 & \omega_1 & 0 \\ 0 & \psi_2 & 0 & 0 & \omega_2 \\ 0 & 0 & \psi_3 & -\frac{\beta_1}{\beta_3}\omega_3 & -\frac{\beta_2}{\beta_3}\omega_3 \\ \omega_1 & 0 & -\frac{\beta_1}{\beta_3}\omega_3 & \chi_1 + \left(\frac{\beta_1}{\beta_3}\right)^2 \chi_3 & \frac{\beta_1\beta_2\chi_3}{\beta_3^2} \\ 0 & \omega_2 & -\frac{\beta_2}{\beta_3}\omega_3 & \frac{\beta_1\beta_2}{\beta_3^2}\chi_3 & \chi_2 + \left(\frac{\beta_2}{\beta_3}\right)^2 \chi_3 \end{bmatrix} \tag{8.64}$$

At this interior stationary point, (see equations (8.55,8.56)), the elements of this matrix can be written as (recall that $G_i^* \alpha_i^* = \gamma_0$):

$$\frac{\psi_i^*}{f''(\gamma_0)} = \beta_i \frac{\gamma_0^2}{(G_i^*)^3} \equiv \beta_i \frac{(\alpha_i^*)^3}{\gamma_0}$$

$$\frac{\omega_i^*}{f''(\gamma_0)} = \beta_i \alpha_i^*$$

and, since

$$\begin{aligned} \chi_i^* &= \\ \beta_i G_i^* f''(\gamma_i^*) - 2\lambda^* (1+\alpha_i^*)^{-3} &= \\ \beta_i \frac{\gamma_0}{\alpha_i^*} f''(\gamma_0) + 2\left[\beta_i f'(\gamma_0)(1+\alpha_i^*)^2\right](1+\alpha_i^*)^{-3} &= \\ \beta_i f''(\gamma_0)\left(\frac{\gamma_0}{\alpha_i^*} + 2\frac{f'(\gamma_0)}{f''(\gamma_0)}(1+\alpha_i^*)^{-1}\right) \end{aligned}$$

it is convenient to set

$$\frac{\chi_i^*}{f''(\gamma_0)} = \beta_i \hat{\chi}_i$$

where, for notational convenience,

$$\hat{\chi}_i := \left(\frac{\gamma_0}{\alpha_i^*} + \frac{2\rho_0}{1+\alpha_i^*}\right) \tag{8.65}$$

is defined; and

$$\rho_0 := \frac{f'(\gamma_0)}{f''(\gamma_0)} \tag{8.66}$$

Hence, at the point being tested, equation(8.64) leads to $\mathbf{M}^{\mathrm{T}} \times \phi 3_{xx} \times \mathbf{M} / f''(\gamma_0) =$

$$\begin{bmatrix} \frac{\beta_1\alpha_1^3}{\gamma_0} & 0 & 0 & \beta_1\alpha_1 & 0 \\ 0 & \frac{\beta_2\alpha_2^3}{\gamma_0} & 0 & 0 & \beta_2\alpha_2 \\ 0 & 0 & \frac{\beta_3\alpha_3^3}{\gamma_0} & -\beta_1\alpha_3 & \beta_2\alpha_3 \\ \beta_1\alpha_1 & 0 & -\beta_1\alpha_3 & \beta_1\hat{\chi}_1 + \frac{\beta_1^2\hat{\chi}_3}{\beta_3} & \frac{\beta_1\beta_2\hat{\chi}_3}{\beta_3} \\ 0 & \beta_2\alpha_2 & -\beta_2\alpha_3 & \frac{\beta_1\beta_2\hat{\chi}_3}{\beta_3} & \beta_2\hat{\chi}_2 + \frac{\beta_2^2\hat{\chi}_3}{\beta_3} \end{bmatrix} \tag{8.67}$$

The objective is to test whether the matrix $\mathbf{M}^{\mathrm{T}} \times \phi_{xx} \times \mathbf{M}$ is negative definite, which would confirm the interior stationary point as a maximizer. It has been shown in chapter 2 that for the class of functions being considered, $f''(\gamma_0)$ is always negative. Therefore, an equivalent test is whether the matrix $\mathbf{M}^{\mathrm{T}} \times \phi_{xx} \times \mathbf{M} / f''(\gamma_0)$ (equation (8.67)) is positive definite.

A matrix is positive definite if and only if each one of its principal minor determinants are positive. It is immediate that the determinants of the first three principal minors of the matrix given in (8.67) ,

$$\left[\frac{\beta_1\alpha_1^3}{\gamma_0}\right], \begin{bmatrix} \frac{\beta_1\alpha_1^3}{\gamma_0} & 0 \\ 0 & \frac{\beta_2\alpha_2^3}{\gamma_0} \end{bmatrix} \text{ and } \begin{bmatrix} \frac{\beta_1\alpha_1^3}{\gamma_0} & 0 & 0 \\ 0 & \frac{\beta_2\alpha_2^3}{\gamma_0} & 0 \\ 0 & 0 & \frac{\beta_3\alpha_3^3}{\gamma_0} \end{bmatrix}$$

are all positive.

The fourth principal minor is

$$\begin{bmatrix} \frac{\beta_1\alpha_1^3}{\gamma_0} & 0 & 0 & \beta_1\alpha_1 \\ 0 & \frac{\beta_2\alpha_2^3}{\gamma_0} & 0 & 0 \\ 0 & 0 & \frac{\beta_3\alpha_3^3}{\gamma_0} & -\beta_1\alpha_3 \\ \beta_1\alpha_1 & 0 & -\beta_1\alpha_3 & \beta_1\hat{\chi}_1 + \frac{\beta_1^2\hat{\chi}_3}{\beta_3} \end{bmatrix}$$

whose determinant is

$$\begin{aligned} \frac{\beta_2\alpha_2^3}{\gamma_0} \begin{vmatrix} \frac{\beta_1\alpha_1^3}{\gamma_0} & 0 & \beta_1\alpha_1 \\ 0 & \frac{\beta_3\alpha_3^3}{\gamma_0} & -\beta_1\alpha_3 \\ \beta_1\alpha_1 & -\beta_1\alpha_3 & \beta_1\hat{\chi}_1 + \frac{\beta_1^2\hat{\chi}_3}{\beta_3} \end{vmatrix} &= \\ \frac{\beta_2\alpha_2^3}{\gamma_0} \left[\left(\beta_1\hat{\chi}_1 + \frac{\beta_1^2\hat{\chi}_3}{\beta_3}\right) \frac{\beta_1\beta_3\alpha_1^3\alpha_3^3}{\gamma_0^2} \right. &+ \\ \left. - \frac{\beta_1^2\beta_3\alpha_1^2\alpha_3^3}{\gamma_0} - \frac{\beta_1^3\alpha_3^2\alpha_1^3}{\gamma_0} \right] &= \\ \frac{\beta_1^2\beta_2\alpha_1^3\alpha_2^3\alpha_3^3}{\gamma_0^2} \left[\frac{\beta_3\hat{\chi}_1 + \beta_1\hat{\chi}_3}{\gamma_0} - \frac{\beta_3}{\alpha_1} - \frac{\beta_1}{\alpha_3} \right] & \end{aligned}$$

The question is then whether

$$(\beta_1\hat{\chi}_3 + \beta_3\hat{\chi}_1) \overset{<}{>} \gamma_0 \left(\frac{\beta_3}{\alpha_1} + \frac{\beta_1}{\alpha_3}\right)$$

i.e, (replacing $\hat{\chi}_i$ with its defining expression (8.65)), whether

$$\begin{aligned} \frac{\beta_1\gamma_0}{\alpha_3}+\frac{2\beta_1\rho_0}{1+\alpha_3}+\frac{\beta_3\gamma_0}{\alpha_1}+\frac{2\beta_3\rho_0}{1+\alpha_1} &\lessgtr \\ \frac{\beta_3\gamma_0}{\alpha_1}+\frac{\beta_1\gamma_0}{\alpha_3} & \\ \text{i.e., whether} & \\ \left(\frac{2\beta_1}{1+\alpha_3}+\frac{2\beta_3}{1+\alpha_1}\right)\rho_0 &\lessgtr 0 \end{aligned} \tag{8.68}$$

Observe that for the class of functions being considered, $\rho_0 = f'(\gamma_0)/f''(\gamma_0)$ is always negative, because $f'(x)$ is positive everywhere, and it has been shown in chapter 2 that $f''(\gamma_0)$ is negative. Therefore, the left-hand side of inequality (8.68) is less than zero, which proves that the determinant is in fact *negative*!

In summary, the first three principal minor determinants are *positive*, while the fourth such determinant is *negative*. This implies that the concerned matrix is indefinite. Therefore, the interior stationary point is neither a local minimizer nor a local maximizer. It is a saddle point.

8.4.3.2 Single-Favorite Boundary Solution (SFBS)

An allocation satisfying FONOC, where every terminal's data rate is less than the highest available value (equations (8.56,8.55)) was found. Unfortunately, this allocation is not the desired maximizer. This indicates that the true maximizer is a non-interior solution to FONOC; i.e., a solution in which one or more terminals operate at the lowest available spreading gain (highest available data rate). In principle, the number of possible non-interior solutions could be very large, of the order of 2^N. A basic rationale is needed to systematically search for these solutions.

A reasonable starting point is to to seek an allocation satisfying FONOC in which only the spreading gain of the "most important" terminal is set at the lowest available value, G_0 (i.e. this terminal operates at the highest available data rate), with other terminals' spreading gains to be determined by the analysis. This is done below by setting $G_N = G_0$, and $\mu_i = 0$ for $1 \le i < N$.

8.4.3.2.1 General form of SFBS The first $N-1$ rows of the vector equation (8.39), and the fact that $\mu_i = 0$ for $1 \le i < N$ has been set, yield

$$G_i\alpha_i = \gamma_0 \text{ for } 1 \le i < N \tag{8.69}$$

with γ_0 as defined by equation (8.14), and shown in figure (8.1).

The bottom half of the vector equation (8.39) leads to:

$$-\lambda = \beta_i f'(G_i^*\alpha_i^*)\,(1+\alpha_i^*)^2 \text{ for } 1 \le i < N \tag{8.70}$$

and

$$-\lambda = \frac{\beta_N}{G_0^2} f'(x)\,(G_0+x)^2 \tag{8.71}$$

with $x := G_0\alpha_N^*$.

Combining equations (8.69 and 8.70) yields

$$\frac{1}{1+\alpha_j^*} = \frac{\sqrt{\beta_j/\beta_i}}{1+\alpha_i^*} \text{ for } 1 \le i\,,\ j < N \tag{8.72}$$

Through equation (8.72), α_i^* $(1 < i < N)$ can be expressed in terms of α_1. This way, the constraint relation (8.7) becomes an equation with only two unknowns, α_1 and α_N. With

$$B_{N-1} := \sum_{j=1}^{N-1} \sqrt{\beta_j} \tag{8.73}$$

substituting equation (8.72) into (8.7) ($\sum_i (1+\alpha_i)^{-1} = N-1$) yields

$$\frac{B_{N-1}}{1+\alpha_1^*} + \frac{G_0}{G_0+x} = N-1 \quad \rightarrow \tag{8.74}$$

$$\frac{G_0+x}{1+\alpha_1^*} = \frac{N-1}{B_{N-1}}x + \frac{N-2}{B_{N-1}}G_0 \quad \rightarrow \tag{8.75}$$

$$\alpha_1^* + 1 = \frac{B_{N-1}}{N-1-(1+\alpha_N^*)^{-1}} \tag{8.76}$$

Equations (8.70, and 8.71) can be combined as :

$$\beta_N f'(x)\,(G_0+x)^2 = G_0^2 f'(\gamma_0)\,(1+\alpha_1^*)^2 \tag{8.77}$$

which can be put (using equation (8.75)) as

$$\left(C_1\frac{x}{G_0} + D_1\right)^2 \frac{f'(x)}{f'(\gamma_0)} = \frac{1}{\beta_N} \tag{8.78}$$

with

$$C_1 = \frac{N-1}{B_{N-1}}\ ;\ D_1 = \frac{N-2}{B_{N-1}} \tag{8.79}$$

Assuming that a meaningful solution to equation (8.78) can be found, denote such solution as δ_0. In terms of δ_0, a complete allocation satisfying FONOC can be identified. By definition, $\delta_0 = G_0\alpha_N^*$ which implies that $\alpha_N^* = \delta_0/G_0$ satisfies FONOC. From α_N^*, equation (8.76) gives immediately α_1^*, and from α_1^* and equation (8.72), each α_i^* $(1 < i < N)$ is obtained. And since each G_i^* $(1 \le i < N)$ must satisfy $G_i^*\alpha_i^* = \gamma_0$ (equation (8.69)), once each α_i^* $(1 < i < N)$ is known, so is the corresponding

G_i^*. The complete allocation is given by:

$$\begin{aligned} G_N^* &= G_0 && (8.80)\\ G_0\alpha_N^* = \gamma_N^* &= \delta_0 && (8.81)\\ \text{for } 1 \le i < N & \\ \alpha_i^* &= \frac{1}{\sqrt{\beta_i}} \frac{B_{N-1}}{N-1-(1+\delta_0/G_0)^{-1}} - 1 && (8.82)\\ G_i^*\alpha_i^* = \gamma_i^* &= \gamma_0 && (8.83) \end{aligned}$$

However, each G_i^* must satisfy $G_i^* \ge G_0$ or $\alpha_i^* \le \gamma_0/G_0$. Equation (8.82) indicates that $\alpha_1 \ge \alpha_i$ for all i. Thus, it suffices that

$$\alpha_1^* = \frac{B_{N-1}}{N-1-(1+\delta_0/G_0)^{-1}} - 1 \le \gamma_0/G_0 \tag{8.84}$$

8.4.3.2.2 Existence of this solution The preceding allocation depends on a solution to the single-variable algebraic equation (8.78). Below, the conditions under which this algebraic equation has solution(s) are examined.

Observe, first, that $C_1x/G_0 + D_1 \le x+1$. This is so, because the left-hand side of this inequality is largest when G_0 and B_{N-1} are smallest (see equations (8.79)). Because of technological limitations, $G_0 \ge 1$ (the highest available data rate cannot exceed the channel's "chip rate"). And, $B_{N-1} = \sum_{j=1}^{N-1}\sqrt{\beta_j} \ge N-1$, since, by construction, $1 = \beta_1 \le \beta_i$ for $\forall i$. Hence, $C_1 \le 1$ and $D_1 \le (N-2)/(N-1) \le 1$. All this implies that $C_1x/G_0 + D_1 \le x+1$.

For the class of functions being considered, the graph of the function $x^2 f'(x)$ is observed to be "bell-shaped", as displayed by figure (8.1), and so is the graph of $(x+1)^2 f'(x)/f'(\gamma_0)$. On the basis of the preceding paragraph, it can be further argued that the function $(C_1x/G_0+D_1)^2 f'(x)/f'(\gamma_0)$ is also bell-shaped. This implies that, if G_0 is "too large", the "peak" of this function may fall below $1/\beta_N$, unless β_N is also "very large". Thus, equation (8.78) may have no solution. On the other hand, when G_0 is sufficiently small and/or β_N is sufficiently large, two values of x, on either side of the peak of the concerned function, say $x_1^* \le x_2^*$, will satisfy equation (8.78). Intuitively, one would expect that the larger of these two values be the best candidate for a maximizer. However, a larger SIR for the favorite terminal leads to a smaller throughput for the non-favored terminals. Thus, with many non-favored terminals and just one favorite, it is possible that the network weighted throughput be higher when the SIR of the favorite terminal is the lower of the two values satisfying eq. (8.78). But this may not be yield a global maximum.

On the other hand, the nth row of equation (8.39) yields the multiplier associated with the constraint $G_0 - G_N \le 0$ as

$$\mu_N = \frac{\delta_0 f'(\delta_0) - f(\delta_0)}{G_0^2/\beta_N} \tag{8.85}$$

It is necessary for a maximizer that $\mu_N \leq 0$. This condition is best interpreted by writing it as

$$\frac{G_0^2}{\beta_N}\mu_N = t^2 \frac{d}{dt}\left(f(t)/t\right)\Big|_{t=\delta_0} \leq 0$$

The development in chapter 2 shows that, for the class of functions f being considered, the derivative of $f(t)/t$ is positive for $t < \gamma_0$, is zero at γ_0, and is negative for $t > \gamma_0$(with γ_0 defined by equation (8.14)). Thus, in order for δ_0 to lead to a maximizer, it is necessary that

$$\delta_0 \geq \gamma_0 \tag{8.86}$$

Thus, even if $1/\beta_N$ falls below the maximal value of the function $(C_1 x/G_0 + D_1)^2 f'(x)/f'(\gamma_0)$, the resulting intersection points may both be less than γ_0, which would violate a necessary condition for a maximizer.

8.4.3.3 A Multi-Favorite Boundary Solution (MFBS)

The remainder of this investigation focuses on the special case in which $\beta_i = 1$ for $i = 1 \ldots N_1$, and $\beta_i = \beta > 1$ otherwise. That is, there are only two possible weights.

The single favorite boundary solution to FONOC discussed in the preceding section may *not* exist, and even if it does exist, it may *not* lead to a global maximizer. This section investigates a more general solution to FONOC in which all the important terminals, N_2, and several ordinary terminals, say $n_1 \leq N_1$, operate at maximal data rate.

8.4.3.3.1 General structure of the solution There are $N_1 - n_1$ non-favored terminals. Thus, $\mu_i = 0$ for $1 \leq i \leq N_1 - n_1$ (see equations (8.40)). Working with the first $N_1 - n_1$rows of the vector equation (8.39), we obtain, for $1 \leq i\ ,\ j \leq N_1 - n_1$,

$$\gamma_i f'(\gamma_i) - f(\gamma_i) = 0 \rightarrow \gamma_i^* \equiv G_i^* \alpha_i^* = \gamma_0 \tag{8.87}$$

with γ_0 defined as the unique positive solution to eq. (8.14).

We also establish by working with the bottom half of the vector equation (8.39) that:

$$-\lambda = f'(G_i^* \alpha_i^*)\left(1 + \alpha_i^*\right)^2 \text{ for } 1 \leq i \leq N_1 - n_1 \tag{8.88}$$

and

$$-\lambda = f'(G_0 \alpha_i^*)\left(1 + \alpha_i^*\right)^2 \text{for } N_1 - n_1 < i \leq N_1 \tag{8.89}$$

and

$$-\lambda = \beta f'(G_0 \alpha_i^*)\left(1 + \alpha_i^*\right)^2 \text{for } N - N_2 \leq i \leq N \tag{8.90}$$

Combining equations (8.87) and (8.88), we obtain

$$\alpha_i^* = \alpha_1^* \text{ for } 1 \le i \le N_1 - n_1 \tag{8.91}$$

For $N_1 - n_1 < i \le N_1$, eq. (8.89) leads to n_1 equations of the form

$$f'(G_0\alpha_i^*)\,(1+\alpha_i^*)^2 = f'(G_0\alpha_j^*)\,\left(1+\alpha_j^*\right)^2$$

Evidently, this equation is satisfied with

$$\alpha_i^* = \alpha_j^* = y/G_0 \text{ for } N_1 - n_1 < i\,,\ j \le N_1 \tag{8.92}$$

A similar analysis of eq. (8.90) leads to

$$\alpha_i^* = \alpha_j^* = x/G_0 \text{ for } N - n_2 < i\,,\ j \le N \tag{8.93}$$

Now, the constraint relation (8.7) ($\sum_i (1+\alpha_i)^{-1} = N-1$) becomes an equation with only three unknowns, α_1, x and y. Substituting eqs. (8.91, 8.92 and 8.93) into (8.7) yields

$$\frac{N_1 - n_1}{1+\alpha_1} + \frac{n_1}{1+\frac{y}{G_0}} + \frac{N_2}{1+\frac{x}{G_0}} = N-1 \tag{8.94}$$

Equations (8.88, 8.89, and 8.90) imply that

$$f'(y)\left(1+\frac{y}{G_0}\right)^2 = \beta f'(x)\left(1+\frac{x}{G_0}\right)^2 \tag{8.95}$$

$$\beta f'(x)\left(1+\frac{x}{G_0}\right)^2 = f'(\gamma_0)\,(1+\alpha_1^*)^2 \tag{8.96}$$

Equation (8.96) provides a closed-form expression for α_1^* in terms of x:

$$\alpha_1^* = \left(1+\frac{x}{G_0}\right)\sqrt{\frac{\beta f'(x)}{f'(\gamma_0)}} - 1 \tag{8.97}$$

The function on the right-hand side of eq. (8.97) takes on values as low as -1, and yields a bell-shaped graph (such as that shown at the top of fig. 8.3). But, physically, α_1 cannot be negative. Thus, the existence of a MFBS in which all the ordinary terminals are active necessitates that the SIR of the important terminals be held within certain interval. This range expands as β grows, but shrinks as G_0 increases. Furthermore, α_1 cannot be too large, either. This is so because, in order to satisfy FONOC, the non-favored terminals must operate with SIR equal to γ_0. Thus the spreading gain for these terminals must equal γ_0/α_1. But if α_1 is large, this ratio may be smaller than G_0, which is the smallest allowable spreading gain. That is, it is necessary that $0 < \alpha_1 \le \gamma_0/G_0$. This further constrains the values of x that can be chosen.

Within the appropriate range, eq. (8.97) allows us to write eq. (8.94) as :

$$\frac{n_1}{1+\frac{y}{G_0}}+\frac{N_1-n_1}{\left(1+\frac{x}{G_0}\right)\sqrt{\frac{\beta f'(x)}{f'(\gamma_0)}}}+\frac{N_2}{1+\frac{x}{G_0}}=N-1 \tag{8.98}$$

Equations (8.95) and (8.98) form a system of two non-linear equations in two unknowns which is, in principle, solvable. Once the appropriate values of x^* and $y^*$are known, α_1^*, the optimal CIR for terminals $1\ldots N-n_1$, can be obtained from eq. (8.97), and the matching spreading gain, from eq. (8.87), as γ_0/α_1^*. Thus, from x^* and y^*, a complete multi-favorite solution to FONOC can be obtained. This solution is discussed below, its possible optimality is addressed.

8.4.3.3.2 Discussion of the MFBS The general structure of this solution to FONOC is very similar to that of the dual-favorite solution (one important and one ordinary terminal operate at highest data rate, and the rest operate with SIR γ_0). The caption of figure 8.3 summarizes much of what can be said about the MFBS. Further insights are given in section 8.4.4.2 through numerical examples.

Generally, there are four intersection points, one in each of the branches of the concerned X-curve. It is clear that, among the four intersection points, the NE one yields the largest throughput for the favorite terminals, since both x (the SIR of the important terminals) and y (the SIR of the favored non-important terminals) are as high as possible. But the non-favored terminals, whose SIR is γ_0, by eq. (8.87), must also be considered. The throughput of each non-favored terminal is obtained as $f(\gamma_0)/G_1 = f(\gamma_0)/(\gamma_0/\alpha_1^*) \propto \alpha_1^*$. α_1^* is obtained from x^* through eq. (8.97), which gives rise to a "bell shaped" graph (see comments immediately following eq. (8.97)). Thus, α_1^* (and hence the throughput of the non-favored terminals) is *decreasing* in x^* beyond a certain value of x^*. Therefore, if the number of non-favored terminals, N_1-n_1, is larger than the number of favorites, N_2+n_1, the NE intersection point may *not* lead to the largest overall weighted throughput.

Moreover, when $n_1=N_1$ so that *all* terminals, whether important or not, operate at maximal data rate, then the U-curve is replaced by a hyperbolic "L-curve", as displayed in fig. 8.3. To see this more clearly, observe that when $n_1=N_1$, we can solve eq. (8.98) for y in terms of x, obtaining:

$$\frac{y}{G_0}=\frac{N_1}{N_1+N_2-1-\frac{N_2}{1+x/G_0}}-1 \tag{8.99}$$

For $x=0$, $y=G_0/(N_1-1)$; and as $x\to\infty$, $y\to -G_0(N_2-1)/(N_1+N_2-1)$.Thus, when all terminals operate at maximal data rate, if $N_2>1$ (several "heavy-weight" terminals), there is an SIR value x beyond which y would have to be negative in order to satisfy the constraint on the power ratios, eq. (8.7). That is, the "L-curve" falls below zero for x sufficiently large. Hence, in this case, x cannot exceed $G_0/(N_2-1)$. Furthermore, for low G_0, the maximum value of y, which is $G_0/(N_1-1)$ could be so low, that the L-curve may intersect only the SW leg of the X-curve, in which case, both x and y are "low", and this would lead to a *minimum*, not a maximum. The message, in this case, is that there are too many "favored" terminals (those operating at the highest data rate); some need to be

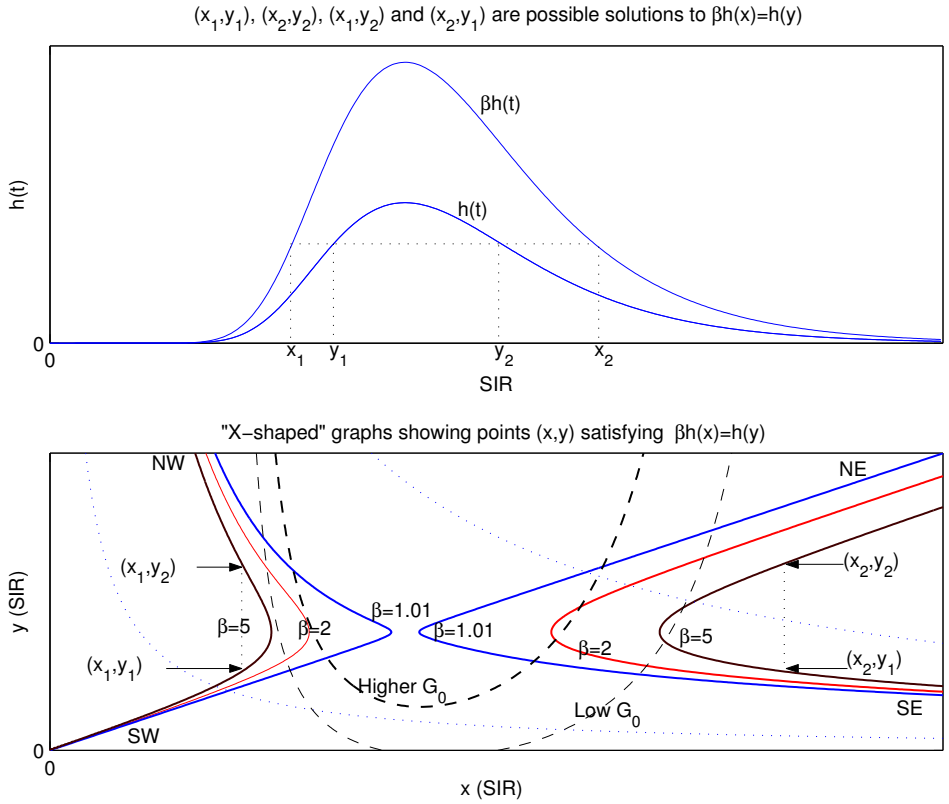

Figure 8.3:
With the SIR of the favored terminals denoted as x (important) and y (ordinary), FONOC requires that $\beta h(x) = h(y)$ (eq. (8.95)). Any of the pairs (x_1,y_1), (x_2,y_2), (x_1,y_2), or (x_2,y_1) (top) satisfies this equation, but may not be feasible. When all such points are plotted, an "X-shaped" graph emerges (NE, NW, SW and SE are directional labels). On the same axes, the U-shaped graph arising from the constraint equation (8.98) is also plotted. The 4 intersection points between the U-shaped and X-shaped graphs for the given (G_0,β) pair lead to feasible solutions to FONOC, provided that the resulting CIR and data rate for the non-favored terminals are also feasible. When G_0 is "large", the "U" lies above the "X" and no intersections exist. In such a case, *all* terminals are set to operate at the highest data rate, and the hyperbolic curves (from eq. (8.99)) replace the U curves. If G_0 is low enough, the hyperbola may *only* intersect the SW leg of the X-curve, which leads to a minimum.

downgraded to “non-favored”. On the other hand, with a large enough G_0, the hyperbola intersects the “Northern legs” of the X. In this case, a maximum results. Thus, when G_0 is large enough, all terminals should operate at the highest available data rate.

8.4.4 Finding the Global Maximizer

8.4.4.1 Solution procedure

In discussing the procedure, for expositional convenience we assume that there is only one important terminal ($N_2 = 1$). The key variable is the number of “favored” terminals (those operating at highest data rate). At least the important terminal must be in this group, with n_1, the number of favored ordinary terminals, possibly being as low as zero, and as high as N_1, the total number of ordinary terminals.

- Set $n_1 = 0$ (Single-favorite). Find, if possible, the 2 positive solutions to eq. (8.78), say x_1^* and x_2^*. Either value can be a FONOC-solving SIR for the favorite terminal, and each leads to a complete allocation. For each of these 2 values, through eq. (8.82) obtain a corresponding α, the FONOC-solving CIR for the ordinary terminals, whose matching spreading gain is γ_0/α. If $\gamma_0/\alpha > G_0$, a complete *feasible* solution to FONOC has been found, and the corresponding weighted throughput can be calculated. Of the 2 solutions to eq. (8.78), the one yielding the highest network weighted throughput should be chosen. It is possible that no single-favorite solution to FONOC exists. In any case, set $n_1 = 1$ and proceed to find a dual-favorite solution .

- For $1 \leq n_1 < N_1$ (multifavorite solution) proceed as follows. Find the solutions (up to four) to the system of equations formed by eq. (8.95) and eq. (8.98). This is the equivalent of finding the four intersections between an X-shaped and a U-shaped graph (fig. 8.3). But not all of these intersections are useful. If the x value is outside certain range, the FONOC-solving CIR of the non-favored terminals, α, may be negative, or its matching spreading gain may be less than G_0. *Each one* of the useful intersections determine a complete solution to FONOC. The SIRs of the favored terminals are x (important) and y (ordinary). The FONOC-solving CIR for the non-favored terminals can be found from eq. (8.97), and the matching spreading gain is γ_0/α. The corresponding weighted network throughput can then be calculated for each feasible solution, and the one leading to the greatest network throughput chosen. If the U curve is “too wide”, meaning that x would make α negative, proceed to the next item, below. Otherwise, increment n_1 and repeat this complete item (draw another U curve for the new n_1), until $n_1 = N_1$.

- For $n_1 = N_1$ (all terminals, important or not, operate at the highest data rate), find the solution to the system of equations formed by eq. (8.95) and eq. (8.99). This is the equivalent of finding the intersections between an X-shaped graph and a hyperbola (fig. 8.3). The SIRs of the important terminal is x and that of the ordinary terminals is y. The matching CIRs are

respectively x/G_0 and y/G_0. Each intersection leads to a feasible solution to FONOC, from which the weighted throughput can be calculated. If the only intersection lies in the SW leg of the X, the all-favored solution is a local minimizer (useless).

- The global maximizer is found among the feasible FONOC-solving allocations already discussed, and is whichever yields the largest weighted throughput.

8.4.4.2 Numerical examples

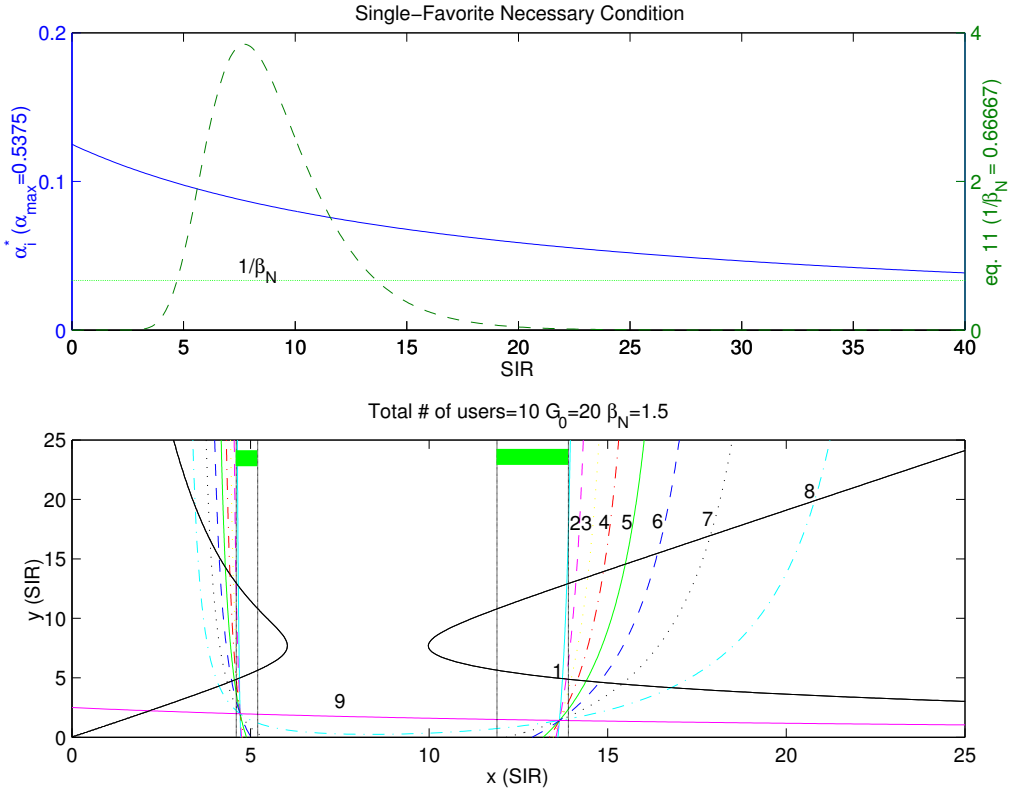

Figure 8.4:
With a moderate $G_0 = 20$ and $\beta = 1.5$, two single-favorite solutions exist, at $x \approx$4.5 and 13 (top). But the best yields only a throughput of 0.12 the chip rate. Fortunately, the bottom subplot shows 4 dual-favorite ($n_1 = 1$) solutions at (13.8, 12.8), (13.7, 4.9), (4.7,12.7) and (4.7,5.0) leading to *weighted* throughput of 0.7, 0.65, 0.05 and 0.015 the chip rate, respectively. The "all favored" solution ($n_1 = 9$) leads to a minimum.

In the examples shown in figures 8.4 , 8.5 and 8.6, the frame-success function is $f(x) = [1 - (1/2)\exp(x/2)]^{80}$, corresponding to non-coherent FSK, no FEC, and packet size of 80 bits. The "preferred" SIR $\gamma_0 = 10.75$ for this FSF. There are 10 terminals, one of which is "important".

The top subplot refers to the "single-favorite" solution (SFBS), with x the SIR of the favorite. The first order optimizing conditions (FONOC) require the SIR of the favorite to be at one of the intersections between the shown bell-shaped curve and the line $1/\beta$ (eq. (8.78)). The hyperbola

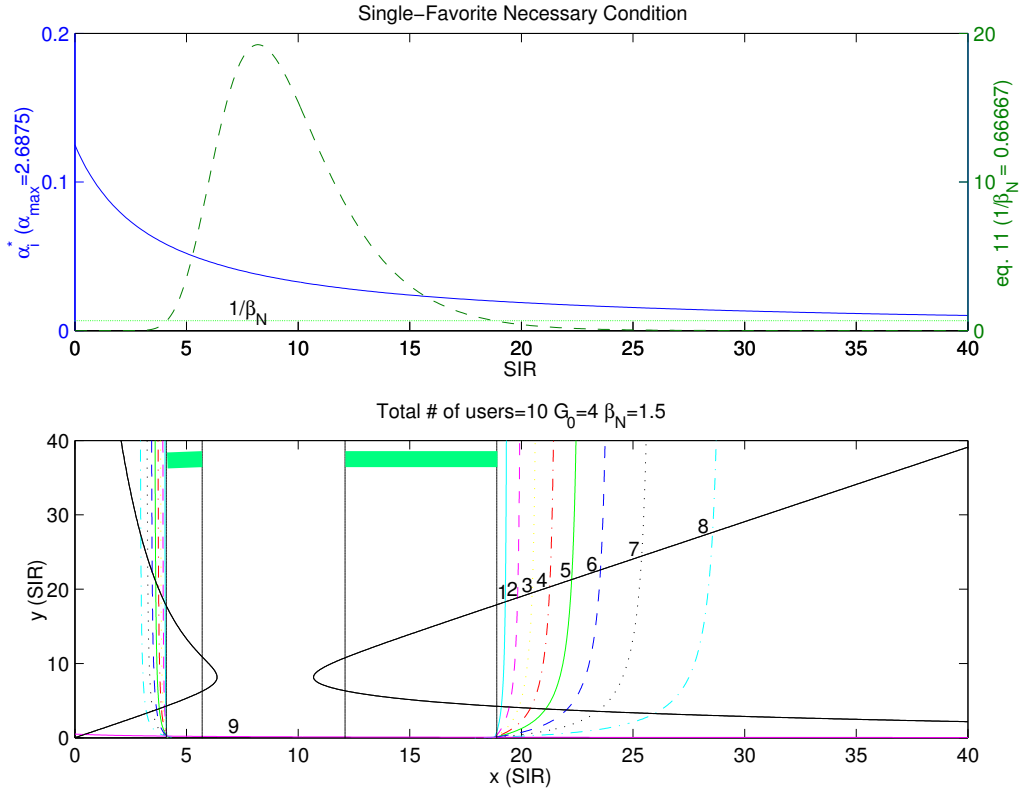

Figure 8.5:
With a small $G_0 = 4$ and a moderate $\beta = 1.5$ the single-favorite solution exists (top), and leads to the maximum. But all the multi-favorite solutions fail (intersections of U and X curves falls outside the acceptable range of x). An "all favored" solution exists (barely visible) but leads to a minimum.

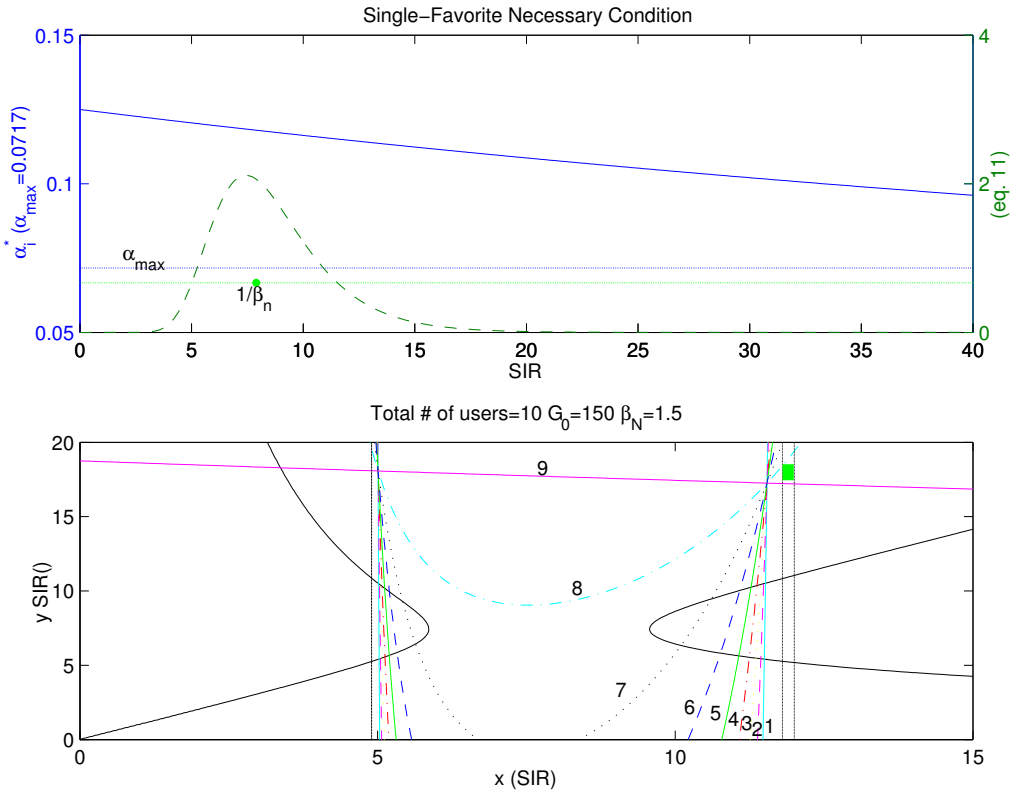

Figure 8.6:
With a high $G_0 = 150$ and $\beta = 1.5$ *no* single-favorite solution is available. Although the bell curve does intercept the line $1/\beta$ (top), for any x, α will be above 0.07, which means its matching spreading gain will fall below G_0. The same problem plagues multi-favorite solutions (U-X intersections) with $1 \leq n_1 \leq 8$, all of which fall outside the acceptable range for x (shown by the thick green lines). However, the "all-favored" solutions ($n_1 = 9$) (intersections of the hyperbola and the X) do exist. The NE intersection leads to the global maximizer.

at the top corresponds to α, the CIR of the non-favorites, as a function of x (eq. (8.82)). If α exceeds $\alpha_{max} = \gamma_0/G_0$, its matching spreading gain γ_0/α is less than G_0, the lowest available. The bottom subplot corresponds to the multi-favorite solutions, in which the important terminal and n_1 ordinary ones operate with the lowest available spreading gain G_0, while the remaining $N_1 - n_1$ ordinary terminals operate with an SIR of γ_0. x and y are respectively the SIR of the important and ordinary terminals operating at the highest data rate ("favored"). The X-graph arises from eq. (8.95), and the U graph from eq. (8.98). U curves are numbered with the chosen n_1. The 4 intersection points between the U and X graphs lead to feasible solutions to FONOC, provided that x lies inside the intervals indicated by the thick green line. Outside these intervals, either the resulting CIR, α, for the non-favored terminals, or its matching data rate is unacceptable.

8.5 Discussion

The optimal allocation of power levels and data rates for terminals transmitting to one base station, in a scenario relevant to 3G CDMA, has been investigated. The objective function is the *weighted* sum of each terminal's throughput. For much of the development, two weights, which admit various interpretations, including levels of importance, "utilities", or monetary prices, are considered (certain results are given for the general case in which there are as many weights as there are terminals) . The properties of the physical layer are embodied in the frame success function (FSF), which gives, in terms of received signal-to-interference ratio (SIR), the probability that a data packet is correctly received. But *no* specific functional form ("equation") is imposed on the FSF. It is assumed that *all that is known* about the FSF is that its graph is "S-shaped", and the analysis follows from properties derived from this shape (a few additional technical assumptions to be discussed below are also made, in order to characterize the solutions to FONOC). Therefore, this analysis applies to many physical layer configurations of practical interest. Each physical layer has a preferred SIR, γ_0, easily identified in the graph of the FSF.

The special case in which only two terminals, one more "important" than the other, share the cell has been thoroughly solved, and the the second-order conditions for a maximum have been verified. The analysis of this special case illustrates clearly the solution procedure, and develops much intuition. The 2-terminal case is separately discussed in section 8.3.5.

The N-terminal analysis focuses on a specific scenario, in which a few "important" terminals share a cell with many "ordinary" terminals. It is presumed that the system can accommodate all the important terminals at the highest available data rate. But it is not clear how many, if any, of the ordinary terminals should be set to operate at this high data rate, and at which rate should operate the others, in order to maximize the cell's weighted throughput. A complete solution procedure is given, which finds for all terminals the data rates (through the corresponding spreading gains) and the transmission power levels (through the corresponding carrier-to-interference ratios) that maximize the cell weighted throughput. Additionally, specific numerical examples are provided and discussed. In the end, terminals end up divided in two groups: favored, which operate at the

highest available data rate, and non-favored, which achieve the preferred SIR of γ_0. It is significant that this SIR is a respectable value. For example, for a simple, but plausible FSF (equation (8.6)), $f(\gamma_0) = 0.83$. Thus, even "non-favored" terminals enjoy reasonable frame-error performance. The set of favored terminals always include all important terminals, and may include some ordinary ones.

In describing the solutions to FONOC, certain assumptions are made concerning the shapes of the graphs arising from some functions of the derivative of the FSF. Specifically, in discussing the two-terminal model, the observation that the graphs of $x f'(x)$ and $x^2 f'(x)$ are "bell curves", as shown in figure 8.1, play a fundamental role. In fact, what is needed is that these graphs be strictly quasi-concave; i.e., these functions must be strictly increasing between zero and the respective positive value where each has its maximum, and be strictly decreasing beyond that point. For the N-terminal analysis, the assumptions made on the derived functions are (i) that the function f is such that:

$$(ax+b)^2 f'(x) \text{ be strictly quasi-concave } \forall x \geq 0 \text{ and } \forall a,b \in [0,1] \tag{8.100}$$

and (ii) that the shapes of the graphs displayed in fig. 8.3 are as shown. Specifically, the graphs of $(x/G_0+1)^2 f'(x)$ must be quasi-concave (which is implied by condition (8.100) with $a = 1/G_0$ and $b = 1$), which leads to the X-shaped graph; and that $f'(x)$ be single-peaked (which is implied by condition (8.100) with $a = 0$ and $b = 1$), and which leads to the U-graph.

However, as of this writing, no formal proof is available showing that for *all* S-curves these graphs are as desired. Technically, this means that the analysis describing the solutions to FONOC applies to the subset of S-curves for which the concerned graphs have the desired properties. Practically, this means that before applying this analysis, the engineer should verify that the frame-success function corresponding to the specific physical-layer of interest is such that the pertinent graphs have the desired shapes. If they don't, this analysis needs to be adapted. Notice, however, that the analysis in previous chapters is not affected by these restrictions.

This model is extended in chapter 9 to consider non-negligible noise, as well as the presence of media-transmitting terminals operating at a fixed data-rate with inflexible SIR requirements. Considering non-negligible noise is important because the noise term may include out-of-cell interference, which is often substantial. Future studies may consider the issues of QoS, fairness, and decentralized implementations, all of which are of practical importance.

Chapter 9

Maximal Data Throughput in the Presence of Power Limited Media Terminals

9.1 Introduction

Modern wireless networks will accommodate simultaneous transceivers operating at very different bit rates. Some of the transceivers may be transferring data, while others transfer media content, such as voice, images, or video. Chapter 8 studies throughput maximization in a model relevant to a single-cell VSG-CDMA system in which each *data* terminal can operate within a range of bit rates, assumed continuous for tractability. This chapter discusses how to extend that model to consider three additional items: (i) Transmission power limits, (ii) non-negligible out-of-cell interference, and (iii) the presence of media-transmitting terminals with fixed bit rates and inflexible SIR requirements.

Power limitations are important for obvious reasons. However, when out-of-cell interference is negligible (system is "interference limited"), the noise term in the SIR expression may be neglected. Then, the power allocation question reduces to finding a vector of carrier-to-interference ratios expressing power ratios between the received power of the terminals. For example, when there are only two terminals, power allocation reduces to finding the optimal ratio between the received power of the two terminals. In theory, the specific power levels are arbitrary, as long as the optimal ratio is maintained. However, when the noise term includes strong out-of-cell interference, the power limitations of the terminals need to be taken explicitly into account. Additionally, there may be media-transmitting terminals operating at fixed bit rates and SIR. From the stand point of the data terminals, these media terminals appear as additional sources of "noise", which decrease the total data throughput.

In this chapter, data terminals continue to be delay-tolerant, with power and data rates that can be assigned at will within specified limits, to maximize the (weighted) throughput. However, the

media-transmitting terminals operate at fixed data rates, and have inflexible SIR requirements. The media terminals may belong to various classes, each identified by its data-rate and SIR pair. The data terminals may also belong to various classes, each identified by how its throughput is weighted by the network. There are two possible weights, which admit various interpretations, including levels of importance or priority among the *data* terminals, "utility" per bit, or monetary prices. The data rates of the data terminals, and the power levels of all terminals are allocated to maximize the sum of the weighted throughputs of each *data* terminals, while respecting the fixed operating conditions of the media terminals.

At the core of this analysis is a frame-success function (FSF) that gives the probability that a data packet is received successfully in terms of the terminal's received signal-to-interference ratio (SIR). This function depends on many physical attributes of the system, such as the modulation technique, the forward error detection scheme, the nature of the channel, and properties of the receiver. No particular algebraic functional form ("equation") is imposed on the FSF. Rather, it is assumed that *all that is known* about this function is that its graph is a smooth S-shaped curve, as displayed in fig. 8.1, and properties derived from this shape form the basis of this analysis. Hence, this analysis should apply to many physical layer configurations of practical interest. Chapter 3 discusses further this modeling approach.

Below, a relatively simple optimization model relevant to uplink data and media transmission in one VSG-CDMA cell is built. This chapter focuses on the special case in which a power-limited media terminal interacts with two data terminals, one of which is more "important" than the other (the model built below can handle a somewhat more general situation than that which is analyzed). The aim is to show that much of the analysis of chapter 8 can still be applied, with relatively minor modifications, to the more complicated and realistic situation of this chapter. The first-order necessary optimizing conditions (FONOC) for the dual class situation of interest are presented, and two possible solutions to FONOC are discussed: one in which only the important data terminal operates at the highest available data rate, and another solution in which both data terminals operate at this rate. This analysis makes clear that the development of chapter 8 can be extended to consider the situation of this chapter with only superficial modifications.

9.2 Problem Formulation

9.2.1 Optimization Model

$$\max_{G_i,\alpha_i} \sum_{i=1}^{N_D} \beta_i T_i(G_i,\alpha_i) \tag{9.1}$$

subject to

$$s_0 := \sum_{i=1}^{N_D} \frac{\alpha_i}{1+\alpha_i} < 1 \tag{9.2}$$

$$G_i \geq G_0 \qquad \{1 \leq i \leq N_D\} \tag{9.3}$$

$$G_3 = \bar{G}_3 \tag{9.4}$$

$$\alpha_3 = \frac{\bar{\gamma}_3}{G_3} \tag{9.5}$$

$$P_i \leq \bar{P}_i \qquad \{1 \leq i \leq N\} \tag{9.6}$$

In this simple model,

1. $N_D = 2$ is the number of *data* terminals, while $N_M = 1$ is the number of *media* terminals. $N = N_D + N_M = 3$.

2. The throughput of *data* terminal i is defined as $R_C T_i(G_i, \alpha_i)$, with

$$T_i(G_i, \alpha_i) := \frac{f(G_i \alpha_i)}{G_i} \tag{9.7}$$

3. $G_i = R_C/R_i$, $i \in \{1, \ldots, N\}$ is the spreading gain of terminal i; i.e., the ratio of the channel's chip rate , R_C to the terminal's data transmission rate R_i (bits per second). $G_0 \geq 1$ is the lowest available spreading gain (determined by the highest available data rate).

4. α_i is the carrier-to-interference ratio (CIR) of the signal from terminal i received at the base station. α_i is defined as,

$$\alpha_i := \frac{P_i h_i}{\sum_{\substack{j=1 \\ j \neq i}}^{N} P_j h_j + \sigma^2} = \frac{Q_i}{\sum_{\substack{j=1 \\ j \neq i}}^{N} Q_j + \sigma^2} \tag{9.8}$$

   with P_i the transmission power of terminal i, h_i its path gain, $h_i P_i := Q_i$ its received power, and σ^2 a representative of the average noise power and, possibly, out-of-cell interference. It can be shown that, with $\sigma^2 > 0$, the CIR's must be such that $\sum \alpha_i/(1+\alpha_i) < 1$ (constraint (9.2)) to ensure that a set of positive received powers exist that produce the given α_i's. (See appendix B, and references [32, 1]). However, some of the resulting power levels may be too high for some terminals. This is discussed below.

5. The product $G_i \alpha_i$, denoted as γ_i, is terminal i's signal to interference (SIR) ratio. For media terminals, a specific SIR value must be provided. For data terminal, the SIR is to be determined optimally, along with the data rates, to maximize the network's weighted throughput. Notice that

$$\alpha_i/(1+\alpha_i) \equiv 1/(1+\alpha_i^{-1}) \equiv 1/(1+G_i/\gamma_i) \tag{9.9}$$

6. Each terminal has an upper bound on its transmission power, $\bar{P}_i$. For convenience, $h_i\bar{P}_i = \bar{Q}_i$ is set. For media terminals, $\hat{h}_i = (1+\bar{G}_i/\bar{\gamma}_i)h_i$ is defined as the terminal's "effective" path gain, because the analysis shows that the terminal with the lowest $\hat{h}_i\bar{P}_i$ has the greatest difficulty in reaching the power level leading to its desired SIR. The greatest limitation to network performance is imposed by the terminal in the worst situation. Because of the inflexible SIR requirement of media terminals, it is less favorable for the cell that the terminal in the worst situation be a media terminal, as opposed to a data terminal, and this is assumed below.

7. $\beta_i \geq 1$ is a weight, which admits various practical interpretations. In the special case discussed in this chapter, $1 = \beta_1 \leq \beta_2 = \beta$

8. There is a frame-success function (FSF), f_S, which gives the probability of the correct reception of a data packet in terms of the received SIR. We assume that *all that is known* about this function is that $f(x) := f_S(x) - f_S(0)$ has the general properties of the generalized "S-curve" discussed in chapter 2 (see fig. 8.1), and that it has a continuous second derivative. Because $f_S(0)$ is very small, the difference between f_S and f is generally negligible. Nevertheless, this correction is made for technical reasons. To provide numerical examples, the FSF corresponding, under suitable assumptions, to non-coherent FSK modulation, with no FEC, and packet size 80, which is given by equation (8.6), is used.

In the development below, an asterisk used as a superscript on a variable denotes a specific value of the variable which satisfies certain optimality condition. Any data terminal operating at maximal data rate is referred to as "favored" or "favorite", and a data terminal in the high-weight class is termed "important", as opposed to "ordinary".

9.2.2 Power Limitations

When constraint (9.2) holds, the resulting received power levels are such that

$$Q_i = \frac{\sigma^2}{1-s_0}\frac{\alpha_i}{1+\alpha_i} \tag{9.10}$$

with

$$s_0 := \sum_{i=1}^{N}\frac{\alpha_i}{1+\alpha_i} \equiv \sum_{i=1}^{N}\frac{1}{1+G_i/\gamma_i} \tag{9.11}$$

(See appendix B, and references [32, 1]).

By observing that

$$\frac{\alpha_i}{1+\alpha_i} + \frac{1}{1+\alpha_i} \equiv 1 \tag{9.12}$$

s_0 can be written as

$$s_0 := \sum_{i=1}^{N}\frac{\alpha_i}{1+\alpha_i} \equiv N - \sum_{i=1}^{N}\frac{1}{1+\alpha_i} \tag{9.13}$$

But with power limitations, some terminals may not be able to reach the power level given by eq. (9.10). To avoid this, the feasibility condition given by inequality (9.2) is modified, as in [32], as follows :

$$\begin{aligned} \forall i, \quad Q_i = \frac{\sigma^2}{1-s_0}\frac{\alpha_i}{\alpha_i+1} \leq h_i\bar{P_i} \quad &\rightarrow \\ \forall i, \quad s_0 \leq 1 - \frac{\sigma^2}{h_i\bar{P_i}}\frac{\alpha_i}{\alpha_i+1} \quad &\rightarrow \\ s_0 \leq 1 - \frac{\sigma^2}{\min_i\{(1+1/\alpha_i)h_i\bar{P_i}\}} \quad &\rightarrow \\ s_0 \leq 1 - \frac{\sigma^2}{\left(1+\bar{G_N}/\bar{\gamma_N}\right)h_N\bar{P_N}} \quad &(9.14) \end{aligned}$$

Inequality 9.14 assumes that terminal N is in the "worst situation". For example, the data terminals may be such that $h_i\bar{P_i} \geq \left(1+\bar{G_N}/\bar{\gamma_N}\right)h_N\bar{P_N}$ for $1 \leq i \leq N_D$. This guarantees that regardless of the optimal choice of α_i, a data terminal will not minimize $(1+1/\alpha_i)h_i\bar{P_i}$.

9.3 Solving the special case

Below, the special case in which the cell is shared by three terminals: an "ordinary" data terminal, whose throughput is weighted by one, an "important" data terminal whose weight is $\beta > 1$, and a media terminal with inflexible data rate and SIR requirements is discussed. Pessimistically, it is assumed that the media terminal also has the most stringent power limitation (for $i \in \{1,2\}$, $h_i\bar{P_i} \geq (1+\bar{G_3}/\bar{\gamma_3})h_3\bar{P_3}$).

9.3.1 Optimization Model Restated

$$\max_{G_i,\alpha_i} \frac{f(G_1\alpha_1)}{G_1} + \beta\frac{f(G_2\alpha_2)}{G_2} \tag{9.15}$$

subject to

$$\begin{aligned} \frac{\alpha_1}{1+\alpha_1} + \frac{\alpha_2}{1+\alpha_2} &\leq 1-\varepsilon_3 && (9.16)\\ G_i &\geq G_0 \; i \in \{1,2\} && (9.17)\\ G_3 &= \bar{G_3} && (9.18)\\ \alpha_3 &= \bar{\gamma_3}/\bar{G_3} && (9.19) \end{aligned}$$

Constraint (9.16) follows from (9.14) with

$$\varepsilon_3 = \left(1+\frac{\sigma^2}{h_3P_3}\right)\frac{1}{1+G_3/\bar{\gamma_3}} \tag{9.20}$$

Some reflexion indicates that constraint (9.16) should be satisfied with equality. Otherwise, the throughput could be increased by raising either α_i, while still satisfying constraint (9.16). However, it is not clear a priori whether either or both of constraints (9.17) should be satisfied with equality.

9.3.2 First-Order Necessary Optimizing Conditions (FONOC)

The Lagrangian corresponding to this problem can be written as

$$\begin{aligned} & T_1(G_1\alpha_1) + \beta T(G_2\alpha_2) \quad + \\ \lambda\left(\sum_{i=1}^{2}\frac{\alpha_i}{1+\alpha_i} - 1 + \varepsilon_3\right) & + \sum_{i=1}^{2}\mu_i(G_0 - G_i) \end{aligned} \tag{9.21}$$

The FONOC can be expressed in vector form, with $\gamma_i = G_i\alpha_i$, as:

$$\begin{bmatrix} \partial T_1(G_1,\alpha_1)/\partial G_1 - \mu_1 \\ \beta\partial T_2(G_2,\alpha_2)/\partial G_2 - \mu_2 \\ f'(\gamma_1) + \lambda(1+\alpha_1)^{-2} \\ \beta f'(\gamma_2) + \lambda(1+\alpha_2)^{-2} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{9.22}$$

with

$$\sum_{i=1}^{2}\frac{\alpha_i}{1+\alpha_i} = 1 - \varepsilon_3 \tag{9.23}$$

$$\mu_1(G_0 - G_1) = 0 \tag{9.24}$$

$$\mu_2(G_0 - G_2) = 0 \tag{9.25}$$

Notice that

$$\frac{\partial T_i(G_i,\alpha_i)}{\partial G_i} = \frac{\gamma_i f'(\gamma_i) - f(\gamma_i)}{G_i^2} \tag{9.26}$$

and from eq. (9.13), condition (9.23) can be equivalently stated as

$$\sum_{i=1}^{2}\frac{1}{1+\alpha_i} = 1 + \varepsilon_3 \tag{9.27}$$

9.3.3 Solving FONOC

9.3.3.1 A single-favorite boundary solution

The development in the preceding chapter suggests the investigation of a solution to FONOC in which the important data terminal operates at maximal data rate ($G_2 = G_0$), with the data rate of the ordinary terminal somewhere within its allowable range (i.e., $\mu_1 = 0$, which allows any $G_1 \geq G_0$ per "complementary slackness" condition (9.24)).

From the top row of the matrix equation (9.22), $\gamma_1 f'(\gamma_1) = f(\gamma_1)$ is obtained, which is an equation of the general form:

$$x f'(x) = f(x) \tag{9.28}$$

With f an S-curve, there is a unique positive value γ_0 which satisfies equation (9.28), which can be seen in figure (8.1) at the tangency point between the graph of f and a straight line from the origin. Therefore,

$$G_1^* \alpha_1^* = \gamma_0 \tag{9.29}$$

Combining eq. (9.29) with the bottom half of the matrix equation (9.22) yields

$$\left(\frac{1+\alpha_2}{1+\alpha_1}\right)^2 \frac{f'(\gamma_2)}{f'(\gamma_0)} = \frac{1}{\beta} \tag{9.30}$$

Equation (9.27) can be written as

$$\frac{1+\alpha_2}{1+\alpha_1} = (1+\varepsilon_3)\alpha_2 + \varepsilon_3 \tag{9.31}$$

Combining eqs. (9.30) and (9.31) yields, with x in place of γ_2, :

$$\left((1+\varepsilon_3)\frac{x}{G_0} + \varepsilon_3\right)^2 \frac{f'(x)}{f'(\gamma_0)} = \frac{1}{\beta} \tag{9.32}$$

In eq. (9.32), all quantities, except for x, are presumed known. Thus, this is a single-variable equation. Notice that $G_0 \geq 2$; and values of $\varepsilon_3$greater than or equal to 1 are useless, because if $\varepsilon_3 \geq 1$ condition (9.22) cannot possibly be satisfied; thus, $(1+\varepsilon_3)(x/G_0) + \varepsilon_3 \leq x+1$. This fact is useful in arguing that $((1+\varepsilon_3)(x/G_0) + \varepsilon_3)^2 f'(x)/f'(\gamma_0)$ has the same "bell-shaped" graph of the function $(x+1)^2 f'(x)$ (fig. 8.1). This implies that, if G_0 is "too large", the "top" of this bell may fall below $1/\beta$, unless β is also "very large". Thus, eq. (9.32) may have no solution. On the other hand, when G_0 is sufficiently small and/or β is sufficiently large, two values of x, on either side of the peak, will satisfy eq. (9.32). Let the larger value, δ_0, be chosen as the FONOC-solving SIR for the important terminal. With this value, α_1 is directly obtained from eq. (9.27), and by plugging the α_1 value into eq. (9.29), G_1 is obtained. Thus, a complete "single-favorite" solution to FONOC is found. However, if the resulting α_1^* is negative, or if $G_1^* < G_0$, this solution is useless, and a "dual-favorite" solution, with both data terminals operating at the highest available data rate, must be sought.

9.3.3.2 Dual-favorite Boundary Solution

In the preceding section, the SFBS, in which only the important terminal operates at the lowest available spreading gain (highest data rate), was considered. It was pointed out that the SFBS may fail to exist depending on the values of the parameters G_0, β. In this section, a "greedy" solution to FONOC, in which both terminals operate at the highest available data rate, is sought.

Working with the last two rows of equation (9.22) it is established, with $x = \gamma_2$ and $y = \gamma_1$, that:

$$f'(y)\left(1+\frac{y}{G_0}\right)^2 = \beta f'(x)\left(1+\frac{x}{G_0}\right)^2 \tag{9.33}$$

Eq. (9.27) can be re-written as

$$\frac{1}{1+x/G_0}+\frac{1}{1+y/G_0} = 1+\varepsilon_3 \tag{9.34}$$

Eqs. (9.33 and 9.34) form a system of two non-linear equations in two unknowns. This system can be solved. Its solution is characterized through fig. 9.1.

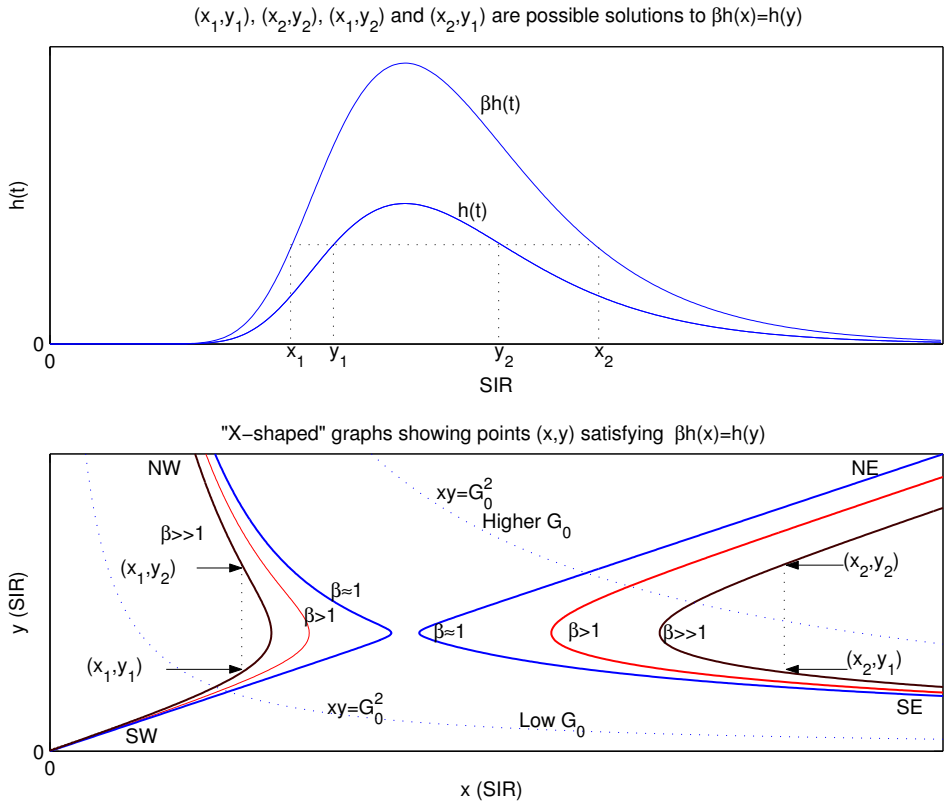

Figure 9.1:
With x and y respectively the SIR of the favorite and sub-favorite terminals, FONOC requires that $\beta h(x) = h(y)$, with $h(t) = f'(t)\,(1+t/G_0)^2$. Any of the pairs (x_1,y_1), (x_2,y_2), (x_1,y_2), or (x_2,y_1) (top) satisfies this equation, but may not be feasible. We plot all such points, which reveals an "X-shaped" graph for each β. NE, NW, SW and SE are directional labels used to identify the "legs" of the X . On the same axes, we plot the hyperbolic curves (dotted) which represent the constraint equation (9.34). When G_0 is low, the hyperbola may *only* intersect the SW leg of the X-curve, which leads to a minimum.

9.4 Discussion

The optimal power levels and data rates for data terminals that share one base station with media terminals, which have fixed bit rates and inflexible SIR requirements, have been investigated. This scenario is relevant to 3G CDMA. The objective is to maximize the *weighted* sum of the *data* terminal's throughput, while honoring QoS commitments made to the media terminals. Two weights, which admit various interpretations, including levels of importance, "utilities", or monetary prices, are considered. The properties of the physical layer are embodied in the frame success function (FSF), which gives, in terms of received signal-to-interference ratio (SIR), the probability that a data packet is correctly received. No specific functional form ("equation") is imposed on the FSF. It is assumed that *all that is known* about the FSF is that its graph is "S-shaped", and the analysis follows from properties derived from this shape (some additional technical assumptions are needed by certain results, as discussed in section 8.5). Therefore, many physical layer configurations of practical interest are accommodated. Each physical layer has a preferred SIR, γ_0, easily identified in the graph of the FSF.

The main conclusion of this chapter is that the analysis in chapter 8, in which no media-transmitting terminals are considered, and where out-of-cell interference is neglected, can be readily adapted to the more general and interesting situation discussed in this chapter. The effect of the media terminals, the out-of-cell interference (noise), and the power limitations of the terminals, combine into a single term, ε_3, that reduces the right-hand-side of the constraint on the carrier-to-interference ratios. Thus, the expression $\sum \alpha_i/(1+\alpha_i) = 1$ becomes $\sum \alpha_i/(1+\alpha_i) = 1-\varepsilon_3$. The objective function, and other constraints remain unchanged. The $\varepsilon_3$term appears, harmlessly, in certain intermediate expressions, but does not alter the shapes of the key graphs describing the solutions, or the fundamental conclusions of the analysis in the preceding chapter. The discussion in section 8.5 applies to the analysis in this chapter. Particularly, the discussed technical limitations imposed by the assumptions on the shapes of the derived graphs also apply to the development herein.

Chapter 10

Conclusions, Limitations and Future Directions

10.1 Retrospective Overview

This work has presented a tractable analytical framework useful to the analysis of resource management issues in the context of wireless communications. Several application have been studied, emphasizing centralized and decentralized resource allocations for data and media communication. Specific applications include decentralized power control making use of “mechanism design” to achieve an efficient allocation, power and data rate assignment for maximal “weighted” cell throughput in a 3G CDMA context, power and coding rate selection for video streaming when video segments have been scalably encoded, and choosing the “right amount” of tolerable media distortion when fidelity is expensive. An appendix addresses capacity questions in a 3G CDMA cellular system, when base station receivers decode cooperatively (macrodiversity). While the cellular third-generation CDMA-based architecture has often been targeted, the fundamental ideas can be transferred to other communication scenarios.

The proposed framework has three key elements: (i) a tractable abstraction of the human sensory system, (ii) a tractable abstraction of the physical layer of a wireless communication link, and (iii) a fundamental technical result. In these 3 elements, a function about which all that is known is that it is an "S-curve", plays a central role. The fundamental result involves the maximization of the ratio $f(x)/x$ with f an S-curve. Without imposing any particular “equation” on the considered functions, the solution to this maximization problem is shown to always exist, be unique, and be graphically describable. A tangent line drawn from the origin to the graph of f specifies the optimal solution. The ratio $f(x)/x$ is also shown to be quasi-concave.

Chapters are largely self-contained, reflecting the fact that this report evolved from individual papers devoted to specific applications. After each chapter, a discussion section is provided, outlining and interpreting the main lessons learned, and discussing some of the limitations. Further comments on results, limitations, and future extensions of a more general nature are made below.

But before, an apparent digression is incurred, by addressing the problem of optimally allocating resources among various subprojects within a personal research program. This discussion provides some insights into the research priorities that led to this document.

10.2 Allocation of effort among research projects

A researcher must optimally allocate a fixed amount of a resource, say time, among a selection of a "long" list of possible projects, part of a research program. The more time he devotes to a given paper, the greater its "utility" ("quality", "impact", usefulness), but the lesser the time left to pursue other projects. What is the "right thing to do"? Finding the right answer necessitates two items : (i) a clear specification of the relation between the utility of a paper and the amount of resource devoted to it, and (ii) a clear criterion defining the goal of this optimization.

One would expect (hope?) that the utility of a paper is increasing with the time spent on it, and that there is a maximum level of utility the paper arising from a given project can reach. The development in chapter 1 suggests that a good hypothesis about the function $U(t)$ giving the quality of the paper as function of the amount of resource devoted to it is that it is some S-curve. A single S-curve may apply to all considered papers because they are all part of the same research program, and this analysis involves a single researcher. As for the optimization criterion, it is obvious that maximizing the total number of completed papers would lead to a large number of "useless" papers, whereas maximizing the quality of each individual paper may result in "too few" papers. A reasonable criterion is then to maximize the "total utility", which is obtained by adding up the utility of each completed paper. That is, if the amount of available resource is T and t units are devoted to each paper, the total utility is $(T/t)U(t)$. Hence, the researcher should allocate to each paper the amount of resource t^* that maximizes $U(t)/t$ ("quality per unit of resource"), which, as chapter 2 shows, is uniquely determined by a tangent to the S-curve U drawn from the origin (unless $t^* > T$ in which case T is the maximizer).

The main challenge to apply the preceding analysis in a real situation may be to estimate the quality/time S-curve. Yet, the analysis provides a valuable and intuitive lesson: the projects have an optimal "stopping time", which defines an "efficient" level of "quality". For a researcher, to spend more "effort" on a given project to increase its "quality" beyond this level is inefficient, in the sense that it reduces the total "impact" of the researcher's efforts. And this intuitive lesson, to a great extent, has guided resource allocation to the various problems analyzed in this work. Thus, none of the chapters are "finished", if this term is taken to mean that no interesting issues of technical or practical importance are left to be explored.

10.3 "Unfinished" business

At the end of each chapter there are comments on the main results, and on limitations and desirable extensions. Below there are additional comments on some high-priority items which, if successfully

pursued in the future, would enhance this research program.

The key technical analysis involving the maximization of the ratio $f(x)/x$ with f an S-curve is, to the best of the author's knowledge, rigorous and finished. However, in several parts of this document, it becomes useful to maximize a ratio of the form $h(x)/x$ where $h(x) = f(\phi(x))$; that is, h is a composite function of an S-curve, and some other monotonic function ϕ. In chapter 1, in the development leading to the quality-rate relation, ϕ is a convex curve. In determining the optimal SIR for video streaming (chapters 1 and 6) ϕ arises as another S-curve. In fact, it is desirable to extend the $f(x)/x$ result to consider the somewhat more general problem of maximizing $f(x)/g(x)$. This ratio can be written as $h(t)/t$, with $t = g(x)$, $h(t) = f(\phi(t))$ and $\phi(t) = g^{-1}(t)$. In these cases, when ϕ is a "well-behaved" monotonic function, it is reasonable and intuitive to expect that the composite function $f(\phi(\cdot))$ retain the S-shape, and numerical experimentation has confirmed it. However, this work does not provide a formal proof of this fact. That is, this work does not prove that the composite function "starts out" convex, and smoothly transitions to convex as it approaches a horizontal asymptote. Formally proving this fact, or specifying the general conditions under which it is true, should be high in an agenda of future research.

Chapter 4 applies a "mechanism" available in the economics literature to achieve an efficient decentralized power allocation among data terminals sharing a CDMA cell. Reference [43], the original economics paper, shows the efficiency of the allocation of this mechanism in a fairly general scenario, and outlines a simple algorithm that leads the terminals to the efficient allocation even when they are not fully informed about the "situation" of each other. Chapter 4 partially characterizes the allocation arising when this mechanism is applied to two terminals in the presence of successive-interference cancellation (SIC), a situation under which one terminal creates interference for the other, but *not* vice-versa. More analytical and numerical work is needed to fully characterize this situation. Additionally, the more common situation in which terminals interfere with each other needs to be explored further.

The video streaming analysis discussed in chapters 1 and 6 assumes, for simplicity, that the channel is "quasi-deterministic", in the sense that the average throughput is treated as a deterministic quantity. This simplification leads to a clear and intuitive result, involving the composite function of two S-curves, one determined by the physical layer, and the other by the human-visual system. A more rigorous analysis which takes explicitly into account the stochastic nature of the channel is desirable. All the media models in chapters 5, 6 and 7 are point-to-point. Extension to a multi-user scenario can be obtained, for example, by utilizing the game theory framework discussed in chapter 4. The key difference lies in the indices to be maximized by the terminals: bits per Joule in chapter 4, versus quality per Joule in 6.

The throughput maximization analysis of chapters 8, and 9 has a number of technical limitations most of which are discussed at the end of chapter 8. Some of the most interesting extensions are mentioned at the end of chapter 8, and include quality-of-service constraints for the data terminals, "fairness" issues, considering many cells, and decentralized implementations. Additionally, in these chapters the weights ($\beta's$) are taken as given. Several optimization problems involving these

coefficients could be set up. If they are interpreted as prices per bit set by the cell administrator, determining these prices is of interest. And if they are interpreted as weights reflecting "priorities" or "importance", since it is advantageous to have a high β, an obvious question is how to assign them and why. For instance, $\beta's$ may be auctioned or otherwise sold. An additional issue is that of the availability of the orthogonal variable-spreading-factor (OVSF) codes employed to implement VSG-CDMA, the technology targeted by the analysis. These codes are limited, and they follow a tree structure such that a long code has a "parent" code of a shorter length. When a code is assigned, none of its "children" can be employed. Recent works in the literature focused on managing these codes include [23, 38]. Chapters 8 and 9 implicitly assume that the desired codes are always available. Introducing the code-availability constraint is of practical interest, although it is a difficult problem to approach analytically, and may lead to combinatorial difficulties.

10.4 Main contributions

The main accomplishments of this work are best determined by others. Nevertheless, it may be appropriate to outline, from an advocative point of view, some of the ideas in this document that should receive priority consideration by anyone seeking its main contributions.

A key analytical tool in this work has been the S-curve. As discussed in chapters 1 and 2, this model has proved to be very useful in many fields, including ecology, biology, engineering, and economics. It appears that all prior studies involving this family of curves rely on specific algebraic formulas, usually associated with certain differential equations. The present work shows that such specific formulas are neither necessary nor helpful. For instance, modeling the frame-success function associated with a wireless communication link as a "formula-free" S-curve (an "abstraction" of the physical layer) yields an analysis that applies to most physical layers of interest.

The S-curve can additionally serve as an abstraction of the human visual system, by capturing the relation between the perceptual quality of a media signal and some objective parameter, such as coding rate or distortion. This approach leads to a "quality-distortion theory", as introduced in chapters 1 and 7. In chapters 1 and 6, the physical and the "human" S-curves "merge" into a composite function, which determines the optimal transmission power (and indirectly the coding rate) of a video streaming system. The S-curve representing the user's perception of quality is unchangeable, from an engineering point of view. But the channel S-curve could be altered via manipulation of communication items such as the modulation scheme. This raises the provocative thought that by "matching" the channel S-curve to the one representing the human element, a communication system tailored to the specific end user would emerge.

Thus, some of the key ideas in this document concern modeling. Modeling is in general a challenging endeavor. Typically the analyst faces a trade-off between the richness or degree of generality of the model and its tractability. If complexity is of no concern, it is usually straightforward to build a very general model of a complicated phenomenon, at the expense of tractability. Conversely, a very tractable model of such a phenomenon can usually be obtained via a long list of simplify-

ing assumptions, if one is not concerned about the generality of the predictions obtained through that model. It is considerably more challenging to build a general model with some constraint on tractability, or to build a tractable model with some constraint on the generality of what can be learned through the model. It is nearly impossible to build a model of a complicated phenomenon that both makes the analysis more tractable than that of previous models, while simultaneously generalizing the applicability of what can be learned with the new model. Yet, a reasonable argument can be made that the models introduced in this work on the basis of the "formula-free" S-curve, both generalize and simplify the corresponding analysis.

For instance, the frame-success function (FSF), which gives the probability that a data packet is received successfully as a function of the terminal's signal-to-interference ratio (SIR) at the receiver, is determined by many attributes of the physical layer, including modulation, FEC scheme, the nature of the channel, and antenna diversity, if any. An exact expression for this function for a realistic model of a wireless communication situation may be prohibitively difficult or impossible to obtain, and even if available, it may be intractable or very inconvenient, and highly dependent on the chosen physical layer configuration. Abstracting this function as an unspecified S-curve clearly makes the analysis much more general, since it considers a large family of physical layer configurations, with the only significant restriction that they give rise to an S-shaped FSF (or to an FSF that is sufficiently close to an S-curve). Since, over a limited domain, nearly all increasing concave curves, convex curves, step functions, and ramps can be closely approximated by an S-curve, assuming that the FSF is an S-curve is a very mild assumption. What is really surprising, and what makes these ideas truly useful, is that this level of generality seems to come with *zero* complexity cost. On the contrary, by focusing on the shape of the FSF, one is able to present clear and specific results which, in some non-trivial situations, can be easily described by the simple artifice of drawing a tangent to the S-curve from the origin (the "knee" of the S-curve).

The technique used to characterize the solutions to certain systems of equations should also be considered. The idea of characterizing these solutions by focusing on the general shapes of the functions involved may not be new, but it is certainly not common in this field. A good example of this is found in chapter 8, in the caption of figure 8.3.

Figure 8.3, repeated for convenience as figure 10.1 in this chapter, corresponds to a situation in which an "important" data terminal share a 3G CDMA cell with several "ordinary" terminals. The important terminal and (possibly) several ordinary terminals are termed "favored" because they operate at the highest available data rate. Any terminal not operating at this data rate, operate at a specific SIR value found at the "knee" of the S-shaped graph of the frame-success function (where a line that goes through the origin meets the S-curve). In figure 10.1, x is the SIR of the important terminal and y the SIR of the favored ordinary terminals. In order to know the values of x and y that satisfy the first-order optimizing necessary conditions (FONOC), a system of two non-linear equations needs to be solved: eqs. (8.95) and (8.98).

One could have stopped the analysis at that point, and proceeded with numerical experimentation. However, this work proceeds as follows: First, it observes that eq. (8.95) is of the form

$\beta h(x) = h(y)$ ($\beta \geq 1$ is the "weight" of the throughput of the important terminal). It further observes that the graph of $h()$ is a "bell curve" as displayed at the top of fig. 10.1. Then, it recognizes that for any pair (x_2, y_2) satisfying this equation, in which both x_2 and y_2 are to the right of the peak of the bell, there is another pair (x_1, y_1) also satisfying this equation, such that both x_1 and y_1 are to the left of the peak. Some reflection indicates that the "mixed" pairs (x_1, y_2) and (x_2, y_1) also satisfy this equation. Thus, for a given β, the graph of all the points that satisfy equation eq. (8.95) must have four regions. In one region, both coordinates of an order pair are "large" (like (x_2, y_2)), in another region both coordinates are "small" (like (x_1, y_1)), and then there are the two regions corresponding to the "mixed" pairs, in which one coordinate is "large" and the other is "small". What has been described is essentially an X-shaped graph, as displayed at the bottom of fig. 10.1 (increasing β has the effect of "pulling apart" the X, as shown).

A similar analysis leads to the conclusion that the graph corresponding to eq. (8.98) is a U-curve (except when all terminals operate at the highest data rate) as shown fig. 10.1. Generally, there will be four intersections between the U and the X, except that under certain choice of parameters the U may lie above the X and no solutions exist. With all terminals at the highest available data rate the U "deforms" into an "L" (a hyperbola) and, under certain choice of parameters, may only intersect the SW leg of the X, in which both x and y are "small", which would lead to a (local) minimizer, as opposed to a maximizer.

Through this geometric exercise, a great deal is learned about the optimizing SIR values (which will generally be in the NE "arm" of the X). Because this analysis only relies on the general shapes of the function involved (X, U, L, "bell", etc), which are derived from the original S-curve, one can be confident that what has been learned will remain valid for most reasonable choices of parameter, as long as the physical layer is such that its frame-success function is an S-curve, a very mild restriction.

While a great deal of effort has been invested in achieving the technical correctness of this document, it would be adventurous to issue any guaranty in this regard. In fact, given the the high technical content of this work, and its depth, breadth, and length, it would be rather surprising if no significant errors are ever found. Yet, there are some basic ideas and techniques in this document that appear to be fundamentally sound, and more importantly, fundamentally useful, regardless of the technical correctness of any specific mathematical expression. While CDMA, the technology of third-generation wireless communication system is often targeted throughout this work, the basic abstractions and techniques presented here are largely technology-neutral. Thus, it is conceivable that aspects of this work will remain useful beyond the lifetimes of the targeted technology, and of those involved in its writing.

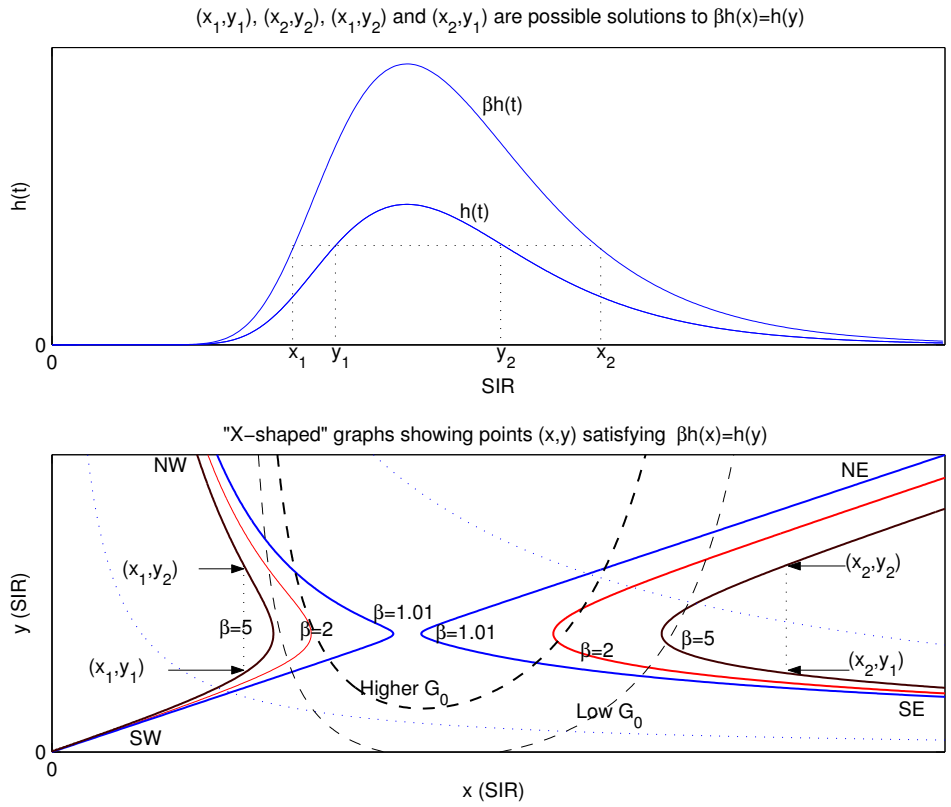

Figure 10.1: Characterizing the solution to a system of non-linear equations through the shapes of key graphs

Appendix A

Some Basic Results on Concavity

Much of this as well as other relevant material can be found in reference [2], in particular in chapter III. The presentation here follows that in the mathematical appendix of reference [21]. However, the material of subsection (A.2.2) is not found in those references, and is developed in full here.

A.1 Concave and convex functions

Consider a function $f : I \to R$, defined on an interval $I \subset \Re$.

Definition: The function f is said to be concave if, $\forall x_1, x_2 \in I$ and $\alpha \in (0,1)$,

$$f(\alpha x_1 + (1-\alpha)x_2) \geq \alpha f(x_1) + (1-\alpha) f(x_2) \tag{A.1}$$

The function f is said to be *strictly* concave if the above inequality holds strictly whenever $x_1 \neq x_2$.

Definition: The function f is said to be (strictly) *convex* if the function $-f$ is (strictly) concave.

A.2 Properties of continuously differentiable concave and convex functions

A.2.1 Tangent line Theorem

The continuously differentiable function $f : I \to R$, defined on an interval $I \subset \Re$, is concave if and only if, $\forall x_1, x_2 \in I$,

$$f(x_2) \leq f(x_1) + f'(x_1) \cdot (x_2 - x_1) \tag{A.2}$$

This function is *strictly* concave if and only if the above inequality holds strictly $\forall (x_1 \neq x_2) \in I$.

The function f is convex if and only if, $\forall x_1, x_2 \in I$,

$$f(x_2) \geq f(x_1) + f'(x_1) \cdot (x_2 - x_1) \tag{A.3}$$

This function is *strictly* convex if and only if the above inequality holds strictly $\forall (x_1 \neq x_2) \in I$.

A.2.2 The Monotonicity of y-intercepts

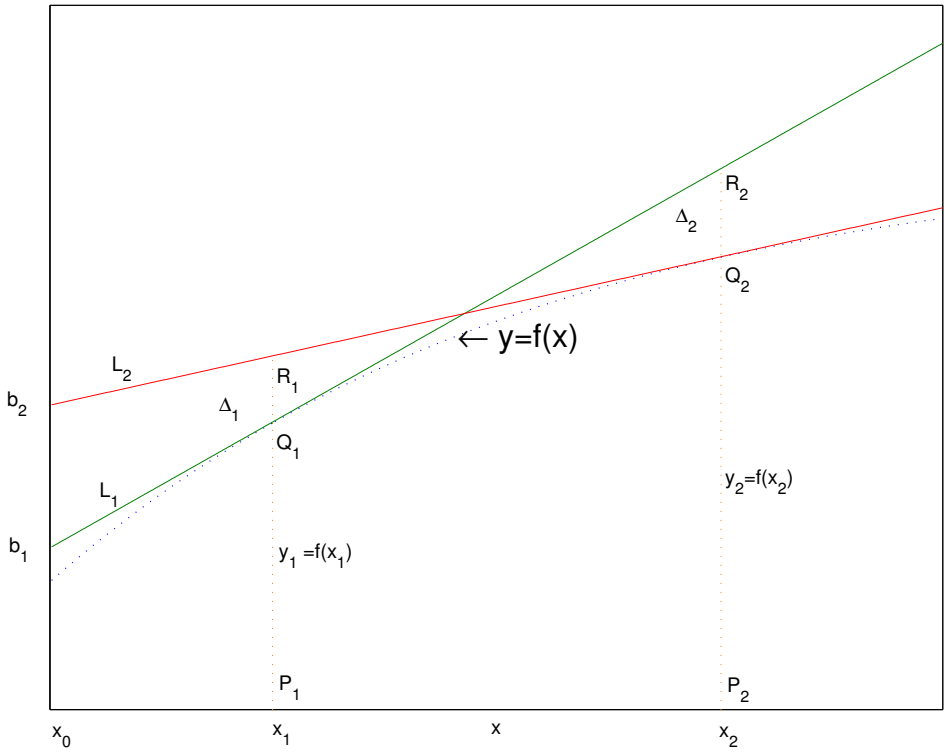

Figure A.1: Increasing Y intercepts

Corollary: Let $f : I \to R$ denote a continuously differentiable *concave* function, defined on an interval $I \subset \Re$. Let x_0, x_1, x_2 be elements of I such that $x_0 < x_1 < x_2$. Then,

$$f(x_2) + (x_0 - x_2) f'(x_2) \geq f(x_1) + (x_0 - x_1) f'(x_1) \tag{A.4}$$

If f is *strictly* concave the above inequality holds strictly.

Proof:

See figure (A.1). In this development, $i \in \{1, 2\}$.

First notice that $g_i(x) = f(x_i) + f'(x_i)(x - x_i)$ denotes the equation of a line tangent at the point (x_i, y_i) $(y_i \doteq f(x_i))$ to the the curve describing the graph of f.

Let $b_i \doteq f(x_i) + (x_0 - x_i) f'(x_i)$.

Thus, b_i is the "height" of tangent line L_i at the abscissa x_0, or its "intercept" with a vertical line drawn at x_0. Hence, inequality (A.4) can be restated as $b_2 > b_1$. In the special case $x_0 = 0$, b_i become the "y-intercept" or ordinate at the origin of the line L_i.

Let $\Delta_1 \doteq g_2(x_1) - y_1$ and $\Delta_2 \doteq g_1(x_2) - y_2$.

Geometrically, Δ_1 is the length of the segment Q_1R_1, which equals the difference between the "height" of the tangent L_2 and the value of the function f, both measured at the abscissa x_1. Δ_2 has an analogous interpretation.

Observe that the points (x_0, b_1), Q_1 and R_2 are all in the line L_1.

Likewise, (x_0, b_2), R_1 and Q_2 are all in the line L_2.
Therefore:

$$\frac{y_1 - b_1}{x_1 - x_0} = \frac{y_2 + \Delta_2 - b_1}{x_2 - x_0} \Rightarrow$$

$$b_1 = \frac{(x_2 - x_0)y_1 - (x_1 - x_0)(y_2 + \Delta_2)}{x_2 - x_1} \tag{A.5}$$

$$\frac{y_2 - b_2}{x_2 - x_0} = \frac{y_1 + \Delta_1 - b_2}{x_1 - x_0} \Rightarrow$$

$$b_2 = \frac{(x_2 - x_0)(y_1 + \Delta_1) - (x_1 - x_0)y_2}{x_2 - x_1} \tag{A.6}$$

Consequently:

$$b_2 - b_1 = \frac{(x_2 - x_0)\Delta_1 + (x_1 - x_0)\Delta_2}{x_2 - x_1} \tag{A.7}$$

By construction, $x_0 < x_1 < x_2$.
By inequality (A.2), both Δ_1 and Δ_2 are non-negative, and both are positive if f is strictly concave. Therefore, the right hand side of equation (A.7) is non-negative, and it is positive, if f is strictly concave.
That is, if f is concave, $b_2 \geq b_1$, and $b_2 > b_1$if f is strictly concave.

Q.E.D.

Given the fact that $-f$ is concave whenever f is convex (see section(A.1)), the following result is immediate:

Corollary: Let $f : I \to R$ denote a continuously differentiable *convex* function, defined on an interval $I \subset \Re$. Let x_0, x_1, x_2 be elements of I such that $x_0 < x_1 < x_2$. Then,

$$f(x_2) + (x_0 - x_2)f'(x_2) \leq f(x_1) + (x_0 - x_1)f'(x_1) \tag{A.8}$$

If f is *strictly* convex the above inequality holds strictly.

Appendix B

Power, Ratios, and Capacity

B.1 From power ratios to power levels : closed-form solution

In certain situations of interest concerning wireless communication networks, rather than dealing with power levels directly, one may wish to choose directly the quantities representing the ratios of each transceiver's power level to the sum of the other transceivers' power levels (plus noise power if applicable). Here we discuss which constraints, besides non-negativity, need to be applied to these power ratios to ensure that they correspond to feasible power levels, and provide a closed-form relation giving the power vector in terms of the ratios.

Specifically, let α_i be defined as :

$$\alpha_i = \frac{P_i h_i}{\sum_{\substack{j=1 \\ j \neq i}}^{N} P_j h_j + \sigma^2} = \frac{Q_i}{\sum_{\substack{j=1 \\ j \neq i}}^{N} Q_j + \sigma^2} \tag{B.1}$$

In this expression, P_i is associated with the transmit power of transceiver i, h_i corresponds to the path gain from transceiver i to the base station, and σ^2 represents the noise power in the base station receiver. $Q_i = P_i h_i$ is then the received power at the base station in the signal transmitted by transceiver i. N represents the number of active users.

Each α_i can be called a transceiver's carrier to interference ratio (CIR). The corresponding signal to interference and noise ratio, SINR is defined as the product $G_i \alpha_i$, with G_i the corresponding "processing gain" or ratio of the channel's "chip rate" to the transceiver's transmission rate.

Notice that implicit in the above formulation is the assumption that a single base station is of interest. Considering multiple base stations complicates the notation, without casting any new light on the problem. Hence, a single base station is considered.

The defining equations for the α_i's (see equation (B.1) above) yield a linear system of equations (see eq. (B.2) below). The α_j's correspond to feasible power ratios, whenever this system can be solved for physically meaningful Q_j's.

One could attack this issue via elementary algebra. However, a matrix algebra approach, centered on the concepts of eigenvalues and eigenvector, is preferred, because it provides more valuable

insights into the structure of the problem. This is not surprising. Eigenvalues and eigenvectors have played a prominent role in the development of power control theory. For instance, see reference [7], which is an influential work.

B.1.1 Problem formulation

The defining equations for the α_i's (see equation (B.1) above) yield a system of equations which can be expressed in matrix form as:

$$\begin{pmatrix} 1 & -\alpha_1 & -\alpha_1 & \cdots & -\alpha_1 \\ -\alpha_2 & 1 & -\alpha_2 & \cdots & -\alpha_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ -\alpha_N & -\alpha_N & -\alpha_N & \cdots & 1 \end{pmatrix} \begin{pmatrix} Q_1 \\ Q_2 \\ \vdots \\ Q_N \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{pmatrix} \sigma^2 \tag{B.2}$$

However, it will prove convenient to divide both sides of each one of the original equations by $1+\alpha_j$ where the index "j" corresponds to the one α_j which appears in the corresponding equation. Thus, for example, the equation corresponding to the second row becomes:

$$-\frac{\alpha_2}{1+\alpha_2}Q_1 + \frac{1}{1+\alpha_2}Q_2 - \frac{\alpha_2}{1+\alpha_2}Q_3 - \cdots - \frac{\alpha_2}{1+\alpha_2}Q_N = \frac{\alpha_2}{1+\alpha_2}\sigma^2 \tag{B.3}$$

Now, for notational convenience, we define:

$$a_k = \frac{\alpha_k}{1+\alpha_k} \tag{B.4}$$

It will prove useful to observe the following trivial algebraic identity:

$$a_k + \frac{1}{1+\alpha_k} = \frac{\alpha_k}{1+\alpha_k} + \frac{1}{1+\alpha_k} = 1 \quad \Rightarrow \quad \frac{1}{1+\alpha_k} = 1 - a_k \tag{B.5}$$

Taking into account (B.4) and (B.5), equation (B.3) can be re-written as

$$-a_2Q_1 + (1-a_2)Q_2 - a_2Q_3 - \cdots - a_2Q_N = a_2\sigma^2$$

After treating all the equations in the system of interest analogously, we can express the system of equations (B.2) as:

$$\left(\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix} - \begin{pmatrix} a_1 & a_1 & a_1 & \cdots & a_1 \\ a_2 & a_2 & a_2 & \cdots & a_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ a_N & a_N & a_N & \cdots & a_N \end{pmatrix} \right) \begin{pmatrix} Q_1 \\ Q_2 \\ \vdots \\ Q_N \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_N \end{pmatrix} \sigma^2 \tag{B.6}$$

Notice that this matrix equation can be expressed as

$$(\mathbf{I}-\mathbf{A})\,\vec{Q}=\vec{a}\sigma^2$$

Above, $\mathbf{I}$ is the $N\times N$ identity matrix, and $\mathbf{A}$ is a strictly positive matrix with each of its columns equal to the vector $\vec{a}$.

B.1.2 Feasibility Condition

According to non-negative matrix theory, the above system has a non-negative solution whenever the Perron eigenvalue of $\mathbf{A}$ is less than 1. See [34, Section 2.1]. Hence, we need to find the largest eigenvalue of $\mathbf{A}$.

By definition, $\lambda,\vec{v}$ are an eigenvalue/eigenvector pair for matrix $\mathbf{A}$ if $\vec{v}\neq\vec{0}$ and they satisfy:

$$\mathbf{A}\vec{v}=\lambda\vec{v} \tag{B.7}$$

But because of the special structure of $\mathbf{A}$ it can be easily verified that, for any vector $\vec{x}$,

$$\mathbf{A}\vec{x}=\left(\sum_{j=1}^{N}x_j\right)\vec{a} \tag{B.8}$$

Comparing eqs. (B.7) and (B.8), it becomes apparent that in order for the scalar/ vector pair $\lambda,\vec{v}$ to satisfy eq. (B.7), it must be that $\lambda=\sum_{j=1}^{N}x_j$ and, either of the following two conditions hold:

i) If $\sum_{j=1}^{N}x_j\neq 0$, $\vec{x}$ must be a multiple of $\vec{a}$ (so that eq. (B.8) is satisfied).

ii) $\vec{x}$ is a non-zero vector such that $\sum_{j=1}^{N}x_j=0$ (which would also satisfy eq. (B.8)).

Notice that, in general, in R^N, one can find $N-1$ linearly independent vectors such that $\sum_{j=1}^{N}x_j=0$

This development completely specifies the eigenvalues and characterizes the eigenvectors of $\mathbf{A}$.

There are only two distinct eigenvalues : $\lambda_1=s\doteq\sum_{j=1}^{N}a_j$ and $\lambda_2=0$, the latter of which has multiplicity $N-1$. By definition, s is the Perron eigenvalue of $\mathbf{A}$.

The eigenvector corresponding to s can be taken to be precisely $\vec{a}$.

In conclusion, the set of values denoted as $\alpha'_j s$ correspond to feasible power ratios whenever $\sum_{j=1}^{N}a_j<1$.

B.1.3 Explicit Solution

One can reach the above conclusion without invoking non-negative matrix theory, by solving explicitly the system of equations of interest: $(\mathbf{I}-\mathbf{A})\,\vec{Q}=\vec{a}\sigma^2$. This can be done with the information obtained through the preceding development. One has to consider separately two cases: $\sigma>0$ and $\sigma=0$.

The case $\sigma > 0$

As discussed above, the right-hand side of the above equation is an eigenvector for the matrix $\mathbf{A}$ corresponding to the eigenvalue $s \doteq \sum_{j=1}^{N} a_j$. This suggests that the relationship between the eigenvalues/eigenvectors pairs of $\mathbf{A}$ and those of the matrix $\mathbf{I} - \mathbf{A}$ be investigated.

In fact, if $\vec{v}$ is and eigenvector for $\mathbf{A}$ corresponding to the eigenvalue λ, then

$$(\mathbf{I} - \mathbf{A})\,\vec{v} = \vec{v} - \lambda\vec{v} = (1-\lambda)\,\vec{v}$$

Thus, $1-\lambda$ and the same $\vec{v}$ are also an eigenvalue/eigenvector pair for $\mathbf{I} - \mathbf{A}$!

In particular, $\vec{a}$ is also an eigenvector for $\mathbf{I} - \mathbf{A}$ with eigenvalue $1 - s$, that is,

$$(\mathbf{I} - \mathbf{A})\,\vec{a} = (1-s)\,\vec{a}$$

From this it follows that, if $s \neq 1$, $\vec{Q} = (\sigma^2/(1-s))\vec{a}$ (which is positive whenever $\vec{a}$ is, and $s < 1$) is the solution of $(\mathbf{I} - \mathbf{A})\,\vec{Q} = \vec{a}\sigma^2$. That is, each component of the power vector must satisfy

$$\frac{\sigma^2}{1-s} a_k \tag{B.9}$$

with $s = \sum_{j=1}^{N} a_j$. This expression is well-defined and physically meaningful whenever $s < 1$.

The case $\sigma = 0$

If noise is negligible, which implies $\sigma = 0$, then in order for the system $(\mathbf{I} - \mathbf{A})\,\vec{Q} = \vec{0}$ to have a non-trivial solution, the determinant of the matrix $\mathbf{I} - \mathbf{A}$ must be zero. This can only happen if $1 - s$, which is the only eigenvalue of $\mathbf{I} - \mathbf{A}$ which is different from one (see preceding discussion about the relationship between the eigenvalues of A and those of $\mathbf{I} - \mathbf{A}$), equals zero; that is, if $s = \sum_{j=1}^{N} a_j = 1$.

In this case the system has infinitely many solutions. It can be verified that any power vector $\vec{Q}$ proportional to $\vec{a}$ is a solution to $(\mathbf{I} - \mathbf{A})\,\vec{Q} = \vec{0}$.

In fact, this has already been established. s and $\vec{a}$ have been shown to be and eigenvalue/eigenvector pair for $\mathbf{A}$. That is, $\mathbf{A}\vec{a} = s\vec{a}$. Therefore, when $s = 1$, $(\mathbf{I} - \mathbf{A})\,\vec{a} = \vec{a} - \vec{a} = \vec{0}$, which confirms that $\vec{Q} = \vec{a}$ is indeed a solution of $(\mathbf{I} - \mathbf{A})\,\vec{Q} = \vec{0}$, as is any $\vec{Q} \propto \vec{a}$.

B.2 Interpretations and Conclusion

Above, the conditions under which each one of a set of positive numbers corresponds to a transceiver's carrier to interference ratio, CIR, have been given. These conditions have been derived by studying the solution of a system of linear equations engendered by the CIR definition (see eq. (B.1)). In fact, a closed-form expression yielding the solution has been given (see eq. (B.9)). The interpretation of these results casts some light on the structure of power control problems, and has some implications

for the modeling of these problems.

Much of the preceding development is centered on some new variables. These variables may, at first glance, seem devoid of physical significance. However, a more deliberate look at them reveals that they, and the conditions given in terms of them, can be interpreted in physically significant manner.

Recall that the a_k's were introduced in eq. (B.4) as

$$a_k = \frac{\alpha_k}{1+\alpha_k}$$

where the α_k's were defined in eq. (B.1) as the received CIR's of certain signals. This means that if the original α_k's are indeed physically meaningful, the a_k's can be expressed in terms of the received power vector as follows:

$$a_k = \frac{\alpha_k}{1+\alpha_k} = \frac{Q_k/I_k}{1+Q_k/I_k} = \frac{Q_k}{I_k+Q_k} = \frac{Q_k}{Q_C} \tag{B.10}$$

Above, I_k is the total interfering power, including noise, experienced by user k, i.e, $I_k = \sum_{\substack{j=1 \\ j\neq k}}^{N} Q_j + \sigma^2$, and Q_C is the total received power, including noise. Thus, our a_k's represent the respective signal's fractional "share" of the total power being received (including noise), or the signal-to-channel ratio, SCR, a physically meaningful quantity.

In fact, a_k can be viewed as a rough "measure" of the channel's "quality" as experienced by user k. If $a_k = 1$, user k's signal power is the only one being received (nor even noise interferes with this signal). This represent an ideal situation, in which any non-negligible amount of power received in this signal will result in error-free transmission. At the other extreme, $a_k = 0$ indicates the worst possible situation from the perspective of user k.

Along these lines, the sum $s = \sum_{j=1}^{N} a_j$ is seen to satisfy

$$s = \sum_{j=1}^{N} a_j = \sum_{j=1}^{N} \frac{Q_j}{Q_C} = \frac{\sum_{j=1}^{N} Q_j}{Q_C} \equiv \frac{\sum_{j=1}^{N} Q_j}{\sum_{j=1}^{N} Q_j + \sigma^2} \tag{B.11}$$

Now, the condition $s < 1$ is discovered to make plenty of sense. If the original α_k's are indeed physically meaningful, as long as $\sigma^2 > 0$, the numerator in the preceding expression, eq. (B.11), is definitively less than the denominator, for which s must indeed be less than 1. And if $\sigma^2 = 0$, $s = 1$ must hold.

From eq. (B.11), it follows that $1 - s$, an expression appearing in eq. (B.9), which gives the power vector in terms of the ratios, represents the noise's fractional "share" of the total received power, σ^2/Q_C. This shows that when the feasibility conditions are satisfied, eq. (B.9) is an identity. That is:

$$\frac{\sigma^2}{1-s} a_k = \frac{\sigma^2}{(\sigma^2/Q_C)} \frac{Q_k}{Q_C} \equiv Q_k$$

Finally, this analysis has some implications for the modeling of the phenomenon of interest. A

critical step in building a mathematical model is to choose an appropriate set of variables. It is well known that some variables help to uncover the underlying structure of the phenomenon of interest and facilitate its analysis, while a different variable choice may hide important interrelations, and complicate the analysis. For example, in the analysis of linear systems, it is often possible to "diagonalize" a matrix representing a linear transformation through a change of coordinates involving the matrix eigenvectors. This new representation can be quite useful in simplifying the analysis.

The development in this note hints that the signal-to-channel ratio (SCR), defined as the ratio of a received signal power to the total received power in the channel (including noise), a quantity with perfectly clear physical significance, may be the "natural" ratio in the analysis of power-control and related phenomena, as opposed to the SINR, which is the ratio "traditionally" favored in the literature.

Both the CIR and the SCR hold the same "information", and the conversion of one to the other is straightforward. However, a candidate SCR vector can be tested for feasibility simply by checking whether the sum of its components is less than 1 (or, if noise is negligible, whether this sum equals one). Likewise, the power vector yielding a desired, feasible SCR vector is directly proportional to the SCR vector, with the constant of proportionality being a simple, physically meaningful function of the sum of the desired SCR's. And if noise is negligible, then the power vector can be taken to be exactly the same as the SCR vector. Hence, the SCR can be a considerably more convenient variable choice than the CIR.

Of course, there is a reason why the CIR has been favored. In the relatively simple AWGN channel, the bit error probability can be shown to be dependent on the signal-to-noise ratio. But the typical wireless channel is considerably more complicated than an AWGN channel. Modeling its bit error probability as determined by the SINR's is a high-level approximation. A similar approximation in terms of the SCR's could be equally justified (or unjustified).

Appendix C

Allocating Limited Power with Elastic Signal-to-Interference Targets

C.1 Introduction

The signal-to-interference ratio (SIR) is a fundamental quality-of-service (QoS) index in the operation of CDMA networks. In many situations, a terminal enters the network intending to perform certain task. This task may require that the frame-error rate be kept under specified limits, and these limits ultimately translates into minimum SIR requirement. Likewise, a terminal may be able to operate at various levels of SIR, but there may be a level which is optimal for the terminal (e.g, this level may maximize the terminal's "utility"). Thus, when a terminal expresses an interest in joining the network, a question that immediately arises is whether the network, under acting conditions, can support the SIR desired/required by the new terminal, without failing to honor previous commitments made to other active terminals. Answering this question is an important resource management issue: admission. This situation is particularly interesting in the context of variable spreading gain (VSG) CDMA, a technique part of 3G standards, in which terminals with dissimilar data rates, share a common "chip rate", but operate with non-identical spreading gains.

The SIR is determined by the power levels of all active terminals, plus random noise, which may actually represent out-of-cell interference. This problem ultimately comes down to determining whether a vector of SIR's is such that there is a "matching" vector of power levels, each meeting appropriate constraints, which produces the desired/required SIR for each terminal. The answer to this question is known, even in the VSG-CDMA context. Each desired SIR, γ_i, can be supported, if they are such that $\sum_i 1/\hat{G}_i < 1$, with $\hat{G}_i = 1 + G_i/\gamma_i$, and G_i the respective spreading gain [1, 32]. However, in this case there is also a specific power level Q_i with which the signal of terminal i must be received. With limited transmission power, a poorly situated terminal may be unable to reach its respective Q_i even while operating at maximal power.

If SIR targets are inflexible, when one or more terminals cannot reach the power level necessary for them to achieve their required SIR, at least one terminal must be refused service or turned off.

The problem, however, becomes more interesting if there are "flexible" terminals, each willing and able to operate below its desired SIR. For example, for a given physical layer, there is an SIR value that maximizes the bits per Joule performance of a data transmitting terminal. But the terminal can also operate at a lower SIR, at the expense of energy efficiency. Likewise, media transmitting terminals could operate at a lower than desired SIR, at the expense of higher media distortion, due to a higher FER. In cases like these, the procedure to be followed to allocate power when some terminals cannot reach the appropriate power level needs to be clarified. This is done in the remainder of this note.

The literature on power control and capacity of CDMA networks is plentiful. Reference [1] is a recent work which discusses many previous publications on this issue, and [10] is an authoritative recent survey. However, to our knowledge, previous works do not address our problem, namely, what to do when terminals cannot reach their desired SIRs but some terminals requirements are flexible. For example, [32] considers maximal received power constraints by imposing conditions on the SIR such that the matching powers are all less than the respective maximal received power. But these conditions can substantially reduce the capacity of the network, when a terminal is severely power limited (for example, because it is in a very bad location), and it is certainly unnecessary if this terminal's SIR requirement is flexible. On the other hand, [30] does touch on our problem, in the context of a power control "game" among data-transmitting terminals. It is that analysis which we clarify and extend in this work.

C.2 Problem Statement

N terminals wish to share a CDMA cell. Out-of-cell interference is included as part of the noise term σ^2. It is immaterial whether or not some of these terminals are not yet active and want to join those already active. Terminal i is characterized by a path loss coefficient h_i to the base station (BS), an upper bound on its transmission power, $\bar{P}_i$, a data transmission rate R_i (which determines a spreading gain $G_i = R_c/R_i$, with R_c the channel's "chip rate"), and its preferences on the SIR space. These preferences are such that it wants the largest feasible SIR in the interval $[\underline{\gamma}_i, \gamma_i]$. Notice that, when energy is limited, a higher SIR is only better up to a point, even in a single-user channel. For instance, given a physical layer, a quasi-concave ("bell shaped") function of the SIR gives the number of bits that a data transmitting terminal can successfully transfer per unit of energy. This means that there is a specific SIR which maximizes bits per Joule. If a terminal operates with SIR that is *lower or higher* than the optimal value, its bits per Joule efficiency suffers [30].

For convenience, we set $h_i\bar{P}_i = \bar{Q}_i$, $\gamma_i/G_i = \alpha_i$, and $1 + G_i/\gamma_i = \hat{G}_i$. We call α_i the carrier to interference ratio (CIR) (the SIR is the product of the CIR by the respective spreading gain), and $\hat{G}_i$ the "effective spreading gain" (unity plus spreading gain per unit of desired SIR). Clearly, $\alpha_i = 1/(\hat{G}_i - 1)$.

We define $\hat{h}_i = \hat{G}_i h_i$ as the terminal's "effective" path gain, because the analysis shows that the terminal with the lowest $\hat{h}_i\bar{P}_i$ has the greatest difficulty to reach the power level leading to its desired

SIR. For expositional convenience, we assume that $\hat{h}_1\bar{P}_1 \geq \hat{h}_2\bar{P}_2 \geq \cdots \geq \hat{h}_N\bar{P}_N$. Thus, terminal 1 is in the "best situation", and terminal N in the "worst situation".

We seek a non-negative vector specifying a received power level for each terminal. This vector must be "optimal" in some reasonable sense. We assume that each terminal values energy, and does not want to spend more energy than it needs to in order to maximize its preferences on the SIR space.

C.3 Solution

C.3.1 When all terminals are power sufficient

Evidently, if there is one feasible power allocation such that, for each i, the SIR of terminal i is its preferred value, γ_i, then we would choose such allocation. Thus, our first task is to investigate conditions under which such allocation is feasible. That is, with $Q_i = h_i P_i$ denoting the *received* power from terminal i, we ask under which conditions a system of N equations of the form:

$$\frac{Q_i}{\sum_{\substack{j=1 \\ j\neq i}}^{N} Q_j + \sigma^2} = \alpha_i \equiv \frac{1}{\hat{G}_i - 1} \tag{C.1}$$

has a non-negative solution, and if so what is it in closed form?

The answer is found in [32], and under slightly more general conditions in [1], and the complete development can be found in appendix B. Equation (C.1) leads to a system of equations:

$$\begin{pmatrix} 1 & -\alpha_1 & \cdots & -\alpha_1 \\ -\alpha_2 & 1 & \cdots & -\alpha_2 \\ \vdots & \vdots & \vdots & \vdots \\ -\alpha_N & -\alpha_N & \cdots & 1 \end{pmatrix} \begin{pmatrix} Q_1 \\ Q_2 \\ \vdots \\ Q_N \end{pmatrix} = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_N \end{pmatrix} \sigma^2 \tag{C.2}$$

One can show that if the condition

$$s_0 := \sum_{k=1}^{N} \frac{\alpha_k}{1+\alpha_k} \equiv \sum_{k=1}^{N} \frac{1}{\hat{G}_k} < 1 \tag{C.3}$$

is satisfied, the system (C.2) has a unique solution, in which each component of the received power vector is given by:

$$Q_k^* = \frac{\sigma^2}{1-s_0}\frac{\alpha_k}{1+\alpha_k} \equiv \frac{\sigma^2}{1-s_0}\frac{1}{\hat{G}_k} \tag{C.4}$$

Evidently, if $\forall i \quad \hat{G}_i = \hat{G}$, then condition (C.3) and equation (C.4) reduce to, respectively, :

$$s_0 = \frac{N}{\hat{G}} < 1 \tag{C.5}$$

$$Q_k^* = \frac{\sigma^2}{\hat{G} - N} \tag{C.6}$$

C.3.2 Some terminals lack sufficient power

C.3.2.1 General strategy

If condition (C.3) is satisfied, and each terminal can reach (at the receiver) the power level given by equation (C.4), then each can achieve its desired SIR by setting its transmit power level at Q_k/h_k. But, evidently, one or more terminals may not be able to reach such level, because of power limitations. Our analysis will show that the terminal with the lowest $\hat{h}_i\bar{P}_i$ (with $\hat{h}_i = \hat{G}_i h_i$) has the greatest difficulty in reaching the power level leading to its desired SIR. We say that this terminal is in the "worst situation". We have assumed, for expositional convenience, that $\hat{h}_1\bar{P}_1 \geq \hat{h}_2\bar{P}_2 \geq \cdots \geq \hat{h}_N\bar{P}_N$. Thus, terminal N is in the "worst situation", followed be terminal $N-1$, and so on down to terminal 1, which is in the "best situation".

After calculating the vector of received power that produces the vector of desired SIRs for all terminals, *infeasible* power levels may result (negative, or positive but too high for some terminals). In this case, we set the terminal in the worst situation, i.e., terminal N, to operate at its maximal power. Then, we re-calculate the power vector necessary for the *other* terminals to achieve their desired SIRs, under this new operating condition. If this new vector (of order $N-1$) is feasible, we stop; otherwise we set the terminal in the second worst situation (terminal $N-1$) to *also* operate at maximal power, and calculate the new vector (of order $N-2$) leading to the desired SIR of the other terminals. If this new vector is feasible, we stop; otherwise, we continue recursively. We end with M terminals operating at maximal power, and each of the remaining ones operating at the power level which allows it to achieve its desired SIR.

C.3.2.2 Capacity cost of serving a terminal in a bad situation

One may be tempted to rule out SIR vectors which demand power levels that are "too high" for any one of the terminals. One could accomplish this by modifying, as in [32], the feasibility condition given by inequality (C.3) as follows :

$$\begin{aligned} \forall k, \quad \frac{\sigma^2}{1-s_0}\frac{1}{\hat{G}_k} \leq h_k\bar{P}_k \rightarrow s_0 \leq 1 - \frac{\sigma^2}{\hat{G}_k h_k \bar{P}_k} \quad \rightarrow \\ \sum_{j=1}^{N} \frac{1}{\hat{G}_j} \leq 1 - \frac{\sigma^2}{\min_k\{\hat{G}_k h_k \bar{P}_k\}} \end{aligned} \tag{C.7}$$

Without power limitations, the capacity of the cell is determined by inequality (C.3). Thus, we can think that, when each terminal is power *unlimited*, the cell capacity is unity. But with power limits, the feasibility inequality becomes (C.7). Thus, $(\hat{G}_k h_k \bar{P}_k/\sigma^2)^{-1}$ can be interpreted as the amount of "capacity" which has to be sacrificed in order to accommodate the power limitation of the terminal in the least favorable situation. The sacrificed capacity is *increasing* in the terminal's

SIR, but *decreasing* in its spreading and path gains, and totally vanishes if its transmission power is unlimited. Notice also that $\hat{G}_k h_k \bar{P}_k$ can be written as $\hat{h}_k \bar{P}_k$. Thus, $\hat{h}_k = \hat{G}_k h_k$ can be called the "effective" path gain. When all terminals have the same transmit power limit, $\hat{h}_k$ determines, which terminal is in the "worst situation". For instance, a terminal in a bad location (low h_i) may do reasonable well if its spreading gain is "large" with respect to its desired SIR .

$\hat{G}_k h_k \bar{P}_k$ could be very small for some terminal, which would happen, for example, when a terminal operating at a high data rate (low G_k) and demanding a high SIR is very far from the BS (low h_k). In such case, the right-hand-side of inequality (C.7) could also be very small, or even negative, which would substantially reduce or totally vanish the set of SIR vectors that can be supported ("capacity region").

C.3.2.3 One terminal at maximal power

For expositional convenience, we have already assumed that terminal N minimizes the RHS of inequality (C.7). First, notice that even if more than one terminal failed to reach the level given by equation (C.4), we should start setting *only* terminal N, at maximal power. By hypothesis, the maximal received power from this terminal is less than that given by equation (C.4). Thus, other terminals will experience less interference in this scenario, than they would have, if terminal N had been able to reach the specified power level. Therefore, it is possible that a terminal which previously could not reach the power level necessitated by its desired SIR, may be able to do so now.

In this scenario, the received power from terminal N, Q_N, is presumed fixed at $h_N \bar{P}_N = \bar{Q}_N$, while others need to be found to satisfy :

$$\frac{Q_i}{\sum_{\substack{j=1 \\ j \neq i}}^{N-1} Q_j + \Sigma_1^2} = \frac{1}{\hat{G}_i} \equiv \alpha_i \tag{C.8}$$

where $\Sigma_1^2 := \bar{Q}_N + \sigma^2$.

Evidently, equation (C.8) leads to a system of equations analogous to (C.2), except that it is of order N-1, and Σ_1^2 replaces σ^2. From the development leading to condition (C.3), the feasibility condition for the existence of a non-negative solution of this new system is:

$$s_1 := \sum_{k=1}^{N-1} \frac{1}{\hat{G}_k} < 1 \tag{C.9}$$

Likewise, if inequality (C.9) is satisfied, a unique solution exists, in which the first N-1 components of the received power vector satisfy:

$$Q_k^* = \frac{\bar{Q}_N + \sigma^2}{1 - s_1} \frac{1}{\hat{G}_k} \tag{C.10}$$

Notice that if inequality (C.3) is satisfied, so is inequality (C.9). But the converse is obviously not

true. This suggests that we can apply the current procedure, even when inequality (C.3) has failed, at the outset.

C.3.2.4 Several terminals maxed out

After proceeding as in section C.3.2.3, we may find that terminal $N-1$, which is in the second worst situation, cannot reach the power level given by equation (C.10). In this case, as discussed in section C.3.2.1, we would set both terminals N and $N-1$ to operate at maximal power, and check whether each of the remaining terminals is able to reach the power level leading to its desired SIR. If this is *not* the case, we would then set terminals $N-2$ through N to operate at maximal power, and verify if each of the other terminals has enough power to achieve its desired SIR. And so on.

The verification step, with M terminals operating at maximal power proceeds as follows. The received power for $i = N-M+1, \ldots, N$ are presumed fixed at $h_i\bar{P_i} = \bar{Q_i}$, while others need to be found to satisfy :

$$\frac{Q_i}{\sum_{\substack{j=1 \\ j\neq i}}^{N-M} Q_j + \Sigma_M^2} = \frac{1}{\hat{G}_i} \equiv \alpha_i \tag{C.11}$$

$$\text{with, } \Sigma_M^2 := \sigma^2 + \sum_{i=N-M+1}^{N} \bar{Q_i}$$

Evidently, equation (C.11) leads to a system of equations analogous to (C.2), except that it is of order N-M, and Σ_M^2 replaces σ^2. From the development leading to condition (C.3), the feasibility condition for the existence of a non-negative solution of this new system is:

$$s_M := \sum_{k=1}^{N-M} \frac{1}{\hat{G}_k} < 1 \tag{C.12}$$

If inequality (C.12) is satisfied, a unique solution exists, in which the first N-M components of the received power vector satisfy:

$$Q_k^* = \frac{\Sigma_M^2}{1-s_M} \frac{1}{\hat{G}_k} \tag{C.13}$$

If $\forall i$, $\hat{G}_i = \hat{G}$, the feasibility condition (C.12) and equation (C.13) become, respectively:

$$s_M = \frac{N-M}{\hat{G}} < 1 \tag{C.14}$$

$$Q_k^* = \frac{\Sigma_M^2}{\hat{G}-N+M} \tag{C.15}$$

C.4 Discussion

In situations of practical interest involving data or media transmission in 3G wireless networks, each terminal may desire a certain optimal SIR, but it may be able and willing to function at sub-optimal

SIR levels. We have presented an analytical procedure to allocate power when some "flexible" terminals cannot reach the power level leading to their optimal SIR, because of power limits and poor location, for example. The procedure is analytical, and the necessary closed-form expressions are provided. In general, we end with M terminals operating at maximal power, and the remaining ones achieving their desired SIRs. M can be as low as zero (each terminal achieves its optimal SIR), or as high as N (no terminal achieves its optimal SIR).

If each of the M maxed out terminals achieves an SIR which is in its acceptable range, the obtained power allocation is perfectly reasonable. It allows each of $N - M$ terminals to operate at its optimal SIR ("satisfied" terminals), while giving each maxed-out terminal an acceptable SIR. Furthermore, this allocation is an "equilibrium", in the sense that no terminal would be better off by *unilaterally* changing its transmission power. This is clear for the satisfied terminals, because any such terminal, by *unilaterally* changing its power, would move its SIR away from its preferred level. And a maxed out terminal would like to but *cannot* increase its power level to raise its SIR closer to its desired value.

On the other hand, if some of the maxed out terminals end up with SIRs that are "too low", it is not clear what should be done. Lowering the SIRs of some/all of the satisfied terminals just enough so that, if possible, the terminal in the worst situation can reach its minimum acceptable SIR seems reasonable, provided that the new SIR for each of the terminals is still acceptable. On the other hand, one could argue that the satisfied terminals should not be sacrificed for no fault of their own, to help poorly situated terminals. These terminals could possibly wait for a better channel condition (due to their movement, for example), without disrupting the satisfaction of better situated terminals.

If a decision is made to turn off maxed out terminals whose SIR (in the final round) are "too low", this must be done sequentially, starting with the terminal in the worst situation. Once this terminal is powered off, the power vector needs to be recalculated, because, with less interference, each terminal needs less received power to achieve a given SIR; thus, the terminal in the second worst situation may now be able to reach its desired SIR, or at least an acceptable SIR, if it could not do so before. If this terminal's SIR is still unacceptable, then it should be turned off also, and the power vector recalculated once again. And so on.

Sacrificing the better situated terminals to help the poorly situated ones is, essentially, a "Robin Hood" scheme (to steal from the rich to help the poor). The appropriateness and fairness of such scheme is more a philosophical issue than an engineering one.

Appendix D

The Capacity of CDMA systems with Multiple Antennas at the Receiver

As described by Hanly [9], macrodiversity is a scheme in which the cellular structure of a wireless communication network is removed and “each mobile...(is)...jointly decoded by all receivers in the network”. Alternatively, one can think of a single-cell network equipped with several receiving antennas, possibly distributed in various locations throughout the cell. Hanly [9] shows that this scheme can significantly increase the capacity of a CDMA wireless communication network.

The macrodiversity capacity results provided by [9] assume that the transmission power of each transmitter contributes to its own interference. This approximation is generally appropriate for a CDMA system in which each transmitter’s spreading gain is “large”, which, normally means that its (pre-spread) “carrier to interference ratio” is “small”.

But modern wireless networks are expected to accommodate simultaneous transceivers operating at a wide range of data rates. “Variable spreading gain" (VSG) CDMA is one of the technologies through which new standards accommodate such multi-rate traffic (see for instance, Nanda, et al.[24]). In a VSG-CDMA system (see I and Sabnani[11]), each transceiver’s spreading gain is determined as the ratio of the common chip rate to the transceiver’s data rate. Thus, high data rate sources generally operate with “low” spreading gains, and “high” carrier-to-interference ratios. Under these conditions, the “self-interference” approximation may not be appropriate.

Explicitly considering transmission power limits, and without recurring to the “self-interference” approximation, this note derives results determining the capacity region of a CDMA cellular network under macrodiversity. The “complexity” of applying the new results is comparable to that of the approximated ones. The analysis is grounded on the Brouwer’s fixed point theorem and the Banach’s contraction mapping principle, two well established mathematical results.

Below, the basic macrodiversity relation is presented, first in the traditional form, and subsequently in matrix form, in terms of convenient new variables. Then, it is shown that the basic macrodiversity capacity question is equivalent to determining whether certain meaningful function has a fixed point. Subsequently, conditions are identified under which the desired solution ex-

ists. Moreover, further conditions are explored under which this solution is unique, and can be determined through an intuitive, well-behaved algorithm. Finally, the results are interpreted and discussed. Space limitations preclude a comprehensive comparison between the new results and those previously available. Nevertheless, some brief contrasting comments are made, highlighting the fact that the new results are less conservative, which can make a significant difference in the throughput of a 3G system.

A mathematical appendix introduces the essential mathematical terminology, and some technical results.

D.1 The macro-diversity framework

D.1.1 Basic relation

Under macro-diversity, the cellular structure is removed and each transmitter is jointly decoded by all "receivers" (base stations, or antennas in a single cell). Hanly [9] argues that, in this situation, a relevant QoS index for terminal i is the product of its spreading gain by α_i, defined as:

$$\alpha_i = \frac{P_i h_{i1}}{\sum_{\substack{j=1 \\ j\neq i}}^{N} P_j h_{j1} + \sigma_1^2} + \cdots + \frac{P_i h_{iK}}{\sum_{\substack{j=1 \\ j\neq i}}^{N} P_j h_{jK} + \sigma_K^2} \tag{D.1}$$

K is the number of "receivers" in the network, and h_{ik} is the "path loss" coefficient in the signal from terminal i when received at k. α_i can be thought of as a desired "carrier to interference ratio" (CIR).

D.1.2 The Capacity question

Conditions are sought under which a given N-vector of positive numbers, $\vec{\alpha} := \left[\begin{array}{ccc} \alpha_1 & \cdots & \alpha_N \end{array} \right]^{\mathrm{t}}$, is such that there exists another N-vector of positive numbers, $\left[\begin{array}{ccc} P_1 & \cdots & P_N \end{array} \right]^{\mathrm{t}}$, satisfying appropriate constraints, and equation (D.1) for each i. If this is the case, the system of N equations like (D.1) has a feasible solution, and the vector of power ratios $\vec{\alpha}$ is said to be in the "capacity region" of the system.

D.1.3 Normalizations and re-formulations

Noise normalization. Let all powers be divided by $\sigma_1^2 + \cdots + \sigma_K^2$. Also, let $\nu_k = \sigma_k^2/(\sigma_1^2 + \cdots + \sigma_K^2)$. Although this normalization introduces no notational change on the power vector, it is understood that henceforth all powers are expressed as multiple of the total noise power $\sigma_1^2 + \cdots + \sigma_K^2$.

Total received power from a given transmitter. Let

$$Q_i := P_i \sum_{k=1}^{K} h_{ik} \tag{D.2}$$

Scaled power. Let $q_i := Q_i/\alpha_i$ (The total received power from terminal i is "scaled" by that terminal's desired CIR α_i).

Relatives losses. Let

$$g_{ik} := \frac{h_{ik}}{\sum_{j=1}^{K} h_{ij}} \tag{D.3}$$

The power at receiver k coming from transmitter i, $P_i h_{ik} = g_{ik}\alpha_i q_i$.

Now, the basic macro-diversity equation can be restated as:

$$\frac{q_i g_{i1}}{\sum_{\substack{j=1 \\ j\neq i}}^{N} \alpha_j q_j g_{j1} + \nu_1} + \cdots + \frac{q_i g_{iK}}{\sum_{\substack{j=1 \\ j\neq i}}^{N} \alpha_j q_j g_{jK} + \nu_K} = 1 \tag{D.4}$$

Notice that $P_i = \alpha_i q_i / \sum_{k=1}^{K} h_{ik}$, which is measured as a multiple of the total noise power $\sigma_1^2 + \cdots + \sigma_K^2$.

D.1.4 Macrodiversity matrix relations

Let

$$Y_{ik}(\vec{q}) := \sum_{\substack{j=1 \\ j\neq i}}^{N} \alpha_j q_j g_{jk} + \nu_k \tag{D.5}$$

$Y_{ik}(\vec{q})$ can be written as the scalar product of vectors as:

$$\begin{bmatrix} g_{1k} & \cdots & g_{i-1,k} & 0 & g_{i+1,k} & \cdots & g_{Nk} \end{bmatrix} \cdot D\vec{q} + \nu_k$$

with

$$D := \begin{bmatrix} \alpha_1 & 0 & 0 & 0 \\ 0 & \alpha_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \alpha_N \end{bmatrix} \tag{D.6}$$

so that,

$$D\vec{q} = \begin{bmatrix} \alpha_1 q_1 \\ \alpha_2 q_2 \\ \vdots \\ \alpha_N q_N \end{bmatrix}$$

It will prove useful to recognize the vectors $\vec{Y}_i(\vec{q}) := \begin{bmatrix} Y_{i1} & \cdots & Y_{iK} \end{bmatrix}^{\mathrm{t}}$.

By "stacking" these interference vectors, one arrives at a "macro-vector" of length NK satisfying:

$$\vec{Y} := \begin{bmatrix} \vec{Y}_1 \\ \vec{Y}_2 \\ \vdots \\ \vec{Y}_{N-1} \\ \vec{Y}_N \end{bmatrix} = \mathcal{G}D \begin{bmatrix} q_1 \\ q_2 \\ \vdots \\ q_{N-1} \\ q_N \end{bmatrix} + \begin{bmatrix} \vec{\nu} \\ \vec{\nu} \\ \vdots \\ \vec{\nu} \\ \vec{\nu} \end{bmatrix} \tag{D.7}$$

where $\mathcal{G}$ is a matrix defined as

$$\begin{bmatrix} \vec{0} & \vec{g}_2 & \cdots & \vec{g}_{N-1} & \vec{g}_N \\ \vec{g}_1 & \vec{0} & \cdots & \vec{g}_{N-1} & \vec{g}_N \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \vec{g}_1 & \vec{g}_2 & \cdots & \vec{0} & \vec{g}_N \\ \vec{g}_1 & \vec{g}_2 & \cdots & \vec{g}_{N-1} & \vec{0} \end{bmatrix} \equiv \begin{bmatrix} \mathcal{G}^1 \\ \mathcal{G}^2 \\ \vdots \\ \mathcal{G}^{N-1} \\ \mathcal{G}^N \end{bmatrix} \tag{D.8}$$

with $\vec{0}$ the zero vector of appropriate length, and

$$\vec{g}_i := \begin{bmatrix} g_{i1} \\ \vdots \\ g_{iK} \end{bmatrix} \quad \vec{\nu} := \begin{bmatrix} \nu_1 \\ \vdots \\ \nu_K \end{bmatrix} \quad \vec{\vec{\nu}} := \begin{bmatrix} \vec{\nu} \\ \vdots \\ \vec{\nu} \end{bmatrix} \tag{D.9}$$

The matrix $\mathcal{G}D$ is some times denoted as $\hat{\mathcal{G}}$. $\mathcal{G}^{ik}$ (respect. $\hat{\mathcal{G}}^{ik}$) may denote the specific *row* of $\mathcal{G}$ (respect. $\hat{\mathcal{G}}$) "matching" Y_{ik}, with $\mathcal{G}^i$ (respect. $\hat{\mathcal{G}}^i$) the corresponding sub-matrix. Thus,

$$Y_{ik} = \mathcal{G}^{ik} \cdot D \cdot \vec{q} + \nu_k \equiv \hat{\mathcal{G}}^{ik} \cdot \vec{q} + \nu_k \tag{D.10}$$

The preceding notational transformations can be clarified by considering the specific case in which there are $N = 3$ transmitters and $K = 2$ receivers. In this case:

$$\vec{q} = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} \; ; \; D = \begin{bmatrix} \alpha_1 & 0 & 0 \\ 0 & \alpha_2 & 0 \\ 0 & 0 & \alpha_3 \end{bmatrix}$$

$$\begin{aligned} \vec{Y}_1 \equiv \begin{bmatrix} Y_{11} \\ Y_{12} \end{bmatrix} &= \begin{bmatrix} 0 & g_{21} & g_{31} \\ 0 & g_{22} & g_{32} \end{bmatrix} \begin{bmatrix} \alpha_1 q_1 \\ \alpha_2 q_2 \\ \alpha_3 q_3 \end{bmatrix} + \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix} \\ &\equiv \begin{bmatrix} \vec{0} & \vec{g}_2 & \vec{g}_3 \end{bmatrix} D\vec{q} + \vec{\nu} \end{aligned}$$

$$\vec{Y}_2 \equiv \begin{bmatrix} Y_{21} \\ Y_{22} \end{bmatrix} = \begin{bmatrix} g_{11} & 0 & g_{31} \\ g_{12} & 0 & g_{32} \end{bmatrix} \begin{bmatrix} \alpha_1 q_1 \\ \alpha_2 q_2 \\ \alpha_3 q_3 \end{bmatrix} + \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix}$$
$$\equiv \begin{bmatrix} \vec{g}_1 & \vec{0} & \vec{g}_3 \end{bmatrix} D\vec{q} + \vec{\nu}$$

$$\vec{Y}_3 \equiv \begin{bmatrix} Y_{31} \\ Y_{32} \end{bmatrix} = \begin{bmatrix} g_{11} & g_{21} & 0 \\ g_{12} & g_{22} & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 q_1 \\ \alpha_2 q_2 \\ \alpha_3 q_3 \end{bmatrix} + \begin{bmatrix} \nu_1 \\ \nu_2 \end{bmatrix}$$
$$\equiv \begin{bmatrix} \vec{g}_1 & \vec{g}_2 & \vec{0} \end{bmatrix} D\vec{q} + \vec{\nu}$$

$$\vec{Y} \equiv \begin{bmatrix} \vec{Y}_1 \\ \vec{Y}_2 \\ \vec{Y}_3 \end{bmatrix} \equiv \begin{bmatrix} Y_{11} \\ Y_{12} \\ Y_{21} \\ Y_{22} \\ Y_{31} \\ Y_{32} \end{bmatrix} =$$
$$\begin{bmatrix} 0 & g_{21} & g_{31} \\ 0 & g_{22} & g_{32} \\ g_{11} & 0 & g_{31} \\ g_{12} & 0 & g_{32} \\ g_{11} & g_{21} & 0 \\ g_{12} & g_{22} & 0 \end{bmatrix} \begin{bmatrix} \alpha_1 q_1 \\ \alpha_2 q_2 \\ \alpha_3 q_3 \end{bmatrix} + \begin{bmatrix} \nu_1 \\ \nu_2 \\ \nu_1 \\ \nu_2 \\ \nu_1 \\ \nu_2 \end{bmatrix} \equiv$$
$$\begin{bmatrix} \vec{0} & \vec{g}_2 & \vec{g}_3 \\ \vec{g}_1 & \vec{0} & \vec{g}_3 \\ \vec{g}_1 & \vec{g}_2 & \vec{0} \end{bmatrix} D\vec{q} + \begin{bmatrix} \vec{\nu} \\ \vec{\nu} \\ \vec{\nu} \end{bmatrix}$$

D.2 A fixed-point problem

Equation (D.4) can now be re-written as:

$$\frac{q_i g_{i1}}{Y_{i1}} + \cdots + \frac{q_i g_{iK}}{Y_{iK}} = 1 \tag{D.11}$$

For a *fixed* interference vector $\vec{Y}$ this equation can be easily solved for q_i, to obtain the vector $\vec{q}$ which would satisfy the system of equations of the form (D.11). This suggests the following approach. For a given $\vec{Y}$, define the transformation:

$$\vec{T}(\vec{q}) := \begin{bmatrix} \left(\frac{g_{11}}{Y_{11}(\vec{q})} + \cdots + \frac{g_{1K}}{Y_{1K}(\vec{q})}\right)^{-1} \\ \vdots \\ \left(\frac{g_{N,1}}{Y_{N,1}(\vec{q})} + \cdots + \frac{g_{N,K}}{Y_{N,K}(\vec{q})}\right)^{-1} \end{bmatrix} \tag{D.12}$$

$\vec{T}(\vec{q})$ yields the power vector under which each transceiver would achieve its desired α_i if the interference vector $\vec{Y}$ remained fixed. Of course, as the power levels are adjusted, a new interference vector results, $\vec{Y}(\vec{q}) = \mathcal{G}D\vec{q} + \vec{\vec{v}}$. This new vector will lead to further power adjustments, and so on, in an iterative manner.

Under the appropriate conditions, this algorithm will "converge" in the sense that a vector $\vec{q}^*$ exists such that $\vec{q}^* = T(\vec{q}^*)$; i.e., $\vec{q}^*$ is a "fixed point" of the mapping $\vec{T}$. These conditions determine the feasibility of the ratios α_i.

D.3 Mathematical results

Several well-known results useful in solving fixed point problems are presented below. Some relevant background material is discussed in a mathematical appendix.

D.3.1 Background material

Let S denote a vector space (for a formal definition of these spaces see [17, pp. 11-12]).

Norms and metrics. A norm, $\|\cdot\|$, on S is a function from S into the non-negative real numbers $\Re_+$ "generalizing" the idea of the "Euclidean length" of a vector. It engenders a "metric" ('distance'), defined as $d(x,y) = \|x-y\|$.

Infinity norm. $\|\cdot\|_\infty$ is defined as

$$\|\vec{x}\|_\infty := \max(|x_1|, |x_2|, \cdots, |x_N|) \tag{D.13}$$

Linear operators. If T is a mapping from a vector space, S_1, into another, S_2, (i.e., $T : S_1 \rightarrow S_2$), it is said to be *linear* if for any $x, y \in S_1$ and $\lambda_1, \lambda_2 \in \Re$, $T(\lambda_1 x + \lambda_2 y) = \lambda_1 T(x) + \lambda_2 T(y)$.

The **operator norm** of a linear operator T is defined as

$$\|T\| := \sup_{\|x\| \neq 0} \frac{\|T(x)\|}{\|x\|} \equiv \sup_{\|x\|=1} \|T(x)\| \tag{D.14}$$

where sup denotes the supremum or least upper bound.

Matrix infinity norm. When a linear operator is expressed as $T(x) = Ax$, with A a suitably dimensioned matrix, and the underlying norm is $\|\cdot\|_\infty$, its "operator norm" is the "maximum absolute

row sum" of A. If a_{ij} denotes the element corresponding to the i^{th} row and j^{th} column of matrix A,

$$\|A\|_\infty := \sup_{\|x\|=1} |Ax| = \max_i \left(\sum_j |a_{ij}| \right) \tag{D.15}$$

Row sum of the product of two non-negative matrices. If A and B are suitably dimensioned non-negative matrices, the row sum of the product, AB, can be obtained as the product $A \cdot \text{rsum}(B)$, with $\text{rsum}(B)$ the vector resulting from the sum of the columns of B.

D.3.2 Brouwer's Fixed Point Theorem

Theorem(Brouwer's): Let $T : S \to S$ be a continuous function from a non-empty, compact, convex set $S \subset \Re^n$ *into* itself. There is a $x_0 \in S$ such that $x_0 = T(x_0)$.

Proof: See [3, p.28].

D.3.3 Banach's result

Contraction Mappings. Let S be a vector space endowed with the norm $\|\cdot\|$. Suppose T is a mapping from S into itself (i.e., $T : S \to S$). If there is a real number λ, $0 \leq \lambda < 1$ such that $\|T(x) - T(y)\| \leq \lambda \|x - y\|$for all $x, y \in S$ then T is said to be a *contraction mapping*.

Successive approximation. For expositional convenience, let $T^m(x)$ for $x \in S$ be defined inductively by $T^0(x) = x$ and $T^{m+1}(x) = T\left(T^m(x)\right)$, with $m \in \{1, 2, \cdots\}$.

Banach's Contraction Mapping Principle: Let S be a closed subset of $\Re^n$. Suppose that T is a mapping from S into itself. If T is a contraction mapping on S, there is a unique vector $x_0 \in S$ such that $x_0 = T(x_0)$. Moreover, x_0 can be obtained by "successive approximation", starting from an arbitrary initial $x \in S$; i.e., for all $x \in S$, $\lim_{m\to\infty} T^m(x) = x_0$.

Furthermore,

$$\|T^m(x) - x_0\| \leq \frac{\lambda^m}{1-\lambda} \|T(x) - x\|$$

Proof: See [13, Theorem 3.1, page 41].

More general versions of this result, and many extensions can be found in many sources, including [13].

Contraction condition for differentiable mappings. If the considered vector space S is *convex* and the considered mapping is such that its derivative $T'(x)$ exists over S, then for any $x_1, x_2 \in S$, and $L := \{x = x_1 + t(x_2 - x_1) : 0 \leq t \leq 1\}$ the mean value inequality holds that

$$\|T(x_1) - T(x_2)\| \leq \sup_{x \in L} \left\|T'(x)\right\| \|x_1 - x_2\| \tag{D.16}$$

Hence, in this situation $\|T'(x)\| \leq \lambda < 1$ implies that T is a contraction mapping on S [17, p. 272].

D.4 Fixed points, and algorithms

D.4.1 From S *into* S

In order for the previously-mentioned results to be applicable to the mapping $\vec{T}(\vec{q})$, it must map vectors from an appropriate set, to vectors *in the same* set.

D.4.1.1 Set of scaled power vectors

In general, any feasible scaled power vector $\vec{q}$ must be in the set $S := \left\{ \vec{q} \in \Re_+^{\mathrm{N}}, \vec{0} \leq \vec{q} \leq \vec{q}^L \right\}$ with $\vec{q}^L$ the "largest" feasible total received (scaled) power vector. If P_i^L is the transmission power limit of transceiver i, $q_i^L = (1/\alpha_i) P_i^L \sum_{k=1}^K h_{ik}$.

This set is closed by definition. It is straightforward to verify that it is also convex.

D.4.1.2 "into" condition

It is immediate that each component $T_i(\vec{q})$ is increasing in each component of $\vec{Y}_i$. And each component of $\vec{Y}_i(\vec{q})$ is increasing in $\vec{q}$. Therefore, to verify that $\vec{T}(\vec{q})$ is in S, the critical value is $\vec{T}(\vec{q}^L)$. Specifically, it is necessary that $T_i(\vec{q}^L) \leq q_i^L$ or that, (see equation (D.12)),

$$\frac{g_{i1} q_i^L}{Y_{i1}(\vec{q}^L)} + \cdots + \frac{g_{iK} q_i^L}{Y_{iK}(\vec{q}^L)} \geq 1 \tag{D.17}$$

where, by equation (D.5), $Y_{ik}(\vec{q}^L) = \sum_{\substack{j=1 \\ j \neq i}}^N \alpha_j q_j^L g_{jk} + \nu_k$.

Recall that $P_i h_{ik} \equiv g_{ik} \alpha_i q_i$. Hence, the preceding condition can be written as:

$$\alpha_i \leq \frac{P_i^L h_{i1}}{\sum_{\substack{j=1 \\ j \neq i}}^N P_j^L h_{j1} + \nu_1} + \cdots + \frac{P_i^L h_{iK}}{\sum_{\substack{j=1 \\ j \neq i}}^N P_j^L h_{jK} + \nu_K} \tag{D.18}$$

It may be reasonable to assume that $P_i^L = P^L \forall i$, and that ν_k / P^L is "very small" as compared to $\sum_{\substack{j=1 \\ j \neq i}}^N h_{jk}$. Then, condition (D.18) becomes:

$$\alpha_i \leq \frac{h_{i1}}{\sum_{\substack{j=1 \\ j \neq i}}^N h_{j1}} + \cdots + \frac{h_{iK}}{\sum_{\substack{j=1 \\ j \neq i}}^N h_{jK}} \tag{D.19}$$

D.4.2 Existence of a fixed point

Proposition: If a vector of desired CIR, $\vec{\alpha}$, is such that condition (D.17) is satisfied – or so is the "neater" condition (D.19), under the mild assumptions under which it is valid – then $\vec{\alpha}$ is feasible.

Proof: The set S of feasible (scaled) power vectors is a closed, bounded and convex subset of $\Re^N$. If condition (D.17) or, when appropriate, (D.19) , is satisfied, the mapping $\vec{T}(\vec{q})$ is into. It is considered self-evident (and can be shown) that this mapping is continuous over the set S. Therefore,

Brouwer's fixed-point theorem applies (see section D.3.2). Hence, $\vec{T}(\vec{q})$ has at least one fixed point. Q.E.D.

However, Brouwer's theorem says nothing about the uniqueness of the solution, or the behavior of the algorithm discussed in section D.3.3.

D.5 Toward a unique fixed point

This section explores conditions under which the norm of the derivative of $\vec{T}(\vec{q})$ is less than one, so that Banach's principle can be applied. In this case, a unique fixed-point exists, and it can be found via a simple, well-behaved algorithm (see section D.3.3).

D.5.1 Derivative of $\vec{T}(\vec{q})$

$\vec{T}'(\vec{q})$ is given by the corresponding "Jacobian" matrix of partial derivatives, where $\partial T_i/\partial q_j$ corresponds to its i^{th} row and j^{th} column. From equation (D.12),

$$T_i(\vec{q}) = \left(\frac{g_{i1}}{Y_{i1}(\vec{q})} + \cdots + \frac{g_{iK}}{Y_{iK}(\vec{q})} \right)^{-1} \tag{D.20}$$

Thus,

$$\frac{\partial T_i}{\partial q_j} = \frac{\partial T_i}{\partial Y_{i1}} \frac{\partial Y_{i1}}{\partial q_j} + \frac{\partial T_i}{\partial Y_{i2}} \frac{\partial Y_{i2}}{\partial q_j} + \cdots + \frac{\partial T_i}{\partial Y_{iK}} \frac{\partial Y_{iK}}{\partial q_j} \tag{D.21}$$

$\partial T_i/\partial Y_{ik}$ is obtained as:

$$g_{ik} Y_{ik}^{-2} \left(\frac{g_{i1}}{Y_{i1}} + \frac{g_{i2}}{Y_{i2}} + \cdots + \frac{g_{iK}}{Y_{iK}} \right)^{-2} \equiv g_{ik} \left(\frac{T_i}{Y_{ik}} \right)^2 \tag{D.22}$$

Additionally, by equation (D.5), $Y_{ik}(\vec{q}) = \sum_{\substack{j=1 \\ j \neq i}}^{N} \alpha_j q_j g_{jk} + \nu_k$. Therefore,

$$\frac{\partial Y_{ik}}{\partial q_j} = \begin{cases} 0 \text{ for } j = i \\ \alpha_j g_{jk} \text{ for } j \neq i \end{cases} \tag{D.23}$$

Replacing equations (D.22) and (D.23) into equation (D.21) one obtains that

$$\partial T_i/\partial q_i \equiv 0 \, \forall i \tag{D.24}$$

and, for $j \neq i$, $\partial T_i/\partial q_j =$

$$\begin{aligned} T_i^2 \left(\frac{g_{i1}}{Y_{i1}^2} \frac{\partial Y_{i1}}{\partial q_j} + \frac{g_{i2}}{Y_{i2}^2} \frac{\partial Y_{i2}}{\partial q_j} + \cdots + \frac{g_{iK}}{Y_{iK}^2} \frac{\partial Y_{iK}}{\partial q_j} \right) &= \\ \alpha_j T_i^2 \left(\frac{g_{i1}}{Y_{i1}^2} g_{j1} + \frac{g_{i2}}{Y_{i2}^2} g_{j2} + \cdots + \frac{g_{iK}}{Y_{iK}^2} g_{jK} \right) &= \\ \alpha_j T_i^2 \sum_{k=1}^{K} \frac{g_{ik} g_{jk}}{Y_{ik}^2} & \end{aligned} \tag{D.25}$$

D.5.2 Norm of $T'(\vec{q})$

By definition, $\left\| \vec{T}'(\vec{q}) \right\|_\infty$ is the maximum absolute row sum of $\vec{T}'(\vec{q})$ (see section D.3.1). In view of equations (D.24) and (D.25), the i^{th} row of $\vec{T}'(\vec{q})$ adds up to

$$\begin{aligned} \sum_{j=1}^{N} \frac{\partial T_i}{\partial q_j} &= T_i^2(\vec{q}) \sum_{\substack{j=1 \\ j \neq i}}^{N} \alpha_j \sum_{k=1}^{K} \frac{g_{ik} g_{jk}}{Y_{ik}^2(\vec{q})} \\ &= T_i^2(\vec{q}) \sum_{k=1}^{K} \frac{g_{ik}}{Y_{ik}^2(\vec{q})} \sum_{\substack{j=1 \\ j \neq i}}^{N} \alpha_j g_{jk} \\ &:= f_{ik}(\vec{q}) \rho_{ik} \end{aligned} \tag{D.26}$$

Observe that $\rho_{ik} := \sum_{j=1}^{N} \alpha_j g_{jk} - \alpha_i g_{ik}$ is the sum of the components of $\hat{\mathcal{G}}^{ik}$, which is the row of the matrix $\mathcal{G}D \equiv \hat{\mathcal{G}}$ associated with Y_{ik} (see equation (D.10)). It represents the parameters in equation (D.26) which can be influenced by limiting the vector $\vec{\alpha}$. For a given $\vec{q}$, the function $f_{ik}(\vec{q}) := T_i^2(\vec{q}) \sum_{k=1}^{K} g_{ik}/Y_{ik}^2(\vec{q})$ is determined by the channel via the various path loss coefficients.

D.5.3 Contraction condition

On the basis of the preceding development, in order for $\left\| \vec{T}'(\vec{q}) \right\|_\infty < 1$ so that $\vec{T}(\vec{q})$ is a contraction, $\vec{\alpha}$ must be such that

$$\max_{\vec{q}} \; f_{ik}(\vec{q}) \rho_{ik} < 1 \;\; \forall i,k \tag{D.27}$$

with ρ_{ik} the sum of the components of $\hat{\mathcal{G}}^{ik}$(see equation (D.10)) and $f_{ik}(\vec{q})$ given by:

$$f_{ik}(\vec{q}) = \frac{\frac{g_{i1}}{Y_{i1}^2} + \cdots + \frac{g_{iK}}{Y_{iK}^2}}{\left(\frac{g_{i1}}{Y_{i1}} + \cdots + \frac{g_{iK}}{Y_{iK}} \right)^2} \tag{D.28}$$

D.5.4 Properties of the Contraction Condition

1. **Well-definedness**. Condition (D.27) is well defined, because f_{ik} is a continuous function, for which it must have a maximum over a closed and bounded set (see sec. (D.4.1.1))

2. $f_{ik} \geq 1$. This is so because f_{ik} is of the form $(\lambda_1\phi(x_1)+\cdots+\lambda_K\phi(x_K))/\phi(\lambda_1 x_1+\cdots+\lambda_K x_K)$ with $\phi(x)=x^2$, $\lambda_i \in [0,1]$, $\sum_i \lambda_i = 1$ and x_i positive. The function $\phi(x)=x^2$ is easily shown to by convex. And for any convex ϕ, Jensen's inequality holds that $\lambda_1\phi(x_1)+\cdots+\lambda_K\phi(x_K) \geq \phi(\lambda_1 x_1+\cdots+\lambda_K x_K)$. (See also section D.5.6).

3. If $\vec{q}$ is such that $Y_{ik}(\vec{q}) = Y_{il}(\vec{q})\ \forall k,l$ then $f_{ik}(\vec{q}) = 1$. This follows directly because $\sum_k g_{ik} = 1$ by definition (see equation (D.3))

4. If each transceiver is "equidistant" to each "receiver" (antenna), in the sense that $h_{ik} = h_{il}\ \forall i,k,l$ then $f_{ik}(\vec{q}) \equiv 1$. This also follows directly because in this case $g_{ik} = g_{il} \equiv 1/K\ \forall i,k,l$ (see equation (D.3)). In this case, the contraction condition (D.27) reduces to $\|\mathcal{G}D\|_\infty \equiv \|\mathcal{G}\vec{\alpha}\|_\infty < 1$

5. In the special case in which K=2, the maximum f_{ik} is attained for the particular $\vec{q}$ which creates the largest "separation" between Y_{i1} and Y_{i2}. (See section D.5.6).

D.5.5 A unique solution and an algorithm to find it

Proposition: If a vector of desired CIR, $\vec{\alpha}$, is such that condition (D.17) , or, when appropriate, condition (D.19), is satisfied, and so is condition (D.27) above, then $\vec{\alpha}$ is feasible. Furthermore, the power vector leading to $\vec{\alpha}$ is unique, and can be obtained via the well-behaved algorithm described in section D.3.3.

Proof: The "power set" S is a closed subset of $\Re^n$ (see section D.4.1.1). If condition (D.17) , or, when appropriate, condition (D.19), is satisfied, the transformation $T(\vec{q})$ is a mapping from S into S. If condition (D.27) is also satisfied, T is a contraction mapping. Therefore, under the hypothesis of this proposition, Banach's principle applies (see section D.3.3). Q.E.D.

D.5.6 Maximum of an interesting ratio

It is of interest to determine a supremum of the form

$$\sup_{0\leq x_0\leq x_1, x_2\leq x_3} \frac{\lambda x_1^2 + (1-\lambda)x_2^2}{(\lambda x_1 + (1-\lambda)x_2)^2} \tag{D.29}$$

where $0 \leq \lambda \leq 1$ is fixed, and $x_1 \leq x_2$ are positive real numbers in certain interval.

The above ratio is a continuous function for which it must necessarily have a maximum over any closed and bounded set.

Also, $x_{12} := \lambda x_1 + (1-\lambda)x_2$ is simply a convex combination ("mixture") of x_1 and x_2; i.e., a point between x_1 and x_2. Likewise, $\lambda x_1^2 + (1-\lambda)x_2^2$ is a "mixture" of x_1^2 and x_2^2, with the same "mixture" parameter λ (see figure (D.1)).

The function $f(x) := x^2$ is easily shown to be convex. And, by definition, any convex function satisfies $\lambda f(x_1) + (1-\lambda) f(x_2) \geq f(\lambda x_1 + (1-\lambda)x_2)$. Therefore, the ratio (D.29) is always greater than or equal to 1.

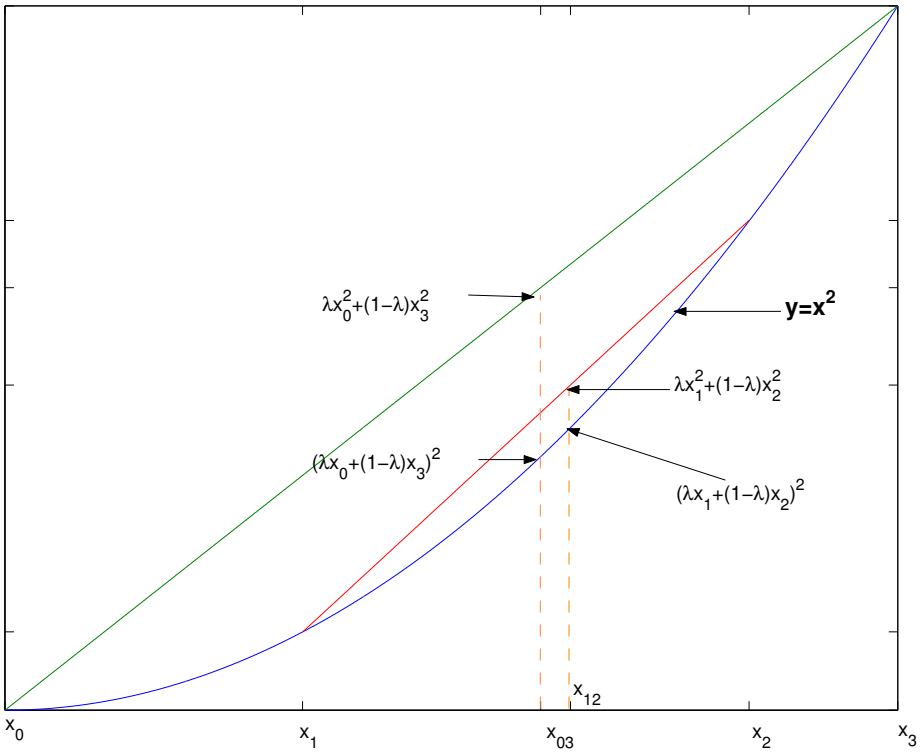

Figure D.1: Maximizing an interesting ratio: $\lambda x_i^2 + (1-\lambda)x_j^2$ versus $(\lambda x_i + (1-\lambda)x_j)^2$

It is straightforward to verify that the first-order optimizing conditions for this ratio are satisfied whenever $x_1 = x_2$. But in this case, the ratio equals 1, which is its smallest possible value. Therefore, the maximum is attained over the boundary of the feasible region; i.e., $x_1 = x_0$ and $x_2 = x_3$ leads to the maximum.

D.6 Discussion

This note provides an answer to the question of whether a certain vector, $\vec{\alpha}$, of positive numbers interpreted as desired "carrier-to-interference ratios" is feasible in a macrodiversity CDMA environment, in the sense that there are feasible power levels which produce the desired ratios. The answer is in the affirmative whenever condition (D.17) is satisfied. Under mild assumptions, this condition takes the simple form $\alpha_i \leq A_i$, with A_i a relatively simple function involving ratios of the various path loss coefficients of the active transceivers. However, not much can be said about the underlying power vector, or the performance of any particular algorithm in finding it.

This note also explores a more elaborate condition, (D.27). *Together* with condition (D.17), condition (D.27) implies that the power vector leading to $\vec{\alpha}$ is unique, and can be found by way of a well-behaved simple algorithm. This algorithm can depart from an arbitrary power vector. It is of the form $x^{n+1} = f(x^n)$ with x^0 arbitrary. A simple expression gives the "error" after a given number of iterations.

In general, condition (D.27) depends on the maximum of a relatively simple function. More

research is needed to determine the practical implications of obtaining this maximum, or a reasonable approximation for it. However, in special cases, in particular when each terminal happens to be "equidistant" from the antennas, this condition reduces to $\|\mathcal{G}\vec{\alpha}\|_\infty < 1$. In words, this condition requires that the "largest weighted average" of the desired α_i's be less than one. The possible weight vectors are the rows of the "relative gains" matrix $\mathcal{G}$. It is significant that each row of this matrix always has at least one element equal to zero, which implies that, in verifying this condition, at most N-1 of the α_i's are simultaneously weighted.

Space limitations preclude a comprehensive contrasting of these results to those originally presented in [9]. Nevertheless, some brief comments will be made.

First, condition (D.17) does not have an obvious "counterpart" in [9]. The result derived in [9] under the "self-interference" approximation is $\sum_{i=1}^{N}\alpha_i < K$, which limits the sum without imposing an individual limit on each term. However, one can make a rough comparison by assuming that condition (D.19) applies and is satisfied by each i, and that each terminal is "equidistant" from each antenna so that $h_{ik} \approx h_{il} \approx h_i \; \forall i,k,l$. This symmetry would practically arise, for example, with $K = 2$, if the two receiving antennas are directly across from each other in opposing sides of a road segment, and each terminal is located along the axis of this segment. When this symmetry exists,

$$\sum_{i=1}^{N}\alpha_i \leq \sum_{i=1}^{N} K\frac{h_i}{\sum_{\substack{j=1\\ j\neq i}}^{N} h_j} \approx K\frac{\sum_{i=1}^{N} h_i}{\sum_{\substack{j=1\\ j\neq i}}^{N} h_j} > K$$

This indicates that condition (D.19) is "less conservative" that the approximated condition from [9].

The more elaborate condition, (D.27), may also be compared, with caution, with the approximated result from [9], by considering, again, the special symmetric situation. In this case, condition (D.27) reduces to $\|\mathcal{G}\vec{\alpha}\|_\infty < 1$, as remarked above. Additionally, each $g_{ik} = 1/K$ (see equation (D.3)). Therefore, the j^{th} row of this matrix has the form $(1/K)\begin{bmatrix} 1 & \cdots & 1 & 0 & 1 & \cdots & 1 \end{bmatrix}$ where the only zero is at the j^{th} position (see equation (D.8)). Hence, the product of the j^{th} row of $\mathcal{G}$ by $\vec{\alpha}$ simply adds all the components of $\vec{\alpha}$ except for α_j and divide the sum by K. For example, with 3 terminals, the second row of $\mathcal{G}$ is $(1/K)\begin{bmatrix} 1 & 0 & 1 \end{bmatrix}$ and the product of this row by $\vec{\alpha}$ equals $(\alpha_1+\alpha_3)/K$. $\|\mathcal{G}\vec{\alpha}\|_\infty$ simply picks out the largest component of the product $\mathcal{G}\vec{\alpha}$. The j^{th} component of $\mathcal{G}\vec{\alpha}$ is a sum of the form $(\sum_{i=1}^{N}\alpha_i - \alpha_j)/K$. Thus, the largest component of $\mathcal{G}\vec{\alpha}$ will be the one that leaves out of the sum the smallest component of $\vec{\alpha}$. For instance, if α_N happens to be the smallest α_i, then $\|\mathcal{G}\vec{\alpha}\|_\infty = (1/K)\sum_{i=1}^{N-1}\alpha_i$. Hence, in the "symmetric" case, the approximated result demands that $\sum_{i=1}^{N}\alpha_i < K$, whereas condition (D.27) only imposes that $\sum_{i=1}^{N-1}\alpha_i < K$ (assuming α_N is the smallest desired CIR).

It is stressed that, in the context of a 3G network, when relatively few high data-rate terminals may be sharing a channel, the less conservative results could make a significant difference. For example, suppose K=1, and that three high data-rate sources wish to share a channel, each demanding a CIR of 2/5. This is plausible in a VSG-CDMA situation (see introduction). The approximated result dictates that only 2 of them can be accommodated, whereas condition (D.27) indicates that all

three can, “with room to spare”. In a 3G environment, leaving, unnecessarily, out even one terminal could be significant, if, as presumed, the additional terminal would have transmitted megabits of data each second.

Literature Cited

[1] S. ATAMAN AND A. WAUTIER, *Perfect power control in a multiservice CDMA cellular system*, IEEE PIMRC, 3 (2002), pp. 1222 –6.

[2] L. BERKOVITZ, *Convexity and optimization in R^n*, J. Wiley, New York, 2002.

[3] K. BORDER, *Fixed Point Theorems with Applications to Economics and Game Theory*, Cambridge Univ. Press, Cambridge, UK, 1985.

[4] R. DANSEREAU AND W. KINSNER, *Psychovisual correlations with multifractal measures for wavelet and wavelet packet progressive image transmission*, Canadian Conf. on Elec. Comp. Eng., 1 (2000), pp. 435 –9.

[5] H. FRENK AND S. SCHAIBLE, *Fractional programming: Introduction and applications*, in Encyclopedia of Optimization, C. Floudas and P. Pardalos, eds., Kluwer Academic Publishers, Dordrecht-Boston-London, 2002.

[6] D. GOODMAN AND N. MANDAYAM, *Power control for wireless data*, IEEE Personal Communications, 7 (2000), pp. 48 –54.

[7] S. GRANDHI, R. VIJAYAN, D. J. GOODMAN, AND J. ZANDER, *Centralized power control in cellular radio systems*, Vehicular Technology, IEEE Transactions on, 42 (1993), pp. 466 –8.

[8] P. GRAY, *Psychology*, Worth Pub., New York, 2002.

[9] S. HANLY, *Capacity and power control in spread spectrum macrodiversity radio networks*, Communications, IEEE Transactions on, 44 (1996), pp. 247–256.

[10] S. HANLY AND D. TSE, *Power control and capacity of spread spectrum wireless networks*, Automatica, 35 (1999), pp. 1987–2012.

[11] C.-L. I AND K. SABNANI, *Variable spreading gain CDMA with adaptive control for true packet switching wireless network*, IEEE ICC, 2 (1995), pp. 725 –730.

[12] H. JI, *Resource Management in Communication Networks via Economic Models*, PhD thesis, Rutgers Univ., New Jersey, 1997.

[13] M. KHAMSI AND W. KIRK, *An Introduction to Metric Spaces and Fixed Point Theory*, Wiley, New York, 2001.

[14] M. KHANSARI AND M. VETTERLI, *Layered transmission of signals over power-constrained wireless channels*, IEEE ICIP, 3 (1995), pp. 380–3.

[15] A. KWASINSKI AND N. FARVARDIN, *Resource allocation for cdma networks based on real-time source rate adaptation*, IEEE ICC, 5 (2003), pp. 3307 – 3311.

[16] J. LEE, R. MAZUMDAR, AND N. SHROFF, *Joint power and data rate allocation for the downlink in multi-class CDMA wireless networks*, Proc. of the 40th Allerton Conference on Communications, Control and Computing, (2002).

[17] D. LUENBERGER, *Optimization by Vector Space Methods*, Wiley, New York, 1969.

[18] C. LUNA AND A. KATSAGGELOS, *Maximizing user utility in video streaming applications*, IEEE ICASSP, 3 (2001), pp. 1465 –1468.

[19] ——, *Maximizing user utility in video streaming applications*, IEEE Trans. on Circ. and Sys. for Video Tech., 13 (2003), pp. 141–8.

[20] A. MACKENZIE AND S. WICKER, *Game theory and the design of self-configuring, adaptive wireless networks*, IEEE Communications Magazine, 39 (2001), pp. 126 –131.

[21] A. MAS-COLELL, M. WHINSTON, AND J. GREEN, *Microeconomic Theory*, Oxford Univ. Press, New York, 1995.

[22] P. MEYER, J. YUNG, AND J. AUSUBEL, *A primer on logistic growth and substitution: The mathematics of the loglet lab software*, Technological Forecasting and Social Change, 61 (1999), pp. 247–71.

[23] T. MINN AND K.-Y. SIU, *Dynamic assignment of orthogonal variable-spreading-factor codes in w-cdma*, IEEE JSAC, 18 (2000), pp. 1429 – 1440.

[24] S. NANDA, K. BALACHANDRAN, AND S. KUMAR, *Adaptation techniques in wireless packet data services*, IEEE Communications Magazine, 38 (2000), pp. 54 –64.

[25] J. NELDER, *The fitting of a generalization of the logistic curve*, Biometrics, 17 (1961), pp. 89–110.

[26] T. OTTOSSON AND A. SVENSSON, *Multi-rate schemes in DS/CDMA systems*, IEEE Vehicular Technology Conference, 2 (1995), pp. 1006 –1010.

[27] J. PONSTEIN, *Seven kinds of convexity*, SIAM Review, 9 (1967), pp. 115–119.

[28] F. RICHARDS, *A flexible growth function for empirical use*, Journal of Experimental Botany, 10 (1959), pp. 290–300.

[29] V. RODRIGUEZ, *An analytical foundation for resource management in wireless communications*, IEEE Globecom, 2 (2003), pp. 898–902.

[30] ——, *Robust modeling and analysis for wireless data resource management*, IEEE WCNC, 2 (2003), pp. 717–722.

[31] V. RODRIGUEZ AND D. GOODMAN, *Improving a utility function for wireless data*, WICAT Tech. Rep. 02-009, Polytechnic Univ., Brooklyn, New York, 2002. http://wicat.poly.edu/reports.

[32] A. SAMPATH, P. KUMAR, AND J. HOLTZMAN, *Power control and resource management for a multimedia CDMA wireless system*, IEEE PIMRC, 1 (1995), pp. 21–5.

[33] C. SARAYDAR, N. MANDAYAM, AND D. GOODMAN, *Efficient power control via pricing in wireless data networks*, Communications, IEEE Transactions on, 50 (2002), pp. 291 –303.

[34] E. SENETA, *Non-Negative Matrices*, Wiley, New York, 1973.

[35] V. SHAH, N. MANDAYAM, AND D. GOODMAN, *Power control for wireless data based on utility and pricing*, IEEE PIMRC, 3 (1998), pp. 1427–32.

[36] J. SHAPIRO, *Embedded image coding using zerotrees coefficients of wavelet*, IEEE Trans. on Signal Proces., 41 (1993), pp. 3445 –62.

[37] C. W. SUNG AND W. S. WONG, *Power control and rate management for wireless multimedia CDMA systems*, IEEE Trans. Commun., 49 (2002), pp. 1215–26.

[38] Y.-C. TSENG AND C.-M. CHAO, *Code placement and replacement strategies for wideband cdma ovsf code tree management*, IEEE Trans. on Mobil Comp., 1 (2002), pp. 293 – 302.

[39] A. TSOULARIS, *Analysis of logistic growth models*, Research Letters in the Information and Mathematical Sciences, 2 (2001), pp. 23–46. Available at `http://www.massey.ac.nz/~wwiims/research/letters/volume2number1/`.

[40] S. ULUKUS AND L. GREENSTEIN, *Throughput maximization in CDMA uplinks using adaptive spreading and power control*, IEEE ISSSTA, 2 (2000), pp. 565 –9.

[41] B. USEVITCH, *A tutorial on modern lossy wavelet image compression: foundations of JPEG 2000*, IEEE Signal Proc. Mag., 18 (2001), pp. 22 –35.

[42] H. VARIAN, *Microeconomic Analysis*, W.W. Norton & Co., New York, 3rd ed., 1992.

[43] ——, *A solution to the problem of externalities when agents are well-informed*, The American Economic Review, 84 (1994), pp. 1278–93.

[44] P. VERHULST, *Notice sur la loi que la population suit dans son accroissement*, Correspondence Mathematique et Physique, 10 (1838), pp. 113–121.

[45] W. VICKERY, *Counterspeculation, auctions and competitive sealed tenders*, Journal of Finance, 16 (1961), pp. 8–37.

[46] Y. WANG, J. OSTERMANN, AND Y. ZHANG, *Video Processing and Communications*, Prentice Hall, Upper Saddle River, New Jersey, 2002.

[47] Q. ZHANG, W. ZHU, Z. JI, AND Y.-Q. ZHANG, *A power-optimized joint source and channel coding for scalable video streaming over wireless channels*, IEEE ISCAS, 5 (2001), pp. 137–140.

[48] M. ZORZI AND R. RAO, *Error control and energy consumption in communications for nomadic computing*, Computing, IEEE Transactions on, 46 (1997), pp. 279 –89.