

CONTENTS

1	Generic Multiuser Detection and Statistical Physics	1
1.1	Introduction	2
1.1.1	Generic Multiuser Detection	2
1.1.2	Single-user Characterization of Multiuser Systems	3
1.1.3	On the Replica Method	4
1.1.4	Statistical Inference Using Practical Algorithms	5
1.1.5	Statistical Physics and Related Problems	5
1.2	Generic Multiuser Detection	6
1.2.1	CDMA/MIMO Channel Model	6
1.2.2	Generic Posterior Mean Estimation	7
1.2.3	Specific Detectors as Posterior Mean Estimators	9
1.3	Main Results: Single-user Characterization	10
1.3.1	Is the Decision Statistic Gaussian?	10
1.3.2	The Decoupling Principle: Individually Optimal Detection	12
1.3.3	Decoupling Principle: Generic Multiuser Detection	17
1.3.4	Justification of Results: Sparse Spreading	19
1.3.5	Well-known Detectors as Special Cases	21
1.4	The Replica Analysis of Generic Multiuser Detection	23
1.4.1	The Replica Method	23
1.4.2	Free Energy	27

ii CONTENTS

1.4.3	Joint Moments	34
1.5	Further Discussion	36
1.5.1	On Replica symmetry	36
1.5.2	On Metastable Solutions	37
1.6	Statistical Physics and the Replica Method	39
1.6.1	A Note on Statistical Physics	39
1.6.2	Multiuser Communications and Statistical Physics	40
1.7	Interference Cancellation	42
1.7.1	Conventional Parallel Interference Cancellation	42
1.7.2	Belief Propagation	42
1.8	Concluding Remarks	46
	Acknowledgements	47
	References	48

CHAPTER 1

GENERIC MULTIUSER DETECTION AND STATISTICAL PHYSICS

DONGNING GUO (NORTHWESTERN UNIVERSITY)
AND
TOSHIYUKI TANAKA (KYOTO UNIVERSITY)

This chapter presents a tutorial of the general framework for the analysis and design of large multiuser systems using statistical physics techniques. For that purpose, multiuser signal detection is first cast as a general problem of Bayesian inference. In particular, generic suboptimal detection is regarded as “optimal” detection for some (probably mismatched) postulated system model. The large-system performance of a broad family of generic detectors is then obtained using the replica method, a powerful tool developed in statistical physics. The central result is a single-user characterization of the multiuser systems, which we call the *decoupling principle*: The virtual subchannel between the input symbol and the detection output for each individual user in the multiuser system is essentially equivalent to a single-user Gaussian channel, where the aggregate effect of the multiaccess interference from all other users is tantamount to a degradation in the signal-to-noise ratio. This degradation, known as the multiuser efficiency, can be found by solving a fixed-point equation. The error probability and spectral efficiency of the multiuser system are thus obtained. The decoupling principle applies to well-known linear detectors including the single-user matched filter, decorrelator, linear MMSE detector, linear interference cancelers, as well as nonlinear ones such as the jointly and individually optimal detectors. Relationships to practical detection schemes using belief propagation are also discussed. For simplicity, the analysis is limited to synchronous systems with additive Gaussian noise.

1.1 INTRODUCTION

Fuelled by the advent and rapid development of cellular telephony, the problem of multiuser signal detection (or separation) has received great attention since the mid-1980s as one of the major avenues towards optimal error performance and spectrum usage in wireless communications. This chapter introduces the concept of *generic multiuser detection* and summarizes some recent advances in the analysis and design of generic detectors using statistical physics techniques.

1.1.1 Generic Multiuser Detection

Consider a multidimensional communication system in which each user randomly generates a “signature vector” and modulates its own (usually error-control coded) symbols onto the signature for transmission. The received signal is the superposition of all users’ signals corrupted by Gaussian noise. With knowledge of all signature vectors, the goal of a multiuser receiver is to reliably recover the information intended for all or a subset of the users. The multiuser channel, best described by a vector model, is very versatile and is widely used in applications that include code-division multiple access (CDMA) as well as certain multiple-input multiple-output (MIMO) systems.

The maximum information rate through a multiuser channel is achieved by jointly optimal decoding, which is prohibitively complex for all but a small user population and codeword length. Hence the tasks of untangling the mutually interfering streams and exploiting the redundancy in the error-control codes are often separated. Oftentimes, the multiuser detector plays the role of a front end, which provides individual stream of (hard or soft) decision statistics to independent single-user decoders.

The simplest meaningful detector ignores the presence of multiaccess interference (MAI) and carries out single-user matched filtering (SUMF), whose error performance is generally very poor. The best probability of error of an uncoded system is achieved by solving a hypothesis testing problem with an exponential number of hypotheses in the number of interfering users [1, 2], which is in general NP-hard [3]. Optimal error performance in asynchronous channels is achieved by more involved sequence detection [4]. A large gap has been demonstrated between the probability of error of the naïve SUMF and that of individually optimal (IO) and jointly optimal (JO) detection. In order to explore the trade-offs between performance and computational complexity, numerous suboptimal detection schemes have been proposed, such as the decorrelator, linear minimum mean-square error (LMMSE) detector, and various interference cancelers.

All of the aforementioned detectors can be derived as some form of optimal detection with heuristic (but untrue) assumptions based on conventional wisdom and practical considerations. For example, the SUMF is optimal assuming the MAI to be Gaussian or absent; the decorrelator provides optimal detection assuming that there is no background noise; and the LMMSE detector maximizes the output signal-to-interference-and-noise ratio (SINR) assuming a Gaussian input. With the exception of decorrelating receivers, the multiuser detector outputs are still contaminated by MAI.

The viewpoint taken in this chapter is that, in general, every suboptimal detector can be regarded as computing an “optimally” detected output given some

“mismatched” system model. This perspective has its origin in the general theory of statistical inference or learning, where a “student” may adopt a probability model that is different from that of the “teacher” (see e.g., [5], for a discussion of such cases). In particular, the so-called *generic multiuser detector* computes the *posterior mean* of the transmitted symbols given the observation based on a postulated probability law of the system. More on this viewpoint will be discussed in Section 1.2.

1.1.2 Single-user Characterization of Multiuser Systems

Using techniques and methodologies originating in statistical physics, two fundamental questions about multiuser systems and generic detection are addressed in this chapter: 1) Given a multiuser detector, how to characterize the (single-user) subchannel between the input and output of each user? 2) Given a multiuser system, what are the achievable information rates by optimal joint decoding and suboptimal single-user decoding, respectively?

The preceding questions have been well studied for linear detection schemes. In fact, the analysis of multiuser communication systems is to a large extent the pursuit of a single-user characterization of the performance. A key performance measure, the *multiuser efficiency*, was introduced in [4, 6] to refer to the signal-to-noise ratio (SNR) degradation of the multiuser detection output relative to single-user performance. The multiuser efficiencies of the SUMF, decorrelator, LMMSE detector and linear interference cancelers at any given SNR were found as functions of the correlation matrix of the spreading sequences (i.e., the signature vectors), which can also be written explicitly in terms of the eigenvalues of the matrix.

The performance of finite-size multiuser systems is often not easy to evaluate, including when averaged over random sequences (e.g., [7–9]). An alternative paradigm for the analysis is to take the large-system limit instead, namely, to study the case where the number of users and the dimension of the channel both tend to infinity with a fixed ratio. A key consequence is that the dependence of performance measures on the spreading sequences vanishes as the system size increases. In the special case of linear detection, the output converges to a Gaussian statistic, which allows the performance to be solely quantified using the output SINR (e.g., [10–12]). It appears that the users are decoupled in such a way that each user experiences a single-user channel with SNR degradation in lieu of MAI.

A major spate of success of large-system analysis is achieved by using random matrix theory, the central dictate of which is that the empirical distributions of the eigenvalues of a random matrix converge to a deterministic distribution as its dimension increases [13, 14]. As a result, the multiuser efficiency of a sufficiently large system can be obtained as an integral with respect to the limiting eigenvalue distribution. Indeed, this random matrix technique is applicable to any performance measure that can be expressed as a function of the eigenvalues, e.g., the multiuser efficiency of the decorrelator [15–17] and the large-system capacity of CDMA channels [18, 19] (see also [20, 21]). Moreover, the large-system multiuser efficiency of the LMMSE detector is found to be the unique solution to the Tse-Hanly fixed-point equation [11] (see also [15] for the special equal-power case). The multiuser system with LMMSE detection admits a single-user characterization, as is also indicated by the notion of effective interference in [11]. It is important to note that

such large-system results are often quite representative of the performance with a moderate to large user population.

Few explicit expressions of the efficiencies in terms of eigenvalues are available beyond the above cases. Little success has been reported in the application of random matrix theory when the detector is nonlinear. It was not until statistical physics techniques were applied to the analysis of multiuser detection that a major breakthrough became possible. Using the so-called replica method, the large-system uncoded minimum bit-error-rate (BER) (hence the optimal multiuser efficiency) and spectral efficiency (the input–output mutual information per dimension) with equal-power binary inputs were first obtained in [22–26] and generalized to the case of arbitrary inputs and powers in [17, 27]. Reference [28] studied the channel capacity under separate decoding and noted that the additive decomposition of the optimum spectral efficiency in [19] holds also for binary inputs. The same formula was conjectured to be valid regardless of the input distribution [29]. The most general framework to date is developed in [27, 30–32], where both joint decoding and generic multiuser detection followed by separate decoding are studied assuming an arbitrary input distribution and flat fading.

The main results of the chapter are presented in Section 1.3. The centerpiece is a single-user characterization of the multiuser system, called the “decoupling principle”, which states that the multiuser channel followed by generic detection is essentially equivalent to a bank of single-user channels, one for each user. The conjecture in [29] is also validated. The decoupling principle carries great practicality and finds convenient uses in finite-size systems where the analytical asymptotic results are a good approximation. It is also found to be applicable to multirate and multicarrier CDMA [33, 34].

1.1.3 On the Replica Method

The replica method, which underlies most of the results in this chapter, was invented in 1975 by S. F. Edwards and P. W. Anderson to study the free energy of disordered magnetic systems, called spin glasses [35]. It has since become a standard technique in statistical physics [36]. Analogies between statistical physics and neural networks, image processing, and communications have gradually been noticed (e.g., [37, 38]), on the basis of which the range of application of the replica method has been expanding. There have been many recent activities applying statistical physics wisdom and the replica method to sparse-graph error-control codes (e.g., [39–43]). The same techniques have also been used to study the capacity of MIMO channels [44, 45]. Among other techniques, mean field theory is used to derive iterative detection algorithms [46, 47].

For the purpose of analytical tractability, we will invoke several assumptions crucial to the replica method and common in the statistical physics literature (see Section 1.4.1.2). Unfortunately, these assumptions have not been fully justified. Thus although there has been some recent progress [48, 49], the mathematical rigor of the general results in this chapter is pending on breakthroughs in those problems. Note, however, that the key results have been rigorously proved in the special case of relatively small load and where the spreading matrix is sparse in some sense by showing the optimality of belief propagation (BP) [50–52]. The technique paves a new avenue for the interpretation and justification of the general results. The replica analysis of generic detection is presented in Section 1.4. Some further discussions

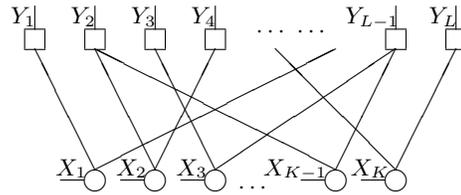


Figure 1.1. A bipartite factor graph describing the probability law of a multiuser system.

on the replica method is found in Section 1.5. Useful statistical physics concepts and methodologies are introduced in Section 1.6.

1.1.4 Statistical Inference Using Practical Algorithms

The input-output relationship of a multiuser systems can in general be fully described (probabilistically) using a bipartite factor graph. As the example shown in Figure 1.1., each edge connects a *symbol node* which represents an input symbol and a *chip node* which represents a component of the output signal. The problem of multiuser detection can be regarded as a statistical inference problem on the graph. As mentioned earlier, performing the inference exactly, e.g., in order to obtain the posterior mean, is computationally hard.

An important family of iterative algorithms for performing the computation approximately is known as *belief propagation*, the formulation of which is attributed to Pearl [53]. In fact, all multiuser detection algorithms discussed in this chapter can be regarded as BP on the factor graph with appropriate heuristic postulates. BP or its variations with linear complexity are especially appealing in practice. Section 1.7 discusses how to design low-complexity algorithms based on BP. In particular, it is shown that parallel interference cancellation (PIC) can be understood as a further simplification of BP.

1.1.5 Statistical Physics and Related Problems

This chapter can be regarded not only as an application of statistical physics ideas and techniques to the communication problem at hand, but also as progress in a much broader research trend, where large-scale problems in various fields are formulated using probability theory and analyzed using statistical physics. The trend can be traced back to the 1980s, where researchers of spin glasses, whose primary objective is to obtain macroscopic characterizations of large disordered systems on the basis of their microscopic specifications, became aware that their methodologies can also be applicable to problems outside statistical physics, such as constraint satisfaction problems [54, 55] and neural networks [56]. Successes in these fields have triggered subsequent applications of statistical physics to various other fields, such as information and communication theory, computation theory, learning and artificial intelligence, etc. One example of the most exciting interplay between these multitude of disciplines is found in recent research activities of sparse-graph error-control codes, where “macroscopic” analysis and “microscopic” algorithm design are concurrently studied, revealing a deep relationship between statistical physics

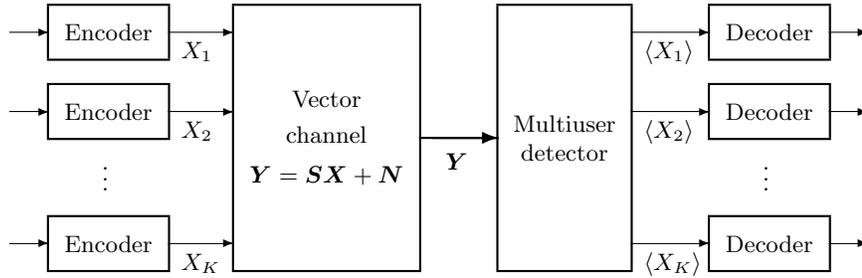


Figure 1.2. Single-user encoding, multiuser channel, and multiuser detection followed by independent single-user decoding.

characterization of a problem and properties of inference algorithms for solving it, as well as demonstrating importance of statistical physics concepts such as phase transition and finite size scaling, in the context of error-control coding [57, 58]. This fertile interdisciplinary field sets the stage for the unique treatment of the multidimensional communication problem described in this chapter.

1.2 GENERIC MULTIUSER DETECTION

In this section, we first describe the multiuser system model considered in this chapter. We then put forth a framework of generic multiuser detection and specialize it to several popular detectors.

1.2.1 CDMA/MIMO Channel Model

Consider the vector channel depicted in Figure 1.2., which in general models a MIMO system. In this chapter, the model describes a real-valued¹ fully-synchronous K -user CDMA system with spreading factor L . Each encoder maps its message into a sequence of channel symbols. All users employ the same type of signaling so that at each interval the K symbols are independent and identically distributed (i.i.d.) random variables with distribution (probability measure) P_X . Let $\mathbf{X} = [X_1, \dots, X_K]^T$ denote the vector of input symbols from the K users in one symbol interval. For notational convenience in the analysis, it is assumed that either a probability density function (pdf) or a probability mass function (pmf) of the distribution P_X exists,² and is denoted by p_X . Let $p_{\mathbf{X}}(\mathbf{x}) = \prod_{k=1}^K p_X(x_k)$ denote the joint (product) distribution.³

Let the instantaneous SNR of user k be denoted by γ_k and $\mathbf{A} = \text{diag}\{\sqrt{\gamma_1}, \dots, \sqrt{\gamma_K}\}$. Denote the spreading sequence of user k by $\mathbf{S}_k = \frac{1}{\sqrt{L}}[S_{1k}, S_{2k}, \dots, S_{Lk}]^T$, where S_{nk} are i.i.d. random variables with zero mean, unit variance and finite moments. The realization of S_{lk} and \mathbf{S}_k are denoted by s_{lk} and \mathbf{s}_k to distinguish from their random counterparts. The $L \times K$ channel “state” matrix is denoted by $\mathbf{S} = [\sqrt{\gamma_1} \mathbf{S}_1, \dots, \sqrt{\gamma_K} \mathbf{S}_K]$. The synchronous flat-fading CDMA channel is

¹Extension to a complex-valued system is straightforward [27].

²Validity of the results in this chapter do not depend on the existence of a pdf or pmf.

³The main results of this chapter extend to cases where the entries of \mathbf{X} are dependent: See [32].

described by:

$$\mathbf{Y} = \sum_{k=1}^K \sqrt{\gamma_k} \mathbf{S}_k X_k + \mathbf{N} \quad (1.1)$$

$$= \mathbf{S}\mathbf{X} + \mathbf{N} \quad (1.2)$$

where $\mathbf{N} \sim \mathcal{N}(0, \mathbf{I})$ is a vector of independent standard Gaussian entries. Without loss of generality we assume P_X to have zero mean and unit variance.

1.2.2 Generic Posterior Mean Estimation

As depicted in Figure 1.2., the multiuser detector front end estimates the transmitted symbols given the received signal and the channel state without using any knowledge of the error-control codes employed by the transmitters. Meanwhile, each single-user decoder only observes the sequence of decision statistics corresponding to one user, and ignores the existence of all other users. By adopting this separate decoding approach, the channel together with the multiuser detector front end is viewed as a bank of coupled single-user channels. Note that the detection output sequence for an individual user is in general not a sufficient statistic for decoding this user's own information.

To capture the intended suboptimal structure, we restrict the capability of the multiuser detector; otherwise the detector could in principle encode the channel state and the received signal (\mathbf{S}, \mathbf{Y}) into a single real number as its output to each user, which is a sufficient statistic for all users. A plausible choice is the (canonical) *posterior mean estimator* (PME), which computes the mean value of the posterior probability distribution $p_{\mathbf{X}|\mathbf{Y},\mathbf{S}}$, hereafter denoted by angle brackets $\langle \cdot \rangle$:

$$\langle \mathbf{X} \rangle = \mathbb{E} \{ \mathbf{X} \mid \mathbf{Y}, \mathbf{S} \}. \quad (1.3)$$

The expectation is taken over the posterior probability distribution $p_{\mathbf{X}|\mathbf{Y},\mathbf{S}}$, which is induced from the input distribution $p_{\mathbf{X}}$ and the conditional Gaussian density function $p_{\mathbf{Y}|\mathbf{X},\mathbf{S}}$ of the channel (1.2) by Bayes' formula:⁴

$$p_{\mathbf{X}|\mathbf{Y},\mathbf{S}}(\mathbf{x}|\mathbf{y}, \mathbf{s}) = \frac{p_{\mathbf{X}}(\mathbf{x})p_{\mathbf{Y}|\mathbf{X},\mathbf{S}}(\mathbf{y}|\mathbf{x}, \mathbf{s})}{\int p_{\mathbf{X}}(\mathbf{x})p_{\mathbf{Y}|\mathbf{X},\mathbf{S}}(\mathbf{y}|\mathbf{x}, \mathbf{s}) \, d\mathbf{x}} \quad (1.4)$$

where the integral shall be replaced by a sum if \mathbf{X} is discrete. Note that, although implicit in notation, $\langle \mathbf{X} \rangle$ is a function of (\mathbf{Y}, \mathbf{S}) , which is dependent on the input \mathbf{X} through (1.2).

Also known as the *conditional mean estimator*, (1.3) achieves the minimum mean-square error for each user, and is therefore the (nonlinear) MMSE detector. We also regard it as a soft-output version of the individually optimal multiuser detector (assuming uncoded transmission). In case of binary antipodal transmission, the posterior mean estimate is consistent in its sign with the individually optimal hard decision. Although this consistency property does not hold for general m -ary

⁴Uppercase letters are usually used for matrices and random variates, while lowercase letters are used for deterministic scalars and vectors. As a compromise, the realization of the spreading matrix \mathbf{S} is denoted as \mathbf{s} . We keep the use of \mathbf{s} to minimum.

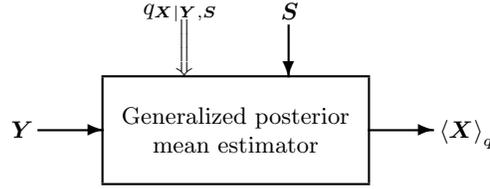


Figure 1.3. Generalized posterior mean estimator.

constellation, (1.3) is optimal in mean-square sense, and is thus a sensible detection output to be used for further decoding.⁵

The PME can be understood as an “informed” optimal estimator which is supplied with the posterior distribution and then computes its mean. A generalization of the canonical PME is conceivable: Instead of informing the estimator of the actual posterior $p_{\mathbf{X}|\mathbf{Y},\mathbf{S}}$, we can supply at will any well-defined conditional distribution $q_{\mathbf{X}|\mathbf{Y},\mathbf{S}}$ as depicted in Figure 1.3.. Given (\mathbf{Y}, \mathbf{S}) , the estimator can nonetheless perform “optimal” estimation based on this postulated measure q . We call this the *generalized posterior mean estimation*, which is conveniently denoted as

$$\langle \mathbf{X} \rangle_q = \mathbb{E}_q \{ \mathbf{X} \mid \mathbf{Y}, \mathbf{S} \} \quad (1.5)$$

where $\mathbb{E}_q\{\cdot\}$ stands for the expectation with respect to the postulated measure q . Suppose $q_{\mathbf{X}|\mathbf{Y},\mathbf{S}}$ is induced from a prior $q_{\mathbf{X}}$ and a conditional distribution $q_{\mathbf{Y}|\mathbf{X},\mathbf{S}}$, then the generalized PME can be expressed as the expectation of \mathbf{X} taken over the postulated posterior probability distribution

$$q_{\mathbf{X}|\mathbf{Y},\mathbf{S}}(\mathbf{x}|\mathbf{y}, \mathbf{s}) = \frac{q_{\mathbf{X}}(\mathbf{x})q_{\mathbf{Y}|\mathbf{X},\mathbf{S}}(\mathbf{y}|\mathbf{x}, \mathbf{s})}{\int q_{\mathbf{X}}(\mathbf{x})q_{\mathbf{Y}|\mathbf{X},\mathbf{S}}(\mathbf{y}|\mathbf{x}, \mathbf{s}) d\mathbf{x}}. \quad (1.6)$$

For brevity, we also refer to (1.5) as the PME. In view of (1.3), the subscript in (1.5) can be dropped if the postulated measure q coincides with the actual one p .

In general, postulating a mismatched measure $q \neq p$ causes degradation in detection performance. Such a strategy may be either due to lack of knowledge of the true statistics or a particular choice that anticipates benefits, such as reduction of computational complexity. In principle, any deterministic estimation strategy can be regarded as a PME since we can always choose to put a unit mass at the desired estimation output given (\mathbf{Y}, \mathbf{S}) , the fact which demonstrates that the concept of PME is generic and versatile. We will see in Section 1.2.3 that by postulating an appropriate measure q , the PME can be particularized to many popular multiuser detectors. The generic representation (1.5) is pivotal here because it allows a unified treatment of a large family of multiuser detectors which results in a simple single-user characterization for all of them.

In this chapter, the posterior $q_{\mathbf{X}|\mathbf{Y},\mathbf{S}}$ supplied to the PME is assumed to be the one that is induced from a postulated CDMA system, where the input follows a

⁵A more sophisticated detector produces the posterior distribution about each input symbol, which generally contains much richer content than point estimates such as the posterior mean. However, as we shall see in this chapter, the posterior distribution is equivalent to a conditional Gaussian distribution in large systems so that the posterior mean suffices.

certain distribution q_X , and the input–output relationship of the postulated channel differs from the actual channel (1.2) by only the noise variance. Precisely, the postulated system is characterized by

$$\mathbf{Y} = \mathbf{S}\mathbf{X}' + \sigma\mathbf{N}' \quad (1.7)$$

where \mathbf{S} is the state matrix of the actual channel (1.2), the components of \mathbf{X}' are i.i.d. with distribution q_X , and \mathbf{N}' is statistically the same as the Gaussian noise \mathbf{N} in (1.2). The postulated input distribution q_X is assumed to have zero mean and unit variance. The posterior $q_{\mathbf{X}|\mathbf{Y},\mathbf{S}}$ is determined by q_X and $q_{\mathbf{Y}|\mathbf{X},\mathbf{S}}$ according to Bayes' formula (1.6). The postulated noise level σ serves as a control parameter. Indeed, the PME so defined is the optimal detector for a postulated multiuser system with its input distribution and noise level different from the actual ones. In general, the postulated channel state could also be different from the actual instance \mathbf{S} , but this is out of the scope of this chapter, as we limit ourselves to studying the (fairly rich) family of multiuser detectors that can be represented as PMEs parameterized by the postulated input and noise level (q_X, σ) .

1.2.3 Specific Detectors as Posterior Mean Estimators

We identify specific choices of the postulated input distribution q_X and noise level σ under which the PME is particularized to well-known multiuser detectors.

The characteristic of the actual channel (1.2) is

$$p_{\mathbf{Y}|\mathbf{X},\mathbf{S}}(\mathbf{y}|\mathbf{x},\mathbf{s}) = (2\pi)^{-\frac{L}{2}} \exp\left[-\frac{1}{2}\|\mathbf{y} - \mathbf{s}\mathbf{x}\|^2\right], \quad (1.8)$$

and that of the postulated channel is

$$q_{\mathbf{Y}|\mathbf{X},\mathbf{S}}(\mathbf{y}|\mathbf{x},\mathbf{s}) = (2\pi\sigma^2)^{-\frac{L}{2}} \exp\left[-\frac{1}{2\sigma^2}\|\mathbf{y} - \mathbf{s}\mathbf{x}\|^2\right]. \quad (1.9)$$

The posterior distribution can be obtained using Bayes' formula (cf. (1.6)) as

$$q_{\mathbf{X}|\mathbf{Y},\mathbf{S}}(\mathbf{x}|\mathbf{y},\mathbf{s}) = \frac{(2\pi\sigma^2)^{-\frac{L}{2}} q_{\mathbf{X}}(\mathbf{x})}{q_{\mathbf{Y}|\mathbf{S}}(\mathbf{y}|\mathbf{s})} \exp\left[-\frac{\|\mathbf{y} - \mathbf{s}\mathbf{x}\|^2}{2\sigma^2}\right] \quad (1.10)$$

where

$$q_{\mathbf{Y}|\mathbf{S}}(\mathbf{y}|\mathbf{s}) = (2\pi\sigma^2)^{-\frac{L}{2}} \mathbb{E}_q \left\{ \exp\left[-\frac{\|\mathbf{y} - \mathbf{S}\mathbf{X}\|^2}{2\sigma^2}\right] \middle| \mathbf{S} = \mathbf{s} \right\} \quad (1.11)$$

and the expectation in (1.11) is taken over $\mathbf{X} \sim q_{\mathbf{X}}$.

1.2.3.1 Linear Detectors Let the postulated input be Gaussian, i.e., q_X is $\mathcal{N}(0, 1)$. The optimal detector (PME) for the postulated model (1.9) with this Gaussian input is a linear filtering of the received signal \mathbf{Y} :

$$\langle \mathbf{X} \rangle_q = [\mathbf{s}^\top \mathbf{s} + \sigma^2 \mathbf{I}]^{-1} \mathbf{s}^\top \mathbf{Y}, \quad (1.12)$$

which corresponds to the LMMSE detector and the decorrelator by choosing $\sigma = 1$ and $\sigma \rightarrow 0$ respectively. If $\sigma \rightarrow \infty$, (1.12) is consistent with the SUMF output:

$$\sigma^2 \langle X_k \rangle_q \longrightarrow \mathbf{s}_k^\top \mathbf{Y}, \quad \text{in } L^2 \text{ as } \sigma \rightarrow \infty. \quad (1.13)$$

1.2.3.2 Optimal Detectors Let the postulated prior distribution q_X be identical to p_X . Let $\sigma \rightarrow 0$, then the probability mass of the distribution $q_{\mathbf{X}|\mathbf{Y},\mathbf{s}}$ is concentrated on a vector that minimizes $\|\mathbf{y} - \mathbf{s}\mathbf{x}\|$, which also maximizes the likelihood function $p_{\mathbf{Y}|\mathbf{X},\mathbf{s}}(\mathbf{y}|\mathbf{x},\mathbf{s})$. The PME $\lim_{\sigma \rightarrow 0} \langle \mathbf{X} \rangle_q$ is thus equivalent to that of jointly optimal (or maximum-likelihood) detection [15]. Alternatively, if $\sigma = 1$, then the postulated measure coincides with the actual measure, i.e., $q_{\mathbf{X}|\mathbf{Y},\mathbf{s}}(\mathbf{x}|\mathbf{y},\mathbf{s}) = p_{\mathbf{X}|\mathbf{Y},\mathbf{s}}(\mathbf{x}|\mathbf{y},\mathbf{s})$. The PME output $\langle \mathbf{X} \rangle$ is the mean of the posterior probability distribution, which is seen as the (soft) individually optimal detector. Also worth mentioning is that, if $\sigma \rightarrow \infty$, the PME reduces to the SUMF.

1.2.3.3 Interference Cancelers Suppose all symbols but X_1 are revealed as $\hat{x}_2, \dots, \hat{x}_K$. The detector can use

$$q_{\mathbf{Y}|X_1,\mathbf{s}}(\mathbf{y}|x_1,\mathbf{s}) \propto p_{\mathbf{Y}|\mathbf{X},\mathbf{s}}(\mathbf{y}|[x_1, \hat{x}_2, \dots, \hat{x}_K]^\top, \mathbf{s}) \quad (1.14)$$

as the postulated channel characteristics in order to estimate X_1 . The resulting PME of X_1 is simply an estimate obtained by matched filtering the received signal after canceling the interference reconstructed from $\hat{x}_2, \dots, \hat{x}_K$. This scheme can be used for all users in either a successive or a parallel manner as well as in multistage fashion (e.g., [59–62] and [Chapter Grant-Rasmussen this book]). As is shown in Section 1.7, interference cancellation is closely related to efficient approximate algorithms for statistical inference in Bayesian networks.

1.3 MAIN RESULTS: SINGLE-USER CHARACTERIZATION

Before burdening the reader with statistical physics concepts and methodologies, we introduce the main results of this chapter and describe the breakthrough in understanding large multiuser systems made possible by the replica analysis.

A *large system* in this chapter refers to the limit that both the number of users and the spreading factor tend to infinity but with their ratio, known as the *system load*, converging to a positive number, i.e., $K/L \rightarrow \beta > 0$. The load β may or may not be smaller than 1. It is also assumed that the SNRs of all users, $\{\gamma_k\}_{k=1}^K$, are i.i.d. with distribution P_γ , hereafter referred to as the *SNR distribution*. All moments of the SNR distribution are assumed to be finite. Clearly, the empirical distributions of the SNRs converge to the same distribution P_γ as $K \rightarrow \infty$. Note that this SNR distribution captures the (flat) fading characteristics of the channel.

Throughout this chapter we consider detection in one symbol interval assuming that the channel state is known by the receiver.

1.3.1 Is the Decision Statistic Gaussian?

Linear multiuser detectors are easy to analyze because of the simple structure of their decision statistics. In general, the detection output is the sum of three independent components: the desired signal, the MAI and the Gaussian noise, i.e., the (normalized) decision statistic for user k is expressed as

$$\langle X_k \rangle = X_k + \sum_{i \neq k} I_i + N_k. \quad (1.15)$$

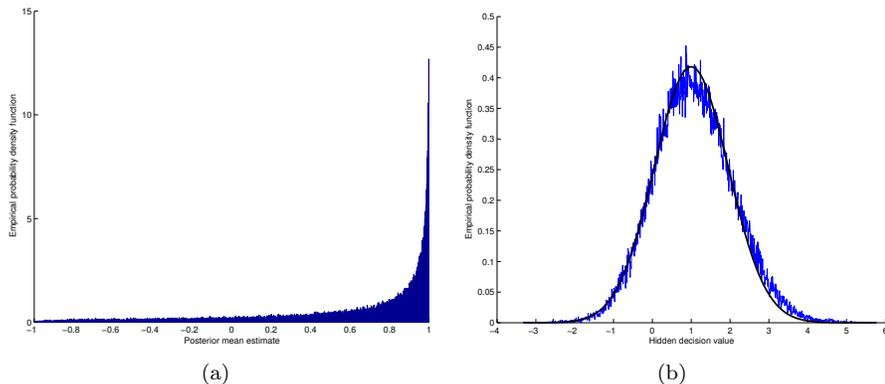


Figure 1.4. The empirical probability density functions of the decision statistics conditioned on +1 being transmitted. The system has 8 users, the spreading factor is 12, and SNR=2 dB. A total of 10,000 trials were recorded. (a) The soft individually optimal detection output. (b) The “hidden” equivalent Gaussian statistic. The asymptotic Gaussian distribution is also plotted for comparison.

The error performance is determined by the statistics of the MAI and the noise. For a sufficiently large system, it is common to assume that the MAI is Gaussian conditioned on the SNRs, so that the performance is quantified as identical to that of a single-user Gaussian channel with the same input but enhanced noise (or, equivalently, degraded SNR). This simple single-user characterization is justified because the MAI converges weakly to a Gaussian random variable, independent of the noise, as $K \rightarrow \infty$ [12].

However, the above analysis does not apply beyond linear detection schemes. The problem here is inherent to nonlinear processing, where the detection output cannot be decomposed as a sum of independent components associated with the desired signal and the unwanted interference respectively. Moreover, the detection output is in general asymptotically non-Gaussian conditioned on the input (consider, e.g., the discrete output of the maximum-likelihood detector in case of binary transmission).

The above difficulty is largely overcome by applying statistical physics methodologies, and in particular the replica method, to the treatment of generic multiuser detection in the large-system regime. Although the output decision statistic of a nonlinear detector cannot be decomposed as (1.15), it converges in the large-system limit to a simple monotone function of a “hidden” Gaussian random variable conditioned on the input X_k , i.e.,

$$\langle X_k \rangle \rightarrow f(Z_k) \quad (1.16)$$

where $Z_k = X_k + W_k$ and W_k is Gaussian and independent of X_k . One may contend that it is always possible to monotonically map a non-Gaussian random variable to a Gaussian one. What is useful (and surprising) here is that 1) the mapping f depends on neither the instantaneous spreading sequences, nor the transmitted symbols which we wish to estimate in the first place; and 2) the statistic Z_k is equal to the desired signal plus an independent Gaussian noise.

By applying an inverse of the function f (which can be readily determined) to $\langle X_k \rangle$, the equivalent conditionally Gaussian statistic Z_k is recovered, so that we

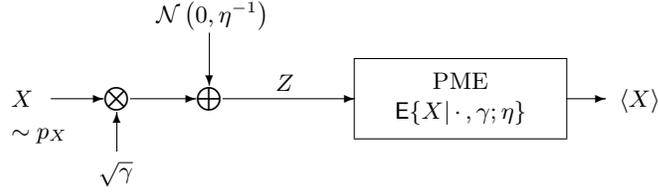


Figure 1.5. The single-user channel and PME.

are back to the familiar ground where the output SINR (defined for the equivalent Gaussian statistic Z_k) completely characterizes the performance for an individual user. We can thus define the multiuser efficiency as the ratio of the output SINR and the input SNR, which is consistent with its original notion in [15].

Example 1 *Figure 1.4.(a) plots the approximate probability density functions obtained from the histogram of the output of the soft individually optimal detector conditioned on +1 being transmitted. Note that negative decision values correspond to decision error; hence the dark area on the negative half plane gives the BER. Since the distribution shown in Figure 1.4.(a) is far from Gaussian, the usual notion of output SINR fails to capture the system performance. In fact, much work in the literature is devoted to evaluating the error performance by Monte Carlo simulation. Figure 1.4.(b) plots the density of the conditionally Gaussian statistic obtained by applying f^{-1} to the non-Gaussian detection output in Figure 1.4.(a). The theoretically predicted Gaussian density function (the smooth curve) is also shown for comparison. The “fit” is remarkable considering that a relatively small system of 8 users with spreading factor 12 is considered. Note that when the multiuser detector is linear, the mapping f is also linear, and (1.16) reduces to (1.15).*

The above example demonstrates the decoupling principle. The asymptotic normality of the decision statistic or its function allows the performance of multiuser systems to be simply characterized by the effective SNR, or SINR, of the detection output. The main claims are formally stated in the following, first for optimal detection (Section 1.3.2) and then for generic multiuser detection (Section 1.3.3). The analysis and discussion of statistical physics techniques are relegated to Sections 1.4 and 1.6.

1.3.2 The Decoupling Principle: Individually Optimal Detection

In order to describe the decoupling result, we first introduce a scalar channel:

$$Z = \sqrt{\gamma} X + \frac{1}{\sqrt{\eta}} N \quad (1.17)$$

where $X \sim p_X$, γ is the channel gain, $N \sim \mathcal{N}(0, 1)$ the additive Gaussian noise independent of X , and $\eta > 0$ the *inverse noise variance*, which is also understood as the degradation of the channel. The conditional distribution associated with the channel (1.17) is

$$p_{Z|X, \gamma; \eta}(z|x, \gamma; \eta) = \sqrt{\frac{\eta}{2\pi}} \exp\left[-\frac{\eta}{2} (z - \sqrt{\gamma} x)^2\right] \quad (1.18)$$

where we generally treat γ as a random variable but η a deterministic parameter. Thus (1.17) is a flat-fading channel. However, since we are interested in a single symbol interval with γ known to the receiver, it is more convenient to refer to (1.17) as a Gaussian channel (for given γ). With $X \sim p_X$ and given η and γ , the input-output mutual information of the channel (1.17) is denoted by $I(X; \sqrt{\eta\gamma}X + N)$. The posterior mean estimate of X given the output Z is⁶

$$\langle X \rangle = \mathbf{E} \{X|Z, \gamma; \eta\}, \quad (1.19)$$

which is an implicit function of Z . The Gaussian channel concatenated with the PME is depicted in Figure 1.5.. Clearly, $\langle X \rangle$ is also the (nonlinear) MMSE estimate, since it achieves the minimum mean-square error:

$$\mathcal{E}_X(\eta\gamma) = \mathbf{E} \{ (X - \langle X \rangle)^2 | \gamma; \eta \}. \quad (1.20)$$

Throughout this chapter, the (decreasing) function $\mathcal{E}_X(a)$ denotes the MMSE of estimating X in Gaussian noise with SNR equal to a .

Consider the individually optimal detection defined by (1.3) and also described in Section 1.2.3.2, where the detection output for user k is the posterior mean estimate $\mathbf{E} \{X_k | \mathbf{Y}, \mathbf{S}\}$. We claim⁷ that, from a single user's perspective, the channel between the input and detection output is asymptotically equivalent to the scalar Gaussian channel (1.17) with an appropriate value of η that is interpreted as the multiuser efficiency.

Claim 1 *In the large-system limit, the distribution of the output $\langle X_k \rangle$ of the individually optimal detector for the multiuser channel (1.2) conditioned on $X_k = x$ being transmitted with SNR $\gamma_k = \gamma$ converges to the distribution of the posterior mean estimate $\langle X \rangle$ of the single-user Gaussian channel (1.17) conditioned on $X = x$ being transmitted, i.e., the posterior cumulative distribution function (cdf)*

$$P_{\langle X_k \rangle | X_k, \gamma_k}(\tilde{x}|x, \gamma) \longrightarrow P_{\langle X \rangle | X, \gamma}(\tilde{x}|x, \gamma) \quad (1.21)$$

for all γ and all x, \tilde{x} where the cdf P_X is continuous.⁸ Here, the optimal multiuser efficiency η is determined from the following fixed-point equation:⁹

$$\eta^{-1} = 1 + \beta \mathbf{E} \{ \gamma \mathcal{E}_X(\eta\gamma) \} \quad (1.22)$$

where the expectation is taken over P_γ . In case (1.22) has more than one solution, η is chosen to minimize¹⁰

$$\mathbf{C}_{\text{joint}} = \beta I(X; \sqrt{\eta\gamma}X + N | \gamma) + \frac{1}{2}[(\eta - 1) \log e - \log \eta]. \quad (1.23)$$

⁶The posterior mean estimate is defined for both the single-user model (e.g., (1.17)) and the multiuser model (e.g., (1.2)) and denoted by the same notation $\langle \cdot \rangle$. The meaning of the notation should be clear from the context.

⁷Since as explained in Section 1.1, rigorous justification for some of the key statistical physics tools (essentially the replica method) is still pending, the key results in this chapter are referred to as claims. Proofs are provided in Section 1.4 based on several assumptions.

⁸If X is a continuous random variable then the cdf is continuous on $(-\infty, \infty)$. If X is discrete then the cdf is continuous at all but a finite or countable number of values.

⁹Because of the way the MMSE is defined, the fixed-point equation is true for arbitrary input distribution P_X which need not have zero mean and unit variance.

¹⁰The base of logarithm is consistent with the unit of information measure throughout unless stated otherwise.

It is important to note that the efficiency η does not depend on any specific SNRs in the large system; rather, it depends only on the distributions P_γ and P_X . The conditional mutual information in (1.23) is obtained as an average over the SNR distribution

$$I(X; \sqrt{\eta\gamma}X + N | \gamma) = \int_0^\infty I(X; \sqrt{\eta t}X + N) dP_\gamma(t). \quad (1.24)$$

The physical meaning of C_{joint} will be clear shortly. Note that the left-hand side (LHS) of (1.21) is a random cdf dependent on the matrix \mathbf{S} . The convergence in (1.21) holds in probability.

The essence of Claim 1 is the following single-user characterization of multiuser systems: From an individual user's viewpoint, the input–output relationship of the multiuser channel and PME is increasingly similar to that under a simple single-user setting as the system becomes large. Indeed, given the (scalar) input and output statistics, it is impossible to distinguish whether the underlying system is in the (large) multiuser or the single-user setting. It is also interesting to note that the (asymptotically) equivalent single-user system takes an analogous structure as the multiuser one (compare Figures 1.2. and 1.5.). Note that the conditionally Gaussian variable Z is not directly available in the multiuser system. Rather, one can process (\mathbf{Y}, \mathbf{S}) to obtain Z as a sufficient statistic for X (see e.g., [52]).

The single-user PME (1.19) is merely a decision function applied to the Gaussian channel output, which can be expressed explicitly as

$$\mathbb{E}\{X | Z = z, \gamma; \eta\} = \frac{p_1(z, \gamma; \eta)}{p_0(z, \gamma; \eta)} \quad (1.25)$$

where we define the following useful functions

$$p_i(z, \gamma; \eta) = \mathbb{E}\{X^i p_{Z|X, \gamma; \eta}(z | X, \gamma; \eta) | \gamma\}, \quad i = 0, 1, \dots \quad (1.26)$$

where the expectation is taken over p_X . Note that $p_0(z, \gamma; \eta) = p_{Z|\gamma; \eta}(z | \gamma; \eta)$. The decision function (1.25) is in general nonlinear.

The MMSE can be computed as¹¹

$$\mathcal{E}_X(\eta\gamma) = 1 - \int \frac{p_1^2(z, \gamma; \eta)}{p_0(z, \gamma; \eta)} dz. \quad (1.27)$$

Solutions to the fixed-point equation (1.22) can in general be found numerically. The conditional mutual information (over $\gamma \sim P_\gamma$) in (1.23) is also easy to compute.

Example 2 *Assume all users take binary antipodal input and the same SNR of 2 dB ($\gamma = 1.585$). Let $\beta = 2/3$. Solving the fixed-point equation (1.22) yields $\eta = 0.69$. Thus, from each user's point of view, if individually optimal detection is employed, the distribution of the decision statistic is identical to that of the posterior mean estimate of the input to a Gaussian channel with SNR equal to $\eta\gamma = 1.098$ (0.41 dB). The distribution of the detection output conditioned on +1 being transmitted is shown in Figure 1.4.(b), which is centered at 1 with a variance of $1/(\eta\gamma) = 0.911$.*

¹¹The integral with respect to z is from $-\infty$ to ∞ . For notational simplicity we omit integral limits in this chapter whenever they are clear from context.

The fixed-point equations (1.22) may have multiple solutions. This is known as phase coexistence in statistical physics. Among those solutions, the (thermodynamically) dominant solution gives the smallest value of C_{joint} , which is in fact the optimal spectral efficiency as we shall discuss in Section 1.4.2.3. This is the solution that carries relevant operational meaning in the communication problem. In general, as the system parameters (such as the load) change, the dominant solution may switch from one of the coexisting solutions to another. This phenomenon is known as *phase transition*.

The decision function (1.25) is one-to-one because of the following, which is easily proved using the Cauchy-Schwartz inequality [27].

Proposition 1 *The decision function (1.25) is strictly monotone increasing in z for all $\gamma, \eta > 0$.*

This monotonicity result is not surprising because larger (smaller) values of channel output is likely to be caused by larger (smaller) values of the input.

In the large-system limit, given the detection output $\langle X_k \rangle$, one can apply the inverse of the decision function to recover an equivalent conditionally Gaussian statistic Z , which is centered at the actual input X_k scaled by $\sqrt{\gamma_k}$ with a variance of η^{-1} . Note that $\eta \in [0, 1]$ from (1.22). It is clear that the MAI is asymptotically equivalent to an enhancement of the noise by η^{-1} , i.e., the effective SNR is reduced by a factor of η , hence the term *multiuser efficiency*. Indeed, in the large-system limit, the multiuser channel with the PME front end can be decoupled into a bank of independent single-user Gaussian channels with the same degradation in each user's SNR.

Corollary 1 *In the large-system limit, the mutual information between input symbol and the output of the individually optimal multiuser detector for each user is equal to the input-output mutual information of the equivalent single-user Gaussian channel with the same input and SNR degraded by η , which is the multiuser efficiency given by Claim 1. That is, conditioned on the input SNR being γ_k for user k ,*

$$I(X_k; \langle X_k \rangle | \mathbf{S}) \rightarrow I(X; \sqrt{\eta \gamma_k} X + N) \quad (1.28)$$

where $X \sim p_X$ and $N \sim \mathcal{N}(0, 1)$ are independent.

The overall spectral efficiency under separate decoding is the average of all users' mutual information multiplied by the load:

$$C_{\text{sep}}(\beta) = \beta I(X; \sqrt{\eta \gamma} X + N | \gamma). \quad (1.29)$$

The optimal spectral efficiency under joint decoding is greater than that under separate decoding (1.29), where the increase is given by the following:

Claim 2 *The spectral efficiency gain of optimal joint decoding over individually optimal detection followed by separate decoding of the multiuser channel (1.2) is determined, in the large-system limit, by the optimal multiuser efficiency as*

$$C_{\text{joint}}(\beta) - C_{\text{sep}}(\beta) = \frac{1}{2} [(\eta - 1) \log e - \log \eta] \quad (1.30)$$

$$= D(\mathcal{N}(0, \eta) \| \mathcal{N}(0, 1)). \quad (1.31)$$

Indeed, the spectral efficiency under joint decoding is given by (1.23).

As a by-product, Müller's conjecture on the mutual information loss [28, 29] is true for arbitrary inputs and SNRs. Incidentally, the loss is identified as a Kullback-Leibler divergence [63] between two Gaussian distributions in (1.31) [27].

Interestingly, the spectral efficiencies under joint and separate decoding are also related by an integral equation, which was originally given in [19, (160)] for the special case of Gaussian inputs.

Theorem 1 *Regardless of the input and SNR distributions,*

$$C_{\text{joint}}(\beta) = \int_0^\beta \frac{1}{\beta'} C_{\text{sep}}(\beta') d\beta'. \quad (1.32)$$

Proof: Since $C_{\text{joint}}(0) = 0$ trivially, it suffices to show

$$\beta \frac{d}{d\beta} C_{\text{joint}}(\beta) = C_{\text{sep}}(\beta). \quad (1.33)$$

By (1.31) and (1.23), it is enough to show

$$\beta \frac{d}{d\beta} I(X; \sqrt{\eta\gamma} X + N | \gamma) + \frac{1}{2} \frac{d}{d\beta} [(\eta - 1) \log e - \log \eta] = 0. \quad (1.34)$$

As the efficiency η is a function of the system load β , (1.34) is equivalent to

$$\frac{d}{d\eta} I(X; \sqrt{\eta\gamma} X + N | \gamma) + \frac{1}{2\beta} (1 - \eta^{-1}) \log e = 0. \quad (1.35)$$

The mutual information and the MMSE in Gaussian channels are related by the following formula [64, Theorem 1],

$$\frac{1}{\log e} \frac{d}{dg} I(X; \sqrt{g} X + N) = \frac{1}{2} \mathcal{E}_X(g), \quad \forall g. \quad (1.36)$$

Thus (1.35) holds as η satisfies the fixed-point equation (1.22). \blacksquare

Theorem 1 is an outcome of the chain rule of mutual information:

$$I(\mathbf{X}; \mathbf{Y} | \mathbf{S}) = \sum_{k=1}^K I(X_k; \mathbf{Y} | \mathbf{S}, X_{k+1}, \dots, X_K). \quad (1.37)$$

The LHS of (1.37) is the total mutual information of the multiuser channel. Each mutual information in the right-hand side (RHS) is a single-user mutual information over the multiuser channel conditioned on the symbols of previously decoded users. As argued below, the limit of (1.37) as $K \rightarrow \infty$ becomes the integral equation (1.32).

Consider a successive interference canceler with PME front ends against yet undecoded users in which reliably decoded symbols are used to reconstruct the interference for cancellation. Since the error probability of decoded symbols vanishes with code block-length, the interference from decoded users are asymptotically completely removed. Assume without loss of generality that the users are decoded

in reverse order, then the PME for user k sees only $k - 1$ interfering users. Hence the performance for user k under such successive decoding is identical to that under multiuser detection with separate decoding in a system with k instead of K users. Nonetheless, the equivalent single-user channel for each user is Gaussian by Claim 1. The multiuser efficiency experienced by user k , $\eta(k/L)$, is a function of the load k/L seen by the PME for user k . By Corollary 1, the single-user mutual information for user k is therefore

$$I\left(X; \sqrt{\eta(k/L)} \gamma_k X + N\right). \quad (1.38)$$

The overall spectral efficiency under successive decoding converges almost surely by the law of large numbers:

$$\frac{1}{L} \sum_{k=1}^K I\left(X; \sqrt{\eta(k/L)} \gamma_k X + N\right) \rightarrow \int_0^\beta I(X; \sqrt{\eta(\beta')} \gamma X + N | \gamma) d\beta' \quad (1.39)$$

which is the RHS of (1.32). This suggests that decoding and stripping users one-by-one in a large system is tantamount to increasing the SNR little-by-little in some intricate way.

Together with Theorem 1, the convergence in (1.39) implies the following:

Corollary 2 *In the large-system limit, successive decoding with an individually optimal detection front end against yet undecoded users achieves the optimal multiuser channel capacity under any constraint on the input.*

Corollary 2 is a generalization of the result that a successive canceler with a linear MMSE front end against undecoded users achieves the capacity of the CDMA channel under Gaussian inputs.¹²

1.3.3 Decoupling Principle: Generic Multiuser Detection

1.3.3.1 A Companion Channel Consider a random transformation $p_{Y|X}$ which characterizes a memoryless channel $X \rightarrow Y$. The problem of Bayesian inference is in general to infer about X given Y based on the posterior probability law $p_{X|Y}$. Under many circumstances, e.g., when $p_{X|Y}$ is not exactly known, inference may be carried out using an alternative law $q_{X|Y}$. For all estimation purposes, it suffices to know the joint probability distribution of (X, Y, X') where $Y \rightarrow X'$ is characterized by $q_{X|Y}$ and X' is independent of X conditioned on Y . Precisely, $p_{XX'|Y}(x, x'|y) = p_{X|Y}(x|y)q_{X|Y}(x'|y)$ for all (x, x', y) . We call the random transformation $q_{X|Y}$ a *companion channel* of the channel $p_{Y|X}$.

The above can be specialized to the current problem. Let $q_{Z|X, \gamma; \xi}$ represent the input-output relationship of a Gaussian channel akin to (1.17), the only difference being that the inverse noise variance is ξ instead of η ,

$$q_{Z|X, \gamma; \xi}(z|x, \gamma; \xi) = \sqrt{\frac{\xi}{2\pi}} \exp\left[-\frac{\xi}{2}(z - \sqrt{\gamma}x)^2\right]. \quad (1.40)$$

Throughout, we choose to explicitly associate η with distribution p and ξ with distribution q for clarity. Similar to that in the multiuser setting, by postulating the

¹²This principle, originally discovered in [65], has been shown with other proofs and in other settings [18, 66–70].

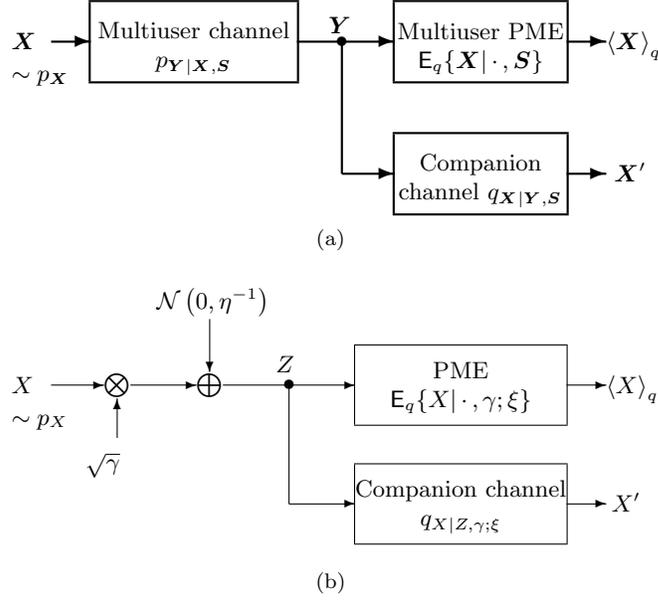


Figure 1.6. (a) The multiuser channel, the (multiuser) PME, and the (multiuser) companion channel. (b) The equivalent single-user Gaussian channel, PME and companion channel.

input distribution to be q_X , a posterior probability distribution $q_{X|Z, \gamma; \xi}$ is induced by q_X and $q_{Z|X, \gamma; \xi}$ using Bayes' formula (cf. (1.6)). Thus we have a single-user companion channel defined by $q_{X|Z, \gamma; \xi}$, which outputs a random variable X' given the channel output Z (Figure 1.6.(b)). A (generalized) single-user PME is defined naturally as:

$$\langle X \rangle_q = E_q \{ X | Z, \gamma; \xi \} = \frac{q_1(Z, \gamma; \xi)}{q_0(Z, \gamma; \xi)} \quad (1.41)$$

where the following functions are defined akin to (1.26):

$$q_i(z, \gamma; \xi) = E_q \{ X^i q_{Z|X, \gamma; \xi}(z|X, \gamma; \xi) | \gamma \}, \quad i = 0, 1, \dots \quad (1.42)$$

where the expectation is taken over q_X . The probability law of the composite system depicted by Figure 1.6.(b) is determined by γ and two parameters η and ξ .

Let us define the mean squared error of the PME as

$$\mathcal{E}(\gamma; \eta, \xi) = E \left\{ \left(X - \langle X \rangle_q \right)^2 \middle| \gamma; \eta, \xi \right\}, \quad (1.43)$$

and also define the variance of the companion channel as

$$\mathcal{V}(\gamma; \eta, \xi) = E \left\{ \left(X' - \langle X \rangle_q \right)^2 \middle| \gamma; \eta, \xi \right\}. \quad (1.44)$$

Note that $\xi = \eta$ if X and X' are i.i.d. given Z .

1.3.3.2 Main Results Consider the multiuser channel (1.2) with input distribution p_X and SNR distribution P_γ . Let its output be fed into the posterior mean estimator (1.5) and a companion channel $q_{\mathbf{X}|\mathbf{Y},\mathbf{S}}$, both parameterized by the postulated input q_X and noise level σ (refer to Figure 1.6.(a)). Let X_k , X'_k , and $\langle X_k \rangle_q$ be the input, the companion channel output and the posterior mean estimate for user k with input SNR γ_k .

Fix $(\beta, P_\gamma, p_X, q_X, \sigma)$. Consider also the single-user Gaussian channel (1.18) with inverse noise variance η and its companion channel depicted in Figure 1.6.(b). Let $X \sim p_X$ be the input to the single-user Gaussian channel, X' be the output of the single-user companion channel parameterized by (q_X, ξ) , and $\langle X \rangle_q$ is the corresponding posterior mean estimate (1.41), with $\gamma = \gamma_k$.

Claim 3 Consider the multiuser and single-user systems described above (also Figure 1.6.).

(a) The joint distribution of $(X_k, X'_k, \langle X_k \rangle_q)$ conditioned on the channel state \mathbf{S} converges in probability as $K \rightarrow \infty$ and $K/L \rightarrow \beta$ to the joint distribution of $(X, X', \langle X \rangle_q)$ with $\gamma = \gamma_k$, i.e., the posterior cdf

$$P_{X_k, X'_k, \langle X_k \rangle_q | \gamma_k}(x, x', \tilde{x} | \gamma) \longrightarrow P_{X, X', \langle X \rangle_q | \gamma}(x, x', \tilde{x} | \gamma) \quad (1.45)$$

in probability for every x, x', \tilde{x} where the cdf P_X is continuous at x, x' , and \tilde{x} .

(b) The parameter η , known as the multiuser efficiency, satisfies together with ξ the coupled equations:

$$\eta^{-1} = 1 + \beta \mathbf{E} \{ \gamma \cdot \mathcal{E}(\gamma; \eta, \xi) \}, \quad (1.46a)$$

$$\xi^{-1} = \sigma^2 + \beta \mathbf{E} \{ \gamma \cdot \mathcal{V}(\gamma; \eta, \xi) \}, \quad (1.46b)$$

where the expectations are taken over P_γ . In case of multiple solutions to (1.46), (η, ξ) is chosen to minimize the free energy expressed as

$$\begin{aligned} \mathcal{F} = & - \mathbf{E} \left\{ \int p_{Z|\gamma; \eta}(z|\gamma; \eta) \log q_{Z|\gamma; \xi}(z|\gamma; \xi) dz \right\} + \frac{1}{2\beta} [(\xi - 1) \log e - \log \xi] \\ & - \frac{1}{2} \log \frac{2\pi}{\xi} - \frac{\xi}{2\eta} \log e + \frac{\sigma^2 \xi (\eta - \xi)}{2\beta \eta} \log e + \frac{1}{2\beta} \log(2\pi) + \frac{\xi}{2\beta \eta} \log e. \end{aligned} \quad (1.47)$$

Claim 3 reveals that, from an individual user's viewpoint, the input-output relationship of the multiuser channel, PME and companion channel is increasingly similar to that under a simple single-user setting as the system becomes large.

Finally, it is straightforward to verify that the decoupling result for individually optimal detection (Claim 1) is a special case of the results for generic detection (Claim 3) with the postulated distribution q identical to the actual distribution p as well as symmetry assumption $\xi = \eta$.

1.3.4 Justification of Results: Sparse Spreading

Claims 1–3 have not been rigorously proved because the underlying replica method is until now an unjustifiable technique. In the following we provide an interpretation of the central fixed-point equation which was first discussed in [31]. In particular, we derive (1.22) under the assumption that interference cancellation based on posterior mean estimates of interfering users is optimal, along with some additional independence assumptions.

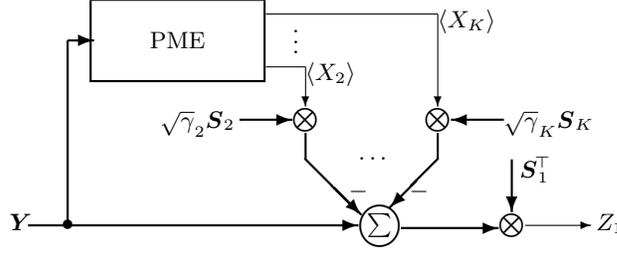


Figure 1.7. A canonical interference canceler equivalent to the single-user channel.

Suppose we construct without loss of generality an estimator for user 1 using interference cancellation as depicted in Figure 1.7. (see also a discussion of interference cancellation in Section 1.7). Let $\langle X_2 \rangle, \dots, \langle X_K \rangle$ be the generalized PME estimates for user 2 through user K . A decision statistic for user 1 can be generated by first subtracting the reconstructed interferences using those estimates and then matched filtering with respect to user 1's spreading sequence:

$$Z_1 = \sqrt{\gamma_1} X_1 + \sum_{k=2}^K \mathbf{S}_1^T \mathbf{S}_k \sqrt{\gamma_k} (X_k - \langle X_k \rangle) + N_1 \quad (1.48)$$

where N_1 is standard Gaussian. We make two specious assumptions:

1. The desired symbol X_1 , the Gaussian noise N_1 , and the residual errors $(X_k - \langle X_k \rangle)$ are independent;
2. The statistic Z_1 is sufficient for achieving the MMSE for X_1 .

By the first assumption, the sum of the residual MAI and Gaussian noise converges to a Gaussian random variable as $K \rightarrow \infty$ by virtue of the central limit theorem. Let the variance of $X_k - \langle X_k \rangle$ be denoted by $V(\gamma_k)$, which depends on γ_k . Then variance of the total interference in (1.48) is

$$1 + \beta \mathbf{E} \{ \gamma V(\gamma) \} \quad (1.49)$$

which implies that the efficiency for user 1 is

$$\eta_1 = \frac{1}{1 + \beta \mathbf{E} \{ \gamma V(\gamma) \}}. \quad (1.50)$$

Evidently, the efficiency is not dependent on the user number and hence identical for all users; the subscript of the efficiency can be dropped. By the second assumption, the mean squared error based on the statistic Z_1 should be equal to the MMSE, which is $\mathcal{E}_X(\eta \gamma_1)$. Note that the same applies to all users, hence $V(\gamma_k) = \mathcal{E}_X(\eta \gamma_k)$. Therefore, formula (1.50) becomes exactly the fixed-point equation (1.22).

Therefore, if the above two assumptions were valid, we would have recovered the fixed-point equation (1.22) in Claim 1. Moreover, we would have constructed a degraded Gaussian channel for user 1 equivalent to the single-user channel as shown in Figure 1.6.(b). We can also argue that every user enjoys the same efficiency since otherwise users with worse efficiency may benefit from users with better efficiency until an equilibrium is reached. Roughly speaking, the PME output is a “fixed-point” of a parallel interference canceler. The multiuser efficiency, in a sense, is the outcome of such an equilibrium.

The above interpretation does not hold in general due to the unjustifiable independence assumption¹³. In particular, $\mathbf{S}_1^\top \mathbf{S}_k$ are not independent, albeit uncorrelated, for all k . Also, $\langle X_k \rangle$ are dependent on the desired signal X_1 and the noise N_1 , which is evident in the special case of linear MMSE detection.

Interestingly, the above argument can be made rigorous in the special case where the spreading matrix \mathbf{S} is *sparse* (or extremely diluted) in some sense (see also [72, 73]). In [50–52], the general formula (1.22) has been justified for binary inputs and arbitrary inputs and SNR respectively with sparse spreading and a relatively small load, which is the first partial proof of (1.22) without resorting to the replica method. The key observation is that, the residual errors are independent over almost all choices of the spreading matrix if the posterior mean estimates are replaced by the (asymptotically equivalent) estimates supplied by parallel interference cancellation, or belief propagation. Interference cancellation and belief propagation are the subject of Section 1.7 where some practical multiuser detection schemes are discussed.

1.3.5 Well-known Detectors as Special Cases

As shown in Section 1.2.3, several well-known multiuser detectors can be regarded as appropriately parameterized PME. Thus many previously known results can be recovered as special cases of the findings in Sections 1.3.2 and 1.3.3.

1.3.5.1 Linear Detectors Let the postulated prior q_X be standard Gaussian so that the multiuser PME represents a linear detector. Since the input Z and output X of the companion channel are jointly Gaussian (refer to Figure 1.6.(b)), the single-user PME is simply a linear attenuator:

$$\langle X \rangle_q = \frac{\xi\sqrt{\gamma}}{1 + \xi\gamma} Z. \quad (1.51)$$

From (1.43), the mean squared error is

$$\mathcal{E}(\gamma; \eta, \xi) = \mathbf{E} \left\{ \left[X_0 - \frac{\xi\sqrt{\gamma}}{1 + \xi\gamma} \left(\sqrt{\gamma} X_0 + \frac{N}{\sqrt{\eta}} \right) \right]^2 \right\} \quad (1.52)$$

$$= \frac{\eta + \xi^2\gamma}{\eta(1 + \xi\gamma)^2}. \quad (1.53)$$

Meanwhile, the variance of X conditioned on Z is independent of Z . Hence the variance (1.44) of the companion channel output is independent of η :

$$\mathcal{V}(\gamma; \eta, \xi) = \frac{1}{1 + \xi\gamma}. \quad (1.54)$$

From Claim 3, one finds that ξ is the solution to

$$\xi^{-1} = \sigma^2 + \beta \mathbf{E} \left\{ \frac{\gamma}{1 + \xi\gamma} \right\}, \quad (1.55)$$

¹³It should be noted, however, that a similar independence argument can also be found in statistical physics literature (see, e.g., [36]). Such independence property is called the “cluster property” in statistical physics [71].

and the multiuser efficiency is determined as

$$\eta = \xi + \xi (\sigma^2 - 1) \left[1 + \beta \mathbf{E} \left\{ \frac{\gamma}{(1 + \xi\gamma)^2} \right\} \right]^{-1} \quad (1.56)$$

which is independent of the input distribution p_X .

Let $\sigma \rightarrow \infty$ so that the PME becomes the matched filter. One finds $\xi\sigma^2 \rightarrow 1$ by (1.55) and consequently, the multiuser efficiency of the matched filter is [15]

$$\eta^{(\text{mf})} = \frac{1}{1 + \beta \mathbf{E} \{\gamma\}}. \quad (1.57)$$

In case $\sigma = 1$, one has the linear MMSE detector. By (1.56), $\eta = \xi$ and by (1.55), the efficiency $\eta^{(\text{lm})}$ is the unique solution to the Tse-Hanly equation [11,18]:

$$\eta^{-1} = 1 + \beta \mathbf{E} \left\{ \frac{\gamma}{1 + \eta\gamma} \right\}. \quad (1.58)$$

By letting $\sigma \rightarrow 0$ one obtains the decorrelator. If $\beta < 1$, then (1.55) gives $\xi \rightarrow \infty$ and $\xi\sigma^2 \rightarrow 1 - \beta$, and the multiuser efficiency is found as $\eta = 1 - \beta$ by (1.56) regardless of the SNR distribution (as shown in [15]). If $\beta > 1$, and assuming the generalized form of the decorrelator as the Moore-Penrose inverse of the correlation matrix [15], then ξ is the unique solution to

$$\xi^{-1} = \beta \mathbf{E} \left\{ \frac{\gamma}{1 + \xi\gamma} \right\} \quad (1.59)$$

and the multiuser efficiency is found by (1.56) with $\sigma = 0$. In the special case of identical SNRs, an explicit expression is found [16,17]

$$\eta^{(\text{dec})} = \frac{\beta - 1}{\beta + \gamma(\beta - 1)^2}, \quad \beta > 1. \quad (1.60)$$

By Claim 3, the mutual information with input distribution p_X for a user with SNR given as γ under linear multiuser detection is $I(X; \langle X \rangle_q) = I(X; \sqrt{\eta\gamma}X + N)$ where $N \sim \mathcal{N}(0,1)$ and η depends on which type of linear detector is in use. By Claim 2, the total spectral efficiency, which is achieved by Gaussian inputs, is expressed in terms of the LMMSE efficiency [19]:

$$\mathbf{C}_{\text{joint}}^{(\text{G})} = \frac{\beta}{2} \mathbf{E} \left\{ \log \left(1 + \eta^{(\text{lm})}\gamma \right) \right\} + \frac{1}{2} \left[\left(\eta^{(\text{lm})} - 1 \right) \log e - \log \eta^{(\text{lm})} \right]. \quad (1.61)$$

1.3.5.2 Optimal Detectors Using the actual input distribution p_X as the postulated prior of the PME results in optimum multiuser detectors. As discussed in Section 1.2.3.2, in case of the jointly optimal detector, the postulated noise level σ is 0, and (1.46) becomes

$$\eta^{-1} = 1 + \beta \mathbf{E} \{ \gamma \cdot \mathcal{E}(\gamma; \eta, \xi) \}, \quad (1.62a)$$

$$\xi^{-1} = \beta \mathbf{E} \{ \gamma \cdot \mathcal{V}(\gamma; \eta, \xi) \}. \quad (1.62b)$$

The parameters can then be solved numerically.

In case of the individually optimal detector, $\sigma = 1$ and $q = p$. The optimal efficiency η is the solution to the fixed-point equation (1.22) given in Claim 1.

It is of practical interest to find the spectral efficiency under the constraint that the input symbols are antipodally modulated as in the popular BPSK. In this case, equally likely prior maximizes the mutual information. The MMSE is

$$\mathcal{E}^{(b)}(\gamma) = 1 - \int \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} \tanh(\gamma - z\sqrt{\gamma}) \, dz, \quad (1.63)$$

where the superscript (b) stands for binary inputs. By Claim 1, the multiuser efficiency $\eta^{(b)}$ is a solution to the fixed-point equation [17]:

$$\frac{1}{\eta} = 1 + \beta \mathbf{E} \left\{ \gamma \left[1 - \int \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} \tanh(\eta\gamma - z\sqrt{\eta\gamma}) \, dz \right] \right\}. \quad (1.64)$$

The channel capacity for a user with binary input, SNR equal to γ and separate decoding is given by [28]

$$\mathbf{C}^{(b)}(\gamma) = - \int \frac{e^{-\frac{z^2}{2}}}{\sqrt{2\pi}} \log \cosh \left(\eta^{(b)}\gamma - z\sqrt{\eta^{(b)}\gamma} \right) \, dz + \eta^{(b)} \gamma \log e. \quad (1.65)$$

The joint-decoding spectral efficiency with binary inputs is thus

$$\mathbf{C}_{\text{joint}}^{(b)} = \beta \mathbf{E} \left\{ \mathbf{C}^{(b)}(\gamma) \right\} + \frac{1}{2} \left[\left(\eta^{(b)} - 1 \right) \log e - \log \eta^{(b)} \right] \quad (1.66)$$

which is also a generalization of an implicit result in [26].

1.4 THE REPLICA ANALYSIS OF GENERIC MULTIUSER DETECTION

This section introduces the replica method and presents the replica analysis of generic multiuser detection which leads to Claims 1–3. We first describe the procedure of the replica method and demonstrate its use with a simple example. We then apply the method to the analysis of the multiuser system and present the calculation of the mutual information in some detail.

1.4.1 The Replica Method

Before describing the replica method, we first revisit the key measures used to characterize the multiuser system, including in particular the input–output mutual information. For convenience, natural logarithms are assumed from this point on.

1.4.1.1 Spectral Efficiency and Detection Performance Consider the multiuser channel, the PME and the companion channel as depicted in Figure 1.6.(a). Fix the input distribution p_X . The key quantity is the spectral efficiency

$$\mathbf{C} = \frac{1}{L} I(\mathbf{X}; \mathbf{Y} | \mathbf{S}), \quad (1.67)$$

which we wish to evaluate. In some cases one may want to evaluate the mutual information $I(\mathbf{X}; \mathbf{Y} | \mathbf{S} = \mathbf{s})$, which is a function of the realization \mathbf{s} of \mathbf{S} . In

such cases one assumes the self-averaging property, in which the random quantity $(1/L)I(\mathbf{X}; \mathbf{Y} | \mathbf{S} = \mathbf{s})$ is assumed to converge to \mathbb{C} as $L \rightarrow \infty$ for almost all realizations of \mathbf{S} . This property has been justified in the special case where q_X is Gaussian [11, 18] as well as in the case of Gaussian spreading sequence and binary input [74].

The spectral efficiency is expressed as

$$\mathbb{C} = \frac{1}{L} \mathbb{E} \left\{ \log \frac{p_{\mathbf{Y}|\mathbf{X},\mathbf{S}}(\mathbf{Y}|\mathbf{X},\mathbf{S})}{p_{\mathbf{Y}|\mathbf{S}}(\mathbf{Y}|\mathbf{S})} \right\} \quad (1.68)$$

$$= -\beta \mathbb{E} \left\{ \frac{1}{K} \log p_{\mathbf{Y}|\mathbf{S}}(\mathbf{Y}|\mathbf{S}) \right\} - \frac{1}{2} \log(2\pi e) \quad (1.69)$$

where the simplification to (1.69) is because $p_{\mathbf{Y}|\mathbf{X},\mathbf{S}}$ given by (1.8) is an L -dimensional Gaussian density. In Section 1.4.1.2 we show that the replica method can be used to calculate the normalized conditional differential entropy in (1.69),

$$\mathbb{E} \left\{ \frac{1}{K} \log p_{\mathbf{Y}|\mathbf{S}}(\mathbf{Y}|\mathbf{S}) \right\} \quad (1.70)$$

which is also referred to as the *free energy* using the physics terminology.

In case of a multiuser detector front end, one is interested in the quality of the detection output for each user, which is completely described by the distribution of the detection output conditioned on the input. Let us focus on an arbitrary user k , and let X_k , $\langle X_k \rangle_q$ and X'_k be the input, the PME output, and the companion channel output, respectively (cf. Figure 1.6.(a)). Instead of the conditional distribution $P_{\langle X_k \rangle_q | X_k}$, we solve a somewhat more ambitious problem: the joint distribution of $(X_k, \langle X_k \rangle_q, X'_k)$ conditioned on the channel state \mathbf{S} in the large-system limit. The replica approach calculates the joint moments

$$\mathbb{E} \left\{ X_k^i (\langle X_k \rangle_q)^j \langle X_k \rangle_q^l \right\}, \quad i, j, l = 0, 1, \dots \quad (1.71)$$

by studying a free-energy-like quantity, as will be discussed in Section 1.4.3. The joint distribution becomes clear once all the moments (1.71) are determined, so does the relationship between the detection output $\langle X_k \rangle_q$ and the input X_k . It turns out that, as stated in Claim 3, the large-system joint distribution of $(X_k, \langle X_k \rangle_q, X'_k)$ is identical to that of the input, PME output and companion channel output associated with a single-user Gaussian channel with the same input distribution but with a degradation in the SNR.

We have distilled the problems under both joint and separate decoding to finding some ensemble averages, namely, the free energy (1.70) and the joint moments (1.71). In order to calculate these quantities, we resort to a powerful technique, the heart of which is sketched in the following.

1.4.1.2 The Replica Method Direct calculation of the differential entropy (free energy) (1.70) is hard. The replica method can be described as the following procedure to that effect:

1. Reformulate the free energy (1.70) as

$$\mathcal{F} = - \lim_{K \rightarrow \infty} \frac{1}{K} \lim_{u \rightarrow 0} \frac{\partial}{\partial u} \log \mathbb{E} \left\{ p_{\mathbf{Y}|\mathbf{S}}^u(\mathbf{Y}|\mathbf{S}) \right\}. \quad (1.72)$$

The equivalence of (1.70) and (1.72) can be verified by noticing that for all positive random variable Θ ,

$$\lim_{u \rightarrow 0} \frac{\partial}{\partial u} \log \mathbb{E} \{\Theta^u\} = \lim_{u \rightarrow 0} \frac{\mathbb{E} \{\Theta^u \log \Theta\}}{\mathbb{E} \{\Theta^u\}} = \mathbb{E} \{\log \Theta\}. \quad (1.73)$$

2. For an arbitrary positive integer u , calculate

$$- \lim_{K \rightarrow \infty} \frac{1}{K} \log \mathbb{E} \left\{ p_{\mathbf{Y}|\mathbf{S}}^u(\mathbf{Y}|\mathbf{S}) \right\} \quad (1.74)$$

by introducing u replicas of the system (hence the name “replica” method).

3. Assuming the resulting expression from Step 2 to be valid for all real-valued u in the vicinity of $u = 0$, take its derivative at $u = 0$ to obtain the free energy (1.72). It is also assumed that the limits in (1.72) can be interchanged.

We note that when analyzing general suboptimal estimators, the free energy is defined as (1.72) with $p_{\mathbf{Y}|\mathbf{S}}(\mathbf{Y}|\mathbf{S})$ replaced by some alternative distribution $q_{\mathbf{Y}|\mathbf{S}}(\mathbf{Y}|\mathbf{S})$ while the expectation remains over the joint probability measure $p_{\mathbf{Y},\mathbf{S}}$.

The rigorous mathematical minds will immediately question the validity of taking Step 3. In particular, the expression obtained for integer values may not be valid for real values in general [75]. In fact, the continuation of the expression to real values is not unique, e.g., $f(u) + \sin(\pi u)$ and $f(u)$ coincide at all integer u for every function f . Nevertheless, as we shall see, the the replica method simply takes the same expression derived for integer values of u , which is natural and straightforward in the problem at hand. The rigorous justification for Step 3 is still an open problem. Surprisingly, this continuation assumption, along with other assumptions—sometimes very intricate—on symmetries of solutions, if necessary (see Section 1.5.1), leads to correct results in all non-trivial cases where the results are known through other rigorous methods. In other cases, the replica method produces results that match well with numerical studies.

1.4.1.3 A Simple Example Before applying the replica method to the much more involved multiuser detection problem, we give a simple example of its application to the analysis of a single-user system. Let

$$\mathbf{Y} = \sqrt{\frac{\gamma}{L}} \mathbf{S}X + \mathbf{N} \quad (1.75)$$

where $X \sim p_X$, $\mathbf{N} \sim \mathcal{N}(0, \mathbf{I})$, and $\mathbf{S} = [S_1, \dots, S_L]^\top$ is a column vector with i.i.d. entries of mean 0 and unit variance. It is easy to see that the channel is equivalent (via matched filtering) to a single-user Gaussian channel with the same SNR. In the following we obtain the mutual information for $L \rightarrow \infty$ using the replica method as a warm-up exercise of the technique.

Similar to (1.69), conditioned on the channel state matrix, the input–output mutual information of (1.75) is

$$I(X; \mathbf{Y}|\mathbf{S}) = -\mathbb{E} \left\{ \log p_{\mathbf{Y}|\mathbf{S}}(\mathbf{Y}|\mathbf{S}) \right\} - \frac{L}{2} \log(2\pi e) \quad (1.76)$$

where $p_{\mathbf{Y}|\mathbf{S}}(\mathbf{y}|\mathbf{S}) = \mathbb{E} \left\{ p_{\mathbf{Y}|\mathbf{S},X}(\mathbf{y}|\mathbf{S}, X) \right\}$. In the following, we evaluate (1.76) using the replica method. The calculation is rather lengthy, while the outcome is quite simple.

The differential entropy can be obtained from

$$\mathbb{E} \{ \log p_{\mathbf{Y}|\mathbf{S}}(\mathbf{Y}|\mathbf{S}) \} = \lim_{u \rightarrow 0} \frac{\partial}{\partial u} \log \mathbb{E} \{ p_{\mathbf{Y}|\mathbf{S}}^u(\mathbf{Y}|\mathbf{S}) \}. \quad (1.77)$$

For any positive integer u , one can introduce u replicas of the original system, and evaluate the moment as follows:

$$\mathbb{E} \{ p_{\mathbf{Y}|\mathbf{S}}^u(\mathbf{Y}|\mathbf{S}) \} = \mathbb{E} \left\{ \int p_{\mathbf{Y}|\mathbf{S}}^{u+1}(\mathbf{y}|\mathbf{S}) \, d\mathbf{y} \right\} \quad (1.78)$$

$$= \mathbb{E} \left\{ \int \prod_{a=0}^u p_{\mathbf{Y}|\mathbf{S},X}(\mathbf{y}|\mathbf{S}, X_a) \, d\mathbf{y} \right\} \quad (1.79)$$

where the integral is over all entries of the vector \mathbf{y} from $-\infty$ to ∞ . Plugging in the Gaussian densities $p_{\mathbf{Y}|\mathbf{S},X}$, the RHS of (1.79) becomes $(2\pi)^{-(u+1)L/2}$ times

$$\begin{aligned} & \mathbb{E} \left\{ \int \prod_{a=0}^u \exp \left[-\frac{1}{2} \sum_{l=1}^L \left(y_l - \sqrt{\frac{\gamma}{L}} S_l X_a \right)^2 \right] \, d\mathbf{y} \right\} \\ &= \mathbb{E} \left\{ \mathbb{E} \left\{ \int \prod_{l=1}^L \exp \left[-\frac{1}{2} \sum_{a=0}^u \left(y_l - \sqrt{\frac{\gamma}{L}} S_l X_a \right)^2 \right] \, d\mathbf{y} \mid \mathbf{X} \right\} \right\} \end{aligned} \quad (1.80)$$

$$= \mathbb{E} \left\{ \left[\mathbb{E} \left\{ \int \exp \left[-\frac{1}{2} \sum_{a=0}^u \left(y - \sqrt{\frac{\gamma}{L}} S X_a \right)^2 \right] \, dy \mid \mathbf{X} \right\} \right]^L \right\} \quad (1.81)$$

where in (1.80) and (1.81) the inner expectation is with respect to the spreading chip(s) conditioned on the symbols \mathbf{X} , and (1.81) is due to symmetry and independence of the L chips. The integral in (1.81) is simply over a Gaussian density, which can be evaluated as

$$\begin{aligned} & \int \exp \left[-\frac{1}{2} \sum_{a=0}^u \left(y - \sqrt{\frac{\gamma}{L}} S X_a \right)^2 \right] \, dy \\ &= \sqrt{\frac{2\pi}{u+1}} \exp \left[\frac{\gamma}{2(u+1)L} \left(S \sum_{a=0}^u X_a \right)^2 - \frac{\gamma}{2L} \sum_{a=0}^u (S X_a)^2 \right]. \end{aligned} \quad (1.82)$$

By (1.81) and (1.82), the RHS of (1.79) becomes

$$\frac{(2\pi)^{-uL/2}}{(u+1)^{L/2}} \mathbb{E} \left\{ \left(\mathbb{E} \left\{ \exp \left[\frac{\gamma S^2 (\sum_{a=0}^u X_a)^2}{2(u+1)L} - \frac{\gamma S^2}{2L} \sum_{a=0}^u X_a^2 \right] \mid \mathbf{X} \right\} \right)^L \right\}. \quad (1.83)$$

Note that the exponent in (1.83) vanishes as $L \rightarrow \infty$. Using $\mathbb{E} \{ S^2 \} = 1$, we have

$$[(2\pi)^u (u+1)]^{\frac{1}{2}} \mathbb{E} \{ p_{\mathbf{Y}|\mathbf{S}}^u(\mathbf{Y}|\mathbf{S}) \} \rightarrow \mathbb{E} \left\{ \exp \left[\frac{\gamma (\sum_{a=0}^u X_a)^2}{2(u+1)} - \frac{\gamma}{2} \sum_{a=0}^u X_a^2 \right] \right\} \quad (1.84)$$

as $L \rightarrow \infty$. The RHS of (1.84) can be rearranged using the unit area property of Gaussian density:¹⁴

$$e^{x^2} = \sqrt{\frac{\eta}{2\pi}} \int \exp \left[-\frac{\eta}{2} z^2 + \sqrt{2\eta} x z \right] \, dz, \quad \forall x, \eta \quad (1.85)$$

¹⁴Equation (1.85) is a variant of the Hubbard-Stratonovich transform [76].

with $\eta = u + 1$ and $x = \sqrt{\frac{\gamma}{2(u+1)}} \sum_{a=0}^u X_a$. The RHS of (1.84) becomes

$$\begin{aligned} & \sqrt{\frac{u+1}{2\pi}} \mathbb{E} \left\{ \int \exp \left[-\frac{1}{2}(u+1)z^2 + \sqrt{\gamma}z \sum_{a=0}^u X_a - \frac{\gamma}{2} \sum_{a=0}^u X_a^2 \right] dz \right\} \\ &= \sqrt{\frac{u+1}{2\pi}} \int \left[\mathbb{E} \left\{ \exp \left[-\frac{1}{2}(z - \sqrt{\gamma}X)^2 \right] \right\} \right]^{u+1} dz. \end{aligned} \quad (1.86)$$

It is convenient to define a random variable $Z = \sqrt{\gamma}X + N$ where N is standard Gaussian. Let us define

$$p_{Z|X}(z|x) = \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{1}{2}(z - \sqrt{\gamma}x)^2 \right], \quad (1.87)$$

and

$$p_Z(z) = \frac{1}{\sqrt{2\pi}} \mathbb{E} \left\{ \exp \left[-\frac{1}{2}(z - \sqrt{\gamma}X)^2 \right] \right\}. \quad (1.88)$$

From (1.84) and (1.86),

$$[(2\pi)^u (u+1)]^{\frac{L-1}{2}} \mathbb{E} \left\{ p_{\mathbf{Y}|\mathbf{S}}^u(\mathbf{Y}|\mathbf{S}) \right\} \rightarrow \int p_Z^{u+1}(z) dz = \mathbb{E} \left\{ p_Z^u(Z) \right\}. \quad (1.89)$$

Therefore, from (1.76) and (1.89),

$$I(X; \mathbf{Y}|\mathbf{S}) = - \lim_{u \rightarrow 0} \frac{\partial}{\partial u} \log \mathbb{E} \left\{ p_{\mathbf{Y}|\mathbf{S}}^u(\mathbf{Y}|\mathbf{S}) \right\} - \frac{L}{2} \log(2\pi e) \quad (1.90)$$

$$\rightarrow - \int p_Z(z) \log p_Z(z) dz - \frac{1}{2} \log(2\pi e) \quad (1.91)$$

$$= h(Z) - h(N) \quad (1.92)$$

$$= I(X; \sqrt{\gamma}X + N). \quad (1.93)$$

It has thus been shown that in the large-dimension limit, the multi-dimensional channel (1.75) has the same mutual information as the scalar Gaussian channel with the same input and SNR as we initially expected.

1.4.2 Free Energy

In the remainder of this section, we present major steps of the replica analysis which lead to Claims 1–3. The outline of this development is as follows. We calculate the free energy (1.70) using (1.72) so that the spectral efficiency under joint decoding is immediate from (1.69). In Section 1.4.3, we show a sketch for calculating the joint moments, (1.71), which lead to the decoupling of the multiuser channel. Some of the calculations are tedious so we omit some details but provide enough clues and intuition so that the reader can connect the dots. For more details we refer the reader to [26, 27, 31].

For an arbitrary positive integer u , we introduce u independent replicas of the companion channel with the same received signal \mathbf{Y} and channel state \mathbf{S} as depicted in Figure 1.8.. The *partition function* of the replicated system, from which we evaluate the free energy (see Section 1.4.1.2), is

$$q_{\mathbf{Y}|\mathbf{S}}^u(\mathbf{y}|\mathbf{s}) = \mathbb{E}_q \left\{ \prod_{a=1}^u q_{\mathbf{Y}|X, \mathbf{S}}(\mathbf{y}|X_a, \mathbf{s}) \right\} \quad (1.94)$$

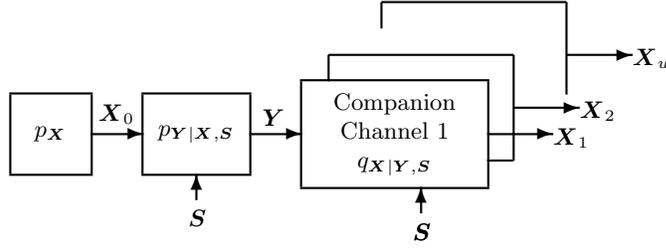


Figure 1.8. The replicas of the companion channel.

where the expectation is taken over the replicated variables $\{X_{ak}|a = 1, \dots, u, k = 1, \dots, K\}$. In (1.94), $X_{ak} \sim q_X$ are i.i.d. since $(\mathbf{Y}, \mathbf{S}) = (\mathbf{y}, \mathbf{s})$ are given. From (1.94),

$$\mathbb{E} \left\{ q_{\mathbf{Y}|\mathbf{S}}^u(\mathbf{Y}, \mathbf{S}) \right\} = \mathbb{E} \left\{ \int p_{\mathbf{Y}|\mathbf{X}, \mathbf{S}}(\mathbf{y}|\mathbf{X}_0, \mathbf{S}) \prod_{a=1}^u q_{\mathbf{Y}|\mathbf{X}, \mathbf{S}}(\mathbf{y}|\mathbf{X}_a, \mathbf{S}) d\mathbf{y} \right\} \quad (1.95)$$

where the expectations are taken over the channel state matrix \mathbf{S} , the original symbol vector \mathbf{X}_0 (i.i.d. entries with distribution p_X), and the replicated symbols $\mathbf{X}_a, a = 1, \dots, u$. For convenience, let $\sigma_0 = 1$ and $\sigma_a = \sigma$ for $a = 1, 2, \dots$. Plugging in (1.8) and (1.9), we have

$$\mathbb{E} \left\{ q_{\mathbf{Y}|\mathbf{S}}^u(\mathbf{Y}, \mathbf{S}) \right\} = \mathbb{E} \left\{ \int \frac{(2\pi\sigma^2)^{-\frac{uL}{2}}}{(2\pi)^{\frac{L}{2}}} \prod_{a=0}^u \exp \left[-\frac{\|\mathbf{y} - \mathbf{S}\mathbf{X}_a\|^2}{2\sigma_a^2} \right] d\mathbf{y} \right\}. \quad (1.96)$$

Note that \mathbf{S} and \mathbf{X}_a are independent in (1.96). Let $\underline{\mathbf{X}} = [\mathbf{X}_0, \dots, \mathbf{X}_u]$. The fact that the L dimensions of the multiuser channel are independent and statistically identical allows the RHS of (1.96) to be written as

$$\mathbb{E} \left\{ \left[(2\pi\sigma^2)^{-\frac{u}{2}} \int \mathbb{E} \left\{ \prod_{a=0}^u \exp \left[-\frac{(y - \overline{\mathbf{S}}\mathbf{A}\mathbf{X}_a)^2}{2\sigma_a^2} \right] \middle| \mathbf{A}\underline{\mathbf{X}} \right\} \frac{dy}{\sqrt{2\pi}} \right]^L \right\} \quad (1.97)$$

where the inner expectation is taken over $\overline{\mathbf{S}} = [S_1, \dots, S_K]$, a row vector of i.i.d. random variables each taking the same distribution as the random chips S_{nk} . It is clear that the original expectation over the growing chip dimension L is replaced by the fixed dimension u of the replicas.

Define the following variables:

$$V_a = \frac{1}{\sqrt{K}} \sum_{k=1}^K \sqrt{\gamma_k} S_k X_{ak}, \quad a = 0, 1, \dots, u. \quad (1.98)$$

Clearly, (1.97) can be rewritten as

$$\mathbb{E} \left\{ q_{\mathbf{Y}|\mathbf{S}}^u(\mathbf{Y}, \mathbf{S}) \right\} = \mathbb{E} \left\{ \exp \left[K G_K^{(u)}(\mathbf{A}\underline{\mathbf{X}}) \right] \right\} \quad (1.99)$$

where

$$G_K^{(u)}(\mathbf{A}\mathbf{X}) = -\frac{u}{2\beta} \log(2\pi\sigma^2) + \frac{1}{\beta} \log \int \mathbb{E} \left\{ \prod_{a=0}^u \exp \left[-\frac{(y - \sqrt{\beta} V_a)^2}{2\sigma_a^2} \right] \middle| \mathbf{A}\mathbf{X} \right\} \frac{dy}{\sqrt{2\pi}}. \quad (1.100)$$

Note that given \mathbf{A} and \mathbf{X} , each V_a is a sum of K weighted i.i.d. random chips. Due to a vector version of the central limit theorem, $\mathbf{V} = [V_0, V_1, \dots, V_u]^\top$ converges to a zero-mean Gaussian random vector as $K \rightarrow \infty$. For $a, b = 0, 1, \dots, u$, define

$$Q_{ab} = \mathbb{E} \{ V_a V_b \mid \mathbf{A}\mathbf{X} \} = \frac{1}{K} \sum_{k=1}^K \gamma_k X_{ak} X_{bk}. \quad (1.101)$$

Although implicit in notation, Q_{ab} is a function of $\{\gamma_k, X_{ak}, X_{bk}\}_{k=1}^K$. The random vector \mathbf{V} can essentially be replaced by a zero-mean Gaussian vector with covariance matrix $\mathbf{Q} = (1/K)\mathbf{X}^\top \mathbf{A}^2 \mathbf{X}$. As a result,

$$\exp \left[G_K^{(u)}(\mathbf{A}\mathbf{X}) \right] = \exp \left[G^{(u)}(\mathbf{Q}) + \mathcal{O}(K^{-1}) \right] \quad (1.102)$$

where the integral of the Gaussian density in (1.100) can be simplified to obtain

$$G^{(u)}(\mathbf{Q}) = -\frac{1}{2\beta} \log \det(\mathbf{I} + \mathbf{\Sigma}\mathbf{Q}) - \frac{1}{2\beta} \log \left(1 + \frac{u}{\sigma^2} \right) - \frac{u}{2\beta} \log(2\pi\sigma^2) \quad (1.103)$$

where $\mathbf{\Sigma}$ is a $(u+1) \times (u+1)$ matrix:¹⁵

$$\mathbf{\Sigma} = \frac{\beta}{\sigma^2 + u} \begin{bmatrix} u & & -\mathbf{e}^\top \\ \hline -\mathbf{e} & (1 + \frac{u}{\sigma^2}) \mathbf{I} - \frac{1}{\sigma^2} \mathbf{e}\mathbf{e}^\top & \end{bmatrix} \quad (1.104)$$

where \mathbf{e} is a $u \times 1$ column vector whose entries are all 1. It is clear that $\mathbf{\Sigma}$ is invariant if two nonzero indexes are interchanged, i.e., $\mathbf{\Sigma}$ is symmetric in the replicas.

By (1.99) and (1.102),

$$\frac{1}{K} \log \mathbb{E} \left\{ q_{\mathbf{Y}|\mathbf{S}}^u(\mathbf{Y}, \mathbf{S}) \right\} = \frac{1}{K} \log \mathbb{E} \left\{ \exp \left[K \left(G^{(u)}(\mathbf{Q}) + \mathcal{O}(K^{-1}) \right) \right] \right\} \quad (1.105)$$

$$= \frac{1}{K} \log \int \exp \left[K G^{(u)}(\mathbf{Q}) \right] d\mu_K^{(u)}(\mathbf{Q}) + \mathcal{O}\left(\frac{1}{K}\right) \quad (1.106)$$

where the expectation over the replicated symbols is rewritten as an integral over the probability measure of the covariance matrix \mathbf{Q} .

1.4.2.1 Large Deviations and Saddle Point Since Q_{ab} given by (1.101) is a sum of independent random variables for each pair (a, b) , the probability measure $\mu_K^{(u)}$ satisfies the large deviations property. By Cramér's Theorem [77, Theorem II.4.1], there exists a rate function $I^{(u)}$ such that the measure $\mu_K^{(u)}$ satisfies

$$-\lim_{K \rightarrow \infty} \frac{1}{K} \log \mu_K^{(u)}(\mathcal{A}) = \inf_{\mathbf{Q} \in \mathcal{A}} I^{(u)}(\mathbf{Q}) \quad (1.107)$$

¹⁵The indexes of all $(u+1) \times (u+1)$ matrices in this chapter start from 0.

for all measurable sets \mathcal{A} of $(u+1) \times (u+1)$ matrices.

Let the moment generating function be defined as

$$M^{(u)}(\tilde{\mathbf{Q}}) = \mathbb{E} \left\{ \exp \left[\gamma \mathbf{X}^\top \tilde{\mathbf{Q}} \mathbf{X} \right] \right\} \quad (1.108)$$

where $\tilde{\mathbf{Q}}$ is a $(u+1) \times (u+1)$ symmetric matrix, $\mathbf{X} = [X_0, X_1, \dots, X_u]^\top$, and the expectation in (1.108) is taken over independent random variables $\gamma \sim P_\gamma$, $X_0 \sim p_X$ and $X_1, \dots, X_u \sim q_X$. The rate of the measure $\mu_K^{(u)}$ is given by the Legendre-Fenchel transform of the cumulant generating function [77]:

$$I^{(u)}(\mathbf{Q}) = \sup_{\tilde{\mathbf{Q}}} \left[\text{tr} \left\{ \tilde{\mathbf{Q}} \mathbf{Q} \right\} - \log M^{(u)}(\tilde{\mathbf{Q}}) \right] \quad (1.109)$$

where the supremum is taken with respect to the symmetric matrix $\tilde{\mathbf{Q}}$.

As the exponential factor in (1.106) is proportional to K , and since we are taking the limit $K \rightarrow \infty$, the integral is dominated by the maximum of the overall effect of the exponent and the rate of the measure on which the integral takes place (the saddle-point method). Precisely, by Varadhan's theorem [77, Theorem II.7.1], the free energy for a given replica number u is

$$\mathcal{F}_u = - \lim_{K \rightarrow \infty} \frac{1}{K} \log \mathbb{E} \left\{ q_{\mathbf{Y}|\mathbf{S}}^u(\mathbf{Y}, \mathbf{S}) \right\} = - \sup_{\mathbf{Q}} \left[G^{(u)}(\mathbf{Q}) - I^{(u)}(\mathbf{Q}) \right] \quad (1.110)$$

where the supremum is over all valid covariance matrices. Plugging in (1.103), (1.108) and (1.109),

$$\mathcal{F}_u = \inf_{\mathbf{Q}} \sup_{\tilde{\mathbf{Q}}} T^{(u)}(\mathbf{Q}, \tilde{\mathbf{Q}}), \quad (1.111)$$

with

$$\begin{aligned} T^{(u)}(\mathbf{Q}, \tilde{\mathbf{Q}}) &= \frac{1}{2\beta} \log \det(\mathbf{I} + \Sigma \mathbf{Q}) + \text{tr} \left\{ \tilde{\mathbf{Q}} \mathbf{Q} \right\} - \log \mathbb{E} \left\{ \exp \left[\gamma \mathbf{X}^\top \tilde{\mathbf{Q}} \mathbf{X} \right] \right\} \\ &+ \frac{1}{2\beta} \log \left(1 + \frac{u}{\sigma^2} \right) + \frac{u}{2\beta} \log (2\pi\sigma^2). \end{aligned} \quad (1.112)$$

For an arbitrary \mathbf{Q} , we first seek the point of zero gradient with respect to $\tilde{\mathbf{Q}}$ and find that for any given \mathbf{Q} , the extremum in $\tilde{\mathbf{Q}}$ satisfies

$$\tilde{\mathbf{Q}} = \frac{\mathbb{E} \left\{ \gamma \mathbf{X} \mathbf{X}^\top \exp \left[\gamma \mathbf{X}^\top \tilde{\mathbf{Q}} \mathbf{X} \right] \right\}}{\mathbb{E} \left\{ \exp \left[\gamma \mathbf{X}^\top \tilde{\mathbf{Q}} \mathbf{X} \right] \right\}}. \quad (1.113)$$

Let $\tilde{\mathbf{Q}}^*(\mathbf{Q})$ denote the solution to (1.113). We then seek the point of zero gradient of $T^{(u)}(\mathbf{Q}, \tilde{\mathbf{Q}}^*(\mathbf{Q}))$ with respect to \mathbf{Q} .¹⁶ By virtue of the zero-gradient condition

¹⁶The following identities are useful:

$$\frac{\partial \log \det \mathbf{Q}}{\partial x} = \text{tr} \left\{ \mathbf{Q}^{-1} \frac{\partial \mathbf{Q}}{\partial x} \right\}, \quad \frac{\partial \mathbf{Q}^{-1}}{\partial x} = -\mathbf{Q}^{-1} \frac{\partial \mathbf{Q}}{\partial x} \mathbf{Q}^{-1}.$$

with respect to $\tilde{\mathbf{Q}}$, one finds that the derivative of $\tilde{\mathbf{Q}}^*(\mathbf{Q})$ with respect to \mathbf{Q} is multiplied by 0 and hence inconsequential. Therefore, the extremum in \mathbf{Q} satisfies

$$\tilde{\mathbf{Q}} = -\beta^{-1} (\mathbf{I} + \Sigma \mathbf{Q})^{-1} \Sigma. \quad (1.114)$$

It is interesting to note from the resulting joint equations (1.113)–(1.114) that the order in which the supremum and infimum are taken in (1.111) can be exchanged. The solution $(\mathbf{Q}^*, \tilde{\mathbf{Q}}^*)$ is in fact a saddle point of $T^{(u)}$. Notice that (1.113) can also be expressed as

$$\mathbf{Q} = \mathbb{E} \left\{ \gamma \mathbf{X} \mathbf{X}^\top \mid \tilde{\mathbf{Q}} \right\} \quad (1.115)$$

where the expectation is over an appropriately defined conditional measure $P_\gamma \times p_{\mathbf{X}|\tilde{\mathbf{Q}},\gamma}$ where

$$p_{\mathbf{X}|\tilde{\mathbf{Q}},\gamma}(\mathbf{x}|\tilde{\mathbf{Q}},\gamma) = p_{\mathbf{X}}(\mathbf{x}) \frac{\exp[\gamma \mathbf{x}^\top \tilde{\mathbf{Q}} \mathbf{x}]}{\mathbb{E} \left\{ \exp[\gamma \mathbf{X}^\top \tilde{\mathbf{Q}} \mathbf{X}] \right\}} \quad (1.116)$$

which is evidently a pdf or pmf. Let $\mathbf{Q}^*(u)$ and $\tilde{\mathbf{Q}}^*(u)$ be the solution to (1.113)–(1.114) as functions of u . The free energy is then found by (1.72) and (1.111).

1.4.2.2 Replica Symmetry Solution Solving joint equations (1.113)–(1.114) directly is prohibitive except in the simplest cases such as q_X being Gaussian. In the general case, suggested by the symmetry in the matrix Σ (1.104), we postulate that the solution to the joint equations satisfies *replica symmetry*, namely, both $\mathbf{Q}^*(u)$ and $\tilde{\mathbf{Q}}^*(u)$ are invariant if two (nonzero) replica indexes are interchanged. In other words, the extremum can be written as

$$\mathbf{Q}^*(u) = \begin{bmatrix} r & m & m & \dots & m \\ m & p & q & \dots & q \\ m & q & p & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & q \\ m & q & \dots & q & p \end{bmatrix}, \quad \tilde{\mathbf{Q}}^*(u) = \begin{bmatrix} c & d & d & \dots & d \\ d & g & f & \dots & f \\ d & f & g & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & f \\ d & f & \dots & f & g \end{bmatrix} \quad (1.117)$$

where r, m, p, q, c, d, f, g are some real functions of u . The validity of the replica symmetry assumption is discussed in Section 1.5.1. Under this symmetry assumption, the problem of seeking the extremum (1.111) over a $(u+1)^2$ -dimensional space (with u also a variable) is reduced to seeking the extremum over several parameters.

The eight parameters (r, m, p, q, c, d, f, g) can be solved from the joint equations (1.113)–(1.114) under replica symmetry assumption. The detailed calculation is omitted. It is interesting to note that the u -dependence of the parameters obtained from the joint equations (1.113)–(1.114) does not contribute to the free energy (1.111) due to the zero gradient conditions. Thus for the purpose of the free energy (1.111), it suffices to find the derivative in (1.111) with $\mathbf{Q}^*(u)$ and $\tilde{\mathbf{Q}}^*(u)$ replaced by their values at $u = 0$, which we simply denote by \mathbf{Q}^* and $\tilde{\mathbf{Q}}^*$. From this point on, with slight abuse of notation, let r, m, p, q, c, d, f, g represent their values at $u = 0$.

Using (1.114) and (1.117), it can be shown that at $u = 0$,

$$c = 0, \quad (1.118a)$$

$$d = \frac{1}{2[\sigma^2 + \beta(p - q)]}, \quad (1.118b)$$

$$f = \frac{1 + \beta(r - 2m + q)}{2[\sigma^2 + \beta(p - q)]^2}, \quad (1.118c)$$

$$g = f - d. \quad (1.118d)$$

The parameters r, m, p, q can be determined from (1.115) by studying the measure $p_{\mathbf{X}, \gamma | \tilde{\mathbf{Q}}}$ under replica symmetry and $u \rightarrow 0$. For that purpose, define two useful parameters with a modest amount of foresight:

$$\eta = \frac{2d^2}{f} \quad \text{and} \quad \xi = 2d. \quad (1.119)$$

The moment generating function (1.108) is evaluated using the property (1.85) with $\eta = 2d^2/f$ and noticing that $c = 0$, $g - f = -d$ to obtain

$$\begin{aligned} M^{(u)}(\tilde{\mathbf{Q}}^*) &= \mathbb{E} \left\{ \sqrt{\frac{\eta}{2\pi}} \int \exp \left[-\frac{\eta}{2} (z - \sqrt{\gamma} X_0)^2 \right] \right. \\ &\quad \times \left. \left[\mathbb{E}_q \left\{ \exp \left[-\frac{\xi}{2} z^2 - \frac{\xi}{2} (z - \sqrt{\gamma} X)^2 \right] \middle| \gamma \right\} \right]^u dz \right\}. \end{aligned} \quad (1.120)$$

It is clear that the limit of (1.120) as $u \rightarrow 0$ is 1, i.e.,

$$\lim_{u \rightarrow 0} \mathbb{E} \left\{ \exp \left[\gamma \mathbf{X}^\top \tilde{\mathbf{Q}}^* \mathbf{X} \right] \right\} = 1. \quad (1.121)$$

Hence (1.113) implies that, as $u \rightarrow 0$, the limit of $Q_{ab}^* = \mathbb{E} \left\{ \gamma X_a X_b \middle| \tilde{\mathbf{Q}}^* \right\}$ is identical to

$$\lim_{u \rightarrow 0} \mathbb{E} \left\{ \gamma X_a X_b \exp \left[\gamma \mathbf{X}^\top \tilde{\mathbf{Q}}^* \mathbf{X} \right] \right\}. \quad (1.122)$$

We apply the transform (1.85) to decouple the cross terms of the form $X_c X_d$ in the exponent in (1.122). In fact all terms unrelated to X_a and X_b integrate to 1 which do not contribute to the limit. More details are found in [27].

1.4.2.3 Single-user Channel Interpretation We now give a useful representation for the parameters r, m, p, q defined in (1.117). Consider $a = 0$ and $b = 1$ for instance. Expanding (1.122), as $u \rightarrow 0$,

$$Q_{01}^* = \mathbb{E} \left\{ \gamma X_0 X_1 \exp \left[\gamma \mathbf{X}^\top \tilde{\mathbf{Q}}^* \mathbf{X} \right] \right\} \quad (1.123)$$

$$\rightarrow \mathbb{E} \left\{ \gamma X_0 \int \sqrt{\frac{\eta}{2\pi}} \exp \left[-\frac{\eta}{2} (z - \sqrt{\gamma} X_0)^2 \right] \right. \quad (1.124)$$

$$\left. \times \frac{X_1 \sqrt{\frac{\xi}{2\pi}} \exp \left[-\frac{\xi}{2} (z - \sqrt{\gamma} X_1)^2 \right]}{\mathbb{E}_q \left\{ \sqrt{\frac{\xi}{2\pi}} \exp \left[-\frac{\xi}{2} (z - \sqrt{\gamma} X_1)^2 \right] \middle| \gamma \right\}} dz \right\}. \quad (1.125)$$

Let two single-user Gaussian channels be defined as in Section 1.3.2, i.e., the input–output relationship of the two channels are described by $p_{Z|X,\gamma;\eta}$ given by (1.18) and $q_{Z|X,\gamma;\xi}$ by (1.40). Assuming that the input distribution to the channel $q_{Z|X,\gamma;\xi}$ is q_X , a posterior probability distribution $q_{X|Z,\gamma;\xi}$ is induced, which defines a companion channel. Let X_0 be the scalar input to the channel $p_{Z|X,\gamma;\eta}$ and $X = X_1$ be the output of the companion channel $q_{X|Z,\gamma;\xi}$. The posterior mean with respect to the measure q , denoted by $\langle X \rangle_q$, is given by (1.41). The Gaussian channel $p_{Z|X,\gamma;\eta}$, the companion channel $q_{X|Z,\gamma;\xi}$ and the PME, all in the single-user setting, are depicted in Figure 1.6.(b). Then, (1.125) can be understood as an expectation over X_0 , X and Z to obtain

$$\mathbb{E} \left\{ \gamma X_0 \int \mathbb{E}_q \{ X | Z = z, \gamma; \xi \} p_{Z|X,\gamma;\eta}(z|X_0, \gamma; \eta) dz \right\} = \mathbb{E} \left\{ \gamma X_0 \langle X \rangle_q \right\}. \quad (1.126)$$

Similarly, (1.122) can be evaluated for all (a, b) yielding together with (1.117):

$$r = \lim_{u \rightarrow 0} Q_{00}^* = \mathbb{E} \left\{ \gamma X_0^2 \right\} = \mathbb{E} \left\{ \gamma \right\}, \quad (1.127a)$$

$$m = \lim_{u \rightarrow 0} Q_{01}^* = \mathbb{E} \left\{ \gamma X_0 \langle X \rangle_q \right\}, \quad (1.127b)$$

$$p = \lim_{u \rightarrow 0} Q_{11}^* = \mathbb{E} \left\{ \gamma X^2 \right\}, \quad (1.127c)$$

$$q = \lim_{u \rightarrow 0} Q_{12}^* = \mathbb{E} \left\{ \gamma (\langle X \rangle_q)^2 \right\}. \quad (1.127d)$$

In summary, the parameters c, d, f, g are given by (1.118) as functions of r, m, p, q , which are in turn determined by the statistics of the two channels (1.18) and (1.40) parameterized by $\eta = 2d^2/f$ and $\xi = 2d$ respectively. It is not difficult to see that

$$r - 2m + q = \mathbb{E} \left\{ \gamma \left(X_0 - \langle X \rangle_q \right)^2 \right\}, \quad (1.128a)$$

$$p - q = \mathbb{E} \left\{ \gamma \left(X - \langle X \rangle_q \right)^2 \right\}. \quad (1.128b)$$

Using (1.118) and (1.119), it can be checked that

$$r - 2m + q = \frac{1}{\beta} \left(\frac{1}{\eta} - 1 \right), \quad \text{and} \quad p - q = \frac{1}{\beta} \left(\frac{1}{\xi} - \sigma^2 \right). \quad (1.129)$$

Under replica symmetry, $G^{(u)}(\mathbf{Q}^*)$ is evaluated using (1.103) and expressed in η and ξ . Together with (1.111) and (1.120), the free energy is found as (1.47), where by (1.128) and (1.129), (η, ξ) satisfies

$$\eta^{-1} = 1 + \beta \mathbb{E} \left\{ \gamma \left(X_0 - \langle X \rangle_q \right)^2 \right\}, \quad (1.130a)$$

$$\xi^{-1} = \sigma^2 + \beta \mathbb{E} \left\{ \gamma \left(X - \langle X \rangle_q \right)^2 \right\}. \quad (1.130b)$$

Because of the saddle-point evaluation in (1.110), in case of multiple solutions to (1.130), (η, ξ) is chosen as the solution that gives the minimum free energy \mathcal{F} . By defining $\mathcal{E}(\gamma; \eta, \xi)$ and $\mathcal{V}(\gamma; \eta, \xi)$ as in (1.43) and (1.44), the coupled equations

(1.118) and (1.127) can be summarized to establish the key fixed-point equations (1.46). It will be shown in Section 1.4.3 that, from an individual user's viewpoint, the multiuser PME and the multiuser companion channel, parameterized by arbitrary (q_X, σ) , have an equivalence as a single-user PME and a single-user companion channel.

1.4.2.4 Spectral Efficiency and Multiuser Efficiency Finally, for the purpose of the total spectral efficiency, we set the postulated measure q to be identical to the actual measure p (i.e., $(q_X, \sigma) = (p_X, 1)$). The inverse noise variances (η, ξ) satisfy joint equations but we choose the replica-symmetric solution $\eta = \xi$. Using the identity

$$C_{\text{joint}} = \beta \mathcal{F}|_{q=p} - \frac{1}{2} \log(2\pi e), \quad (1.131)$$

the total spectral efficiency is

$$C_{\text{joint}} = -\beta \mathbf{E} \left\{ \int p_{Z|\gamma;\eta}(z|\gamma;\eta) \log p_{Z|\gamma;\eta}(z|\gamma;\eta) dz \right\} - \frac{\beta}{2} \log \frac{2\pi e}{\eta} + \frac{1}{2}(\eta - 1 - \log \eta), \quad (1.132)$$

where η satisfies

$$\eta + \eta \beta \mathbf{E} \left\{ \gamma \left[1 - \int \frac{[p_1(z, \gamma; \eta)]^2}{p_{Z|\gamma;\eta}(z|\gamma;\eta)} dz \right] \right\} = 1. \quad (1.133)$$

The optimal spectral efficiency of the multiuser channel is thus found.

We remark that the essence of the replica method here is its capability of converting a difficult expectation (e.g., of a logarithm) with respect to a given large system to an expectation of a simpler form with respect to the replicated system. Quite different from conventional techniques is the emphasis of large systems and symmetry from the beginning, where the central limit theorem and large deviations help to calculate the otherwise intractable quantities.

1.4.3 Joint Moments

Consider the multiuser Gaussian channel, the PME and the companion channel depicted in Figure 1.6.(a). The joint moments (1.71) are of interest here. For simplicity, we first study joint moments of the input and the companion channel output, which can be obtained as expectations under the replicated system [31, Lemma 3.1]:

$$\mathbf{E} \left\{ X_{0k}^i X_k^j \right\} = \mathbf{E} \left\{ X_{0k}^i X_{mk}^j \right\}, \quad m = 1, \dots, u. \quad (1.134)$$

It is then straightforward to calculate (1.71) by following the same procedure.

In [27], it is shown that the moments (1.134) can be obtained as

$$\lim_{u \rightarrow 0} \frac{\partial}{\partial h} \frac{1}{\alpha_1 K} \log \mathbf{E} \left\{ Z^{(u)}(\mathbf{Y}, \mathbf{S}, \mathbf{X}_0; h) \right\} \Big|_{h=0} \quad (1.135)$$

where $\alpha_1 \in (0, 1)$ and

$$Z^{(u)}(\mathbf{y}, \mathbf{s}, \mathbf{x}_0; h) = \mathbf{E}_q \left\{ \exp \left[h \sum_{k=1}^{K_1} x_{0k}^i X_{mk}^j \right] \prod_{a=1}^u \exp \left[-\frac{\|\mathbf{y} - \mathbf{s} \mathbf{X}_a\|^2}{2\sigma^2} \right] \right\} \quad (1.136)$$

where $K_1 = \alpha_1 K$ and we assume that the SNRs of the first K_1 users are equal to γ . Regarding (1.136) as a partition function for some random system allows the same techniques in Section 1.4.2 to be used to write

$$\lim_{K \rightarrow \infty} \frac{1}{K} \log \mathbb{E} \left\{ Z^{(u)}(\mathbf{Y}, \mathbf{S}, \mathbf{X}_0; h) \right\} = \sup_{\mathbf{Q}} \left[\beta^{-1} G^{(u)}(\mathbf{Q}) - I^{(u)}(\mathbf{Q}; h) \right] \quad (1.137)$$

where $G^{(u)}(\mathbf{Q})$ is given by (1.103) and the rate $I^{(u)}(\mathbf{Q}; h)$ is found as

$$I^{(u)}(\mathbf{Q}; h) = \sup_{\tilde{\mathbf{Q}}} \left[\text{tr} \left\{ \tilde{\mathbf{Q}} \mathbf{Q} \right\} - (1 - \alpha_1) \log M^{(u)}(\tilde{\mathbf{Q}}) - \alpha_1 \log M^{(u)}(\tilde{\mathbf{Q}}, \gamma; h) \right] \quad (1.138)$$

where $M^{(u)}(\tilde{\mathbf{Q}})$ is defined in (1.108), and

$$M^{(u)}(\tilde{\mathbf{Q}}, \gamma; h) = \mathbb{E} \left\{ \exp \left[h X_0^i X_m^j \right] \exp \left[\gamma \mathbf{X}^\top \tilde{\mathbf{Q}} \mathbf{X} \right] \middle| \gamma \right\}. \quad (1.139)$$

From (1.137) and (1.138), taking the derivative in (1.135) with respect to h at $h = 0$ leaves only one term

$$\left. \frac{\partial}{\partial h} \log M^{(u)}(\tilde{\mathbf{Q}}, \gamma; h) \right|_{h=0} = \frac{\mathbb{E} \left\{ X_0^i X_m^j \exp \left[\gamma \mathbf{X}^\top \tilde{\mathbf{Q}} \mathbf{X} \right] \middle| \gamma \right\}}{\mathbb{E} \left\{ \exp \left[\gamma \mathbf{X}^\top \tilde{\mathbf{Q}} \mathbf{X} \right] \middle| \gamma \right\}}. \quad (1.140)$$

Since

$$Z^{(u)}(\mathbf{Y}, \mathbf{S}, \mathbf{X}_0; h) \Big|_{h=0} = q_{\mathbf{Y}|\mathbf{S}}^u(\mathbf{Y}, \mathbf{S}), \quad (1.141)$$

the $\tilde{\mathbf{Q}}^*$ which satisfies (1.140) and gives the supremum in (1.138) at $h \rightarrow 0$ is exactly the $\tilde{\mathbf{Q}}^*$ which gives the supremum of (1.109), which is replica-symmetric by assumption. By introducing the parameters (η, ξ) as in Section 1.4.2, and by definition of q_i and p_i in (1.42) and (1.26) respectively, (1.140) can be further evaluated as

$$\frac{\int \left(\sqrt{\frac{2\pi}{\xi}} e^{-\frac{\xi z^2}{2}} \right)^u p_i(z, \gamma; \eta) q_0^{u-1}(z, \gamma; \xi) q_j(z, \gamma; \xi) dz}{\int \left(\sqrt{\frac{2\pi}{\xi}} e^{-\frac{\xi z^2}{2}} \right)^u p_0(z, \gamma; \eta) q_0^u(z, \gamma; \xi) dz}. \quad (1.142)$$

Taking the limit $u \rightarrow 0$, one has from (1.134)–(1.142) that as $K \rightarrow \infty$,

$$\frac{1}{K_1} \sum_{k=1}^{K_1} \mathbb{E} \left\{ X_{0k}^i X_{mk}^j \right\} \rightarrow \int p_i(z, \gamma; \eta) \frac{q_j(z, \gamma; \xi)}{q_0(z, \gamma; \xi)} dz. \quad (1.143)$$

Let $X_0 \sim p_X$ be the input to the scalar Gaussian channel $p_{Z|X, \gamma; \eta}$ and Z be its output (see Figure 1.6.(b)). Let X be the output of the companion channel with Z as its input. Then $X_0 - Z - X$ is a Markov chain. The RHS of (1.143) is

$$\int p_0(z, \gamma; \eta) \frac{p_i(z, \gamma; \xi)}{p_0(z, \gamma; \xi)} \frac{q_j(z, \gamma; \xi)}{q_0(z, \gamma; \xi)} dz = \mathbb{E} \left\{ \mathbb{E} \left\{ X_0^i \middle| Z \right\} \mathbb{E} \left\{ X^j \middle| Z \right\} \right\}. \quad (1.144)$$

Letting $K_1 \rightarrow 1$ (thus $\alpha_1 \rightarrow 0$)¹⁷ so that the requirement that the first K_1 users take the same SNR becomes unnecessary, we have shown by (1.134), (1.143) and

¹⁷To be precise, this step requires a more delicate treatment, since the saddle-point evaluation involved in our calculation only captures terms of order $O(K)$ in the exponent. It has been shown that the result remains the same even if we take $O(1)$ terms into consideration [78].

(1.144) that for every SNR distribution and every user $k \in \{1, \dots, K\}$

$$\mathbb{E} \left\{ X_{0k}^i X_k^j \right\} \rightarrow \mathbb{E} \left\{ X_0^i X^j \right\} \quad \text{as } K \rightarrow \infty. \quad (1.145)$$

We assume that the joint distribution $P_{X_{0k} X_k}$ is determinate, i.e., uniquely determined by the joint moments¹⁸. Therefore, for every user k , the joint distribution of the input X_{0k} to the multiuser channel and the output X_k of the multiuser companion channel converges to the joint distribution of the input X_0 to the single-user Gaussian channel $p_{Z|X, \gamma; \eta}$ and the output X of the single-user companion channel $q_{X|Z, \gamma; \xi}$.

Applying the same methodology as developed thus far, one can also calculate the joint moments (1.71) to obtain¹⁹

$$\mathbb{E} \left\{ X_{0k}^i X_k^j \langle X_k \rangle_q^l \right\} \rightarrow \mathbb{E} \left\{ X_0^i X^j \langle X \rangle_q^l \right\} \quad (1.146)$$

where $\langle X \rangle_q$ is the single-user PME output as seen in Figure 1.6.(b), which is a function of the Gaussian channel output Z . Again, assuming the determinacy, the joint distributions of $(X_{0k}, X_k, \langle X_k \rangle_q)$ converge to that of $(X_0, X, \langle X \rangle_q)$. Indeed, from the viewpoint of user k , the multiuser setting is equivalent to the single-user setting in which the SNR suffers a degradation η (compare Figures 1.6.(b) and 1.6.(a)). Hence we have justified the decoupling principle and Claim 3.

In the large-system limit, the transformation from the input X_{0k} to the multiuser detection output $\langle X_k \rangle_q$ is nothing but a single-user Gaussian channel $p_{Z|X, \gamma; \eta}$ concatenated with a decision function (1.41). The decision function is one-to-one due to Proposition 1 and hence inconsequential from both detection- and information-theoretic viewpoints.

We now conclude that the equivalent single-user channel is an additive Gaussian noise channel with input SNR γ and noise variance η^{-1} as depicted in Figure 1.6.(b). Claim 3 follows, which implies Claims 1 and 2 in the special case that the postulated measure q is identical to the actual measure p . Curiously, this decoupling result is identical to what is obtained using parallel interference cancellation in Section 1.3.4 with invalid independence assumption, or using belief propagation in the special case of sparse spreading matrix [52].

1.5 FURTHER DISCUSSION

1.5.1 On Replica symmetry

The validity of the replica symmetry assumption can be checked by calculating the Hessian of $[G^{(u)}(\mathbf{Q}) - I^{(u)}(\mathbf{Q})]$ at the replica symmetric supremum [37]. If all the eigenvalues of the Hessian associated with modes that break replica symmetry are negative at the replica symmetric supremum, then the solution is stable against

¹⁸Note that the determinacy does not necessarily hold in general (the moment problem [79, p. 227], [80]), even though all distributions of finite support and most discrete and continuous distributions of practical interest (e.g., Gaussian distribution) are determinate. Sufficient conditions for a multidimensional distribution to be determinate are given in [81, 82]. In particular, if the marginals are determinate, the joint distribution is also determinate [81].

¹⁹Note the change of notation: X is replaced by X_0 which corresponds to the 0-th replica and X' is replaced by X .

perturbations which break replica symmetry. If not, then the solution is said to suffer from the de Almeida-Thouless (AT) instability, which has been named after two physicists who first performed the stability analysis on a spin glass model [83].

For the basic prescription of the AT stability analysis, we ask the reader to see [26], where one will find detailed description of the analysis for the equal-power binary input case. Essentially the same analysis can be performed in the generic case discussed in this chapter, and the result is summarized as follows.

Claim 4 (*AT stability*) *A replica-symmetric solution is stable against replica symmetry breaking (RSB) if the following inequality holds.*

$$-\beta\xi^2 + \left[\mathbb{E} \left\{ \gamma (\langle X^2 \rangle_q - (\langle X \rangle_q)^2) \right\} \right]^{-1} < 0 \quad (1.147)$$

The LHS of (1.147) is the eigenvalue of the perturbation modes (the so-called “replicon” modes) that determines the AT stability. One can numerically check the AT stability condition in order to see if a particular numerical solution of the replica symmetric fixed-point equations (1.46) is stable against replica symmetry breaking. In the equal-power binary case, AT instability may actually be observed, although not always, when the postulated noise level σ is less than 1.

When the AT stability is violated for a solution with replica symmetry, it means that \mathbf{Q} at the true supremum (1.110) should lack the replica symmetry, and \mathbf{Q} with broken symmetry will give us even larger values of $[G^{(u)} - I^{(u)}]$. A systematic way of improving replica-symmetric solutions has been proposed for spin glass models, which consists of considering a series of symmetry breaking schemes, the so-called 1-step RSB, 2-step RSB, etc. A preliminary study on the equal-power binary CDMA problem suggests that consideration of the 1-step RSB alters the solutions only slightly [84]. We thus expect that the analysis with replica symmetry assumption provides us with quantitatively accurate enough picture even if the assumption is not valid.

1.5.2 On Metastable Solutions

As we have briefly mentioned in Section 1.4.2.3, the fixed-point equations (1.46) determining (η, ξ) may have multiple solutions. Since we are interested in obtaining the true supremum of $[G^{(u)} - I^{(u)}]$, what we have to do is to compare the values of the free energy \mathcal{F} of those solutions, and to pick up the one that minimizes the free energy. Then we can safely discard the other solutions, since they seem not to have any operational meaning. Or, do they?

They do, if practical (suboptimal) schemes for obtaining the PME are considered. In order to understand benefits of considering solutions other than the one giving the true supremum, it should be a good idea to exploit an analogy with a ‘magnet.’ Typical magnetic materials respond to externally applied magnetic field by expressing magnetization. Magnetization curves (Figure 1.9.(a)) represent how the magnetization depends on the external magnetic field. At high enough temperature the magnetization depends monotonically on, and uniquely determined by, the external field (Figure 1.9.(a), monotone curve). On the other hand, at low temperature (lower than the so-called Curie temperature), the magnetization may take multiple values within a certain range of the external field (Figure 1.9.(a), the S-shaped curve). In particular, non-zero magnetization will be observed even

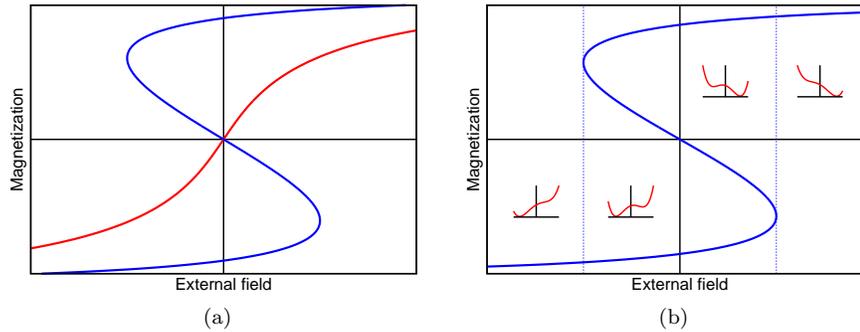


Figure 1.9. Typical magnetization curves of ferromagnets. (a) Magnetization curve at high (monotone curve) and low (S-shaped curve) temperatures. (b) Magnetization curve at low temperature. Insets show the free energy profiles in the corresponding regions.

when the external field is absent. It is called spontaneous magnetization, and is the theory underlying the phenomenon such as seen about a magnet. The structural change of the magnetization curve with temperature is an example of the phase transition.

A simple mathematical model of magnetism can explain the phase transition analytically. The analysis consists of evaluating the free energy in essentially the same manner as the analysis in this chapter, that is, with the fixed-point method. At the low-temperature (“ferromagnetic”) phase, the fixed-point equations may have three solutions. When it is the case, the free energy, as a function of magnetization, has a structure as shown in the insets of Figure 1.9.(b). Thus the true solution in the mathematical sense is the one that minimizes the free energy, and is given by the topmost and the lowermost branches in Figure 1.9.(b) when the external field is positive and negative, respectively. The solution is called a (globally) stable solution. The solution in the topmost branch with negative external fields, and the one in the lowermost branch with positive external fields, only locally minimizes the free energy. They are called “metastable” solutions. The last solution in the middle branch, called an unstable solution, locally maximizes the free energy.

The significance of the metastable solutions is manifested in a phenomenon called *hysteresis*. In the low-temperature condition, the magnetization we observe may not be the one corresponding to the true solution (the globally stable solution). Depending on history, the system may take a state corresponding to a metastable solution, the fact that explains the hysteresis.

Essentially the same description applies to the multiuser detection problem as well, as can be seen in Figure 1.10. Whereas the globally optimum solution switches from the uppermost branch to the lowermost one, the numerical results obtained using suboptimal belief propagation algorithms (see Section 1.7.2) seem to follow the uppermost branch even though it is not the true solution on the shoulder at the RHS of the S-shaped curve. The phenomenon can be ascribed to the fact that the multiuser detection algorithm does not know the true detection results initially: Indeed, the system may easily get trapped in the metastable solution with large error probability, due to the “history” effect caused by the initial configurations being far away from the true detection results.

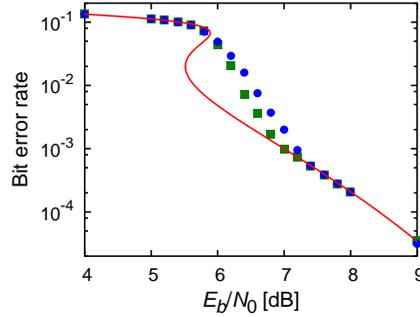


Figure 1.10. Comparison of BER between replica analysis of the individually optimal detector (solid curve) and numerical experiments on the BP-based multiuser detection algorithm described in Section 1.7.2.3 (dots: $L = 2,000$, squares: $L = 4,000$). System load β is 1.6. BPSK data modulation, as well as perfect power control is assumed.

1.6 STATISTICAL PHYSICS AND THE REPLICA METHOD

So far, we have worked with the mathematical aspect of the replica method only and avoided physical concepts. In retrospect, it is enlightening to draw an equivalence between multiuser communications and many-body problems in statistical physics, which also provides the underlying rationale for applying the replica theory in the first place.

1.6.1 A Note on Statistical Physics

Consider the physics of a macroscopic system, which typically consists of 10^{20} or more particles. Let the microscopic configuration of the system be described by a vector \mathbf{x} . The configuration of the system evolves over time according to the laws of physics. However, in view of the enormous degrees of freedom of such a macroscopic system, it is practically impossible to track the time evolution of the configuration. On the other hand, we are most interested in the macroscopic properties of the system, not the detailed configuration of the humongous number of particles. Elegantly, statistical physics introduces a probabilistic description of the system. Let $p(\mathbf{x})$ denote the probability that the system is found in configuration \mathbf{x} . Assuming the system is at thermal equilibrium in contact with a heat bath, statistical physics states that $p(\mathbf{x})$ is given by the so-called *Gibbs-Boltzmann distribution*:

$$p(\mathbf{x}) = Z^{-1} \exp \left[-\frac{1}{T} H(\mathbf{x}) \right] \quad (1.148)$$

where $H(\mathbf{x})$ denotes the *Hamiltonian*, i.e., the function that associates each configuration \mathbf{x} to its energy, where

$$Z = \sum_{\mathbf{x}} \exp \left[-\frac{1}{T} H(\mathbf{x}) \right] \quad (1.149)$$

is the *partition function* normalizing $p(\mathbf{x})$, and where the parameter $T > 0$ denotes the *temperature* of the system.

The Gibbs–Boltzmann distribution can also be characterized as the solution to a constrained optimization problem, in which the *entropy* (disorder) of the system

$$\mathcal{S} = - \sum_{\mathbf{x}} p(\mathbf{x}) \log p(\mathbf{x}) \quad (1.150)$$

is maximized under the constraint that the *energy* of the system is fixed to be

$$\mathcal{E} = \sum_{\mathbf{x}} p(\mathbf{x}) H(\mathbf{x}). \quad (1.151)$$

This optimization problem can be solved using the Lagrange multiplier method. Using (1.150) and (1.151), the probability distribution $p(\mathbf{x})$ is found to be equal to (1.148), where the Lagrange multiplier $1/T$, serving as the inverse temperature, is determined by the energy constraint (1.151).

Generally speaking, statistical physics is a theory that studies macroscopic properties (e.g., pressure, magnetization) of such a system starting from the Hamiltonian by taking the above probabilistic viewpoint. For the system shown above, it is not the most probable configuration (the ground state which has the minimum energy), but those configurations with energy close to \mathcal{E} that contribute to the physics of the system. Indeed, such configurations form the “typical set,” which determines macroscopic properties of the system. From the mathematical point of view, one can regard that statistical physics provides a framework of statistical theory regarding probability models with huge degrees of freedom. This view is fundamental in linking statistical physics with various problems in information and communication theory.

One particularly useful macroscopic quantity of the thermodynamic system is the *free energy*:

$$\mathcal{F} = \mathcal{E} - T \mathcal{S}. \quad (1.152)$$

Using (1.148)–(1.151), the free energy at equilibrium can also be expressed as

$$\mathcal{F} = -T \log Z. \quad (1.153)$$

Indeed, at thermal equilibrium, the temperature and energy of the system remain constant, the entropy is the maximum possible, and the free energy is at its minimum. The free energy is often the starting point for calculating macroscopic properties of a thermodynamic system. For example, the energy \mathcal{E} and the entropy \mathcal{S} are obtained by differentiating \mathcal{F} ; $\mathcal{E} = \partial(\mathcal{F}/T)/\partial(1/T)$ and $\mathcal{S} = -\partial\mathcal{F}/\partial T$ hold, respectively.

1.6.2 Multiuser Communications and Statistical Physics

1.6.2.1 Equivalence of Multiuser Systems and Spin Glasses In order to take advantage of the statistical physics methodologies, we equate the multiuser communication problem to an artificial thermodynamic system, called spin glass. A *spin glass* is a system consisting of many directional spins, in which the interaction of the spins is determined by the so-called *quenched random variables* whose values are determined by the realization of the spin glass. An example is a system consisting of molecules with magnetic spins that evolve over time, while the positions of the molecules that determine the amount of interactions are random (disordered) but

remain fixed for each concrete instance. In the probabilistic context, the quenched variables are simply the random variables we condition on to calculate the expectation values of the performance measures. We then average over the quenched variables in order to obtain the average performance. Let the quenched random variables be denoted by (\mathbf{Y}, \mathbf{S}) . The system can be understood as K random spins sitting in quenched randomness $(\mathbf{Y}, \mathbf{S}) = (\mathbf{y}, \mathbf{s})$, and its statistical physics described as in Section 1.6.1 with a parameterized Hamiltonian $H_{\mathbf{y}, \mathbf{s}}(\mathbf{x})$.

Suppose the temperature $T = 1$ and that the Hamiltonian is defined as

$$H_{\mathbf{y}, \mathbf{s}}(\mathbf{x}) = \frac{\|\mathbf{y} - \mathbf{s}\mathbf{x}\|^2}{2\sigma^2} - \log q_{\mathbf{X}}(\mathbf{x}) + \frac{L}{2} \log(2\pi\sigma^2), \quad (1.154)$$

then the Gibbs–Boltzmann distribution, the configuration distribution of the spin glass at equilibrium, is given by (1.10) and its corresponding partition function by (1.11) (cf. (1.148) and (1.149)). Precisely, the probability that the transmitted symbol is $\mathbf{X} = \mathbf{x}$ under the postulated model, given the observation $\mathbf{Y} = \mathbf{y}$ and the channel state $\mathbf{S} = \mathbf{s}$, is equal to the probability that the spin glass is found at configuration \mathbf{x} , given quenched random variables $(\mathbf{Y}, \mathbf{S}) = (\mathbf{y}, \mathbf{s})$.

The characteristics of the system is encoded in the quenched randomness (\mathbf{Y}, \mathbf{S}) . In the communication channel described by (1.2), (\mathbf{Y}, \mathbf{S}) takes a specific distribution, i.e., the distributions of the received signal and channel state matrix according to the prior and conditional distributions that underlie the “original” spins.

The free energy of the thermodynamic (or communication) system normalized by the number of users is

$$-\frac{T}{K} \log Z(\mathbf{Y}, \mathbf{S}) = -\frac{1}{K} \log q_{\mathbf{Y}|\mathbf{S}}(\mathbf{Y}|\mathbf{S}) \quad (1.155)$$

where we assume $T = 1$. If we assume self-averaging for the per-user free energy (1.155), it converges in probability to its expected value over the distribution of the quenched random variables (\mathbf{Y}, \mathbf{S}) in the large-system limit $K \rightarrow \infty$, which is denoted by \mathcal{F} ,

$$\mathcal{F} = -\lim_{K \rightarrow \infty} \mathbb{E} \left\{ \frac{1}{K} \log q_{\mathbf{Y}|\mathbf{S}}(\mathbf{Y}|\mathbf{S}) \right\}. \quad (1.156)$$

Hereafter, by the free energy we refer to its large-system limit (1.156).

The reader should be cautioned that for disordered systems, thermodynamic quantities may or may not be self-averaging [85]. Buttressed by numerical examples and associated results using random matrix theory, as well as recent progress [74], the self-averaging property is assumed to hold in this work.

The self-averaging property resembles the asymptotic equipartition property (AEP) in information theory [63]. An important consequence is that a macroscopic quantity of a thermodynamic system, which is a function of a large number of random variables, may become increasingly predictable from merely a few parameters independent of the realization of the quenched randomness as the system size grows without bound.

In view of (1.69) and (1.156), the large-system spectral efficiency of the multiuser system is affine in the free energy with a postulated measure q identical to the actual measure p :

$$\mathbb{C} \rightarrow \beta \mathcal{F}|_{q=p} - \frac{1}{2} \log(2\pi e). \quad (1.157)$$

Indeed, the replica analysis presented in Section 1.4 was developed based on this observation.

1.7 INTERFERENCE CANCELLATION

1.7.1 Conventional Parallel Interference Cancellation

So far we have discussed, via the single-user characterization, theoretical structure and information transmission capability of the CDMA channels, equipped with various multiuser posterior mean estimators. Straightforward computation of the PME requires K -dimensional integration (or summation if X_k 's are discrete), the computational complexity of which generally grows exponentially in K . This is practically hard (see Section 1.1.1).²⁰

The interference cancellation is a heuristic idea for reducing complexity by sub-optimal processing. Suppose we are given a realization \mathbf{s} of the channel state matrix \mathbf{S} . The basic observation is that the matched filter output for user k is decomposed into three terms:

$$\mathbf{s}_k^\top \mathbf{Y} = \|\mathbf{s}_k\|^2 \sqrt{\gamma_k} X_k + \sum_{k' \neq k} \mathbf{s}_k^\top \mathbf{s}_{k'} \sqrt{\gamma_{k'}} X_{k'} + \mathbf{s}_k^\top \mathbf{N} \quad (1.158)$$

The first is the “signal” term that is proportional to the desired symbol X_k . The second is the MAI consisting of the symbols of the remaining users. The third is the noise term. In order for a good estimation, we want the interference and noise terms to be small. Whereas essentially nothing can be done in order to eliminate the noise term, one could think of reducing the interference because, in the context of multiuser detection, we want to estimate not only X_k but $\{X_{k'}, k' \neq k\}$ as well, which means that we will certainly have some estimates for the latter, and that these estimates could be used to reconstruct and then cancel the interference. For example, if the estimates $\{\hat{x}_{k'}, k' \neq k\}$ are good enough, the “interference cancellation”

$$z_k = \mathbf{s}_k^\top \left(\mathbf{y} - \sum_{k' \neq k} \mathbf{s}_{k'} \sqrt{\gamma_{k'}} \hat{x}_{k'} \right), \quad (1.159)$$

would give us a quantity that is almost free of the interference.

The parallel interference cancellation (PIC), also referred to as the multistage detector [61], is the idea of performing the interference cancellation in stages and in parallel. It is formulated as

$$z_k^t = \mathbf{s}_k^\top \mathbf{y} - \sum_{k' \neq k} \mathbf{s}_k^\top \mathbf{s}_{k'} \sqrt{\gamma_{k'}} \hat{x}_{k'}^t \quad (1.160)$$

$$\hat{x}_k^{t+1} = f_k(z_k^t) \quad (1.161)$$

where $f_k(\cdot)$ is a decision function for user k , which may be defined on the basis of a postulated channel characteristics such as (1.14). Initialization of PIC is typically done by setting $\{\hat{x}_k^0\}$ with a computationally simple estimator, such as a linear detector.

1.7.2 Belief Propagation

1.7.2.1 Application of Belief Propagation to Multiuser Detection We next turn our focus to a systematic method for approximate computation of the posterior means.

²⁰The Gaussian-prior case is an exception, in which the PMEs are calculated algebraically, yielding a linear detector (1.12).

Posterior mean estimation is also important in researches of artificial intelligence: How one can represent uncertainties surrounding an intelligent agent is an important issue in artificial intelligence, and one might think it natural to use probability models to handle the uncertainties. Pearl's proposal of BP [53] has provided us with a unified approach to calculating posterior means, provided that a probability model is represented as a graphical model. As mentioned in Section 1.1.4, the multiuser system can be described by a bipartite graph shown in Figure 1.1., where a symbol node X_k and a chip node Y_l are connected by an edge if $s_{lk} \neq 0$. The task of a multiuser detector is to infer the symbols based on the observation at the chip nodes, to which the framework of BP is applicable. However, BP gives exact posterior means only for a limited class of probability models (i.e., those that do not contain cycles), and it generally provides approximate posterior means. BP has nevertheless been regarded as very important because the decoding algorithms of many capacity-achieving error-control coding, e.g., the turbo decoding algorithm for turbo codes and the sum-product algorithm for sparse-graph codes, turn out to be instances of BP [86, 87]. In view of such outstanding success of BP in error-control coding, one might think it worthwhile to consider application of BP to the multiuser detection.

It is possible, at least in principle, to apply BP to the multiuser detection problem, which yields the following procedure iterating the "Horizontal" and "Vertical" steps until convergence is achieved:

Input: Channel output \mathbf{y} , channel state \mathbf{s} , prior $q_X(x)$.

Initialization: Set $t := 0$, and

$$\pi_{lk}^0(x_k) = q_X(x_k), \quad l = 1, \dots, L; \quad k = 1, \dots, K. \quad (1.162)$$

Main Iterations:

for $t = 0$ to maximum number of iterations **do**

 "Horizontal" step:

$$\rho_{lk}^{t+1}(x_k) = \int \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_l - (\mathbf{s}\mathbf{x})_l)^2}{2\sigma^2}\right] \prod_{k' \neq k} [\pi_{lk'}^t(x_{k'}) dx_{k'}] \quad (1.163)$$

 "Vertical" step:

$$\pi_{lk}^{t+1}(x_k) = \alpha_{lk} q_X(x_k) \prod_{l' \neq l} \rho_{l'k}^{t+1}(x_k) \quad (1.164)$$

 where α_{lk} is the normalization coefficient so that $\pi_{lk}^{t+1}(x_k)$ is a pmf.

end for

Output: After convergence is achieved, calculate

$$\pi_k(x_k) = \alpha_k q_X(x_k) \prod_{l=1}^L \rho_{lk}(x_k), \quad (1.165)$$

which gives an approximate marginal distribution of X_k .

return $\pi_k(x)$, for all $k = 1, \dots, K$ and all x .

The difficulty in applying BP to the user detection resides in the Horizontal step, where one has to perform $(K - 1)$ -dimensional integration, whose computational complexity grows exponentially in K .

1.7.2.2 Conventional Parallel Interference Cancellation as Approximate BP Here we discuss a simple heuristic approximation [88, 89] to alleviate the computational difficulty, which, interestingly, leads to the conventional parallel interference cancellation scheme of Section 1.7.1. Their heuristics consists of two basic ideas: The first idea is to replace the quantities $\{\pi_{lk}^t(x_k); l = 1, \dots, N\}$, computed in the vertical steps, with

$$\pi_k^t(x_k) = \alpha_k q_X(x_k) \prod_{l=1}^N \rho_{lk}^t(x_k) \propto \pi_{lk}^t(x_k) \rho_{lk}^t(x_k). \quad (1.166)$$

The approximation of $\pi_{lk}^t(x_k)$ with $\pi_k^t(x_k)$ is expected to be quite good asymptotically, because the factor $\rho_{lk}^t(x_k)$ is one among the N factors. The second idea is, instead of evaluating the expectation with respect to $\prod_{k' \neq k} \pi_{lk'}^t(x_{k'})$ in the horizontal steps (1.163), to plug the expectations $m_{k'}^t$ of $X_{k'}$ with respect to $\pi_{k'}^t(x_{k'})$ into the exponent of the integrand. The approximation gives

$$\rho_{lk}^{t+1}(x_k) \propto \exp \left[\frac{1}{\sigma^2} \sqrt{\frac{\gamma_k}{L}} u_l^t s_{lk} x_k - \frac{\gamma_k}{2\sigma^2} \frac{s_{lk}^2}{L} x_k^2 \right], \quad (1.167)$$

and

$$\pi_k^{t+1}(x_k) \propto q_X(x_k) \exp \left[\frac{\sqrt{\gamma_k} x_k}{\sigma^2} \mathbf{s}_k^\top \mathbf{u}^t - \frac{\gamma_k x_k^2}{2\sigma^2} \|\mathbf{s}_k\|^2 \right], \quad (1.168)$$

where $\mathbf{u}^t = [u_1, \dots, u_L]^\top$, $u_l^t = y_l - L^{-1/2} \sum_{k' \neq k} \sqrt{\gamma_{k'}} s_{lk'} m_{k'}^t$. Under the random spreading assumption and in the large-system regime, $\|\mathbf{s}_k\|^2 = L^{-1} \sum_{l=1}^L s_{lk}^2$ can safely be regarded as being equal to 1 due to the law of large numbers, so that (1.168) is represented as an update rule in terms of $\{m_k^t\}$, as

$$m_k^{t+1} = f \left(\mathbf{s}_k^\top \mathbf{y} - \sum_{k' \neq k} \mathbf{s}_k^\top \mathbf{s}_{k'} \sqrt{\gamma_{k'}} m_{k'}^t, \gamma_k; \frac{1}{\sigma^2} \right). \quad (1.169)$$

The function $f(z, \gamma; \xi)$ is the decision function $\mathbb{E}_q \{X | Z = z, \gamma; \xi\}$ induced by the single-user channel with q_X and $q_{Z|X, \gamma; \xi}$ as defined in (1.40). Equation (1.169) is nothing but the conventional PIC algorithm discussed in Section 1.7.1, demonstrating that the conventional PIC algorithm can be regarded as an approximation to the BP algorithm.

1.7.2.3 BP-based Parallel Interference Cancellation Algorithm We now discuss a more sophisticated approximation scheme [46]. Let us first rewrite the quantity $(\mathbf{s}\mathbf{x})_l$ as

$$(\mathbf{s}\mathbf{x})_l = \frac{1}{\sqrt{L}} \sum_{k' \neq k} \sqrt{\gamma_{k'}} s_{lk'} x_{k'} + \sqrt{\frac{\gamma_k}{L}} s_{lk} x_k \equiv (\mathbf{s}\mathbf{x})_{l \setminus k} + \sqrt{\frac{\gamma_k}{L}} s_{lk} x_k. \quad (1.170)$$

Noting that the second term of the rightmost side of (1.170) can be regarded small compared with the first term as L becomes large, one can expand the integrand

of (1.163) as

$$\begin{aligned} \exp\left[-\frac{(y_l - (\mathbf{s}\mathbf{x})_l)^2}{2\sigma^2}\right] &= \exp\left[-\frac{(y_l - (\mathbf{s}\mathbf{x})_{l\setminus k})^2}{2\sigma^2}\right] \\ &\times \left(1 + \frac{1}{\sigma^2} \sqrt{\frac{\gamma_k}{L}} s_{lk} x_k [y_l - (\mathbf{s}\mathbf{x})_{l\setminus k}] \right. \\ &\left. + \frac{\gamma_k}{2\sigma^4 L} \left(s_{lk} x_k [y_l - (\mathbf{s}\mathbf{x})_{l\setminus k}]\right)^2 + O(L^{-3/2})\right). \end{aligned} \quad (1.171)$$

The key observation is that the distributions $\{\pi_{l'k'}^t(x_{k'})\}$ affects the LHS of (1.163) only through the distribution of $(\mathbf{s}\mathbf{X})_{l\setminus k}$, which can be regarded as a Gaussian random variable in the large-system regime due to the central-limit theorem, since it is a weighted sum of the independent random variables $X_{k'} \sim \pi_{l'k'}^t$; $k' \neq k$. The mean and variance are

$$\mu_{lk}^t = \frac{1}{\sqrt{L}} \sum_{k' \neq k} \sqrt{\gamma_{k'}} s_{lk'} m_{lk'}^t, \quad (1.172)$$

and

$$C_{lk}^t = \frac{1}{L} \sum_{k' \neq k} \gamma_{k'} s_{lk'}^2 V_{lk'}^t, \quad (1.173)$$

respectively, where m_{lk}^t and V_{lk}^t are the mean and the variance of $X_k \sim \pi_{lk}^t$. Note that μ_{lk}^t is an estimate, based on $\{\pi_{l'k'}^t\}$, of the MAI component in y_l for user k , and that C_{lk}^t represents uncertainty of the estimate, quantifying magnitude of residual MAI component after the cancellation of MAI with μ_{lk}^t . Retaining terms up to order of L^{-1} , and calculating the Gaussian integral, the horizontal step can be represented as follows:

$$\rho_{lk}^{t+1}(x_k) \propto \exp\left[\sqrt{\frac{\gamma_k}{L}} \frac{s_{lk}}{\sigma^2 + C_{lk}^t} x_k - \frac{\gamma_k s_{lk}^2}{2L(\sigma^2 + C_{lk}^t)} x_k^2\right]. \quad (1.174)$$

Introducing the parametrization

$$\rho_{lk}^t(x_k) \propto \exp\left[\sqrt{\frac{\gamma_k}{L}} \theta_{lk}^t x_k - \frac{\gamma_k}{2L} \Xi_{lk}^t x_k^2\right], \quad (1.175)$$

the horizontal and vertical steps are represented as

$$\theta_{lk}^{t+1} = \frac{s_{lk}(y_l - \mu_{lk}^t)}{\sigma^2 + C_{lk}^t} \quad \text{and} \quad \Xi_{lk}^{t+1} = \frac{s_{lk}^2}{\sigma^2 + C_{lk}^t}, \quad (1.176)$$

and

$$\pi_{lk}^t(x_k) = \alpha_{lk} q_X(x_k) \exp\left[\sqrt{\gamma_k} x_k \left(\frac{1}{\sqrt{L}} \sum_{l' \neq l} \theta_{l'k}^t\right) - \frac{\gamma_k x_k^2}{2} \left(\frac{1}{L} \sum_{l' \neq l} \Xi_{l'k}^t\right)\right], \quad (1.177)$$

respectively. The mean m_{lk}^t and the variance V_{lk}^t are to be calculated from $\pi_{lk}^t(x_k)$. Equations (1.172), (1.173), (1.176), (1.177) define an approximate BP algorithm for user detection. This algorithm has polynomial-order computational complexity per iteration, as opposed to the exponential-order complexity of the original BP.

The final results are to be read out, after convergence is achieved, as statistics of the distributions

$$\pi_k(x_k) = \alpha_k q_X(x_k) \exp \left[\sqrt{\gamma_k} x_k \left(\frac{1}{\sqrt{L}} \sum_{l=1}^L \theta_{lk}^* \right) - \frac{\gamma_k x_k^2}{2} \left(\frac{1}{L} \sum_{l=1}^L \Xi_{lk}^* \right) \right], \quad (1.178)$$

where θ_{lk}^* and Ξ_{lk}^* denote the respective quantities at the equilibrium.

One might ask how good the algorithm performs. The key observations to the question are that one can regard $\pi_{lk}^t(x_k)$, as given by (1.177), as a posterior distribution with the prior q_X and the Gaussian channel $q_{Z|X,\gamma;\xi}(z_{lk}^t|x_k, \gamma_k; \xi_{lk}^t)$ (see (1.40)) with appropriately chosen parameters,

$$\xi_{lk}^t = \frac{1}{L} \sum_{l' \neq l} \Xi_{l'k}^t \quad \text{and} \quad z_{lk}^t = \frac{1}{\xi_{lk}^t \sqrt{L}} \sum_{l' \neq l} \theta_{l'k}^t, \quad (1.179)$$

and that one can apply the density evolution [90] idea to analyze macroscopic dynamical behaviors of the BP-based algorithm [89, 91], which is motivated by its great success in the analysis of BP-based decoding algorithms of sparse-graph codes. Basically, the density evolution describes time evolution of the distributions of the “messages” $(\theta_{lk}^t, \Xi_{lk}^t)$ and (m_{lk}^t, V_{lk}^t) . When applied to the approximate BP algorithm introduced above, it turns out that the distributions of z_{lk}^t and ξ_{lk}^t are relevant. Relying on a heuristic argument (which can be justified only in case of sparse spreading [50, 51]), one finds, under random spreading and in the large-system limit, that ξ_{lk}^t becomes deterministic and independent of l or k , and that

$$(\xi^{t+1})^{-1} = \sigma^2 + \beta \mathbf{E} \{ \gamma \mathcal{V}^t \} \quad (1.180)$$

holds, where $\mathcal{V}^t \approx V_{lk}^t$ denotes the variance of π_{lk}^t , and where we dropped the indexes lk from ξ_{lk}^t due to the asymptotic independence. As for z_{lk}^t , one can regard it as following a zero-mean Gaussian distribution, and the variance, denoted here by $(\eta^t)^{-1}$, turns out to satisfy

$$(\eta^{t+1})^{-1} = 1 + \beta \mathbf{E} \{ \gamma \mathcal{E}^t \}, \quad (1.181)$$

where \mathcal{E}^t denotes the mean squared error of the estimate m_{lk}^t . Comparing the density evolution formulas (1.180) and (1.181) with the fixed-point equations of the replica analysis (1.46), one observes that stationarity condition of the density evolution formulas coincide with the fixed-point equations for arbitrary inputs, which has been proved only in case of sparse spreading [51], as is pointed out in Sections 1.1.3 and 1.3.4. On the theoretical side, the coincidence suggests an interesting and not yet fully understood link between the replica analysis and the BP-based algorithm. As for the application side, on the other hand, it suggests that, under the random spreading, the BP-based algorithm performs as predicted by the replica analysis in the large-system limit, and can thus be “asymptotically optimal.”

1.8 CONCLUDING REMARKS

This chapter presents a simple characterization of the large-system performance of multiuser detection under arbitrary input and SNR distribution (and/or flat

fading). A broad family of multiuser detectors is studied under the umbrella of posterior mean estimators, which includes well-known detectors such as the matched filter, decorrelator, linear MMSE detector, maximum likelihood (jointly optimal) detector, and the individually optimal detector.

A key conclusion is the decoupling of a multiuser channel concatenated with a generic multiuser detector front end. It is found that the detection output for each user is a deterministic function of a “hidden” Gaussian statistic centered at the transmitted symbol. Hence the single-user channel seen at the multiuser detection output is equivalent to a Gaussian channel conditioned on the input SNR in which the overall effect of MAI is a degradation in the effective SNR. The degradation factor, known as the multiuser efficiency, is the solution to a pair of coupled fixed-point equations, and can be easily computed numerically if not analytically.

Another set of results, tightly related to the decoupling principle, lead to general formulas for the large-system spectral efficiency of multiuser channels expressed in terms of the multiuser efficiency, both under joint and separate decoding.

Turning to algorithmic issues, the chapter also discusses application of belief propagation. It is shown that the conventional parallel interference cancellation is an approximate BP, and that more systematic approximation leads to a variant of PIC, whose performance is expected to be asymptotically optimal in the large-system limit. The suggested relation between density evolution formulas of the algorithm and the fixed-point equations obtained by the replica analysis, which has not been fully explored yet, might be of help interpreting the replica results further.

From a practical viewpoint, this chapter presents new results on the efficiency of CDMA communication under arbitrary user powers and input signaling such as PSK and QAM. The results in this chapter allow the performance of multiuser detection to be characterized by a single parameter, the multiuser efficiency. Thus, the results offer convenient performance measures and valuable insights in the design and analysis of multiuser systems, e.g., in power control [92].

The linear system in our study also models MIMO channels under various circumstances. The results can thus be used to evaluate the output SINR or spectral efficiency of high-dimensional MIMO channels (such as multiple-antenna systems) with arbitrary signaling and various detection techniques. Some of the results in this chapter have been generalized to MIMO channels with spatial correlation at both transmitter and receiver sides [93], as well as to MIMO-CDMA channels [94].

Acknowledgements

We are grateful to the editor and anonymous reviewer for their helpful comments. Portions of the results included in this chapter originated during DG’s Ph.D. study under the advice of Professor Sergio Verdú, to whom DG is very grateful. DG would also like to thank Dr. Chih-Chun Wang for useful discussions. TT would like to thank Mr. Keigo Takeuchi of Kyoto University, for his helpful discussion and constructive comments on the drafts, and Professor Mihai Putinar, University of California at Santa Barbara, for providing information about recent development of the moment problem.

This work is supported in part by the National Science Foundation CAREER Award CCF-0644344, the DARPA IT-MANET Grant W911NF-07-1-0028, and

Grant-in-Aid for Scientific Research on Priority Areas (Nos. 14084209 and 18079010), MEXT, Japan.

REFERENCES

1. D. Horwood and R. Gagliardi, "Signal design for digital multiple access communications," *IEEE Trans. Commun.*, vol. 23, pp. 985–995, May 1975.
2. K. S. Schneider, "Optimum detection of code division multiplexed signals," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 15, no. 1, pp. 181–185, Jan. 1979.
3. S. Verdú, "Computational complexity of optimum multiuser detection," *Algorithmica*, vol. 4, no. 3, pp. 303–312, 1989.
4. S. Verdú, "Minimum probability of error for asynchronous Gaussian multiple-access channels," *IEEE Trans. Inform. Theory*, vol. 32, no. 1, pp. 85–96, Jan. 1986.
5. M. Opper and W. Kinzel, "Statistical mechanics of generalization," in *Models of Neural Networks III: Association, Generalization, and Representation*, E. Domany, J. L. van Hemmen, and K. Schulten, Eds. Springer, 1996.
6. S. Verdú, "Minimum probability of error for asynchronous multiple access communication systems," in *Proc. IEEE Military Communications Conference*, vol. 1, Nov. 1983, pp. 213–219.
7. J. S. Lehnert and M. B. Pursley, "Error probabilities for binary direct-sequence spread-spectrum communications with random signature sequences," *IEEE Trans. Commun.*, vol. 35, pp. 87–98, Jan. 1987.
8. D. Guo, L. K. Rasmussen, and T. J. Lim, "Linear parallel interference cancellation in long-code CDMA multiuser detection," *IEEE J. Selected Areas Commun.*, vol. 17, pp. 2074–2081, Dec. 1999.
9. U. Madhow and M. L. Honig, "On the average near-far resistance for MMSE detection of direct sequence CDMA signals with random spreading," *IEEE Trans. Inform. Theory*, vol. 45, no. 6, pp. 2039–2045, Sep. 1999.
10. M. B. Pursley, "Performance evaluation for phase-coded spread-spectrum multiple-access communication—Part I: System analysis," *IEEE Trans. Commun.*, vol. 25, no. 8, pp. 795–799, Aug. 1977.
11. D. N. C. Tse and S. V. Hanly, "Linear multiuser receivers: Effective interference, effective bandwidth and user capacity," *IEEE Trans. Inform. Theory*, vol. 45, no. 2, pp. 641–657, Mar. 1999.
12. D. Guo, S. Verdú, and L. K. Rasmussen, "Asymptotic normality of linear multiuser receiver outputs," *IEEE Trans. Inform. Theory*, vol. 48, no. 12, pp. 3080–3095, Dec. 2002.
13. Z. D. Bai, "Methodologies in spectral analysis of large dimensional random matrices, a review," *Statistica Sinica*, vol. 9, no. 3, pp. 611–677, Jul. 1999.
14. A. M. Tulino and S. Verdú, "Random matrix theory and wireless communications," *Foundations and Trends in Communications and Information Theory*, vol. 1, no. 1, pp. 1–182, 2004.
15. S. Verdú, *Multiuser Detection*. Cambridge University Press, 1998.
16. Y. C. Eldar and A. M. Chan, "On the asymptotic performance of the decorrelator," *IEEE Trans. Inform. Theory*, vol. 49, no. 9, pp. 2309–2313, Sep. 2003.

17. D. Guo and S. Verdú, "Multiuser detection and statistical mechanics," in *Communications, Information and Network Security*, V. Bhargava, H. V. Poor, V. Tarokh, and S. Yoon, Eds. Kluwer Academic Publishers, 2002, ch. 13, pp. 229–277.
18. S. Verdú and S. Shamai, "Spectral efficiency of CDMA with random spreading," *IEEE Trans. Inform. Theory*, vol. 45, no. 2, pp. 622–640, Mar. 1999.
19. S. Shamai and S. Verdú, "The impact of frequency-flat fading on the spectral efficiency of CDMA," *IEEE Trans. Inform. Theory*, vol. 47, no. 4, pp. 1302–1327, May 2001.
20. A. J. Grant and P. D. Alexander, "Random sequence multisets for synchronous code-division multiple-access channels," *IEEE Trans. Inform. Theory*, vol. 44, no. 7, pp. 2832–2836, Nov. 1998.
21. P. B. Rapajic and D. Popescu, "Information capacity of a random signature multiple-input multiple-output channel," *IEEE Trans. Commun.*, vol. 48, no. 8, pp. 1245–1248, Aug. 2000.
22. T. Tanaka, "Analysis of bit error probability of direct-sequence CDMA multiuser demodulators," in *Advances in Neural Information Processing Systems*, T. K. Leen et al., Ed. The MIT Press, 2001, vol. 13, pp. 315–321.
23. —, "Average-case analysis of multiuser detectors," in *Proc. IEEE Int. Symp. Inform. Theory*. Washington, D.C. USA, Jun. 2001, p. 287.
24. —, "Performance analysis of neural CDMA multiuser detector," in *Proc. INNS-IEEE International Joint Conference on Neural Networks*. Washington, D.C., USA, Jul. 2001, pp. 2832–2837.
25. —, "Statistical mechanics of CDMA multiuser demodulation," *Europhysics Letters*, vol. 54, no. 4, pp. 540–546, 2001.
26. —, "A statistical mechanics approach to large-system analysis of CDMA multiuser detectors," *IEEE Trans. Inform. Theory*, vol. 48, no. 11, pp. 2888–2910, Nov. 2002.
27. D. Guo and S. Verdú, "Randomly spread CDMA: Asymptotics via statistical physics," *IEEE Trans. Inform. Theory*, vol. 51, no. 6, pp. 1982–2010, Jun. 2005.
28. R. R. Müller and W. H. Gerstaecker, "On the capacity loss due to separation of detection and decoding," *IEEE Trans. Inform. Theory*, vol. 50, no. 8, pp. 1769–1778, Aug. 2004.
29. R. R. Müller, "On channel capacity, uncoded error probability, ML-detection and spin glasses," in *Proc. Workshop on Concepts in Information Theory*. Breisach, Germany, 2002, pp. 79–81.
30. D. Guo and S. Verdú, "Replica analysis of CDMA spectral efficiency," in *Proc. IEEE Inform. Theory Workshop*. Paris, France, 2003.
31. D. Guo, "Gaussian channels: Information, estimation and multiuser detection," Ph.D. dissertation, Department of Electrical Engineering, Princeton University, 2004.
32. T. Tanaka, "Replica analysis of performance loss due to separation of detection and decoding in CDMA channels," in *Proc. IEEE Int. Symp. Inform. Theory*. Seattle, WA, USA, 2006, pp. 2368–2372.
33. D. Guo, "Performance of synchronous multirate CDMA via statistical physics," in *Proc. IEEE Int. Symp. Information Theory*. Adelaide, Australia, Sep. 2005.
34. —, "Performance of multicarrier CDMA in frequency-selective fading via statistical physics," *IEEE Trans. Inform. Theory*, vol. 52, no. 4, pp. 1765–1774, Apr. 2006.
35. S. F. Edwards and P. W. Anderson, "Theory of spin glasses," *Journal of Physics F: Metal Physics*, vol. 5, pp. 965–974, 1975.

36. M. Mézard, G. Parisi, and M. A. Virasoro, *Spin Glass Theory and Beyond*. World Scientific, 1987.
37. H. Nishimori, *Statistical Physics of Spin Glasses and Information Processing: An Introduction*, ser. Number 111 in International Series of Monographs on Physics. Oxford University Press, 2001.
38. N. Sourlas, “Spin-glass models as error-correcting codes,” *Nature*, vol. 339, no. 6227, pp. 693–695, Jun. 1989.
39. Y. Kabashima and D. Saad, “Statistical mechanics of error-correcting codes,” *Europhysics Letters*, vol. 45, no. 1, pp. 97–103, 1999.
40. A. Montanari, “Turbo codes: The phase transition,” *European Physical Journal B*, vol. 18, pp. 121–136, 2000.
41. T. Murayama, Y. Kabashima, D. Saad, and R. Vicente, “Statistical physics of regular low-density parity-check error-correcting codes,” *Physical Review E*, vol. 62, no. 2, pp. 1577–1591, 2000.
42. A. Montanari and N. Sourlas, “The statistical mechanics of turbo codes,” *European Physical Journal B*, vol. 18, no. 1, pp. 107–119, 2000.
43. Y. Kabashima and D. Saad, “Statistical mechanics of low-density parity-check codes,” *J. Phys. A: Math. Gen.*, vol. 37, pp. R1–R43, 2004.
44. A. L. Moustakas, S. H. Simon, and A. M. Sengupta, “MIMO capacity through correlated channels in the presence of correlated interferers and noise: A (not so) large N analysis,” *IEEE Trans. Inform. Theory*, vol. 49, no. 10, pp. 2545–2561, Oct. 2003.
45. R. R. Müller, “Channel capacity and minimum probability of error in large dual antenna array systems with binary modulation,” *IEEE Trans. Signal Processing*, vol. 51, pp. 2821–2828, Nov. 2003.
46. Y. Kabashima, “A CDMA multiuser detection algorithm on the basis of belief propagation,” *J. Phys. A: Math. Gen.*, vol. 36, pp. 11 111–11 121, 2003.
47. T. Fabricius and O. Winther, “Correcting the bias of subtractive interference cancellation in CDMA: Advanced mean field theory,” Informatics and Mathematical Modelling, Technical University of Denmark, Tech. Rep., 2003.
48. M. Talagrand, “Rigorous results for mean field models for spin glasses,” *Theoretical Computer Science*, vol. 265, pp. 69–77, Aug. 2001.
49. —, *Spin Glasses: A Challenge for Mathematicians*. Springer, 2003.
50. A. Montanari and D. Tse, “Analysis of belief propagation for non-linear problems: The example of CDMA (or: How to prove Tanaka’s formula),” in *Proc. IEEE Inform. Theory Workshop*. Punta del Este, Uruguay, Mar. 2006, pp. 122–126.
51. D. Guo and C.-C. Wang, “Multiuser detection of sparsely spread CDMA,” *IEEE J. Select. Areas Commun., Special Issue on Multiuser Detection for Advanced Communication Systems and Networks*, May 2008.
52. —, “Random sparse linear systems observed via arbitrary channels: A decoupling principle,” in *Proc. IEEE Int. Symp. Inform. Theory*. Nice, France, Jun. 2007.
53. J. Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988, revised 2nd printing.
54. M. Mézard and G. Parisi, “Replicas and optimization,” *J. Physique Lett.*, vol. 46, no. 17, pp. L-771–L-778, Sep. 1985.
55. Y. Fu and P. W. Anderson, “Application of statistical mechanics to NP-complete problems in combinatorial optimisation,” *J. Phys. A: Math. Gen.*, vol. 19, no. 9, pp. 1605–1620, Jun. 1986.

56. D. J. Amit, H. Gutfreund, and H. Sompolinsky, "Storing infinite numbers of patterns in a spin-glass model of neural networks," *Phys. Rev. Lett.*, vol. 55, no. 14, pp. 1530–1533, Sep. 1985.
57. M. Mézard and A. Montanari, *Information, Physics and Computation*. Oxford Univ. Press, 2008.
58. R. Richardson and R. Urbanke, *Modern Coding Theory*. Cambridge Univ. Press, 2008.
59. P. Patel and J. Holtzman, "Analysis of simple successive interference cancellation scheme in a DS/CDMA," *IEEE J. Select. Areas Commun.*, vol. 12, pp. 796–807, Jun. 1994.
60. L. K. Rasmussen, T. J. Lim, and A.-L. Johansson, "A matrix-algebraic approach to successive interference cancellation in CDMA," *IEEE Trans. Commun.*, vol. 48, no. 1, pp. 145–151, Jan. 2000.
61. M. K. Varanasi and B. Aazhang, "Multistage detection in asynchronous code-division multiple-access communications," *IEEE Trans. Commun.*, vol. 38, pp. 509–519, Apr. 1990.
62. D. Guo, L. K. Rasmussen, S. Sun, and T. J. Lim, "A matrix-algebraic approach to linear parallel interference cancellation in CDMA," *IEEE Trans. Commun.*, vol. 48, pp. 152–161, Jan. 2000.
63. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley, 2006.
64. D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inform. Theory*, vol. 51, no. 4, pp. 1261–1282, Apr. 2005.
65. M. K. Varanasi and T. Guess, "Optimum decision feedback multiuser equalization with successive decoding achieves the total capacity of the Gaussian multiple-access channel," in *Proc. Asilomar Conf. on Signals, Systems and Computers*. Monterey, CA, USA, Nov. 1997, pp. 1405–1409.
66. P. Rapajic, M. Honig, and G. Woodward, "Multiuser decision-feedback detection: Performance bounds and adaptive algorithms," in *Proc. IEEE Int. Symp. Inform. Theory*. Cambridge, MA USA, Aug. 1998, p. 34.
67. S. L. Ariyavisitakul, "Turbo space-time processing to improve wireless channel capacity," *IEEE Trans. Commun.*, vol. 48, no. 8, pp. 1347–1359, Aug. 2000.
68. R. R. Müller, "Multiuser receivers for randomly spread signals: Fundamental limits with and without decision-feedback," *IEEE Trans. Inform. Theory*, vol. 47, no. 1, pp. 268–283, Jan. 2001.
69. T. Guess and M. K. Varanasi, "An information-theoretic framework for deriving canonical decision-feedback receivers in Gaussian channels," *IEEE Trans. Inform. Theory*, vol. 51, no. 1, pp. 173–187, Jan. 2005.
70. G. D. Forney, Jr., "Shannon meets Wiener II: On MMSE estimation in successive decoding schemes," in *Proc. 42nd Allerton Conf. Commun., Control, and Computing*. Monticello, IL, USA, 2004. [Online]. Available: arXiv:cs/0409011 [cs.IT]
71. D. Ruelle, *Statistical Mechanics: Rigorous Results*. Imperial College Press and World Scientific Publishing, 1999.
72. M. Yoshida and T. Tanaka, "Analysis of sparsely-spread CDMA via statistical mechanics," in *Proc. IEEE Int. Symp. Inform. Theory*. Seattle, WA, USA, 2006, pp. 2378–2382.

73. J. Raymond and D. Saad, "Sparsely spread CDMA—a statistical mechanics-based analysis," *J. Phys. A: Math. Theor.*, vol. 40, pp. 12 315–12 333, 2007.
74. S. B. Korada and N. Macris, "On the concentration of the capacity for a code division multiple access system," in *Proc. IEEE Int. Symp. Inform. Theory*. Nice, France, Jun. 2007.
75. T. Tanaka, "Moment problem in replica method," *Interdisciplinary Information Sciences*, vol. 13, no. 1, pp. 17–23, 2007.
76. J. Hubbard, "Calculation of partition functions," *Physics Review Letters*, vol. 3, no. 2, pp. 77–78, 1959.
77. R. S. Ellis, *Entropy, Large Deviations, and Statistical Mechanics*, ser. A series of comprehensive studies in mathematics. Springer-Verlag, 1985, vol. 271.
78. K. Nakamura and T. Tanaka, "Microscopic analysis for decoupling principle of linear vector channel," submitted to *IEEE Int. Symp. Info. Theory*, 2008. [Online]. Available: arXiv:0801.4198 [cs.IT]
79. W. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd ed. John Wiley & Sons, Inc., 1971, vol. II.
80. N. I. Akhiezer, *The Classical Moment Problem*. Oliver and Boyd Ltd., 1965.
81. L. C. Petersen, "On the relation between the multidimensional moment problem and the one-dimensional moment problem," *Math. Scand.*, vol. 51, pp. 361–366, 1982.
82. B. Fuglede, "The multidimensional moment problem," *Expo. Math.*, vol. 1, pp. 47–65, 1983.
83. J. R. L. de Almeida and D. J. Thouless, "Stability of the Sherrington-Kirkpatrick solution of a spin glass model," *Journal of Physics A: Mathematical and Physical*, vol. 11, pp. 983–990, 1978.
84. T. Uezu, M. Yoshida, T. Tanaka, and M. Okada, "Statistical mechanical analysis of CDMA multiuser detectors—AT stability and entropy of the RS solution, and 1RSB solution," *Progress of Theoretical Physics Supplement*, vol. 157, pp. 254–257, 2005.
85. F. Comets, "The martingale method for mean-field disordered systems at high temperature," in *Mathematical Aspects of Spin Glasses and Neural Networks*, A. Bovier and P. Picco, Eds. Birkhäuser, 1998, pp. 91–113.
86. R. J. McEliece, D. J. C. MacKay, and J.-F. Cheng, "Turbo decoding as an instance of Pearl's 'belief propagation' algorithm," *IEEE J. Select. Areas Commun.*, vol. 16, no. 2, pp. 140–152, Feb. 1998.
87. D. J. C. MacKay, "Good error-correcting codes based on very sparse matrices," *IEEE Trans. Inform. Theory*, vol. 45, no. 2, pp. 399–431, Mar. 1999, errata, 47(5):2101, July 2001.
88. T. Tanaka, "Density evolution for multistage CDMA multiuser detector," in *Proc. IEEE Int. Symp. Inform. Theory*. Lausanne, Switzerland, 2002, p. 23.
89. T. Tanaka and M. Okada, "Approximate belief propagation, density evolution, and statistical neurodynamics for CDMA multiuser detection," *IEEE Trans. Inform. Theory*, vol. 51, no. 2, pp. 700–706, Feb. 2005.
90. T. J. Richardson and R. L. Urbanke, "The capacity of low-density parity-check codes under message-passing decoding," *IEEE Trans. Inform. Theory*, vol. 47, pp. 599–618, Feb. 2001.
91. T. Ikehara and T. Tanaka, "Decoupling principle in belief-propagation-based CDMA multiuser detection algorithm," in *Proc. IEEE Int. Symp. Inform. Theory*. Nice, France, 2007, pp. 2081–2085.

92. F. Meshkati, D. Guo, H. V. Poor, and S. C. Schwartz, "A unified approach to energy-efficient power control in large CDMA systems," *IEEE Trans. Wireless Commun.*, 2008, to appear.
93. C. K. Wen, P. Ting, and J.-T. Chen, "Asymptotic analysis of mimo wireless systems with spatial correlation at the receiver," *IEEE Trans. Commun.*, vol. 54, no. 2, pp. 349–363, Feb. 2006.
94. K. Takeuchi, T. Tanaka, and T. Yano, "Asymptotic analysis of general multiuser detectors in MIMO DS-CDMA channels," *IEEE J. Select. Areas Commun.*, May 2008.