UNIVERSITY OF CALIFORNIA
Santa Barbara

A Deterministic Annealing Framework for Global Optimization of Delay-Constrained Communication and Control Strategies

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Electrical and Computer Engineering

by

Mustafa Said Mehmetoglu

Committee in Charge:

Professor Kenneth Rose, Chair

Professor Shivkumar Chandrasekaran

Professor João Hespanha

Professor Yasamin Mostofi

September 2016

The Dissertation of
Mustafa Said Mehmetoglu is approved:

Professor Shivkumar Chandrasekaran

Professor João Hespanha

Professor Yasamin Mostofi

Professor Kenneth Rose, Committee Chairperson

August 2016

A Deterministic Annealing Framework for Global Optimization of

Delay-Constrained Communication and Control Strategies

by

Mustafa Said Mehmetoglu

To my family, for their love and support

Acknowledgements

I would like to thank Professor Kenneth Rose for his excellent guidance and encouragement throughout my PhD. His extensive knowledge and deep insight into the subject have helped me immensely. I was able to gain a strong understanding of the fundamentals and carry out all the research of my PhD thanks to his excellent teaching of subject matters and research techniques. I have always been impressed by his insight and intellect, and will always remember our intellectually stimulating discussions on research.

I would like to thank all the committee members, Prof. Shivkumar Chandrasekaran, Prof. João Hespanha and Prof. Yasamin Mostofi for reading and reviewing my thesis. I have been fortunate to have taken several wonderful classes during my studies from highest caliber faculty members, and I am thankful to all of them for offering and teaching those classes.

I would like to thank my main collaborator and mentor Emrah for his help and active involvement in my research, and for insightful discussions. I am grateful to all of my lab mates including Kumar, Tejaswi and Mehdi for interesting and thought provoking discussions.

I have had many good friends and roommates in Santa Barbara and I am thankful to all of them for making Santa Barbara a second home for me. Their friendship and support helped made my stay enjoyable.

I have had wonderful parents who always supported me and encouraged me to continue during hard times of my PhD. I am grateful to my father, Idris, for giving me the vision to pursue the highest degree of study in such an elite university. This achievement would have been impossible without his guidance. I thank my mother, Gulsum, for her unconditional love and support for me.

Finally, I would like to thank to my love and wonderful wife, Esra, for leaving everything behind and coming with me to Santa Barbara, and supporting me as I work towards my degree. I hope that we can spend the rest of our lives together and have a wonderful marriage.

Curriculum Vitæ

Mustafa Said Mehmetoglu

Education

2013 Master of Science in Electrical and Computer Engineering, University of California at Santa Barbara, USA.

2011 Bachelor of Science in Electrical Engineering, Bilkent University, Ankara, Turkey.

Experience

2011-2016 Graduate Student Researcher, University of California, Santa Barbara.

2011-2012 Teaching Assistant, University of California, Santa Barbara.

2010 Intern, EEM Elevator Systems, Turkey.

Publications

- M. S. Mehmetoglu, E. Akyol, and K. Rose, "Deterministic Annealing Optimization for Witsenhausen's and Related Decentralized Stochastic Control Problems", submitted to *IEEE Transactions on Automatic Control*

- M. S. Mehmetoglu, E. Akyol, and K. Rose, "Analog Multiple Descriptions: A Zero Delay Source-Channel Coding Approach", Proc. of *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016

- M. Mehmetoglu, E. Akyol, and K. Rose, "Deterministic Annealing Based Optimization for Zero-Delay Source-Channel Coding in Networks", *IEEE Transactions on Communications*, 2015

- M. S. Mehmetoglu, E. Akyol, and K. Rose, "Optimization of zero-delay mappings for distributed coding by deterministic annealing", Proc. of *IEEE In-*

ternational Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014

- M. S. Mehmetoglu, E. Akyol, and K. Rose, "A Deterministic Annealing Approach to Witsenhausen's Counterexample", Proc. of *IEEE International Symposium on Information Theory (ISIT)*, 2014

- M. S. Mehmetoglu, E. Akyol, and K. Rose, "A Deterministic Annealing Approach to Optimization of Zero-delay Source-Channel Codes", Proc. of *IEEE Information Theory Workshop (ITW)*, Sep. 2013

Abstract

A Deterministic Annealing Framework for Global Optimization of Delay-Constrained Communication and Control Strategies

Mustafa Said Mehmetoglu

This dissertation is concerned with the problem of global optimization of delay constrained communication and control strategies. Specifically, the objective is to obtain optimal encoder and decoder functions that map between the source space and the channel space, to minimize a given cost functional. The cost surfaces associated with these problems are highly complex and riddled with local minima, rendering gradient descent based methods ineffective. This thesis proposes and develops a powerful non-convex optimization method based on the concept of deterministic annealing (DA) – which is derived from information theoretic principles with analogies to statistical physics, and was successfully employed in several problems including vector quantization, classification and regression. DA has several useful properties including reduced sensitivity to initialization and strong potential to avoid poor local minima. DA-based optimization methods are developed here for the following fundamental communication problems: the Wyner-Ziv setting where only a decoder has access to side information, the distributed setting where independent encoders transmit over independent channels to a central de-

coder, and analog multiple descriptions setting which is an extension of the well known source coding problem of multiple descriptions. Comparative numerical results are presented, which show strict superiority of the proposed method over gradient descent based optimization methods as well as prior approaches in literature. Detailed analysis of the highly non-trivial structure of obtained mappings is provided.

The thesis further studies the related problem of global optimization of controller mappings in decentralized stochastic control problems, including Witsenhausen's celebrated 1968 counter-example. It is well-known that most decentralized control problems do not admit closed-form solutions and require numerical optimization. An optimization method is developed, based on DA, for a class of decentralized stochastic control problems. Comparative numerical results are presented for two test problems that show strict superiority of the proposed method over prior approaches in literature, and analyze the structure of obtained controller functions.

Contents

List of Figures

Chapter 1

Introduction

Shannon's information theory [63], which has paved the way for the modern communication age, has various shortcomings as we advance to more complicated emerging networks. The classical communication theory usually assumed point-to-point communications, allowed infinite block length (hence long delay) and unbounded complexity of source and channel coding. Digital communications have proliferated due to advanced source compression and error control techniques despite the aforementioned shortcomings, namely, substantial delay and complexity. On the other hand, various emerging communication applications are complex networks that have strict delay and complexity constraints. The problem of obtaining the optimal coding schemes at finite delay is therefore an important open problem with considerable practical implications [64, 38, 19, 11, 67, 54, 30].

As an example practical setting where strict delay and resource constraints are present, consider neural activity monitoring, where neural sensors implanted into the body are used as an interface to the nervous system [62]. Neural implants are employed to explore neuronal networks and the inextricable links between environmental stimuli and neuronal signaling, behavior and control [62]. The system can be used for advanced research and treatment of neurodegenerative diseases. As most practical uses are highly interactive, the communication network is strictly delay limited. In order to avoid damage to live tissue due to heating, the power consumption needs to be extremely constrained [50]. Moreover, the sensors are extremely small, making only relatively simple circuitry feasible [61]. Digital systems entail long delay and require complex digital circuitry that occupies space and dissipates heat, and are therefore not suitable for neural activity monitoring. This type of extreme application strongly motivates the theory and methods for delay and complexity-constrained networking that we pursue in this work.

1.1 Optimality of Uncoded Transmission

While it is well known that finite-delay coding schemes do not achieve the asymptotic bounds in general (see, e.g., [63, Theorem 21] or [22]), zero-delay communication is in fact optimal in some cases. An example would be communicating a binary uniform source over a binary symmetric channel with Hamming

distortion metric. Direct (uncoded) transmission of the binary sequence over the channel achieves optimal performance [21]. Similarly, optimal transmission of a Gaussian source over a channel with additive white Gaussian noise, under a power constraint and mean squared error distortion measure, can be achieved by simply transmitting the source values directly, with proper scaling to achieve power constraint [26]. In fact, a more general condition of when such uncoded transmission of a discrete memoryless source over a discrete memoryless channel is optimal is given in [22]. These results motivate an approach referred to as joint source-channel coding (JSCC). JSCC is an effective method to address the problems of long delay and high complexity of the separation strategy. Recently, there has been growing interest in utilizing zero-delay mappings in network applications, see, e.g., [16, 39] for coding over multiple access channels, [69, 17, 70] for distributed coding of correlated sources and [14, 4] for analog multiple description coding. However, there are no known methods to find optimal low-delay JSCC strategies for general networks.

1.2 Numerical Optimization

Until recently, there have been two main approaches to numerical optimization of the mappings: i) Optimization of the parameter set of a structured mapping [35, 71, 54, 30]. The performance of this approach is limited to the parametric

form (structure) assumed. For example, in [9] saw-tooth like structure is assumed for the mapping in the Wyner-Ziv setting and parameters of such mappings are optimized. ii) Design based on power constrained channel optimized vector quantization where a discretized version of the problem is tackled using tools developed for vector quantization [19, 18, 37].

In this thesis, we optimize encoder and decoder mappings to achieve good zero-delay communication strategies for real-time networking applications. Our approach builds on recent prior work in our lab [2] where the problem is studied in the original functional domain, i.e., without any discretization in the problem formulation and without any assumption of a parametrized mapping. Necessary conditions for optimality of mappings were derived, noting that while such conditions have theoretical value, they generally identify local optima. They are practically useless in the case of highly complex cost surfaces. In other words, simple greedy methods that are based on iterative imposition of necessary conditions of optimality tend to get trapped in poor local minima. In [2], "noisy channel relaxation" (NCR) [20] was employed to mitigate this problem inherent to such optimization problems. As we show in this work, while NCR is reasonably effective for simple settings, using more advanced non-convex optimization tools improves the performance significantly in sophisticated network scenarios.

1.3 Deterministic Annealing

In this dissertation, we propose a method based on a powerful non-convex optimization framework, *deterministic annealing*, to numerically approach globally optimal zero-delay mappings in network scenarios. Deterministic annealing (DA) is derived within a probabilistic framework where the main idea is to introduce controlled randomization into the optimization process, yet deterministically optimize the appropriate expectation functionals. The application-specific cost is minimized at successive stages of decreasing randomness, and a nonrandom solution is ultimately obtained while avoiding many poor local minima. Based on information theoretic principles with analogies to statistical physics, DA has been successfully used in non-convex optimization problems including clustering [57], vector quantization [58], regression [55] and more (see review in [56]).

We note that DA has been traditionally used in discrete settings such as quantizer optimization, and integrating DA within the mapping optimization framework in here poses a significant challenge. There are many important advantages of the proposed DA-based method compared to gradient descent based methods, including ability to avoid poor local minima and independence from initialization; and optimization in the original functional domain without any discretization or simplifying assumptions. Our approach improves significantly over prior approaches, some of which are NCR based.

1.4 Decentralized Stochastic Control

Decentralized control systems have multiple controllers designed to collaboratively achieve a common objective while taking actions based on their individual observations. No controller, in general, has direct access to the observations of the other controllers. This makes the design of optimal decentralized control systems a challenging problem. One of the most studied structures, termed "linear quadratic Gaussian" (LQG), involves linear dynamics, quadratic cost functions and Gaussian variables. Since in the case of centralized LQG problems, the optimal mappings are linear, it was naturally conjectured that linear control mappings remain optimal in decentralized settings. However, Witsenhausen proposed an example of a decentralized LQG control problem, commonly referred to as Witsenhausen's counter-example (WCE), for which he provided a simple non-linear control strategy that outperforms all linear strategies [73]. The problem has been viewed as a benchmark in stochastic networked control, see [81] for a detailed treatment.

We observe that the problem of finding optimal controller mappings in decentralized control is similar to finding optimum communication techniques in our zero-delay network settings. Consequently, we develop a general non-convex optimization method, inspired by the approaches we developed in the communication setting, which is suitable for a class of decentralized control problems. We present

comparative numerical results for two test problems that show strict superiority of the proposed method over prior approaches in literature.

Chapter 2

Side Information Setting

2.1 Introduction

In this chapter, we focus on what is perhaps the most simple networking communication problem that we refer to as the side information setting. The sender wishes to communicate the source over a discrete memoryless channel to a receiver which decodes the source under a minimum mean squared error (MSE) distortion measure. The receiver, but not the sender, has access to some side information that is correlated with the source. This setting has been studied extensively theoretically [66, 77, 1] and the optimal asymptotically achievable performance is known [78]. For practical coding approaches for this setting, see, e.g., [52].

We propose an optimization method based on deterministic annealing (DA), to numerically approach globally optimal zero-delay mappings in this setting. Having a powerful optimization method at hand, we analyze the structure of experimentally obtained mappings and investigate some conjectures made in prior work. For instance, one such claim was on the structure of optimal mappings in the side information setting, for which our results provide contradictory experimental evidence.

2.2 Preliminaries

2.2.1 Notations

Let $\mathbb{R}$, $\mathbb{N}$, and $\mathbb{R}^+$ denote the respective sets of real numbers, natural numbers, and positive real numbers. We represent scalars and random variables with lowercase and uppercase letters (e.g., x and X), column vectors and random column vectors with boldface lowercase and uppercase letters (e.g., $\boldsymbol{x}$ and $\boldsymbol{X}$), respectively. $\|\cdot\|$ denotes L_2 norm operator. Let $\mathbb{E}(\cdot)$ and $\mathbb{P}(\cdot)$ denote the expectation and probability operators, respectively. The probability density function of the random variable X is $f_X(x)$. Let ∇ and ∇_x denote the gradient and partial gradient with respect to x, respectively. Let $f^{'}(x) = \frac{df(x)}{dx}$ denote the first-order derivative of the continuously differentiable function f. The Gaussian density

with mean $\boldsymbol{\mu}$ and covariance matrix R is denoted as $\mathcal{N}(\boldsymbol{\mu}, R)$. We use natural logarithms which, in general, may be complex, and the integrals are, in general, Lebesgue integrals.

2.2.2 Problem Definition

In the side information setting, given in Figure 2.1, side information $\boldsymbol{Z} \in \mathbb{R}^{m_2}$ is available to the decoder, while source $\boldsymbol{X} \in \mathbb{R}^{m_1}$ is mapped to a channel input by the encoding function $\boldsymbol{g} : \mathbb{R}^{m_1} \to \mathbb{R}^p$ and transmitted over the channel with additive noise $\boldsymbol{N} \in \mathbb{R}^p$. The received channel output $\boldsymbol{Y} = \boldsymbol{g}(\boldsymbol{X}) + \boldsymbol{N}$ and side information $\boldsymbol{Z}$ are mapped to the estimate $\hat{\boldsymbol{X}}$ by the decoding function $\boldsymbol{w} : \mathbb{R}^p \times \mathbb{R}^{m_2} \to \mathbb{R}^{m_1}$. The problem is to find optimal mappings $\boldsymbol{g}, \boldsymbol{w}$, where optimality is in the sense that they minimize MSE

$$D(\boldsymbol{g}, \boldsymbol{w}) = \mathbb{E}\{\|\boldsymbol{X} - \hat{\boldsymbol{X}})\|^2\}, \tag{2.1}$$

subject to some power constraint on the encoder

$$P(\boldsymbol{g}) = \mathbb{E}\{\|\boldsymbol{g}(\boldsymbol{X})\|^2\} \leq P_E \tag{2.2}$$

where $P_E > 0$ is the specified encoder power level. Simple time-sharing arguments show that D is a convex functional of P, hence the solution is achieved at $P = P_E$ (see [3] for details.) Converting to Lagrangian formulation, we define the following

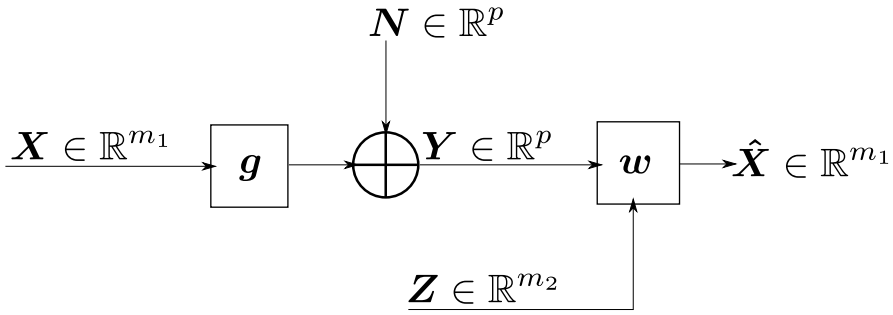


Figure 2.1: The side information setting.

cost function to be minimized

$$J = D(\boldsymbol{g}, \boldsymbol{w}) + \lambda P(\boldsymbol{g}) \tag{2.3}$$

where λ is a Lagrange multiplier corresponding to the power constraint on the encoder (we suppressed the dependence of J on $\boldsymbol{g}$ and $\boldsymbol{w}$).

2.2.3 Prior Work: Necessary Conditions of Optimality and Greedy Descent Algorithms

Here, we summarize the relevant contributions of prior work (see [2] for more details). Let the encoder $\boldsymbol{g}$ be fixed. Then, the optimal decoder is the MSE estimator of $\boldsymbol{X}$ given $\boldsymbol{Z} = \boldsymbol{z}$ and $\boldsymbol{Y} = \boldsymbol{y}$:

$$\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{z}) = \mathbb{E}\{\boldsymbol{X}|\boldsymbol{y}, \boldsymbol{z}\}. \tag{2.4}$$

Expanding the expressions for expectation and applying Bayes' rule, the optimal decoder can be written in terms of known quantities as

$$\boldsymbol{w}(\boldsymbol{y},\boldsymbol{z}) = \frac{\int \boldsymbol{x}\, f_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{z})\, f_{\boldsymbol{N}}(\boldsymbol{y}-\boldsymbol{g}(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}}{\int\ f_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x},\boldsymbol{z})\, f_{\boldsymbol{N}}(\boldsymbol{y}-\boldsymbol{g}(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}}, \tag{2.5}$$

where we used the fact that $f_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y},\boldsymbol{x}) = f_{\boldsymbol{N}}(\boldsymbol{y}-\boldsymbol{g}(\boldsymbol{x}))$. For optimality of $\boldsymbol{g}$, assuming the decoder $\boldsymbol{w}$ is fixed, a necessary condition is

$$\nabla_{\boldsymbol{g}} J(\boldsymbol{g},\boldsymbol{w}) = 0, \tag{2.6}$$

where

$$\begin{aligned}\nabla_{\boldsymbol{g}} J(\boldsymbol{g},\boldsymbol{w}) = &\lambda f_{\boldsymbol{X}}(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x}) \\ &-\mathbb{E}\{\boldsymbol{w}'(\boldsymbol{g}(\boldsymbol{x})+\boldsymbol{N},\boldsymbol{Z})(\boldsymbol{x}-\boldsymbol{w}(\boldsymbol{g}(\boldsymbol{x})+\boldsymbol{N},\boldsymbol{Z}))\},\end{aligned} \tag{2.7}$$

and $\boldsymbol{w}'$ denotes the Jacobian of $\boldsymbol{w}$ with respect to its first argument.

Remark 2.2.1. *Note that in the case of jointly Gaussian sources and Gaussian channel(s) with matched source-channel dimensions, linear mappings satisfy the necessary conditions of optimality, however, they are highly suboptimal. As we will see, careful optimization obtains considerably better mappings that are far from linear.*

Iteratively alternating between the imposition of individual necessary conditions of optimality will successively decrease the Lagrangian cost until a stationary point is reached. We refer to this method as "greedy descent". There is no

reason to expect that a greedy descent algorithm will converge to the globally optimal solution. In fact, experiments show severe issues of local optima and strong dependence on initialization of such methods. As a remedy, the noisy channel relaxation (NCR) method of [20] was embedded in the algorithm in [2], i.e., the descent method was run at gradually decreasing levels of λ, wherein the result at each level serves as initialization for the next level of λ (see [20] for details). While such simple relaxations are effective in simple communication settings, the networked problem we consider here requires a stronger optimization approach.

2.2.4 Asymptotically Achievable Limits

It is insightful to consider asymptotic bounds, which are obtained at infinite delay, while keeping in mind that the problem we consider is delay limited. Let $R(D)$ and $C(P)$ denote the source rate-distortion function and channel capacity, respectively. According to Shannon's source and channel coding theorems, the source can be compressed to $R(D)$ bits (per source sample) at distortion level D, and that $C(P)$ bits can be transmitted over the channel (per channel use) with arbitrarily low probability of error (see, e.g., [12]). The optimal coding scheme is the tandem combination of the optimal source and channel coding schemes, hence, by setting

$$R(D) = C(P), \tag{2.8}$$

one obtains a lower bound on the distortion of any source-channel coding scheme. For simplicity, we derive the expressions for the "optimum performance theoretically attainable" (OPTA) for Gaussian scalar source and noise. The channel capacity with additive white Gaussian noise is given by

$$C(P) = \frac{1}{2}\log\left(1 + \frac{P}{\sigma_N^2}\right), \tag{2.9}$$

where P is the transmission power and σ_N^2 is the noise variance.

For our setting, OPTA can be obtained by equating Wyner-Ziv rate distortion function [78] to the channel capacity. The Wyner-Ziv rate distortion function of X, when Z serves as side information, and $(X, Z) \sim \mathcal{N}(\mathbf{0}, R_{X,Z})$ where $R_{X,Z} = \sigma_X^2 \left[\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right]$ and σ_X^2, ρ are the variance and correlation coefficient, respectively, with $|\rho| \leq 1$ is:

$$R(D) = \max\left(0, \frac{1}{2}\log\frac{(1-\rho^2)\sigma_X^2}{D}\right). \tag{2.10}$$

We plug (2.10) and (2.9) in (2.8) to obtain

$$D_{OPTA} = \frac{(1-\rho^2)\sigma_X^2}{\left(1 + \frac{P_T}{\sigma_N^2}\right)}. \tag{2.11}$$

2.3 Proposed Method

2.3.1 Overview

In this section, we develop the DA based method for the optimization of encoder and decoder mappings. Since the decoder is given in closed form, the method focuses on optimizing the encoder mapping. We first partition the input space of the encoder into partition cells and assign a local model to each of the cells. Next, the encoder output is made probabilistic by randomizing the partitions, i.e., input points are assigned to each local model according to some probability distribution. We then propose an optimization process where the (random) encoder is optimized (along with the decoder) while constraining the Shannon entropy. By gradually reducing the entropy to 0, we obtain the desired mappings.

2.3.2 Derivation of proposed method

We consider piecewise functions which approximate the desired mappings by partitioning the space and matching a simple local model to each region. Piecewise functions consist of two components: a space partition and a parametric local model per partition cell. First, the source space $\mathbb{R}^m$ is partitioned into $\mathcal{K}$ regions (cells) denoted $\mathcal{R}_k^m$. Each cell $\mathcal{R}_k^m$ has an associated function $\boldsymbol{g}_k$ which is parametrized (affine, lattice, etc.) and the parameter set is denoted by Λ_k. Thus,

the encoding function can be written as

$$\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{g}_k(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \mathcal{R}_k^m \text{ and for } k = 1, \ldots, \mathcal{K} \tag{2.12}$$

In (2.12), the selection of local model index k is deterministic for a given realization of $\boldsymbol{X}$, i.e., the output of the encoder only depends on $\boldsymbol{X}$. To derive a DA based approach, we introduce a random variable, K, that corresponds to random selection of index k. In other words, let the encoder randomly select the local model index k when it receives an input $\boldsymbol{x}$, according to the value of a random variable that we call K. For a given realization of $\boldsymbol{X}$, the output of the encoder is now given in probability as

$$\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{g}_k(\boldsymbol{x}) \text{ with probability } p_{K|\boldsymbol{X}}(k|\boldsymbol{x}). \tag{2.13}$$

The conditional probability $p_{K|\boldsymbol{X}}(k|\boldsymbol{x})$ is referred to as *association probability*, in the sense that it represents the probability of input point $\boldsymbol{x}$ belonging to cell $\mathcal{R}_k^m$ (thus, the source space partition is now random). The probability distribution that we introduce (and optimize) is $p_{K|\boldsymbol{X}}$ (not the joint $p_{\boldsymbol{X},K}$) since the input distribution is given in the problem statement and is therefore fixed. The MSE cost and transmission power are still calculated as in (2.1) and (2.2), though the expectation is now taken over K in addition to what was done before. Let us rewrite (2.3) accounting for K:

$$J = \int_{\mathbb{R}^{m_1}} \left[\sum_{k=1}^{\mathcal{K}} J_k(\boldsymbol{x}) p_{K|\boldsymbol{X}}(k|\boldsymbol{x}) \right] f_{\boldsymbol{X}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \tag{2.14}$$

where

$$J_k(\boldsymbol{x}) = \mathbb{E}\{\|\boldsymbol{x} - \boldsymbol{w}(\boldsymbol{g}_k(\boldsymbol{x}) + \boldsymbol{N}, \boldsymbol{Z})\|^2\} + \lambda \|\boldsymbol{g}_k(\boldsymbol{x})\|^2. \tag{2.15}$$

We have the following lemma:

Lemma 2.3.1. *The minimum of (2.14) is achieved by hard probabilities, i.e., for given* $\boldsymbol{x}$, $p_{K|\boldsymbol{X}}(k|\boldsymbol{x}) = 1$ *for* k *that minimize* $J_k(\boldsymbol{x})$.

Proof. Let us fix Λ_k and $\boldsymbol{w}$, and consider optimizing (2.14) with respect to $p_{K|\boldsymbol{X}}$. It is clear that the optimal $p_{K|\boldsymbol{X}}$ will implement 'hard' associations, that is, every point $\boldsymbol{x}$ will be fully associated with the local model that makes the minimum contribution to cost. □

According to Lemma 2.3.1, the generalized search space of random encoders have the same global minimum as the original problem. Although this is desirable eventually, in order to avoid poor local optima we impose and control the level of randomness, i.e., we introduce a constraint on the randomness of the encoder, which is measured by the Shannon entropy. The total entropy of the encoder is given by $H(\boldsymbol{X}, K) = H(\boldsymbol{X}) + H(K|\boldsymbol{X})$ and since $H(\boldsymbol{X})$ is constant (predetermined by the source), the entropic quantity of interest is the conditional entropy $H(K|\boldsymbol{X})$. This is also intuitively justified in the sense that the randomness we

introduced into the problem is precisely captured by $p_{K|\boldsymbol{X}}$, hence can be measured and controlled by $H(K|\boldsymbol{X})$. We denote the randomness of the solution by H and define it as $H \triangleq H(K|\boldsymbol{X})$ where

$$H(K|\boldsymbol{X}) = -\mathbb{E}\{\log p_{K|\boldsymbol{X}}\}. \tag{2.16}$$

The problem is now recast as minimization of the expected cost with respect to parameters of local models, association probabilities and decoder, subject to a constraint on the level of randomness of the system, i.e.,

$$\begin{aligned} \underset{\Lambda_1,\ldots,\Lambda_{\mathcal{K}},p(1|\boldsymbol{x}),\ldots,p(\mathcal{K}|\boldsymbol{x}),\boldsymbol{w}}{\text{minimize}} \quad & J, \\ \text{subject to} \quad & H \geq H_0, \end{aligned}$$

where J is defined in (2.14) and H_0 specifies the minimum requirement on the entropy level. This constrained optimization problem can be reformulated by introducing Lagrange parameter $T \in \mathbb{R}^+$ to obtain the Lagrangian

$$F = J - TH, \tag{2.17}$$

to be minimized. There are two important extremal points of this Lagrangian. First, for $T \to \infty$, the minimum F is obtained by maximizing the entropy, which is achieved by uniform association probabilities: $p_{K|\boldsymbol{X}}(k|\boldsymbol{x}) = 1/\mathcal{K}$ for all k and $\boldsymbol{x}$. Consequently, all local models equally account for all points and are identical once optimized, or effectively, there is a single *distinct* local model. Secondly, in the limit $T \to 0$, minimizing F corresponds to minimizing J directly, which

produces a deterministic encoder. This intuitive observation can be verified by the expression for optimal $p_{K|\boldsymbol{X}}(k|\boldsymbol{x})$ given in Section 2.3.4.

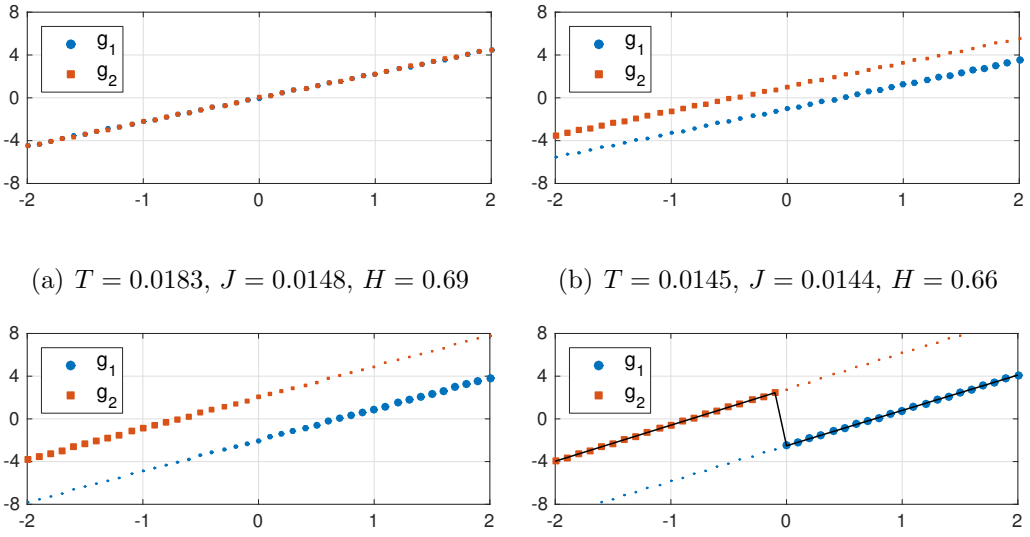
Although DA is derived from information theoretic principles, it is motivated by and has strong analogies to annealing processes in statistical physics. The Lagrangian functional in (2.17) can be viewed as the Helmholtz free energy of a corresponding physical system, where J is the thermodynamic energy and H is the entropy of the system, and Lagrange parameter T is the "temperature". This analogy suggests the possibility of implementing an annealing process, where the temperature is gradually lowered while the system is kept at thermal equilibrium, i.e., the free energy is at minimum. The annealing process is started at a high temperature (highly random mappings) where, in fact, the entropy is maximized (single local model). This minimum is then tracked at successively lower temperatures (which corresponds to lower levels of entropy) as the system undergoes a sequence of phase transitions through which the model complexity (the number of distinct local models) grows. As the temperature approaches zero, the physical system converges to ground state (global minimum of the energy). Similarly, as $T \to 0$, we obtain a hard (nonrandom) mapping while avoiding poor local minima. We note, however, that DA method does not guarantee to find the globally optimum solution in general, only when certain continuity conditions are satisfied by the phase transitions.

2.3.3 Deterministic Annealing

The optimization method starts at a high value of T and gradually lowers it while minimizing F at each step. At high temperature, there is effectively a single distinct local model. As the temperature is decreased, a bifurcation point is reached where the current solution is no longer a minimum, so that there exists a better solution with a higher number of distinct local models. Intuitively, at this temperature, the current solution is a saddle point where multiple local models are coincident (i.e., their parameters are same) and in order to move to a better solution, it is necessary to perturb the local models. Such bifurcations are referred to as "phase transitions" and the corresponding temperatures are called "critical temperatures"[1].

We present an example simulation in Figure 2.2 that illustrates the basics of the method, including phase transitions. Here the sources and channel are scalar, i.e.,$m = n = 1$, g_k are selected as affine and $\mathcal{K} = 2$. When T is large, there is a single distinct local model. As we lower T, the system goes through a phase transition where the two local models split from each other (after a slight perturbation). The corresponding value of T is referred to as the first critical temperature. Note how entropy (H) is traded for reduction in cost (J).

[1]We omit the derivation of critical temperatures in this thesis, see [56] for phase transition analysis in the general DA setting.



(a) $T = 0.0183$, $J = 0.0148$, $H = 0.69$

(b) $T = 0.0145$, $J = 0.0144$, $H = 0.66$

(c) $T = 0.0094$, $J = 0.0120$, $H = 0.40$

(d) $T = 0.0001$, $J = 0.0097$, $H = 0.0$

Figure 2.2: The evolution of the encoder in the algorithm is demonstrated. The two models are shown by dotted lines and the sizes of dots are relative to the probability association at that input point. The line in (d) is the deterministic encoder obtained. $\mathcal{K} = 2$.

Mappings with more than 2 local models can be obtained by starting with a larger $\mathcal{K}$. However, a computationally more efficient method that we employ here is as follows: We start with 1 local model and keep only the distinct local models, but duplicate and perturb them at each temperature. The duplicates will merge at every iteration until a critical temperature is reached, and will split into distinct models at a phase transition.

Although our method is derived in the general, continuous source and channel domain, in practical simulations we sample the source and noise distributions to allow numerical computation of integrals. The sampling is not "inherent" to the derived method and, in fact, can be adjusted during the algorithm run. We emphasize that this is in contrast with prior quantizer design based methods that are entirely formulated in a discrete setting.

The practical algorithm is initialized with a single local model. Since T must be set higher than the first critical temperature, we simply choose T large enough that during the first couple of temperatures, duplicated local models merge back, i.e., no phase transitions are observed. As the temperature is gradually lowered, we track the minimum, i.e., find the association probabilities $p_{K|\boldsymbol{X}}(k|\boldsymbol{x})$, local model parameters Λ_k and decoder $\boldsymbol{w}$ that minimize the Lagrangian F. As demonstrated, the system will go through phase transitions during which the number of local models, $\mathcal{K}$, increases. We stop when T is near 0 and perform "zero entropy

iteration", i.e., associate every source point with the "best" local model to obtain deterministic encoder. We accordingly give a brief sketch of the practical method in Algorithm 1. In Step 6, we employed an exponential cooling schedule. Update equations for Step 3 are given in the next section.

2.3.4 Update Equations

The central part of the method is the minimization of free energy (F) by iteratively updating the association probabilities, local model parameters and decoders. The following theorem states the update equations for association probabilities.

Theorem 2.3.2. *At any temperature T, minimum free energy F is achieved when association probabilities are in the form of Gibbs distribution given as:*

$$p(k|\boldsymbol{x}) = \frac{e^{-J_k(\boldsymbol{x})/T}}{\sum_{k'} e^{-J_{k'}(\boldsymbol{x})/T}} \quad \forall k, \tag{2.18}$$

where $J_k(\boldsymbol{x})$ is given by (2.15).

Proof. We write the Lagrangian cost in (2.17) as

$$F = \int_{\mathbb{R}^{m_1}} \left[\sum_{k=1}^{\mathcal{K}} J_k(\boldsymbol{x}) p_{K|\boldsymbol{X}}(k|\boldsymbol{x}) \right] f_{\boldsymbol{X}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + T \sum_k \int_{\boldsymbol{x}} p(k|\boldsymbol{x}) \log p(k|\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \tag{2.19}$$

where $J_k(\boldsymbol{x})$ is given in (2.15). From (2.19) it can be seen that F is convex in $p(k|\boldsymbol{x})$, since first term is linear and second term is convex in $p(k|\boldsymbol{x})$. To find the

Algorithm 1 Proposed DA-Based Method

Inputs: *Involved distributions, desired local model type,* λ, α, ϵ, Δ_F, T_{min}, $\Delta_{\boldsymbol{g}}$.

Outputs: *Optimized* $\boldsymbol{g}, \boldsymbol{w}$.

Initialization: $T = T_{max}$, $\mathcal{K} = 1$, *randomly chosen* $\boldsymbol{g}_1$. $J_{old} = J_{initial}$.

1. ***Duplication:***

   For each $\boldsymbol{g}_i$, *create an identical local model* $\boldsymbol{g}_j$.

   $p(i|\boldsymbol{x}) \leftarrow \frac{p(i|\boldsymbol{x})}{2}$ *and* $p(j|\boldsymbol{x}) \leftarrow \frac{p(i|\boldsymbol{x})}{2}$.

   $\mathcal{K} \leftarrow 2\mathcal{K}$.

2. ***Perturbation:***

   For each parameter $\phi_k \in \boldsymbol{\Lambda}_k$, $\phi_k \leftarrow \phi_k + \epsilon R$, *where* R *is standard Gaussian random variable.*

3. ***Thermal Equilibrium:***

   Compute F *and set* $F_{old} \leftarrow F$.

   3.1. *Compute optimal* $\boldsymbol{w}$ *using (2.26).*

   3.2. *Compute optimal* $p(k|\boldsymbol{x})$, $\forall k$ *using (2.18).*

   3.3. *Optimize* Λ_k, $\forall k$ *using (2.24).*

   3.4. *Compute* F. *If* $\frac{F - F_{old}}{F_{old}} \leq \Delta_F$, *go to Step 4, otherwise* $F_{old} \leftarrow F$ *and go to Step 3.1.*

4. **Model Size:**

   If $d(\Lambda_i, \Lambda_j) < \Delta_{\boldsymbol{g}}$, where $d(\cdot, \cdot)$ is euclidean distance, remove $\boldsymbol{g}_j$ and set $p(i|\boldsymbol{x}) \leftarrow p(i|\boldsymbol{x}) + p(j|\boldsymbol{x})$, $\forall i, j$.

   $\mathcal{K} \leftarrow$ New model size.

5. **Stopping:**

   Stop if $T \leq T_{min}$, otherwise go to Step 6.

6. **Cooling:**

   $T \leftarrow T * \alpha$.

   Go to Step 1.

---

minimum, we set $\nabla_{p(k|\boldsymbol{x})} F = 0$:

$$J_k(\boldsymbol{x}) + T \log p(k|\boldsymbol{x}) + T = 0, \tag{2.20}$$

which yields

$$p(k|\boldsymbol{x}) = Ce^{-(J_k(\boldsymbol{x})-T)/T}. \tag{2.21}$$

The normalizing factor C is to ensure that

$$\sum_k p(k|\boldsymbol{x}) = 1. \tag{2.22}$$

Plugging (2.21) in (2.22), we have

$$C = \frac{1}{\sum_{k'} e^{-(J_{k'}(\boldsymbol{x})-T)/T}}. \tag{2.23}$$

Plugging (2.23) in (2.21) yields (2.18). □

Remark 2.3.3. *Theorem 2.3.2 is analogous to the principle of minimal free energy in statistical physics. A fundamental principle in statistical physics states that the minimum free energy is achieved when the system is at thermal equilibrium, at which point it is governed by Gibbs distribution.*

The evolution of association probabilities, $p(k|\boldsymbol{x})$, during the annealing process can be observed from how (2.18) is changing with T. The following corollary confirms the intuitive explanation we provided earlier.

Corollary 2.3.4. *As $T \to \infty$ (at a high temperature) the system is governed by uniform association probabilities and the entropy is maximum. As $T \to 0$, the associations become deterministic and the entropy is* 0.

The optimal local model parameters cannot be obtained in closed form, hence we perform gradient descent search. A local model parameter $\phi_k \in \Lambda_k$ is updated according to

$$\phi_k \leftarrow \phi_k - \varphi \frac{\partial F}{\partial \phi_k} \tag{2.24}$$

where φ is selected by line search and the gradient can be obtained as

$$\frac{\partial F}{\partial \phi_k} = \frac{\partial J}{\partial \phi_k} = \int_{\boldsymbol{x}} f_{\boldsymbol{X}}(\boldsymbol{x}) p(k|\boldsymbol{x}) \frac{\partial J_k(\boldsymbol{x})}{\partial \phi_k} \mathrm{d}\boldsymbol{x}. \tag{2.25}$$

The derivative $\frac{\partial J_k(\boldsymbol{x})}{\partial \phi_k}$ is calculated numerically. The optimal decoder can be derived similar to (2.5):

$$\boldsymbol{w}(\boldsymbol{y}, \boldsymbol{z}) = \frac{\int \boldsymbol{x}\, f_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x}, \boldsymbol{z}) \sum_k f_{\boldsymbol{N}}(\boldsymbol{y} - \boldsymbol{g}_k(\boldsymbol{x})) p(k|\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}}{\int f_{\boldsymbol{X},\boldsymbol{Z}}(\boldsymbol{x}, \boldsymbol{z}) \sum_k f_{\boldsymbol{N}}(\boldsymbol{y} - \boldsymbol{g}_k(\boldsymbol{x})) p(k|\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}}. \tag{2.26}$$

2.3.5 Design Complexity

Due to difficulties in estimating the time required for gradient descent, exact comparison of computational complexity of numerical optimization methods (including the method presented here and others referred to in Section 2.2.3) is difficult and depends on the actual source-channel distributions as well as choice of various algorithm parameters. On the other hand, optimization of parametrized mappings (e.g., in [17]) is faster, but requires knowing the structure of a good solution, which can be obtained by methods such as the one presented here. In our experiments, the time required for DA was on the same order as that of NCR, albeit with a higher constant. Thus, better performance is obtained at the expense of slight increase in complexity.

2.4 Experimental Results

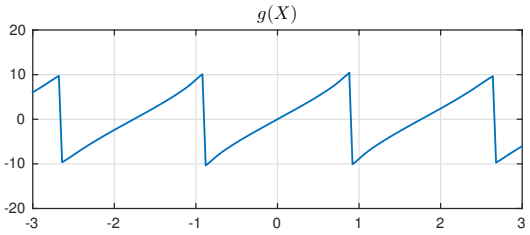
While the proposed algorithm is general and directly applicable to any choice of source and channel dimensions, for conciseness of the results section, we as-

sume that sources and channels are scalar. In this case, the encoder mapping is denoted as $g : \mathbb{R} \to \mathbb{R}$ and the local model functions g_k are selected as affine. In principle, the set of g_k can be chosen from any parametric model. Choosing a more complex model, such as a higher order polynomial, can potentially improve the performance of the algorithm, albeit with increased computational complexity. For the exponential cooling schedule, we set $\alpha = 0.95$, i.e., $T \leftarrow T * 0.95$. The performance of the proposed method is assessed by comparisons to the optimal affine solution, greedy method and NCR-based method, as well as OPTA (for reference only, as OPTA requires infinite delay). For the NCR based method, we decrease λ exponentially as $\lambda_{new} = \lambda_{old} * 0.8$ in 50 steps to the desired value.

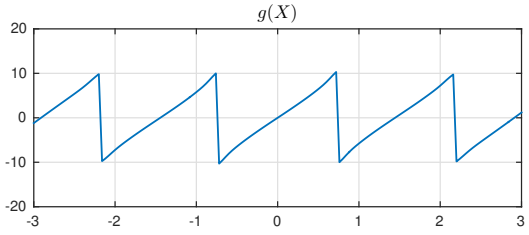
The noise signals in all examples are chosen as independent zero-mean Gaussians with unit variance, i.e., $N \sim \mathcal{N}(0, 1)$. For numerical computations we impose bounded support (-5σ to $+5\sigma$), i.e., we neglect tails of infinite support distributions in the examples.

We first give examples for the Gaussian case, where the source and side information are jointly Gaussian, distributed according to $\mathcal{N}(\boldsymbol{\mu}, R)$ where $\boldsymbol{\mu} = [0, 0]$, $R = \left[\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right]$, and $|\rho| < 1$ is the correlation coefficient between source and side information. We define $\text{SNR} = 10 \log_{10}(1/D)$ and $\text{CSNR} = 10 \log_{10}(P(g))$.

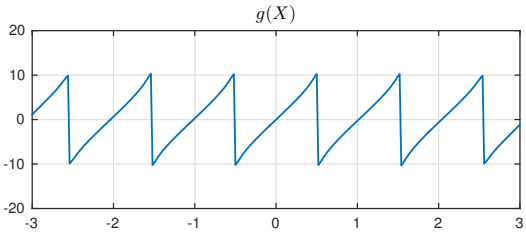
Example mappings are given in Figure 2.3. We first note that the central characteristics observed in digital Wyner-Ziv mappings are captured by the obtained



(a) SNR=21.2 dB, CSNR=15.0 dB, $\rho = 0.97$



(b) SNR=23.0 dB, CSNR=15.0 dB, $\rho = 0.98$

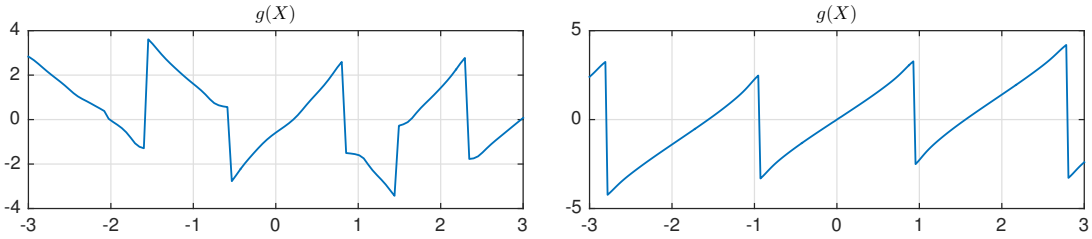


(c) SNR=26.0 dB, CSNR=15.0 dB, $\rho = 0.99$

Figure 2.3: Example encoder mappings, generated by DA, for the decoder side information setting, jointly Gaussian source and side information.

mappings as noted before (see, e.g., [2, 37]), in the sense of many-to-one mappings, where multiple source intervals are mapped to the same channel interval. We refer to each one-to-one section in these mappings as a "bin", in Figure 2.3a there are 5 bins in the interval shown (the meaning of bin here is different than in digital Wyner-Ziv mappings). The uncertainty about the source interval is resolved (significantly decreased) by the decoder using the side information. Since all variables are Gaussian and distortion measure is MSE, it is intuitively intriguing to investigate whether the optimal mappings have any parametric form or structure to be exploited in the design stage. For example, since in the absence of decoder side information optimal mappings are well known to be linear, one can expect to see linear mappings in each bin. In fact, such parametric form was explicitly assumed in [9], and it was reported the optimized parametric mappings perform very close to the results obtained via NCR in [2]. Our numerical results demonstrate that each bin is non-linear as some nonlinearity can be observed especially near the ends of each bin, as opposed to the conjecture in [2].

From Figure 2.3 we see how the width of bins depends on the correlation between the source and side information. It can be seen that at higher correlation the bins are narrower. This is intuitively expected since, as the correlation increases, so does the benefit of side information in terms of distinguishing different bins.



(a) SNR=20.9 dB, CSNR=14.3 dB, $\rho = 0.98$ (b) SNR=21.7 dB, CSNR=14.3 dB, $\rho = 0.98$

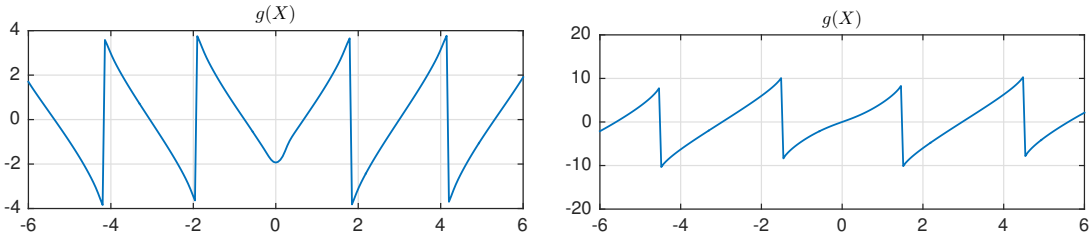
Figure 2.4: Two results by NCR for side information setting. In (a) the bins do not have the optimal shape that was obtained by DA and in (b) the discontinuity points are not optimal.

To exploit this capability, the encoder narrows the bins, which in turn reduces the power $\mathbb{E}\{g^2(X_1)\}$.

To illustrate the improvement of DA over NCR in the encoding mappings themselves, we present two mappings obtained by NCR in Figure 2.4. We emphasize that the performance of NCR depends on initial mappings, initial noise level and the noise-relaxation schedule. This dependence is illustrated in Figure 2.4, where in one case the shape of bins are different then those in DA and sub-optimal, and in the other the points of discontinuity are not optimal.

We also give an example with a different source distribution, Gaussian mixture, in Figure 2.5:

$$(X_1, X_2) \sim \left(\frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_1, R) + \frac{1}{2}\mathcal{N}(\boldsymbol{\mu}_2, R)\right) \tag{2.27}$$



(a) SNR=30 dB, CSNR=16 dB

(b) SNR=36 dB, CSNR=23 dB

Figure 2.5: Example encoder mappings, generated by DA, for Gaussian mixture distribution, side information setting.

where $\boldsymbol{\mu}_1 = [-3, -3]$, $\boldsymbol{\mu}_2 = [3, 3]$ and $R = \left[\begin{smallmatrix} 1 & 0.95 \\ 0.95 & 1 \end{smallmatrix}\right]$. This distribution has two Gaussian "nodes" centered far from each other at $x = -3$ and $x = 3$. From an intuitive point of view, the optimum encoder can be viewed as two Wyner-Ziv like encoders, occupying the negative and positive halves of real line and both centered at the node centers. It is clear that for several source and channel distributions, optimal encoding mappings are many-to-one, i.e., this property is not unique to the Gaussian distribution.

The comparative performance results for different optimization techniques is given in Figure 2.6 for correlation coefficient $\rho = 0.99$. Since NCR performance depends on the initial conditions, we ran the NCR algorithm several times with different conditions and pick the mappings with best performance. Results from the greedy method are also presented in order to illustrate the abundance of locally optimum points and the difficulty of the optimization problem. Note that the

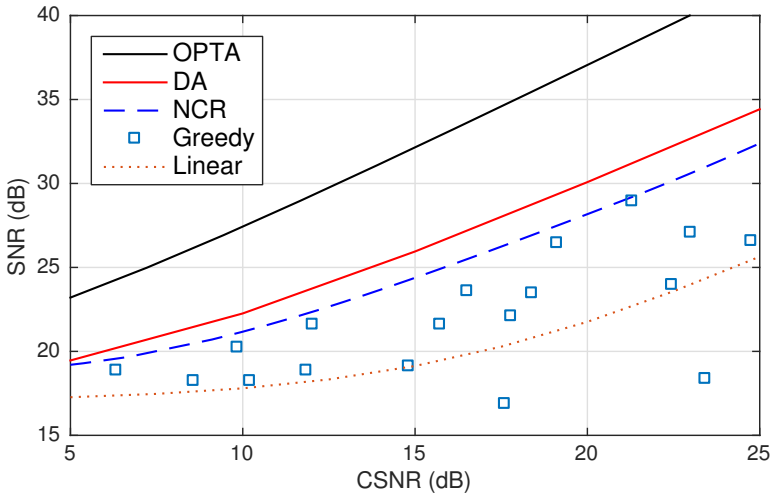


Figure 2.6: The performance comparison for the side information setting, the proposed method versus the noisy relaxation (NCR), greedy optimization and the linear mappings. $\rho = 0.99$.

proposed method is independent of the initialization and only run once. We also present the performance of OPTA as benchmark while noting that it is asymptotic and may require infinite delay. The performance of linear encoder and decoder is plotted as well, since it is also a local minimum (see Remark 2.2.1). It is important to note that the linear solution performs significantly worse than the non-linear mappings obtained.

Chapter 3

Distributed Coding

3.1 Introduction

In this chapter, we move to a more involved network setting that we refer to as the distributed coding. This setting, shown in Figure 3.1, involves distributed (separate) coding and transmission of two correlated sources to a central decoder that reconstructs individual sources. As an example practical scenario, consider a sensor network where sensor measurements are correlated, but sensors encode and communicate their measurements separately due to physical constraints. The distributed coding problem has been studied extensively, see, e.g., [66, 72].

The setting can easily be seen as an extension of the side information setting discussed in the previous chapter, such that each channel's output is used as side

information when decoding for the other channel. In fact, if the decoder's goal is to reconstruct one of the sources only (say $\boldsymbol{X}_1$), then this problem is referred to as "coding with a helper" where the second encoder ($\boldsymbol{g}_2$) provides "coded side information" to the decoder [76, 1].

If the decoder wants to estimate a function of the sources, the setting is referred to as "function computation problem". This is of interest for certain applications such as a wireless sensor network deployed in order to compute a function of the measurements [27, 23, 51, 47, 49].

We extend the method introduced in the previous chapter to this setting, to approach optimal encoder and decoder mappings. Our results strictly improve over recent competing approaches, as well as prior approaches in literature. Several practically important observations are made regarding the functional properties of the optimal mappings.

3.2 Preliminaries

3.2.1 Notations

Let $\mathbb{R}$, $\mathbb{N}$, and $\mathbb{R}^+$ denote the respective sets of real numbers, natural numbers, and positive real numbers. We represent scalars and random variables with lowercase and uppercase letters (e.g., x and X), column vectors and random column

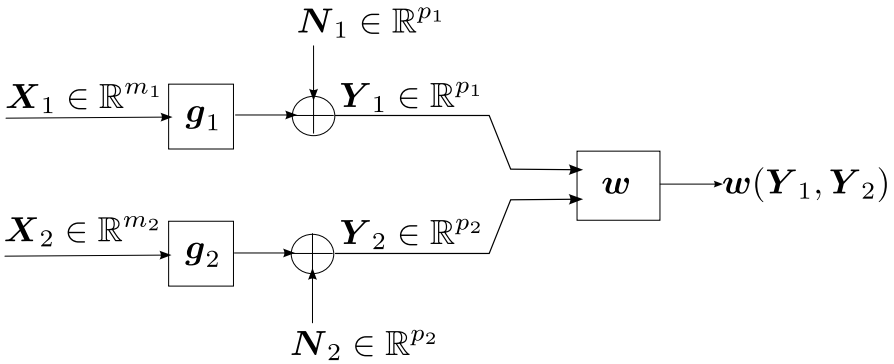


Figure 3.1: The distributed coding problem.

vectors with boldface lowercase and uppercase letters (e.g., $\boldsymbol{x}$ and $\boldsymbol{X}$), respectively. $\|\cdot\|$ denotes L_2 norm operator. Let $\mathbb{E}(\cdot)$ and $\mathbb{P}(\cdot)$ denote the expectation and probability operators, respectively. The probability density function of the random variable X is $f_X(x)$. Let ∇ and ∇_x denote the gradient and partial gradient with respect to x, respectively. Let $f'(x) = \frac{df(x)}{dx}$ denote the first-order derivative of the continuously differentiable function f. The Gaussian density with mean $\boldsymbol{\mu}$ and covariance matrix R is denoted as $\mathcal{N}(\boldsymbol{\mu}, R)$. We use natural logarithms which, in general, may be complex, and the integrals are, in general, Lebesgue integrals.

3.2.2 Problem Definition

The distributed coding setting, given in Figure 3.1, has two sources $\boldsymbol{X}_1 \in \mathbb{R}^{m_1}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{m_2}$ mapped to some channel input by the encoding functions $\boldsymbol{g}_i$:

$\mathbb{R}^{m_i} \to \mathbb{R}^{p_i}$, and the decoder receives $\boldsymbol{Y_i} = \boldsymbol{g}_i(\boldsymbol{X}_i) + \boldsymbol{N}_i$ for $i = 1, 2$. In general, the decoder might have two type of objectives. In the first one, the decoder aims to reconstruct each source with minimum distortion. The decoder is defined as $\boldsymbol{w} : \mathbb{R}^{p_1} \times \mathbb{R}^{p_2} \to \mathbb{R}^{m_1} \times \mathbb{R}^{m_2}$ as it maps the received channel outputs to the estimates $\hat{\boldsymbol{X}}_i$ for $i = 1, 2$. For this case, we define distortion as

$$D(\boldsymbol{g}_1, \boldsymbol{g}_2, \boldsymbol{w}) = \mathbb{E}\{\|\boldsymbol{X}_1 - \hat{\boldsymbol{X}}_1\|^2 + \eta\|\boldsymbol{X}_2 - \hat{\boldsymbol{X}}_2\|^2\} \tag{3.1}$$

where $\eta \in \mathbb{R}^+$ is a given weight coefficient. The case of $\eta = 0$ corresponds to the "coding with a helper" problem as mentioned in Section 3.1. The second type of objective is the function computation. Denoting the desired function as $\boldsymbol{\gamma}(\boldsymbol{X}_1, \boldsymbol{X}_2) : \mathbb{R}^{m_1} \times \mathbb{R}^{m_2} \to \mathbb{R}^r$, the decoder is defined as $\boldsymbol{w} : \mathbb{R}^{p_1} \times \mathbb{R}^{p_2} \to \mathbb{R}^r$ and the cost is given by

$$D(\boldsymbol{g}_1, \boldsymbol{g}_2, \boldsymbol{w}) = \mathbb{E}\{\|\boldsymbol{\gamma}(\boldsymbol{X}_1, \boldsymbol{X}_2) - \boldsymbol{w}(\boldsymbol{Y}_1, \boldsymbol{Y}_2)\|^2\}. \tag{3.2}$$

The problem, for both cases, is to find the mappings $\boldsymbol{g}_1, \boldsymbol{g}_2, \boldsymbol{w}$ that minimize the overall distortion (which is given in (3.1) or (3.2) depending on the objective) subject to power constraints on the encoders, which can be in two forms: Individual power constraints given by

$$P(\boldsymbol{g}_i) = \mathbb{E}\{\|\boldsymbol{g}_i(\boldsymbol{X}_i)\|^2\} \leq P_{T,i} \text{ for } i = 1, 2. \tag{3.3}$$

or a total power constraint of the form

$$\sum_{i=1}^{2} P(\boldsymbol{g}_i) \leq P_T, \tag{3.4}$$

which offers the additional degree of freedom of optimizing power allocations to the encoders. For optimization purposes, we similarly define the following Lagrangian functional as the objective cost to be minimized

$$J = D + \sum_{i=1}^{2} \lambda_i P(\boldsymbol{g}_i), \tag{3.5}$$

where $\lambda_i \in \mathbb{R}^+$, $i = 1, 2$, are Lagrange multipliers to impose the individual power constraints on the encoders in the first case. The total power constraint case corresponds to the special case of (3.5) with $\lambda_1 = \lambda_2 = \lambda$, i.e., the Lagrangian cost to minimize is

$$J = D + \lambda[P(\boldsymbol{g}_1) + P(\boldsymbol{g}_2)], \tag{3.6}$$

where λ controls the total power.

Necessary conditions of optimality can be derived for this problem, see [2].

3.2.3 Asymptotically Achievable Limits

For simplicity, we derive OPTA for quadratic Gaussian distributed source coding for sources $(X_1, X_2) \sim \mathcal{N}(\mathbf{0}, R_{X_1,X_2})$ where $R_{X_1,X_2} = \sigma_X^2 \left[\begin{smallmatrix} 1 & \rho \\ \rho & 1 \end{smallmatrix}\right]$ with $|\rho| \leq 1$.

The complete rate distortion region satisfies the following inequalities [68]:

$$R_1 \geq \frac{1}{2}\log^+\left(\frac{1-\rho^2+\rho^2 2^{-2R_2}}{D_1}\right) \tag{3.7}$$

$$R_2 \geq \frac{1}{2}\log^+\left(\frac{1-\rho^2+\rho^2 2^{-2R_1}}{D_2}\right) \tag{3.8}$$

$$R_1+R_2 \geq \frac{1}{2}\log^+\left(\frac{(1-\rho^2)\beta(D_1,D_2)}{2D_1D_2}\right) \tag{3.9}$$

where $\log^+ x = \max(0, \log x)$ and

$$\beta(D_1,D_2) = 1+\sqrt{1+\frac{4\rho^2 D_1 D_2}{(1-\rho^2)^2}}. \tag{3.10}$$

We set $R_i = C(P_i)$ for $i = 1, 2$, where $C(P)$ is given in (2.9) to obtain OPTA.

3.3 Method for Distributed Coding

Although the method described in the previous chapter can be used for optimizing the distributed encoders separately (within separate annealing processes), we found that such a method fails to avoid poor local minima as it fails to account for interaction between encoder optimizations. Instead, we develop a method here that optimizes the (random) encoders and decoders within a single annealing process. The resulting annealing method is an extension of the previous one with higher complexity due to the distributed nature of the problem.

We have two independent sets of partitions of input source space: $\mathcal{K}_1$ cells represented by $\mathcal{R}^m_{k_1}$ and $\mathcal{K}_2$ cells represented by $\mathcal{R}^m_{k_2}$. We define both encoders in

this setting as

$$\boldsymbol{g}_i(\boldsymbol{x}_i) = \boldsymbol{g}_{k_i}(\boldsymbol{x}_i) \text{ for } \boldsymbol{x}_i \in \mathbb{R}^m_{k_i},\ i = 1, 2. \tag{3.11}$$

Following the same procedure of randomization, we define random variables K_1 and K_2 along with association probabilities:

$$p(k_i|\boldsymbol{x}_i) \triangleq \mathbb{P}\{\boldsymbol{x}_i \in \mathbb{R}^m_{k_i}\}, \quad \forall k_i, \boldsymbol{x}_i, \text{ for } i = 1, 2. \tag{3.12}$$

The cost is to be minimized subject to the constraint on the joint entropy of the system. Noting that $K_1 \leftrightarrow X_1 \leftrightarrow X_2 \leftrightarrow K_2$ form a Markov chain by construction, we express the joint entropy as

$$H(\boldsymbol{X}_1,K_1,\boldsymbol{X}_2,K_2)=H(\boldsymbol{X}_1,\boldsymbol{X}_2)+H(K_1|\boldsymbol{X}_1)+H(K_2|\boldsymbol{X}_2). \tag{3.13}$$

Since $H(\boldsymbol{X}_1, \boldsymbol{X}_2)$ is a constant determined by the sources, we define $H \triangleq H(K_1|\boldsymbol{X}_1)+ H(K_2|\boldsymbol{X}_2)$ where

$$H(K_i|\boldsymbol{X}_i) = \mathbb{E}\{\log p(K_i|\boldsymbol{X}_i)\} \text{ for } i = 1, 2, \tag{3.14}$$

and the free energy of the system is, again, given by $F = J - TH$.

The algorithm sketch is similar to the side information setting and is not reproduced here in its entirety. Since we optimize both encoders within the same annealing process, the same operations in Algorithm 1 in Chapter 2 are performed for both encoders, sequentially.

The following theorem presents the optimal association probabilities for the distributed setting. The proof follows similar steps as in the proof of Theorem 2.3.2 and omitted for brevity.

Theorem 3.3.1. *At any temperature, minimum free energy (F) is achieved when the system is governed by Gibbs distribution given as:*

$$p(k_i|\boldsymbol{x}_i) = \frac{e^{-J_{k_i}(\boldsymbol{x}_i)/T}}{\sum\limits_{k_i'} e^{-J_{k_i'}(\boldsymbol{x}_i)/T}} \quad \textit{for } i = 1, 2 \tag{3.15}$$

where

$$\begin{aligned} J_{k_i}(\boldsymbol{x}_i) = & \mathbb{E}\{\|\boldsymbol{X}_1 - \hat{\boldsymbol{X}}_1\|^2 + \eta\|\boldsymbol{X}_2 - \hat{\boldsymbol{X}}_2\|^2 | \boldsymbol{X}_i = \boldsymbol{x}_i, K_i = k_i\} \\ & + \lambda_i \boldsymbol{g}_{k_i}^2(\boldsymbol{x}_i) \end{aligned} \tag{3.16}$$

if the cost is defined as in (3.1), and

$$\begin{aligned} J_{k_i}(\boldsymbol{x}_i) = & \mathbb{E}\{\|\boldsymbol{\gamma}(\boldsymbol{X}_1, \boldsymbol{X}_2) - \boldsymbol{w}(\boldsymbol{Y}_1, \boldsymbol{Y}_2)\|^2 | \boldsymbol{X}_i = \boldsymbol{x}_i, K_i = k_i\} \\ & + \lambda_i \boldsymbol{g}_{k_i}^2(\boldsymbol{x}_i) \end{aligned} \tag{3.17}$$

if the cost is defined as in (3.2).

Proof. Proven following the same steps in the proof of Theorem 2.3.2. □

The parameters of local models can be optimized through gradient descent search. Optimal decoding is achieved similarly as $\hat{\boldsymbol{X}}_i = \mathbb{E}\{\boldsymbol{X}_i|\boldsymbol{y}_1, \boldsymbol{y}_2\}$ for $i = 1, 2$ for first type of objective, and $\boldsymbol{w}(\boldsymbol{y}_1, \boldsymbol{y}_2) = \mathbb{E}\{\gamma(\boldsymbol{X}_1, \boldsymbol{X}_2)|\boldsymbol{y}_1, \boldsymbol{y}_2\}$ for the second

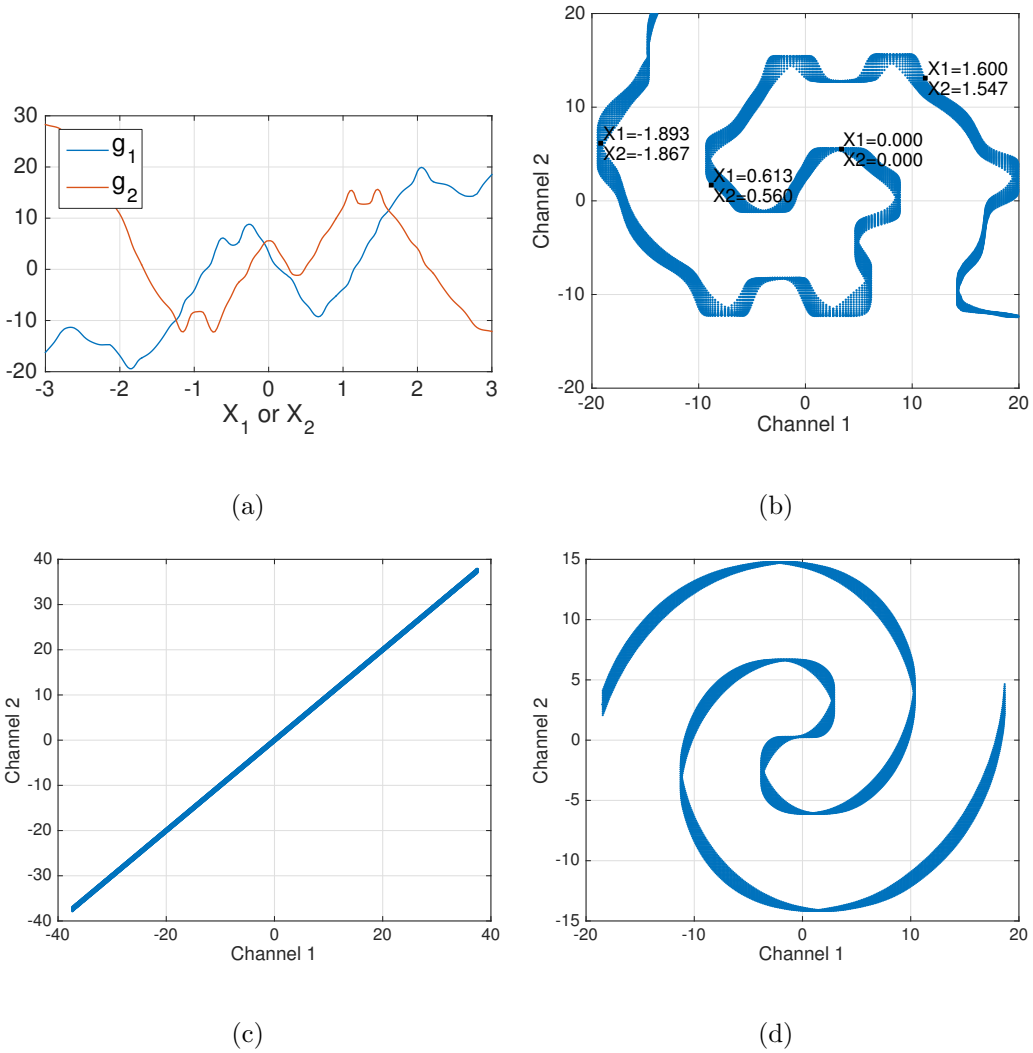
type. Both expressions can be written in terms of known quantities similar to that in (2.26).

3.4 Experimental Results

While the proposed algorithm is general and directly applicable to any choice of source and channel dimensions, for conciseness of the results section, we assume that sources and channels are scalar. The details of the simulations are the same as in Section 2.4.

In these experiments the sources are jointly Gaussian with unit variance and their correlation coefficient is denoted by ρ. We first analyze the case of individual reconstructions, where the cost is as defined in (3.1). The weighing coefficient η in (3.1) is taken as 1.

The encoding mappings observed are many-to-one, where an example is given in Figure 3.2a to gain intuition into the workings of these coding schemes. From Figure 3.2a, where both encoders are plotted together, we see that in different source intervals, one of the mappings is many-to-one while the other one is one-to-one (usually linear). For instance, in the interval $X \in [-0.3, 0.5]$, g_1 is approximately linear while g_2 is many-to-one. Intuitively, in each of these intervals, one channel is used as side information to reduce the uncertainty about the interval of other source.



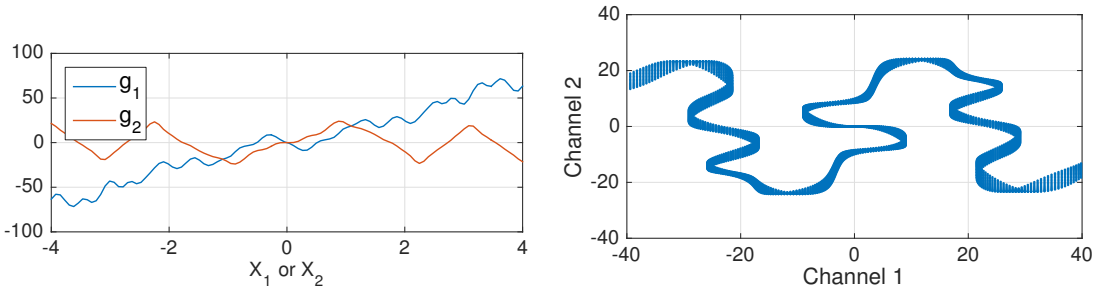
(a) (b)

(c) (d)

Figure 3.2: Example encoding scheme for distributed coding setting with $\rho = 0.999$. In (a), g_1 and g_2 are plotted. In (b) we see how the channel space is filled. In (c) the channel space filling for linear solution is shown. In (d) a typical Archimedean spiral used in literature is shown.

In order to gain further intuition into the workings of this non-linear strategy, we analyze how the channel space is filled, which is illustrated in Figure 3.2b with some example source pairs shown. For this plot, we consider only practically relevant source pairs, i.e., for close values of X_1 and X_2, as otherwise they are extremely unlikely to occur. This mapping has the same characteristics with that of Archimedean spirals used in literature (example plotted in Figure 3.2d), in the sense that more likely source values are mapped closer to the origin and the mapping continues outwards in a circular fashion, to fill the channel space while preserving transmission power. In fact, spirals are suggested since they have this characteristic. Although our mappings have the same characteristic, they are far different from a spiral. Channel space filling for linear encoders, which satisfies the necessary conditions of optimality and are therefore a locally optimum solution, is shown in Figure 3.2c where their inefficient power consumption can be seen as most low-power channel values are left unused. This helps understand why linear encoders are sub-optimal.

Spiral-like channel filling as shown in Figure 3.2a may sometimes be sub-optimal. The channel space can be filled in a different way, especially in case of unequal transmission powers. In Figure 3.3, we provide such mappings where we still see the same characteristics mentioned earlier, but the channel space is filled differently.



(a) (b)

Figure 3.3: An example, obtained by DA, with different transmission power constraints on encoders. (a) Both encoders are plotted together. (b) Channel space filling is shown. Although similar characteristics are observed, the channel space is filled differently.

In [17], the authors noted that for $0 < \rho < 0.95$, their structured solutions do not improve over linear solutions at high CSNR. We provide an example of non-linear scheme in Figure 3.4 for $\rho = 0.9$ that improves over linear solution. For lower correlations our method produces linear solutions. This can be explained by considering the channel space filling. As the correlation is lowered, the strips shown in Figure 3.2a and Figure 3.3b become wider. Thus, at lower correlations the strip is too wide to be twisted and bent into the channel space by a non-linear mapping, making linear as the only possibly option. Based on these experiments, we reach to a similar conclusion that optimal mappings are non-linear only at high correlation - albeit our method offers non-linear gains over a larger range of ρ values.

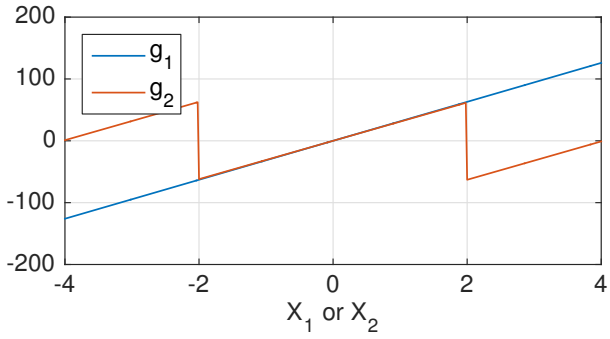
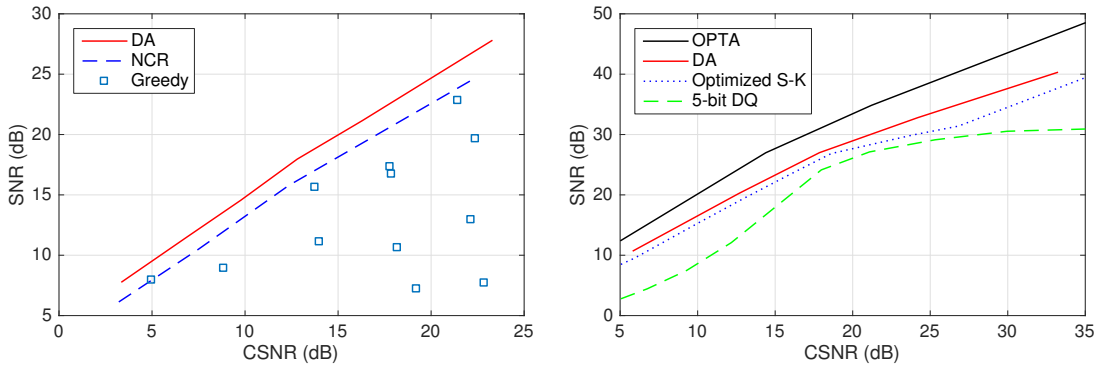


Figure 3.4: Non-linear solution that improves over linear for $\rho = 0.9$. CSNR = 29 dB, SNR = 29.82 dB. Linear solution at same CSNR achieves SNR = 29.60 dB.

Performance comparison of different numerical optimization techniques (DA, NCR and greedy descent with random initialization) for total power allocation case ($\lambda_1 = \lambda_2$) is provided in Figure 3.5a where we define SNR $= 10\log_{10}(2/D)$ (average distortion in dB) and CSNR $= 10\log_{10}((P(g_1) + P(g_2))/2)$ (average power in dB). We note that since individual powers are not constrained, different transmission powers are allowed in this comparison for all methods.

We also provide comparison to other coding schemes found in the literature. In [17], the authors analyze parametric mappings of two types, spirals and sawtooth mappings, in distributed coding setting and compare to distributed quantizer scheme analyzed in [69]. In Figure 3.5b, we provide comparison with our results to the ones reported in [17] for the same setting. In this comparison, the same power allocation is enforced for both encoders. As expected, mappings op-



(a) (b)

Figure 3.5: (a) Performance comparison of different numerical optimization methods for distributed coding setting with the constraint on total transmission power. $\rho = 0.99$. (b) Performance comparison for distributed coding setting with other approached found in literature. Optimized S-K refers to performance of structured mappings in [17] (spirals and sawtooth mappings) and 5-bit DQ is from [69]. 5-bit DQ is optimized for 18 dB CSNR. $\rho = 0.999$.

timized in function space perform better than parametric mappings which only approximately model optimal mappings as demonstrated in Figure 3.2.

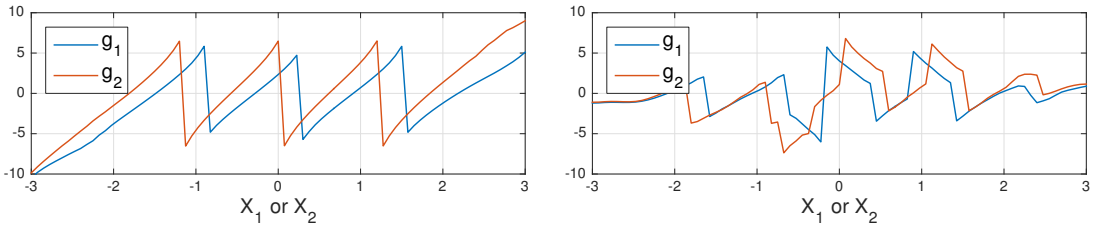
We finally take a look at the function computation problem for which the distortion is defined in (3.2). As a test case, we employed the difference function, $\gamma = X_1 - X_2$. Encoder mappings optimized with DA are given in Figure 3.6a. Both sources are mapped in many-to-one fashion with no way to resolve the uncertainty about the source interval. This is unlike previous mappings, where the uncertainty about source interval is resolved by side information, i.e., the other channel would locally act as side information. In the case of difference function, the actual values are not needed, thus, both sources are mapped in many-to-one

Table 3.1: Performance of Obtained Mappings for Difference Function

Method	CSNR_1 (dB)	CSNR_2 (dB)	SNR (dB)
DA	19.9	21.4	27.3
Linear-1	19.9	21.4	17.0
Linear-2	28.9	30.4	27.2
NCR	19.9	21.5	24.0

fashion. Nevertheless, the decoder is able to estimate the difference of sources accurately.

We give performance comparison in Table 3.1 where $\text{CSNR}_i = 10\log_{10}(P(g_i))$ for $i = 1, 2$ and $\text{SNR} = 10\log_{10}(1/D)$. DA achieves 10 dB higher SNR than the linear solution with the same power allocation, whereas the linear solution that achieves the same SNR requires 9 dB more power for each channel. Although the improvements depend on the problem parameters, these results nevertheless demonstrate that there are significant gains in utilizing non-linear encoder functions instead of linear ones. DA performance is better than NCR as well, as the shape of encoders are better optimized as can be seen in comparison in Figure 3.6.



(a) (b)

Figure 3.6: Example solutions obtained for function computation problem, where $\gamma = X_1 - X_2$. (a) DA result (b) NCR result. CSNR and SNR values are in Table 3.1.

Chapter 4

Analog Multiple Descriptions Coding

4.1 Introduction

The problem of multiple descriptions coding (MDC) - posed by Gersho, Witsenhausen, Wolf, Wyner, Ziv and Ozarow at the 1979 IEEE Information Theory Workshop - is a long standing open problem in source coding. The problem can be described as follows. Suppose we want to send a description of a stochastic process to a receiver through a communication network. There is a chance that the description will be lost. Therefore, we send two descriptions, and hope that one of them will reach the destination. Each description should be individually

good, since the description that is received is not known a-priori. If both are received, we then want to reconstruct the original process with minimum distortion using both descriptions. The difficulty of the problem lies in the fact that for individually good descriptions, we should make both descriptions close to the original process, hence the descriptions must be significantly correlated. However, in that case, when both descriptions are received, the second description contributes little to the reconstruction beyond what first description conveys. This tension yields a tradeoff between the quality of individual descriptions and the central reconstruction, which is the main subject of the MDC problem.

It is important to note that the MDC problem is not merely an isolated intellectual curiosity. Practical coding solutions, inspired by information-theoretic MD encoding schemes, have been extensively pursued for image, video and audio compression and transmission over packet networks, see [28] for an overview of MDC.

Note, however, that many digital MDC schemes incur long delay and complexity. In the presence of a channel with known statistics and a strict delay constraint, the MDC problem can be addressed by joint source-channel coding (JSCC) approaches based on zero-delay analog mappings. The main objective of this chapter is to design the zero-delay encoding/decoding mappings that minimize a given cost function under channel cost constraints. While the proposed

method is applicable to more general scenarios, we particularly focus on a setting that involves a zero-mean Gaussian source, two additive white Gaussian noise (AWGN) channels, the mean-squared error (MSE) distortion and the transmission power constraints as channel costs. We obtain the optimal JSCC mappings for analog MDC problem by numerical optimization. In general, such optimization problems pose significant challenges due to highly complex cost surfaces that render simple descent based methods useless. Here, we adopt the optimization paradigm we developed earlier for the analog MDC problem. To the best of our knowledge, this is the first attempt to obtain numerically optimized mappings for the general analog MDC problem.

The analog MDC problem has recently been considered in [14, 4], where 2-to-1 encoding functions that map two source symbols to a single channel symbol (primarily used in 2:1 bandwidth compression in zero-delay JSSC problems [31]) were modified to obtain good numerical performance. An important difference between our work and this prior work is that our approach is not limited to any parametric function, or a particular problem setting a priori, and is hence applicable to any scenario. Prior approaches are limited to very specific settings, such as 2:1 bandwidth compression, where the bandwidth of each channel is one half of the source bandwidth. This limitation was indeed recognized in a follow-up work in [15] where the approach was extended to some other integer valued

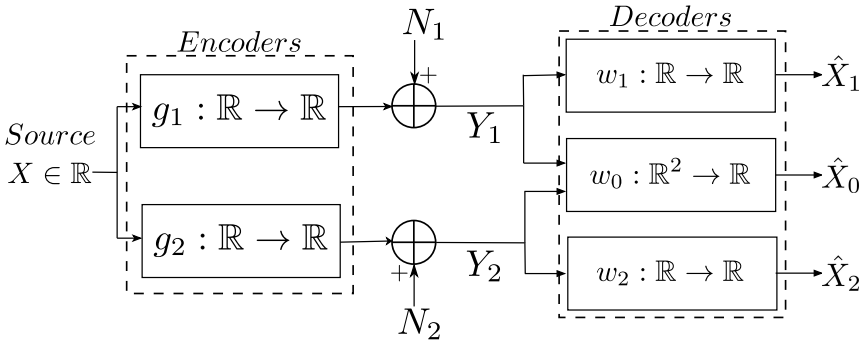


Figure 4.1: Analog multiple descriptions coding.

bandwidth settings (2:M, where M is positive integer), while noting that this extension is limited to specific bandwidth ratios. A related feature of the approach we propose here is that it is optimized for a given bandwidth and any channel SNR, which enables finding solutions for any bandwidth ratio including, of course, the 2:1 setting considered in prior work.

4.2 Problem Definition

The problem setting we consider is depicted in Figure 4.1. A Gaussian source $X \sim N(0,1)$ is mapped to channel symbols by mappings $g_i : \mathbb{R} \to \mathbb{R}$ for $i = 1, 2$. The channel noise variables are $N_i \sim N(0,1)$, $i = 1, 2$ where N_i is statistically independent of X. The receiver is modeled as three decoders that estimates the source from the channel outputs $Y_i = g_i(X) + N_i$, $i = 1, 2$. Each *side decoder* $w_i : \mathbb{R} \to \mathbb{R}$, $i = 1, 2$, estimates the source from corresponding Y_i, whereas the

central decoder $w_0 : \mathbb{R}^2 \to \mathbb{R}$ reconstructs the source using both channel outputs. We define the individual decoder MSE distortions $D_i \triangleq \mathbb{E}\{(X - \hat{X}_i)^2\}$, $i = 0, 1, 2$. We assume on-off channel model where a channel might fail with probability $\epsilon \ll 1$. The overall distortion to be minimized is then defined as:

$$D \triangleq (1 - \epsilon)D_0 + \epsilon(D_1 + D_2). \tag{4.1}$$

We optimize the encoder and decoder mappings to minimize (4.1) under a constraint on transmission powers defined as $P_i \triangleq \mathbb{E}\{g_i^2(X)\}$, $i = 1, 2$. For optimization purposes, we follow the procedure in prior chapters and define the following Lagrangian as objective cost function to be minimized:

$$J = D + \lambda(P_1 + P_2). \tag{4.2}$$

4.3 Information Theoretic Bounds

The information-theoretic MDC problem has been completely solved for the case of Gaussian source and MSE distortion (see the references in [28] for details). Adopting this solution to source-channel coding settings, the optimum performance theoretically achievable (OPTA) is given as follows: For D_1 and D_2 we have

$$\sigma_X^2 \geq D_i \geq \sigma_X^2(1 + P_i)^{-\beta_i}, i = 1, 2, \tag{4.3}$$

where β_i is the bandwidth ratio on channel i for $i = 1, 2$, i.e., the number of channel symbols used per source symbol. Given D_1 and D_2, the achievable central distortion is given as

$$D_0 = \begin{cases} \nu & \text{if } D_1 + D_2 > \sigma_X^2(1+\nu) \\ \nu\phi & \text{otherwise} \end{cases} \tag{4.4}$$

where

$$\nu = \sigma_X^2(1+P_1)^{-\beta_1}(1+P_2)^{-\beta_2}, \tag{4.5}$$

$$\phi = \frac{1}{1-\left(\sqrt{\left(1-\frac{D_1}{\sigma_X^2}\right)\left(1-\frac{D_2}{\sigma_X^2}\right)}-\sqrt{\frac{D_1D_2}{\sigma_X^4}-\nu}\right)^2}. \tag{4.6}$$

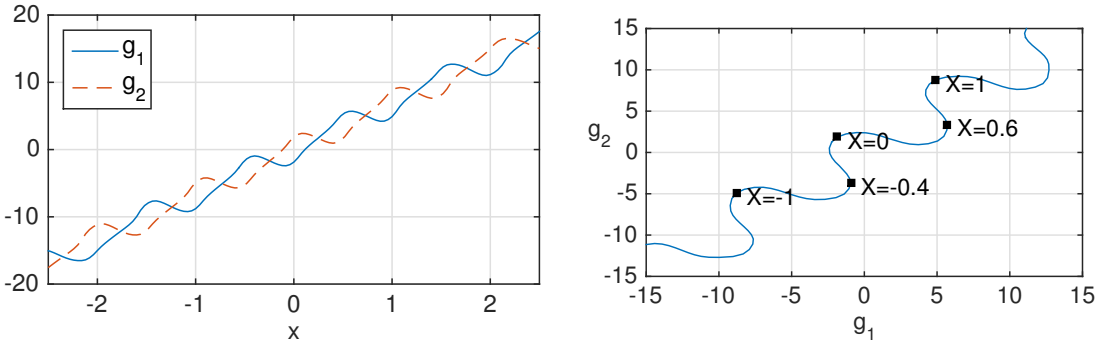
Remark 4.3.1. *We re-emphasize that the information theoretic bounds (OPTA) assume infinite delay encoding and decoding, while the problem here is formulated in limited-delay setting. Hence, in general OPTA constitutes a loose bound and may not be achievable by limited-delay schemes.*

4.4 Overview of Optimization Method

In Chapter 2 it was shown that even the simple network problem of decoder with side information has a non-convex cost surface riddled with local minima, making greedy descent-based techniques suboptimal and highly sensitive to initialization. Accordingly, a non-convex optimization method was proposed, based

on the ideas of DA, to mitigate poor local minima. It is reasonable to expect the challenge due to local minima to be more severe in the more complicated setting that we consider here. We therefore adapt our optimization method to this problem to obtain the desired mappings. Consider the setting in Chapter 3 which is distributed coding of two correlated sources and communicating them over two independent channels to a central decoder. The method we introduced there can be adapted to the MDC problem, by considering a single source instead of two correlated sources and introducing side decoders to be optimized. The adaptation is a straightforward modification and therefore the details are not included here.

The important advantage of DA-based optimization for analog MDC problem is that it is applicable to the general problem setting and makes no assumptions on the distributions or objective cost functions. It is also adaptable to various network topologies, bandwidth ratios and channel models, as demonstrated in this chapter. This is in contrast to prior approach [14] which only applies to a very specific setting. Furthermore, our approach still improves over the results in [14] as shown in this chapter.



(a) (b)

Figure 4.2: An example of proposed JSCC mappings that achieve zero-delay MD coding. (a) g_1 and g_2. In (b), g_1 vs. g_2 is plotted to show how channel space is filled. Each point on the curve corresponds to a source value, example points are given.

4.5 Experimental Results

4.5.1 Zero Delay Analog MD Mappings

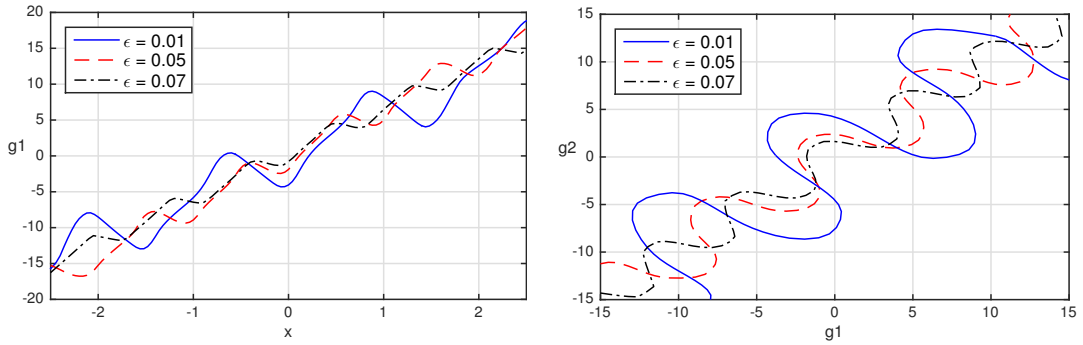
In [14], since communication channels are assumed 2:1, known bandwidth reduction mappings are used. However, such heuristics are not available in general. Consider, for example, 1:1 setting: the communication channels are point-to-point, and for the Gaussian case we consider here it is known that the optimal solution for a single channel is linear encoding [26]. However, as we show in this section, linear solutions cannot exploit the diversity of parallel channels well, and nonlinear mappings that significantly improve over linear ones exist.

The mappings we propose are shown in Figure 4.2. We first notice that mappings g_1 and g_2 are not only nonlinear, they are in fact many-to-one, in the sense

that multiple source points are mapped to the same channel value. This introduces uncertainty about the source interval at the side decoders. Many-to-one mappings have been found for analog network problems as shown in prior chapters. In those examples, the decoder is able to reduce the uncertainty about the source interval by using additional information. However, in our case the side decoders are unable to do so, and this introduces some distortion. Although counterintuitive, and perhaps a poor solution for point-to-point setting, these mappings achieve better performance compared to the linear solution in the MDC setting, and are currently the best known mappings.

In Figure 4.2b, we map g_1 vs. g_2 to show how channel space is filled for communication with the central decoder, which can be considered as bandwidth expansion case. The channel space is filled in a way that seeks a compromise between bandwidth expansion mappings and linear mappings, since the overall distortion in (4.1) is a compromise between distortion at side decoders (where the best mappings would be linear) and central decoder.

Figure 4.3 demonstrates the behavior of these mappings when channel failure probability ϵ is varied. We plot encoders (only g_1) and channel curves for three different ϵ values under the same transmission power constraints. As ϵ decreases, the distortion at the side decoders become less dominant. Consequently, the uncertainty about source interval increases at each channel since encoders map



(a) (b)

Figure 4.3: We present how mappings change as the channel failure probability ϵ is changed. (a) The change in g_1 (b) The change in channel space.

bigger source intervals to the same channel values as seen in Figure 4.3a, resulting in higher distortion at side decoders. On the other hand, the overall system *approaches* an analog bandwidth expansion system, resulting in better filling of channel space as seen in Figure 4.3b.

We present the performance of the proposed mappings in Figure 4.4, where $\text{SNR} = 10\log_{10}(1/D)$, D is defined in (4.1). We use equal transmission power on both channels, and define $\text{CSNR} = 10\log_{10}(P_1) = 10\log_{10}(P_2)$. The performance is compared to OPTA as well as the heuristic choice of the linear encoding scheme. Several observations are made here: First, our mappings are able to follow OPTA with a relatively constant gap, by trading D_0 and D_1, D_2 as ϵ changes, whereas the linear scheme does not offer the same flexibility since it essentially minimizes D_1 and D_2 irrespective of D_0. Secondly, as $\epsilon \to 0$, the objective becomes that

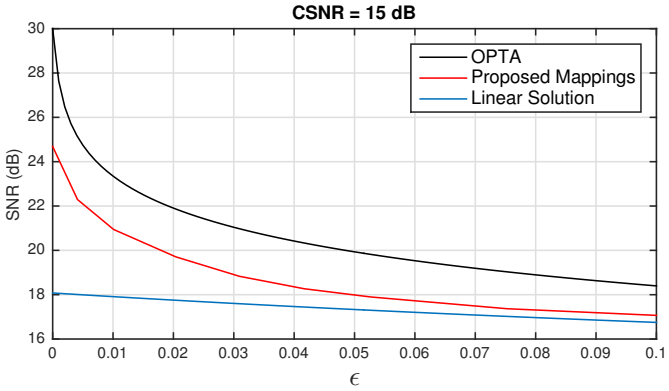
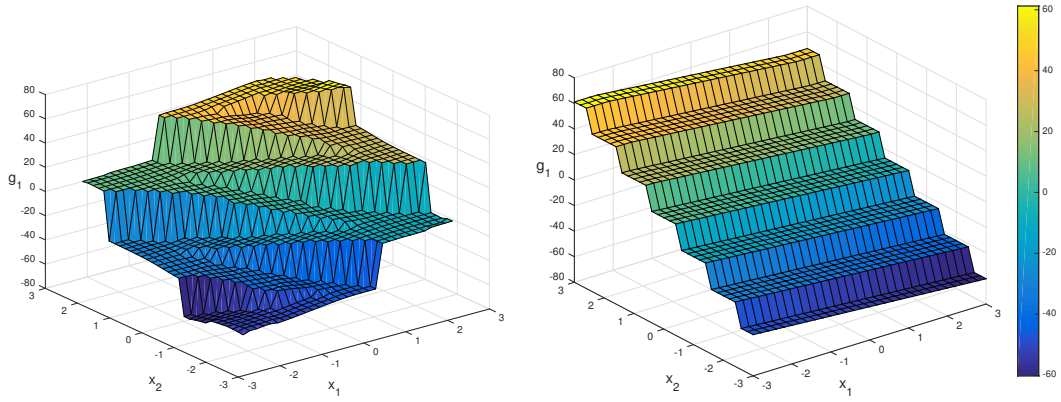


Figure 4.4: The performance of proposed mappings, SNR vs. channel failure probability ϵ, where SNR$= 10\log_{10}(1/D)$ and CSNR$= 10\log_{10}(P_1) = 10\log_{10}(P_2) = 15dB$.



(a) (b)

Figure 4.5: Mappings for 2:1 case. (a) Proposed mappings. (b) Mappings used in prior approach.

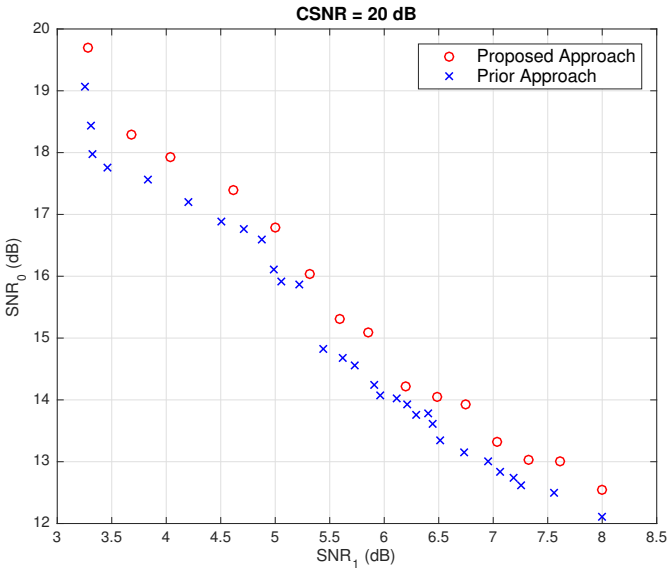


Figure 4.6: Performance of the proposed 2:1 mappings. Prior approach is reported in [14].

of minimizing D_0 directly, which would make the system equivalent to 1:2 bandwidth expansion communication. Our mappings approach the performance of the best-known bandwidth expansion mappings as reported in [31]. The analysis of connections between the mappings obtained in this chapter and bandwidth expansion problem is left for future research.

4.5.2 2:1 Analog MD Mappings

Although this paper is mainly focused on zero-delay 1:1 mappings as explained in the previous section, here we provide a preliminary set of results on optimal mappings for the 2:1 setting that was considered in prior work [14], for comparison

purposes. Our mappings are as given in Figure 4.5a, where the encoder $g_1 : \mathbb{R}^2 \to \mathbb{R}$ is shown (g_2 is similar but rotated by $\pi/2$).

In [14], two types of mappings are considered, and the somewhat more efficient one is shown in Figure 4.5b. The "sloped-steps" in this mapping extend through the X_1 direction, which results in suboptimal encoding of X_2, in the sense that its encoding is effectively a 5-level quantization. In our mappings, the sloped-steps extend diagonally in the $X_1 = -X_2$ direction, resulting in both sources being encoded efficiently. Moreover, the steps merge towards the ends rather than being separate.

The proposed mappings achieve better performance as can be seen in Figure 4.6, where we have $\mathrm{SNR}_1 = 10\log_{10}(1/D_1)$, $\mathrm{SNR}_0 = 10\log_{10}(1/D_0)$ and $\mathrm{CSNR} = 10\log_{10}(P_1) = 10\log_{10}(P_2)$. For given SNR_1, our gains in SNR_0 vary from 0.5 dB to 1 dB. Note that, since the prior approach employs mappings with a fixed structure, its performance is arbitrary and is relatively better at some points. Hence it has varying performance compared to our approach which optimize mappings without making assumptions on the structure. Further analysis on the structure of these mappings is currently under investigation.

Chapter 5

Decentralized Control

5.1 Introduction

This chapter studies the problem of global optimization of controller mappings in decentralized stochastic control problems including Witsenhausen's celebrated 1968 counter-example (WCE). It is well known that most decentralized problems do not admit closed-form solutions and require numerical optimization. Decentralized control systems such as WCE arise in many practical applications, and numerous variations on WCE have been studied in the literature (see, e.g., [5, 7, 10, 24, 25, 29, 34, 44, 59, 60]). In general, linear control strategies are not optimal for decentralized control LQG systems, except when the system admits some specific information structures (see, e.g., [33, 53, 80]). It is well-understood

that if the information structure in a decentralized control problem is nonclassical, as in the case of WCE, non-linear strategies may widely outperform optimal linear strategies. Finding the optimal mappings for such problems is usually a difficult task unless they admit an explicit (and often as simple as linear) solution, see [5] for a set of problems that include both tractable and intractable examples.

Recent research efforts have focused on developing efficient numerical methods for decentralized control problems [8, 13, 32, 40], specifically for WCE [6, 36, 43, 45]. Some of the existing methods rely on simplifying properties of WCE, such as monotonicity, and are therefore not easily generalizable [13, 32, 42]. Moreover, methods that require analytical derivation for each particular setting are not fully automated [40, 42].

In this chapter, we build on our methods introduced in previous chapters and develop a general optimization method for decentralized stochastic control problems. Our method does not rely on the simplifying properties of a particular setting or analytical derivations. We note that deterministic annealing (DA) has been successfully used in various other problems in control theory including dynamic coverage control problems [65, 79, 41] and cluster analysis in control systems [48], however, the method introduced here is general and applicable for a class of decentralized control problems.

We demonstrate the DA-based method on two specific problems. We first analyze the numerically "over-mined" WCE problem. We then study a more involved variation on WCE, introduced in [44], which includes an additional noisy channel over which the two controllers can communicate. The second controller, therefore, has access to some side information which is controlled by the first controller. We refer to this setting as the "side channel problem" motivated by the class of "decoder side information" problems as discussed in Chapter 2. It has been demonstrated in [44] that non-linear strategies may outperform the best linear strategies, however, the question of how to approach the optimal solution remains open.

Having a powerful optimization method at hand, we analyze the structure of experimentally obtained mappings. For instance, Wu and Verdú have shown [75], using tools from optimal transport and functional properties of estimation over a Gaussian channel, that the solution of WCE must have real analytic left inverse. An important practical consequence of this result is that a piecewise linear function cannot be optimal. Our numerical results demonstrate that the "steps" in obtained mappings show small deviations from linear, experimentally confirming this theoretical finding.

5.2 Problem Definition

5.2.1 Notation

Let $\mathbb{R}$, $\mathbb{E}(\cdot)$ and $\mathbb{P}(\cdot)$ denote the set of real numbers, the expectation and probability operators, respectively. We represent random variables and their realizations with uppercase and lowercase letters (e.g., X and x), respectively. Let X_i^j denote $X_i, \ldots, X_j$. The probability density function of the random variable X is $f_X(x)$. The Gaussian density with mean μ and standard deviation σ is denoted as $\mathcal{N}(\mu, \sigma^2)$. We use natural logarithms which, in general, may be complex, and the integrals are, in general, Lebesgue integrals.

5.2.2 General Problem Definition

Formally, we consider a discrete-time stochastic control problem with non-classical information pattern involving n controllers, and assume that the order of control actions is fixed in advance, i.e., the system is sequential [74]. Following the problem definition in [74], let $(\Omega, \mathcal{B}, \mathcal{P})$ be a probability space, where Ω denotes the random quantities involved in the system such as initial input, and (U_i, Σ_i), for $i = 1, \ldots, n$, are measurable spaces with U_i denoting the set of control actions. Controller mappings (functions) are denoted by $g_i : \Omega \times U_1 \times U_2 \ldots U_{i-1} \to U_i$, for $i = 1, \ldots, n$. For convenience, we denote the input

set of g_i by $X_i = \Omega \times U_1 \times U_2 \ldots U_{i-1}$. The system is then defined by the following set of equations

$$u_i = g_i(x_i),\ i = 1, \ldots, n. \tag{5.1}$$

The sequential property ensures that the controllers are indexed in such a way that the action of controller g_k does not depend on the actions of g_{k+1}^n, for all k.

Let f be a real-valued and bounded measurable function of ω, u_1^n on $(\Omega, \mathcal{B})$, i.e., f is a random variable. The problem objective is to find the set of functions g_1^n that minimize the value of the cost function J:

$$J = \mathbb{E}\{f(\omega, u_1^n)\}. \tag{5.2}$$

5.3 Proposed Method

Let us denote the space of controller input x_i by $\mathbb{R}_{x_i}$, for $i = 1, \ldots, n$. Assume there exists a partition of $\mathbb{R}_{x_i}$ into $\mathcal{M}_i > 0$ disjoint regions denoted by $\mathbb{R}_{i,m_i}$ $(m_i = 1, \ldots, \mathcal{M}_i)$:

$$\bigcup_{m_i=1}^{\mathcal{M}_i} \mathbb{R}_{i,m_i} = \mathbb{R}_{x_i}. \tag{5.3}$$

Note that each value of x_i belongs to exactly one of the partition regions, referred to as a deterministic (non-random) partition.

We begin our formulation by imposing a piecewise structure on the controller mappings. Consider the structured mapping g_i, for $i = 1, \ldots, n$, written as

$$g_i(x_i) = g_{i,m_i}(x_i) \quad \text{for } x_i \in \mathbb{R}_{i,m_i}. \tag{5.4}$$

Each $g_{i,m_i}(x_i)$ is a parametric function referred to as "local model". Effectively, each of the mappings g_i is defined with a structure determined by two components: a space partition where regions are denoted by $\mathbb{R}_{i,m_i}$ and a parametric local model per partition cell, i.e., $g_{i,m_i}(x_i)$ for $\mathbb{R}_{i,m_i}$. The number of local models (partition regions) for mapping g_i is $\mathcal{M}_i$. The local models can take any prescribed form such as linear, quadratic or Gaussian and we let $\Lambda(g_{i,m_i})$ denote the parameter set for local model g_{i,m_i}.

We follow our approach in prior chapters and introduce controlled randomization into the problem formulation. We replace the deterministic partition of space by a random partition, i.e., we associate every input point (x_i) with partition regions *in probability*. To this end, we introduce random variables M_i, for $i = 1, \ldots, n$, whose realization is the partition index m_i. We define the *association probabilities* as conditional distribution on the partition index given the input:

$$p_i(m_i|x_i) = \mathbb{P}\{x_i \in \mathbb{R}_{i,m_i}\} = \mathbb{P}\{g_i(x_i) = g_{i,m_i}(x_i)\}, \tag{5.5}$$

for $i = 1, \ldots, n$. Our viewpoint is that we consider $\mathbb{R}_{i,m_i}$ as regular regions, with the exact membership of an input point to a region being the outcome of a random experiment.

Consequently, the mappings are now random, in the sense that the output of controller g_i for an input x_i is given in probability as

$$g_i(x_i) = g_{i,m_i}(x_i) \quad \text{with probability } \; p_i(m_i|x_i). \tag{5.6}$$

By construction, and due to the sequential property, we have that given X_i, M_i is independent of the random variables M_1^{i-1}.

The expectation in (5.2) is now taken over X_i^n and M_i^n. Let us rewrite it for a fixed value of i as follows:

$$J = \int_{x_i} \sum_{m_i=1}^{\mathcal{M}_i} \mathbb{E}\{f(\omega, u_1^n)|m_i, x_i\} p_i(m_i|x_i) f_{x_i}(x_i) \mathrm{d}x_i \tag{5.7}$$

where $\mathbb{E}\{f(\omega, u_1^n)|m_i, x_i\}$ can be viewed as the cost of associating x_i with local model g_{i,m_i}. As in Lemma 2.3.1, assuming fixed local model parameters, optimizing (5.7) with respect to $p_i(m_i|x_i)$ would clearly produce deterministic mappings, since the minimum is achieved by setting $p_i(m_i|x_i) = 1$ for the pair $\{m_i, x_i\}$ for which $\mathbb{E}\{f(\omega, u_1^n)|m_i, x_i\}$ is minimum. Therefore, the ultimate objective of obtaining optimum deterministic controllers is preserved as the random encoders share the same global minimum as deterministic ones. However, direct optimization of the cost with respect to $p_i(m_i|x_i)$ results in poor local minima. Instead,

we minimize (5.2) at prescribed levels of *randomness*, which we measure by the Shannon entropy. The joint entropy of the system can be written

$$H(X_1^n, M_1^n) = H(X_1) + \sum_{i=2}^{n} H(X_i|M_1^{i-1}, X_1^{i-1}) + H(M_1|X_1) + \sum_{i=2}^{n} H(M_i|X_i, M_1^{i-1}, X_1^{i-1}). \tag{5.8}$$

It is easy to see from conditional independence arguments that the conditional entropies in the second term in the righthand side of (5.8) can be simplified to $H(X_i|X_1^{i-1})$, and those in the last term to $H(M_i|X_i)$. Thus the first two terms of (5.8) are fixed and determined by the problem statement (the joint distribution of X_1^n). We therefore discard the first two fixed terms of (5.8), rearrange the remaining terms, to obtain a conveniently compact measure of randomness defined as

$$H \triangleq \sum_{i=1}^{n} H(M_i|X_i). \tag{5.9}$$

The conditional entropy $H(M_i|X_i)$ is given by

$$H(M_i|X_i) = -\int_{x_i} \sum_{m_i=1}^{\mathcal{M}_i} p(m_i|x_i) \log p(m_i|x_i) f_{X_i}(x_i) \mathrm{d}x_i. \tag{5.10}$$

Accordingly, we construct the Lagrangian

$$F = J - TH \tag{5.11}$$

as the objective function to be minimized, where J is given in (5.2), H is given in (5.9) and T is the Lagrange multiplier associated with the entropy constraint.

The practical method is similar to Algorithm 1 in Chapter 2, however we perform the operations in each step for the controllers sequentially, i.e., from g_1 to g_n.

Remark 5.3.1. *Our method is derived without recourse to discretization. Although practical simulations involve sampling of the continuous space during numerical computations of integrals, this is in contrast to methods that are entirely formulated in discrete settings.*

Remark 5.3.2. *Critical temperatures can be derived analytically if, for the problem considered, phase transitions are of "continuous" nature, in the sense that tracked minimum becomes a saddle point at the exact critical temperature. The condition for saddle point can be obtained using variational calculus, see [56] for phase transition analysis in DA. Our experiments indicate that, at least for the test cases considered in this paper, phase transitions are not continuous. While pre-calculating the critical temperature may enable a numerical speed up of the annealing process, it is not necessary to implementing the practical algorithm. Hence, the derivation and characteristics of phase transitions are kept outside the scope of this work.*

Theorem 5.3.3. *At any temperature T, the optimal $p_i(m_i|x_i)$ that minimize (5.11) is given by*

$$p_i(m_i|x_i) = \frac{e^{-\mathbb{E}\{f(\omega,u_1^n)|m_i,x_i\}/T}}{\sum_{m_i} e^{-\mathbb{E}\{f(\omega,u_1^n)|m_i,x_i\}/T}} \tag{5.12}$$

Proof. Proof follows similar steps to those for Theorem 2.3.2. □

The optimal $p_i(m_i|x_i)$ that minimize (5.11) can be derived in closed form. Plugging (5.7) and (5.10) in (5.11), we find the optimal $p_i(m_i|x_i)$ as

$$p_i(m_i|x_i) = \frac{e^{-\mathbb{E}\{f(\omega,u_1^n)|m_i,x_i\}/T}}{\sum_{m_i} e^{-\mathbb{E}\{f(\omega,u_1^n)|m_i,x_i\}/T}} \tag{5.13}$$

Optimization of parameters in $\Lambda(g_{m_i})$ can be done using any standard method. Typically, a variant of gradient descent is used when closed form expressions cannot be obtained.

5.4 Applications of the Proposed Method

5.4.1 Witsenhausen's Counter-example

Problem Description

Let X_0 and W be Gaussian random variables with distributions $\mathcal{N}(0, \sigma_{X_0}^2)$ and $\mathcal{N}(0, 1)$, respectively. WCE is a 2-stage control problem with controllers

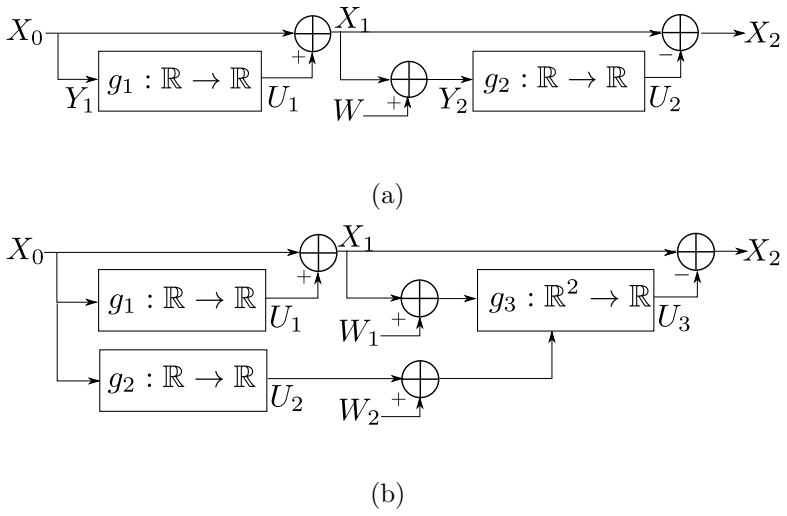


Figure 5.1: Settings used for testing the proposed method (a) Witsenhausen's counter-example (b) Side channel problem.

$g_1 : \mathbb{R} \to \mathbb{R}$ and $g_2 : \mathbb{R} \to \mathbb{R}$, defined by the following equations:

$$U_1 = g_1(X_0), \quad U_2 = g_2(X_1 + W),$$
$$X_1 = X_0 + U_1, \; X_2 = X_1 - U_2. \tag{5.14}$$

The schematic representation is given in Figure 5.1a. The objective is to minimize the cost

$$J = \mathbb{E}\{k^2 U_1^2 + X_2^2\}. \tag{5.15}$$

For convenience, we define $f_1(X_0) = g_1(X_0) + X_0$.

Some properties of $f_1(\cdot)$ are known, including the property of symmetry about the origin (thus, positive half is enough to describe a given solution) [73]. Witsenhausen has provided the following solution that outperforms the optimal linear

solution for a given set of problem parameters ($k = 0.2$, $\sigma_{X_0} = 5$):

$$f_1(x_0) = \sigma_{X_0} \operatorname{sgn}(x_0) \tag{5.16}$$

where $\operatorname{sgn}(\cdot)$ is the signum function. Since there is a single "step" in the positive half of real line, this solution is referred to as a "1-step" solution. Improved solutions that appeared in literature utilize 2.5, 3, 3.5 and 4-step functions (an $x.5$ step function has a step that straddles the origin). Moreover, the latter solutions made improvements by using slightly sloped steps rather than constant ones.

Although in standard application of DA-based method we randomize all controllers, for computational efficiency, we restrict the randomization to only g_1 as

$$g_1(x_0) = g_{1,m_1}(x_0) \quad \text{with probability } \; p_1(m_1|x_0) \tag{5.17}$$

and numerically compute (update) g_2 by using the fact that optimal g_2 given g_1 is

$$g_2(Y_2) = \mathbb{E}\{X_1|Y_2\} \tag{5.18}$$

where $Y_2 = X_1 + W$.

For this particular problem, we use linear local models given by

$$g_{1,m_1}(x_0) = a_{1,m_1} x_0 + b_{1,m_1}. \tag{5.19}$$

while noting that optimal g_1 must have analytic left inverse and hence cannot be piecewise linear [75]. Nevertheless, the minimal cost can be approached arbitrarily

closely by piecewise linear functions [75]. Thus, for numerical algorithms, linear models are sufficient.

Results for WCE

Preliminary results on the application to WCE appeared in [46]. We first provide results for the standard benchmark case where $k = 0.2$, $\sigma_{X_0} = 5$ that was used in many papers in literature. The annealing process is illustrated in Figure 5.2, where evolution of the mapping can be seen (only the positive half is shown thanks to the symmetry property). The obtained mapping is referred to as a "sloped 5-step" solution. At high temperature, there is only one local model, thus, the function is 1-step. As the temperature is lowered, the solution undergoes phase transitions, revealing more steps for the mapping function. In this work we calculated the solution with 5 steps. Although more steps possibly exist, improvement to cost is numerically insignificant with additional steps. Some earlier results from the literature are given in Table 5.1, where it can be seen that our method produced the minimum cost achieved to date.

Another benchmark case, $k = 0.63$, was suggested in [29] as potentially being more relevant for confirming the high gains of optimal non-linear mappings. Our resulting mapping for $k = 0.63$, $\sigma_{X_0} = 5$ is given in Figure 5.3a, which is a 6-step solution. Our numerical results suggest that the gain over linear solution is smaller

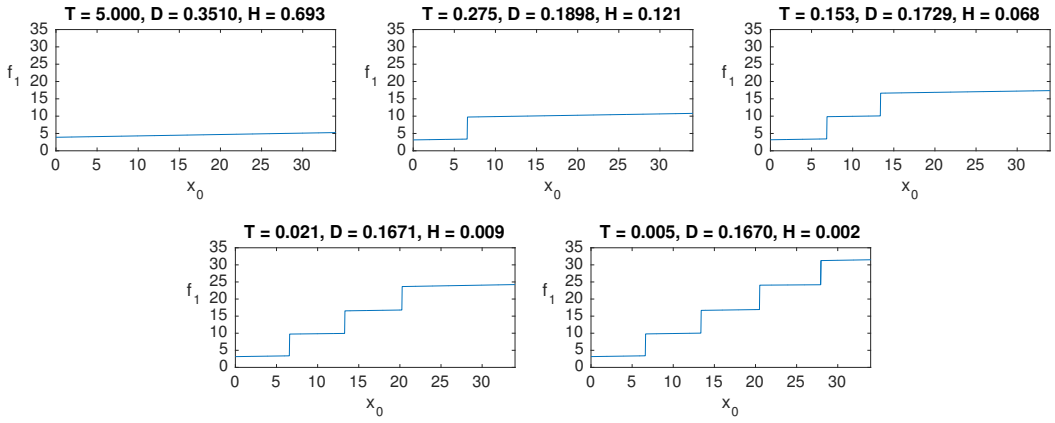
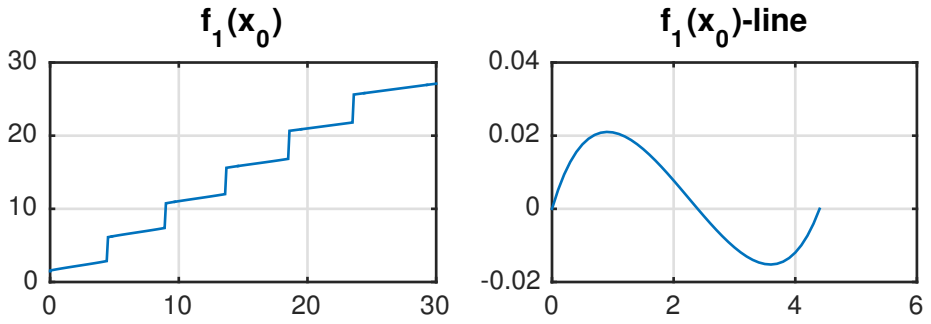


Figure 5.2: Evolving graph of $f_1(x_0)$ in WCE during various phases of the annealing process. We note that mapping is actually random during algorithm run. Here, for demonstration, we fully associate every x_0 with g_{1,m_1} for which $p_1(m_1|x_0)$ is largest.

Table 5.1: Results for WCE

Solution	Cost
Optimal linear Solution	0.96
1-step, Witsenhausen [73]	0.404253
2-step,[13]	0.190
Sloped 2.5 - step, [6]	0.1701
Sloped 3.5 - step, [42]	0.1673132
Sloped 3.5 - step, [43]	0.1670790
Sloped 4 - step, [36]	0.16692462
Sloped 5 - step, our result	0.16692291



(a) (b)

Figure 5.3: Numerical result for WCE in the case of $k = 0.63$, $\sigma_{X_0} = 5$. (a) 6-step solution. (b) The deviation of the first step in $f_1(x_0)$ from a straight line between the end points of the step.

compared to the standard benchmark case above: $J = 0.844$ for the solution in Figure 5.3a whereas cost associated with the optimal linear mapping is $J = 0.961$.

These numerical results illustrate an important theoretical result as well. In [75] authors proved that the optimal f_1 must have analytic left inverse and therefore cannot be piecewise linear, which was believed to be the case due to numerical results (see, e.g., [42]). Our numerical results indicate that steps are in fact non-linear, as shown in Figure 5.3b. The steps become non-linear during the final stages of the algorithm as multiple local models appear to form a single step. To the best of our knowledge, this is the first numerical result illustrating non-linearity of the steps.

5.4.2 Side Channel Problem

Problem Description

Let X_0 be a Gaussian random variable with distribution $\mathcal{N}(0, \sigma^2_{X_0})$, and W_1, W_2 be independent Gaussian random variables, both with a distribution $\mathcal{N}(0, 1)$. The system is defined by the following equations:

$$\begin{aligned} &U_1 = g_1(X_0),\ U_2 = g_2(X_0),\ U_3 = g_3(X_1 + W_1, U_2 + W_2), \\ &X_1 = X_0 + U_1, \quad X_2 = X_1 - U_3. \end{aligned} \tag{5.20}$$

The problem is to optimize the cost function

$$J = \mathbb{E}\{k^2 U_1^2 + X_2^2\} \tag{5.21}$$

for given σ_{X_0} and positive parameter k, subject to a power constraint on U_2:

$$b_{SNR} = \mathbb{E}\{U_2^2\} \tag{5.22}$$

where b_{SNR} is the specified power level. We again define $f_1(X_0) = g_1(X_0) + X_0$.

This problem setting is illustrated in Figure 5.1b and was introduced in [44]. It can be seen as a generalization of WCE with an additional communication channel between the controllers, i.e., a non-linear function of input X_0 is communicated by g_2 to the controller denoted by $g_3 : \mathbb{R}^2 \to \mathbb{R}$. The non-linear mappings analyzed in [44], which widely outperform the best linear solution in a large range of b_{SNR}, are such that both f_1 and g_2 are staircase functions of x_0.

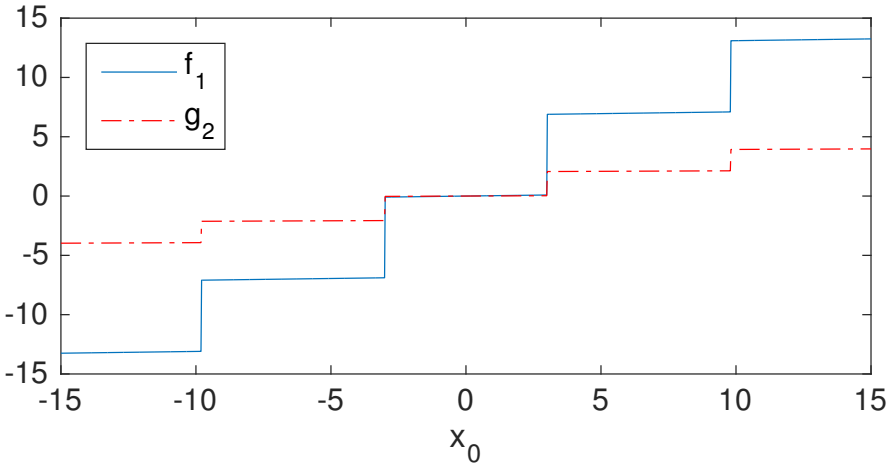


Figure 5.4: Mappings suggested in [44] for the side channel problem, where both mappings suggested are staircase functions.

The non-linear mappings analyzed in [44], which widely outperform the best linear solution in a large range of b_{SNR}, are depicted in Figure 5.4. These are similar to the mapping g_1 in WCE, both g_1 and g_2 are staircase functions of x_0.

Results for Side Channel Problem

The original problem is to minimize (5.21) subject to the constraint in (5.22). We follow the standard approach in optimization theory and convert this constrained problem to unconstrained Lagrangian formulation:

$$J = \mathbb{E}\{k^2 U_1^2 + X_2^2 + \lambda U_2^2\} \tag{5.23}$$

where λ is chosen to satisfy the power constraint (5.22) with equality. In the experiments, we used the standard benchmark parameters that were used for the

Table 5.2: Cost Comparison Table for Side Channel Problem

b_{SNR}	Linear Cost	J^M ([44])	J^*	$(J^M - J^*)/J^M$
0	0.960	0.185	0.167	0.10
2.6	0.696	0.149	0.079	0.47
4.7	0.432	0.101	0.040	0.60
5.9	0.344	0.081	0.026	0.68
9.0	0.203	0.052	0.012	0.77

original WCE, that is, $k = 0.2$ and $\sigma_{X_0} = 5$. We have varied λ to obtain results at different b_{SNR}.

In Table 5.2 we compare the cost of our solutions (denoted by J^*) to the ones given in [44] (denoted by J^M), and the best linear mappings. Significant cost reductions can be observed. The relative improvement over the solution of [44] is listed in the last column.

Remark 5.4.1. *When $b_{SNR} = 0$, the problem degenerates to WCE, thus the cost is 0.1669, the best known to date.*

We present several mappings obtained by our method in Figure 5.5. Some interesting features of these mappings are observed. The mappings f_1 are staircase functions with constant steps similar to the ones obtained for the original WCE problem, however, the steps get smaller and increase in number as the side channel SNR increases; that is, $f_1(x_0)$ approaches x_0. Note that the control cost term in (5.21), $\mathbb{E}\{k^2 U_1^2\}$, achieves its minimum when $g_1 = 0$, i.e., $f_1(x_0) = x_0$. This is, however, not optimal due to the estimation error at the second stage. Intuitively,

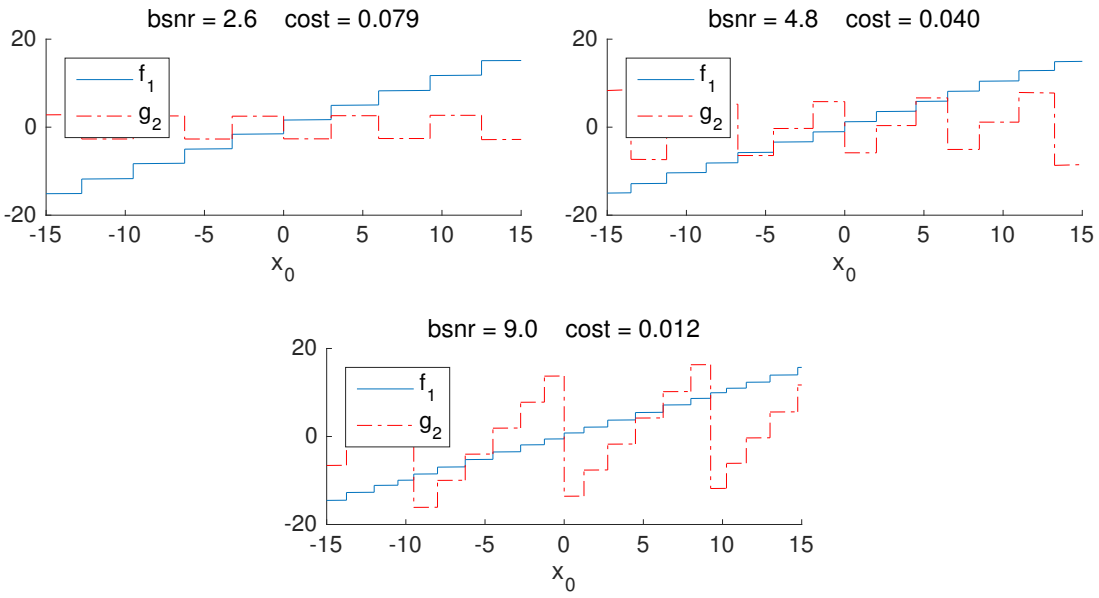


Figure 5.5: Example mappings we obtained for the side channel variation problem. The first controller is plotted at various SNR levels.

as the second controller has access to better side information (i.e. at higher SNR), the estimation error is decreased and as observed in Figure 5.5, $f_1(x_0)$ approaches x_0. The relative improvement in cost, given in Table 5.2, increases with SNR, which is consistent with the above observation.

The mappings for the side channel, g_2, are irregular and the overall shape varies with SNR. This observation, together with the above for f_1, suggests that the mappings f_1 and g_2 are not scale invariant. The discontinuities in f_1 and g_2 coincide as expected, as the discontinuities in side information signal those in f_1 to g_3.

5.5 Advantages of Proposed Method

There are several improvements of the method proposed here over existing methods in literature.

1. It is derived in the original, continuous domain, without discretization. The continuous space is sampled during numerical computation of integrals only. This is in contrast with many prior methods such as those in [32, 36, 43] that are entirely formulated in a discrete setting.

2. Our method is based on DA, a powerful non-convex optimization framework. DA has been successfully used as a remedy to the problem of poor local minima in non-convex optimization problems [56], and is shown to outperform competing methods such as "noisy channel relaxation" (NCR) (see, e.g., earlier chapters). Although NCR performs well for the simple setting of WCE [36], it is susceptible to get trapped in local minima in more involved settings. We again note that DA has strong potential to avoid poor local minima but does not guarantee convergence to globally optimum solution.

3. From its DA foundation, our method directly inherits notably useful properties including reduced sensitivity to initialization. The authors in [6] had to experiment with a large number of initial weight vectors to obtain the result included in Table 5.1.

4. We do not make any assumptions about the controller mappings. Methods presented in [13, 32, 42] benefit from the monotonicity of optimal mapping in WCE. The results in Figure 5.5 demonstrate that monotonicity may not hold for optimal mappings in the general setting.

5. The method is fully automated and does not require analytical derivations or manual interventions during algorithm run. This is in contrast to the method presented in [42] which requires analytical work during the procedure.

6. The method is applicable to a broad class of stochastic control problems. Many prior methods require non-trivial work in order to generalize, for instance, [43] proposes a method that is generalizable, but it requires conversion to a potential game problem.

Chapter 6

Conclusions

This thesis is mainly concerned with optimal encoding and decoding rules in emerging communication networks that are characterized by very low delay requirements and constrained resources. Our general approach is to numerically optimize encoder and decoder mappings using deterministic annealing (DA) based optimization. DA has several useful properties, the most critical being a strong potential to avoid poor local minima. This aspect of DA makes it suitable for these problems where cost surfaces are riddled with local minima, rendering gradient based methods insufficient.

6.1 Main Contributions

In Chapter 2, we studied the problem of finding globally optimal encoder and decoder pairs in zero delay source-channel coding, focusing on decoder side information setting. We developed a method based on DA to approach global optimality. The numerical results show that, by using carefully optimized non-linear (and in many cases many-to-one) mappings, significant gains can be obtained over linear solutions, which are optimal in point-to-point settings (for the specific case of Gaussians under MSE). Simulation results demonstrate the performance of the proposed algorithm, which consistently outperform greedy optimization methods and noisy channel relaxation.

In Chapter 3, we adapted our optimization method introduced in Chapter 2 for optimizing distributed zero-delay codes. Our results are superior to the more ad hoc method of noisy channel relaxation, as well as prior approaches in the literature. The obtained mappings exhibit properties that are reminiscent of digital Wyner-Ziv mappings.

Chapter 4 is concerned with the zero-delay multiple descriptions coding (MDC) problem. Using an adaptation of our non-convex optimization method, we propose schemes based on joint source-channel coding that provide good performance for different configurations of side and central distortions. It is demonstrated that our approach outperforms its known competitors. Obtained mappings exhibit

counter-intuitive features such as many-to-one mappings for zero-delay MDC that outperform the more natural choice of linear mappings.

Finally, in Chapter 5, we proposed a general optimization method for distributed control problems, whose solutions are known to be non-linear, and demonstrated its effectiveness on two problems from the literature. The first problem is the celebrated benchmark problem known as Witsenhausen's counter-example, for which our approach obtained the best known cost value. As a second test case we focused on the side channel setting introduced in [44], where it is motivated as a two stage noise cancellation problem. The mappings obtained are highly nontrivial, offer considerably improved performance, and raise interesting questions about the functional properties of optimal mappings in decentralized control, which are the focus of ongoing research.

6.2 Future Directions

- **Extension to other settings:** There are other network settings such as multiple access channel that are left for future work. Our method can be extended to such settings and their optimal mappings can be analyzed.

- **Analysis of the structure of obtained mappings:** More detailed study of the structure of obtained mappings, such as analytic expressions that approximate numerical results, is left as future work.

- **Theoretical results on the structure of optimal mappings:** Similar to the results for optimal mappings in WCE, such as symmetry around origin, theoretical findings about the structure of optimal mappings can make numerical optimization methods faster as they reduce the search space.

- **Fundamental limits:** Shannon's information theoretic bounds are achievable, in general, only by allowing infinite delay and arbitrarily high complexity. They are therefore not applicable to the delay constrained networks that we considered in this work. One of the challenges of future work is to calculate achievable limits in constrained delay and complexity.

Bibliography

[1] R. Ahlswede and J. Korner. Source coding with side information and a converse for degraded broadcast channels. *Information Theory, IEEE Transactions on*, 21(6):629–637, 1975.

[2] E. Akyol, K. Rose, and T. Ramstad. Optimized analog mappings for distributed source-channel coding. In *Data Compression Conference (DCC)*, pages 159–168, March 2010.

[3] E. Akyol, K. Viswanatha, K. Rose, and T. Ramstad. On zero delay source-channel coding. *IEEE Transactions on Information Theory*, 60(12):7473–7489, 2014.

[4] I. Alustiza, P.M. Crespo, and B. Beferull-Lozano. Analog multiple description joint source-channel coding based on lattice scaling. *IEEE Transactions on Signal Processing*, 63(12):3046–3061, 2015.

[5] T. Başar. Variations on the theme of the Witsenhausen counterexample. In *47th IEEE Conference on Decision and Control Proceedings (CDC)*, pages 1614–1619. IEEE, 2008.

[6] M. Baglietto, T. Parisini, and R. Zoppoli. Numerical solutions to the Witsenhausen counterexample by approximating networks. *Automatic Control, IEEE Transactions on*, 46(9):1471–1477, 2001.

[7] R. Bansal and T. Başar. Stochastic teams with nonclassical information revisited: When is an affine law optimal? *Automatic Control, IEEE Transactions on*, 32(6):554–559, 1987.

[8] L. Bao, M. Skoglund, and K.H. Johansson. Encoder decoder design for event-triggered feedback control over bandlimited channels. In *American Control Conference, 2006*, pages 4183–4188, 2006.

[9] X. Chen and E. Tuncel. Zero-delay joint source-channel coding using hybrid digital-analog schemes in the Wyner-Ziv setting. *IEEE Transactions on Communications*, 62(2):726–735, 2014.

[10] C. Choudhuri and U. Mitra. On Witsenhausen's counterexample: The asymptotic vector case. In *Information Theory Workshop (ITW), 2012 IEEE*, pages 162–166, 2012.

[11] S.Y. Chung. *On the construction of some capacity approaching coding schemes*. PhD thesis, Massachusetts Institute of Technology, 2000.

[12] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. J.Wiley New York, 1991.

[13] M. Deng and Y. Ho. An ordinal optimization approach to optimal control problems. *Automatica*, 35(2):331 – 338, 1999.

[14] A. Erdozain, P.M. Crespo, and B. Beferull-Lozano. Multiple description analog joint source-channel coding to exploit the diversity in parallel channels. *IEEE Transactions on Signal Processing*, 60(11):5880–5892, 2012.

[15] A. Erdozain, P.M. Crespo, and B. Beferull-Lozano. Optimum distortion exponent in parallel fading channels by using analog joint source-channel coding schemes. In *Data Compression Conference (DCC)*, pages 277–286, 2012.

[16] P. Floor, A. Kim, N. Wernersson, T. Ramstad, M. Skoglund, and I. Balasingham. Zero-delay joint source-channel coding for a bivariate Gaussian on a Gaussian MAC. *IEEE Transactions on Communications*, 60(10):3091–3102, 2012.

[17] P.A. Floor, A.N. Kim, T.A. Ramstad, and I. Balasingham. Zero delay joint source channel coding for multivariate Gaussian sources over orthogonal Gaussian channels. *Entropy*, 15(6):2129–2161, 2013.

[18] P.A. Floor, T.A. Ramstad, and N. Wernersson. Power constrained channel optimized vector quantizers used for bandwidth expansion. In *4th International Symposium on Wireless Communication Systems (ISWCS)*, pages 667–671, Oct 2007.

[19] A. Fuldseth and T.A. Ramstad. Bandwidth compression for continuous amplitude channels based on vector approximation to a continuous subset of the source signal space. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 4, pages 3093–3096, Apr 1997.

[20] S. Gadkari and K. Rose. Robust vector quantizer design by noisy channel relaxation. *IEEE Transactions on Communications*, 47(8):1113–1116, 1999.

[21] A. E. Gamal and Y. Kim. *Network Information Theory*. Cambridge University Press, 2011.

[22] M. Gastpar, B. Rimoldi, and M. Vetterli. To code, or not to code: Lossy source-channel communication revisited. *IEEE Transactions on Information Theory*, 49(5):1147–1158, 2003.

[23] A. Giridhar and P.R. Kumar. Toward a theory of in-network computation in wireless sensor networks. *IEEE Communications Magazine*, 44(4):98–107, 2006.

[24] G. Gnecco and M. Sanguineti. New insights into Witsenhausen's counterexample. *Optimization Letters*, 6(7):1425–1446, 2012.

[25] G. Gnecco, M. Sanguineti, and M. Gaggero. Suboptimal solutions to team optimization problems with stochastic information structure. *SIAM Journal on Optimization*, 22(1):212–243, 2012.

[26] T. Goblick Jr. Theoretical limitations on the transmission of data from analog sources. *IEEE Transactions on Information Theory*, 11(4):558–567, 1965.

[27] M. Goldenbaum and S. Stanczak. Robust analog function computation via wireless multiple-access channels. *IEEE Transactions on Communications*, 61(9):3863–3877, 2013.

[28] V.K. Goyal. Multiple description coding: Compression meets the network. *Signal Processing Magazine*, 18(5):74–93, 2002.

[29] P. Grover, Se Yong Park, and A. Sahai. Approximately optimal solutions to the finite-dimensional Witsenhausen counterexample. *Automatic Control, IEEE Transactions on*, 58(9):2189–2204, 2013.

[30] F. Hekland, P.A. Floor, and T.A. Ramstad. Shannon-Kotelnikov mappings in joint source-channel coding. *IEEE Transactions on Communications*, 57(1):94–105, 2009.

[31] F. Hekland, P.A. Floor, and T.A. Ramstad. Shannon-Kotel'nikov mappings in joint source-channel coding. *IEEE Transactions on Communications*, 57(1):94–105, 2009.

[32] Y.-C. Ho and J.T. Lee. Granular optimization: An approach to function optimization. In *Decision and Control, 2000. Proceedings of the 39th IEEE Conference on*, volume 1, pages 103–111 vol.1, 2000.

[33] Y.C. Ho and K.C. Chu. Team decision theory and information structures in optimal control problems–Part I. *Automatic Control, IEEE Transactions on*, 17(1):15–22, 1972.

[34] Y.C. Ho and K.C. Chu. Information structure in dynamic multi-person control problems. *Automatica*, 10(4):341 – 351, 1974.

[35] Y. Hu, J. Garcia-Frias, and M. Lamarca. Analog joint source-channel coding using non-linear curves and MMSE decoding. *IEEE Transactions on Communications*, 59(11):3016–3026, 2011.

[36] J. Karlsson, A. Gattami, T. Oechtering, and M. Skoglund. Iterative source-channel coding approach to Witsenhausen's counterexample. In *American Control Conference (ACC), 2011*, pages 5348–5353. IEEE, 2011.

[37] J. Karlsson and M. Skoglund. Optimized low delay source channel relay mappings. *IEEE Transactions on Communications*, 58(5):1397–1404, 2010.

[38] V.A. Kotelnikov. *The theory of optimum noise immunity.* New York: McGraw-Hill, 1959.

[39] J. Kron, F. Alajaji, and M. Skoglund. Low-delay joint source-channel mappings for the Gaussian MAC. *IEEE Communications Letters*, 18(2):249–252, 2014.

[40] A. Kulkarni and T. Coleman. An optimizer's approach to stochastic control problems with nonclassical information structures. *Automatic Control, IEEE Transactions on*, PP(99):1–1, 2014.

[41] A. Kwok and S. Martinez. A distributed deterministic annealing algorithm for limited-range sensor coverage. *Control Systems Technology, IEEE Transactions on*, 19(4):792–804, 2011.

[42] J.T. Lee, E. Lau, and Y.-C. Ho. The Witsenhausen counterexample: a hierarchical search approach for nonconvex optimization problems. *Automatic Control, IEEE Transactions on*, 46(3):382–397, 2001.

[43] N. Li, J. Marden, and J. Shamma. Learning approaches to the Witsenhausen counterexample from a view of potential games. In *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*, pages 157–162. IEEE, 2009.

[44] N.C. Martins. Witsenhausen's counter example holds in the presence of side information. In *Decision and Control, 2006 45th IEEE Conference on*, pages 1111–1116, 2006.

[45] W.M. McEneaney and S.H. Han. Optimization formulation and monotonic solution method for the Witsenhausen problem. *Automatica*, 55:55 – 65, 2015.

[46] M. Mehmetoglu, E. Akyol, and K. Rose. A deterministic annealing approach to witsenhausen's counterexample. In *2014 IEEE International Symposium on Information Theory (ISIT)*, pages 3032–3036, June 2014.

[47] V. Misra, V. Goyal, and L. Varshney. Distributed scalar quantization for computing: High-resolution analysis and extensions. *IEEE Transactions on Information Theory*, 57(8):5298–5325, 2011.

[48] M. Morozkov, O. Granichin, Z. Volkovich, and X. Zhang. Fast algorithm for finding true number of clusters. Applications to control systems. In *Control and Decision Conference (CCDC), 2012 24th Chinese*, pages 2001–2006, 2012.

[49] B. Nazer and M. Gastpar. Computation over multiple-access channels. *IEEE Transactions on Information Theory*, 53(10):3498–3516, 2007.

[50] R. H. Olsson and K. D. Wise. A three-dimensional neural recording microsystem with implantable data compression circuitry. *IEEE Journal of Solid-State Circuits*, 40(12):2796–2804, Dec 2005.

[51] A. Orlitsky and J.R. Roche. Coding for computing. *IEEE Transactions on Information Theory*, 47(3):903–917, 2001.

[52] S. S. Pradhan and K. Ramchandran. Distributed source coding using syndromes (discus): design and construction. *IEEE Transactions on Information Theory*, 49(3):626–643, Mar 2003.

[53] R. Radner. Team decision problems. *The Annals of Mathematical Statistics*, 33(3):857–881, 09 1962.

[54] T.A. Ramstad. Shannon mappings for robust communication. *Telektronikk*, 98(1):114–128, 2002.

[55] A.V Rao, D. Miller, K. Rose, and A. Gersho. A deterministic annealing approach for parsimonious design of piecewise regression models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(2):159–173, 1999.

[56] K. Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proceedings of the IEEE*, 86(11):2210–2239, 1998.

[57] K. Rose, E. Gurewitz, and G. Fox. Statistical mechanics and phase transitions in clustering. *Physical Review Letters*, 65(8):945–948, 1990.

[58] K. Rose, E. Gurewitz, and G. Fox. Vector quantization by deterministic annealing. *IEEE Transactions on Information Theory*, 38(4):1249–1257, 1992.

[59] M. Rotkowitz and S. Lall. A characterization of convex problems in decentralized control∗. *Automatic Control, IEEE Transactions on*, 51(2):274–286, Feb 2006.

[60] N. Saldi, S. Yüksel, and T. Linder. Finite Model Approximations and Asymptotic Optimality of Quantized Policies in Decentralized Stochastic Control. *ArXiv e-prints*, 2015.

[61] Dongjin Seo, Ryan M. Neely, Konlin Shen, Utkarsh Singhal, Elad Alon, Jan M. Rabaey, Jose M. Carmena, and Michel M. Maharbiz. Wireless recording in the peripheral nervous system with ultrasonic neural dust. *Neuron*, 91(3):529–539, 2016/08/29.

[62] Mijail D. Serruya, Nicholas G. Hatsopoulos, Liam Paninski, Matthew R. Fellows, and John P. Donoghue. Brain-machine interface: Instant neural control of a movement signal. *Nature*, 416(6877):141–142, 03 2002.

[63] C. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(1):379–423, 1948.

[64] C.E. Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, 1949.

[65] P. Sharma, S.M. Salapaka, and C.L. Beck. Entropy-based framework for dynamic coverage and clustering problems. *Automatic Control, IEEE Transactions on*, 57(1):135–150, 2012.

[66] D. Slepian and J. Wolf. Noiseless coding of correlated information sources. *IEEE Transactions on Information Theory*, 19(4):471–480, 1973.

[67] V.A. Vaishampayan and S.I.R. Costa. Curves on a sphere, shift-map dynamics, and error control for continuous alphabet sources. *IEEE Transactions on Information Theory*, 49(7):1658–1672, 2003.

[68] A.B. Wagner, S. Tavildar, and P. Viswanath. Rate region of the quadratic Gaussian two-encoder source-coding problem. *IEEE Transactions on Information Theory*, 54(5):1938–1961, 2008.

[69] N. Wernersson, J. Karlsson, and M. Skoglund. Distributed quantization over noisy channels. *IEEE Transactions on Communications*, 57(6):1693–1700, 2009.

[70] N. Wernersson and M. Skoglund. Nonlinear coding and estimation for correlated data in wireless sensor networks. *IEEE Transactions on Communications*, 57(10):2932–2939, 2009.

[71] N. Wernersson, M. Skoglund, and T. Ramstad. Polynomial based analog source channel codes. *IEEE Transactions on Communications*, 57(9):2600–2606, 2009.

[72] H. Witsenhausen. The zero-error side information problem and chromatic numbers (corresp.). *IEEE Transactions on Information Theory*, 22(5):592–593, Sep 1976.

[73] H.S. Witsenhausen. A counterexample in stochastic optimum control. *SIAM Journal on Control and Optimization*, 6(1):131–147, 1968.

[74] H.S. Witsenhausen. A standard form for sequential stochastic control. *Mathematical systems theory*, 7(1):5–11, 1973.

[75] Y. Wu and S. Verdu. Witsenhausen's counterexample: A view from optimal transport theory. In *Decision and Control and European Control Conference (CDC-ECC), 2011 50th IEEE Conference on*, pages 5732–5737, 2011.

[76] A. Wyner. A theorem on the entropy of certain binary sequences and applications–ii. *IEEE Transactions on Information Theory*, 19(6):772–777, Nov 1973.

[77] A. Wyner. On source coding with side information at the decoder. *Information Theory, IEEE Transactions on*, 21(3):294–300, 1975.

[78] A. Wyner and J. Ziv. The rate-distortion function for source coding with side information at the decoder. *IEEE Transactions on Information Theory*, 22(1):1–10, 1976.

[79] Yunwen Xu, S.M. Salapaka, and C.L. Beck. Clustering and coverage control for systems with acceleration-driven dynamics. *Automatic Control, IEEE Transactions on*, 59(5):1342–1347, 2014.

[80] S. Yuksel. Stochastic nestedness and the belief sharing information pattern. *Automatic Control, IEEE Transactions on*, 54(12):2773–2786, 2009.

[81] S. Yüksel and T. Başar. *Stochastic Networked Control Systems: Stabilization and Optimization under Information Constraints*. Springer, 2013.