Observation Uncertainty in Gaussian Sensor Networks

Anand D. Sarwate

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2006-3
http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-3.html

January 23, 2006

Acknowledgement

I'd like to thank my advisor, Professor Michael Gastpar, for guiding this work, as well as Professor Anant Sahai for useful feedback on the draft. I had helpful discussions about this work with Bobak Nazer, Dan Hazen, and Krishnan Eswaran.

This work was supported by an NDSEG Fellowship from the United States Department of Defense, and the National Science Foundation under award CCF-0347298.

Finally, thanks to Elizabeth Foster-Shaner for her infinite patience.

Observation Uncertainty in Gaussian Sensor Networks

by

Anand Dilip Sarwate

S.B. Electrical Engineering (Massachusetts Institute of Technology), 2002
S.B. Mathematics (Massachusetts Institute of Technology), 2002

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

Master of Science

in

Engineering - Electrical Engineering and Computer Sciences

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, BERKELEY

Committee in charge:

Professor Michael Gastpar, Chair
Professor Anant Sahai

Fall 2005

The thesis of Anand Dilip Sarwate is approved.

Chair	Date
	Date

University of California, Berkeley

Fall 2005

Observation Uncertainty in Gaussian Sensor Networks

by

Anand Dilip Sarwate

Abstract

Observation Uncertainty in Gaussian Sensor Networks

by

Anand Dilip Sarwate

Master of Science in Engineering - Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Michael Gastpar, Chair

The term "sensor network" encompasses a wide range of engineering systems with dramatically different characteristics. We consider a specific class of sensor networks whose objective is to reconstruct a source at a central terminal. Our objective in this thesis is to quantify the asymptotic error in reconstructing the source as the number of data sources, sensors, and model complexity increases. We consider three types of estimation systems – unconstrained estimators for vector Gaussian sources that are allowed direct access to the sensor observations, estimators for discrete sources that receive information via rate constrained links from the sensors, and estimators for scalar Gaussians whose input is the output of a multiple-access channel.

We first establish bounds on the optimal estimator performance of these networks using a centralized estimator with access to all of the sensor observations. We assume the observations are noisy linear functions of the source and are thus specified by a matrix. Because the asymptotic error depends only on the spectral properties of this matrix, we can use tools from matrix analysis to give bounds on the spectrum and error in terms of the entries of the matrix for a number of different scenarios. Finally, we look at the case where the matrix is partially unknown. In some

cases we can estimate the matrix directly from the data and in others we must minimize the worst mismatch distortion.

These problems can also be looked at in a more information-theoretic framework. We look at a lossless distributed source coding problem in which the joint distribution of the sources is partially unknown. Although for any finite number of sensors standard multi-terminal source codes can easily be adapted to handle the model uncertainty across time, we show a rate penalty is incurred if the number of sensors and blocklength go to ∞ simultaneously. This represents one kind of tradeoff between delay and complexity for the scaling behavior of these systems.

Finally, we look at the case where the sensors must communicate their observations across an additive white Gaussian noise multiple-access channel. With a known correlation structure, the optimal error converges to 0 as $1/M$, where M is the number of sensors. However, a simple feedback scheme using K bits broadcast to all sensors can provide a distortion that scales to 0 as $M^{-K/(K+2)}$. We conjecture that providing similar feedback to an optimal source code will not improve the performance beyond that of our protocol.

Professor Michael Gastpar
Thesis Committee Chair

I dedicate this thesis to my parents, Dilip V. Sarwate and Sandhya D. Sarwate, and to my brother, Sanjiv D. Sarwate. Without their love and support I would never have gotten this far.

Contents

Contents ii

List of Figures iv

Acknowledgements v

1 Prelude: a model for sensor networks 1

1.1 A toy example . . . 2

1.2 Problem descriptions and main results . . . 3

1.2.1 Discrete sources and distribution uncertainty . . . 4

1.2.2 Gaussian sources and fading observations . . . 5

1.3 Where we are going . . . 8

2 Scherzo: centralized estimation from linear models in AWGN 10

2.1 Linear estimation error via matrix spectra : a review . . . 11

2.1.1 Problem statement . . . 11

2.1.2 Bounded energy observations . . . 17

2.1.3 Observation models from LTI filters . . . 21

2.1.4 Random matrices with iid entries . . . 26

2.2 Estimation with fading observations . . . 27

2.2.1 Slow fading . . . 28

2.2.2 Fast fading . . . 30

2.3 Our toy example . . . 33

3 Rondo: source coding with uncertainty 35

3.1 Slepian-Wolf coding over a class of distributions . . . 36
3.2 Multi-terminal coding with many terminals . . . 42
3.3 The example revisited . . . 44

4 Finale: fading observations and alignment **46**
4.1 Uncertainty in observations . . . 47
4.1.1 Fading observations : a general model . . . 47
4.1.2 Scalar multiplicative fading . . . 50
4.2 Existing schemes . . . 51
4.2.1 Separate source and channel coding . . . 52
4.2.2 Uncoded transmission . . . 53
4.3 A simple feedback framework . . . 55
4.3.1 A single bit of feedback for sign fading . . . 56
4.3.2 Example: feedback from a beacon sensor . . . 59
4.3.3 Many bits of feedback . . . 61
4.3.4 Feedback for bounded scalar fading . . . 64
4.3.5 A conjecture for the CEO problem with limited feedback . . . 65
4.4 Other directions and our example . . . 66

5 Coda **70**

Bibliography **72**
References . . . 72

List of Figures

1.1 The distributed source coding problem. Each terminal views one component of a correlated source and encodes it into a rate-limited message. The decoder uses all the messages to reconstruct the sources. 4

1.2 The network considered in Chapter 4. 7

2.1 General diagram for remote observation via a known matrix A. 12

2.2 A (contrived) example of a circulant network with $M = L$. The black circles on the inner ring are sources observed by an outer ring of sensors represented by the gray circles. 15

2.3 An example of a network on a line. Imagine the two lines are actually the same line. The squares mark the locations of the sources, and the circles mark the locations of the sensors. The sensor observations are the superposition of the impulse response of the filter centered at the sensor locations. The upsampling ratio N is shown via the dotted lines. 22

2.4 Filtering model for remote sensing. 22

2.5 Wiener filter solution for spatial filtering. 23

3.1 Source coding for a class of sources. 37

3.2 Binary correlated source with one remote component. Conditioned on the first component, the decoder can figure out if the second one was complemented or not. . . . 42

4.1 Sensor network with fading observations. The function A can be arbitrary, but we will generally assume that it is a linear transformation. 48

4.2 Gaussian network with fading observations. 50

4.3 MAP rules for perfect feedback and perfect sign feedback. The plus is the noisy observation $(u_m[0], u_m[-1])$. Under perfect information, $\hat{A}_m = 1$, whereas with only sign information the *expected probability of success* is maximized when $\hat{A}_m = -1$. . 62

Acknowledgements

I'd like to thank my advisor, Professor Michael Gastpar, for helping me through this work, as well as Professor Anant Sahai for providing useful feedback on the draft with such short notice. I had many helpful discussions about this work with my fellow students, especially Bobak Nazer, Dan Hazen, and Krishnan Eswaran.

The work in this thesis was supported by an NDSEG Fellowship, which is sponsored by the United States Department of Defense, as well as by the National Science Foundation under award CCF-0347298.

Finally, thanks to Elizabeth Foster-Shaner for her infinite patience.

Chapter 1

Prelude: a model for sensor networks

Consider the following hypothetical scenario: many sensors are placed around the watershed of a city in order to monitor contaminant levels in the ground water. These contaminants may have been introduced, for example, by illegal dumping of waste. The goal of the network is to measure the concentration levels and report this information back to a monitoring station. Every sensor can only measure the concentration of a single chemical that is the by-product of several types of contaminants, so the observation of an individual sensor may not be very informative. The sensors have a small processor, a wireless radio to communicate with the monitoring station, and limited battery power. The engineering problem is to design an efficient system for tracking the contaminant levels over time.

This is a problem of data-gathering and estimation using a wireless sensor network. We are interested in the theoretical bounds on the estimation error at the central observer and what happens to these bounds as the number of sources and sensors increase. In order to accurately address these questions we must have a model of remote sensing that is both rich enough to capture the problems specific to this application and simple enough to be amenable to theoretical analysis. In general, the complexity of real-world sensing scenarios is not accurately reflected in the models studied by theoreticians. Even though the physics of the observation mechanism may be well-understood, the resulting model may be intractable. Different modeling techniques used

on the observation and communication halves of the problem may cause difficulties in merging the two. Finally, those results which can be proved may yield little insight into engineering tradeoffs or may be so tailored to a specific situation as to be ungeneralizable.

In this thesis we will investigate a very specific class of data-gathering sensor networks. We will introduce structured uncertainty into the mapping between the observed variables at the sensors (e.g. concentration levels of a chemical by-product) and an underlying data source of interest (e.g. concentration levels of contaminants). This modeling uncertainty is different from the uncertainty caused by noise in the observations; it is uncertainty about how the observations are related *to each other* rather than their reliability.

A sensor network designed for estimation will have different performance limits depending on the constraints on communication among the sensors and between the sensors and the base station. Correspondingly, we look at three different scenarios: centralized estimation, lossless multi-terminal source coding, and estimation over a shared additive multiple-access channel. We will will first discuss a toy example to show what we mean by observation uncertainty and then describe our three problems and main results.

1.1 A toy example

Suppose $\{S_1[n]\}$ and $\{S_2[n]\}$ are a pair of iid discrete-time Gaussian random processes with mean 0 and variance σ_1^2 and σ_2^2, respectively. These two processes represent different sources that we would like to estimate using a sensor network. The network consists of M sensors, each of which observes a discrete-time process $\{U_j[n]\}$ for $j = 1, 2, \ldots M$. These processes are given by the equation

$$U_j[n] = \begin{bmatrix} A_{1j} & A_{2j} \end{bmatrix} \begin{bmatrix} S_1[n] \\ S_2[n] \end{bmatrix} + W_j[n] \tag{1.1}$$

where $\{\{W_j[n]\} : j = 1, 2, \ldots M\}$ is a collection of independent iid Gaussian processes with mean 0 and variance σ_W^2. The pair $[A_{1j}\ A_{2j}]$ is equal to [0 1] or [1 0] equiprobably, but does not change

over time. Gathering the equations into a matrix we have

$$U = AS + W \ . \tag{1.2}$$

This models the case where each sensor observes exactly one of the two sources through noise.

This example gives an idea for what we mean by observation uncertainty. Each sensor does not know *a priori* which source it is observing. Another view is that the covariance of the matrix A is uncertain or that the joint distribution of U is uncertain. If $\sigma_1 \neq \sigma_2$ then each sensor can compute its own empirical variance and make an estimate of which source it is observing with exponentially small probability of error. In this scenario the modeling uncertainty is resolvable at the sensors given a sufficient amount of observed data.

However, if $\sigma_1 = \sigma_2$ then we must come up with something more clever. Depending on the access an the estimator has to the sensor's information, information about the correlation between the sensors may become costly as M increases. If the estimator has direct access to the sensor observations, it can try to sort the sensors by measuring their correlation and then estimate the sources separately. If the sensors must compress their observations before transmitting them, they may include some overhead to allow the estimator to do this sorting. This overhead could be avoided if the sensors can communicate between themselves, as we shall see.

1.2 Problem descriptions and main results

In this section we will describe two different frameworks for the sensor network problem as well as what we mean by uncertainty in observations. For sources taking values in a discrete set, we assume that each sensor observes a different source directly but that the joint distribution of all the sources is unknown. For Gaussian sources, we assume that the sensor observations are a noisy linear transformation of the sources. In all cases we assume a discrete-time model for the source process as well as any communication channels.

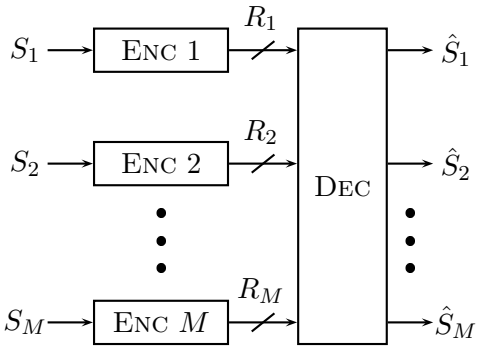

Figure 1.1. The distributed source coding problem. Each terminal views one component of a correlated source and encodes it into a rate-limited message. The decoder uses all the messages to reconstruct the sources.

1.2.1 Discrete sources and distribution uncertainty

In Chapter 3 we address the problem of multi-terminal source coding with distribution uncertainty. The picture is shown in Figure 1.1. We assume that the source to be estimated is a tuple $\mathbf{S} = (S_1, S_2, \ldots, S_M)$, where each component $S_m \in \mathcal{S}_m$ and $\mathcal{S}_m$ are finite sets. These sources have some joint distribution $P(\mathbf{S})$ that is known to lie in a set of distributions Λ. Sensor m observes the sequence $S_m^n = (S_m[1], S_m[2] \ldots S_m[n])$ of source samples and maps it into one of 2^{nR_m} possible messages. The goal is to find the set of *rate tuples* $(R_1, \ldots, R_M)$ such that the decoder can recover the original source sequences with a probability of error that goes to 0 as n goes to ∞.

We will assume that the true distribution $P(\mathbf{S})$ cannot be estimated from the marginal distributions of the sensors, and that Λ consists of these "indistinguishable" distributions. In this case, the set of rates is limited by the worst-case distributions in the class Λ [7]. We give a construction via binning in the style of Cover and Thomas [6] for this result, which gives an explicit characterization of the (negligible) overhead needed to compensate for the class Λ. The overhead is the form $\log |\Lambda| n^{-1} \log n$, which corresponds to rate needed to communicate the joint type of a distribution in Λ.

Since the focus of this thesis is on scaling behaviors, we are interested in the case where $M \to \infty$

as well as $n \to \infty$. Because Λ is a function of M, the complexity of our model may increase exponentially in the number of sensors. For a fixed M, we can write the asymptotic excess rate as $n \to \infty$ as $\log |\Lambda_M| n^{-1} \log n$. However, if M increases simultaneously with the blocklength, and Λ_M grows exponentially in M, for a fixed n this excess rate may not converge to 0. Taking the blocklength as a proxy for the processing delay and $|\Lambda_M|$ as a proxy for the model complexity, we can show that if M grows faster than $\frac{n}{\log n}$ the region of achievable rates must shrink to accommodate the information communicated to the decoder about the joint distribution.

1.2.2 Gaussian sources and fading observations

In contrast to the discrete case, for continuous sources we will assume that the number of sources L is smaller than the number of sensors M. We model the underlying source as an iid Gaussian process $\{\mathbf{S}[n]\}$ taking values in $\mathbb{R}^L$. At each time $\mathbf{S}[n]$ is jointly Gaussian with mean 0 and variance $\sigma_S^2 I$. We view this as L independent sources which we would like to estimate to minimize an expected squared-error criterion.

The observed signal $U_m[n]$ at sensor m is given by the equation

$$U_m[n] = A_m(\{\mathbf{S}[k] : k \leq n\}) + W_m[n] \ . \tag{1.3}$$

This follows from the Wold decomposition, which says that we can decompose the process U_m conditioned on $\mathbf{S}$ into a deterministic part (a function of $\mathbf{S}$) and an additive stochastic process W_m, which we view as noise. To make things even more simple, we will assume that W_m is iid across time and space according to some probability measure μ_W.

By observation uncertainty, we mean the $\{A_m\}$ are themselves random variables that take values in a set of functions. For example, suppose that there is only one source and that the sensor observation U_m is a noisy weighted average of the previous $N+1$ time samples of the source:

$$U_m[n] = \sum_{k=0}^{N} A_m[k] S[n-k] + W_m[n] = (A_m * S)[n] + W_m[n] \ . \tag{1.4}$$

Although this is just a linear filter, we may not know the filter coefficients exactly; they may depend on the sensor's physical location with respect to the source. We can either treat the filter $A_m[n]$

as a unknown parameter or (in Bayesian style) as a random variable with some prior distribution $p_A(\cdot)$.

Let us collect the functions A_m into a vector A that takes values in a set of functions $\mathcal{A}$. The choice of $\mathcal{A}$ will reflect the degree of uncertainty in the observation function. If the realization of A is known to the sensors, then we can condition on A and find the average behavior by taking the expectation with respect to $p_A(\cdot)$. A more interesting case is when A is unknown to the sensors, so that we must design strategies robust to the choice of A.

We call this model *fading observations* in analogy to fading communication channels, which are used to model wireless links with multipath interference. We identify two major distinctions – fast and slow fading. In fast fading, the realization of A changes at every time step. In our filtering example above, this would mean that the filter used to compute $U_m[n]$ is different from the filter used to compute $U_m[n+1]$. Fast fading may occur as a result of source or sensor mobility or may have to do with the physics of the quantity measured by the sensors. In slow fading A is chosen once and fixed for all time, albeit unknown to the sensors. Again, the choice of slow versus fast fading models is application dependent.

In both Chapter 2 and 4 we will assume that A is a matrix in $\mathbb{R}^{M\times L}$. Conditioned on knowing A, the sensor observations are jointly Gaussian, so the optimal centralized estimator is linear. In Chapter 2 we review MMSE estimation for Gaussian random variables and express the error in terms of the singular values of A. We can then use results from matrix analysis to analyze the distortion for different constraints on the entries of A and the estimator. In the case where A is unknown but slow-fading, we show that a centralized estimator can in some cases estimate A and then do the same MMSE estimation as before. However, for fast-fading A the optimal strategy is less clear. We examine the case when the estimator must be linear. In the case where the fading distribution $p_A(\cdot)$ is known, we can find the best linear estimator. In the case where $p_A(\cdot)$ is unknown, we formulate the problem as finding a linear estimator to minimize the worst mismatch.

In Chapter 4 we assume the sensors must communicate their observations over the additive white Gaussian noise channel shown in Figure 1.2. Rather than the rate constraints of Chapter 3,

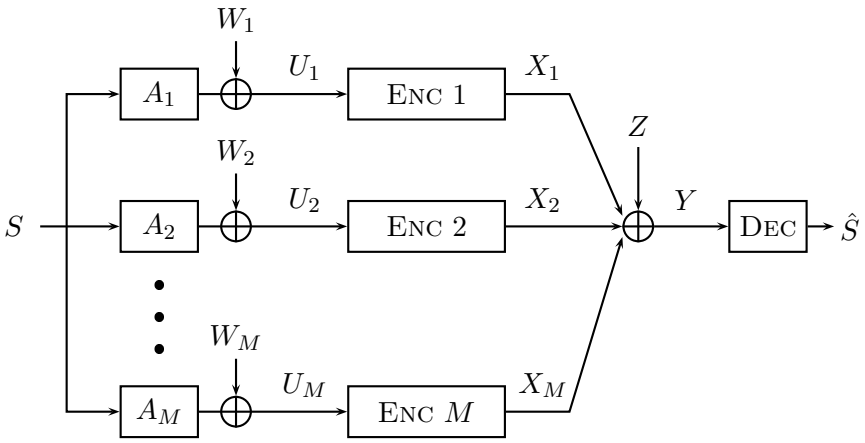

Figure 1.2. The network considered in Chapter 4.

we assume the sensors' communication is power-limited:

$$E\left[|X_m[n]|^2\right] \leq P \ . \tag{1.5}$$

We assume there is a single source ($L = 1$) and a slow-fading matrix A that is a vector with iid entries according to some bounded zero-mean distribution.

In the absence of fading, two strategies have been proposed in the literature. The off-the-shelf solution is to have the sensors use the optimal distributed lossy compression scheme and then transmit their compressed messages losslessly across the communication channel. The lowest distortion achievable for a rate R is proportional to $1/R$ [28] and the highest rate achievable across the channel is proportional to $\log M$, so the end-to-end distortion is $1/\log M$. However, if the sensors simply transmit their observations raw [18] but scaled up to the power constraint of the channel, the distortion scales like $1/M$. We call the former scheme *separation-based transmission* and the latter *uncoded transmission*.

How does the slow-fading model affect the performance of these two strategies? The key problem is that the sensors cannot estimate locally the sign of their observation. We show that this type of sign-ambiguity, which is detrimental to centralized encoders, renders the uncoded transmission

scheme useless – on average, the sensors cancel themselves out and the received signal-to-noise ratio goes to 0 as $M \to \infty$. Thus the distortion does not improve at all with more sensors.

Recall that in the case of centralized estimation with slow fading, the estimator could sometimes disambiguate between the different fading possibilities. How much information do the sensors need to perform a similar hypothesis test in this scenario? We propose sharing the signal of one sensor (a "beacon") with all the others. An open conjecture is that this extra side information does not affect the scaling rate of the optimal distributed source code. We show that even if the beacon's signal is quantized to 1 bit, K samples of this side information is sufficient to give the uncoded transmission protocol an distortion scaling rate of $O(M^{-K/(K+2)})$. If the beacon transmits every time slot, the scaling rate will approach the optimal $O(M^{-1})$.

1.3 Where we are going

Our interest in this thesis is on the performance achievable as the number of sensors tends to infinity. The benefits of looking at the asymptotic performance are twofold. Firstly, because the models we use are gross simplifications of real networks, a tight characterization of the performance is impossible, so scaling behaviors may give more insight than small system designs. A more aesthetic benefit is that many of the expressions have "nice" limiting behavior. However, there is a danger to considering only the asymptotic picture: intuitions valid for finite networks break down in the limit. To see this, consider the network density. To increase the number of sensors, we can let the area covered by the network expand to keep density constant or keep the area fixed and let the density go to infinity. In the latter case, the distance between sensors will become very small, and the aggregate signal-to-noise ratio for a single source across the sensors may tend to infinity.

For the sensor networks studied here we presume the existence of a centralized decoder that wishes to aggregate the information sent by the sensors in order to reconstruct the source or function of the source. The constraints on the decoder's access to the sensors' observations are the motivation for the three problems studied in the remaining chapters of this thesis. In the next chapter we will look at centralized estimation, where the decoder has full access to the sensor

observations. In Chapter 3 we will look at lossless compression with rate-limited communication to the decoder. Finally, in chapter 4 we will look at a joint source-channel communication scenario.

Chapter 2

Scherzo: centralized estimation from linear models in AWGN

We turn first to estimators that have direct access to the sensor observations. Specifically, we will review MMSE estimation of Gaussian sources viewed through linear functions with additive white Gaussian noise (AWGN). In the case where the source is a vector and the functions are memoryless, the observation process is simply multiplication by a matrix. The bulk of this chapter is a review on using matrix spectra to express the estimation error and using bounds on eigenvalues to characterize observation matrices. In particular, we can quantify the effects of sensor density, dynamic range, and other engineering parameters on the limiting behavior of these systems. We will then look at cases when the matrix characterizing the observations is not known *a priori*. In the case where the matrix is unknown but not time-varying, we can first estimate the matrix and then build an optimal linear estimator. If the matrix is time-varying, we give a two characterizations of the performance of linear estimators depending on if the distribution of the time variation is known.

2.1 Linear estimation error via matrix spectra : a review

As a review, we will first describe a generic matrix model for the observations and derive conditions on the entries of the matrix for the error to converge to a constant. We then examine a specific example of this kind of matrix arising from upsampling and spatially filtering an underlying source. Finally, we describe the performance of central estimators under multiplication by a random matrix and leverage results on the spectral convergence to calculate the asymptotic mean-squared error. Our objective is to unify several different ways of generating linear models and analyze them all via the asymptotic spectra of the associated matrices, using well-known tools from matrix analysis.

2.1.1 Problem statement

The structured observation models we would like to consider are motivated by different geometric assumptions on the sources and sensors. However, the induced mathematical model is the same in all cases, and is illustrated by the block diagram in Figure 2.1. For a problem with M sensors and L sources, we are interested in the estimation error as $M \to \infty$. There are two cases of interest: M/L constant and $M/L \to \infty$. In the former we will see that the sampling density M/L will appear in the asymptotic error expressions. In the latter, we are mostly interested in whether the error converges to 0 and if so, how fast.

The source generates an independent, identically distributed (iid) sequence of vectors $S[k] \in \mathbb{R}^L$:

$$\mathbf{S} = \{S[k] : k > 0\} \tag{2.1}$$

where at each time k the vector $S[k]$ is a jointly Gaussian vector with mean 0 and covariance $\sigma_S^2 I$. These source vectors are multiplied by a matrix $A \in \mathbb{R}^{M \times L}$, called the *observation matrix*, and noise is added to form the sensor observation vector

$$U[k] = A \cdot S[k] + W[k] \tag{2.2}$$

where $\{W[k] : k > 0\}$ are iid Gaussian random vectors with mean 0 and covariance $\sigma_W^2 I$. We call equation (2.2) the *observation process*.

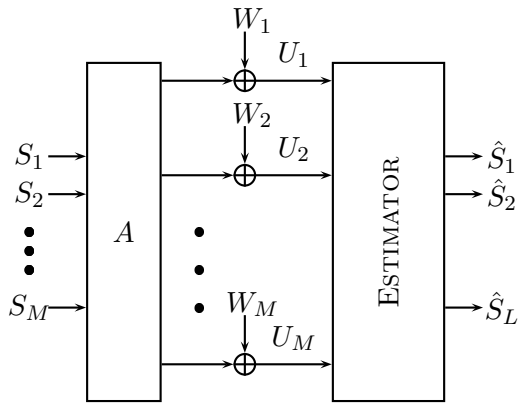

Figure 2.1. General diagram for remote observation via a known matrix A.

A *memoryless centralized estimator* for this problem is a function

$$f : \mathbb{R}^M \to \mathbb{R}^L \tag{2.3}$$

that takes an observation vector U and creates a source estimate $\hat{S} = f(U)$. The estimation error is measured by taking the expectation of a distortion function $d(S, \hat{S})$:

$$D = E[d(S, \hat{S})] \tag{2.4}$$

In our case we measure distortion by mean-squared error:

$$d(S, \hat{S}) = \frac{1}{L}\|S - \hat{S}\|^2 \ . \tag{2.5}$$

We consider memoryless estimators because they are optimal when the source is memoryless as well. However, as we will see later, this may not be the case when the matrix A changes over time.

In our simple case it is well known that the optimal estimator is linear, so that

$$\hat{S} = f(U) = F \cdot U = F(AS + W) \ . \tag{2.6}$$

The following well-known proposition gives the error for the optimal estimator in terms of the singular values of A.

Proposition 1. *Let $\{\alpha_j : j = 1, 2, \dots L\}$ be the singular values of A. Then the MMSE is given by*

$$D = \frac{1}{L} \sum_{i=1}^{L} \frac{\sigma_S^2 \sigma_W^2}{\alpha_i^2 \sigma_S^2 + \sigma_W^2} \ . \tag{2.7}$$

Proof. The estimator can be written in terms of the covariance and cross correlation matrices of the various vectors. Let

$$\Sigma_S = E[SS^T] = \sigma_S^2 I \tag{2.8}$$

$$\Sigma_W = E[WW^T] = \sigma_S^2 I \tag{2.9}$$

$$\Sigma_{US} = E[US^T] = A\Sigma_S \tag{2.10}$$

$$\Sigma_{SU} = \Sigma_{US}^T = \Sigma_S^T A^T \tag{2.11}$$

$$\Sigma_U = E[UU^T] = A\Sigma_S A^T + \Sigma_W \tag{2.12}$$

Then

$$F = \Sigma_{SU} \Sigma_U^{-1} \ . \tag{2.13}$$

We can calculate the expected squared error:

$$\begin{aligned} E[\|S - \hat{S}\|^2] &= E[\|S - \Sigma_{SU}\Sigma_U^{-1} U\|^2] \\ &= \operatorname{tr}\left[E[(S - \Sigma_{SU}\Sigma_U^{-1}U)(S - \Sigma_{SU}\Sigma_U^{-1}U)^T]\right] \\ &= \operatorname{tr}\left[\Sigma_S - \Sigma_{SU}\Sigma_U^{-1}\Sigma_{US}\right] \\ &= \sigma_S^2 L - \operatorname{tr}\left[\sigma_S^4 A^T (AA^T\sigma_S^2 + \sigma_W^2)^{-1} A\right] \\ &= \sigma_S^2 \left(L - \sigma_S^2 \operatorname{tr}\left[AA^T(AA^T\sigma_S^2 + \sigma_W^2)^{-1}\right]\right) \ . \end{aligned}$$

Let $A = U\Lambda_A V^T$ be the singular value decomposition of A, where U and V are orthogonal and Λ_A is the matrix of singular values of A with $(\Lambda_A)_{ii} = \alpha_i$. Now:

$$\begin{aligned} E[\|S - \hat{S}\|^2] &= \sigma_S^2 \left(L - \sigma_S^2 \operatorname{tr}\left[U\Lambda_A^2 U^T (U\Lambda_A^2 U^T \sigma_S^2 + \sigma_W^2)^{-1}\right]\right) \\ &= \sigma_S^2 \left(L - \sum_{i=1}^{L} \frac{\alpha_i^2 \sigma_S^2}{\alpha_i^2 \sigma_S^2 + \sigma_W^2}\right) \\ &= \sum_{i=1}^{L} \frac{\sigma_S^2 \sigma_W^2}{\alpha_i^2 \sigma_S^2 + \sigma_W^2} \ . \end{aligned}$$

□

As we can see from this equation, the optimal centralized estimation error is only a function of the singular values of the observation matrix A. Finding the asymptotic behavior of these singular values will yield the corresponding error bounds in this chapter.

Suppose that that we are interested in estimating a scalar source so that $L = 1$ and M is allowed to grow to ∞. In this case A is a vector and $\alpha_1^2 = \|A\|^2$, so

$$D = \frac{\sigma_S^2 \sigma_W^2}{\sigma_S^2 \|A\|^2 + \sigma_W^2} \ . \tag{2.14}$$

If the entries of A are all bounded away from 0 then as M increases $\|A\|^2 \to \infty$ and the distortion scales to 0 as $M \to \infty$. We will return to this estimation problem in Chapter 4. We now close with some specialized examples.

Example : Circulant network

Suppose that A has the following structure:

$$A = \begin{bmatrix} a_0 & a_{M-1} & \cdots & a_{M-L+1} \\ a_1 & a_0 & \cdots & a_{M-L+2} \\ a_2 & a_1 & \cdots & a_{M-L+3} \\ \vdots & \vdots & \vdots & \vdots \\ a_{M-1} & a_{M-2} & \cdots & a_{M-L} \end{bmatrix} \tag{2.15}$$

That is, each column of the observation matrix is a cyclic shift of the previous column. This may happen if the sensors are placed in a circle around a second circle of sources [18]. In a far-field approximation this may be a reasonable model. The singular values of A are the eigenvalues of $B = A^T A$, which has a special structure:

$$B = \begin{bmatrix} b_0 & b_1 & \cdots & b_{L-1} \\ b_{L-1} & b_0 & \cdots & b_{L-2} \\ \vdots & \vdots & \vdots & \vdots \\ b_1 & b_2 & \cdots & b_0 \end{bmatrix} \tag{2.16}$$

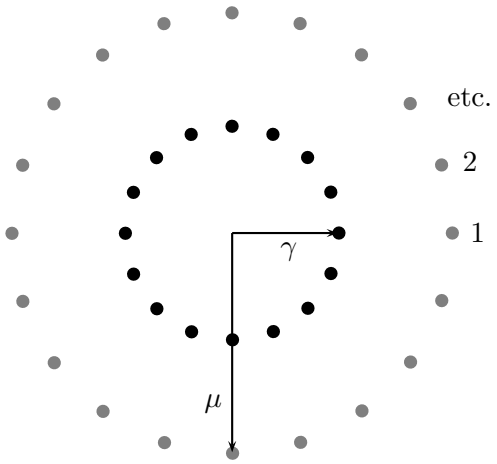

Figure 2.2. A (contrived) example of a circulant network with $M = L$. The black circles on the inner ring are sources observed by an outer ring of sensors represented by the gray circles.

This is a *circulant matrix*. A useful property of circulant matrices is that they are diagonalized by the discrete Fourier Transform (DFT) matrix [20]. The singular values are the DFT coefficients of the sequence $\{a_m\}$. Thus we can write the distortion as:

$$D = \frac{1}{L}\sum_{i=1}^{L} \frac{\sigma_S^2 \sigma_W^2}{\beta_i^2 \sigma_S^2 + \sigma_W^2} \ . \tag{2.17}$$

where

$$\beta_i = \sum_{l=0}^{L-1} b_l e^{-j2\pi \frac{l}{L}} \ , \tag{2.18}$$

the DFT of the first row of B.

For a physical example of a sensor network, consider the diagram shown in Figure 2.1.1. In this example $M = L$, with a circle of sensors surrounding a circle of sources[1]. Suppose furthermore that the sensor observations can be written as:

$$U_m = \sum_{l=1}^{L} \ell(d(m,l)) S_l \ , \tag{2.19}$$

where $d(m,l)$ is the distance from sensor m to source l and $\ell(\cdot)$ is an attenuation (path-loss) that is a function of the distance. If γ and μ are the radii of the inner and outer circles respectively,

[1]Of course, we could have the positions of the sources and sensors interchanged, so that the sensors form a circular array, much like a panopticon [16].

then we can write this as

$$U_m = \sum_{l=1}^{L} \ell\left(\left(\gamma^2 + \mu^2 - 2\gamma\mu\cos\left(2\pi\frac{|m-l|}{M}\right)\right)^{1/2}\right) S_l \ . \tag{2.20}$$

This has the desired form. By choosing a model for $\ell(\cdot)$, we can evaluate the spectra and hence the asymptotic performance for centralized estimators with this topology.

Our interest in circulant matrices is not because they provide an accurate model for sensor network observations, but because they characterize the limiting behavior of Toeplitz matrices, which arise in the LTI filtering framework later in this chapter.

Example : block models

In some instances we may be able to break our estimation problem into independent parts and write the overall distortion as the average of the different components. Consider an observation matrix A which can be written in a block form:

$$A = \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_K \end{bmatrix} . \tag{2.21}$$

Let the block A_k have M_k rows and L_k columns. Then the covariance matrix can also be written in block form

$$B = A^T A = \begin{bmatrix} A_1^T A_1 & 0 & \cdots & 0 \\ 0 & A_2^T A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & A_K^T A_K \end{bmatrix} . \tag{2.22}$$

The eigenvalues for this matrix are the union of the singular values for each A_k, $1 \le k \le K$, so

$$D = \frac{1}{L}\sum_{k=1}^{K}\sum_{i=1}^{L_k} \frac{\sigma_S^2\sigma_W^2}{\beta_{i,k}^2\sigma_S^2 + \sigma_W^2} \ , \tag{2.23}$$

where $\beta_{i,k}$ is the i-th eigenvalue of $A_k^T A_k$.

2.1.2 Bounded energy observations

We now constrain the matrix A to model a scenario in which the sensors can receive limited power and each source emits limited energy. By this we mean that the sequence of coefficients $\{a_{ml}\}$ is square summable over l for each fixed m and square summable over m for each fixed l. It is intuitive that in this case the the sources cannot be recovered perfectly because the signal to noise ratio for each source is bounded. As we noted in the last chapter, the assumptions here fit with the expanding network view of scaling.

Let $L_n = L_0 n$ and $M_n = M_0 n$ be the number of sources and sensors respectively for a problem at scale n. Let $\{a_{ml} : l > 0, m > 0\}$ be a 2-dimensional array of real numbers and define $A^{(n)} = \{a_{ml} : 1 \leq l \leq L_n, 1 \leq m \leq M_n\}$ be the observation matrix at scale n.

Bounded row and column norms

The constraints we impose are the following:

$$\|A^{(n)}\|_1 = \max_{1 \leq l \leq \infty} \sum_{m=1}^{\infty} |a_{ml}| < \varepsilon_C \tag{2.24}$$

$$\|A^{(n)}\|_\infty = \max_{1 \leq m \leq \infty} \sum_{l=1}^{\infty} |a_{ml}| < \varepsilon_R \tag{2.25}$$

These are the *maximum column sum* and *maximum row sum* norms, respectively. The first constraint bounds the total contribution of a source to all the sensor's observations while the second bounds the contribution of the all the sources to a sensor's observation. Under these two constraints, we can bound the asymptotic distortion in terms of the constants ε_C and ε_R.

Proposition 2. *If the observation matrix A satisfies the bounds in (2.24) and (2.25), then the asymptotic distortion for a centralized estimator is bounded away from* 0.

Proof. We will use the row and column sum bounds to prove a bound on the maximum singular value of the matrix $A^{(n)}$ that is independent of n. The singular values of $A^{(n)}$ are the eigenvalues

of the $L_n \times L_n$ matrix $B = (A^{(n)})^T A^{(n)}$. Consider the column sum of B:

$$\sum_{j=1}^{L_m} B_{ij} = \sum_{j=1}^{L_n} \sum_{k=1}^{M_n} a_{ki} a_{kj} \leq \sum_{k=1}^{M_n} |a_{ki}| \sum_{j=1}^{L_n} |a_{kj}| \leq \sum_{k=1}^{M_n} |a_{ki}| \varepsilon_R \leq \varepsilon_C \varepsilon_R$$

This bound holds for all n and i so we have $\|B\|_1 < \infty$. The largest eigenvalue of B is upper bounded by any matrix norm on B [23, p. 297], so $\alpha_i \leq \varepsilon_C \varepsilon_R$ for all i.

Turning to the distortion expression in (2.7) we can easily bound the distortion away from 0:

$$D \geq \frac{1}{L_n} \sum_{i=1}^{L_n} \frac{\sigma_S^2 \sigma_W^2}{\varepsilon_C \varepsilon_R \sigma_S^2 + \sigma_W^2} = \frac{\sigma_S^2 \sigma_W^2}{\varepsilon_C \varepsilon_R \sigma_S^2 + \sigma_W^2} > 0 \ . \tag{2.26}$$

Since this bound is independent of n, the asymptotic distortion is also greater than 0. □

Simply having unbounded row or column norms is insufficient for the distortion to converge to 0. For example, we could have a matrix A whose rank is only 1:

$$A = \begin{pmatrix} 1 & 2^{-1} & 2^{-2} & 2^{-3} & \cdots \\ 1 & 2^{-1} & 2^{-2} & 2^{-3} & \cdots \\ 1 & 2^{-1} & 2^{-2} & 2^{-3} & \cdots \\ 1 & 2^{-1} & 2^{-2} & 2^{-3} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{pmatrix} \ . \tag{2.27}$$

For this choice of A, every row is summable and every column is not, but all but 1 of the singular values are 0, so the distortion does not converge to 0.

Toeplitz matrices

In some specific cases, we can obtain a closed form solution for the limit of the centralized distortion as the number of sources and sensors tend to infinity. One particular case is that of *Toeplitz* covariance matrices. Let $\{b_k : -\infty < k < \infty\}$ be a sequence of real numbers that satisfies

$$\sum_{k=-\infty}^{\infty} |b_k| = \beta < \infty \ . \tag{2.28}$$

The matrix B can be thought of as the covariance matrix of a wide-sense stationary random process. So if the sensor observations are wide-sense stationary across *space* we would get a matrix with this

structure. The corresponding observation matrix A can be thought of as a linear space-invariant filter, as discussed in the sequel. It is important to realize that the oversampling ratio M_0/L_0 is hidden in the matrix B, so that in the expressions given below we cannot simply change the value of M_0/L_0 to lower the error.

The Fourier transform of $\{b_k\}$ is

$$B(\omega) = \lim_{k\to\infty} \sum_{j=-k}^{k} a_k e^{-jk\omega} \ . \tag{2.29}$$

Let $K_n = \frac{1}{2}(M_n - 1)$. Suppose that the matrix $B = A^{(n)}(A^{(n)})^T$ is a Toeplitz matrix whose first row is $(b_0, \ldots, b_{K_n})$ and whose first column is $(b_0, \ldots, b_{-K_n})$. Since the (i,j)-th entry of B is the correlation between sensors i and j, this gives us the wide-sense stationarity across space.

The celebrated Grenander-Szegö Theorem [21, pp. 64–65] on the distribution of eigenvalues of Toeplitz forms gives us a convenient expression for the limit. The theorem states that for any function F continuous on the support of the eigenvalues, the average of the function converges to a limit:

$$\lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} F(\alpha_j) = \frac{1}{2\pi} \int_{-\pi}^{\pi} F(B(\omega)) d\omega \ . \tag{2.30}$$

In our case we have

$$\lim_{n\to\infty} \frac{1}{M_n} \sum_{j=1}^{M_n} \frac{\sigma_S^2 \sigma_W^2}{\alpha_j^{(n)} \sigma_S^2 + \sigma_W^2} = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\sigma_S^2 \sigma_W^2}{B(\omega)\sigma_S^2 + \sigma_W^2} d\omega \ . \tag{2.31}$$

where α_j^n is the j-th eigenvalue of B. Note that B is rank-deficient with $M_n - L_n$ eigenvalues equal to 0. We can rewrite the left side of the equation

$$\lim_{n\to\infty} \left(\frac{M_n - L_n}{M_n} \sigma_S^2 + \frac{1}{M_n} \sum_{j=1}^{L_n} \frac{\sigma_S^2 \sigma_W^2}{\alpha_j^{(n)} \sigma_S^2 + \sigma_W^2} \right) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\sigma_S^2 \sigma_W^2}{B(\omega)\sigma_S^2 + \sigma_W^2} d\omega \ . \tag{2.32}$$

Since $L_n/M_n = L_0/M_0$ is a constant, we have

$$\begin{aligned} D &= \lim_{n\to\infty} \frac{1}{L_n} \sum_{j=1}^{L_n} \frac{\sigma_S^2 \sigma_W^2}{\alpha_j^{(n)} \sigma_S^2 + \sigma_W^2} \\ &= \sigma_S^2 \left(1 - \frac{M_0}{L_0} \left(1 - \frac{1}{2\pi} \int_{-\pi}^{\pi} \frac{\sigma_S^2 \sigma_W^2}{B(\omega)\sigma_S^2 + \sigma_W^2} d\omega \right) \right) \ . \end{aligned} \tag{2.33}$$

This expression quantifies the benefit of the sampling density M_0/L_0 on the asymptotic distortion. Note, however, that the support of $B(\omega)$ depends on M_0/L_0; the more we oversample, the more "pinched" the spectrum becomes.

Example : harmonic decay

Suppose

$$b_n = \frac{\sin \omega_c n}{\pi n} \ , \tag{2.34}$$

which is just a sinc function. Its Fourier transform is a box

$$B(\omega) = \mathbf{1}_{(|\omega|<\omega_c)} \ . \tag{2.35}$$

So the integral breaks into two terms:

$$\begin{aligned} D &= \frac{M_0}{L_0}\frac{1}{2\pi}\left(2\sigma_S^2(\pi-\omega_c) + 2\frac{\sigma_S^2\sigma_W^2}{\sigma_S^2+\sigma_W^2}\omega_c\right) - \left(\frac{M_0}{L_0}-1\right)\sigma_S^2 \\ &= \sigma_S^2\left(1 - \frac{M_0}{L_0}\frac{1}{\pi}\frac{\sigma_S^2}{\sigma_S^2+\sigma_W^2}\omega_c\right) \end{aligned}$$

This shows the effect of the bandwidth of this spatial lowpass filter on the estimation error – the higher the bandwidth the smaller the error.

Example : exponential decay

Suppose that $b_n = \beta^{|n|}$, a two-sided exponential decay. Its Fourier transform is

$$B(\omega) = \frac{1}{|1-\beta e^{j\omega}|} = \frac{1}{1+\beta^2-2\beta\cos\omega} \ . \tag{2.36}$$

The distortion can be written as

$$D = \sigma_S^2\left(1 - \frac{M_0}{L_0}\left(1 - \frac{1}{2\pi}\int_{-\pi}^{\pi}\frac{\sigma_W^2(1+\beta^2-2\beta\cos\omega)}{\sigma_S^2+\sigma_W^2(1+\beta^2-2\beta\cos\omega)}d\omega\right)\right)$$

We have to turn to a table of integrals to solve this. Let us write:

$$\frac{\sigma_W^2(1+\beta^2-2\beta\cos\omega)}{\sigma_S^2+\sigma_W^2(1+\beta^2-2\beta\cos\omega)} = \frac{\sigma_W^2(1+\beta^2)-2\sigma_W^2\beta\cos\omega}{\sigma_S^2+\sigma_W^2(1+\beta^2)-2\sigma_W^2\beta\cos\omega} = \frac{C_1+C_2\cos\omega}{C_3+C_2\cos\omega}$$

Then [19, TI(341)] gives two possibilities depending on the value of $C_3^2 - C_2^2$:

$$\begin{aligned} C_3^2 - C_2^2 &= (\sigma_W^2(1+\beta^2)^2 + \sigma_S^2)^2 - 4\beta^2\sigma_W^2 \\ &= \sigma_W^4 \left(\left(\frac{\sigma_S^2}{\sigma_W^2} \right)^2 + 2(1+\beta^2)\frac{\sigma_S^2}{\sigma_W^2} + (1-\beta^2)^2 \right) \\ &> 0 \ . \end{aligned}$$

Thus:

$$\begin{aligned} \int_{-\pi}^{\pi} \frac{C_1 + C_2\cos\omega}{C_3 + C_2\cos\omega} &= \left[\omega + (C_1 - C_3) \left(\frac{2}{\sqrt{C_3^2 - C_2^2}} \arctan \frac{\sqrt{C_3^2 - C_2^2}\tan\frac{\omega}{2}}{C_3 + C_2} \right) \right]_{-\pi}^{\pi} \\ &= 2\pi\frac{C_2}{C_2} + \frac{2(C_1 - C_3)}{\sqrt{C_3^2 - C_2^2}} \lim_{\theta\to\pi} 2\arctan \frac{\sqrt{C_3^2 - C_2^2}\tan\frac{\theta}{2}}{C_3 + C_2} \\ &= 2\pi - \frac{2\pi\sigma_S^2}{\sqrt{C_3^2 - C_2^2}} \end{aligned}$$

So the distortion is

$$D = \sigma_S^2 \left(1 - \frac{M_0}{L_0}\frac{\sigma_S^2}{\sigma_W^2} \cdot \left(\frac{1}{\left(\frac{\sigma_S^2}{\sigma_W^2}\right)^2 + 2(1+\beta^2)\frac{\sigma_S^2}{\sigma_W^2} + (1-\beta^2)^2} \right)^{1/2} \right) \ .$$

This gives a more complicated relationship between the decay factor β, the oversampling ratio, and the signal to noise ratio σ_S^2/σ_W^2.

The most important class of Toeplitz models comes from modeling the sensor observations as the output of a linear time-invariant (LTI) system driven by the source observations.

2.1.3 Observation models from LTI filters

Having established that the bounded energy conditions in the previous section limit the performance of centralized estimators, we now turn to a situation in which the best centralized estimator may be partially decentralized. The model we choose is one of estimating a spatially distributed source through an LTI filter. A simple physical model for this can be made by assuming the sensors and sources to be located on a line, as shown in Figure 2.3. A diagram of the sampling situation

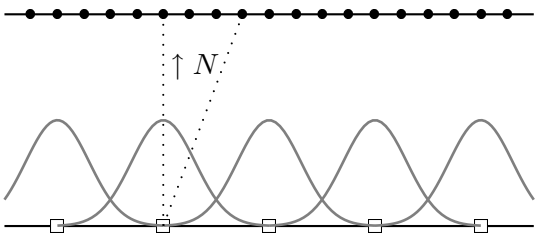

Figure 2.3. An example of a network on a line. Imagine the two lines are actually the same line. The squares mark the locations of the sources, and the circles mark the locations of the sensors. The sensor observations are the superposition of the impulse response of the filter centered at the sensor locations. The upsampling ratio N is shown via the dotted lines.

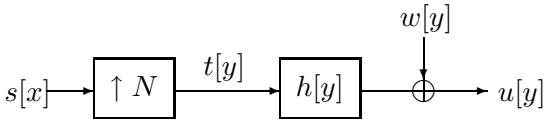

Figure 2.4. Filtering model for remote sensing.

is shown in Figure 2.4. The source is upsampled and filtered by a linear space-invariant filter $h[y]$ and each sensor observes a noisy sample of the upsampled and filtered source.

Let us write the filter as

$$h[y] = \sum_{i=-\infty}^{\infty} h_i \delta[y - i] \ . \tag{2.37}$$

We can write the sequence $\{a_{ml}\}$ in terms of the filter coefficients. We assume that the filter is absolutely summable:

$$\sum_{y=-\infty}^{\infty} |h[y]| = \bar{h} < \infty \ , \tag{2.38}$$

which implies that it is stable and its Fourier transform $H(e^{j\omega})$ exists [30].

The source sequence $s[x]$ is first upsampled by a factor N that corresponds to the ratio M/L in the previous section. This upsampled signal is then filtered by $h[y]$ and noise is added. We can

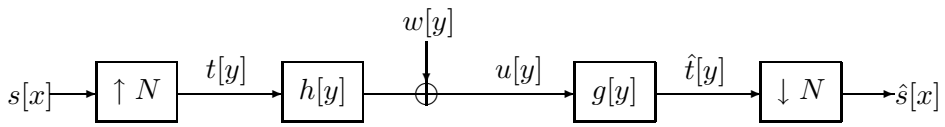

Figure 2.5. Wiener filter solution for spatial filtering.

rewrite the filter as an infinite matrix A (see [35, p. 72]):

$$A = \begin{pmatrix} & \vdots & \vdots & \\ \cdots & h[0] & h[-N] & \cdots \\ \cdots & \vdots & \vdots & \cdots \\ \cdots & h[N-1] & h[-1] & \cdots \\ \cdots & h[N] & h[0] & \cdots \\ & \vdots & \vdots & \end{pmatrix} . \tag{2.39}$$

The matrix AA^T is the autocorrelation matrix of the process $u[y]$. It is Toeplitz and generated by the autocorrelation function $R_u[y] = E[u[y+x]u[x]]$. Therefore the asymptotic distortion is given by equation (2.33) in the previous section. The MMSE estimator for these observations is a linear operator $G(\cdot)$ which, when applied to $u[y]$, yields an estimate $\hat{s}[x]$ that is closest to $u[y]$ in the mean-square sense:

$$G = \underset{G}{\operatorname{argmin}}\, E\left[\|s[x] - Gu[y]\|^2\right] . \tag{2.40}$$

The solution is given in the following proposition.

Proposition 3. *The MMSE estimator for the filtered observation model in Figure 2.4 is a cascade of a non-causal Wiener filter for $t[y]$ followed by downsampling by N.*

Proof. We will show that the proposed system shown in Figure 2.5 is in fact the MMSE estimator. Let $\hat{t}[y]$ be the output of the Wiener filter for $t[y]$ given $u[y]$ and $\hat{s}[x] = \hat{t}[y/N]$ for $y = xN$ and 0 otherwise. By the orthogonality property, the Wiener filter satisfies the following condition:

$$E[(t[y] - \hat{t}[y])\hat{t}[y]] = 0. \tag{2.41}$$

Downsampling will not change this relationship – we can substitute $y = xN$ to get

$$E[(s[x] - \hat{s}[x])\hat{s}[x]] = 0 \ . \tag{2.42}$$

The system in Figure 2.5 is linear and the estimation error is uncorrelated with the original signal. Since all of the signals are jointly Gaussian, the error is in fact independent of the original signal, so this estimator must be the MMSE estimator for $s[x]$. □

To calculate the filter and corresponding estimation error, define the following correlation functions and spectra:

$$\begin{aligned} r_{tu}[y] &= E[t[y+z]u[z]] & (2.43)\\ R_{tu}(e^{j\omega}) &= \sum_{y=-\infty}^{\infty} r_{tu}[y]e^{-j\omega y} & (2.44)\\ r_t[y] &= E[t[y+z]t[z]] & (2.45)\\ R_t(e^{j\omega}) &= \sum_{y=-\infty}^{\infty} r_t[y]e^{-j\omega y}. & (2.46) \end{aligned}$$

The power spectrum of the non-causal Wiener filter is given by

$$G(e^{j\omega}) = \frac{R_{tu}(e^{j\omega})}{R_{tt}(e^{j\omega})}. \tag{2.47}$$

The infinite matrix corresponding to this Wiener filter is the MMSE estimation matrix. The estimation error is given by (2.33) with $B(\omega) = R_t(e^{j\omega})$.

Why do we bother with this filtering perspective? It provides us with a compact description of the estimator (in this case a linear filter) and a means of constructing it that does not rely on multiplying larger and larger matrices. An additional benefit is that Wiener filters can be designed with constraints on the number of nonzero taps. In a sensor network scenario, we interpret this as a constraint on the number of sensors that can collaborate. Specifically, we can constrain the

estimation matrix G to be 0 outside some set of diagonals:

$$G = \begin{pmatrix} & \vdots & \vdots & & & & & \\ \cdots & g_{1,1} & g_{1,2} & \cdots & g_{1,K} & 0 & 0 & \cdots \\ \cdots & g_{2,1} & g_{2,2} & \cdots & g_{1,K} & g_{2,K+1} & 0 & \cdots \\ \cdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots \\ \cdots & g_{K,1} & g_{K,2} & \cdots & g_{K,K} & g_{K,K+1} & g_{K,K+2} & \cdots \\ \cdots & 0 & g_{K+1,2} & \cdots & g_{K+1,K} & g_{K+1,K+1} & g_{K+1,K+2} & \cdots \\ \cdots & 0 & 0 & \cdots & g_{K+2,K} & g_{K+2,K+1} & g_{K+2,K+2} & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \end{pmatrix} . \tag{2.48}$$

Here we have constrained the matrix to operate on at most K steps in either direction.

The solution in this case is relatively simple [22, p. 102]. Define the vectors and matrices

$$\mathbf{u}_y = (u[y-K], u[y-K+1], \ldots, u[y+K])^T \tag{2.49}$$

$$\mathbf{g}_y = (g[y-K], g[y-K+1], \ldots, g[y+K])^T \tag{2.50}$$

$$\mathbf{p}_y = E[\mathbf{u}[y]t[y]] \tag{2.51}$$

$$\mathbf{R}_y = E[\mathbf{u}_y \mathbf{u}_y^H] \tag{2.52}$$

The solution is then:

$$\mathbf{g}_y = \mathbf{R}_y^{-1} \mathbf{p}_y \ . \tag{2.53}$$

We follow it by the downsampling operation as before. Because all of the processes are wide-sense stationary, this distributed solution can be computed offline.

The previous analysis highlights an important problem in trying to make models more accurate. By enforcing a realistic collaboration constraint, we appear to still have a clean and easy-to-compute optimal partially-decentralized solution for the estimation problem. Unfortunately, real sensors are unlikely to be positioned evenly on a line, so the filter will not be space-invariant and the observations will not be wide-sense stationary by index.

2.1.4 Random matrices with iid entries

Let us now suppose that the entries of the observation matrix are independent and identically distributed random variables. One of the deeper results of random matrix theory is that the eigenvalue distribution of large random matrices is the same regardless of the distribution chosen for the entries. Therefore the results we can obtain for random matrix models of sensor observations rely only on the assumption of independent entries. The geometric view of the network here is that the sources and sensors are both located in the same area but *a priori* very little is known about the coefficient between any particular source and sensor. However, once the network is deployed, the relevant coefficients can be measured or approximated. What we wish to know is the probability of achieving a certain asymptotic distortion.

The primary tool we will use is the Marčenko-Pastur law [25] as it is cited in [2, pp. 620-621]. We reproduce it here for completeness:

Theorem 1. *Suppose that $p/n \to y \in (0, \infty)$. If the entries of an $p \times n$ complex matrix A are iid with mean 0 and variance σ^2 and $B = \frac{1}{n}AA^H$, then the empirical spectral distribution (ESD) of B tends (as $p, n \to \infty$) to a limiting distribution with density*

$$p_y(x) = \begin{cases} \frac{1}{2\pi xy\sigma^2}\sqrt{(b-x)(x-a)}, & \text{if } a \le x \le b \\ 0, & \text{otherwise,} \end{cases} \tag{2.54}$$

and a point mass $1 - 1/y$ at the origin if $y > 1$ where $a = a(y) = \sigma^2(1 - \sqrt{y})^2$ and $b = b(y) = \sigma^2(1 + \sqrt{y})^2$.

The important feature of distribution is that at almost all of the nonzero normalized singular values of A will lie in a region bounded away from the origin, so unnormalized singular values scale to infinity like n. This in turn will force the distortion to 0 for our problem.

Proposition 4. *Suppose that $\{a_{ml}\}$ is an array of iid random variables with mean 0 and variance σ^2. Then the asymptotic distortion converges to 0 as $M \to \infty$.*

Proof. Because the singular values of A and A^T are identical, the distribution of the singular values converges to the Marčenko-Pastur law with $y = L_0/M_0 < 1$. All of the eigenvalues in the asymptotic

limit are greater than $a(y)n = n\sigma^2(1 - \sqrt{L_0/M_0})^2$. Thus the distortion is:

$$\begin{aligned} D &= \lim_{n\to\infty} \frac{1}{L_0 n} \sum_{i=1}^{L_0 n} \frac{\sigma_S^2 \sigma_W^2}{\alpha_i^2 \sigma_S^2 + \sigma_W^2} \\ &\leq \lim_{n\to\infty} \frac{1}{L_0 n} \sum_{i=1}^{L_0 n} \frac{\sigma_S^2 \sigma_W^2}{a(y) n \sigma_S^2 + \sigma_W^2} \\ &= \lim_{n\to\infty} \frac{\sigma_S^2 \sigma_W^2}{L_0 (a(y) n \sigma_S^2 + \sigma_W^2)} \\ &= 0 \ . \end{aligned}$$

Thus the distortion converges to 0 □

Indeed, the smallest eigenvalue converges to $a(y)$ almost surely [2, p. 635], so the convergence is even stronger. This result is not surprising in view of the results of the first section, which stated that bounded row and column sum variances are sufficient to force the distortion to a non-zero value. The previous proposition has unrealistic assumptions for many practical systems – the received energy is unbounded in expectation and the observation gains must be iid. However, since the Marčenko-Pastur law is the limiting distribution for all distributions with zero mean and variance σ^2, we can relax the condition on identical distribution. The sufficient condition on the collection $\{a_{ml}\}$ of independent random variables with mean zero and common variance σ^2 is given by [2, p. 623]. Suppose that for any $\delta > 0$,

$$\frac{1}{\delta L_n M_n} \sum_{l=0}^{L_n} \sum_{m=0}^{M_n} E\left[|a_{ml}|^2 \mathbf{1}_{(|a_{ml}| > \delta\sqrt{n})} \right] \to 0 \ . \tag{2.55}$$

If this condition is satisfied, the singular values again converge to the Marčenko-Pastur law and the distortion will again converge to 0. What these results suggest is that if the observations have bounded energy, the distortion is finite and positive, but in the "average case" of unbounded energy, the distortion will converge to 0.

2.2 Estimation with fading observations

While estimation from known matrices is an interesting topic, a more accurate model of the observations may come from the fading observations framework mentioned in the first chapter. Here

the matrix A takes values in in a set $\mathcal{A}$ and we must design an estimator that is robust across this class. In slow fading, A is chosen once and fixed for all time, and in fast fading A is time-varying. We will take up these two cases in turn. Slow fading turns out to be equivalent to the analysis in the previous section because the matrix A can (within reason) be estimated from the observations themselves. The fast fading case is more difficult and we restrict our analysis to linear estimators.

2.2.1 Slow fading

Suppose that $\mathcal{A} = \{A_1, A_2, \dots A_K\}$ and that A is chosen from $\mathcal{A}$. Let $\Delta(A_j|A_k)$ be the mismatch error for a memoryless estimator for A_j when $A = A_k$:

$$\Delta(A_j|A_k) = E\left[\frac{1}{L}\left\|S - \sigma_S^2 A_j^T(\sigma_S^2 A_j A_j^T + \sigma_W^2 I)^{-1}(A_k S + W)\right\|^2\right] . \tag{2.56}$$

We break the estimation into two parts: we first find A from the sequence of observations $U[1], U[2], \dots U[n]$ and then use the MMSE estimator assuming those are the true statistics. The estimation problem can be seen as a hypothesis test between the different candidates in $\mathcal{A}$. For simplicity, we will assume a uniform prior on the set $\mathcal{A}$.

A crucial property that we will need is that each $A_j \in \mathcal{A}$ induces a different joint distribution for the observations U. As an example, suppose $+A \in \mathcal{A}$ and $-A \in \mathcal{A}$. Then the statistical properties of the observations are identical under both hypotheses and there will be no way to tell them apart. Consequently, an estimator built for $+A$ will make $\Delta(-A \mid +A)$ very large. In order to avoid these complications, we will always assume that $\mathcal{A}$ is separable in the sense that $p(U|A_j) \neq p(U|A_k)$ for $j \neq k$.

Suppose that $K = 2$ so that we have a binary hypothesis test. We can bound the probability of error in our hypothesis test following [10, Sec. 3.4]. Let

$$\hat{T}_n = \frac{1}{n}\sum_{j=1}^{n} \log \frac{p_U(U[j]|A = A_1)}{p_U(U[j]|A = A_2)} \tag{2.57}$$

be the normalized observed log likelihood ratio. A well-known result states that the best estimate $\hat{A}$ can be found by comparing $\hat{T}_n$ to a threshold.

Lemma 1. *In the slow fading case with discrete $\mathcal{A}$, if $A = A_k$ the distortion converges to the MMSE distortion as the sample size goes to ∞:*

$$\lim_{n\to\infty} \Delta(\hat{A}(\{U[j] : j = 1, 2, \dots n\})|A_k) \to \Delta(A_k|A_k) \tag{2.58}$$

Proof. Let $\alpha = P(\hat{A} \neq A_k)$. Then

$$\Delta(\{U[j] : j = 1, 2, \dots n\}|A_k) = (1-\alpha)\Delta(A_k|A_k) + \alpha\Delta(A_j|A_k) \tag{2.59}$$

By Corollary 3.4.6 in [10], the probability of error satisfies a large deviations bound so that $\alpha \leq \exp(-n\beta)$ for some constant $\beta > 0$. Since $\Delta(A_j|A_k)$ is bounded, taking the limit on both sides completes the proof. □

This clearly extends to finite K. For centralized estimators, this type of slow fading is uninteresting because it can be disambiguated with exponentially small probability of error. Let us now consider the case where $\mathcal{A}$ is not finite, so that a simple hypothesis test may no longer be sufficient. A simple approach is to compute the sample covariance matrix:

$$\hat{\Sigma}_U = \frac{1}{n}\sum_{j=1}^{n} U[j]U[j]^T \ . \tag{2.60}$$

Unfortunately, the sample covariance is very sensitive to outliers in the data [1]. Several methods for *robust covariance estimation* have been proposed in the statistics literature [1], [5], [26], [40], and depending on the nature of the set $\mathcal{A}$, some will be better than others. For example, if $\mathcal{A}$ is very structured and constrained, the EM approach of [5] may be effective. The worst-case distortion can then be expressed as:

$$D_{max} = \lim_{n\to\infty} \sup_{B\in\mathcal{A}} D(\hat{A}(\{U[j] : j = 1, 2, \dots n\})|B) \ . \tag{2.61}$$

Although slow fading for centralized estimators seems straightforward, estimating A from the marginals at each sensor may prove to be impossible, as we will see in Chapter 4. Similarly, if the sensors are limited in their ability to communicate with each other, computationally intensive covariance estimation procedures that require significant inter-sensor communication may be infeasible.

2.2.2 Fast fading

We now turn to fast fading, in which the observation matrix $A[n]$ varies over time. We will assume that $A[n]$ is iid with some distribution $p_A(\cdot)$ on a finite set $\mathcal{A} = \{A_1, A_2, \ldots, A_K\}$. If $A[n]$ was known to the estimator, it would simply use the MMSE estimator for each A_k:

$$D_{opt} = \sum_{k=1}^{K} p_A(A_k)\Delta(A_k|A_k) \ . \tag{2.62}$$

However, our interest is in the case where $A[n]$ is unknown.

Suppose first that $p_A(\cdot)$ is not known to the estimator. Following the previous section, can we estimate $p_A(\cdot)$? The answer is yes, as long as K is not too large. Consider the average covariance matrix of the observations:

$$\bar{\Sigma}_U = \sum_{k=1}^{K} p_A(A_k)(\sigma_S^2 A_k A_k^T + \sigma_W^2 I) \tag{2.63}$$

The sample covariance matrix of the observation vectors will converge to $\bar{\Sigma}_U$, with the same caveats about outliers as before. Alternatively, we can use the robust methods mentioned earlier to estimate $\bar{\Sigma}_U$. Since $\bar{\Sigma}_U$ is just a linear matrix-valued function with coefficients $\{p_A(A_k)\}$, we can estimate $\{p_A(A_k)\}$ from $\bar{\Sigma}_U$ as long as $K < (M^2 + M)/2$, the dimension of the set of possible covariance matrices. Clearly there are many convergence as well as numerical stability issues in performing this estimation.

Let us instead look at the case where we are forced to choose a fixed memoryless estimation matrix G. For a particular distribution $p_A(\cdot)$ and G we have an optimization problem over the error functions

$$D(p_A, G) = \frac{1}{L} E\left[(S - \hat{S})^T (S - \hat{S})\right] = \frac{1}{L} E\left[(S - G(AS + W))^T (S - G(AS + W))\right] \ , \tag{2.64}$$

where the expectation is taken over A, S, and W.

We can view this as a game in which one player chooses a matrix G to minimize the distortion and the other chooses a distribution p_A to maximize the distortion. Suppose that p_A is known. Then the worst case distortion for this estimator is given by the solution to the following optimization

problem:

$$\sup_{p_A} \inf_{G} D(p_A, G) \ . \tag{2.65}$$

In the case where p_A is unknown, the first player must choose G and reveal that choice to the second player. The relevant quantity is

$$\inf_{G} \sup_{p_A} D(p_A, G) \ . \tag{2.66}$$

The interpretation of this is that for each choice of G there is a "worst-case" distribution p_A, and that we will choose the G that induces the smallest average distortion for the worst-case p_A.

Consider first the case where A is chosen iid across time according to a distribution p_A that is known to the estimator, leading to (2.65). Then

$$\begin{aligned} D(p_A, G) &= \frac{1}{L} \operatorname{tr}\left[\sigma_S^2 I - 2\sigma_S^2 G\bar{A} + G(\sigma_S^2 \Sigma_A + \sigma_W^2) G^T\right] \\ &= \frac{1}{L}\left(L\sigma_S^2 - 2\sigma_S^2 \operatorname{tr}(G\bar{A}) + \sigma_S^2 \operatorname{tr}(G^T G \Sigma_A) + \sigma_W^2 \operatorname{tr}(G^T G)\right) \ , \end{aligned} \tag{2.67}$$

where $\bar{A} = E[A]$ and $\Sigma_A = E[AA^T]$. We need to minimize this over G. Let us first consider the scalar case where $L = M = 1$.

$$\frac{\partial}{\partial G} D(p_A, G) = -2\sigma_S^2 \bar{A} + 2\sigma_S^2 \Sigma_A G + 2\sigma_W^2 G \tag{2.68}$$

$$G = \frac{\bar{A}\sigma_S^2}{\Sigma_A \sigma_S^2 + \sigma_W^2} \ . \tag{2.69}$$

Unfortunately, this analysis does not easily extend to the matrix case. The problem is that $\bar{A}\bar{A}^T \neq \Sigma_A$ in general, so the trick of taking the singular value decomposition used before is no longer valid.

Suppose we take the orthogonality condition for least-squares estimation:

$$\operatorname{tr}\left(E\left[(S - GU)U^T G^T\right]\right) = 0 \tag{2.70}$$

This gives:

$$\operatorname{tr}(\sigma_S^2 \bar{A}^T G^T) = \operatorname{tr}(G(\sigma_S^2 \Sigma_A + \sigma_W^2 I) G^T) \tag{2.71}$$

If we just "guess"

$$G = \sigma_S^2 \bar{A}^T (\sigma_S^2 \Sigma_A + \sigma_W^2 I)^{-1} \ , \tag{2.72}$$

then (perhaps unsurprisingly) we get equality in (2.71). Thus the best estimator is given by (2.72).

Our estimation error is therefore given by

$$\inf_G D(p_A, G) = \frac{1}{L}\left(L\sigma_S^2 - 2\sigma_S^2 \operatorname{tr}\left(\bar{A}^T(\sigma_S^2\Sigma_A + \sigma_W^2 I)^{-1}\bar{A}\right) + \sigma_S^2 \operatorname{tr}\left(\bar{A}^T(\sigma_S^2\Sigma_A + \sigma_W^2 I)^{-1}\bar{A}\right)\right) \quad (2.73)$$

$$= \sigma_S^2 - \frac{1}{L}\sigma_S^2 \operatorname{tr}\left(\bar{A}^T(\sigma_S^2\Sigma_A + \sigma_W^2 I)^{-1}\bar{A}\right) \quad (2.74)$$

We can immediately deduce some interesting consequences of this result. Firstly, if the convex closure of $\mathcal{A}$ contains the 0 matrix, the distortion can be forced to $L\sigma_S^2$ by choosing the p_A that makes $\bar{A} = 0$. Secondly, since our estimator is only a function of the statistics of the A process, it can be constructed "on the fly" based on the sensor observations.

We now turn to the reversed scenario, where a linear estimator must be chosen offline, and then the worst-case p_A is selected, leading to (2.66). Let us first consider the scalar version of the problem. The distortion function is given by (2.67):

$$D(p_A, G) = E\left[S^2 - 2GAS^2 + G^2(A^2S^2 + W^2)\right] \quad (2.75)$$

$$= \sigma_S^2 - 2G\mu_A\sigma_S^2 + G^2(\sigma_A^2\sigma_S^2 + \sigma_W^2)\ , \quad (2.76)$$

where μ_A and σ_A are the mean and variance of A. Note that this distortion only depends on these parameters of A. For any choice of G, we can evaluate $D(A_i, G)$ for every choice of A_i. The worst case p_A will concentrate all its mass on the A_i's that maximize the distortion.

The vector case is no different – for any linear estimator G, one or more of the possible observation matrix values A will maximize the distortion. Thus the set of possible G's is partitioned by this "worst-case" function. Let $\Pi_j = \{G : D(A_j, G) = \min_{A_i} D(A_i, G)\}$. We can choose the best G via the following:

$$\inf_G \sup_{p_A} D(p_A, G) = \min\left\{\inf_{G\in\Pi_j} D(A_j, G) : A_j \in \mathcal{A}\right\} \quad (2.77)$$

Unfortunately, we cannot at this time come up with a nice characterization of this problem under general conditions, so we will close with a scalar example.

Example : spike source

Let us take the pathological example for which $\mathcal{A} = \{1, 1000\}$ and $\sigma_S^2 = \sigma_W^2 = 1$. We partition the set of possible estimators G into those for which $A = 1$ is worst and those for which $A = 1000$ is worst. The dividing point will be when they are equal:

$$\sigma_S^2 - 2G\sigma_S^2 + G^2(\sigma_S^2 + \sigma_W^2) = \sigma_S^2 - 2000G\sigma_S^2 + G^2(10^6\sigma_S^2 + \sigma_W^2) \tag{2.78}$$

$$G = \frac{2(10^3 - 1)}{10^6 - 1} = 1.998 \ \times \ 10^{-3} \ . \tag{2.79}$$

So choosing G to be at this dividing point will lead to the lowest worst-case distortion:

$$D = \sigma_S^2 - 2\sigma_S^2 G + (\sigma_S^2 + \sigma_W^2)G^2 \approx 0.996 \ , \tag{2.80}$$

which is not much better than "guessing." A slightly less pathological example may be to take $\mathcal{A} = \{1, 2\}$:

$$G = \frac{2(2 - 1)}{4 - 1} = \frac{2}{3} \tag{2.81}$$

$$D \approx 0.556 \tag{2.82}$$

For $A = 1$ the optimal distortion is 0.5, so the loss is a little over 11%.

2.3 Our toy example

We now return to the canonical example with which we ended the previous chapter. The first step in our analysis will be to find the spectrum of the observation matrix:

$$AA^T = \begin{bmatrix} B_1 & 0 \\ 0 & B_2 \end{bmatrix} \ , \tag{2.83}$$

where B_i is the number of sensors observing source i. A centralized estimator could compute these numbers approximately by taking a very large sample covariance matrix. The best linear estimator based on knowledge of the B_i would be given by (2.62)

$$D = \frac{1}{2}\frac{\sigma_S^2\sigma_W^2}{B_1\sigma_S^2 + \sigma_W^2} + \frac{1}{2}\frac{\sigma_S^2\sigma_W^2}{B_2\sigma_S^2 + \sigma_W^2} \ . \tag{2.84}$$

However, this presupposes that the estimator can determine the true matrix A. To do this it needs significantly more than just the numbers B_1 and B_2. A naive and computationally crippling computation could allow the estimator to perform pairwise tests to divide the sensors into two groups. The error in these tests is exponentially small in the number of samples, so the sorting would be accurate for any finite M. Since the focus of this thesis is not on computational feasibility, we leave this as an open problem and assume that the true matrix A can indeed be estimated.

The asymptotics of this problem are relatively uninteresting – if each sensor is equally likely to observe S_1 as S_2, then both B_1 and B_2 will converge to $M/2$ as $M \to \infty$. In scaling law parlance we would say that the total distortion goes to 0 as $1/M$. The reason for this bland analysis is that the estimator is both data-dependent and centralized. As we shall see, the problem becomes significantly more complicated once distributed computation and communication enter into the picture.

Chapter 3

Rondo: source coding with uncertainty

In this chapter we will look at lossless distributed source reconstruction using the tools of information theory. Observation uncertainty for this problem takes the form of an unknown joint distribution for the sources. In what follows we assume the reader is familiar with the basics of information theory as described in [6], for example. In the next section we will describe the Slepian-Wolf problem and its extension to uncertain joint distributions. This does not provide any difficulty in the proof – any rates that are achievable for all sources in the class are achievable using a modified Slepian-Wolf code. The main contribution of this chapter comes in section 5, where we look at a source coding system in which the blocklength and number of sensors increases simultaneously. By linking the two we can bound how fast the blocklength must grow in order to accommodate more sensors for a fixed error probability.

Information theory views the problem as one of encoding the sensor observations into a discrete set of indices. Given a sequence of source observations U^n, we create an encoding map $\mathcal{U}^n \to \{1, 2, \ldots, N\}$ and a decoding map $\{1, 2, \ldots, N\} \to \hat{\mathcal{S}}^n$ in order to minimize some some distortion function $d(s^n, \hat{s}^n)]$. In the noiseless case, the sensor observations U^n are simply the source values

S^n. For lossless source coding, the distortion measure is error probability, so that $\hat{\mathcal{S}} = \mathcal{S}$ and

$$d(s^n, \hat{s}^n) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(s_i \neq \hat{s}^i) \tag{3.1}$$

The objective to create a sequence of codes, indexed by n, such that $P_e = E[d(s^n, \hat{s}^n)] \to 0$ as $n \to \infty$. Thus in the limit, the reconstruction is "perfect." In order to do this, we will upper bound the block error $\mathbf{1}(s^n \neq \hat{s}^n)$.

3.1 Slepian-Wolf coding over a class of distributions

In this section we will discuss observation uncertainty in lossless source coding problems. The type of observation uncertainty that we will address is that of *distributional uncertainty*. In the point-to-point case this leads to the well-addressed problem of constructing universal source codes. In the multi-terminal case the famous theorem of Slepian and Wolf provides the springboard for coding schemes that can handle multiple distributions. The idea is shown in Figure 3.1. We will look at how the number of terminals and hence modeling complexity relate to the blocklength, which will motivate our analysis in the next section.

The source-coding community has already addressed uncertainty in the underlying source distribution. For the point-to-point or centralized problem, many *universal* source codes exist for discrete S, using the Lempel-Ziv algorithm [42], context-tree weighting [41], or the Burrows-Wheeler Transform [4], [15], for example. These source codes will compress the source S at a rate that converges to the true entropy $H(S)$ without knowing a priori what that entropy is. In one very simple universal source code, the encoder observes a block of n symbols s^n and transmits the *type* $T(s^n)$ of the block along with a compressed version using a compression algorithm that assumes the block is distributed according to $T(s^n)$. The overhead for transmitting the type is negligible compared to the blocklength, and the rate required by this scheme also converges $H(S)$. This universal scheme is more in the spirit of the multi-terminal problem we will investigate next.

Unfortunately, universal schemes are difficult to extend to multiuser settings. Consider the problem of two terminals observing correlated sources S_1 and S_2. If the statistics are known to the

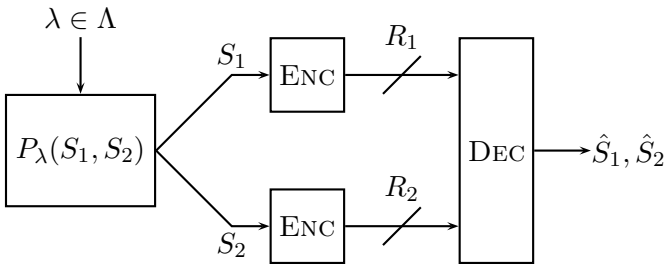

Figure 3.1. Source coding for a class of sources.

encoders and decoder, then the set of rates at which the sources can be communicated losslessly to a destination was found by Slepian and Wolf [34]. An exercise in Csiszár and Körner [8, Exercise 3.1.6] (more generally in [7, Theorem 2]) gives an example of how to construct a single code for all correlated sources that achieves a certain error exponent. To obtain that exponent, the encoder is obligated to raise the rate at which it operates to that of a worst-case source. This is in contrast to the point-to-point schemes, where the coding algorithms converge to the minimum rate needed for the particular source.

Another coding scheme that can achieve these worst-cast points is the sequential binning scheme proposed by Draper, Chang and Sahai [13]. In their framework, unknown statistics are not a problem as long as the target rate pair is in the achievable region for the source, and they provide error exponents for their scheme that can in some cases are better than those for the corresponding block codes. Baron, Khojastepour, and Baraniuk [3] examined the notion of redundancy rates for fixed block-length coding, which captures the excess rate needed to account for distributed coding in the non-asymptotic regime. However, in order to analyze the universality of their scheme over different distributions, they adopt the linked-encoder framework of Oohama [27]. Other strategies to gain universality [12], [24] propose a feedback link from the decoder.

However, in some instances the encoder and decoder may have some limited knowledge of the joint distribution of the sources, and a code in the style of Csiszár and Körner is more reasonable. We will show that it is possible to lower the rates required to the worst source in the class. The code

in [7] is non-constructive and uses a minimum entropy decoder. Here we give an explicit binning construction in the style of [6] using a decoder that looks for jointly typical sequences. For the sake of completeness, we include some standard definitions first.

Definition 1. *A discrete memoryless correlated source is a tuple of random variables* $\mathbf{S} = (S_1, S_2, \ldots, S_m)$ *with variable* S_j *taking values in a finite set* $\mathcal{S}_j$*. The variables are jointly distributed with some distribution* $P(\mathbf{S})$ *independently and identically across time. A class of sources is a collection of joint distributions* $P_\lambda(\mathbf{S})$ *indexed by* $\lambda \in \Lambda$*. Entropies calculated under* P_λ *are also given subscript* λ*, e.g.* $H_\lambda(S_1, S_2)$*,* $H_\lambda(S_1|S_2)$*, etc.*

Definition 2. *A* (n, R_1, R_2) *distributed source code for a class of sources* (S, T) *is a tuple* (ϕ_1, ϕ_2, ψ) *of maps with*

$$\phi_1 : \mathcal{S}^n \to \{1, 2, \ldots 2^{nR_1}\} \tag{3.2}$$

$$\phi_2 : \mathcal{T}^n \to \{1, 2, \ldots 2^{nR_2}\} \tag{3.3}$$

$$\psi : \{1, 2, \ldots 2^{nR_1}\} \times \{1, 2, \ldots 2^{nR_2}\} \to \mathcal{S}^n \times \mathcal{T}^n \ . \tag{3.4}$$

The probability of error for this code under distribution $P_\lambda(S, T)$ *for* $\lambda \in \Lambda$ *is*

$$P_{e,\lambda}^{(n)} = P_\lambda(\psi(\phi_1(S^n), \phi_2(T^n)) \neq (S^n, T^n)) \tag{3.5}$$

Definition 3. *A rate pair* (R_1, R_2) *is achievable for a class of sources* $\{P_\lambda : \lambda \in \Lambda\}$ *if there exists a sequence of* (n, R_1, R_2) *distributed source codes such that* $P_{e,\lambda}^{(n)} \to 0$ *for all* $\lambda \in \Lambda$*. The achievable rate region is the closure of the set of achievable rates.*

Given any Slepian-Wolf rate region $\mathcal{R}$, we can show that all sources whose rate regions lie within $\mathcal{R}$ are achievable using a single code.

Proposition 5. *Let* α_1*,* α_2*, and* α_3 *be positive real constants. Consider the class of sources* $\{P_\lambda : \lambda \in \Lambda\}$ *for the random variables* (S_1, S_2) *that satisfy:*

$$\alpha_1 > H_\lambda(S_1|S_2) \tag{3.6}$$

$$\alpha_2 > H_\lambda(S_2|S_1) \tag{3.7}$$

$$\alpha_3 > H_\lambda(S_1, S_2) \tag{3.8}$$

Then the set of achievable rates is given by $\{(R_1, R_2) : R_1 \geq \alpha_1, R_2 \geq \alpha_2, R_1 + R_2 \geq \alpha_3)$.

Proof. The converse is simple. Fix $\epsilon > 0$ suppose rate $R_1 = \alpha_1 - \epsilon$ was in the achievable rate region. Then it must be in the rate region for all sources in the class. But there exists some source in the class such that $H(S_1|S_2) = \alpha_1 - \epsilon/2$, so R_1 is not an achievable rate for this source, which is a contradiction. Identical arguments can be made for the other inequalities.

The proof of achievability is nearly identical to that in [6, §14.4.1], with a slight modification. The key fact is that for a fixed block length n, there are only a polynomial number of types, so the number of typical sequences over all sources in the class is dominated by the source with the largest entropy in the class. The rate penalty for using polynomially more bins in the Slepian-Wolf code goes to zero with the blocklength.

Let (R_1, R_2) be an achievable rate pair. We will show that the pair $(R_1+\delta, R_2+\delta)$ is achievable using binning for any $\delta > 0$. Fix a block length n.

1. Assign to each sequence $s_1^n \in {\mathcal{S}_\infty}^n$ an index in $\{1, 2, \ldots 2 \cdot 2^{nR_1}\}$, chosen uniformly. Assign to each sequence $s_2^n \in {\mathcal{S}_\in}^n$ an index in $\{1, 2, \ldots 2 \cdot 2^{nR_2}\}$, also chosen uniformly. These are our encoders ϕ_1 and ϕ_2.

2. The messages that the users send are the bin indices of their respective source sequences. Let the messages be m_1 and m_2.

3. Decode (s_1^n, s_2^n) if $\phi(s_1^n) = m_1$, $\phi(s_2^n) = m_2$, and $(s_1^n, s_2^n) \in A_{\epsilon,\lambda}^{(n)}$, the ϵ-typical sets under P_λ.

We must now bound the probability of error. For a fixed P_λ, the coding scheme above has sufficient rate by the Slepian-Wolf theorem. The new error events center around what happens when a pair (s_1^n, s_2^n) that is jointly-typical with respect to P_λ is instead decoded as a pair that is jointly typical with respect to P_μ. In what follows, we assume that (S_1^n, S_2^n) is chosen according to P_λ and that $\lambda \neq \mu$. The errors are then:

1. There is a $\hat{s}_1^n$ such that $\phi_1(\hat{s}_1^n) = \phi_1(S_1^n)$ and $(\hat{s}_1^n, S_2^n) \in A_{\epsilon,\mu}^{(n)}$ for some μ,

2. There is a $\hat{s}_2^n$ such that $\phi_2(\hat{s}_2^n) = \phi_2(S_2^n)$ and $(S_1^n, \hat{s}_2^n) \in A_{\epsilon,\mu}^{(n)}$ for some μ,

3. There is a pair $(\hat{s}_1^n, \hat{s}_2^n) \neq (S_1^n, S_2^n)$ such that $\phi_1(\hat{s}_1^n) = \phi_1(S_1^n)$, $\phi_2(\hat{s}_2^n) = \phi_2(S_2^n)$, and $(\hat{s}_1^n, \hat{s}_2^n) \in A_{\epsilon,\mu}^{(n)}$ for some μ.

To bound these events 1–3 we must turn to our type arguments. Let $\mathcal{P}_n$ denote all types of denominator n. There are at most $(n+1)^{|\mathcal{S}_1|+|\mathcal{S}_2|}$ such types. The number of jointly typical sequences across all classes is then

$$|A_{\epsilon,\Lambda}^{(n)}| = \left| \bigcup_{\mu \in \mathcal{P}_n \cap \Lambda} A_{\epsilon,\mu}^{(n)} \right|$$

The size of the jointly typical sets is then:

$$|A_{\epsilon,\Lambda}^{(n)}(S_1, S_2)| \leq \sum_{\mu \in \mathcal{P}_n \cap \Lambda} 2^{n(H_\mu(S_1,S_2)+\epsilon)} \tag{3.9}$$

$$\leq (n+1)^{|\mathcal{S}_1|+|\mathcal{S}_2|} \left(\sup_{\mu \in \mathcal{P}_n \cap \Lambda} 2^{n(H_\mu(S_1,S_2)+\epsilon)} \right) \tag{3.10}$$

$$\leq \sup_{\mu \in \mathcal{P}_n \cap \Lambda} 2^{n(H_\mu(S_1,S_2)+\epsilon+n^{-1}(|\mathcal{S}_1|+|\mathcal{S}_2|)\log(n+1))} \tag{3.11}$$

$$\leq \sup_{\mu \in \mathcal{P}_n \cap \Lambda} 2^{n(H_\mu(S_1,S_2)+\epsilon+\delta_n)} \tag{3.12}$$

where $\delta_n \to 0$ as $n \to \infty$. Similarly,

$$|A_{\epsilon,\Lambda}^{(n)}(S_1|S_2)| \leq \sup_{\mu \in \mathcal{P}_n \cap \Lambda} 2^{n(H_\mu(S_1|S_2)+\epsilon+\delta_n)} |A_{\epsilon,\Lambda}^{(n)}(S_2|S_1)| \quad \leq \sup_{\mu \in \mathcal{P}_n \cap \Lambda} 2^{n(H_\mu(S_2|S_1)+\epsilon+\delta_n)} . \tag{3.13}$$

For any fixed λ, the probability of the decoding errors above are:

$$\begin{aligned}
P_{e,1} &\leq \sum_{(s_1^n, s_2^n)} P_\lambda(s_1^n, s_2^n) \sum_{\mu \in \mathcal{P}_n \cap \Lambda} \sum_{\hat{s}_1^n \neq s_1^n,\ (\hat{s}_1^n, s_2^n) \in A_{\epsilon,\mu}^{(n)}} P(\phi_1(\hat{s}_1^n) = \phi_1(s_1^n)) \\
&= \sum_{(s_1^n, s_2^n)} P_\lambda(s_1^n, s_2^n) 2^{-nR_1} |A_{\epsilon,\Lambda}^{(n)}(S_1|s_2^n)| \\
&\leq 2^{-nR_1} \sup_{\mu \in \mathcal{P}_n \cap \Lambda} 2^{n(H_\mu(S_1|S_2)+\epsilon+\delta_n)} \\
P_{e,2} &\leq 2^{-nR_2} \sup_{\mu \in \mathcal{P}_n \cap \Lambda} 2^{n(H_\mu(S_2|S_1)+\epsilon+\delta_n)}
\end{aligned}$$

$$P_{e,3} \leq 2^{-n(R_1+R_2)} \sup_{\mu \in \mathcal{P}_n \cap \Lambda} 2^{n(H_\mu(S_1,S_2)+\epsilon+\delta_n)}$$

Since $\mathcal{P}_n$ is dense in Λ, as $n \to \infty$ we have decoding errors of the types above if

$$R_1 < \sup_{\mu \in \Lambda} H_\mu(S_1|S_2)$$

$$R_2 < \sup_{\mu \in \Lambda} H_\mu(S_2|S_1)$$

$$R_1 + R_2 < \sup_{\mu \in \Lambda} H_\mu(S_1, S_2) \ .$$

This establishes the result. □

The theorem says that there exists a Slepian-Wolf code that achieves all rates common to the rate regions of all sources in the class. In some cases the class may be small and all sources have the same rate region, so the rates are tight. As an example, consider a memoryless correlated source (S_1, S_2), with each component taking values in the set $\{-1, 1\}$ with distribution Bernoulli(1/2). However, their product $S_1 S_2$ may have distribution Bernoulli(α) or Bernoulli($1-\alpha$) for some fixed α. Call these two joint distributions P_α and P_β respectively. Note that $H_b(\alpha) = H_b(1-\alpha)$, where $H_b(\cdot)$ is the binary entropy function. The Slepian-Wolf region is the set of rates (R_1, R_2) such that

$$R_1 > H_b(\alpha) \tag{3.14}$$

$$R_2 > H_b(\alpha) \tag{3.15}$$

$$R_1 + R_2 > 1 + H_b(\alpha) \ , \tag{3.16}$$

where $H_b(\cdot)$ denotes the binary entropy function.

For a correlated source with many components, a similar result can be shown to Proposition 5. We omit the proof as it is nearly identical to that shown above.

Proposition 6. *Let $\{\alpha_\sigma : \sigma \subseteq \{1, 2, \ldots m\}\}$ be a set of positive real numbers. Let $\mathbf{S}(\sigma) = \{S_j : j \in \sigma\}$. Consider the class of sources $\{P_\lambda : \lambda \in \Lambda\}$ for the random variables $(S_1, S_2, \ldots, S_m)$ that satisfy:*

$$\alpha_\sigma > H(\mathbf{S}(\sigma)|\mathbf{S}(\sigma^c)) \tag{3.17}$$

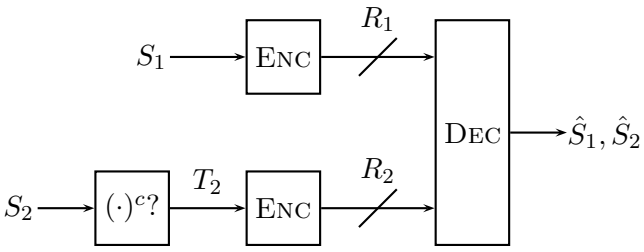

Figure 3.2. Binary correlated source with one remote component. Conditioned on the first component, the decoder can figure out if the second one was complemented or not.

Then the set of achievable rates is given by $\{(R_1, R_2, \ldots, R_k) : R(\sigma) \geq \alpha_\sigma)\ \forall \sigma\}$*, where*

$$R(\sigma) = \sum_{j \in \sigma} R_j \ . \tag{3.18}$$

3.2 Multi-terminal coding with many terminals

Consider a slight variation of the previous example, as shown in Figure 3.2. The source (S_1, S_2) have the same marginals as before, but the product $S_1 S_2$ is Bernoulli(α). The source S_1 is viewed directly by the first terminal, but the second source S_2 is viewed remotely through a channel may or may not flip all of the bits, forming the observed sequence T_2. This induces the same class of sources on (S_1, T_2) as considered above. Using the same coding scheme, we can reconstruct S_1 and T_2 exactly. From the empirical statistics of a long block-length sample we can recover the single bit that determines the particular distribution in the class. This "extra bit" corresponds to the δ_n term in the proof above. There, approximately $\log n$ bits of information about the joint distribution of the source are transmitted in addition to the source sequences.

We now want to look at what happens as the number of sources gets larger. If there are m sources, the previous scheme would use $m \log n$ bits about the joint distribution. One way of thinking about the blocklength n is as a processing delay. If the number of sources is allowed to grow with the processing delay, so that m and n are going to ∞ *together*, the penalty may become

non-negligible. Suppose that $|\mathcal{S}_j| = K$ for all j. For any fixed number m of sources, the number of types of denominator n is upper bounded by $(n+1)^{mK}$. As m gets larger, the number of types grows exponentially with m, which causes a larger rate penalty using the Slepian-Wolf system. Increasing m requires using a larger blocklength to achieve the same error probability. We would like to capture this intuition. To do this, we will fix the ratio $(m \log n)/n$. and consider then a sequence of $(n(m), R_1, \ldots R_m)$ codes.

Definition 4. *A rate sequence $\{R_i\}_{i=1}^{f(n)}$ is achievable under scaling $f(n)$ if there exists a sequence of $(n, R_1, \ldots, R_{f(n)})$ source codes whose probability of error $\epsilon_n \to 0$ as $n \to \infty$.*

This definition forces a tradeoff between the number of sensors and the blocklength of the corresponding universal code. If the modeling complexity increases exponentially in the number of terminals, we may incur a rate penalty. However, if the number of models is only polynomial in the terminals, the exponential error probability from the block code can easily encompass the modeling complexity as well. Our main result is the following.

Proposition 7. *Let $\{\alpha_\sigma : \sigma \subseteq \{1, 2, \ldots m\}\}$ be a set of positive real numbers. Let $\mathbf{S}(\sigma) = \{S_j : j \in \sigma\}$. Consider the class of sources $\{P_\lambda : \lambda \in \Lambda\}$ for the random variables $(S_1, S_2, \ldots, S_m)$ that satisfy:*

$$\alpha_\sigma > H(\mathbf{S}(\sigma)|\mathbf{S}(\sigma^c)) \tag{3.19}$$

Then the sum-rate for achievable tuples under scaling $n/\log n$ is bounded:

$$\lim_{m\to\infty} \sum_{j=1}^{m} R_j \geq \lim_{m\to\infty} H(S_1, S_2, \ldots, S_m) + c_1 \tag{3.20}$$

Proof. We can simply look at the entropy of the source. For a fixed m and n, consider jointly encoding the source $(S_1, S_2, \ldots, S_m)$. This requires $nH(S_1, S_2, \ldots, S_m)$ bits plus $H(\mathcal{P}_n \cap \Lambda)$. There are $c_2 n^{mc_3}$ sources in $\mathcal{P}_n \cap \Lambda$, so we need $c_4 m \log n$ additional bits. Dividing by the block length n:

$$\lim_{m\to\infty} \sum_{j=1}^{m} R_j \geq \lim_{m\to\infty} \geq H(S_1, S_2, \ldots, S_m) + c_4 \frac{m \log n}{n} \ . \tag{3.21}$$

□

Note that if m and n grow at the same rate, then the rate penalty increases logarithmically in the block length. In some sense we can view this as a density-delay tradeoff for source coding systems of this type. For very large systems, the delay required to amortize the modeling uncertainty is also large.

3.3 The example revisited

Although we only discussed lossless coding problems in this chapter, the ideas can provide some intuition for our example. In the case of distributed compression, the problem does not naively decompose into that of partitioning the sensors into two groups and using a CEO code [28] for each group. However, a little more thought shows that the *only* information needed by the terminals to do the encoding is the number of sensors observing the same source as them. However, the decoder must know which source is observed by each sensor in order to do the joint decoding required by the CEO code.

In the case where the matrix A is known, the distortion can be bounded by

$$D \leq \frac{\sigma_S^2}{\frac{\sigma_S^2}{\sigma_W^2} B_1 (1 - \exp(-2R_1/B_1)) + 1} + \frac{\sigma_S^2}{\frac{\sigma_S^2}{\sigma_W^2} B_2 (1 - \exp(-2R_2/B_2)) + 1} , \tag{3.22}$$

where the rates R_1 and R_2 are the sum rates for the terminals observing S_1 and S_2 respectively.

How much information is needed to enable this performance? In order to determine which source they are observing, each sensor can compare its own source sequence to a common "pilot" sequence, perhaps observed by another sensor. This pilot sequence could be the signs of the first ρ samples observed at the sender. By measuring the empirical correlation between the pilot and their own signal, each sensor can decide with high probability to which group they belong. In addition, they would need the numbers B_1 and B_2. In terms of bits, we have

$$\rho + \log B_1 + \log B_2 < \rho + 2 \log M \text{ bits} \tag{3.23}$$

Since these bits must be shared by all the sensors, it would be sufficient to broadcast all this

information to them before transmission. This is precisely the intuition behind the scheme described in the next chapter.

Chapter 4

Finale: fading observations and alignment

In this chapter we look at a joint source-channel coding problem with uncertainty in the observation process. We call these models *fading observation models* in analogy to fading channels in wireless communications. Much of this work has appeared in some form in [32], [33]. In a sense, this chapter will deal with a part of the interface between the previous two problems. The communication channel imposes constraints on our encoding strategies. In the language of Chapter 2, these may be constraints on the covariance of the estimation matrix. In terms of the coding problems in Chapter 3, there may be a rate constraint on our codes. However the underlying question is this: what should the sensors send in order to best help the estimator, given that they must share the available communications resources?

At first glance, it would seem that minimizing the redundancy in the *messages* sent by the sensors would provide the decoder with the most information for its estimate. More precisely, the sensors would use an optimal CEO source code [28] followed by a capacity-achieving channel code. The decoder could then decode the compressed source observations and use those to estimate the source. The problem with this approach is that the multiple-access channel is a bottleneck for the rate. If the communication power available grows linearly with the number of sensors, the capacity

grows only logarithmically. The end-to-end distortion for this strategy scales to zero like $1/\log M$ for a Gaussian source observed in Gaussian noise with an additive white Gaussian noise channel.

The other approach is to let the sensors collaborate in communicating their observations. The uncoded transmission strategy adopted by Gastpar and Vetterli [17] is one such collaboration method. However, all of the sensors must know the joint statistics of the observation process as well as the communication channel. In the case where there is observation uncertainty, the problem of uncoded may be significantly more difficult, as we have seen in the case of centralized estimators.

In this chapter we consider the slow-fading models examined at the end of Chapter 2. Although we cannot prove that there is a strict penalty in the CEO code, the uncoded transmission scheme fails completely in the presence of fading. We exhibit a simple feedback protocol that can bootstrap the uncoded transmission scheme into a regime with a distortion that scales like $M^{-1/3}$, better than that of separate source and channel coding without fading. We also conjecture that the CEO code with similar feedback still exhibits the same $1/\log M$ scaling.

4.1 Uncertainty in observations

We first describe a general class of sensor observation models and then a specific example that will dominate the bulk of the analysis in this chapter.

4.1.1 Fading observations : a general model

The model we propose is similar to that studied at the end of Chapter 2. For clarity, we review some notation. We use an uppercase letter S for a random variable, a lowercase s for its realization, and $P_S(s)$ for its distribution. We write an independent, identically distributed (iid) sequence of random variables indexed by n as $\{S[n]\}_{n=1}^{\infty}$ or S^n. If $S[n]$ is vector valued, we denote the m-th component of $S[n]$ by $S_m[n]$.

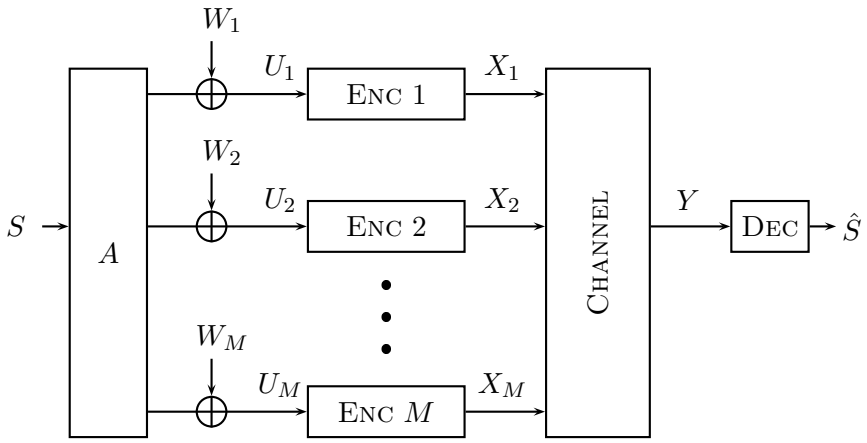

Figure 4.1. Sensor network with fading observations. The function A can be arbitrary, but we will generally assume that it is a linear transformation.

A source generates a sequence of iid symbols

$$S^n = \{S[n]\}_{n=1}^{\infty} \tag{4.1}$$

at each time n according to $P_S(s)$. The M sensors observe the the source through the *observation function* $A(\cdot)$, corrupted by noise W that is iid across sensors and time with distribution $P_W(w)$:

$$U^n = A(S^n) + W^n \ . \tag{4.2}$$

The observation function $A(\cdot)$ is a random variable taking values in a set of functions $\mathcal{A}$ according to some known distribution $P_A(a)$. The choice of $\mathcal{A}$ depends on the specifics of the sensors' design and reflects the relationship between the quantities of interest and the actual observed variables. We do not assume the sensors know the function $A(\cdot)$ after it is chosen. The fast-fading situation with centralized estimation was dealt with in Chapter 2, and we do not address it in the context of joint source-channel coding; instead we will focus on the slow-fading case.

The model generalizes others in the literature. We can view the source as being observed through an input channel as in [11], where the known channel is replaced by a fading channel. If $A(\cdot)$ is the identity map with probability one, we reduce to the the CEO problem [37]. If S is a jointly Gaussian vector and $A(\cdot)$ is a random matrix, we have the source model studied by

Viswanath [36]. If A is deterministic and the communication channel is also modeled by matrix multiplication, we have the network studied in [18]. Typical channel fading models model the fading as multiplicative because of experimental evidence on multipath interference. While the main example we study in this chapter is multiplicative, it is for reasons of expediency rather than the appropriateness of the model. As usual, we will treat sources and noises that are jointly Gaussian.

In this network, the communication channel is added into the picture. We model it as a general multiple-access channel with transition probabilities $p(y|x_1, x_2, \ldots x_M)$. The sensors encode their observations independently into channel inputs X^n. The specific realization of A is not known to the encoders, but the prior distribution $P_A(a)$ is known. The communication channel may have a cost function ρ which the inputs must satisfy with some constraint P in the following sense:

$$E\left[\rho\left(X[n]\right)\right] \leq P \ . \qquad \forall n \tag{4.3}$$

The receiver uses Y to form an estimate $\hat{S}^n = \{\hat{S}[n]\}_{n=1}^{\infty}$ that minimizes a distortion measure $d(S^n, \hat{S}^n)$. In general, the decoder also does not know the realization a of A, so the distortion will depend on this realization. We write the end-to-end distortion $D(A, M, P)$ that is achieved as

$$D(A, M, P) = E_S\left[d\left(S^n, \hat{S}^n\right)\right] \ . \tag{4.4}$$

The goal of our coding scheme is to minimize the expected value of D.

In order to express out scaling results, we use standard asymptotic notation. We say that a function $f(M)$ scales as fast as $g(M)$ or $f(M) = \Omega(g(M))$ if there exists a constant c_l such that

$$f(M) \geq c_l g(M) \ . \tag{4.5}$$

We say that a function $f(M)$ scales as slow as $g(M)$ or $f(M) = O(g(M))$ if there exists a constant c_u such that

$$f(M) \leq c_u g(M) \ . \tag{4.6}$$

We write $f = O(\max\{g_1(M), g_2(M)\})$ if there is a constant c_u such that

$$f(M) \leq c_u \max\{g_1(M), g_2(M)\} \ . \tag{4.7}$$

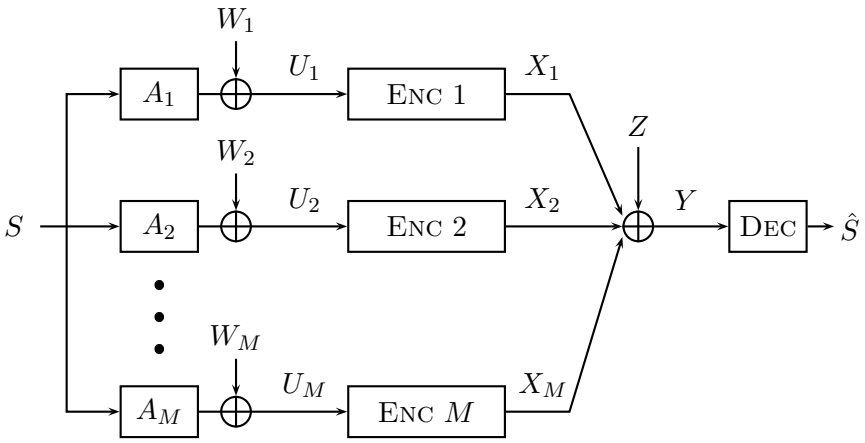

Figure 4.2. Gaussian network with fading observations.

By limiting ourselves to these asymptotic characterizations, we conveniently ignore many factors which affect smaller networks. However, we feel that this analysis is useful in characterizing achievable performance limits.

4.1.2 Scalar multiplicative fading

The first question to ask for sensors observing faded observations is to what extent they can determine the fading process. Consider the case where each sensor receives $A_m(S[n]) + W_m[n]$, where the A_m are drawn iid from some distribution over a finite set $\mathcal{A}$. The sensors can estimate their own marginal distribution locally. If two fading functions $a_1, a_2 \in \mathcal{A}$ induce different marginal distributions on U_m at sensor m, then the sensor can in theory discriminate between them based on the empirical statistics. The problem introduced by fading observations in this setting is from *different* fading functions inducing the *same* marginal distribution on U_m. For point-to-point systems, this is similar to the rate-distortion problem considered by [9]. In order to collaborate, the sensors must disambiguate between the fading processes which could induce their local distribution. We call this problem one of *alignment*.

Again, we turn to Gaussian sources and Gaussian noise. The model we consider is shown

in Figure 4.2. The source S is a Gaussian random variable with mean 0 and variance σ_S^2. The observation noises $\{W_m\}_{m=1}^M$ are iid zero-mean Gaussian random variables with mean σ_W^2. The multiple-access channel is a standard memoryless additive white Gaussian noise channel with a sum-power constraint $\sum E[|X_m|^2] \leq MP$ on the inputs. The channel noise Z is Gaussian with zero mean and variance σ_S^2. The distortion measure is mean-squared error:

$$d(S^n, \hat{S}^n) = \lim_{N\to\infty} \frac{1}{N} E\left[\sum_{n=1}^{N} \left|S[n] - \hat{S}[n]\right|^2\right] . \tag{4.8}$$

Consider a Gaussian network in which the observation fading functions $\{A_m(\cdot)\}$ are multiplication by scalars $\{A_m\}$ satisfying the following:

$$\{A_m\} \text{ are iid with distribution } p_A(a) \tag{4.9}$$

$$p_A(a) = p_A(-a) \tag{4.10}$$

$$|A_m| < \nu \quad \text{a.s.} \tag{4.11}$$

We call this *bounded real scalar fading*. A specific example that we will use is when the $\{A_m\}$ are iid and equiprobable on the set $\{-1, +1\}$. In the next section we will examine the performance of coding strategies on this kind of source fading.

4.2 Existing schemes

The two schemes we examine for this joint-source channel coding problems are uncoded transmission and separation-based coding. Both of these schemes have already been analyzed in the absence of fading [17]. In the separation-based approach, the problem is decomposed into that of distributedly compressing the source into independent bitstreams and then encoding those bitstreams over the multiple-access channel. On the other hand, the sensors could simply forward their observations and use the fact that the channel's summing operation partially computes the MMSE estimate – this is what we call uncoded transmission.

4.2.1 Separate source and channel coding

In this section we derive the asymptotic performance for separation-based compression of sources with distributed encoding. Let $R_S(D)$ be the rate-distortion function for the source S with distortion limit D. Note that $R(D)$ is vector-valued and represents a rate tuple $(R_1, \ldots R_M)$. Let $C(P)$ be the capacity function for the multiple-access channel $p(y|x_1, \ldots x_M)$ under the cost constraint $E[\rho(x_1, \ldots x_M)] \leq P$. Note that $C(P)$ is also a tuple of achievable rates $(R_1^{(c)}, \ldots R_M^{(c)})$ for reliable communication across the multiple-access channel.

Suppose we have a distributed source code for S with rates $(r_1, \ldots r_M)$. If $R(D) = r$ and $r < C(P)$ then we can compress the source and transmit the compressed messages reliably across the channel. Thus if $R(D) < C(P)$ component-wise, we can achieve distortion D across the channel. If $r > C(P)$ in any component, then the rate tuple generated by the source code cannot be communicated reliably across the channel.

Let $R_{tot} = \sum R_m$. If all of the signs $\{A_m\}$ are known, the sum-rate for source coding using a CEO source code in the limit as $M \to \infty$ is given by [28, Equation (6)]:

$$D(R_{tot}, M) = \frac{\sigma_S^2}{\frac{\sigma_S^2}{\sigma_W^2} M(1 - \exp(-2R_{tot}/M)) + 1} \; . \tag{4.12}$$

The sum capacity of the Gaussian multiple-access channel is upper-bounded by the case when the messages may be dependent. The total power is MP, so:

$$R_{tot} \leq \frac{1}{2} \log \left(1 + \frac{M^2 P}{\sigma_S^2} \right) \tag{4.13}$$

Substituting, the achievable distortion is lower bounded by

$$D(M) \geq \frac{\sigma_S^2 \sigma_W^2}{\sigma_W^2 + \sigma_S^2 M \left(1 - \left(\frac{\sigma_S^2}{\sigma_S^2 + M^2 P} \right)^{1/M} \right)} \; . \tag{4.14}$$

Taking the limit as $M \to \infty$ gives the $1/\log M$ behavior described at the beginning of this chapter.

4.2.2 Uncoded transmission

In the uncoded transmission scheme, each sensor scales its own observation to meet the power constraint of the channel. Define the constant η as

$$\eta = \sqrt{\frac{P}{\sigma_S^2 + \sigma_W^2}} \ . \tag{4.15}$$

Then

$$X_m[n] = \eta U_m[n] = \eta(A_m S[n] + W_m[n]) \ . \tag{4.16}$$

The received signal is then

$$Y[n] = \eta \left(\sum_{m=1}^{M} A_m S[n] + \sum_{m=1}^{M} W_m[n] \right) + Z[n] \tag{4.17}$$

The MMSE estimate of S given Y will depend on the random variable $B = \sum_{m=1}^{M} A_m$. For a fixed B define

$$L(B, M) = \frac{\sigma_S^2 \sigma_W^2}{\frac{B^2}{M + (\sigma_S^2/\sigma_W^2)\eta^{-2}} \sigma_S^2 + \sigma_W^2} \ . \tag{4.18}$$

The expected distortion is then

$$D_{unc}(MP) = E_B \left[L(B, M) \right] \ . \tag{4.19}$$

If all the gains A_m are equal to 1, then $B = M$ and the function $L(B, M) = O(M^{-1})$. A more interesting case is when we have bounded real scalar fading:

Proposition 8. *For the Gaussian network with fading observations satisfying (4.9)–(4.11), the distortion scales like* $\Omega(1)$ *using the uncoded transmission scheme.*

Proof. Note that the sensors can each determine the magnitude of their fading coefficients $\{A_m\}$ by computing the marginal density function of their observations. For sensor m, the density of U_m is

$$p_{U_m}(u_m) = \frac{1}{\sqrt{2\pi(A_m^2 \sigma_S^2 + \sigma_W^2)}} \exp\left(-\frac{1}{2(A_m^2 \sigma_S^2 + \sigma_W^2)} u_m^2 \right) \ . \tag{4.20}$$

This density is identical for $A_m = \pm a_m$. The sensors apply the gains

$$\eta_m = \sqrt{\frac{P}{A_m^2 \sigma_S^2 + \sigma_W^2}} \tag{4.21}$$

to get the distortion in (4.19). Note that

$$\sum_{m=1}^{M} \frac{P}{\sigma_S^2 A_m^2 + \sigma_W^2} + \sigma_S^2 \geq M \frac{P}{\sigma_S^2 \nu^2 + \sigma_W^2} + \sigma_S^2 \geq M\mu^2 \tag{4.22}$$

for some $\mu > 0$.

Thus we can lower bound the distortion by

$$D(M) \geq \frac{\sigma_S^2}{\frac{B}{\sqrt{M}}\mu + 1} \ . \tag{4.23}$$

Now note that $B = \sum_{m=1}^{M} \eta_m A_m$, and

$$E[\eta_m A_m] = E\left[\sqrt{\frac{PA_m^2}{A_m^2\sigma_S^2 + \sigma_W^2}} \operatorname{sgn} A_m\right] = 0 \tag{4.24}$$

$$E[\eta_m^2 A_m^2] = E\left[\frac{PA_m^2}{A_m^2\sigma_S^2 + \sigma_W^2}\right] = \sigma_B^2 < \infty \ . \tag{4.25}$$

So by the central limit theorem [14], $BM^{-1/2}$ converges to a Gaussian random variable with mean 0 and variance σ_B^2.

We can now write the expected distortion in the limit:

$$\lim_{M\to\infty} E[D(M)] \geq \lim_{M\to\infty} E\left[\frac{\sigma_S^2}{(BM^{-1/2}\mu)^2 + 1}\right] \tag{4.26}$$

$$= E\left[\lim_{M\to\infty} \frac{\sigma_S^2}{(BM^{-1/2}\mu)^2 + 1}\right] \tag{4.27}$$

$$= \int \frac{\sigma_S^2}{\xi^2 + 1} d\xi \tag{4.28}$$

$$> 0 \ , \tag{4.29}$$

where ξ is normally distributed with mean 0 and variance $\sigma_B^2\mu^2$. Thus the expected distortion does not converge to 0 as $M \to \infty$, so $D(M) = \Omega(1)$. □

We can also give a result that is slightly stronger only on technical grounds.

Proposition 9. *Suppose the fading observation functions in the Gaussian model are multiplication by random variables $\{A_m\}$ taking values in $\{-1, 1\}$. Suppose the $\{A_m\}$ are exchangeable and*

$$\lim_{M\to\infty} M^{-1} \sum A_m = 0 \quad \text{a.s.} \tag{4.30}$$

$$\lim_{M\to\infty} P\left(\frac{1}{\sqrt{M}} \left|\sum A_m\right| > \epsilon\right) \leq K_\epsilon < 1 \ . \tag{4.31}$$

Then uncoded transmission yields an expected distortion that scales with M like $\Omega(1)$.

Proof. Let $H = (B > \sqrt{M}\epsilon)$. In this case,

$$P(H^c) \geq 1 - K_\epsilon > 0 \tag{4.32}$$

We can find a lower bound on the distortion by only taking the expectation over H^c:

$$\begin{aligned} D_{unc}(MP) &\geq E_\Gamma \left[L(M - 2\Gamma, M)\right] \\ &\geq L(\sqrt{M}\epsilon^2, M)\left(K_\epsilon\right) \end{aligned}$$

As $M \to \infty$, this function converges to a constant, so $D_{unc}(M) = \Omega(1)$. □

4.3 A simple feedback framework

The problem with the uncoded transmission scheme is that the sensors' observations are not aligned, so blindly forwarding their observations causes sufficient interference to void the collaborate gain from the coherent addition of the source observations. Another way of viewing this is that any choice of gains is a choice of estimator, and that every estimator performs poorly in expectation over the distribution of A. This lack of alignment is caused by the zero-mean distribution we chose for the observation gains. We now present a scheme that uses a small amount of extra information that recovers some of the performance for uncoded transmission in the unknown-sign model. We show that with a single bit broadcast to all the sensors we can make the distortion converge to 0 as $O(M^{-1/3})$, and for K bits we get $O(M^{-K/(K+2)})$. We generally refer to the extra information used as *feedback*, although we emphasize that it need not come from the end receiver in the sensor

network. Indeed, an interesting case is when the feedback comes from one of the other sensors in the network.

The method we propose to align the sensors is to divide the operation of the network into two phases – a *discovery phase* and a *transmission phase*. We define the time axis so that the discovery phase is at times $n \leq 0$ and the transmission phase is at times $n > 0$. In the discovery phase sensor m forms an estimate $\hat{A}_m$ of its own observation gain A_m based on its own observations and some feedback. Sensor m then forwards $\eta \hat{A}_m U_m$ to compensate for its observation gain. If a sufficient fraction of the sensors are aligned, the distortion will converge to 0 rapidly as the number of sensors increases.

4.3.1 A single bit of feedback for sign fading

In this section we consider a discovery phase of only one time step. Let $S_0 = S[0]$ be the value of the source during the discovery phase on which we base our feedback signal. Let $f(S_0)$ be a feedback signal that is broadcast to all of the sensors. The function $f(\cdot)$ may be stochastic – for example, it may be a noisy observation of the source. We give examples of specific forms of feedback and decision rules in the next section.

After receiving $f(S_0)$, each sensor m forms an estimate $\hat{A}_m = g_m(f(S_0), U_m)$ of its observation gain. Conditioned on a value $S_0 = s_0$, the sensor observations are independent and identically distributed, so the events of successful estimation of the observation gains are independent from sensor to sensor. By the exchangeability of the gains A_m, each sensor should attempt to maximize their probability of success, and will adopt the same decision rule $g_m(\cdot) = g(\cdot)$. Let $v(s_0)$ denote the probability that a sensor correctly estimates its own observation gain. Let $\hat{A}_m = g(f(S_0), U_m)$. Upon making their decisions, the sensors then form the channel inputs

$$X_m[n] = \eta \hat{A}_m U_m[n] \ . \tag{4.33}$$

Let Γ denote the number of sensors for which $\hat{A}_m = A_m$. Conditioned on the value of S_0, the event of each sensor being aligned correctly depends only on the noise value at that sensor, and hence is

an independent Bernoulli random variable with parameter $v(s_0)$, so Γ is a binomial random variable with parameters $(M, v(s_0))$.

The distortion achievable after the feedback is given by

$$E[D \mid f(S_0)] = E_{S_0}\left[\sum_{k=0}^{M} L(M-2k, M) P(\Gamma = k)\right] \tag{4.34}$$

where the expectation is taken over S_0 and

$$P(\Gamma = k) = \binom{M}{k} v(s_0)^k (1 - v(s_0))^{M-k} . \tag{4.35}$$

For equation (4.34) to converge to 0, each term in the summation must converge to 0 as $M \to \infty$. The convergence is in turn dependent on S_0 and the decision rule $g(\cdot)$ by which the sensors estimate their alignment. Intuitively, if S_0 is close to zero, it will be difficult to determine A_m and thus the probability $v(S_0)$ that sensor m aligns correctly will be close to $1/2$. We therefore wish to find a decision rule $g(\cdot)$ that minimizes the probability of error for each sensor, or, alternately, that maximizes the probability of correct alignment

In order for $D(M) = O(M^{-1})$, the random variable Γ must be bounded away from $1/2$ with probability one over the possible values of S_0. Assume that $\lim_{M\to\infty} \Gamma > 1/2 + \epsilon$ with probability one. By the strong law of large numbers, $\Gamma/M \to v(S_0)$ with probability one, which implies $v(S_0) > 1/2 + \epsilon$ with probability one, which will not be true in general. However, if we allow ϵ to scale with M as well, we can recover some of the scaling rate, as shown in the next proposition.

Proposition 10. *Suppose the feedback function $f(\cdot)$ and decision rule $g(\cdot)$ are chosen such that there exist constants $\epsilon_1 > 0$, $\epsilon_2 > 0$, and functions $\alpha(M)$ and $\beta(M)$ such that*

$$\lim_{M\to\infty} \alpha(M) = 0 \tag{4.36}$$

$$\lim_{M\to\infty} \beta(M) = 0 \tag{4.37}$$

$$\lim_{M\to\infty} M\beta(M)^2 = \infty \tag{4.38}$$

$$(|S_0| \geq \epsilon_1 \alpha(M)) \implies v(S_0) \geq \frac{1}{2} + \epsilon_2 \beta(M) . \tag{4.39}$$

Then as $M \to \infty$, the feedback/decision pair (f, g) achieves a distortion $D(M)$ in the transmission

phase satisfying

$$D(M) = O\left(\max\left\{\alpha(M), \frac{1}{M\beta(M)^2}\right\}\right) \ . \tag{4.40}$$

For the network with sign fading.

Remark. Equation (4.39) says that if the source sample S_0 on which we base our feedback is large enough, e.g. at least $\epsilon_1\alpha(M)$, then the probability of successful alignment is at least $\epsilon_2\beta(M)$ more than $1/2$.

Proof. Suppose that (4.36) – (4.39) hold. Let H be the event $(|S_0| \geq \epsilon_1\alpha(M))$. Let Γ be a binomial random variable with parameters $(M, v(s_0))$ so that $E[\Gamma] = Mv(s_0)$. We have the following bound:

$$P\left(|\Gamma - Mv(s_0)| > \frac{M\beta(M)\epsilon_2}{\sqrt{2}}\right) \tag{4.41}$$

$$\leq \exp\left(-\frac{1}{2}\epsilon_2^2 M\beta(M)^2\right) \ . \tag{4.42}$$

We can also write the following bounds, using the assumptions in (4.39):

$$P(H^c) = \frac{1}{\sqrt{2\pi\sigma_S^2}}\int_{-\epsilon_1\alpha(M)}^{\epsilon_1\alpha(M)} e^{-x^2/2\sigma_S^2}dx \tag{4.43}$$

$$\leq \sqrt{\frac{2}{\pi\sigma_S^2}}\epsilon_1\alpha(M) \tag{4.44}$$

$$v\left(S_0\,|H\right) \geq \frac{1}{2} + \epsilon_2\beta(M) \ . \tag{4.45}$$

We evaluate the distortion separately on the events

$$F_1 = H^c$$

$$F_2 = H \cap (\Gamma > (\epsilon_2/2)M\beta(M))$$

$$F_3 = H \cap (\Gamma \leq (\epsilon_2/2)M\beta(M)) \ .$$

On H^c we upper bound the distortion by letting $\Gamma = M/2$. On H, we have equation (4.45), so we can let $\Gamma = M/2$ on F_2 and $\Gamma = \frac{1}{2} + (\epsilon_2/2)M\beta(M)$ on F_3. Putting this together and noting that

the distortion must be less than σ_S^2, we rewrite (4.34) as :

$$\begin{aligned} & E_{S_0}\left[L(M-2\Gamma,M)(1_{F_1}+1_{F_2}+1_{F_3})\right] \\ \leq\ & \sigma_S^2\left(P(H^c)+P\left(\Gamma>(\epsilon_2/2)\mid H\right)\right)+L(M\epsilon_2\beta(M)/2,M)P\left((\Gamma\leq(\epsilon_2/2)\mid H\right) \\ \leq\ & \sigma_S^2\left(\sqrt{\frac{2}{\pi\sigma_S^2}}\epsilon_1\alpha(M)+\exp\left(-\frac{1}{2}\epsilon^2M\beta(M)^2\right)\right)+L(M\epsilon_2\beta(M)/2,M)\ . \end{aligned}$$

The first term converges to zero with the slower of $\alpha(M)$ and $\exp(-M\beta(M)^2)$. The second term converges to zero as $O(M^{-1}\beta(M)^{-2})$. Since $\beta(M)^{-2}$ is sub-linear, we can ignore this latter term, so the distortion is $O(\max\{\alpha(M), M^{-1}\beta(M)^{-2}\})$. □

The bounds for this proof depend only one the relationship between the values of S_0 and the success probability $v(S_0)$. The latter depends only on the noise distribution, and thus it is possible to treat non-Gaussian noises, although in this case the destination's linear estimator will not necessarily be the MMSE estimator.

Proposition 11. *Suppose the fading observation functions satisfy the conditions of Proposition 9. Then the K-bit feedback scheme outlined above achieves a distortion that scales like $O(M^{-K/(K+2)})$.*

4.3.2 Example: feedback from a beacon sensor

Example: perfect feedback

To gain further insight, let us consider an idealized case in which $f(\cdot)$ is the identity function, so that the sensors get to know the source sample exactly; we call this *perfect feedback*. We emphasize that this feedback is only available during the discovery phase and not for all time, and that we are assuming that the discovery phase is one sample long.

Conditioned on the value of $S_0 = s_0$, each sensor is left with the problem of detecting antipodal signals $\pm s_0$ in the presence of Gaussian noise with a uniform prior. The maximum a priori probability (MAP) rule for this problem is a threshold test at 0; for $s_0 > 0$, if $U_m > 0$ then $\hat{A}_m = 1$, and

for $s_0 < 0$ if $U_m > 0$ then $\hat{A}_m = -1$. The probability of success for this rule is

$$v_p(s_0) = 1 - Q\left(\frac{|s_0|}{\sigma_w}\right) \ , \tag{4.46}$$

where

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_x^\infty e^{-y^2/2} dy \ . \tag{4.47}$$

The success probability, conditioned on $H = |S_0| \geq \epsilon_M$, is

$$v_p\left(S_0 \mid H\right) \geq \frac{1}{2} + \frac{\epsilon_M}{2\sqrt{2\pi\sigma_W^2}} \ . \tag{4.48}$$

Proposition 10 tell us that the distortion scales faster than $\max\{\epsilon_M, M^{-1}\epsilon_M^{-2}\}$. Equating these two scaling rates, we can set $\epsilon_M = O(M^{-1/3})$ to yield a scaling rate of $O(M^{-1/3})$.

Suppose instead that the sensors do not receive S_0, but instead a one bit quantization of S_0, or $f(S_0) = \text{sgn}(S_0)$. Since the the MAP rule with full knowledge of S_0 was a threshold test at 0 for all values of S_0, the MAP rule for this case is the same. Another way to phrase the decision rule is if $f(S_0) = \text{sgn}(U_m)$ then $\hat{A}_m = 1$, otherwise $\hat{A}_m = -1$. We again condition on the value of S_0, which gives the the same success probabilities as (4.46) and (4.48). Therefore the distortion with this one bit of feedback is also $O(M^{-1/3})$. It is interesting to note that in this case the achievable distortion scaling does not depend on the "richness" of the actual available feedback.

Example: one bit of feedback

Suppose instead that each sensor is given access to the sign of the signal received at an extra sensor b:

$$f(S_0) = G_b = \text{sgn}(A_b S_0 + W_b) \ . \tag{4.49}$$

Sensor m must then decide if $A_m = A_b$ or $A_m \neq A_b$. Again, we condition on $S_0 = s_0$. The probability that $G_b = A_b$ is given by

$$P\left(G_b = A_b \,|S_0 = s_0\right) = 1 - Q\left(\frac{|s_0|}{\sigma_w}\right) \ . \tag{4.50}$$

Since we cannot exactly know the signs A_b and A_m, we assume without loss of generality that $A_b = 1$ and attempt to distinguish between the two hypotheses $A_m = A_b$ and $A_m = -A_b$.

Suppose we have full knowledge of $U_b = A_b S_0 + W_b$. Under the hypothesis $A_m = A_b$, the pair $(A_b s_0 + W_0,\, A_m s_0 + W_m)$ is jointly Gaussian with mean (s_0, s_0). Under the hypothesis $A_m = -A_b$ they are jointly Gaussian with mean $(s_0, -s_0)$. The decision rule in this case is again a threshold test on the line $U_m = 0$. If U_m and U_b have the same sign, then sensor m guesses $\hat{A}_m = \hat{A}_b = 1$, and if they have different sign it guesses $\hat{A}_m = -\hat{A}_b = -1$. This rule is again indifferent to the value of S_0, as in the perfect feedback case.

The decision rule for A_m that maximizes a posteriori probability of the observations is therefore given by

$$\hat{A}_m = \begin{cases} g(G_b, U_m) = 1 & \text{if } G_b = \text{sgn}(U_m) \\ g(G_b, U_m) = -1 & \text{if } G_b \neq \text{sgn}(U_m)\ , \end{cases} \tag{4.51}$$

regardless of the value of S_0. The probability of success is given by

$$v_n(s_0) = 1 - 2Q\left(\frac{|s_0|}{\sigma_w}\right) Q\left(\frac{-|s_0|}{\sigma_w}\right)\ . \tag{4.52}$$

So the conditional probability of success is:

$$v_n\left(S_0 \,\middle|\, |S_0| > \frac{\epsilon_M}{2}\right) \geq \frac{1}{2} + \frac{\epsilon_M}{\sqrt{2\pi\sigma_W^2}} + \frac{\epsilon_M^2}{2\pi\sigma_W^2}\ . \tag{4.53}$$

The ϵ_M term is dominant as $M \to \infty$ in this conditional probability, the same as in the perfect feedback case. From our previous analysis, we can see that $D(M) = O(M^{-1/3})$.

The noisy feedback example models a situation in which one sensor acts as a "beacon" by broadcasting the sign of its observation to the other sensors. In a scaling sense, the sign of the noisy observation is "as good" as knowing the sign of the source sample exactly, although the constants in the convergence become worse as the noise becomes more severe.

4.3.3 Many bits of feedback

We now consider the effect of lengthening the discovery phase by allowing the feedback involve more than one sample of the source. Let $S_0 = S[0]$ and $S_{-1} = S[-1]$ be the two source samples on which we base our feedback $f(S_0, S_{-1})$. Conditioned on a realization (s_0, s_{-1}) of these two

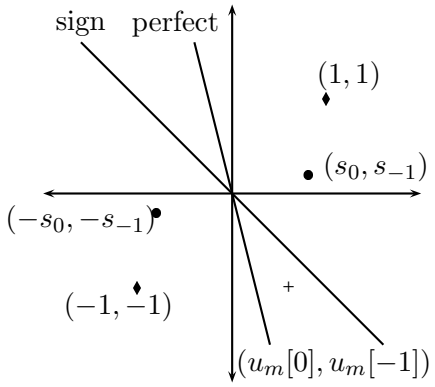

Figure 4.3. MAP rules for perfect feedback and perfect sign feedback. The plus is the noisy observation $(u_m[0], u_m[-1])$. Under perfect information, $\hat{A}_m = 1$, whereas with only sign information the *expected probability of success* is maximized when $\hat{A}_m = -1$.

samples, the sensor observations are again independent, so each sensor should seek to maximize its own probability of success. To illustrate the effect of adding more feedback, we consider the perfect feedback of Section 4.3.2 to simplify the expressions. A similar analysis can be carried out for the noisy feedback case.

Suppose our feedback is the pair (S_0, S_{-1}). Conditioned on the values of S_0 and S_{-1}, sensor m's observations $(U_m[0], U_m[-1])$ are jointly Gaussian with mean (s_0, s_{-1}) under the hypothesis $A_m = 1$ and mean $(-s_0, -s_{-1})$ under the hypothesis $A_m = -1$. The problem is again the same as that of detecting antipodal signals in the presence of Gaussian noise, and the MAP rule is a threshold test shown in Figure 4.3. The probability of success for this rule is

$$v_p(s_0, s_1) = 1 - Q\left(\frac{\sqrt{s_0^2 + s_{-1}^2}}{\sigma_w}\right) . \tag{4.54}$$

Let H be the event $(|S_0| \leq \epsilon\alpha(M), |S_{-1}| \leq \epsilon\alpha(M))$, and H^c its complement. Then we have:

$$\begin{aligned} v_p\left(S_0, S_{-1} \mid H^c\right) &\geq 1 - Q\left(\frac{\sqrt{2}\epsilon\alpha(M)}{\sigma_w}\right) && (4.55) \\ &\geq \frac{1}{2} + \epsilon\alpha(M)\frac{1}{2\sqrt{\pi\sigma_W^2}} . && (4.56) \end{aligned}$$

Since S_0 and S_{-1} are independent, $P(H)$ is proportional to $\alpha(M)^2$. The analysis in Proposition

10 implies that $D(M) = O(\max\{\alpha(M)^2, M\alpha(M)^{-2}\})$. Setting these two terms equal gives $\alpha(M) = M^{-4}$ so $D(M) = O(M^{-1/2})$.

Extending the above analysis in a standard way shows that with K samples of feedback we get distortion $D(M) = O(M^{-K/(K+2)})$. By choosing K arbitrarily large, we get closer to the optimal rate of $O(M^{-1})$.

Suppose we only get the signs of S_0 and S_{-1}, so that $f(S_0, S_{-1}) = (\text{sgn}(S_0), \text{sgn}(S_{-1}))$. The threshold test in the MAP rule for perfect feedback depends on the actual values of (s_0, s_{-1}), as opposed to the threshold test when $K = 1$. Thus the scaling result above does not immediately follow. Sensor m must then determine, based on the observation pair $(U_m[0], U_m[-1])$, whether $A_m = 1$ or $A_m = -1$ in a way that maximizes its probability of making a correct decision. We would like for the probability of success to be greater than 1/2.

Under the two hypotheses, the pair $(U_m[0],\ U_m[-1])$ is jointly Gaussian with mean (s_0, s_{-1}) for $A_m = 1$, mean $(-s_0, -s_{-1})$ for $A_m = -1$ and covariance $\sigma_W^2 I$. The likelihood of observing $(U_m[0], U_m[-1])$ is the expectation of the conditional likelihood over all source pairs (s_0, s_{-1}) that could have generated $f(s_0, s_{-1})$. Because of the symmetry in the distribution of (S_0, S_{-1}), the likelihoods under the two hypotheses are equal on the line orthogonal to the vector $f(s_0, s_{-1})$ so the MAP estimate is a threshold on that line.

To illustrate the difference between the MAP estimate for the case of perfect feedback versus the case of sign feedback, consider Figure 4.3. For a fixed (s_0, s_{-1}), the probability of error is given by the probability that the noise exceeds the distance from the point (s_0, s_{-1}) to the threshold in the direction orthogonal to the decision boundary:

$$v_p(s_0, s_{-1}) = 1 - Q\left(\frac{|s_0| + |s_{-1}|}{\sqrt{2}\sigma_w}\right) \tag{4.57}$$

This differs from equation (4.54) by a shift in the norm inside the $Q(\cdot)$ function; with perfect knowledge of the samples we have a $\mathcal{L}_2$ norm, and with only the sign we have $\mathcal{L}_1$. This explains why the tests and errors were the same in the case when $K = 1$.

Let H be the event that $|S_0|, |S_{-1}| \leq \epsilon\alpha(M)$ and H^c its complement. Then we can bound the

probability of success:

$$v_p\left(s_0, s_{-1}\,|H^c\right) \geq \frac{1}{2} + \frac{\epsilon}{\sqrt{\pi\sigma_W^2}}\alpha(M) \tag{4.58}$$

Thus from our previous analysis, $D(M) = O(M^{-1/2})$ as in the case when we have perfect knowledge of the source samples.

4.3.4 Feedback for bounded scalar fading

As an extension to these results on sign fading, we can modify the scheme for bounded real scalar fading. In order to prevent excess noise, we can set a threshold of ϵ. All sensors which estimate their gain $|A_m| < \epsilon$ do not transmit anything. Since the gains are iid this affects at a constant fraction $P(|A_m| < \epsilon)$ of the sensors almost surely as $M \to \infty$. Consider the case of 1-bit feedback. Suppose the beacon broadcasts the sign of its observation G_b at time 1 to all the other sensors:

$$G_b = \text{sgn}(U_0[1]) = \text{sgn}(A_b S[1] + W_0[1]) \tag{4.59}$$

We call this signal the feedback function $f(U_b)$

Sensor m then checks if $G_b = G_m$ (defined analogously). If so, it estimates $\text{sgn}(A_m) = \text{sgn}(A_b)$, otherwise it estimates $\text{sgn}(A_m) = -\text{sgn}(A_b)$. Call this decision rule $g(G_b, U_m)$. This rule is the maximum a posteriori probability (MAP) rule for detecting the sign of A_m. Denote by $v_m(S[1])$ the probability of successfully estimating $\text{sgn}(A_m)$ at sensor m using this rule. Conditioned on $S[1]$, all the sensor observations are independent. We have already assumed that $A_m > \epsilon$, so $v_m(S[1]) > 1/2 + \delta$ for all m. We have the same proposition relating the scaling rate to the to the success probability:

Proposition 12. *Suppose the feedback function $f(\cdot)$ and decision rule $g(\cdot)$ are chosen such that there exist constants $\epsilon_1 > 0$, $\epsilon_2 > 0$, and functions $\alpha(M)$ and $\beta(M)$ such that*

$$\lim_{M\to\infty} \alpha(M) = 0 \tag{4.60}$$

$$\lim_{M\to\infty} \beta(M) = 0 \tag{4.61}$$

$$\lim_{M\to\infty} M\beta(M)^2 \quad = \quad \infty \tag{4.62}$$

$$(|S[1]| \geq \epsilon_1\alpha(M)) \implies v(S[1]) \geq \frac{1}{2} + \epsilon_2\beta(M) \ . \tag{4.63}$$

Then as $M \to \infty$, the feedback/decision pair (f, g) achieves a distortion $D(M)$ in the transmission phase satisfying

$$D(M) = O\left(\max\left\{\alpha(M), \frac{1}{M\beta(M)^2}\right\}\right) \ . \tag{4.64}$$

for the Gaussian network with bounded real scalar fading.

We then have, by application of the previous proposition, that a single bit of feedback in the mode we have described biases the sum of the forwarded observations into a regime that scales faster than $\sqrt{M}$, which then avoids the central-limit behavior of the misaligned situation. It is clear from these examples that the situations in which the fading process $\{A_m\}$ is problematic are when A_m is zero-mean and symmetric. In these cases the sign of A_m plays a key role. This can be thought of as a phase uncertainty introduced by the fading observations.

4.3.5 A conjecture for the CEO problem with limited feedback

The obvious question to ask at this point is how separate source and channel coding is affected by the introduction of similar feedback capabilities. Although we cannot provide a definitive answer at this time, recent results on the CEO problem [29], [31], [38], [39] suggest that the feedback signal will not affect the scaling behavior of the distortion-rate function. We will briefly sketch an argument along these lines, following the very recent work of Wagner [39].

We can bound the performance of a CEO code with a beacon by changing the model slightly. Let us suppose there are $M + 1$ sensors numbered $0, 1, 2, \ldots M$ and let U_m be defined as before. Now suppose $A_0 = 1$ and let the input to each sensor's encoder be the pair (U_m, U_b) for $m = 1, 2, \ldots M$. We also give the signal U_b to the decoder. Thus the entire problem has been reduced to a "conditional CEO" problem with all terminals sharing a side information signal U_b, and the observed variables at the encoders are conditionally independent given the pair (S, U_b). By evaluating the bound in [39] we can obtain a new set of achievable rates for this problem.

It is intuitively plausible that this coding problem is no different from the original CEO problem save for an extra conditioning on U_b that will only affect the distortion achievable at a given rate by a constant. Furthermore it is clear that this problem will give a lower bound on the distortion-rate function for the case with feedback because we do not even assume a rate-limitation on the side information and even provide it to the decoder. We leave the proof of this conjecture for future work.

4.4 Other directions and our example

We now turn to Gaussian networks in which the source is observed through a class of linear filters. Locally to the sensors, there are two sources of misalignment in these networks: phase uncertainty, and delay uncertainty. The former refers to relative phase differences between sensors with the same power spectrum, and is a generalization of the problem in the previous section. The latter refers to integer delays in the observation of the source. This may be caused by propagation delays or a lack of clock synchronization.

Suppose the observation ensemble $\mathcal{A}$ is a set of filters $\{A^{(k)}[n]\}$ such that the power spectrum of the filtered source $A^{(k)}[n] * S[n]$ is nowhere zero. Each sensor receives the source filtered by one of the filters in $\mathcal{A}$, where $A_m[n]$ is chosen uniformly from $\mathcal{A}$. The sensors can attempt to empirically estimate their observation function's power spectral density and compensate for it by whitening their observations. For example, if $\mathcal{A} = \{\pm 1 \pm \frac{1}{2}z^{-1}\}$, then the power spectral density for sensor m could be

$$U_m(e^{j\omega}) = \sigma_W^2 + \left(1 + \frac{1}{4}\right)\sigma_S^2 + \sigma_S^2 \cos\omega \tag{4.65}$$

for $A_m(z) = \pm(1 + \frac{1}{2}z^{-1})$ or

$$U_m(e^{j\omega}) = \sigma_W^2 + \left(1 + \frac{1}{4}\right)\sigma_S^2 - \sigma_S^2 \cos\omega \tag{4.66}$$

for $A_m(z) = \pm(1 - \frac{1}{2}z^{-1})$.

In this example, some of the filters have the same power spectrum but different phases. The ambiguity in phase, which in this case is again a sign change, cannot be distinguished by the sensors

locally. If feedback were provided in the same way as in the previous section, we might be able to align the sensors to enable coherent communication. In this case, one sensor from each of the two spectral classes would broadcast the sign of its observation – by comparing signs we can achieve the same alignment as before.

More generally, consider a arbitrary collection of causal finite impulse response (FIR) filters $\mathcal{A}$ with identical power spectra, and which all have nonzero impulse response at 0. The sensors cannot distinguish between these filters locally. Suppose the set of indistinguishable filters $\mathcal{A}$ depends on $L+1$ source samples:

$$A(e^{j\omega}) = \sum_{k=0}^{L} a_k e^{-j\omega k} \ . \tag{4.67}$$

Then the power spectrum of the sensor m's observed process is

$$U_m(e^{j\omega}) = \sum_{k=0}^{L} \left(\sum_{l=0}^{L} a_k a_{k-l} \right) \cos k\omega + \sigma_W^2 \ . \tag{4.68}$$

However, the only way for two filters $A_1[n]$ and $A_2[n]$ to induce the same power spectrum $U(e^{j\omega})$ is for the coefficients of $\cos k\omega$ to all be equal. This in turn implies that $a_{1k} = -a_{2k}$ for all k, so the ambiguity is at most between a filter and its negative.

The results of the previous section show that phase misalignment can be catastrophic for the uncoded transmission scheme, but a small amount of feedback can recover some of the performance and outperform the best separation-based approach if the ambiguity takes the form of a sign shift. For ensembles of real filters, we have shown that indeed this ambiguity is at most a sign shift. To help align the sensors with a given filter, we use the same beacon strategy described earlier. This allows a sufficient fraction of the sensors to align themselves with high probability, allowing for further processing to enable coherent communication.

The second source of ambiguity that can arise with linear filters is in the absolute delay. Suppose $\mathcal{A} = \{1, z^{-1}\}$, so that some sensors observe the source with unit delay. Suppose furthermore that $P_A(1) = P_A(z^{-1}) = 1/2$. If the sensors forward their observations uncoded, the received signal $Y[n]$ is given by:

$$Y[n] = \eta \left(B_0 S[n] + B_1 S[n-1] + \sum_{m=1}^{M} W_m[n] \right) + Z[n] \ , \tag{4.69}$$

where B_0 is the number of sensors with zero delay, B_1 is the number sensors with unit delay, and η is

$$\eta = \sqrt{\frac{P}{\sigma_S^2 + \sigma_W^2}} \ . \tag{4.70}$$

The destination can use a smoothing filter to estimate the source sequence. Using a Wiener filter, the power spectrum of the error is given by

$$\varepsilon(e^{j\omega}) = \frac{(M\eta^2\sigma_W^2 + \sigma_S^2)\sigma_S^2}{(B_0^2 + B_1^2 + 2B_0B_1\cos\omega)\eta^2\sigma_S^2 + M\eta^2\sigma_W^2 + \sigma_S^2} \ . \tag{4.71}$$

This error function is different from that in the sign-mismatch case, but also suffers from the effects of phase mismatch from the $\cos\omega$ term. However, a feedback scheme which can bias the sensor's estimate of their own delay can shift the scaling rate in denominator of the integral.

Returning to our canonical example, we can note that it is quite similar to the delay example above. If there are two channel uses per source sample, the sensors can use a feedback scheme to estimate which source they are observing and slot their transmissions accordingly. Unfortunately, as long as a fraction of sensors are *misaligned*, their contributions will result in a coherent interference that scales as fast as the correctly aligned sensors. To be more precise, we would hope to get a distortion of the form

$$D = E\left[\frac{\sigma_S^2\sigma_W^2}{\frac{B_0^2}{B_0+(\sigma_S^2/\sigma_W^2)\eta^{-2}}\sigma_S^2 + \sigma_W^2} + \frac{\sigma_S^2\sigma_W^2}{\frac{B_1^2}{B_1+(\sigma_S^2/\sigma_W^2)\eta^{-2}}\sigma_S^2 + \sigma_W^2}\right] \tag{4.72}$$

$$= 2\frac{\sigma_S^2\sigma_W^2}{\frac{(M/2)^2}{(M/2)+(\sigma_S^2/\sigma_W^2)\eta^{-2}}\sigma_S^2 + \sigma_W^2} \tag{4.73}$$

With the feedback, a portion β of the sensors will remain misaligned (in either direction), so we would have in expectation:

$$D = 2\frac{\sigma_S^2\sigma_W^2}{\frac{M(1/2-\beta)^2\sigma_W^2}{(M/2)\sigma_W^2+M^2\beta^2\sigma_S^2+\sigma_S^2\eta^{-2}}\sigma_S^2 + \sigma_W^2} \tag{4.74}$$

One way around this is to allow the amount of feedback allowed to increase with M. We leave this for future work.

These examples suggest that phase uncertainty can render uncoded schemes useless, but a little bit of feedback goes a long way in terms of scaling rate. For complex signals with continuous

phase differences, the efficacy of this one-shot feedback scheme may be more limited. However, modifying the scheme so that the beacon opportunistically waits for a good signal may allow a tradeoff between the length of the discovery period and the achievable distortion scaling.

Chapter 5

Coda

The current surge in research attention on sensor networks is fueled by the development of practical platforms for implementing a wide range of applications. These applications range from military surveillance, environmental and industrial monitoring, and traffic regulation to "smart homes," commercial robotics, and biomedical sensing. In many cases the network is used to gather data in order to estimate some underlying variable. Because of the inherent unreliability in sensor placement and physical modeling, robust estimation systems are needed in order to deal with real-world applications. In this thesis we took a simple sensor network model and introduced structured uncertainty in the observation model. This structured uncertainty took the form of an unknown correlation between the observed variables. Our objectives were to examine the performance of existing estimators and coding schemes and to propose new tradeoffs and protocols to improve the performance of these schemes in the large-network limit.

The data-gathering sensor networks we study have three main behaviors – an communication, distributed processing, and estimation. We examined these in the reverse order. From the perspective of centralized estimation, we looked at linear observation models with bounds on norms, structural constraints, and random distributions and used the same asymptotic spectral expansions to find the performance of estimators for each type. On the distributed coding front, we looked at universal Slepian-Wolf codes and showed that the order in which the limits are taken in looking

at large block-length codes is important. Finally, we looked at the effect of introducing a simple communication channel into our sensor network. Although phase uncertainty rendered the uncoded transmission protocol useless, we exhibited a simple one-time feedback protocol that can bias the network to recapture some of the performance loss.

The expectation among engineers is that an optimal algorithm should also be robust to perturbations in the modeling assumptions. In sensor networks these perturbations may not be small deviations but instead larger structural uncertainties. Therefore robust models and protocols for sensor networks should incorporate some of this uncertainty or provide a mechanism by which it can be resolved. In this thesis we attempted to provide some steps towards analyzing the effect of this uncertainty for Gaussian networks. Hopefully some of the problems ideas studied here will provide some insight into real network designs.

References

[1] Ammann, L. P. Robust singular value decompositions : A new approach to projection pursuit. *Journal of the American Statistical Association 88*, 422 (1993), 505–514.

[2] Bai, Z. Methodologies in spectral analysis of large dimensional random matrices, a review. *Statistica Sinica 9* (1999), 611–677.

[3] Baron, D., Khojastepour, M. A., and Baraniuk, R. G. Redundancy Rates of Slepian-Wolf Coding. In *43rd Annual Allerton Conference on Communication, Control and Computation* (Monticello, IL, September 2004).

[4] Burrows, M., and Wheeler, D. A block sorting lossless data compression algorithm. Tech. Rep. Technical Report 124, Digital Equipment Corporation, 1994.

[5] Calvin, J. A., and Dykstra, R. L. REML Estimation of Covariance Matrices With Restricted Parameter Spaces. *Journal of the American Statistical Association 90*, 429 (1995), 321–329.

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

[7] Csiszár, I., and Körner, J. G. Towards a general theory of source networks. *IEEE Transactions on Information Theory 26* (1980), 155–165.

[8] Csiszár, I., and Körner, J. G. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Akadémi Kiadó, Budapest, 1982.

[9] Dembo, A., and Weissman, T. The minimax distortion redundancy in noisy source coding. *IEEE Transactions on Information Theory 49*, 11 (2003), 3020–3030.

[10] Dembo, A., and Zeitouni, O. *Large Deviations Techniques and Applications*. Springer, New York, 1998.

[11] Dobrushin, R. L., and Tsybakov, B. S. Information transmission with additional noise. *IEEE Transactions on Information Theory 8* (1962), 293–304.

[12] Draper, S. Universal Incremental Slepian-Wolf Coding. In *43rd Annual Allerton Conference on Communication, Control and Computation* (Monticello, IL, September 2004).

[13] Draper, S. C., Chang, C., and Sahai, A. Sequential Random Binning for Streaming Distributed Source Coding. In *IEEE International Symposium on Information Theory Proceedings (ISIT 2005)* (2005), pp. 1406–1410.

[14] Durrett, R. *Probability: Theory and Examples*, 2nd ed. Duxbury, Belmont, CA, 1995.

[15] Effros, M., Visweswariah, K., Kulkarni, S., and Verdu, S. Universal lossless source coding with the Burrows-Wheeler transform. *IEEE Transactions on Information Theory 48*, 5 (2002), 1061–1081.

[16] Foucault, M. *Discipline and Punish : The Birth of the Prison*. Vintage Books, New York, 1995.

[17] Gastpar, M., and Vetterli, M. Source-channel communication in sensor networks. In *Information Processing in Sensor Networks 2003* (Palo Alto, California, April 2003), F. Zhao and L. Guibas, Eds., Springer, pp. 162–177.

[18] Gastpar, M., and Vetterli, M. Power-bandwidth-distortion scaling laws for sensor networks. In *Information Processing in Sensor Networks 2004* (Berkeley, California, April 2004), ACM.

[19] Gradshteyn, I. S., and Ryzhik, I. M. *Table of Integrals, Series, and Products*, 6th ed. Academic Press, San Diego, 2000.

[20] Gray, R. M. Toeplitz and circulant matrices : a review. Originally written 1971, August 2005.

[21] Grenander, U., and Szegö, G. *Toeplitz Forms and Their Applications*. University of California Press, Berkeley, 1958.

[22] Haykin, S. *Adaptive Signal Processing*, 4th ed. Prentice-Hall, Upper Saddle River, NJ, 2002.

[23] Horn, R. A., and Johnson, C. R. *Matrix Analysis*. Cambridge University Press, Cambridge, 1987.

[24] Jaggi, S., and Effros, M. Universal multiple access source coding. *IEEE Transactions on Information Theory* (to appear).

[25] Marčenko, V., and Pastur, L. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR - Sbornik 1*, 4 (1967), 457–483.

[26] Maronna, R. A., and Yohai, V. J. The Behavior of the Stahel-Donoho Robust Multivariate Estimator. *Journal of the American Statistical Association 90*, 429 (1995), 330–341.

[27] Oohama, Y. Universal coding for correlated sources with linked encoders. *IEEE Transactions on Information Theory 42*, 3 (1996), 837–847.

[28] Oohama, Y. The Rate-Distortion Function for the Quadratic Gaussian CEO Problem. *IEEE Transactions on Information Theory 44*, 3 (1998), 1057–1070.

[29] Oohama, Y. Rate-distortion theory for Gaussian multiterminal source coding systems with several side informations at the decoder. *IEEE Transactions on Information Theory 51*, 7 (2005), 2577–2593.

[30] Oppenheim, A. V., Shafer, R. W., and Buck, J. R. *Discrete-Time Signal Processing*. Prentice Hall, Upper Saddle River, NJ, 1999.

[31] Prabhakaran, V., Tse, D., and Ramchandran, K. Rate region of the quadratic Gaussian CEO problem. In *IEEE International Symposium on Information Theory Proceedings (ISIT 2004)* (2004), p. 117.

[32] Sarwate, A. D., and Gastpar, M. Estimation from misaligned observations with limited feedback. In *39th Conference on Information Sciences and Systems (CISS 2005)* (Baltimore, MD, March 2005).

[33] Sarwate, A. D., and Gastpar, M. Fading observation alignment via feedback. In *4th International Symposium on Information Processing in Sensor Networks (IPSN 2005)* (Los Angeles, CA, April 2005).

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

[35] Vetterli, M., and Kovačević, J. *Wavelets and Subband Coding*. Prentice Hall PTR, Upper Saddle River, New Jersey, 1995.

[36] Viswanath, P. *Sum Rate of a Class of Gaussian Multiterminal Source Coding Problems*. DIMACS. American Mathematical Society, Providence, RI, 2004, pp. 43–60.

[37] Viswanathan, H., and Berger, T. The quadratic Gaussian CEO problem. *IEEE Transactions on Information Theory 43*, 5 (1997), 1549–1559.

[38] Wagner, A. B., and Anantharam, V. An improved outer bound for the multiterminal source coding problem. In *IEEE International Symposium on Information Theory Proceedings (ISIT 2005)* (2005), pp. 1406–1410.

[39] Wagner, A. B., and Anantharam, V. An infeasibility result for the multiterminal source-coding problem. Submitted to IEEE Transactions on Information Theory, November 2005.

[40] Wang, N., and Raftery, A. E. Nearest Neighbor Variance Estimation (NNVE) : Robust Covariance Estimation via Nearest-Neighbor Cleaning. *Journal of the American Statistical Association 97*, 460 (2002), 994–1019.

[41] Willems, F. M. J., Shtarkov, Y. M., and Tjalkens, T. J. The context-tree weighting method: basic properties. *IEEE Transactions on Information Theory* (1995), 653–664.

[42] Ziv, J., and Lempel, A. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory 23*, 3 (1977), 337–343.