1

Distributed Estimation Under Bandwidth and Energy Constraints

Alejandro Ribeiro, Ioannis D. Schizas, Jin-Jun Xiao, Georgios B. Giannakis and Zhi-Quan Luo

University of Minnesota

In parameter estimation problems a sequence of observations $\{\mathbf{x}(n)\}_{n=1}^{N}$ is used to estimate a random or deterministic parameter of interest $\mathbf{s}$. Optimal estimation exploits the statistical correlation between $\mathbf{x}(n)$ and $\mathbf{s}$ that is described either by the joint probability distribution function (pdf) $p(\mathbf{x}(n), \mathbf{s})$ when $\mathbf{s}$ is assumed random; or by a family of observation pdfs $p(\mathbf{x}(n); \mathbf{s})$ parameterized by $\mathbf{s}$ when $\mathbf{s}$ is assumed deterministic. The optimal estimator function producing an estimate $\hat{\mathbf{s}}$ for a given set of observations $\{\mathbf{x}(n)\}_{n=1}^{N}$ is different for random and deterministic parameters. It also depends on the joint pdf $p(\mathbf{x}(n), \mathbf{s})$ (or family of pdfs $p(\mathbf{x}(n); \mathbf{s})$) and the degree of knowledge about them; i.e., whether they are known, dependent on some other (nuisance) parameters, or completely unknown (Kay 1993).

The distributed nature of a wireless sensor network (WSN) implies that observations are collected at different sensors and consequently it dictates that between collection and estimation a communication is present. If bandwidth and power were unlimited, the $\mathbf{x}(n)$ observations could be conveyed with arbitrary accuracy and, intuitively, no major impact would be expected. However, bandwidth and power *are* limited, and the seemingly innocuous communication stage turns out to have a significant impact on the design of optimal estimators and their performance assessed by the estimator variance. On the one hand, if digital communications are to be employed individual observations have to be quantized, transforming the estimation problem into that of estimating $\mathbf{s}$ using a set of quantized observations – certainly different from estimating $\mathbf{s}$ using the original analog-amplitude observations. On the other hand, since components of the (vector) observation $\mathbf{x}(n)$ are

Wireless Sensor Networks: Signal Processing and Communications Perspectives. Edited by A. Swami, Q. Zhao, Y. Hong, and L. Tong

typically correlated, bandwidth and power constraints can be effected by transmitting vectors $\mathbf{y}(n)$ with smaller dimensionality than that of $\mathbf{x}(n)$.

As the discussion in the previous paragraph suggests, the distributed nature of observations couples with stringent bandwidth and power constraints so that estimation in WSNs requires: i) a means of combining local sensor observations in order to reduce their dimensionality while keeping the estimation MSE as small as possible; ii) quantization of the combined observations prior to digital transmission; and iii) construction of estimators based on the quantized digital messages. While addressing these issues jointly is challenging, the present chapter describes recent advances pertaining to all these three requirements.

1.1 Distributed Quantization-Estimation

Consider a WSN consisting of N sensors deployed to estimate a scalar deterministic parameter s. The n^{th} sensor observes a noisy version of s given by

$$x(n) = s + w(n), \qquad n \in [0, N-1] , \tag{1.1}$$

where $w(n)$ denotes zero-mean noise with pdf $p_w(w)$, that is known possibly up to a finite number of unknown parameters. We further assume that $w(n_1)$ is independent of $w(n_2)$ for $n_1 \neq n_2$; i.e., noise variables are independent across sensors.

Due to bandwidth limitations, the observations $x(n)$ have to be quantized and estimation of s can only be based on these quantized values. We will henceforth think of quantization as the construction of a set of indicator variables

$$b_k(n) = \mathbf{1}\{x(n) \in B_k(n)\}, \qquad k \in [1, K] , \tag{1.2}$$

taking the value 1 when $x(n)$ belongs to the region $B_k(n) \subset \mathbf{R}$, and 0 otherwise. Estimation of s will rely on this set of *binary* random variables $\{b_k(n), k \in [1, K]\}_{n=0}^{N-1}$. The latter are Bernoulli distributed with parameters $q_k(n)$ satisfying

$$q_k(n) := \Pr\{b_k(n) = 1\} = \Pr\{x(n) \in B_k(n)\}. \tag{1.3}$$

In the ensuing sections, we will present the Cramér-Rao Lower Bound (CRLB) to benchmark the variance of all unbiased estimators $\hat{s}$ constructed using the binary observations $\{b_k(n), k \in [1, K]\}_{n=0}^{N-1}$. We will further show that it is possible to find maximum likelihood estimators (MLEs) that (at least asymptotically) can achieve the CRLB. Finally, we will reveal that the CRLB based on $\{b_k(n), k \in [1, K]\}_{n=0}^{N-1}$ can come surprisingly close to the clairvoyant CRLB based on $\{x(n)\}_{n=0}^{N-1}$ in certain applications of practical interest.

1.2 Maximum Likelihood Estimation

Let us start by assuming that $p_w(w)$ is known and let $F_w(u) := \int_u^\infty p_w(w)\, dw$ denote the complementary cumulative distribution function (CCDF) of the noise. With the pdf known it suffices to rely on a single region $B_1(n)$ in (1.2) to generate a single bit $b_1(n)$ per sensor, using a threshold τ_c common to all N sensors: $B_1(n) := B_c = (\tau_c, \infty)$, $\forall n$. Based on these binary observations, $b_1(n) := \mathbf{1}\{\mathbf{x}(n) \in (\tau_c, \infty)\}$ received from all N sensors, the fusion center (FC) seeks estimates of s.

An expression for the MLE of s follows readily from the following argument. Using (1.3), we can express the Bernoulli parameter as

$$q_1 = \int_{\tau_c - s}^{\infty} p_w(w) dw = F_w(\tau_c - s). \tag{1.4}$$

On the other hand, it is well known that the MLE of q_1 is given by $\hat{q}_1 = N^{-1}\sum_{n=0}^{N-1} b_1(n)$ (Kay 1993, p. 200). These two facts combined with the invariance property of MLE (Kay 1993, p. 173), readily yield the MLE of s as (Ribeiro and Giannakis 2006a):

$$\hat{s} = \tau_c - F_w^{-1}\left(\frac{1}{N}\sum_{n=0}^{N-1} b_1(n)\right) . \tag{1.5}$$

It can be further shown that the CRLB on the variance of any unbiased estimator $\hat{s}$ based on $b_1(n)_{n=0}^{N-1}$ is (Ribeiro and Giannakis 2006a)

$$\mathrm{var}(\hat{s}) \geq \frac{1}{N}\frac{F_w(\tau_c - s)[1 - F_w(\tau_c - s)]}{p_w^2(\tau_c - s)} \; := \; B(s) \; . \tag{1.6}$$

If the noise is Gaussian and we define the *σ-distance* between the threshold τ_c and the (unknown) parameter s as $\Delta_c := (\tau_c - s)/\sigma$, then (1.6) reduces to

$$B(s) \;=\; \frac{\sigma^2}{N}\frac{2\pi Q(\Delta_c)[1 - Q(\Delta_c]}{e^{-\Delta_c}} \;:=\; \frac{\sigma^2}{N}D(\Delta_c), \tag{1.7}$$

with $Q(u) := (1/\sqrt{2\pi})\int_u^\infty e^{-w^2/2}\,dw$ denoting the Gaussian tail probability function.

The bound $B(s)$ is the variance of $\bar{x} := N^{-1}\sum_{n=0}^{N-1} x(n)$, scaled by the factor $D(\Delta_c)$ – recall that $\mathrm{var}(\bar{x}) = \sigma^2/N$ (Kay 1993, p.31). Optimizing $B(s)$ with respect to Δ_c, yields the optimum at $\Delta_c = 0$ and the minimum CRLB as

$$B_{\min} = \frac{\pi}{2}\frac{\sigma^2}{N}. \tag{1.8}$$

Eq. (1.8) reveals something unexpected: relying on a single bit per $x(n)$, the estimator in (1.5) incurs a minimal (just a $\pi/2$ factor) increase in its variance relative to the clairvoyant $\bar{x}$ which relies on the unquantized data $x(n)$. But this minimal loss in performance corresponds to the ideal choice $\Delta_c = 0$, which implies $\tau_c = s$ and requires perfect knowledge of the unknown s for selecting the quantization threshold τ_c. How do we select τ_c and how much do we lose when the unknown s lies anywhere in $(-\infty, \infty)$, or when s lies in $[S_1, S_2]$, with S_1, S_2 finite and known a priori? Intuition suggests selecting the threshold as close as possible to the unknown parameter s. This can be realized with an iterative estimator $\hat{s}^{(i)}$, which can be formed as in (1.5), using $\tau_c^{(i)} = \hat{s}^{(i-1)}$, the parameter estimate from the previous $(i-1)^{st}$ iteration.

But in the batch formulation considered herein, selecting τ_c is challenging; and a closer look at $B(s)$ in (1.6) will confirm that the loss can be huge if $\tau_c - s \gg 0$. Indeed, as $\tau_c - s \to \infty$ the denominator in (1.6) goes to zero faster than its numerator, since F_w is the integral of the non-negative pdf p_w; and thus, $B(s) \to \infty$ as $\tau_c - s \to \infty$. The implication of the latter is twofold: i) since it shows up in the CRLB, the potentially high variance of estimators based on quantized observations is inherent to the possibly severe bandwidth limitations of the problem itself and is not unique to a particular estimator; ii) for any choice of τ_c, the fundamental performance limits in (1.6) are dictated by the end points $\tau_c - S_1$ and $\tau_c - S_2$ when s is confined to the interval $[S_1, S_2]$. On the other hand, how successful the τ_c selection is depends on the dynamic range $|S_1 - S_2|$ which makes sense because the latter affects the error incurred when quantizing $x(n)$ to $b_1(n)$. Notice that in such joint quantization-estimation problems one faces two sources of error: quantization and noise. To account for both, the proper figure of merit for estimators based on binary observations is what we will term quantization signal-to-noise ratio (Q-SNR):

$$\gamma := \frac{|S_1 - S_2|^2}{\sigma^2}; \tag{1.9}$$

Notice that contrary to common wisdom, the smaller Q-SNR is, the easier it becomes to select τ_c judiciously. Furthermore, the variance increase in (1.6) relative to the variance of the clairvoyant $\bar{x}$ is smaller, for a given σ. This is because as the Q-SNR increases the problem becomes more difficult in general, but the rate at which the variance increases is smaller for the CRLB in (1.6) than for $\text{var}(\bar{x}) = \sigma^2/N$.

1.2.1 Known Noise pdf with Unknown Variance

Perhaps more common than a perfectly known pdf is the case when the noise pdf is known except for its variance $\text{E}[w^2(n)] = \sigma^2$. Introducing the standardized variable $v(n) := w(n)/\sigma$ we write the signal model as

$$x(n) = s + \sigma v(n). \tag{1.10}$$

Let $p_v(v)$ and $F_v(v) := \int_v^\infty p_v(u)du$ denote the known pdf and CCDF of $v(n)$. Note that according to its definition, $v(n)$ has zero mean, $\text{E}[v^2(n)] = 1$, and the pdfs of v and w are related by $p_w(w) = (1/\sigma)p_v(w/\sigma)$. Note also that all two parameter pdfs can be standardized likewise.

To estimate s when σ is also unknown while keeping the bandwidth constraint to 1 bit per sensor, we divide the sensors in two groups each using a different region (i.e., threshold) to define the binary observations:

$$B_1(n) := \begin{cases} (\tau_1, \infty) := B_1, & \text{for } n = 0, \ldots, (N/2) - 1 \\ (\tau_2, \infty) := B_2, & \text{for } n = (N/2), \ldots, N. \end{cases} \tag{1.11}$$

That is, the first $N/2$ sensors quantize their observations using the threshold τ_1, while the remaining $N/2$ sensors rely on the threshold τ_2. Without loss of generality, we assume $\tau_2 > \tau_1$.

The Bernoulli parameters of the resultant binary observations can be expressed as [c.f. (1.4)]:

$$q_1(n) := \begin{cases} F_v\left[\frac{\tau_1 - s}{\sigma}\right] & := q_1 \quad \text{for } n = 0, \ldots, (N/2) - 1, \\ F_v\left[\frac{\tau_2 - s}{\sigma}\right] & := q_2 \quad \text{for } n = (N/2), \ldots, N. \end{cases} \tag{1.12}$$

Given the noise independence across sensors, the MLEs of q_1, q_2 can be found, respectively, as

$$\hat{q}_1 = \frac{2}{N} \sum_{n=0}^{N/2-1} b_1(n), \qquad \hat{q}_2 = \frac{2}{N} \sum_{n=N/2}^{N-1} b_1(n). \tag{1.13}$$

Mimicking (1.5), we can invert F_v in (1.12) and invoke the invariance property of MLEs to obtain the MLE $\hat{s}$ in terms of $\hat{q}_1$ and $\hat{q}_2$. This estimator is given in the following proposition along with its CRLB (Ribeiro and Giannakis 2006b).

Proposition 1.2.1 *Consider estimating s in* (1.10), *based on binary observations constructed from the regions defined in* (1.11).

(a) *The MLE of s is*

$$\hat{s} = \frac{F_v^{-1}(\hat{q}_2)\tau_1 - F_v^{-1}(\hat{q}_1)\tau_2}{F_v^{-1}(\hat{q}_2) - F_v^{-1}(\hat{q}_1)}, \tag{1.14}$$

with F_v^{-1} denoting the inverse function of F_v, and $\hat{q}_1$, $\hat{q}_2$ given by (1.13).

(b) *The variance of any unbiased estimator of s, $\text{var}(\hat{s})$, based on $\{b_1(n)\}_{n=0}^{N-1}$ is bounded by*

$$B(s) := \frac{2\sigma^2}{N} \left(\frac{\Delta_1 \Delta_2}{\Delta_2 - \Delta_1}\right)^2 \left[\frac{q_1(1-q_1)}{p_v^2(\Delta_1)\Delta_1^2} + \frac{q_2(1-q_2)}{p_v^2(\Delta_2)\Delta_2^2}\right] \tag{1.15}$$

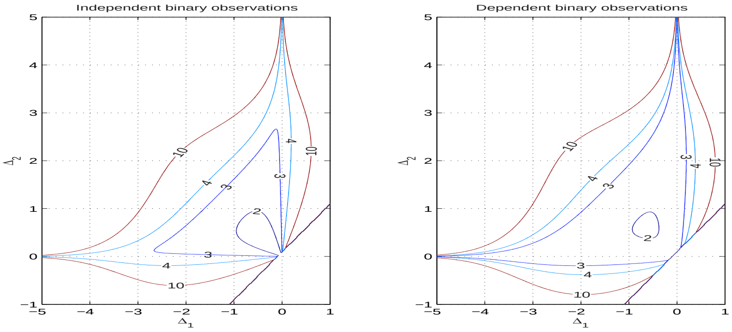

Figure 1.1 Per bit CRLB when the binary observations are independent and dependent, respectively. In both cases, the variance increase with respect to the sample mean estimator is small when the σ-distances are close to 1, being slightly better for the case of dependent binary observations (Gaussian noise).

where q_k is given by (1.12)*, and*

$$\Delta_k := \frac{\tau_k - s}{\sigma}, \quad k = 1, 2, \tag{1.16}$$

is the σ-distance between s and the threshold τ_k.

Eq. (1.15) is reminiscent of (1.6), suggesting that the variances of the estimators they bound are related. This implies that even when the known noise pdf contains unknown parameters, the variance of $\hat{s}$ can come close to the variance of the clairvoyant estimator $\bar{x}$, provided that the thresholds τ_1, τ_2 are chosen close to s relative to the noise standard deviation (so that Δ_1, Δ_2, and $\Delta_2 - \Delta_1$ in (1.16) are ≈ 1). For the Gaussian pdf, Fig. 1.1 shows the contour plot of $B(s)$ in (1.15) normalized by $\sigma^2/N := \mathrm{var}(\bar{x})$. Notice that in the low Q-SNR regime $\Delta_1, \Delta_2 \approx 1$, and the relative variance increase $B(s)/\mathrm{var}(\bar{x})$ is less than 3. This is illustrated by the simulations shown in Fig. 1.2 for two different sets of σ-distances, Δ_1, Δ_2, corroborating the values predicted by (1.15) and the fact that the performance loss with respect to the clairvoyant sample mean estimator, $\bar{x}$, is indeed small.

Dependent binary observations

In the previous subsection, we restricted the sensors to transmit only 1 bit per $x(n)$ datum, and divided the sensors in two classes each quantizing $x(n)$ using a different threshold. A related approach is to let each sensor use two thresholds:

$$\begin{aligned} B_1(n) &:= B_1 = (\tau_1, \infty), \quad n = 0, 1, \ldots, N-1, \\ B_2(n) &:= B_2 = (\tau_2, \infty), \quad n = 0, 1, \ldots, N-1 \end{aligned} \tag{1.17}$$

where $\tau_2 > \tau_1$. We define the per sensor vector of binary observations $\mathbf{b}(n) := [b_1(n), b_2(n)]^T$, and the vector Bernoulli parameter $\mathbf{q} := [q_1(n), q_2(n)]^T$, whose components are as in (1.12).

Note the subtle differences between (1.11) and (1.17). While each of the N sensors generates 1 binary observation according to (1.11), each sensor creates 2 binary observations as per (1.17). The total number of bits from all sensors in the former case is N, but in the latter $N \log_2 3$, since our constraint $\tau_2 > \tau_1$ implies that the realization $\mathbf{b} = (0, 1)$ is impossible. In addition, all bits in the former case are independent, whereas correlation is present in the latter since $b_1(n)$ and $b_2(n)$ come from the same $x(n)$. Even though one would expect this correlation to complicate matters, a

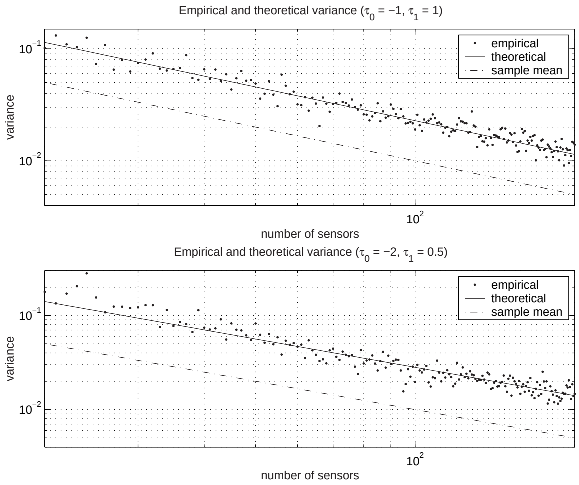

Figure 1.2 Noise of unknown power estimator. The simulation corroborates the close to clairvoyant variance prediction of (1.15) ($\sigma = 1$, $s = 0$, Gaussian noise).

property of the binary observations defined as per (1.17), summarized in the next lemma, renders estimation of s based on them feasible.

Lemma 1.2.2 *The MLE of* $\mathbf{q} := (q_1(n), q_2(n))^T$ *based on the binary observations* $\{\mathbf{b}(n)\}_{n=0}^{N-1}$ *constructed according to* (1.17) *is given by*

$$\hat{\mathbf{q}} = \frac{1}{N}\sum_{n=0}^{N-1}\mathbf{b}(n). \tag{1.18}$$

Interestingly, (1.18) coincides with (1.13), proving that the corresponding estimators of s are identical; i.e., (1.14) yields also the MLE $\hat{s}$ even in the correlated case. However, as the following proposition asserts, correlation affects the estimator's variance and the corresponding CRLB (Ribeiro and Giannakis 2006b).

Proposition 1.2.3 *Consider estimating* s *in* (1.10)*, when* σ *is unknown, based on binary observations constructed from the regions defined in* (1.17)*. The variance of any unbiased estimator of* s*,* $\mathrm{var}(\hat{s})$*, based on* $\{b_1(n), b_2(n)\}_{n=0}^{N-1}$ *is bounded by*

$$B_D(s) := \frac{\sigma^2}{N}\left(\frac{\Delta_1\Delta_2}{\Delta_2 - \Delta_1}\right)^2\left[\frac{q_1\,(1-q_1)}{p_v^2(\Delta_1)\Delta_1^2} + \frac{q_2\,(1-q_2)}{p_v^2(\Delta_2)\Delta_2^2} - \frac{q_2\,(1-q_1)}{p_v(\Delta_1)p(\Delta_2)\Delta_1\Delta_2}\right], \tag{1.19}$$

where the subscript D *in* $B_D(s)$ *is used as a mnemonic for the dependent binary observations this estimator relies on [c.f.* (1.15)*].*

Unexpectedly, (1.19) is similar to (1.15). Actually, a fair comparison between the two requires compensating for the difference in the total number of bits used in each case. This can be accomplished by introducing the per-bit CRLBs for the independent and correlated cases respectively,

$$C(s) = NB(s), \quad C_D(s) = N\log_2(3)B_D(s)\ , \tag{1.20}$$

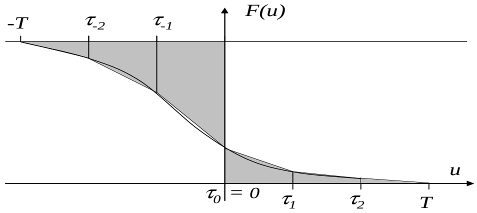

Figure 1.3 When the noise pdf is unknown numerically integrating the CCDF using the trapezoidal rule yields an approximation of the mean.

which lower bound the corresponding variances achievable by the transmission of a single bit.

Evaluation of $C(s)/\sigma^2$ and $C_D(s)/\sigma^2$ follows from (1.15), (1.19) and (1.20) and is depicted in Fig. 1.1 for Gaussian noise and σ-distances Δ_1, Δ_2 having amplitude as large as 5. Somewhat surprisingly, both approaches yield very similar bounds with the one relying on dependent binary observations being slightly better in the achievable variance; or correspondingly, in requiring a smaller number of sensors to achieve the same CRLB.

1.3 Unknown noise pdf

In certain applications it may not be reasonable to assume knowledge about the noise pdf $p_w(w)$. These cases require *nonparametric* approaches as the one pursued in this section.

We assume that $p_w(w)$ has zero mean so that s in (1.1) is identifiable. Let $p_x(x)$ and $F_x(x)$ denote the pdf and CCDF of the observations $x(n)$. As s is the mean of $x(n)$, we can write

$$s := \int_{-\infty}^{+\infty} x p_x(x)\, dx = -\int_{-\infty}^{+\infty} x \frac{\partial F_x(x)}{\partial x}\, dx = \int_0^1 F_x^{-1}(v)\, dv\ , \tag{1.21}$$

where in establishing the second equality we used the fact that the pdf is the negative derivative of the CCDF, and in the last equality we introduced the change of variables $v = F_x(x)$. But note that the integral of the inverse CCDF can be written in terms of the integral of the CCDF as (see also Fig. 1.3)

$$s = -\int_{-\infty}^{0} [1 - F_x(u)]\, du + \int_0^{+\infty} F_x(u)\, du, \tag{1.22}$$

allowing one to express the mean s of $x(n)$ in terms of its CCDF. To avoid carrying out integrals with infinite range, let us assume that $x(n) \in (-T, T)$ which is always practically satisfied for T sufficiently large, so that we can rewrite (1.22) as

$$s = \int_{-T}^{T} F_x(u)\, du\ -\ T. \tag{1.23}$$

Numerical evaluation of the integral in (1.23) can be performed using a number of known techniques. Let us consider an ordered set of interior points $\{\tau_k\}_{k=1}^{K}$ along with end-points $\tau_0 = -T$ and $\tau_{K+1} = T$. Relying on the fact that $F_x(\tau_0) = F_x(-T) = 1$ and $F_x(\tau_{K+1}) = F_x(T) = 0$, application of the trapezoidal rule for numerical integration yields (see also Fig. 1.3)

$$s = \frac{1}{2}\sum_{k=1}^{K} (\tau_{k+1} - \tau_{k-1}) F_x(\tau_k)\ -\ T + e_a, \tag{1.24}$$

with e_a denoting the approximation error. Certainly, other methods like Simpson's rule, or the broader class of Newton-Cotes formulas, can be used to further reduce e_a.

Whichever the choice, the key is that binary observations constructed from the region $B_k := (\tau_k, \infty)$ have Bernoulli parameters

$$q_k := \Pr\{x(n) > \tau_k\} = F_x(\tau_k). \tag{1.25}$$

Inserting the nonparametric estimators $\hat{F}_x(\tau_k) = \hat{q}_k$ in (1.24), our parameter estimator when the noise pdf is unknown takes the form:

$$\hat{s} = \frac{1}{2}\sum_{k=1}^{K} \hat{q}_k(\tau_{k+1} - \tau_{k-1}) \; - \; T. \tag{1.26}$$

Since $\hat{q}_k$'s are unbiased, (1.24) and (1.26) imply that $\mathrm{E}(\hat{s}) = s + e_a$. Being biased, the proper performance indicator for $\hat{s}$ in (1.26) is the mean squared error (MSE), not the variance.

Maintaining the bandwidth constraint of 1 bit per sensor (i.e., $K = 1$), we divide the N sensors in K subgroups containing N/K sensors each, and define the regions

$$B_1(n) := B_k = (\tau_k, \infty), \;\; n = (k-1)(N/K), \ldots, k(N/K) - 1; \tag{1.27}$$

Region $B_1(n)$ will be used by sensor n to construct and transmit the binary observation $b_1(n)$. Herein, the unbiased estimators of q_k are

$$\hat{q}_k = \frac{1}{(N/K)} \sum_{n=(k-1)(N/K)}^{k(N/K)-1} b_1(n), \quad k = 1, \ldots, K, \tag{1.28}$$

and are used in (1.26) to estimate s. It is easy to verify that $\mathrm{var}(\hat{q}_k) = q_k(1-q_k)/(N/K)$, and that $\hat{q}_{k_1}$ and $\hat{q}_{k_2}$ are independent for $k_1 \neq k_2$.

The resultant MSE, $\mathrm{E}[(s - \hat{s})^2]$, can be bounded as follows (Ribeiro and Giannakis 2006b).

Proposition 1.3.1 *Consider $\hat{s}$ given by* (1.26)*, with $\hat{q}_k$ as in* (1.28)*. Assume that for T sufficiently large and known $p_x(x) = 0$, for $|x| \geq T$, the noise pdf has bounded derivative $\dot{p}_w(u) := \partial p_w(w)/\partial w$; and define $\tau_{\max} := \max_k\{\tau_{k+1} - \tau_k\}$ and $\dot{p}_{\max} := \max_{u \in (-T,T)} \{\dot{p}_w(u)\}$. The MSE is given by*

$$\mathrm{E}[(s - \hat{s})^2] = |e_a|^2 + \mathrm{var}(\hat{s}), \tag{1.29}$$

with the approximation error e_a and $\mathrm{var}(\hat{s})$, satisfying

$$|e_a| \leq \frac{T\dot{p}_{\max}}{6}\tau_{\max}^2, \tag{1.30}$$

$$\mathrm{var}(\hat{s}) = \sum_{k=1}^{K} \frac{(\tau_{k+1} - \tau_{k-1})^2}{4} \frac{q_k(1-q_k)}{N/K}, \tag{1.31}$$

with $\{\tau_k\}_{k=1}^K$ a grid of thresholds in $(-T, T)$ and $\{q_k\}_{k=1}^K$ as in (1.25)*.*

Note from (1.31) that the larger contributions to $\mathrm{var}(\hat{s})$ occur when $q_k \approx 1/2$, since this value maximizes the coefficients $q_k(1 - q_k)$; for a symmetric noise pdf, this happens when the thresholds satisfy $\tau_k \approx s$ [c.f. (1.25)]. Thus, as with the case where the noise pdf is known, when s belongs to an a priori known interval $[s_1, s_2]$, this knowledge must be exploited in selecting thresholds around the likeliest values of s.

On the other hand, note that the $\mathrm{var}(\hat{s})$ term in (1.29) will dominate $|e_a|^2$ because $|e_a|^2 \propto \tau_{\max}^4$ as per (1.30). To clarify this point, consider an equispaced grid of thresholds with $\tau_{k+1} - \tau_k = \tau = \tau_{\max}$, $\forall k$, such that $\tau_{\max} = 2T/(K+1) < 2T/K$. Using the (loose) bound $q_k(1-q_k) \leq 1/4$, the MSE is bounded by [c.f. (1.29) - (1.31)]

$$\mathrm{E}[(s-\hat{s})^2] < \frac{4T^6 \dot{p}_{\max}^2}{9K^4} + \frac{T^2}{N}. \tag{1.32}$$

The bound in (1.32) is minimized by selecting $K = N$, which amounts to having *each sensor use a different region* to construct its binary observation. In this case, $|e_a|^2 \propto N^{-4}$ and its effect becomes practically negligible. Moreover, most pdfs have relatively small derivatives; e.g., for the Gaussian pdf we have $\dot{p}_{\max} = (2\pi e \sigma^4)^{-1/2}$. The integration error can be further reduced by resorting to a more powerful numerical integration method, although its difference with respect to the trapezoidal rule will not have noticeable impact in practice.

Since $K = N$, the selection $\tau_{k+1} - \tau_k = \tau$, $\forall k$, yields

$$\hat{s} \;=\; \tau \sum_{n=0}^{N-1} b_1(n) - T \;=\; T\left[\frac{2}{N+1}\sum_{n=0}^{N-1} b_1(n) - 1\right], \tag{1.33}$$

that *does not require knowledge of the threshold* used to construct the binary observations at the FC of a WSN. This feature allows each sensor to randomly select its threshold without using values pre-assigned by the FC; see also (Luo 2005a) for related random quantization algorithms which also yielded *universal* (in the noise variance) parameter estimators based on severely quantized WSN data.

Remark 1 While $e_a^2 \propto T^6$ seems to dominate $\mathrm{var}(\hat{s}) \propto T^2$ in (1.32), this is not true for the operational low-to-medium Q-SNR range for distributed estimators based on binary observations. This is because the support $2T$ over which $F_x(x)$ in (1.23) is non-zero depends on σ and the dynamic range $|S_1 - S_2|$ of the parameter s. And as the Q-SNR decreases, $T \propto \sigma$. But since $\dot{p}_{\max} \propto \sigma^{-2}$, $e_a^2 \propto \sigma^2/N^4$ which is negligible when compared to the term $\mathrm{var}(\hat{s}) \propto \sigma^2/N$.

Apart from providing useful bounds on the finite-sample performance, eqs. (1.30), (1.31), and (1.32) establish asymptotic optimality of the $\hat{s}$ estimators in (1.26) and (1.33) as summarized in the following:

Corollary 1.3.2 *Under the assumptions of Propositions 1.3.1 and the conditions: i) $\tau_{\max} \propto K^{-1}$; and ii) $T^2/N, T^6/K^4 \to 0$ as $T, K, N \to \infty$, the estimators $\hat{s}$ in* (1.26) *and* (1.33) *are asymptotically (as $K, N \to \infty$) unbiased and consistent in the mean-square sense.*

The estimators in (1.26) and (1.33) are consistent even if the support of the data *pdf* is infinite, as long as we guarantee a proper rate of convergence relative to the number of sensors and thresholds.

Remark 2 To compare the estimators in (1.5) and (1.33), consider that $s \in [S_1, S_2] = [-\sigma, \sigma]$, and that the noise is Gaussian with variance σ^2, yielding a Q-SNR $\gamma = 4$. No estimator can have variance smaller than $\mathrm{var}(\bar{x}) = \sigma^2/N$; however, for the (medium) $\gamma = 4$ Q-SNR value they can come close. For the known pdf estimator in (1.5), the variance is $\mathrm{var}(\hat{s}) \approx 2\sigma^2/N$. The unknown pdf estimator in (1.33) requires an assumption about the essentially non-zero support of the Gaussian pdf. If we suppose that the noise pdf is non-zero over $[-2\sigma, 2\sigma]$, the corresponding variance becomes $\mathrm{var}(\hat{s}) \approx 9\sigma^2/N$. The penalties due to the transmission of a single bit per sensor with respect to $\bar{x}$ are approximately 2 and 9. While the increasing penalty is expected as the uncertainty about the noise pdf increases, the relatively small loss is rather unexpected.

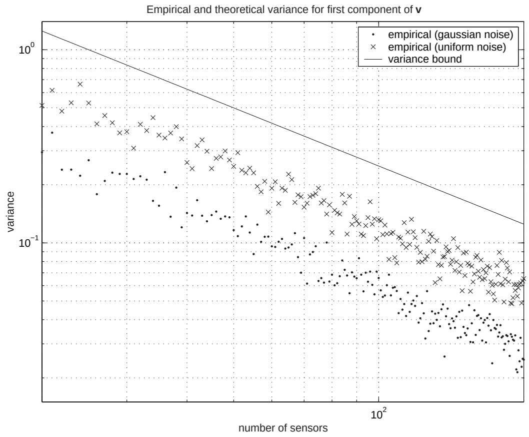

Figure 1.4 The variance of the estimators in (1.5) and (1.33) are close to the sample mean estimator variance ($\sigma^2 := \mathrm{E}[w^2(n)] = 1, T = 3, s \in [-1, 1]$).

Fig. 1.4 depicts theoretical bounds and simulated variances for the estimators (1.5) and (1.33) for an example Q-SNR $\gamma = 4$. The sample mean estimator variance, $\mathrm{var}(\bar{x}) = \sigma^2/N$, is also depicted for comparison purposes. The simulations corroborate the implications of Remark 3, reinforcing the assertion that for low to medium Q-SNR problems quantization to a single bit per observation leads to minimal losses in variance performance. Note that for this particular example, the unknown pdf variance bound, (1.32), overestimates the variance by a factor of roughly 1.2 for the uniform case and roughly 2.6 for the Gaussian case.

1.3.1 Lower bound on the MSE

In Section 1.2 we derived the CRLB offering the fundamental *lower* bound on the achievable variance and the MLE that approaches this bound as N increases. In contrast, (1.32) is an *upper* bound on the MSE of the estimator in (1.33). The counterpart of the CRLB for estimation based on binary observations when the pdf is unknown is a lower bound in the MSE achievable by any estimator.

To obtain this bound we start from the CRLB when the noise pdf is known that we introduced in (1.6). We then maximize this CRLB with respect to the noise pdf and the local quantization rules to obtain a lower bound on the MSE performance of any estimator when the pdf is unknown. The result is summarized in the following proposition (Xiao *et al.* 2005a).

Proposition 1.3.3 *Consider the signal model in* (1.1)*; $x(n)$ observations belonging to the interval $(-T, T)$; i.e., $x(n) \in [-T, T]$; and let each sensor communicate one binary observation $b(n)$ as per* (1.2)*. Then, for any estimator $\hat{s}$ of s relying on $\{b(n)\}_{n=0}^{N-1}$ there exists a noise pdf such that*

$$\mathrm{E}[(s - \hat{s})^2] \geq \frac{T^2}{4N}. \tag{1.34}$$

Proposition 1.3.3 implies that no estimator based on quantizated samples down to a single bit per sensor can attain an MSE smaller than $T^2/4N$. Comparing (1.32) with (1.34) we deduce that the estimator in (1.33) is optimal up to a constant factor of 4.

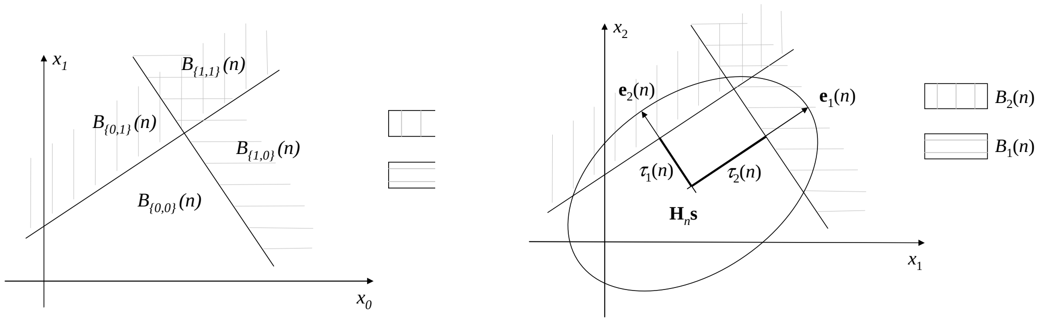

Figure 1.5 (Left): The vector of binary observations $\mathbf{b}$ takes on the value $\{y_1, y_2\}$ if and only if $x(n)$ belongs to the region $B_{\{y_1,y_2\}}$; (Right): Selecting the regions $B_k(n)$ perpendicular to the covariance matrix eigenvectors results in independent binary observations.

1.4 Estimation of Vector Parameters

Consider now the case of a physical phenomenon characterized by a set of p parameters that we lump in to the vector $\mathbf{s} := [s_1, \ldots, s_p]^T$. As before, we wish to find $\mathbf{s}$, by deploying a WSN composed of N sensors $\{S_n\}_{n=0}^{N-1}$, with each sensor observing $\mathbf{s}$ through a linear transformation

$$\mathbf{x}(n) = \mathbf{H}_n\mathbf{s} + \mathbf{w}(n), \tag{1.35}$$

where $\mathbf{x}(n) := [x_1(n), \ldots, x_M(n)]^T \in \mathbf{R}^M$ is the measurement vector at sensor S_n, $\mathbf{w}(n) \in \mathbf{R}^M$ is zero-mean additive noise with pdf $p_{\mathbf{w}}(\mathbf{w})$ and the matrices $\mathbf{H}_n \in \mathbf{R}^{M\times P}$.

As in (1.2), we define the binary observation $b_k(n)$ as the indicator function of $\mathbf{x}(n)$ belonging to the region $B_k(n) \subset \mathbf{R}^M$:

$$b_k(n) = \mathbf{1}\{\mathbf{x}(n) \in B_k(n)\}, \qquad k \in [1, K]\ , \tag{1.36}$$

We then define the per sensor vector of binary observations $\mathbf{b}(n) := [b_1(n), \ldots, b_K(n)]^T$, and note that since its entries are binary, realizations $\mathbf{y}$ of $\mathbf{b}(n)$ belong to the set

$$\mathcal{B} := \{\boldsymbol{\beta} \in \mathbf{R}^K \mid [\boldsymbol{\beta}]_k \in \{0,1\},\ k \in [1, K]\}, \tag{1.37}$$

where $[\boldsymbol{\beta}]_k$ denotes the k^{th} component of $\boldsymbol{\beta}$. With each $\boldsymbol{\beta} \in \mathcal{B}$ and each sensor we now associate the region

$$\mathbf{B}_{\boldsymbol{\beta}}(n) := \bigcap_{[\boldsymbol{\beta}]_k=1} B_k(n) \bigcap_{[\boldsymbol{\beta}]_k=0} \bar{B}_k(n), \tag{1.38}$$

where $\bar{B}_k(n)$ denotes the set-complement of $B_k(n)$ in $\mathbf{R}^M$. Note that the definition in (1.38) implies that $x(n) \in \mathbf{B}_\beta(n)$ if and only if $\mathbf{b}(n) = \boldsymbol{\beta}$; see also Fig. 1.5 (Left) for an illustration in $\mathbf{R}^2$ ($M = 2$). The corresponding probabilities are

$$q_\beta(n) := \Pr\{\mathbf{b}(n) = \boldsymbol{\beta}\} = \int_{\mathbf{B}_{\boldsymbol{\beta}}(n)} p_{\mathbf{w}}[\mathbf{u} - \mathbf{H}_n\mathbf{s}]\, d\mathbf{u}. \tag{1.39}$$

Using definitions (1.39) and (1.37), we can write the pertinent log-likelihood function as

$$L(\mathbf{s}) = \sum_{n=0}^{N-1} \sum_{\boldsymbol{\beta}\in\mathcal{B}} \delta(\mathbf{b}(n) - \boldsymbol{\beta}) \ln q_{\boldsymbol{\beta}}(n), \tag{1.40}$$

and the MLE of $\mathbf{s}$ as

$$\hat{\mathbf{s}} = \arg\max_{\mathbf{s}} L(\mathbf{s}) \ . \tag{1.41}$$

The nonlinear search needed to obtain $\hat{\mathbf{s}}$ could be challenging. Fortunately, as the following proposition asserts, under certain conditions that are usually met in practice, $L(\mathbf{s})$ is concave which implies that computationally efficient search algorithms can be invoked to find its global maximum (Ribeiro and Giannakis 2006b).

Proposition 1.4.1 *If the MLE problem in* (1.41) *satisfies the conditions:*

[c1] *The noise pdf $p_{\mathbf{w}}(\mathbf{w})$ is log-concave (Boyd and Vandenberghe 2004, p.104)*

[c2] *The regions $B_k(n)$ are chosen as half-spaces.*

then $L(\mathbf{s})$ in (1.40) *is a concave function of* $\mathbf{s}$.

Note that [c1] is satisfied by common noise pdfs, including the multivariate Gaussian (Boyd and Vandenberghe 2004, p.104). On the other hand, [c2] places a constraint in the regions defining the binary observations, which is simply up to the designer's choice. The merits of having a concave log-likelihood function are summarized in the following remark.

Remark 3 The numerical search needed to obtain $\hat{\mathbf{s}}$ could be challenged either by the multimodal nature of $L(\mathbf{s})$ or by numerical ill-conditioning caused by e.g., saddle points. But when the log-concavity conditions in Proposition 1.4.1 are satisfied, computationally efficient search algorithms like e.g., Newton's method are guaranteed to converge to the global maximum (Boyd and Vandenberghe 2004, Chap. 2).

1.4.1 Colored Gaussian Noise

Analyzing the performance of the MLE in (1.41) is only possible asymptotically (as N or SNR go to infinity). Notwithstanding, when the noise is Gaussian, simplifications render variance analysis tractable and lead to interesting guidelines for constructing the estimator $\hat{\mathbf{s}}$.

Restrict $p_{\mathbf{w}}(\mathbf{w})$ to the class of multivariate Gaussian pdfs, and let $\mathbf{C}(n)$ denote the noise covariance matrix at sensor n. Assume that $\{\mathbf{C}(n)\}_{n=0}^{N-1}$ are known and let $\{(\mathbf{e}_m(n), \sigma_m^2(n))\}_{m=1}^{M}$ be the set of eigenvectors and associated eigenvalues

$$\mathbf{C}(n) = \sum_{m=1}^{M} \sigma_m^2(n)\mathbf{e}_m(n)\mathbf{e}_m^T(n). \tag{1.42}$$

For each sensor, we define a set of $K = M$ regions $B_k(n)$ as half-spaces whose borders are hyperplanes perpendicular to the covariance matrix eigenvectors; i.e.,

$$B_k(n) = \{\mathbf{x} \in \mathbf{R}^M \mid \mathbf{e}_k^T(n)\mathbf{x} \geq \tau_k(n)\}, \quad k = 1, \dots, K = M, \tag{1.43}$$

Fig (1.5) (Right) depicts the regions $B_k(n)$ in (1.43) for $M = 2$. Note that since each entry of $\mathbf{x}(n)$ offers a distinct scalar observation, the selection $K = M$ amounts to a bandwidth constraint of 1 *bit per sensor per dimension*.

The rationale behind this selection of regions is that the resultant binary observations $b_k(n)$ are independent, meaning that $\Pr\{b_{k_1}(n)b_{k_2}(n)\} = \Pr\{b_{k_1}(n)\}\ \Pr\{b_{k_2}(n)\}$ for $k_1 \neq k_2$. As a result, we have a total of MN independent binary observations to estimate $\mathbf{s}$.

Herein, the Bernoulli parameters $q_k(n)$ take on a particularly simple form in terms of the Gaussian tail function

$$q_k(n) = \int_{\mathbf{e}_k^T(n)\mathbf{u} \geq \tau_k(n)} p_{\mathbf{w}}(\mathbf{u} - \mathbf{H}_n\mathbf{s})\, d\mathbf{u} = Q\left(\frac{\tau_k(n) - \mathbf{e}_k^T(n)\mathbf{H}_n\mathbf{s}}{\sigma_k(n)}\right) := Q[\Delta_k(n)], \tag{1.44}$$

where we introduced the *σ-distance* between $\mathbf{H}_n\mathbf{s}$ and the corresponding threshold $\Delta_k(n) := [\tau_k(n) - \mathbf{e}_k^T(n)\mathbf{H}_n\mathbf{s}]/\sigma_k(n)$.

Due to the independence among binary observations we have $p(\mathbf{b}(n)) = \prod_{k=1}^{K} [q_k(n)]^{b_k(n)} [1 - q_k(n)]^{1-b_k(n)}$, leading to

$$L(\mathbf{s}) = \sum_{n=0}^{N-1}\sum_{k=1}^{K} b_k(n)\ln q_k(n) + [1 - b_k(n)]\ln[1 - q_k(n)], \tag{1.45}$$

whose NK *independent* summands replace the $N2^K$ *dependent* terms in (1.40).

Since the regions $B_k(n)$ are half-spaces, Proposition 1.4.1 applies to the maximization of (1.45) and guarantees that the numerical search for the $\hat{\mathbf{s}}$ estimator in (1.45) is well-conditioned and will converge to the global maximum. More important, it will turn out that these regions render finite sample performance analysis of the MLE in (1.41), tractable. In particular, it is possible to derive a closed-form expression for the Fisher Information Matrix (FIM) (Kay 1993, p.44), as we outline next; see (Ribeiro and Giannakis 2006b) for detailed derivations.

Proposition 1.4.2 *The FIM,* $\mathbf{I}$, *for estimating* $\mathbf{s}$ *based on the binary observations obtained from the regions defined in* (1.43)*, is given by*

$$\mathbf{I} = \sum_{n=0}^{N-1} \mathbf{H}_n^T \left[\sum_{k=1}^{K} \frac{e^{-\Delta_k^2(n)}\mathbf{e}_k(n)\mathbf{e}_k^T(n)}{2\pi\sigma_k^2(n)Q(\Delta_k(n))[1 - Q(\Delta_k(n))]}\right]\mathbf{H}_n. \tag{1.46}$$

Inspection of (1.46) shows that the variance of the MLE in (1.41) depends on the signal function containing the parameter of interest (via $\mathbf{H}_n$), the noise structure and power (via the eigenvalues and eigenvectors), and the selection of the regions $B_k(n)$ (via the σ-distances). Among these three factors only the last one is inherent to the bandwidth constraint, the other two being common to the estimator that is based on the original $\mathbf{x}(n)$ observations.

The last point is clarified if we consider the FIM $\mathbf{I}_x$ for estimating $\mathbf{s}$ given the unquantized vector $\mathbf{x}(n)$. This matrix can be shown to be ((Ribeiro and Giannakis 2006b, Appendix. D)),

$$\mathbf{I}_x = \sum_{n=0}^{N-1} \mathbf{H}_n^T \left[\sum_{m=1}^{M} \frac{\mathbf{e}_m(n)\mathbf{e}_m^T(n)}{\sigma_m^2(n)}\right]\mathbf{H}_n^T. \tag{1.47}$$

If we define the equivalent noise powers as

$$\rho_k^2(n) := \frac{2\pi Q(\Delta_k(n))[1 - Q(\Delta_k(n))]}{e^{-\Delta_k^2(n)}}\sigma_k^2(n), \tag{1.48}$$

we can rewrite (1.46) in the form

$$\mathbf{I} = \sum_{n=0}^{N-1} \mathbf{H}_n^T \left[\sum_{k=1}^{K} \frac{\mathbf{e}_k(n)\mathbf{e}_k^T(n)}{\rho_k^2(n)}\right]\mathbf{H}_n^T, \tag{1.49}$$

which except for the noise powers has form identical to (1.47). Thus, comparison of (1.49) with (1.47) reveals that from a performance perspective, *the use of binary observations is equivalent to an*

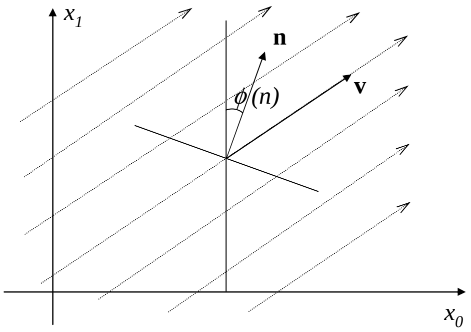

Figure 1.6 The vector flow $\mathbf{v}$ incises over a certain sensor capable of measuring the normal component of $\mathbf{v}$.

increase in the noise variance from $\sigma_k^2(n)$ to $\rho_k^2(n)$, while the rest of the problem structure remains unchanged. Since we certainly want the equivalent noise increase to be as small as possible, minimizing (1.48) over $\Delta_k(n)$ calls for this distance to be set to zero, or equivalently, to select thresholds $\tau_k(n) = \mathbf{e}_k^T(n)\mathbf{H}_n\mathbf{s}$. In this case, the equivalent noise power is

$$\rho_k^2(n) = \frac{\pi}{2}\sigma_k^2(n). \tag{1.50}$$

Surprisingly, even in the vector case a judicious selection of the regions $B_k(n)$ can result in a very small penalty ($\pi/2$) in terms of the equivalent noise increase. Similar to Section 1.2, we can thus claim that while requiring the transmission of 1 bit per sensor per dimension, the variance of the MLE in (1.41), based on $\{\mathbf{b}(n)\}_{n=0}^{N-1}$, yields a variance close to the clairvoyant estimator's variance – which is based on $\{\mathbf{x}(n)\}_{n=0}^{N-1}$ – for low-to-medium Q-SNR problems.

Example 1.4.3 *Suppose we wish to estimate a vector flow using incidence observations. With reference to Fig. 1.6, consider the flow vector* $\mathbf{v} := (v_0, v_1)^T$, *and a sensor positioned at an angle* $\phi(n)$ *with respect to a known reference direction. We will rely on a set of so called incidence observations* $\{x(n)\}_{n=0}^{N-1}$ *measuring the component of the flow normal to the corresponding sensor*

$$x(n) := \langle \mathbf{v}, \mathbf{n} \rangle + w(n) = v_0 \sin[\phi(n)] + v_1 \cos[\phi(n)] + w(n), \tag{1.51}$$

where $\langle , \rangle$ *denotes inner product,* $w(n)$ *is zero-mean AWGN, and* $n = 0, 1, \ldots, N-1$ *is the sensor index. The model* (1.51) *applies to the measurement of hydraulic fields, pressure variations induced by wind and radiation from a distant source (Mainwaring et al. 2002).*

Estimating $\mathbf{v}$ *fits the framework presented in this section requiring the transmission of a single binary observation per sensor,* $b_1(n) = \mathbf{1}\{x(n) \geq \tau_1(n)\}$. *The FIM in* (1.49) *is easily found to be*

$$\mathbf{I} = \sum_{n=0}^{N-1} \frac{1}{\rho_1^2(n)} \begin{pmatrix} \sin^2[\phi(n)] & \sin[\phi(n)]\cos[\phi(n)] \\ \sin[\phi(n)]\cos[\phi(n)] & \cos^2[\phi(n)] \end{pmatrix}. \tag{1.52}$$

Furthermore, since $x(n)$ *in* (1.51) *is linear in* $\mathbf{v}$ *and the noise pdf is log-concave (Gaussian) the log-likelihood function is concave as asserted by Proposition 1.4.1.*

Suppose that we are able to place the thresholds optimally as implied by $\tau_1(n) = v_0 \sin[\phi(n)] + v_1 \cos[\phi(n)]$, *so that* $\rho_1^2(n) = (\pi/2)\sigma^2$. *If we also make the reasonable assumption that the angles are random and uniformly distributed,* $\phi(n) \sim U[-\pi, \pi]$, *then the average FIM turns out to be:*

$$\bar{\mathbf{I}} = \frac{2}{\pi\sigma^2} \begin{pmatrix} N/2 & 0 \\ 0 & N/2 \end{pmatrix}. \tag{1.53}$$

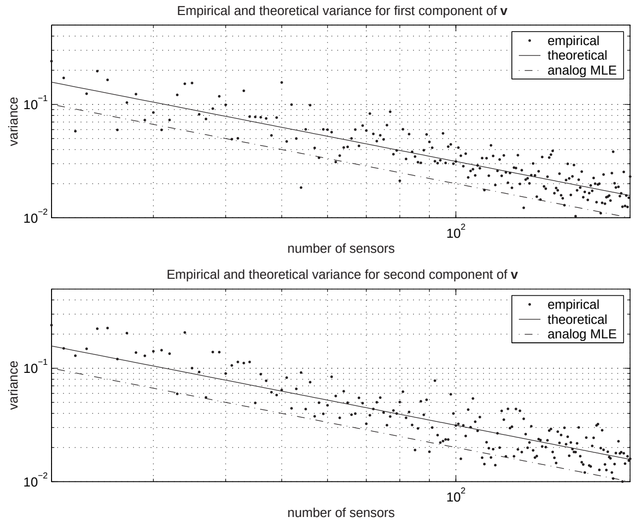

Figure 1.7 Average variance for the components of $\mathbf{v}$. The empirical as well as the bound (1.54) are compared with the analog observations based MLE ($\mathbf{v} = (1, 1)$, $\sigma = 1$).

But according to the law of large numbers $\mathbf{I} \approx \bar{\mathbf{I}}$*, and the estimation variance will be approximately*

$$\mathrm{var}(v_0) = \mathrm{var}(v_1) = \frac{\pi\sigma^2}{N}. \tag{1.54}$$

Fig. 1.7 depicts the bound (1.54)*, as well as the simulated variances* $\mathrm{var}(\hat{v}_0)$ *and* $\mathrm{var}(\hat{v}_1)$ *in comparison with the clairvoyant MLE based on* $\{x(n)\}_{n=0}^{N-1}$*, corroborating our analytical expressions.*

1.5 Maximum a Posteriori Probability Estimation

The parameter of interest $\mathbf{s}$ was so far assumed deterministic. Consequently, the MLE was considered as the optimum estimator and the CRLB as the ultimate performance limit. An alternative formulation is to use available a priori knowledge to model $\mathbf{s}$ as a random vector parameter with a priory pdf $p_{\mathbf{s}}(\mathbf{s})$, estimate $\mathbf{s}$ using a maximum a posteriori (MAP) probability estimator, and regard the MSE as the performance indicator. We will show in this section that despite the different formulation we can obtain results similar to those described in Section 1.4.

Let us recall the observation model in (1.35), denote the mean of $\mathbf{s}$ as $\mathrm{E}(\mathbf{s}) := \boldsymbol{\mu}_{\mathbf{s}}$ and suppose the noise vector is white and Gaussian i.e., $\mathrm{E}[\mathbf{w}(n)\mathbf{w}^T(n)] = \mathrm{diag}[\sigma_1^2(n), \ldots, \sigma_M^2(n)]$. In this case, we write $\mathbf{H}_n := [\mathbf{h}_{n1}, \ldots, \mathbf{h}_{nM}]^T$ and define the (independent) binary observations $\mathbf{b}(n) := [b_1(n), \ldots, b_M(n)]$ as

$$b_k(n) := \mathbf{1}\{x_k(n) > \mathbf{h}_{nk}^T \boldsymbol{\mu}_{\mathbf{s}}\}\ , \tag{1.55}$$

for $k \in [1, M]$. The resemblance with the problem of Section 1.4 is clear and not surprisingly the following proposition holds true (Sha *et al.* 2005).

Proposition 1.5.1 *Consider a vector parameter* $\mathbf{s}$*, with log-concave prior distribution* $p_{\mathbf{s}}(\mathbf{s})$*,the model in* (1.35) *with* $p_{\mathbf{w}}(\mathbf{w})$ *white Gaussian with* $\mathrm{E}[\mathbf{w}(n)\mathbf{w}^T(n)] = \mathrm{diag}[\sigma_1^2(n), \ldots, \sigma_M^2(n)]$*; and* *binary messages* $\{\mathbf{b}(n)\}_{n=0}^{N-1}$ *as in* (1.55). *Then, if we define the per sensor log-likelihood* $L_n(\mathbf{s})$ *as*

$$L_n(\mathbf{s}) = \sum_{k=1}^{M} \ln Q\left(\frac{b_k(n)\mathbf{h}_{nk}^T\left[\boldsymbol{\mu}_{\mathbf{s}} - \mathbf{s}\right]}{\sigma_k(n)}\right). \tag{1.56}$$

(a) *The MAP estimator of* $\mathbf{s}$ *based on* $\{\mathbf{b}(n)\}_{n=0}^{N-1}$ *is given by*

$$\hat{\mathbf{s}}_{\text{MAP}} = \arg\max\left[\sum_{n=0}^{N-1} L_n(\mathbf{s})\right] + \ln[p_{\mathbf{s}}(\mathbf{s})] := \arg\max L(\mathbf{s}). \tag{1.57}$$

(b) *The log-likelihood* $L(\mathbf{s})$ *is a concave function of* $\mathbf{s}$.

Proposition 1.5.1 establishes that at least for white Gaussian noise the comments in Remark 3 carry over to MAP based parameter estimation. In fact, Proposition 1.5.1 has been established under much more general assumptions, including the case of colored Gaussian noise (Sha *et al.* 2005).

1.5.1 Mean-Squared Error

For estimation of random parameters bounds on the MSE can be obtained by computing the pertinent Fisher Information Matrix (FIM) $\mathbf{J}$ that can be expressed as the sum of two parts (Van Trees 1968, p. 84):

$$\mathbf{J} = \mathbf{J}_D + \mathbf{J}_P, \tag{1.58}$$

where $\mathbf{J}_D$ represents information obtained from the data, and $\mathbf{J}_P$ captures *a priori* information. The MSE of the i^{th} component of $\mathbf{s}$ is bounded by the i^{th} diagonal element of $\mathbf{J}$; i.e.,

$$\text{MSE}(\hat{s}_i) \geq \left[\mathbf{J}^{-1}\right]_{ii}. \tag{1.59}$$

Also, note that for any FIM, $[\mathbf{J}^{-1}]_{ii} \geq 1/[\mathbf{J}]_{ii}$ (Kay 1993). This property yields a different bound on $\text{MSE}(\hat{s}_i)$

$$\text{MSE}(\hat{s}_i) \geq \frac{1}{[\,\mathbf{J}\,]_{ii}}, \tag{1.60}$$

which is easier to compute although not tight in general.

The following proposition provides a bound (exact value) on $[\mathbf{J}]_{ii}$ when binary (analog-amplitude) observations are used (Sha *et al.* 2005).

Proposition 1.5.2 *Consider the signal model in (1.35) with* $\mathbf{w}(n)$ *white Gaussian with covariance matrix* $\text{E}[\mathbf{w}(n)\mathbf{w}^T(n)] = \text{diag}[\sigma_1^2(n), \ldots, \sigma_M^2(n)]$ *and Gaussian prior distribution with covariance* $\text{E}[\mathbf{s}\mathbf{s}^T] = \mathbf{C}_{\mathbf{s}}$. *Write (1.35) componentwise as* $x_k(n) = \mathbf{h}_{nk}^T\mathbf{s} + w_k(n)$. *Then, the* i^{th} *diagonal element of the FIM* $\mathbf{J}$ *in* (1.58) *satisfies:*

(a) *when binary observations as in (1.55) are used*

$$[\,\mathbf{J}\,]_{ii} \geq \frac{2}{\pi}\sum_{n=0}^{N-1}\sum_{k=1}^{M}\frac{h_{nki}^2}{\sigma_k(n)\sqrt{\sigma_k^2(n) + \mathbf{h}_{nk}^T\mathbf{C}_{\mathbf{s}}\mathbf{h}_{nk}}} + \left[\mathbf{C}_{\mathbf{s}}^{-1}\right]_{ii} \tag{1.61}$$

(b) *when analog-amplitude observations are used*

$$[\mathbf{J}_{\text{CV}}]_{ii} = \sum_{n=0}^{N-1}\sum_{k=1}^{M}\frac{h_{nki}^2}{\sigma_k^2(n)} + \left[\mathbf{C}_{\mathbf{s}}^{-1}\right]_{ii}. \tag{1.62}$$

Comparing (1.61) with (1.62) the analogy with the result in Proposition 1.4.2 becomes clear. Indeed, we can define the equivalent noise powers as

$$\rho_k^2(n) = \frac{\pi}{2}\sigma_w^2\sqrt{1 + \frac{\mathbf{h}_{nk}^T\mathbf{C_s}\mathbf{h}_{nk}}{\sigma_k^2(n)}} \tag{1.63}$$

so that we can express the bound in (1.61) as

$$[\mathbf{J}_{\mathrm{CV}}]_{ii} = \sum_{n=0}^{N-1}\sum_{k=1}^{M}\frac{h_{nki}^2}{\rho_k^2(n)} + \left[\mathbf{C}_\mathbf{s}^{-1}\right]_{ii}. \tag{1.64}$$

As in the case of deterministic parameters, the effect of quantization in MSE is equivalent to a noise power increase from $\sigma_k^2(n)$ to $\rho_k^2(n)$ [c.f. (1.62) and (1.64)]. In the case of random signals, the average SNR of the observations $x_k(n)$ is well defined and given by $\gamma_{nk} := \mathbf{h}_{nk}^T\mathbf{C_s}\mathbf{h}_{nk}/\sigma_k^2(n)$. Using the latter and (1.64), we infer that the equivalent noise increase is

$$\mathcal{L}_k(n) := \frac{\rho_k^2(n)}{\sigma_k^2(n)} = \frac{\pi}{2}\sqrt{1+\gamma_{nk}}. \tag{1.65}$$

Note that as $\gamma_{nk} \to 0$, the information loss $\mathcal{L}_k(n) \to \pi/2$ corroborating the results in Section 1.4 for deterministic parameter estimation. In any event, it is worth re-iterating the remarkable fact that for low to medium SNR γ, the equivalent noise increase $\mathcal{L}_K$ is small.

1.6 Dimensionality Reduction for Distributed Estimation

In this section, we consider linear distributed estimation of random signals when the sensors observe and transmit analog-amplitude data. Consider the WSN depicted in Fig. 1.8, comprising N sensors linked with an FC. Each sensor, say the nth one, observes an $M_n \times 1$ vector $\mathbf{x}_n$ that is correlated with a $p \times 1$ random signal of interest $\mathbf{s}$. Through a $k_n \times M_n$ fat matrix $\mathbf{C}_n$, each sensor transmits a compressed $k_n \times 1$ vector $\mathbf{C}_n\mathbf{x}_n$, using e.g., multicarrier modulation with one entry riding per subcarrier. Low-power and bandwidth constraints at the sensors encourage transmissions with $k_n \ll M_n$, while linearity in compression and estimation are well motivated by low-complexity requirements. Furthermore, we assume that:

- **(a1)** No information is exchanged among sensors, and each sensor-FC link comprises a $k_n \times k_n$ full rank fading multiplicative channel matrix $\mathbf{D}_n$ along with zero-mean additive FC noise $\mathbf{z}_n$, which is uncorrelated with $\mathbf{x}_n$, $\mathbf{D}_n$, and across channels; i.e., noise covariance matrices satisfy $\mathbf{\Sigma}_{z_{n_1}z_{n_2}} = \mathbf{0}$ for $n_1 \neq n_2$. Matrices $\{\mathbf{D}_n, \mathbf{\Sigma}_{z_nz_n}\}_{n=0}^{N-1}$ are available at the FC.
- **(a2)** Data $\mathbf{x}_n$ and the signal of interest $\mathbf{s}$ are zero-mean with full rank auto- and cross-covariance matrices $\mathbf{\Sigma}_{ss}$, $\mathbf{\Sigma}_{sx_n}$ and $\mathbf{\Sigma}_{x_{n_1}x_{n_2}}$ $\forall\, n_1, n_2 \in [0, N-1]$, all of which are available at the FC.

In multicarrier links, full rank of the channel matrices $\{\mathbf{D}_n\}_{n=0}^{N-1}$ is ensured if sensors do not transmit over subcarriers with zero channel gain. Matrices $\{\mathbf{D}_n\}_{n=0}^{N-1}$ can be acquired via training, and likewise the signal and noise covariances in (a1) and (a2) can be estimated via sample averaging as usual. With multicarrier (and generally any orthogonal) sensor access, the noise uncorrelatedness across channels is also well justified. Notice that unlike (Gastpar *et al.* 2004; Luo *et al.* 2005; Zhang *et al.* 2003; Zhu *et al.* 2005), we neither confine ourselves to a linear signal-plus-noise model $\mathbf{x}_n = \mathbf{H}_n\mathbf{s} + \mathbf{w}_n$, nor we invoke any assumption on the distribution (e.g., Gaussianity) of $\{\mathbf{x}_n\}_{n=0}^{N-1}$ and $\mathbf{s}$. Equally important, we do not assume ideal channel links.

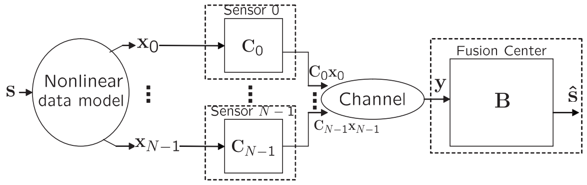

Figure 1.8 Distributed setup for estimating a random signal $\mathbf{s}$

Sensors transmit over orthogonal channels so that the FC separates and concatenates the received vectors $\{\mathbf{y}_n(\mathbf{C}_n) = \mathbf{D}_n\mathbf{C}_n\mathbf{x}_n + \mathbf{z}_n\}_{n=0}^{N-1}$, to obtain the $\sum_{n=0}^{N-1} k_n \times 1$ vector

$$\mathbf{y}(\mathbf{C}_0, \ldots, \mathbf{C}_{N-1}) = \text{diag}(\mathbf{D}_0\mathbf{C}_0, \ldots, \mathbf{D}_{N-1}\mathbf{C}_{N-1})\mathbf{x} + \mathbf{z}, \tag{1.66}$$

Left multiplying $\mathbf{y}$ by a $p \times (\sum_{n=0}^{N-1} k_n)$ matrix $\mathbf{B}$, we form the linear estimate $\hat{\mathbf{s}}$ of $\mathbf{s}$. For a prescribed power P_n per sensor, our problem is to obtain under (a1)-(a2) MSE optimal matrices $\{\mathbf{C}_n^o\}_{n=0}^{N-1}$ and $\mathbf{B}^o$; i.e., we seek (tr denotes matrix trace)

$$\begin{aligned}(\mathbf{B}^o, \{\mathbf{C}_n^o\}_{n=0}^{N-1}) = \arg\min\nolimits_{\mathbf{B}, \{\mathbf{C}_n\}_{n=0}^{N-1}} E[\|\mathbf{s} - \mathbf{B}\mathbf{y}(\mathbf{C}_0, \ldots, \mathbf{C}_{N-1})\|^2],\\ \text{s. to} \quad \text{tr}(\mathbf{C}_n\mathbf{\Sigma}_{x_n x_n}\mathbf{C}_n^T) \leq P_n, \quad n \in \{0, \ldots, N-1\}.\end{aligned} \tag{1.67}$$

1.6.1 Decoupled Distributed Estimation-Compression

We consider first the case where $\mathbf{\Sigma}_{x_n x_m} \equiv \mathbf{0}, \forall n \neq m$, which shows up e.g., when matrices $\{\mathbf{H}_n\}_{n=0}^{N-1}$ in the linear model $\mathbf{x}_n = \mathbf{H}_n\mathbf{s} + \mathbf{w}_n$ are mutually uncorrelated and also uncorrelated with $\mathbf{w}_n$. Then, the multi-sensor optimization task in (1.67) reduces to a set of N decoupled problems. Specifically, it is easy to show that the cost function in (1.67) can be written as (Schizas *et al.* 2005b)

$$J(\mathbf{B}, \{\mathbf{C}_n\}_{n=0}^{N-1}) = \sum_{n=0}^{N-1} E[\|\mathbf{s} - \mathbf{B}_n(\mathbf{D}_n\mathbf{C}_n\mathbf{x}_n + \mathbf{z}_n)\|^2] - (N-1)\text{tr}(\mathbf{\Sigma}_{ss}) \tag{1.68}$$

where $\mathbf{B}_n$ is the $p \times k_n$ submatrix of $\mathbf{B} := [\mathbf{B}_0 \ldots \mathbf{B}_{N-1}]$. As the nth non-negative summand depends only on $\mathbf{B}_n$ and $\mathbf{C}_n$, the MSE optimal matrices are given by

$$\begin{aligned}(\mathbf{B}_n^o, \mathbf{C}_n^o) &= \arg\min\nolimits_{\mathbf{B}_n, \mathbf{C}_n} E[\|\mathbf{s} - \mathbf{B}_n(\mathbf{D}_n\mathbf{C}_n\mathbf{x}_n + \mathbf{z}_n)\|^2],\\ &\text{s. to} \quad \text{tr}(\mathbf{C}_n\mathbf{\Sigma}_{x_n x_n}\mathbf{C}_n^T) \leq P_n, \quad n \in \{0, \ldots, N-1\}.\end{aligned} \tag{1.69}$$

Since the cost function in (1.69) corresponds to a single-sensor setup ($N = 1$), we will drop the subscript n for notational brevity and write $\mathbf{B}_n = \mathbf{B}, \mathbf{C}_n = \mathbf{C}, \mathbf{x}_n = \mathbf{x}, \mathbf{z}_n = \mathbf{z}, P_n = P$ and $k_n = k$. The Lagrangian for minimizing (1.68) can be easily written as:

$$\begin{aligned}J(\mathbf{B}, \mathbf{C}, \mu) = J_o + \text{tr}(\mathbf{B}\mathbf{\Sigma}_{zz}\mathbf{B}^T) + \mu[\text{tr}(\mathbf{C}\mathbf{\Sigma}_{xx}\mathbf{C}^T) - P]\\ + \text{tr}[(\mathbf{\Sigma}_{sx} - \mathbf{B}\mathbf{D}\mathbf{C}\mathbf{\Sigma}_{xx})\mathbf{\Sigma}_{xx}^{-1}(\mathbf{\Sigma}_{xs} - \mathbf{\Sigma}_{xx}\mathbf{C}^T\mathbf{D}^T\mathbf{B}^T)],\end{aligned} \tag{1.70}$$

where $J_o := \text{tr}(\mathbf{\Sigma}_{ss} - \mathbf{\Sigma}_{sx}\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xs})$ is the minimum attainable MMSE for linear estimation of $\mathbf{s}$ based on $\mathbf{x}$.

In what follows, we derive a simplified form of (1.70) the minimization of which will provide closed-form solutions for the MSE optimal matrices $\mathbf{B}^o$ and $\mathbf{C}^o$. Aiming at this simplification, consider the SVD $\mathbf{\Sigma}_{sx} = \mathbf{U}_{sx}\mathbf{S}_{sx}\,\mathbf{V}_{sx}^T$, and the eigen-decompositions $\mathbf{\Sigma}_{zz} = \mathbf{Q}_z\mathbf{\Lambda}_z\mathbf{Q}_z^T$ and $\mathbf{D}^T\mathbf{\Sigma}_{zz}^{-1}\mathbf{D}$

$= \mathbf{Q}_{zd}\mathbf{\Lambda}_{zd}\mathbf{Q}_{zd}^T$, where $\mathbf{\Lambda}_{zd} := \text{diag}(\lambda_{zd,1} \cdots \lambda_{zd,k})$ and $\lambda_{zd,1} \geq \cdots \geq \lambda_{zd,k} > 0$. Notice that $\lambda_{zd,i}$ captures the SNR of the ith entry in the received signal vector at the FC. Further, define $\mathbf{A} := \mathbf{Q}_x^T\mathbf{V}_{sx}\mathbf{S}_{sx}^T\ \mathbf{S}_{sx}\mathbf{V}_{sx}^T\mathbf{Q}_x$ with $\rho_a := \text{rank}(\mathbf{A}) = \text{rank}(\mathbf{\Sigma}_{sx})$, and $\mathbf{A}_x := \mathbf{\Lambda}_x^{-1/2}\mathbf{A}\mathbf{\Lambda}_x^{-1/2}$ with corresponding eigen-decomposition $\mathbf{A}_x = \mathbf{Q}_{ax}\mathbf{\Lambda}_{ax}\mathbf{Q}_{ax}$, where $\mathbf{\Lambda}_{ax} = \text{diag}(\lambda_{ax,1}, \cdots, \lambda_{ax,\rho_a}, 0, \cdots, 0)$ and $\lambda_{ax,1} \geq \ldots \geq \lambda_{ax,\rho_a} > 0$. Moreover, let $\mathbf{V}_a := \mathbf{\Lambda}_x^{-1/2}\mathbf{Q}_{ax}$ denote the invertible matrix which simultaneously diagonalizes the matrices $\mathbf{A}$ and $\mathbf{\Lambda}_x$. Since matrices ($\mathbf{Q}_{zd}, \mathbf{Q}_x, \mathbf{V}_a, \mathbf{U}_{sx}, \mathbf{\Lambda}_{zd}, \mathbf{Q}_{zd}, \mathbf{D}, \mathbf{\Sigma}_{zz}$) are all invertible, for every matrix $\mathbf{C}$ (or $\mathbf{B}$) we can clearly find a unique matrix $\mathbf{\Phi}_C$ (correspondingly $\mathbf{\Phi}_B$) that satisfies:

$$\mathbf{C} = \mathbf{Q}_{zd}\mathbf{\Phi}_C\mathbf{V}_a^T\mathbf{Q}_x^T, \qquad \mathbf{B} = \mathbf{U}_{sx}\mathbf{\Phi}_B\mathbf{\Lambda}_{zd}^{-1}\mathbf{Q}_{zd}^T\mathbf{D}^T\mathbf{\Sigma}_{zz}^{-1}, \tag{1.71}$$

where $\mathbf{\Phi}_C := [\phi_{c,ij}]$ and $\mathbf{\Phi}_B$ have sizes $k \times M$ and $p \times k$, respectively. Using (1.71), the Lagrangian in (1.70) becomes

$$\begin{aligned} J(\mathbf{\Phi}_C, \mu) = \ & J_o + \text{tr}(\mathbf{\Lambda}_{ax}) + \mu(\text{tr}(\mathbf{\Phi}_C\mathbf{\Phi}_C^T) - P) \\ & -\text{tr}\left((\mathbf{\Lambda}_{zd}^{-1} + \mathbf{\Phi}_C\mathbf{\Phi}_C^T)^{-1}\mathbf{\Phi}_C\mathbf{\Lambda}_{ax}\mathbf{\Phi}_C^T\right). \end{aligned} \tag{1.72}$$

Applying the well known Karush-Kuhn-Tucker (KKT) conditions (e.g., (Boyd and Vandenberghe 2004, Ch. 5)) that must be satisfied at the minimum of (1.72), it can be shown that the matrix $\mathbf{\Phi}_C^o$ minimizing (1.72), is diagonal with diagonal entries (Schizas *et al.* 2005b)

$$\phi_{c,ii}^o = \begin{cases} \pm\sqrt{\left(\frac{\lambda_{ax,i}}{\mu^o\lambda_{zd,i}}\right)^{1/2} - \frac{1}{\lambda_{zd,i}}}, & 1 \leq i \leq \kappa \\ 0, & \kappa + 1 \leq i \leq k \end{cases} \tag{1.73}$$

where κ is the maximum integer in $[1, k]$ for which $\{\phi_{c,ii}^o\}_{i=1}^{\kappa}$ are strictly positive, or, rank($\mathbf{\Phi}_C^o$) $= \kappa$; and μ^o is chosen to satisfy the power constraint $\sum_{i=1}^{\kappa}(\phi_{c,ii}^o)^2 = P$ as

$$\mu^o = \frac{(\sum_{i=1}^{\kappa}(\lambda_{ax,i}\lambda_{zd,i}^{-1})^{1/2})^2}{(P + \sum_{i=1}^{\kappa}\lambda_{zd,i}^{-1})^2}. \tag{1.74}$$

When $k > \rho_a$, the MMSE remains invariant (Schizas *et al.* 2005b); thus, it suffices to consider $k \in [1, \rho_a]$. Summarizing, it has been established that:

Proposition 1.6.1 *Under (a1), (a2), and for $k \leq \rho_a$, the matrices minimizing $J(\mathbf{B}_{p\times k}, \mathbf{C}_{k\times M}) = E[\|\mathbf{s} - \mathbf{B}_{p\times k}(\mathbf{D}\mathbf{C}_{k\times M}\mathbf{x} + \mathbf{z})\|^2]$, subject to tr($\mathbf{C}_{k\times M}\mathbf{\Sigma}_{xx}\ \mathbf{C}_{k\times M}^T) \leq P$, are:*

$$\begin{aligned} \mathbf{C}^o &= \mathbf{Q}_{zd}\mathbf{\Phi}_C^o\mathbf{V}_a^T\mathbf{Q}_x^T, \\ \mathbf{B}^o &= \mathbf{\Sigma}_{sx}\mathbf{Q}_x\mathbf{V}_a\mathbf{\Phi}_C^{o\,T}\left(\mathbf{\Phi}_C^o\mathbf{\Phi}_C^{o\,T} + \mathbf{\Lambda}_{zd}^{-1}\right)^{-1}\mathbf{\Lambda}_{zd}^{-1}\mathbf{Q}_{zd}^T\mathbf{D}^T\mathbf{\Sigma}_{zz}^{-1}, \end{aligned} \tag{1.75}$$

where $\mathbf{\Phi}_C^o$ is given by (1.73), and the corresponding Lagrange multiplier μ^o is specified by (1.74). The MMSE is

$$J_{\min}(k) = J_o + \sum_{i=1}^{\rho_a}\lambda_{ax,i} - \sum_{i=1}^{k}\frac{\lambda_{ax,i}(\phi_{c,ii}^o)^2}{\lambda_{zd,i}^{-1} + (\phi_{c,ii}^o)^2}. \tag{1.76}$$

According to Proposition 1.6.1, the optimal weight matrix $\mathbf{\Phi}_C^o$ in $\mathbf{C}^o$ distributes the given power across the entries of the pre-whitened vector $\mathbf{V}_a^T\mathbf{Q}_x\mathbf{x}$ at the sensor in a waterfilling-like manner so as to balance channel strength and additive noise variance at the FC with the degree of dimensionality reduction that can be afforded. It is worth mentioning that (1.73) dictates a minimum power per sensor. Specifically, in order to ensure that rank($\mathbf{\Phi}_C^o$) $= \kappa$ the power must satisfy

$$P > \frac{\sum_{i=1}^{\kappa}(\lambda_{ax,i}\lambda_{zd,i}^{-1})^{1/2}}{\sqrt{\lambda_{ax,\kappa}\lambda_{zd,\kappa}}} - \sum_{i=1}^{\kappa}\lambda_{zd,i}^{-1}. \tag{1.77}$$

The optimal matrices in Proposition 1.6.1 can be viewed as implementing a two-step scheme, where: i) $\mathbf{s}$ is estimated based on $\mathbf{x}$ at the sensor using the LMMSE estimate $\hat{\mathbf{s}}_{LM} = \mathbf{\Sigma}_{sx}\mathbf{\Sigma}_{xx}^{-1}\mathbf{x}$; and ii) compress and reconstruct $\hat{\mathbf{s}}_{LM}$ using the optimal matrices $\mathbf{C}^o$ and $\mathbf{B}^o$ implied by Proposition 1.6.1 after replacing $\mathbf{x}$ with $\hat{\mathbf{s}}_{LM}$. For this estimate-first compress-afterwards (EC) interpretation, (Schizas *et al.* 2005b) have proved that:

Corollary 1.6.2 *For $k \in [1, \rho_a]$, the $k \times M$ matrix in (1.75) can be written as $\mathbf{C}^o = \hat{\mathbf{C}}^o\mathbf{\Sigma}_{sx}\ \mathbf{\Sigma}_{xx}^{-1}$, where $\hat{\mathbf{C}}^o$ is the $k \times p$ optimal matrix obtained by Proposition 1.6.1 when $\mathbf{x} = \hat{\mathbf{s}}_{LM}$. Thus, the EC scheme is MSE optimal in the sense of minimizing (1.68).*

Another interesting feature of the EC scheme implied by Proposition 1.6.1 is that the MMSE $J_{\min}(k)$ is non-increasing with respect to the reduced dimensionality k, given a limited power budget per sensor. Specifically, (Schizas *et al.* 2005b) have shown that that:

Corollary 1.6.3 *If $\mathbf{C}_{k_1 \times M}^o$ and $\mathbf{C}_{k_2 \times M}^o$ are the optimal matrices determined by Proposition 1.6.1 with $k_1 < k_2$, under the same channel parameters $\lambda_{zd,i}$ for $i = 1, \ldots, k_1$, and common power P, the MMSE in (1.76) is non-increasing; i.e., $J_{\min}(k_1) \geq J_{\min}(k_2)$ for $k_1 < k_2$.*

Notice that Corollary 1.6.3 advocates the efficient power allocation that the EC scheme performs among the compressed components. To assess the difference in handling noise effects, it is useful compare the EC scheme with the methods in (Zhu *et al.* 2005) and (Zhang *et al.* 2003), which we abbreviate as C′E and C″E because they perform compression (C) followed by estimation (E). Although C′E and C″E have been derived under ideal link conditions, they can be modified them here to account for $\mathbf{D}_n$. The comparisons will further include an option we term CE, which compresses first the data and reconstructs them at the FC using $\mathbf{C}^o$ and $\mathbf{B}^o$ found by (1.75) after setting $\mathbf{s} = \mathbf{x}$, and then estimates $\mathbf{s}$ based on the reconstructed data vector $\hat{\mathbf{x}}$. For benchmarking purposes, we also depict J_o, achieved when estimating $\mathbf{s}$ based on uncompressed data transmitted over ideal links. Fig. 1.9 (Left) depicts the MMSE versus k for J_o, EC, CE, C′E and C″E for a linear model $\mathbf{x} = \mathbf{H}\mathbf{s} + \mathbf{w}$, where $M = 50$ and $p = 10$. The matrices $\mathbf{H}, \mathbf{\Sigma}_{ss}$ and $\mathbf{\Sigma}_{ww}$, are selected randomly such that tr($\mathbf{H}\mathbf{\Sigma}_{ss}\mathbf{H}^T$)/tr($\mathbf{\Sigma}_{ww}$) $= 2$, while $\mathbf{s}$ and $\mathbf{w}$ are uncorrelated. We set $\mathbf{\Sigma}_{zz} = \sigma_z^2\mathbf{I}_k$, and select P such that $10\log_{10}(P/\sigma_z^2) = 7$dB. As expected J_o benchmarks all curves, while the worst performance is exhibited by C′E. Albeit suboptimal, CE comes close to the optimal EC. Contrasting it with the increase C″E exhibits in MMSE beyond a certain k, we can appreciate the importance of coping with noise effects. This increase is justifiable since each entry of the compressed data in C″E is allocated a smaller portion of the given power as k grows. In EC however, the quality of channel links and the available power determine the number of the compressed components, and allocate power optimally among them.

1.6.2 Coupled Distributed Estimation-Compression

In this section, we allow the sensor observations to be correlated. Because $\mathbf{\Sigma}_{xx}$ is no longer block diagonal, decoupling of the multi-sensor optimization problem cannot be effected in this case. The pertinent MSE cost is [c.f. (1.67)]

$$J(\{\mathbf{B}_n, \mathbf{C}_n\}_{n=0}^{N-1}) = E[\|\mathbf{s} - \textstyle\sum_{n=0}^{N-1}\mathbf{B}_n(\mathbf{D}_n\mathbf{C}_n\mathbf{x}_n + \mathbf{z}_n)\|^2]. \tag{1.78}$$

Minimizing (1.78) does not lead to a closed-form solution and incurs complexity that grows exponentially with N (Luo *et al.* 2005). For this reason, we resort to iterative alternatives which converge at least to a stationary point of the cost in (1.78). To this end, let us suppose temporarily that matrices $\{\mathbf{B}_l\}_{l=0,l\neq n}^{N-1}$ and $\{\mathbf{C}_l\}_{l=0,l\neq n}^{N-1}$ are fixed and satisfy the power constraints tr$(\mathbf{C}_l\mathbf{\Sigma}_{x_l x_l}\mathbf{C}_l^T) = P_l$, for $l = 0, \ldots, N-1$ and $l \neq n$. Upon defining the vector $\bar{\mathbf{s}}_n := \mathbf{s} - \sum_{l=0,l\neq n}^{N-1}(\mathbf{B}_l\mathbf{D}_l\mathbf{C}_l\mathbf{x}_l + \mathbf{B}_l\mathbf{z}_l)$ the cost in (1.78) becomes

$$J(\mathbf{B}_n, \mathbf{C}_n) = E[\|\bar{\mathbf{s}}_n - \mathbf{B}_n\mathbf{D}_n\mathbf{C}_n\mathbf{x}_n - \mathbf{B}_n\mathbf{z}_n\|^2] \ , \tag{1.79}$$

which being a function of $\mathbf{C}_n$ and $\mathbf{B}_n$ only, falls under the realm of Proposition 1.6.1. This means that when $\{\mathbf{B}_l\}_{l=0,l\neq n}^{N-1}$ and $\{\mathbf{C}_l\}_{l=0,l\neq n}^{N-1}$ are given, the matrices $\mathbf{B}_n$ and $\mathbf{C}_n$ minimizing (1.79) under the power constraint tr$(\mathbf{C}_n\mathbf{\Sigma}_{x_n x_n}\mathbf{C}_n^T) \leq P_n$ can be directly obtained from (1.75), after setting $\mathbf{s} = \bar{\mathbf{s}}_n$, $\mathbf{x} = \mathbf{x}_n$, $\mathbf{z} = \mathbf{z}_n$ and $\rho_a =$ rank$(\mathbf{\Sigma}_{\bar{s}_n x_n})$ in Proposition 1.6.1. The corresponding auto- and cross- covariance matrices needed must also be modified as $\mathbf{\Sigma}_{ss} = \mathbf{\Sigma}_{\bar{s}_n\bar{s}_n}$ and $\mathbf{\Sigma}_{sx_n} = \mathbf{\Sigma}_{\bar{s}_n x_n}$. The following result can thus be established for coupled sensor observations:

Proposition 1.6.4 *If (a1) and (a2) are satisfied, and $k_n \leq$ rank$(\mathbf{\Sigma}_{\bar{s}_n x_n})$, then for given matrices $\{\mathbf{B}_l\}_{l=0,l\neq n}^{N-1}$ and $\{\mathbf{C}_l\}_{l=0,l\neq n}^{N-1}$ satisfying tr$(\mathbf{C}_l\mathbf{\Sigma}_{x_l x_l}\mathbf{C}_l^T) = P_l$, the optimal $\mathbf{B}_n^o$ and $\mathbf{C}_n^o$ matrices minimizing $E[\|\mathbf{s} - \sum_{l=0}^{N-1}\mathbf{B}_l(\mathbf{D}_l\mathbf{C}_l\mathbf{x}_l + \mathbf{z}_l)\|^2]$ are provided by Proposition 1.6.1, after setting $\mathbf{x} = \mathbf{x}_n$, $\mathbf{s} = \bar{\mathbf{s}}_n$ and applying the corresponding covariance modifications.*

Proposition 1.6.4 suggests Algorithm 1 for distributed estimation in the presence of fading and FC noise. Notice that Algorithm 1 belongs to the class of block coordinate descent iterative schemes.

Algorithm 1 :

```
Initialize randomly the matrices $\{\mathbf{C}_n^{(0)}\}_{n=0}^{N-1}$ and $\{\mathbf{B}_n^{(0)}\}_{n=0}^{N-1}$, such that tr$(\mathbf{C}_n^{(0)}\mathbf{\Sigma}_{x_n x_n}\mathbf{C}_n^{(0)^T}) = P_n$.
$i = 0$
repeat
  $i = i + 1$
  for $n = 0, N-1$ do
    Given the matrices $\mathbf{C}_0^{(i)}, \mathbf{B}_0^{(i)}, \ldots, \mathbf{C}_{n-1}^{(i)}, \mathbf{B}_{n-1}^{(i)}, \mathbf{C}_{n+1}^{(i-1)}, \mathbf{B}_{n+1}^{(i-1)}, \ldots, \mathbf{C}_{N-1}^{(i-1)}, \mathbf{B}_{N-1}^{(i-1)}$ determine $\mathbf{C}_n^{(i)}, \mathbf{B}_n^{(i)}$ via Proposition 1.6.1
  end for
until $|\mathrm{MSE}^{(i)} - \mathrm{MSE}^{(i-1)}| < \epsilon$ for given tolerance $\epsilon$
```

At every step n during the ith iteration, it yields the optimal pair of matrices $\mathbf{C}_n^o, \mathbf{B}_n^o$, treating the rest as given. Thus, the MSE$^{(i)}$ cost per iteration is non-increasing and the algorithm always converges to a stationary point of (1.78). Beyond its applicability to possibly non-Gaussian and nonlinear data models, it is the only available algorithm for handling fading channels and generally colored FC noise effects in distributed estimation.

Next, we illustrate through a numerical example the MMSE performance of Algorithm 1 in a 3-sensor setup using the same linear model as in Section 1.6.1, while setting $M_0 = M_1 = 17$ and $M_2 = 16$. FC noise $\mathbf{z}_n$ is white with variance $\sigma_{z_n}^2$. The power P_n and variance $\sigma_{z_n}^2$ are chosen such that $10\log_{10}(P/\sigma_{z_n}^2) = 13$dB, for $n = 0, 1, 2$, and $\epsilon = 10^{-3}$. Fig. 1.9 (Right) depicts the MMSE as a function of the total number $k_{tot} = \sum_{n=0}^{2} k_n$ of compressed entries across sensors for: i) a centralized EC setup for which a single (virtual) sensor ($N = 1$) has available the data vectors of all three

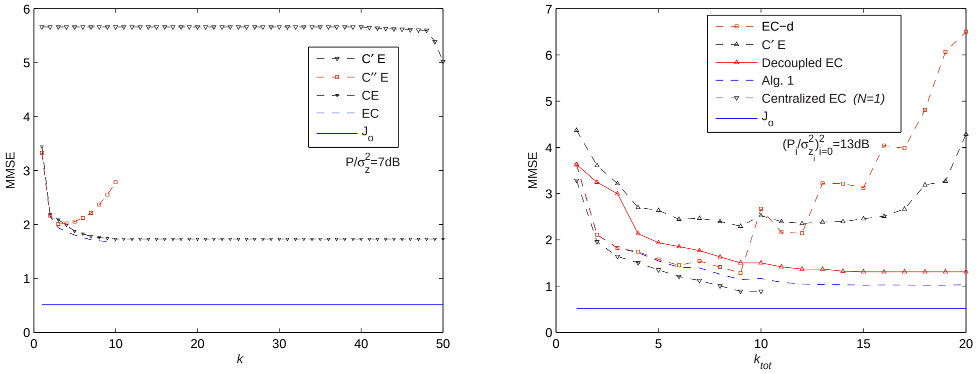

Figure 1.9 MMSE comparisons versus k for a centralized, $L = 1$ (Left), and a distributed 3-sensor setup (Right).

sensors; ii) the estimator returned by Algorithm 1; iii) the decoupled EC estimator which ignores sensor correlations; iv) the C′E estimator and v) an iterative estimator developed in (Schizas *et al.* 2005b), denoted here as EC-d, which similar to C′E accounts for fading but ignores noise. Interestingly, the decentralized Algorithm 1 comes very close to the hypothetical single-sensor bound of the centralized EC estimator, while outperforming the decoupled EC one.

1.7 Distortion-Rate Analysis

In contrast to the previous section, here we consider digital-amplitude data transmission (bits) from the sensors to the FC. In such a setup, all the sensors must adhere to a rate constraint. In order to determine the minimum possible distortion (MSE) between the signal of interest and its estimate at the FC, under encoding rate constraints, we perform Distortion-Rate (D-R) analysis and determine bounds for the D-R function.

Fig. 1.10 (Left) depicts a WSN comprising N sensors that communicate with an FC. Each sensor, say the nth, observes an $M_n \times 1$ vector $\mathbf{x}_n(t)$ which is correlated with a $p \times 1$ random signal (parameter vector) of interest $\mathbf{s}(t)$, where t denotes discrete time. Similar to (Oohama 1998; Pandya *et al.* 2004; Viswanathan and Berger 1997), we assume that:

(a3) No information is exchanged among sensors and the links with the FC are noise-free.

(a4) The random vector $\mathbf{s}(t)$ is generated by a stationary Gaussian vector memoryless source with $\mathbf{s}(t) \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{ss})$; the sensor data $\{\mathbf{x}_n(t)\}_{n=0}^{N-1}$ adhere to the linear-Gaussian model $\mathbf{x}_n(t) = \mathbf{H}_n\mathbf{s}(t) + \mathbf{w}_n(t)$, where $\mathbf{w}_n(t)$ denotes additive white Gaussian noise (AWGN); i.e., $\mathbf{w}_n(t) \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$; noise $\mathbf{w}_n(t)$ is uncorrelated across sensors, time and with $\mathbf{s}$; and $\mathbf{H}_n$ as well as (cross-) covariance matrices $\mathbf{\Sigma}_{ss}$, $\mathbf{\Sigma}_{sx_n}$ and $\mathbf{\Sigma}_{x_nx_m}$ are known $\forall\, n, m \in \{0, \ldots, N-1\}$.

Notice that (a3) assumes that sufficiently strong channel codes are used; while whiteness of $\mathbf{w}_n(t)$ and the zero-mean assumptions in (a4) are made without loss of generality. The linear model in (a4) is commonly encountered in estimation and in a number of cases it even accurately approximates non-linear mappings; e.g., via a first-order Taylor expansion in target tracking applications. Although confining ourselves to Gaussian vectors $\mathbf{x}_n(t)$ is of interest on its own, it can be shown, similarly to (Berger 1971, p. 134), that the D-R functions obtained for Gaussian data bound from above their counterparts for non-Gaussian sensor data $\mathbf{x}_n(t)$.

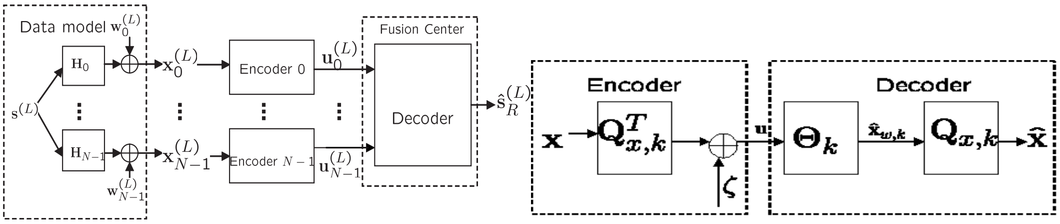

Figure 1.10 (Left): Distributed setup.; (Right): Test channel for $\mathbf{x}$ Gaussian in a point-to-point link.

Blocks $\mathbf{x}_n^{(L)} := \{\mathbf{x}_n(t)\}_{t=1}^L$, comprising L consecutive time instantiations of the vector $\mathbf{x}_n(t)$, are encoded per sensor to yield each encoder's output $\mathbf{u}_n^{(L)} = \mathbf{f}_n^{(L)}(\mathbf{x}_n^{(L)})$, $n = 0, \ldots, N-1$. These outputs are communicated through ideal orthogonal channels to the FC. There, $\mathbf{u}_n^{(L)}$'s are decoded to obtain an estimate of $\mathbf{s}^{(L)} := \{\mathbf{s}(t)\}_{t=1}^L$ denoted as $\hat{\mathbf{s}}_R^{(L)}(\mathbf{u}_0^{(L)}, \ldots, \mathbf{u}_{N-1}^{(L)}) = \mathbf{g}_R^{(L)}(\mathbf{x}_0^{(L)}, \ldots, \mathbf{x}_{N-1}^{(L)})$, since $\mathbf{u}_n^{(L)}$ is a function of $\mathbf{x}_n^{(L)}$. The rate constraint is imposed through a bound on the cardinality of the range of the sensor encoding functions, i.e., the cardinality of the range of $\mathbf{f}_n^{(L)}$ must be no greater than 2^{LR_n}, where R_n is the available rate at the encoder of the nth sensor. The sum rate satisfies the constraint $\sum_{n=0}^{N-1} R_n \leq R$, where R is the total available rate shared by the N sensors. This setup is precisely the vector Gaussian CEO problem in its most general form without any restrictions in the number of observations and the number of parameters (Berger *et al.* 1996). Under this rate constraint, we want to determine the minimum possible MSE distortion $(1/L)\sum_{t=1}^L E[\|\mathbf{s}(t) - \hat{\mathbf{s}}_R(t)\|^2]$ for estimating $\mathbf{s}$ in the limit of infinite block-length L. When $N = 1$, a single-letter information theoretic characterization is known for the latter, but no simplification is known for the distributed multi-sensor scenario.

1.7.1 Distortion-Rate for Centralized Estimation

Let us first specify the D-R function for estimating $\mathbf{s}$(t) in a *single-sensor* setup. The single-letter characterization of the D-R function in this setup allow us to drop the time index. Here, all $\{\mathbf{x}_n\}_{n=0}^{N-1}$ $:= \mathbf{x}$ are available to a single sensor, and $\mathbf{x} = \mathbf{H}\mathbf{s} + \mathbf{w}$. We let $\rho := \text{rank}(\mathbf{H})$ denote the rank of matrix $\mathbf{H}$. The D-R function in such a scenario provides a lower (non-achievable) bound on the MMSE that can be achieved in a multi-sensor distributed setup, where each $\mathbf{x}_n$ is observed by a different sensor. Existing works treat the case $M = p$ (Sakrison 1968; Wolf and Ziv 1970), but here we look for the D-R function regardless of M, p, in the linear-Gaussian model framework.

D-R Analysis for Reconstruction

The D-R function for encoding a vector $\mathbf{x}$, with pdf $p(\mathbf{x})$, using rate R at an individual sensor, and reconstructing it (in the MMSE sense) as $\hat{\mathbf{x}}$ at the FC, is given by (Cover and Thomas 1991, p. 342):

$$D_x(R) = \min_{p(\hat{\mathbf{x}}|\mathbf{x})} E_{p(\hat{\mathbf{x}},\mathbf{x})}[\|\mathbf{x} - \hat{\mathbf{x}}\|^2], \quad \text{s. to } I(\mathbf{x}; \hat{\mathbf{x}}) \leq R \tag{1.80}$$

where $\mathbf{x} \in \mathbb{R}^M$ and $\hat{\mathbf{x}} \in \mathbb{R}^M$, and the minimization is w.r.t. the conditional pdf $p(\hat{\mathbf{x}}|\mathbf{x})$. Let $\mathbf{\Sigma}_{xx} = \mathbf{Q}_x \mathbf{\Lambda}_x \mathbf{Q}_x^T$ denote the eigenvalue decomposition of $\mathbf{\Sigma}_{xx}$, where $\mathbf{\Lambda}_x = \text{diag}(\lambda_{x,1} \cdots \lambda_{x,M})$ and $\lambda_{x,1} \geq \cdots \geq \lambda_{x,M} > 0$.

For $\mathbf{x}$ Gaussian, $D_x(R)$ can be determined by applying rwf to the pre-whitened vector $\mathbf{x}_w := \mathbf{Q}_x^T \mathbf{x}$ (Cover and Thomas 1991, p. 348). For a prescribed rate R, it turns out that $\exists\ k$ such that the first k entries $\{\mathbf{x}_w(i)\}_{i=1}^k$ of $\mathbf{x}_w$ are encoded and reconstructed independently from each other using rate $\{R_i = 0.5 \log_2 (\lambda_{x,i}/d(k,R))\}_{i=1}^{k}$, where $d(k,R) = \left(\prod_{i=1}^{k} \lambda_{x,i}\right)^{1/k} 2^{-2R/k}$ with $R = \sum_{i=1}^{k} R_i$; and the last $M-k$ entries of $\mathbf{x}_w$ are assigned no rate; i.e., $\{R_i = 0\}_{i=k+1}^{M}$. The corresponding MMSE for encoding $\mathbf{x}_w(i)$, the ith entry of $\mathbf{x}_w$, under a rate constraint R_i, is $D_i = E[\|\mathbf{x}_w(i) - \hat{\mathbf{x}}_w(i)\|^2] = d(k,R)$ when $i = 1, \ldots, k$; and $D_i = \lambda_{x,i}$ when $i = k+1, \ldots, M$. The resultant MMSE (D-R function) is

$$D_x(R) = E[\|\mathbf{x} - \hat{\mathbf{x}}\|^2] = E[\|\mathbf{x}_w - \hat{\mathbf{x}}_w\|^2] = kd(k,R) + \textstyle\sum_{i=k+1}^{M} \lambda_{x,i}. \tag{1.81}$$

Especially for $d(k,R)$, it follows that $\max(\{\lambda_{x,i}\}_{i=k+1}^{M}) \leq d(k,R) < \min\{\lambda_{x,1}, \ldots, \lambda_{x,k}\}$. Intuitively, $d(k,R)$ is a threshold distortion determining which entries of $\mathbf{x}_w$ are assigned with nonzero rate. The first k entries of $\mathbf{x}_w$ with variance $\lambda_{x,i} > d(k,R)$ are encoded with non-zero rate, but the last $M-k$ ones are discarded in the encoding procedure (are set to zero).

Associated with the rwf principle is the so called test channel; see e.g., (Cover and Thomas 1991, p. 345). The encoder's MSE optimal output is $\mathbf{u} = \mathbf{Q}_{x,k}^T \mathbf{x} + \boldsymbol{\zeta}$, where $\mathbf{Q}_{x,k}$ is formed by the first k columns of $\mathbf{Q}_x$, and $\boldsymbol{\zeta}$ models the distortion noise that results due to the rate-constrained encoding of $\mathbf{x}$. The zero-mean AWGN $\boldsymbol{\zeta}$ is uncorrelated with $\mathbf{x}$ and its diagonal covariance matrix $\boldsymbol{\Sigma}_{\zeta\zeta}$ has entries $[\boldsymbol{\Sigma}_{\zeta\zeta}]_{ii} = \lambda_{x,i} D_i/(\lambda_{x,i} - D_i)$. The part of the test channel that takes as input $\mathbf{u}$ and outputs $\hat{\mathbf{x}}$, models the decoder. The reconstruction $\hat{\mathbf{x}}$ of $\mathbf{x}$ at the decoder output is

$$\hat{\mathbf{x}} = \mathbf{Q}_{x,k} \boldsymbol{\Theta}_k \mathbf{u} = \mathbf{Q}_{x,k} \boldsymbol{\Theta}_k \mathbf{Q}_{x,k}^T \mathbf{x} + \mathbf{Q}_{x,k} \boldsymbol{\Theta}_k \boldsymbol{\zeta}, \tag{1.82}$$

where $\boldsymbol{\Theta}_k$ is a diagonal matrix with non-zero entries $[\boldsymbol{\Theta}_k]_{ii} = (\lambda_{x,i} - D_i)/\lambda_{x,i}$, $i = 1, \ldots, k$.

D-R Analysis for Estimation

The D-R function for estimating a source $\mathbf{s}$ given observation $\mathbf{x}$ (where the source and observation are probabilistically drawn from the joint pdf $p(\mathbf{x}, \mathbf{s})$) with rate R at an individual sensor, and reconstructing it (in the MMSE sense) as $\hat{\mathbf{x}}$ at the FC is given by (Berger 1971, p. 79)

$$D_s(R) = \min_{p(\hat{\mathbf{s}}_R|\mathbf{x})} E_{p(\hat{\mathbf{s}}_R, \mathbf{s})}[\|\mathbf{s} - \hat{\mathbf{s}}_R\|^2], \quad \text{s. to } I(\mathbf{x}; \hat{\mathbf{s}}_R) \leq R \tag{1.83}$$

where $\mathbf{s} \in \mathbb{R}^p$ and $\hat{\mathbf{s}}_R \in \mathbb{R}^p$, and the minimization is w.r.t. the conditional pdf $p(\hat{\mathbf{s}}_R|\mathbf{x})$. In order to achieve the D-R function, one might be tempted to first compress $\mathbf{x}$ by applying rwf at the sensor, without taking into account the data model relating $\mathbf{s}$ with $\mathbf{x}$, and subsequently use the reconstructed $\hat{\mathbf{x}}$ to form the MMSE conditional expectation estimate $\hat{\mathbf{s}}_{ce} = E[\mathbf{s}|\hat{\mathbf{x}}]$ at the FC. An alternative option would be to first form the MMSE estimate $\hat{\mathbf{s}} = E[\mathbf{s}|\mathbf{x}]$, encode the latter using rwf at the sensor, and after decoding at the FC, obtain the reconstructed estimate $\hat{\mathbf{s}}_{ec}$. Referring as before the former option as *Compress-Estimate* (CE), and to the latter as *Estimate-Compress* (EC), we are interested in determining which one yields the smallest MSE under a rate constraint R. Another interesting question is whether any of the CE and EC schemes enjoys MMSE optimality (i.e., achieves (1.83)). With subscripts ce and ec corresponding to these two options, let us also define the errors $\tilde{\mathbf{s}}_{ce} := \mathbf{s} - \hat{\mathbf{s}}_{ce}$ and $\tilde{\mathbf{s}}_{ec} := \mathbf{s} - \hat{\mathbf{s}}_{ec}$.

For CE, we depict in Fig. 1.11 (Top) the test channel for encoding $\mathbf{x}$ via rwf, followed by MMSE estimation of $\mathbf{s}$ based on $\hat{\mathbf{x}}$. Suppose that when applying rwf to $\mathbf{x}$ with prescribed rate R, the first k_{ce} components of $\mathbf{x}_w$ are assigned with non-zero rate and the rest are discarded. The MMSE optimal encoder's output for encoding $\mathbf{x}$ is $\mathbf{u}_{ce} = \mathbf{Q}_{x,k_{ce}}^T \mathbf{x} + \boldsymbol{\zeta}_{ce}$. The covariance matrix of $\boldsymbol{\zeta}_{ce}$ has diagonal entries $[\boldsymbol{\Sigma}_{\zeta_{ce}\zeta_{ce}}]_{ii} = \lambda_{x,i} D_i^{ce}/(\lambda_{x,i} - D_i^{ce})$ for $i = 1, \ldots, k_{ce}$, where $D_i^{ce} := E[(\mathbf{x}_w(i) - \hat{\mathbf{x}}_w(i))^2]$. Since $D_i^{ce} = \left(\prod_{i=1}^{k_{ce}} \lambda_{x,i}\right)^{1/k_{ce}} 2^{-2R/k_{ce}}$ when $i = 1, \ldots, k_{ce}$ and $D_i^{ce} = \lambda_{x,i}$, when $i = k_{ce} + 1, \ldots, M$, the reconstructed $\hat{\mathbf{x}}$ in CE is [c.f. (1.82)]:

$$\hat{\mathbf{x}} = \mathbf{Q}_{x,k_{ce}} \boldsymbol{\Theta}_{ce} \mathbf{Q}_{x,k_{ce}}^T \mathbf{x} + \mathbf{Q}_{x,k_{ce}} \boldsymbol{\Theta}_{ce} \boldsymbol{\zeta}_{ce}, \tag{1.84}$$

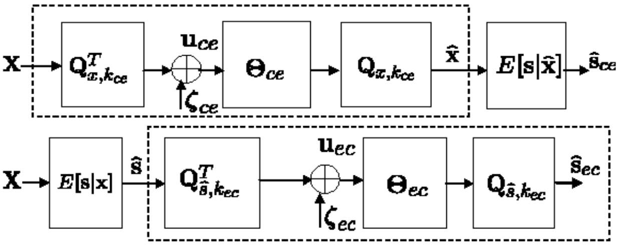
$\mathbf{x}$
$\mathbf{Q}_{x,k_{ce}}^T$
$\mathbf{u}_{ce}$
$\boldsymbol{\zeta}_{ce}$
$\boldsymbol{\Theta}_{ce}$
$\mathbf{Q}_{x,k_{ce}}$
$\hat{\mathbf{x}}$
$E[\mathbf{s}|\hat{\mathbf{x}}]$
$\hat{\mathbf{s}}_{ce}$
$\mathbf{x}$
$E[\mathbf{s}|\mathbf{x}]$
$\hat{\mathbf{s}}$
$\mathbf{Q}_{\hat{s},k_{ec}}^T$
$\mathbf{u}_{ec}$
$\boldsymbol{\zeta}_{ec}$
$\boldsymbol{\Theta}_{ec}$
$\mathbf{Q}_{\hat{s},k_{ec}}$
$\hat{\mathbf{s}}_{ec}$

Figure 1.11 (Top): Test channel for the CE scheme.; (Bottom): Test channel for the EC scheme.

where $[\boldsymbol{\Theta}_{ce}]_{ii} = (\lambda_{x,i} - D_i^{ce})/\lambda_{x,i}$, for $i = 1, \ldots, k_{ce}$. Letting $\check{\mathbf{x}} := \mathbf{Q}_x^T \hat{\mathbf{x}} = [\check{\mathbf{x}}_1^T \;\; \mathbf{0}_{1\times(M-k_{ce})}]^T$, with $\check{\mathbf{x}}_1 := \boldsymbol{\Theta}_{ce}\mathbf{Q}_{x,k_{ce}}^T \mathbf{x} + \boldsymbol{\Theta}_{ce}\boldsymbol{\zeta}_{ce}$, we have for the MMSE estimate $\hat{\mathbf{s}}_{ce} = E[\mathbf{s}|\hat{\mathbf{x}}]$

$$\hat{\mathbf{s}}_{ce} = E[\mathbf{s}|\mathbf{Q}_x^T \hat{\mathbf{x}}] = E[\mathbf{s}|\check{\mathbf{x}}_1] = \boldsymbol{\Sigma}_{s\check{x}_1}\boldsymbol{\Sigma}_{\check{x}_1\check{x}_1}^{-1}\check{\mathbf{x}}_1, \tag{1.85}$$

since $\mathbf{Q}_x^T$ is unitary and the last $M - k_{ce}$ entries of $\check{\mathbf{x}}$ are useless for estimating $\mathbf{s}$. It has been shown in (Schizas *et al.* 2005a) that the covariance matrix $\boldsymbol{\Sigma}_{\tilde{s}_{ce}\tilde{s}_{ce}} := E[(\mathbf{s} - \hat{\mathbf{s}}_{ce})(\mathbf{s} - \hat{\mathbf{s}}_{ce})^T] = \boldsymbol{\Sigma}_{ss} - \boldsymbol{\Sigma}_{s\check{x}_1}\boldsymbol{\Sigma}_{\check{x}_1\check{x}_1}^{-1}\boldsymbol{\Sigma}_{\check{x}_1 s}$ of $\tilde{\mathbf{s}}_{ce}$ is

$$\boldsymbol{\Sigma}_{\tilde{s}_{ce}\tilde{s}_{ce}} = \boldsymbol{\Sigma}_{ss} - \boldsymbol{\Sigma}_{sx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xs} + \boldsymbol{\Sigma}_{sx}\mathbf{Q}_x\boldsymbol{\Delta}_{ce}\mathbf{Q}_x^T\boldsymbol{\Sigma}_{xs}, \tag{1.86}$$

where $\boldsymbol{\Delta}_{ce} := \text{diag}\left(D_1^{ce}\lambda_{x,1}^{-2} \cdots D_N^{ce}\lambda_{x,M}^{-2}\right)$.

In Fig. 1.11 (Bottom) we depict the test channel for the EC scheme. The MMSE estimate $\hat{\mathbf{s}} = E[\mathbf{s}|\mathbf{x}]$ is followed by the test channel that results when applying rwf to a pre-whitened version of $\hat{\mathbf{s}}$, with rate R. Let $\boldsymbol{\Sigma}_{\hat{s}\hat{s}} = \mathbf{Q}_{\hat{s}}\boldsymbol{\Lambda}_{\hat{s}}\mathbf{Q}_{\hat{s}}^T$ be the eigenvalue decomposition for the covariance matrix of $\hat{\mathbf{s}}$, where $\boldsymbol{\Lambda}_{\hat{s}} = \text{diag}(\lambda_{\hat{s},1} \cdots \lambda_{\hat{s},p})$ and $\lambda_{\hat{s},1} \geq \cdots \geq \lambda_{\hat{s},p}$. Suppose now that the first k_{ec} entries of $\hat{\mathbf{s}}_w = \mathbf{Q}_{\hat{s}}^T\hat{\mathbf{s}}$ are assigned with non-zero rate and the rest are discarded. The MSE optimal encoder's output is given by $\mathbf{u}_{ec} = \mathbf{Q}_{\hat{s},k_{ec}}^T\hat{\mathbf{s}} + \boldsymbol{\zeta}_{ec}$, and the estimate $\hat{\mathbf{s}}_{ec}$ is

$$\hat{\mathbf{s}}_{ec} = \mathbf{Q}_{\hat{s},k_{ec}}\boldsymbol{\Theta}_{ec}\mathbf{Q}_{\hat{s},k_{ec}}^T\hat{\mathbf{s}} + \mathbf{Q}_{\hat{s},k_{ec}}\boldsymbol{\Theta}_{ec}\boldsymbol{\zeta}_{ec}, \tag{1.87}$$

where $\mathbf{Q}_{\hat{s},k_{ec}}$ is formed by the first k_{ec} columns of $\mathbf{Q}_{\hat{s}}$. For the $k_{ec} \times k_{ec}$ diagonal matrices $\mathbf{s}_{ec}$ and $\boldsymbol{\Sigma}_{\zeta_{ec}\zeta_{ec}}$ we have $[\mathbf{s}_{ec}]_{ii} = (\lambda_{\hat{s},i} - D_i^{ec})/\lambda_{\hat{s},i}$ and $[\boldsymbol{\Sigma}_{\zeta_{ec}\zeta_{ec}}]_{ii} = \lambda_{\hat{s},i}D_i^{ec}/(\lambda_{\hat{s},i} - D_i^{ec})$, where $D_i^{ec} := E[(\hat{\mathbf{s}}_w(i) - \hat{\mathbf{s}}_{ec,w}(i))^2]$, and $\hat{\mathbf{s}}_{ec,w} := \mathbf{Q}_{\hat{s}}^T\hat{\mathbf{s}}_{ec}$. Recall also that $D_i^{ec} = \left(\prod_{i=1}^{k_{ec}}\lambda_{\hat{s},i}\right)^{1/k_{ec}} 2^{\frac{-2R}{k_{ec}}}$ when $i = 1, \ldots, k_{ec}$ and $D_i^{ec} = \lambda_{\hat{s},i}$, for $i = k_{ec} + 1, \ldots, p$. Upon defining $\boldsymbol{\Delta}_{ec} := \text{diag}\left(D_1^{ec} \cdots D_p^{ec}\right)$, the covariance matrix of $\tilde{\mathbf{s}}_{ec}$ is given by (Schizas *et al.* 2005a)

$$\boldsymbol{\Sigma}_{\tilde{s}_{ec}\tilde{s}_{ec}} = \boldsymbol{\Sigma}_{ss} - \boldsymbol{\Sigma}_{sx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xs} + \mathbf{Q}_{\hat{s}}\boldsymbol{\Delta}_{ec}\mathbf{Q}_{\hat{s}}^T. \tag{1.88}$$

The MMSE associated with CE and EC is given, respectively, by [c.f. (1.86) and (1.88)]

$$\begin{aligned} D_{ce}(R) &:= \text{tr}(\boldsymbol{\Sigma}_{\tilde{s}_{ce}\tilde{s}_{ce}}) = J_o + \epsilon_{ce}(R), \\ D_{ec}(R) &:= \text{tr}(\boldsymbol{\Sigma}_{\tilde{s}_{ec}\tilde{s}_{ec}}) = J_o + \epsilon_{ec}(R), \end{aligned} \tag{1.89}$$

where $\epsilon_{ce}(R) := \text{tr}(\boldsymbol{\Sigma}_{sx}\mathbf{Q}_x\boldsymbol{\Delta}_{ce}\mathbf{Q}_x^T\boldsymbol{\Sigma}_{xs})$, $\epsilon_{ec}(R) := \text{tr}(\mathbf{Q}_{\hat{s}}\boldsymbol{\Delta}_{ec}\mathbf{Q}_{\hat{s}}^T)$, and the quantity $J_o := \text{tr}(\boldsymbol{\Sigma}_{ss} - \boldsymbol{\Sigma}_{sx}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\Sigma}_{xs})$ is the MMSE achieved when estimating $\mathbf{s}$ based on $\mathbf{x}$, without source encoding ($R \to \infty$). Since J_o is common to both EC and CE it is important to compare $\epsilon_{ce}(R)$ with $\epsilon_{ec}(R)$ in order to determine which estimation scheme achieves the smallest MSE. The following proposition provides such an asymptotic comparison:

Proposition 1.7.1 *If $R > R_{th} := \frac{1}{2}\max\{\log_2\left(\left(\prod_{i=1}^{\rho} \lambda_{x,i}\right)/\sigma^{2\rho}\right), \log_2\left(\left(\prod_{i=1}^{\rho} \lambda_{\hat{s},i}\right) / (\lambda_{\hat{s},\rho})^{\rho}\right)\}$, then it holds that $\epsilon_{ce}(R) = \gamma_1 2^{-2R/M}$ and $\epsilon_{ec}(R) = \gamma_2 2^{-2R/\rho}$, where γ_1 and γ_2 are constants.*

An immediate consequence of Proposition 1.7.1 is that the MSE for EC converges as $R \to \infty$ to J_o with rate $O(2^{-2R/\rho})$. The MSE of CE converges likewise, but with rate $O(2^{-2R/M})$. For the typical case $M > \rho$, EC approaches the lower bound J_o faster than CE, implying correspondingly a more efficient usage of the available rate R. This is intuitively reasonable since CE compresses $\mathbf{x}$, which contains the noise $\mathbf{w}$. Since the last $M - \rho$ eigenvalues of $\mathbf{\Sigma}_{xx}$ equal the noise variance σ^2, part of the available rate is consumed to compress the noise. On the contrary, the MMSE estimator $\hat{\mathbf{s}}$ in EC suppresses significant part of the noise. For the special case of a scalar data model ($M = p = 1$) it has been shown (Schizas *et al.* 2005a) that $D_{ec}(R) = D_{ce}(R)$, while for the vector and matrix models ($M > 1$ and/or $p > 1$) we have determined appropriate threshold rates R_{th} have been determined such that $D_{ce}(R) > D_{ec}(R)$ for $R > R_{th}$.

If the SNR is defined as SNR $= \text{tr}(\mathbf{H}\mathbf{\Sigma}_{ss}\mathbf{H}^T)/M\sigma^2$, it is possible to compare the MMSE when estimating $\mathbf{s}$ using the CE and EC schemes; see Fig. 1.12 (Left). With $\mathbf{\Sigma}_{ss} = \sigma_s^2\mathbf{I}_p$, $p = 4$ and $M = 40$, we observe that beyond a threshold rate, the distortion of EC converges to J_o faster than that of CE, which corroborates Proposition 1.7.1.

The analysis so far raises the question whether EC is MSE optimal. We have seen that this is the case when estimating $\mathbf{s}$ with a given rate R without forcing any assumption about M and p. A related claim has been reported in (Sakrison 1968; Wolf and Ziv 1970) for $M = p$. The extension to $M \neq p$ established in (Schizas *et al.* 2005a) can be summarized as follows:

Proposition 1.7.2 *The D-R function when estimating $\mathbf{s}$ based on $\mathbf{x}$ can be expressed as*

$$D_s(R) = \min_{\substack{p(\hat{\mathbf{s}}_R|\mathbf{x})\\ I(\mathbf{x};\hat{\mathbf{s}}_R)\leq R}} E[\|\mathbf{s} - \hat{\mathbf{s}}_R\|^2] = E[\|\tilde{\mathbf{s}}\|^2] + \min_{\substack{p(\hat{\mathbf{s}}_R|\hat{\mathbf{s}})\\ I(\hat{\mathbf{s}};\hat{\mathbf{s}}_R)\leq R}} E[\|\hat{\mathbf{s}} - \hat{\mathbf{s}}_R\|^2], \tag{1.90}$$

where $\hat{\mathbf{s}} = \mathbf{\Sigma}_{sx}\mathbf{\Sigma}_{xx}^{-1}\mathbf{x}$ is the MMSE estimator, and $\tilde{\mathbf{s}}$ is the corresponding MMSE.

Proposition 1.7.2 reveals that the optimal means of estimating $\mathbf{s}$ is to first form the optimal MMSE estimate $\hat{\mathbf{s}}$ and then apply optimal rate-distortion encoding to this estimate. The lower bound on this distortion when $R \to \infty$ is $J_o = E[\|\tilde{\mathbf{s}}\|^2]$, which is intuitively appealing. The D-R function in (1.90) is achievable, because the rightmost term in (1.90) corresponds to the D-R function for reconstructing the MMSE estimate $\hat{\mathbf{s}}$ which is known to be achievable using random coding; see e.g., (Berger 1971, p. 66).

1.7.2 Distortion-Rate for Distributed Estimation

Let us now consider the D-R function for estimating $\mathbf{s}$ in a multi-sensor setup, under a total available rate R which has to be shared among all sensors. Because analytical specification of the D-R function in this case remains intractable, we will present an alternating algorithm that numerically determines an achievable upper bound for it. Combining this upper bound with the non-achievable lower bound corresponding to an equivalent single-sensor setup, and applying the MMSE optimal EC scheme, will provide a region where the D-R function lies in. For simplicity in exposition, we confine ourselves to a two-sensor setup, but the results apply to any finite $N > 2$.

To this end, consider the following single-letter characterization of the upper bound on the D-R function:

$$\bar{D}(R) = \min_{\{p(\mathbf{u}_n|\mathbf{x}_n)\}_{n=0}^{1},\hat{\mathbf{s}}_R} E_{p(\mathbf{s},\{\mathbf{u}_n\}_{n=0}^{1})}[\|\mathbf{s} - \hat{\mathbf{s}}_R\|^2], \quad \text{s. to } I(\mathbf{x}; \{\mathbf{u}_n\}_{n=0}^{1}) \leq R, \tag{1.91}$$

where the minimization is w.r.t. $\{p(\mathbf{u}_n|\mathbf{x}_n)\}_{n=0}^{1}$ and $\hat{\mathbf{s}}_R := \hat{\mathbf{s}}_R(\mathbf{u}_0, \mathbf{u}_1)$. Achievability of $\bar{D}(R)$ can be established by readily extending to the vector case the scalar results in (Chen *et al.* 2004). To carry out the minimization in (1.91), we will develop an alternating scheme whereby $\mathbf{u}_1$ is treated as side information that is available at the decoder when optimizing (1.91) w.r.t. $p(\mathbf{u}_0|\mathbf{x}_0)$ and $\hat{\mathbf{s}}_R(\mathbf{u}_0, \mathbf{u}_1)$. The side information $\mathbf{u}_1$ is considered as the output of an optimal rate-distortion encoder applied to $\mathbf{x}_1$ for estimating $\mathbf{s}$, without taking into account $\mathbf{x}_0$. Since $\mathbf{x}_1$ is Gaussian, the side information will have the form (c.f. subsection 1.7.1) $\mathbf{u}_1 = \mathbf{Q}_1\mathbf{x}_1 + \boldsymbol{\zeta}_1$, where $\mathbf{Q}_1 \in \mathbb{R}^{k_1 \times M_1}$ and $k_1 \leq M_1$, due to the rate constrained encoding of $\mathbf{x}_1$. Recall that $\boldsymbol{\zeta}_1$ is uncorrelated with $\mathbf{x}_1$ and Gaussian; i.e., $\boldsymbol{\zeta}_1 \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\zeta_1\zeta_1})$.

Based on $\boldsymbol{\psi} := [\mathbf{x}_0^T \;\; \mathbf{u}_1^T]^T$, the optimal estimator for $\mathbf{s}$ is the MMSE one: $\hat{\mathbf{s}} = E[\mathbf{s}|\boldsymbol{\psi}] = \boldsymbol{\Sigma}_{s\psi}\boldsymbol{\Sigma}_{\psi\psi}^{-1}$ $\boldsymbol{\psi} = \mathbf{L}_0\mathbf{x}_0 + \mathbf{L}_1\mathbf{u}_1$, where $\mathbf{L}_0$, $\mathbf{L}_1$ are $p \times M_0$ and $p \times k_1$ matrices so that $\boldsymbol{\Sigma}_{s\psi}\boldsymbol{\Sigma}_{\psi\psi}^{-1} = [\mathbf{L}_0 \;\; \mathbf{L}_1]$. If $\tilde{\mathbf{s}}$ is the corresponding MSE, then $\mathbf{s} = \hat{\mathbf{s}} + \tilde{\mathbf{s}}$, where $\tilde{\mathbf{s}}$ is uncorrelated with $\boldsymbol{\psi}$ due to the orthogonality principle. Noticing also that $\hat{\mathbf{s}}_R(\mathbf{u}_0, \mathbf{u}_1)$ is uncorrelated with $\tilde{\mathbf{s}}$ because it is a function of $\mathbf{x}_0$ and $\mathbf{u}_1$, we have $E[\|\mathbf{s} - \hat{\mathbf{s}}_R(\mathbf{u}_0, \mathbf{u}_1)\|^2] = E[\|\hat{\mathbf{s}} - \hat{\mathbf{s}}_R(\mathbf{u}_0, \mathbf{u}_1)\|^2] + E[\|\tilde{\mathbf{s}}\|^2]$, or,

$$E[\|\mathbf{s} - \hat{\mathbf{s}}_R(\mathbf{u}_0, \mathbf{u}_1)\|^2] = E[\|\mathbf{L}_0\mathbf{x}_0 - (\hat{\mathbf{s}}_R(\mathbf{u}_0, \mathbf{u}_1) - \mathbf{L}_1\mathbf{u}_1)\|^2] + E[\|\tilde{\mathbf{s}}\|^2]. \tag{1.92}$$

Clearly, it holds that $I(\mathbf{x}; \mathbf{u}_0, \mathbf{u}_1) = R_1 + I(\mathbf{x}_0; \mathbf{u}_0) - I(\mathbf{u}_1; \mathbf{u}_0)$, where $R_1 := I(\mathbf{x}; \mathbf{u}_1)$ is the rate consumed to form the side information $\mathbf{u}_1$ and the rate constraint in (1.91) becomes $I(\mathbf{x}; \mathbf{u}_0, \mathbf{u}_1) \leq R \Leftrightarrow I(\mathbf{x}_0; \mathbf{u}_0) - I(\mathbf{u}_1; \mathbf{u}_0) \leq R - R_1 := R_0$. The new signal of interest in (1.92) is $\mathbf{L}_0\mathbf{x}_0$; thus, $\mathbf{u}_0$ has to be a function of $\mathbf{L}_0\mathbf{x}_0$. Using also the fact that $\mathbf{x}_0 \rightarrow \mathbf{L}_0\mathbf{x}_0 \rightarrow \mathbf{u}_0$ constitutes a Markov chain, it is possible to obtain from (1.91) the D-R upper bound (Schizas *et al.* 2005a):

$$\bar{\bar{D}}(R_0) = E[\|\tilde{\mathbf{s}}\|^2] + \min_{\substack{p(\mathbf{u}_0|\mathbf{L}_0\mathbf{x}_0), \hat{\mathbf{s}}_R \\ I(\mathbf{L}_0\mathbf{x}_0;\mathbf{u}_0) - I(\mathbf{u}_0;\mathbf{u}_1) \leq R_0}} E[\|\mathbf{L}_0\mathbf{x}_0 - \tilde{\mathbf{s}}_{R,01}(\mathbf{u}_0, \mathbf{u}_1)\|^2], \tag{1.93}$$

where $\tilde{\mathbf{s}}_{R,01}(\mathbf{u}_0, \mathbf{u}_1) := \hat{\mathbf{s}}_R(\mathbf{u}_0, \mathbf{u}_1) - \mathbf{L}_1\mathbf{u}_1$. Through (1.93) we can determine an achievable D-R region, having available rate R_0 at the encoder and side information $\mathbf{u}_1$ at the decoder. Since $\mathbf{x}_0$ and $\mathbf{u}_1$ are jointly Gaussian, the Wyner-Ziv result applies (Wyner and Ziv 1976), which allows one to consider that $\mathbf{u}_1$ is available both at the decoder and the encoder. This, in turn, permits re-writing the (1.93) as (Schizas *et al.* 2005a)

$$\bar{\bar{D}}(R_0) = \min_{\substack{p(\hat{\mathbf{s}}_{R,01}|\tilde{\mathbf{s}}_0) \\ I(\tilde{\mathbf{s}}_0;\hat{\mathbf{s}}_{R,01}) \leq R_0}} E[\|\tilde{\mathbf{s}}_0 - \hat{\mathbf{s}}_{R,01}(\mathbf{u}_0, \mathbf{u}_1)\|^2] + E[\|\tilde{\mathbf{s}}\|^2], \tag{1.94}$$

where $\hat{\mathbf{s}}_{R,01}(\mathbf{u}_0, \mathbf{u}_1) = \hat{\mathbf{s}}_R(\mathbf{u}_0, \mathbf{u}_1) - \mathbf{L}_1\mathbf{u}_1 - E[\mathbf{L}_0\mathbf{x}_0|\mathbf{u}_1]$ and $\tilde{\mathbf{s}}_0 = \mathbf{L}_0\mathbf{x}_0 - E[\mathbf{L}_0\mathbf{x}_0|\mathbf{u}_1]$.

Notice that (1.94) is the D-R function for reconstructing the MSE $\tilde{\mathbf{s}}_0$ with rate R_0. Since $\tilde{\mathbf{s}}_0$ is Gaussian, we can readily apply rwf to the pre-whitened $\mathbf{Q}_{\tilde{s}_0}^T\tilde{\mathbf{s}}_0$ for determining $\bar{\bar{D}}(R_0)$ and the corresponding test channel that achieves $\bar{\bar{D}}(R_0)$. Through the latter, and considering the next eigenvalue decomposition $\boldsymbol{\Sigma}_{\tilde{s}_0\tilde{s}_0} = \mathbf{Q}_{\tilde{s}_0}$ diag$(\lambda_{\tilde{s}_0,1} \cdots \lambda_{\tilde{s}_0,p})\mathbf{Q}_{\tilde{s}_0}^T$, it follows that the first encoder's output that minimizes (1.91) has the form:

$$\mathbf{u}_0 = \mathbf{Q}_{\tilde{s}_0,k_0}^T\mathbf{L}_0\mathbf{x}_0 + \boldsymbol{\zeta}_0 = \mathbf{Q}_0\mathbf{x}_0 + \boldsymbol{\zeta}_0, \tag{1.95}$$

where $\mathbf{Q}_{\tilde{s}_0,k_0}$ denotes the first k_0 columns of $\mathbf{Q}_{\tilde{s}_0}$, k_0 is the number of $\mathbf{Q}_{\tilde{s}_0}^T\tilde{\mathbf{s}}_0$ entries that are assigned with non-zero rate, and $\mathbf{Q}_0 := \mathbf{Q}_{\tilde{s}_0,k_0}^T\mathbf{L}_0$. The $k_0 \times 1$ AWGN $\boldsymbol{\zeta}_0 \sim \mathcal{N}\ (\mathbf{0}, \boldsymbol{\Sigma}_{\zeta_0\zeta_0})$ is uncorrelated with $\mathbf{x}_0$. Additionally, we have $[\boldsymbol{\Sigma}_{\zeta_0\zeta_0}]_{ii} = \lambda_{\tilde{s}_0,i}D_i^0/(\lambda_{\tilde{s}_0,i} - D_i^0)$, where $D_i^0 = \left(\prod_{i=1}^{k_0}\lambda_{\tilde{s}_0,i}\right)^{1/k_0}$ $2^{-2R_0/k_0}$, for $i = 1, \ldots, k_0$, and $D_i^0 = \lambda_{\tilde{s}_0,i}$ when $i = k_0 + 1, \ldots, p$. This way, we are able to determine also $p(\mathbf{u}_0|\mathbf{x}_0)$. The reconstruction function has the form:

$$\begin{aligned}\hat{\mathbf{s}}_R(\mathbf{u}_0, \mathbf{u}_1) = \mathbf{Q}_{\tilde{s}_0,k_0}\boldsymbol{\Theta}_0\mathbf{u}_0 + \mathbf{L}_0\boldsymbol{\Sigma}_{x_0u_1}\boldsymbol{\Sigma}_{u_1u_1}^{-1}\mathbf{u}_1 + \mathbf{L}_1\mathbf{u}_1 \\ - \mathbf{Q}_{\tilde{s}_0,k_0}\boldsymbol{\Theta}_0\mathbf{Q}_{\tilde{s}_0,k_0}^T\mathbf{L}_0\boldsymbol{\Sigma}_{x_0u_1}\boldsymbol{\Sigma}_{u_1u_1}^{-1}\mathbf{u}_1,\end{aligned} \tag{1.96}$$

where $[\mathbf{\Theta}_0]_{ii} = \lambda_{\tilde{s}_0,i} D_i^0 / (\lambda_{\tilde{s}_0,i} - D_i^0)$, and the MMSE is $\bar{\bar{D}}(R_0) = \sum_{j=1}^{p} D_j^0 + E[\|\tilde{\mathbf{s}}\|^2]$.

The approach in this subsection can be applied in an alternating fashion from sensor to sensor in order to determine appropriate $p(\mathbf{u}_n|\mathbf{x}_n)$, for $n = 0, 1$, and $\hat{\mathbf{s}}_R(\mathbf{u}_0, \mathbf{u}_1)$ that at best globally minimize (1.93). The conditional pdfs can be determined by finding the appropriate covariances $\mathbf{\Sigma}_{\zeta_n \zeta_n}$. Furthermore, by specifying the optimal $\mathbf{Q}_0$ and $\mathbf{Q}_1$, characterization of the encoders' structure is obtained. In Fig. 1.12 (Right), we plot the non-achievable lower bound which corresponds to one sensor having available the entire $\mathbf{x}$ and using the optimal EC scheme. Moreover, we plot an achievable D-R upper bound determined by letting the n-th sensor form its local estimate $\hat{\mathbf{s}}_n = E[\mathbf{s}|\mathbf{x}_n]$, and then apply optimal rate-distortion encoding to $\hat{\mathbf{s}}_n$. If $\hat{\mathbf{s}}_{R,0}$ and $\hat{\mathbf{s}}_{R,1}$ are the reconstructed versions of $\hat{\mathbf{s}}_0$ and $\hat{\mathbf{s}}_1$, respectively, then the decoder at the FC forms the final estimate $\hat{\mathbf{s}}_R = E[\mathbf{s}|\hat{\mathbf{s}}_{R,0}, \hat{\mathbf{s}}_{R,1}]$. We also plot the achievable D-R region determined numerically by Algorithm 2. For each rate, the smallest distortion is recorded after 500 executions of the algorithm simulated with $\mathbf{\Sigma}_{ss} = \mathbf{I}_p, p = 4$, and $M_0 = M_1 = 20$, at SNR $= 2$. We observe that the algorithm provides a tight upper bound of the achievable D-R region, which combined with the non-achievable lower bound (solid line) effectively reduces the 'uncertainty region' where the D-R function lies.

Algorithm 2 :

```
Initialize $\mathbf{Q}_0^{(0)}, \mathbf{Q}_1^{(0)}, \mathbf{\Sigma}_{\zeta_0\zeta_0}^{(0)}, \mathbf{\Sigma}_{\zeta_1\zeta_1}^{(0)}$ by applying optimal D-R encoding to each sensor's test channel independently. For a total rate $R$, generate $J$ random increments $\{r(m)\}_{m=0}^{J}$, such that $0 \le r(m) \le R$ and $\sum_{m=0}^{M} r(m) = R$. Set $R_0(0) = R_1(0) = 0$.
for $j = 1,J$ do
  Set $R(j) = \sum_{l=0}^{j} r(l)$
  for $n = 0,1$ do
    $\bar{n} = |n-1|$ %The complementary index
    $R_0(j) = I(\mathbf{x}; \mathbf{u}_{\bar{n}}^{(j)})$
    We use $\mathbf{Q}_{\bar{n}}^{(j-1)}, \mathbf{\Sigma}_{\zeta_{\bar{n}}\zeta_{\bar{n}}}^{(j-1)}, R(j), R_0(j)$ to determine $\mathbf{Q}_n^{(j)}, \mathbf{\Sigma}_{\zeta_n\zeta_n}^{(j)}$ and $\bar{\bar{D}}(R_n(j))$
  end for
  Update matrices $\mathbf{Q}_l^{(j)}, \mathbf{\Sigma}_{\zeta_l\zeta_l}^{(j)}$ that result the smallest distortion $\bar{\bar{D}}(R_l(j))$, with $l \in [0, 1]$
  Set $R_l(j) = R(j) - I(\mathbf{x}; \mathbf{u}_{\bar{l}}^{(j)})$ and $R_{\bar{l}}(j) = I(\mathbf{x}; \mathbf{u}_{\bar{l}}^{(j)})$.
end for
```

1.7.3 D-R Upper Bound via Convex Optimization

In this subsection we outline an alternative approach which relies on convex optimization techniques to obtain numerically an upper bound of the D-R region (Xiao *et al.* 2005b). The idea is to calculate the Berger-Tung achievable D-R region (Berger 1977) for the vector Gaussian CEO problem, and subsequently determine the minimum sum rate $R_\Sigma = \sum_{n=0}^{N-1} R_n$ such that the estimation MSE satisfies $\mathrm{tr}(E[(\mathbf{s} - \hat{\mathbf{s}}_R)(\mathbf{s} - \hat{\mathbf{s}}_R)^T]) < D$, where $\hat{\mathbf{s}}_R = E[\mathbf{s}|\{\mathbf{u}_n\}_{n=0}^{N-1}]$ and D is the desired upper bound on the distortion. The Berger-Tung achievable region is calculated after having the encoders' output to have in form $\mathbf{u}_n = \mathbf{x}_n + \boldsymbol{\zeta}_n$, where $\boldsymbol{\zeta}_n \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_{\zeta_n\zeta_n})$ are independent of $\mathbf{x}_n$, for $n = 0, 1, \ldots, N-1$. Furthermore, the sum rate can be expressed as a function of $\mathbf{H}_n$ and $\mathbf{\Sigma}_{\zeta_n\zeta_n}$ (Xiao *et al.* 2005b)

$$R_\Sigma = 0.5 \log \left(\det \left(\mathbf{I}_p + \sum_{n=0}^{N-1} \mathbf{H}_n^T (\mathbf{I}_{M_n} + \mathbf{\Sigma}_{\zeta_n\zeta_n})^{-1} \mathbf{H}_n \right) \prod_{n=0}^{N-1} \det(\mathbf{I}_{M_n} + \mathbf{\Sigma}_{\zeta_n\zeta_n}^{-1}) \right).$$

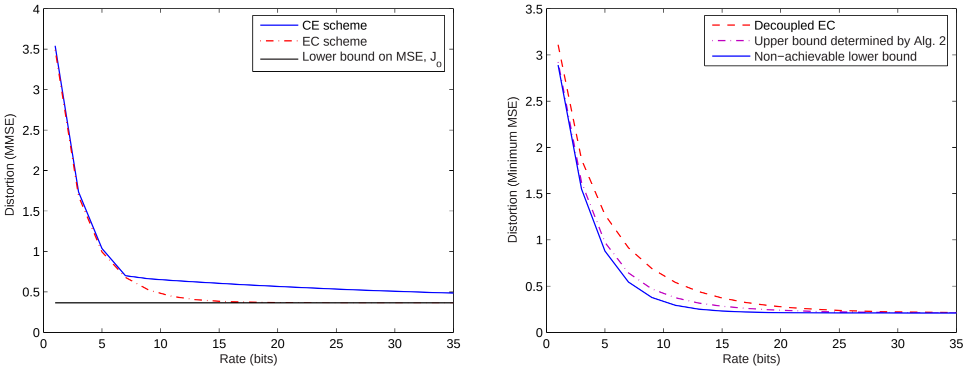

Figure 1.12 (Left): D-R region for EC and CE at SNR $= 2$; (Right): Distortion-rate bounds for estimating $\mathbf{s}$ in a two-sensor setup.

The D-R upper bound is obtained as the optimal solution of the following minimization problem ($\succeq$ denotes positive semidefiniteness)

$$\min_{\{\boldsymbol{\Sigma}_{\zeta_n\zeta_n}\}_{n=0}^{N-1}} R_{\Sigma}, \text{ s. to } \boldsymbol{\Sigma}_{\zeta_n\zeta_n} \succeq \mathbf{0}, \text{ tr}(\boldsymbol{\Sigma}_{\tilde{s}_R\tilde{s}_R}) \leq D, \tag{1.97}$$

where $\boldsymbol{\Sigma}_{\tilde{s}_R\tilde{s}_R} := E[(\mathbf{s} - \hat{\mathbf{s}}_R)(\mathbf{s} - \hat{\mathbf{s}}_R)^T] = (\mathbf{I}_p + \sum_{n=0}^{N-1} \mathbf{H}_n^T(\mathbf{I}_{M_n} + \boldsymbol{\Sigma}_{\zeta_n\zeta_n})^{-1}\mathbf{H}_n)^{-1}$.

Although, the minimization problem in (1.97) is not convex, (Xiao *et al.* 2005b) has shown that (1.97) is equivalent to the following convex formulation:

$$\min_{\boldsymbol{\Sigma}_{\tilde{s}_R\tilde{s}_R}, \{\boldsymbol{\Sigma}_{\zeta_n\zeta_n}\}_{n=0}^{N-1}} -0.5 \log\det(\boldsymbol{\Sigma}_{\tilde{s}_R\tilde{s}_R}) + 0.5 \sum_{n=0}^{N-1} \log\det(\mathbf{I}_{M_n} + \boldsymbol{\Sigma}_{\zeta_n\zeta_n}^{-1}), \tag{1.98}$$

$$\text{s. to tr}(\boldsymbol{\Sigma}_{\tilde{s}_R\tilde{s}_R}) \leq D, \boldsymbol{\Sigma}_{\zeta_n\zeta_n} \succeq \mathbf{0}, \boldsymbol{\Sigma}_{\tilde{s}_R\tilde{s}_R} \succeq \mathbf{0}, (\mathbf{I}_p + \sum_{n=0}^{N-1} \mathbf{H}_n^T(\mathbf{I}_{M_n} + \boldsymbol{\Sigma}_{\zeta_n\zeta_n})^{-1}\mathbf{H}_n)^{-1} \preceq \boldsymbol{\Sigma}_{\tilde{s}_R\tilde{s}_R},$$

which is solved numerically using the interior point method (Boyd and Vandenberghe 2004).

1.8 Closing Comments

We considered distributed estimation using wireless sensor networks and demonstrated that under limited resources the seemingly unrelated problems of dimensionality reduction, compression, quantization and estimation are actually intertwined due to the distributed nature of sensor networks.

We started with parameter estimation under severe bandwidth constraints that were adhered to by *quantizing* each sensor's observation to one or a few bits. By jointly accounting for the unique quantization-estimation tradeoffs present, these bit(s) per sensor were first used to derive distributed maximum likelihood estimators (MLEs) for scalar mean-location estimation in the presence of generally non-Gaussian noise when the noise pdf is completely known; subsequently, when the pdf is known except for a number of unknown parameters; and finally, when the noise pdf is unknown. In all three cases, the resulting estimators turned out to exhibit comparable variances that can come surprisingly close to the variance of the clairvoyant estimator which relies on unquantized observations. This happens when the SNR capturing both quantization and noise effects assumes low-to-moderate values. Analogous claims were established for practical generalizations in the multivariate and colored noise cases for distributed estimation of vector deterministic and random parameters. Therein, MLE and MAP estimators were formed via numerical search but the log-likelihoods were proved to be concave thus ensuring fast convergence to the unique global maximum.

We also pursued a related but distinct approach where the bandwidth constraint is adhered to by reduced-dimensionality observations. We dealt with non-ideal channel links that are characterized by multiplicative fading and additive noise. When data across sensors are uncorrelated, we presented global MSE optimal schemes in closed-form and proved that they implement estimation followed by compression per sensor. For correlated sensor observations, we outlined a block coordinate descent algorithm which guarantees convergence at least to a stationary point of the associated mean-square error cost. The optimal estimators allocate properly the prescribed power following a waterfilling-like principle. Fundamental MSE limits were finally studied through the D-R function for estimating a random vector in a single-sensor setup, where an estimate-first compress-afterwards approach was turns out to be optimal. An alternating algorithm was also outlined for determining numerically a D-R upper bound in the distributed multi-sensor setup. Using this upper bound in conjunction with the non-achievable lower bound, determined through the single-sensor D-R function, yielded a tight region, where the D-R function for distributed estimation lies in.

1.9 Further Reading

The problem of estimation based on quantized observations was studied in early works by (Gubner 1993) and (Lam and Reibman 1993) and revisited in the context of distributed estimation using WSNs in (Papadopoulos *et al.* 2001). The material on Sections 1.1 – 1.4 is based on results from (Ribeiro and Giannakis 2006a) and (Ribeiro and Giannakis 2006b), while the material in Section 1.5 has been reported in (Sha *et al.* 2005). When the noise pdf is unknown, the problem of estimation based on severely quantized data has been also studied by (Luo 2005a), (Luo 2005b) and (Luo and Xiao 2005) where the notion of universal estimators was introduced. A recent extension of the material covered in these sections to state estimation of dynamical stochastic processes can be found in (Ribeiro *et al.* 2007).

Distributed estimation via dimensionality reduction has been also considered in (Zhu *et al.* 2005), (Gastpar *et al.* 2004) and (Zhang *et al.* 2003) for ideal channel links and/or Gaussian data models. Detailed derivations of what was presented in Section 1.6 can be found in (Schizas *et al.* 2005b). When it comes to rate constrained distributed estimation D-R bounds for the Gaussian CEO setup, results are due to (Oohama 1998) and (Chen *et al.* 2004) when $M = p$. The results in Section 1.7 are from (Schizas *et al.* 2005a) and (Xiao *et al.* 2005b).

A different approach to reduce communication costs in distributed estimation is to allow communication between one-hop neighbors only and let the sensors converge to a common estimate. In (Xiao and Boyd 2004) estimation is considered tantamount to convergence to the steady state distribution of a Markov chain. In (Schizas *et al.* 2006) estimation is shown equivalent to distributed optimization of a convex argument. A related approach can be found in (Barbarossa and Scuttari 2006) where the WSN is modelled as a network of coupled oscillators. A different estimation approach using hidden Markov fields is reported in (Dogandžić 2006).

Bibliography

S. Barbarossa and G. Scutari, *Sensor networks with decentralized maximum likelihood estimation capabilities through self-synchronization of locally coupled oscillators* . IEEE Trans. on Signal Processing, Dec. 2005 (submitted).

T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression.* Prentice Hall, 1971.

T. Berger, *Multiterminal Source Coding,* in Lectures Presented at CISM Summer School on the Info. Theory Approach to Comm., July 1977.

T. Berger, Z. Zhang, and H. Viswanathan, *The CEO problem,* IEEE Transactions on Information Theory, vol. 42, pp. 406–411, May 1996.

S. Boyd and L. Vandenberghe, *Convex Optimization.* Cambridge University Press, 2004.

J. Chen, X. Zhang, T. Berger, and S. B. Wicker, *An Upper Bound on the Sum-Rate Distortion Function and Its Corresponding Rate Allocation Schemes for the CEO Problem,* IEEE Journal on Selected Areas in Communications, pp. 406–411, August 2004.

T. Cover and J. Thomas, *Elements of Information Theory.* John Wiley and Sons, 2nd edition ed., 1991.

A. Dogandžić and B. Zhang, *Distributed Estimation and Detection for Sensor Networks Using Hidden Markov Random Field Models* . IEEE Trans. on Signal Processing, vol. 54, pp. 3200–3215, August 2006.

M. Gastpar, P. L. Draggoti, and M. Vetterli, *The distributed Karhunen-Loève transform,* IEEE Transactions on Information Theory, submitted Nov. 2004, available at http://www.eecs.berkeley.edu/∼gastpar/.

J. Gubner, *Distributed Estimation and Quantization,* IEEE Transactions on Information Theory, vol. 39, pp. 1456–1459, 1993.

S. M. Kay, *Fundamentals of Statistical Signal Processing - Estimation Theory.* Prentice Hall, 1993.

W. Lam and A. Reibman, *Quantizer design for decentralized systems with communication constraints,* IEEE Transactions on Communications, vol. 41, pp. 1602–1605, Aug. 1993.

Z.-Q. Luo, *An isotropic universal decentralized estimation scheme for a bandwidth constrained ad hoc sensor network,* IEEE Journal on Selected Areas in Communications, vol. 23, pp. 735–744, April 2005.

Z.-Q. Luo, *Universal Decentralized Estimation in a Bandwidth Constrained Sensor Network,* IEEE Transactions on Information Theory, vol. 51, pp. 2210–2219, June 2005.

Z.-Q. Luo, G. B. Giannakis, and S. Zhang, *Optimal linear decentralized estimation in a bandwidth constrained sensor network,* in Proc. of the Intl. Symp. on Info. Theory, pp. 1441–1445, Adelaide, Australia, Sept. 4-9 2005.

Z.-Q. Luo and J.-J. Xiao, *Decentralized estimation in an inhomogeneous sensing environment,* IEEE Transactions on Information Theory, vol. 51, pp. 3564 –3575, October 2005.

A. Mainwaring, D. Culler, J. Polastre, R. Szewczyk, and J. Anderson, *Wireless sensor networks for habitat monitoring,* in Proc. of the 1st ACM International Workshop on Wireless Sensor Networks and Applications, vol. 3, pp. 88–97, Atlanta, Georgia, 2002.

Y. Oohama, *The Rate-Distortion Function for the Quadratic Gaussian CEO Problem,* IEEE Transactions On Information Theory, pp. 1057–1070, May 1998.

A. Pandya, A. Kansal, G. Pottie, and M. Srivastava, *Fidelity and Resource Sensitive Data Gathering,* in Proc. of the 42nd Allerton Conference, Allerton, IL, September 2004.

H. Papadopoulos, G. Wornell, and A. Oppenheim, *Sequential signal encoding from noisy measurements using quantizers with dynamic bias control,* IEEE Transactions on Information Theory, vol. 47, pp. 978–1002, 2001.

A. Ribeiro and G. B. Giannakis, *Bandwidth-Constrained Distributed Estimation for Wireless Sensor Networks, Part I: Gaussian Case,* IEEE Transactions on Signal Processing, vol. 54, pp. 1131–1143, March 2006.

A. Ribeiro and G. B. Giannakis, *Bandwidth-Constrained Distributed Estimation for Wireless Sensor Networks, Part II: Unknown pdf,* IEEE Transactions on Signal Processing, vol. 54, pp. 2784–2796, July 2006.

A. Ribeiro, G. B. Giannakis, and S. I. Roumeliotis, *SOI-KF: Distributed Kalman Filtering with Low-Cost Communications using the Sign Of Innovations,* IEEE Transactions on Signal Processing, 2007, to appear.

D. J. Sakrison, *Source encoding in the presence of random disturbance,* IEEE Transactions on Information Theory, pp. 165–167, January 1968.

I. D. Schizas, G. B. Giannakis, and N. Jindal, *Distortion-Rate Analysis for Distributed Estimation with Wireless Sensor Networks,* IEEE Transactions On Information Theory, submitted December 2005. available at http://spincom.ece.umn.edu/.

I. D. Schizas, G. B. Giannakis, and Z.-Q. Luo, *Distributed estimation using reduced dimensionality sensor observations,* IEEE Transactions on Signal Processing, submitted November 2005. available at http://spincom.ece.umn.edu/.

I. D. Schizas, A. Ribeiro, and G. B. Giannakis, *Distributed Estimation with Ad Hoc Wireless Sensor Networks,* Proc. of XIV European Sign. Proc. Conf., Florence, Italy, Sept. 4-8, 2006.

A. F. Shah, A. Ribeiro, and G. B. Giannakis, *Bandwidth-Constrained MAP Estimation for Wireless Sensor Networks,* Conference Record of the Thirty-Ninth Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, October 28 - November 1, 2005, Pages: 215 - 219.

H. L. Van Trees, *Detection, Estimation, and Modulation Theory.* John Wiley and Sons, first ed., 1968.

H. Viswanathan and T. Berger, *The Quadratic Gaussian CEO Problem, IEEE Transactions on Information Theory*, pp. 1549–1559, September 1997.

J. Wolf and J. Ziv, *Transmission of noisy information to a noisy receiver with minimum distortion,* IEEE Transactions on Information Theory, pp. 406–411, July 1970.

A. Wyner and J. Ziv, *The Rate-Distortion Function for Source Coding with Side Information at the Decoder,* IEEE Trans. on Info. Theory, pp. 1–10, January 1976.

J.-J. Xiao, Z.-Q. Luo and G. B. Giannakis, *Performance bounds for the rate-constrained universal decentralized estimators in sensor networks,* in Proc. of the IEEE Workshop on Signal Processing Advances in Wireless Communications, New York, NY, 5-8 June 2005, pp. 126–130.

J.-J. Xiao and Z.-Q. Luo, *Optimal rate and power allocation in Gaussian vector CEO problem,* in IEEE Int. Workshop Comp. Advances in Multi-Sensor Adaptive Processing, Puerto Vallarta, Mexico, 13-15 December 2005

L. Xiao and S. Boyd, *Fast Linear Iterations for Distributed Averaging,* Systems and Control Letters, vol. 53, pp. 65 78, 2004.

K. Zhang, X. R. Li, P. Zhang, and H. Li, *Optimal linear estimation fusion–Part VI: Sensor data compression,* in Proc. of the Intl. Conf. on Info. Fusion, pp. 221–228, Queensland, Australia 2003.

Y. Zhu, E. Song, J. Zhou, and Z. You, *Optimal dimensionality reduction of sensor data in multisensor estimation fusion,* IEEE Transactions on Signal Processing, vol. 53, pp. 1631–1639, May 2005.