

# Linköping University Post Print

## Statistical results for system identification based on quantized observations

Fredrik Gustafsson and Rickard Karlsson

N.B.: When citing this work, cite the original article.

Original Publication:

Fredrik Gustafsson and Rickard Karlsson, Statistical results for system identification based on quantized observations, 2009, Automatica, (45), 12, 2794-2801.

<http://dx.doi.org/10.1016/j.automatica.2009.09.014>

Copyright: Elsevier Science B.V., Amsterdam.

<http://www.elsevier.com/>

Postprint available at: Linköping University Electronic Press

<http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-52878>

# Statistical Results for System Identification based on Quantized Observations

Fredrik Gustafsson<sup>1</sup> Rickard Karlsson<sup>2</sup>

---

## Abstract

System identification based on quantized observations requires either approximations of the quantization noise, leading to suboptimal algorithms, or dedicated algorithms tailored to the quantization noise properties. This contribution studies fundamental issues in estimation that relate directly to the core methods in system identification. As a first contribution, results from statistical quantization theory are surveyed and applied to both moment calculations (mean, variance etc) and the likelihood function of the measured signal. In particular, the role of adding dithering noise at the sensor is studied. The overall message is that tailored dithering noise can considerably simplify the derivation of optimal estimators. The price for this is a decreased signal to noise ratio, and a second contribution is a detailed study of these effects in terms of the Cramér-Rao lower bound. The common *additive uniform noise* approximation of quantization is discussed, compared, and interpreted in light of the suggested approaches.

---

## 1 Introduction

System identification provides a unified theory for how to estimate parameters in a system based on observed input-output data Ljung (1999). This theory applies to all kind of model structures as long as the noise can be described as filtered white (zero mean and independent over time) noise, which is not the case of quantization noise which is signal dependent. That is, there is a need to modify the classical system identification algorithms and theory in the case the measurements are quantized.

As a motivating example, quantization of sensor data is a problem of increasing importance in applications on networked systems. In a sensor network, each node consists of a sensor, a simple processor and a communication means. Such nodes have limited bandwidth leading to quantization effects. Further, they are in many cases energy constrained. It is an interesting fact that computations are several orders of magnitude less energy demanding than communication in state of the art technology. For instance, the sensor node processor proposed in Necchi *et al.* (2006) consumes 2.7 pJ per instruction, which can be compared to short range low rate com-

munication using the Zigbee IEEE 802.15.4 standard, which consumes about 15  $\mu$ J per transmitted bit of information. That is, about 5 million instructions can be performed at the sensor node at the same cost as transmitting one bit of information. The conclusion is that quite advanced pre-processing at the sensor node is possible to mitigate the effects of the quantization implied by the energy induced bandwidth constraint.

The literature on classical quantization is summarized in Oppenheim and Schaffer (1975). Here, it is described as a problem for finite computation precision in microprocessors, and this is basically a deterministic analysis of round-off errors. The foundation for statistical quantization theory was laid by Widrow and co-workers, see Widrow *et al.* (1996); Widrow and Kollar (2008) for some more recent surveys. They interpreted quantization as area sampling. This leads to a continuous *probability density function* (PDF), interpolating the discrete PDF for the quantized observations between the quantization levels. This enables the use of powerful analysis tools as the *characteristic function* (CF), which is a Fourier transformed PDF. From this, the theoretical moments can be computed and compared to the sample moments. Also, the likelihood function can be expressed as a function of the likelihood for unquantized observations. One further result in Widrow *et al.* (1996); Karlsson (2005b) is the interpretation of quantization as one term corresponding to *additive uniform noise* (AUN), which is the standard *ad-hoc* approximation, and one alias term. This interpretation has many parallels to the sampling theorem.

Dithering is a multi-purpose method to mitigate the ef-

---

\* This work was supported by the the Swedish Research Council. A short version of this manuscript has been submitted to the IFAC Symposium on System Identification, SYSID, 2009.

<sup>1</sup> Dept. of Elec. Eng., Linköping University, Sweden. E-mail: fredrik@isy.liu.se

<sup>2</sup> NIRA Dynamics AB, Sweden. E-mail: rickard.karlsson@niradynamics.se

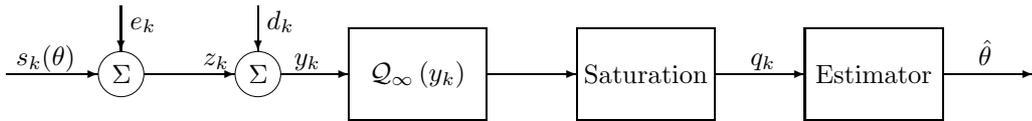


Fig 1. The sensor measures  $z_k$  consisting of signal  $s_k(\theta)$  and noise  $e_k$ . The sensor can then generate dithering noise  $d_k$  added to each measurement before the signal is quantized and saturated and communicated in the network. The receiver gets  $q_k$  from which inference about the parameter  $\theta$  is made.

fects of quantization. The idea is to add artificial noise to the observed signal before quantization. In a sensor network application, this is done at the sensor node before transmission. Web searches reflect a large number of patents utilizing dithering in hardware such as analog to digital converters. One of the first theoretical studies of dithering is Schuchman (1964), where a class of dithering noises is proposed that avoids the alias terms in all moment formulas. That is, with such dithering, the AUN assumption holds for the moments. The statistical relations of the quantization error are surveyed in Wannamaker *et al.* (2000).

However, the quantization theory cited above does not address the case where the signal to be quantized contains a deterministic signal component, which may be parameterized in a vector of unknown parameters. There are a few publications in the area of system identification based on quantized sensor data. For instance, Wang *et al.* (2003) studies the problem of binary sensors and Wang and Yin (2007) suggests an algorithm for estimating the gain in dynamic systems with a thorough analysis, and some extensions to other problems than gains are indicated. Still, general results appear to be lacking. In general, the approximation of quantization as *additive uniform noise* (AUN) can be used to derive suboptimal algorithms.

The purpose with this contribution is to apply the classical quantization theory to some basic system identification algorithms and analysis methods. Let  $\theta$  denote the unknown parameters and  $q_k$  the available observations from the sensor. A block diagram for the problem is depicted in Fig 1. The following problems will be studied:

- *Prediction error methods* (PEM), *nonlinear least squares* (NLS) and *weighted least squares* (WLS) methods require the mean and variance of  $q_k$  to be known as a function of  $\theta$ . Methods based on higher order statistics Nandi (1999) rely on knowledge of more theoretical moments of the signal.
- The *maximum likelihood* (ML) estimate requires the likelihood function  $p(\theta|q_{1:N})$  to be known, where  $q_{1:N}$  denotes the set  $\{q_k\}_{k=1}^N$ .
- The *Cramér-Rao lower bound* (CRLB) gives a lower bound on the estimation covariance for any unbiased estimator, including the ones from NLS, WLS and ML methods above.

It is the purpose of this contribution to provide specific expressions for the moments of  $q_k$ , the likelihood function for  $\theta$  and the CRLB.

The outline is as follows. Section 2 surveys quantization noise and some of its fundamental properties. Section 3 provides some results on the use of dithering noise. Section 4 first summarizes some known properties of moment calculations using quantized signals, then novel explicit expressions are presented for the first moments with and without dithering, respectively. Section 5 gives the likelihood function for the system parameters given quantized observations, and Section 6 focuses on the CRLB.

## 2 Quantization Model

Quantization is here defined according to the *midriser* convention in Lipshitz *et al.* (1992) as

$$q = \mathcal{Q}_m(y) = \begin{cases} -m\Delta + \frac{\Delta}{2}, & y < -m\Delta, \\ \Delta \lfloor \frac{y}{\Delta} \rfloor + \frac{\Delta}{2}, & -m\Delta \leq y < m\Delta, \\ m\Delta - \frac{\Delta}{2}, & y \geq m\Delta. \end{cases} \quad (1)$$

Here,  $\lfloor x \rfloor$  denotes the integer smaller than or equal to  $x$ . This definition covers sign quantization  $\mathcal{Q}_1(y) = \text{sign}(y)$ , with the implicit assumption that  $\Delta = 2$  in this case. The quantized value can be represented with an integer  $i = -m, -m+1, \dots, m-1$ . The dithering noise  $d_k$  is added in the sensor to the measurement  $z_k$  before the quantization, and the *non-subtractive dithering* (NSD) convention is used (the dithering noise is not subtracted from  $q_k$ ). The quantizer can be split into one unsaturated quantizer followed by a saturation block as illustrated in Fig 1, and the analysis can be split accordingly. For convenience,  $\mathcal{Q}_\infty(y) = \Delta \lfloor \frac{y}{\Delta} \rfloor + \frac{\Delta}{2}$  is defined as the un-saturated quantization function, in which case the discrete likelihood function is given by

$$\text{Prob}(q = i\Delta + \frac{\Delta}{2}) = \int_{i\Delta}^{(i+1)\Delta} p_y(y) dy. \quad (2)$$

This integral can equivalently be defined as  $p_y \star p_u(i\Delta + \Delta/2)$ , where  $\star$  denotes convolution. Here,  $p_u$  is the PDF

of the standard uniform noise defined as

$$p_u(u) = \begin{cases} \frac{1}{\Delta}, & -\frac{\Delta}{2} \leq u \leq \frac{\Delta}{2}, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

As shown in for instance Sripad and Snyder (1977); Widrow *et al.* (1996); Wannamaker *et al.* (2000); Widrow and Kollar (2008), a continuous likelihood that provides the same moment formulas as (2) can be defined as

$$p_q(q) = \sum_{i=-\infty}^{\infty} \delta\left(q - i\Delta + \frac{\Delta}{2}\right) p_y \star p_u(q). \quad (4)$$

This is referred to as *area sampling* in Widrow *et al.* (1996). The quantized measurement  $q$  contains the following uncertainties, as can be interpreted from (4):

- (1) The uncertainty in the unquantized signal  $y$ .
- (2) The first order effect (the term with  $i = 0$ ) of the quantization level  $\Delta$  which will be referred to as the *additive uniform noise* (AUN) effect, with the corresponding variance  $\Delta^2/12$ .
- (3) Remaining higher order effects from the terms with  $i \neq 0$ , which will be interpreted as PDF *aliasing* because of the similarity of (4) with Poisson's summation formula in sampling theory.
- (4) Saturation effects when  $|y| > m\Delta$ , not revealed in (4).

Let  $\Phi_y(\omega)$  denote the *characteristic function* (CF) for  $p_y(y)$ , see Gut (1995); Stuart and Ord (1994). It is defined as the Fourier transform (with sign-reversed frequency)

$$\Phi_y(\omega) = \mathcal{F}\{p_y(y)\} = \mathbb{E}(e^{j\omega y}) = \int_{-\infty}^{\infty} e^{j\omega y} p_y(y) dy. \quad (5)$$

For instance, the CF of a uniform distribution (3) is

$$\Phi_u(\omega) = \text{sinc}(\Delta\omega/2) = \frac{\sin(\Delta\omega/2)}{\Delta\omega/2}. \quad (6)$$

Standard transform properties applied to (4) imply that the CF for  $q$  is

$$\Phi_q(\omega) = \sum_{i=-\infty}^{\infty} \Phi_y\left(\omega + i\frac{2\pi}{\Delta}\right) \text{sinc}\left(\frac{\Delta(\omega + i\frac{2\pi}{\Delta})}{2}\right). \quad (7)$$

Now, if the signal satisfies the band-limited condition in Widrow *et al.* (1996); Karlsson (2005b), defined by

$$\Phi_y(\omega) = 0, \quad |\omega| \geq \pi/\Delta, \quad (8)$$

the terms corresponding to  $i \neq 0$  in (7) will not contribute to the sum, and the convolution–multiplication duality in transform theory applied to (7) implies that the CF of the unquantized distribution can be recovered exactly as

$$\Phi_y(\omega) = \begin{cases} \frac{\Phi_q(\omega)}{\text{sinc}(\Delta\omega/2)}, & |\omega| < \frac{\pi}{\Delta}, \\ 0, & |\omega| \geq \frac{\pi}{\Delta}. \end{cases} \quad (9)$$

That is,  $p_y(y)$  can be recovered by an inverse Fourier transformation. This result is commonly referred to as *Quantization Theorem I* (QT I). This, and many related results are surveyed in the excellent monograph Widrow and Kollar (2008). Note that the classical statistical quantization theory does not include a deterministic signal component, and does not directly apply to the case in Fig 1.

### 3 Dithering

Consider now the case in Fig 1, where  $y$  consists of a signal component, measurement noise and an optional dithering noise.

The signal  $s_k(\theta)$  is assumed uncorrelated over time for two reasons. The first one is to avoid the associated problems with correlation that quantization gives rise to. The second one is that predictive coding is a very good idea before transmitting quantized data over a network, see Agüero *et al.* (2007).

If  $s(\theta) + e$  is not band-limited, a band-limited dithering noise can be added to the signal before quantization, and it follows from basic transform properties and the definition (8) that the total signal  $y = s(\theta) + e + d$  will be band-limited. One such distribution is proposed in Karlsson (2005b) and Appendix A in Widrow and Kollar (2008), and it is summarized below:

- The PDF is defined for any integer  $k > 0$  as

$$p_d(d) = c \text{sinc}^{2k}(\pi d/(2k\Delta)) = c_k \frac{\sin^{2k}\left(\frac{\pi d}{2k\Delta}\right)}{\left(\frac{\pi d}{2k\Delta}\right)^{2k}}, \quad (10)$$

$$c_k = \int \text{sinc}^{2k}(\pi d/(2k\Delta)) dd, \quad (11)$$

where  $c_k$  is the normalization constant.

- All odd moments are zero, and  $\mathbb{E}(d^m) = \infty$  for all  $m \geq 2k$ .
- For  $k = 3$ , the following properties hold:

$$c_k = 0.3030, \quad J^{1/2}(p_d) = 1.28\Delta, \quad (12)$$

$$\mathbb{E}(d^2) = 1.66\Delta^2, \quad \mathbb{E}(d^4) = 8.27\Delta^4. \quad (13)$$

Here,  $J^{1/2}(p_d)$  denotes the square root of Fisher's information matrix for the dithering noise with PDF  $p_d$ .

It is shown in Widrow *et al.* (1996); Karlsson (2005b) that there is a duality between sampling and quantization. Quantizing signals that are not band-limited according to (8) give rise to aliasing similar to Poisson's summation formula in sampling theory. Adding band-limited noise takes the role of anti-alias filtering, after which perfect reconstruction of the signal's PDF or continuous waveform, respectively, is possible.

## 4 Moment-Based Estimation

### 4.1 Moment Formulas

One useful property of the CF is that all higher order moments can be calculated from it. This follows from the Taylor expansion  $\Phi_y(\omega) = \mathbb{E}(e^{j\omega y}) = 1 + j\omega\mathbb{E}(y) - \frac{1}{2!}\omega^2\mathbb{E}(y^2) + \dots$ . Hence,

$$\mathbb{E}(y^r) = \frac{1}{j^r} \frac{d^r}{d\omega^r} \Phi_y(\omega) \Big|_{\omega=0}. \quad (14)$$

Since the CF needs to be correct only close to the origin to compute all moments, aliasing is here tolerated as long as there is no folding in (7) at  $\omega = 0$ . That is, a sufficient condition for moment reconstruction is that  $\Phi_y(\omega) = 0$  for  $|\omega| > 2\pi/\Delta$ . This is known as Quantization Theorem II (QT II), see Widrow *et al.* (1996); Widrow and Kollar (2008). The excellent survey Wannamaker *et al.* (2000) provides the following properties:

- In NSD quantization, the quantization error is defined as  $\varepsilon = z - q$  and this error is neither statistically independent of the input or uniformly distributed for arbitrary input distributions  $p_z(z)$ , no matter how the dithering PDF  $p_d(d)$  is chosen.
- In *subtractive dithering* (SD) quantization, on the other hand, the quantization error is defined as  $\varepsilon = y - q = z - (q - d)$ , and it is both statistically independent of the input and uniformly distributed for arbitrary input distributions  $p_z(z)$ , if the dither noise satisfies the Schuchman's condition in Schuchman (1964)

$$\Phi_d(2\pi k/\Delta) = 0, \quad \forall k \neq 0. \quad (15)$$

Note that the Schuchman condition is satisfied for the definition of band-limited noise (8). Put in other words, the class of band-limited PDF's is a subset of the class of PDF's that satisfy the Schuchman condition.

- For NSD quantization, the moments of the quantization error  $\varepsilon$  can be recovered as

$$\mathbb{E}(\varepsilon^r) = \left(\frac{j}{2\pi}\right)^r \frac{d^r}{d\omega^r} \Phi_y(\omega) \Big|_{\omega=0} \quad (16a)$$

if

$$\frac{d^r}{d\omega^r} \Phi_y(\omega) \Big|_{\omega=2\pi k/\Delta} = 0, \quad k \neq 0. \quad (16b)$$

Further,  $\varepsilon$  is independent of the signal  $z$ .

- A dithering PDF that satisfies this condition up to moment  $r$  is

$$\Phi_d(\omega) = \text{sinc}^{r+1}(\omega). \quad (17)$$

Compare this definition to the one in (10). Generating dither noise can be implemented by summing  $r + 1$  uniformly distributed samples.

We will in the following sections apply this theory to compute the moments of  $q_k$  and provide some simple and illustrative examples.

### 4.2 Gaussian Noise without Dithering

The following theorem is a special case of the general result in Section 5.4 in Widrow *et al.* (1996), here applied to Gaussian measurement noise to illustrate the complicated expressions in moment-based estimation implied by quantization.

**Theorem 1 (Moments of  $Q_\infty(e)$ )** Consider the case  $q = Q_\infty(e)$ , where  $e$  is zero mean Gaussian noise with variance  $\sigma^2$ . The variance can for small  $\sigma/\Delta$  be approximated by

$$\mathbb{E}(q^2) \approx \underbrace{\sigma^2}_{\text{GN}} + \underbrace{\frac{\Delta^2}{12}}_{\text{AUN}} - \underbrace{\left(4\sigma^2 + \frac{\Delta^2}{\pi^2}\right)}_{\text{alias}} \cdot e^{-2\pi^2 \frac{\sigma^2}{\Delta^2}}. \quad (18a)$$

The fourth order moment is given by

$$\begin{aligned} \mathbb{E}(q^4) \approx & \underbrace{3\sigma^4}_{\text{GN}} + \underbrace{\frac{\Delta^4}{80} + \frac{\sigma^2\Delta^2}{2}}_{\text{AUN}} \\ & + \underbrace{\left(-\frac{\Delta^4}{2\pi^2} - 2\sigma^2\Delta^2 + \frac{6\Delta^2\sigma^2}{\pi^2} + \frac{3\Delta^4}{\pi^4} + \frac{32\sigma^6\pi^2}{\Delta^2}\right)}_{\text{alias}} \cdot e^{-2\pi^2 \frac{\sigma^2}{\Delta^2}}. \end{aligned} \quad (18b)$$

**PROOF.** The CF of  $y = e$ , where  $e$  here denotes the Gaussian noise, is given by

$$\begin{aligned} \Phi_y(\omega) &= \mathbb{E}(e^{j\omega y}) = \int_{-\infty}^{\infty} e^{j\omega v} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}v^2} dv \\ &= e^{-\frac{(\omega\sigma)^2}{2}}. \end{aligned} \quad (19)$$

Applying the moment formula (14) to the terms corresponding to  $i = -2, \dots, 2$  in (7) using (21) and the chain

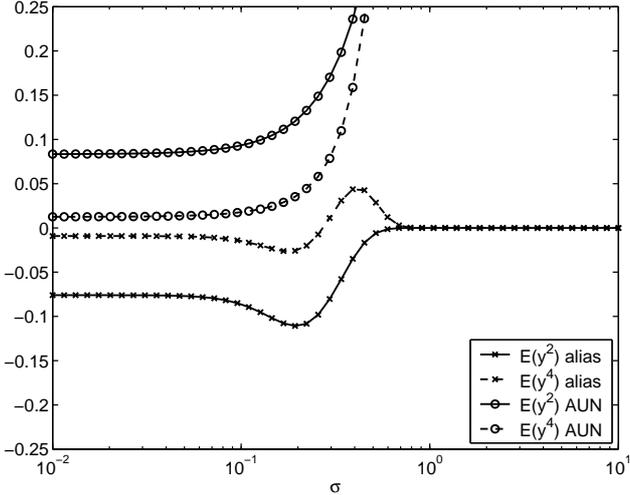


Fig 2. The AUN and alias part of (18), for quantized Gaussian noise using different  $\sigma$ , with quantization level  $\Delta = 1$ .

rule now gives the result. The terms corresponding to  $|i| > 2$  give rise to the higher order terms.

The first line in each equation in (18) describes the Gaussian noise and AUN effect ( $i = 0$ ), while the second line corresponds to terms due to aliasing. Fig 2 illustrates the dependence of Gaussian noise standard deviation  $\sigma$  for the case of  $\Delta = 1$ . As can be seen from (18a), the alias term is negligible when  $\Delta \ll \sigma$ , and the critical region occurs around  $\Delta \approx 3\sigma$ . Another interesting fact is that the AUN and alias contributions almost cancel out when  $\Delta \gg 3\sigma$  (note that the dominating term in (18a) in this case is  $\Delta^2(1/12 - 1/\pi^2 + 1/(4\pi^2)) \approx 0$ ).

A more interesting case appears when there is a deterministic signal component. This case is less treated in literature, and the next theorem provides an exact and novel expression for the second order moment.

### Theorem 2 (Moments of $\mathcal{Q}_\infty(s(\theta) + e)$ )

The second moment of  $q = \mathcal{Q}_\infty(s(\theta) + e)$ , where  $e$  is Gaussian distributed, is given by

$$\mathbb{E}(q^2) = s^2(\theta) + \sigma^2 + \frac{\Delta^2}{12} + \sum_{k=1}^{\infty} (-1)^k e^{-\frac{2k^2\pi^2\sigma^2}{\Delta^2}} \quad (20)$$

$$\left[ \left( 4\sigma^2 + \frac{\Delta^2}{k^2\pi^2} \right) \cos\left( 2k\pi \frac{s(\theta)}{\Delta} \right) + \frac{2\Delta s(\theta)}{k\pi} \sin\left( 2k\pi \frac{s(\theta)}{\Delta} \right) \right]$$

**PROOF.** Similar to the proof of Theorem 1, the result can be derived by applying the moment formula (14) to

$$\Phi_y(\omega) = \mathbb{E}(e^{j\omega y}) = \int_{-\infty}^{\infty} e^{j\omega(s+v)} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}v^2} dv$$

$$= e^{j\omega s} e^{-\frac{(\omega\sigma)^2}{2}}. \quad (21)$$

using (21) and the chain rule.

Note that (18a) corresponds to the first terms in (20) for the special case  $s(\theta) = 0$ . This special case of (20) appears in Sripad and Snyder (1977).

### 4.3 Gaussian Noise with Dithering

Adding dithering noise, that either satisfies the Schuchman condition or the stronger band-limited condition, leads to substantially simplified equations, as the following theorem shows.

### Theorem 3 (Moments of $\mathcal{Q}_\infty(s(\theta) + e + d)$ )

The first non-zero moments of  $q = \mathcal{Q}_\infty(s(\theta) + e + d)$ , where  $e$  is Gaussian distributed and  $d$  satisfies the Schuchman condition (15), are given by

$$\mathbb{E}(q^2) = s^2(\theta) + \underbrace{\sigma^2}_{\text{GN}} + \underbrace{\mathbb{E}(d^2)}_{\text{DN}} + \underbrace{\frac{\Delta^2}{12}}_{\text{AUN}}, \quad (22a)$$

$$\mathbb{E}(q^4) = s^4(\theta) + 6s^2(\theta)\mathbb{E}(q^2) + \underbrace{3\sigma^4}_{\text{GN}} + \underbrace{\mathbb{E}(d^4)}_{\text{DN}} + \underbrace{\sigma^2\mathbb{E}(d^2)}_{\text{DN} \times \text{GN}} + \underbrace{\frac{\Delta^4}{80}}_{\text{AUN}} + \underbrace{\frac{\sigma^2\Delta^2}{2}}_{\text{AUN} \times \text{GN}}. \quad (22b)$$

**PROOF.** The bandlimited noise implies that the AUN assumption holds. This means that  $q_k = s(\theta) + e_k + d_k$  is an equivalent signal model regarding the moments, and the result follows from standard properties of the Gaussian and uniform distributions.

The theoretical implication of adding band-limited dithering noise is that the terms denoted alias in (18) disappear, while the dithering noise contribution appears instead. These latter terms are known functions of the dithering noise and independent of  $s(\theta)$ .

### 4.4 Moment Based Estimation

A *moment-based estimator*, see Kay (1993), for  $\theta$  is defined as a nonlinear equation system obtained by equating the sample moments with their theoretical values, resulting in the relations

$$\frac{1}{N} \sum_{k=1}^N q_k \approx \mathbb{E}(y) = f_1(\theta), \quad (23a)$$

$$\frac{1}{N} \sum_{k=1}^N q_k^2 \approx \mathbb{E}(y^2) = f_2(\theta), \quad (23b)$$

$$\frac{1}{N} \sum_{k=1}^N q_k^3 \approx \mathbb{E}(y^3) = f_3(\theta), \quad (23c)$$

and so on. Solving this nonlinear system of equations yield the moment estimate of  $\theta$ .

Consider first the case without dithering. A moment-based estimator of the Gaussian noise variance can be derived using either (18a) or (18b). Just solve one of these equation for  $\sigma^2$ , where the left-hand side is replaced with the corresponding observed sample moment. Since these expressions take both the AUN and alias effect of quantization into account, the estimate will be unbiased (neglecting the higher order terms). However, the estimator is a nonlinear implicit function of the moment and an iterative search algorithm is needed.

Consider next the case with dithering noise satisfying the Schuchman condition. For instance, assume the bandlimited noise in (10) is used for  $k = 3$ . A moment-based estimator of noise variance is given by

$$\widehat{\sigma^2} = \frac{1}{N} \sum_{i=1}^N q_i^2 - 1.66\Delta^2 - \frac{\Delta^2}{12}. \quad (24)$$

That is, a closed form and simple expression is obtained.

The conclusion is that the moment formula (14), Poisson's summation formula, and the CF of the dithering noise provide all information needed to derive an unbiased moment based estimator. However, the use of dithering noise substantially simplifies the numerical computation of the estimate.

## 5 Maximum Likelihood Estimation

### 5.1 Likelihood Formulas

The *maximum likelihood estimate* (MLE) for the parameters  $\theta$  given a set of quantized observations  $q_{1:N}$  is given by

$$\hat{\theta} = \arg \max_{\theta} \prod_{k=1}^N p_{q|\theta}(q_k) \quad (25a)$$

where  $\arg \max_{\theta}$  means the maximizing argument. The discrete likelihood function corresponding to (2) is

$$\text{Prob}(q = i\Delta + \frac{\Delta}{2} | \theta) = \int_{i\Delta}^{(i+1)\Delta} p_{s(\theta)+e|\theta} \star p_d(y) dy. \quad (25b)$$

The continuous likelihood function that corresponds to the above at the quantization levels is similarly to (4) given by

$$p_{q|\theta} = \sum_{i=-\infty}^{\infty} \delta\left(q - i\Delta + \frac{\Delta}{2}\right) p_{s(\theta)+e|\theta} \star p_d \star p_u(q). \quad (25c)$$

The key point is that in case of bandlimited dithering noise, the alias terms in the sum disappear and the ML estimator simplifies to

$$\hat{\theta} = \arg \max_{\theta} \prod_{k=1}^N p_{s(\theta)+e|\theta} \star p_d \star p_u(q). \quad (26)$$

That is, numerical algorithms for the standard system identification case can be reused, where the unquantized likelihood  $p_{s(\theta)+e|\theta}$  is replaced with a smoothed version.

In the next section, the special case of an unknown scalar mean,  $s(\theta) = \theta$ , is studied in detail. This case gives particular simple numerical algorithms, and the case will be used to illustrate the estimation bounds in the following sections.

### 5.2 MLE for the Mean in the Gaussian Case

It is rather straightforward to derive the ML estimators for the mean  $s(\theta) = \theta$  given quantized observations  $q_k = \mathcal{Q}_m(\theta + e_k)$  in the Gaussian case. First, for binary observations, form the log-likelihood as

$$\begin{aligned} \log p(q_{1:N} | \theta) &= \log \prod_{i=1}^N p(q_i | \theta) = \sum_{i=1}^N \log p(q_i | \theta) = \\ &= N_- \log \varrho(-\theta/\sigma) + N_+ \log(1 - \varrho(-\theta/\sigma)), \end{aligned} \quad (27)$$

where  $N_-$  and  $N_+$  denote the number of terms with  $q_i = -1$  and  $q_i = +1$  respectively, so that  $N_- + N_+ = N$ . Here,  $\varrho(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$  denotes the standard Gaussian distribution function. Maximizing the expression by differentiation yields

$$\frac{N_+}{N_-} = \frac{1 - \varrho(-\hat{\theta}^{\text{ML}}/\sigma)}{\varrho(-\hat{\theta}^{\text{ML}}/\sigma)}. \quad (28)$$

Hence

$$\varrho(-\hat{\theta}^{\text{ML}}/\sigma) = \frac{N_-}{N_- + N_+} = \frac{N_-}{N}. \quad (29)$$

Since the left hand side is a monotone and increasing function, the estimate,  $\hat{\theta}^{\text{ML}}$ , can be found with a simple line search. For more information on sign quantizers, see for instance Host-Madsen and Händel (2000), where the ML and CRLB are calculated for estimating the frequency of a sinusoidal in noise.

Second, for multi-level quantization, the log-likelihood is

$$\log p(q_{1:N}|\theta) = \sum_{i=1}^N \log p(q_i|\theta) = \sum_{j=-m}^m N_j \log p_j(\theta), \quad (30)$$

where  $N_j$  is the number of occurrences of each  $q_i = -j\Delta + \Delta/2$ , so that  $\sum_j N_j = N$ . The ML estimate is here found numerically by searching for maximum of (30). Here  $p_j(\theta)$  is given by (A.6) for the case  $y_k = \mathcal{Q}_m(\theta + e_k)$ . A numerical illustration will be given in Section 6.2.

## 6 Cramér-Rao Lower Bounds

In the sequel, the analysis involves gradients of scalar functions or vector valued functions. The gradient is defined as:

$$\nabla_{\theta} g^T(\theta) = \begin{pmatrix} \frac{\partial g_1}{\partial \theta_1} & \cdots & \frac{\partial g_m}{\partial \theta_1} \\ \vdots & & \vdots \\ \frac{\partial g_1}{\partial \theta_n} & \cdots & \frac{\partial g_m}{\partial \theta_n} \end{pmatrix}, \quad g: \mathbb{R}^n \mapsto \mathbb{R}^m. \quad (31a)$$

Further, the Laplacian for the scalar function  $g(\theta, \nu)$  with  $\theta \in \mathbb{R}^n, \nu \in \mathbb{R}^m$  is defined as

$$\Delta_{\nu}^{\theta} g(\theta, \nu) = \nabla_{\nu} (\nabla_{\theta} g(\theta, \nu))^T, \quad g: \mathbb{R}^n \times \mathbb{R}^m \mapsto \mathbb{R}. \quad (31b)$$

For an unbiased estimator, the expected mean equals the true parameter,  $\mathbb{E}(\hat{\theta}) = \theta$ . The CRLB, Cramér (1946); Kay (1993); Lehmann (1983), is then given by

$$\text{Cov}(\theta - \hat{\theta}) = \mathbb{E}((\theta - \hat{\theta})(\theta - \hat{\theta})^T) \succeq J^{-1}(\theta), \quad (32a)$$

$$J(\theta) = \mathbb{E}(-\Delta_{\theta}^{\theta} \log p(y|\theta)), \quad (32b)$$

where  $J(\theta)$  denotes the *Fisher information matrix* (FIM) in the measurement  $y$  regarding the stochastic variable  $\theta$ . Also note that an equivalent representation of the information, Kay (1993), is

$$J(\theta) = \mathbb{E}(\nabla_{\theta} \log p(y|\theta) (\nabla_{\theta} \log p(y|\theta))^T). \quad (33)$$

For the case with independent measurements  $y_i, i = 1, \dots, M$ , the information is given as

$$J(\theta) = \sum_{i=1}^M J^{(i)}(\theta), \quad (34)$$

due to the additivity of information, assuming that  $J^{(i)}$  is the information for measurement  $i$ .

### 6.1 CRLB for Mean Estimation

Consider now the problem of estimating  $\theta$  from the quantized measurements  $q = \mathcal{Q}_m(\theta + e)$ . Explicit expressions for the information for Gaussian noise are derived in the sequel, and it is demonstrated that the AUN assumption,

$$J_{\text{approx}}^{-1}(\theta) = \sigma^2 + \frac{\Delta^2}{12}, \quad (35)$$

in many cases is quite misleading.

Using band-limited dithering noise, the property (34) gives the CRLB for  $y = \mathcal{Q}_m(\theta + e + d)$  as (without approximation)

$$J_{\text{dithering}}^{-1}(\theta) = \sigma^2 + \frac{\Delta^2}{12} + J(p_d), \quad (36)$$

where  $J(p_d)$  is the FIM for the mean of the dithering distribution  $p_d$ . This value can be tabulated for each distribution, for instance as done in (12).

The information without dithering depends on  $\theta$  and includes saturation effects. From now on, saturation effects in the quantization will be taken into account. First, the sign quantizer is given for its simplicity, and then the general multi-level case is treated. Here, the focus is now on exact calculations using the FIM in (32b) and the equivalent representation via (33). Other recent work on quantization, focusing on the exact CRLB expression can be found in for instance Landes (2005); Karlsson (2005b).

First, the Fisher information for the sign quantizer is derived.

**Theorem 4 (Fisher Information for sign quantizer)**  
Consider the sign quantizer

$$q = \mathcal{Q}_1(\theta + e) = \text{sign}(\theta + e), \quad e \in \mathcal{N}(0, \sigma^2). \quad (37)$$

The Fisher information is

$$J_1(\theta) = \frac{e^{-\frac{\theta^2}{\sigma^2}}}{2\pi\sigma^2} \frac{1}{(1 - \varrho(-\theta/\sigma))\varrho(-\theta/\sigma)}, \quad (38)$$

where  $\varrho(\theta) \triangleq \text{Prob}(\Theta < \theta)$  denotes the Gaussian distribution function.

**PROOF.** See Appendix A.

The sign quantizer can be generalized to the multi-level quantization case.

**Theorem 5 (Multi-level quantization)** Consider the multi-level quantizer.

$$q = \mathcal{Q}_m(\theta + e), \quad e \in \mathcal{N}(0, \sigma^2). \quad (39)$$

The Fisher information is

$$\begin{aligned} J_m(\theta) = & \frac{\left(-\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{-m\Delta-\theta}{\sigma}\right)^2}\right)^2}{\rho\left(\frac{-m\Delta-\theta}{\sigma}\right)} \\ & + \sum_{j=-m+1}^{m-1} \frac{\left(-\frac{1}{\sqrt{2\pi}\sigma} \left(e^{-\frac{1}{2}\left(\frac{(j+1)\Delta-\theta}{\sigma}\right)^2} - e^{-\frac{1}{2}\left(\frac{j\Delta}{\sigma}\right)^2}\right)\right)^2}{\rho\left(\frac{(j+1)\Delta-\theta}{\sigma}\right) - \rho\left(\frac{j\Delta-\theta}{\sigma}\right)} \\ & + \frac{\left(\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{m\Delta-\theta}{\sigma}\right)^2}\right)^2}{1 - \rho\left(\frac{m\Delta-\theta}{\sigma}\right)} \end{aligned} \quad (40)$$

where  $\rho(\theta) \triangleq \text{Prob}(\Theta < \theta)$  denotes the Gaussian distribution function.

**PROOF.** See Appendix A.

### 6.2 Illustrative Example

Consider the multi-level quantizer  $q = \mathcal{Q}_m(\theta + e)$ , with  $m = 3, \Delta = 0.5$ , using the midriser convention. The noise is assumed independent and  $e \in \mathcal{N}(0, 0.14^2)$ . Here,  $\Delta$  is chosen such that  $\frac{\Delta^2}{12} \approx \text{Var}(e)$ .

In Fig 3, the CRLB and the standard deviation for the ML-estimate using 1000 Monte Carlo simulations are presented, as a function of the true value  $\theta$ , using  $N = 100$  measurements. The AUN approximation is also plotted, assuming additive noise  $q = \mathcal{Q}_m(\theta + e) \approx \theta + e + u$ ,  $\text{Var}(u) = \frac{\Delta^2}{12}$ . That is, the horizontal line shows  $\sigma_{\text{approx}}^2 = \frac{1}{N}(\text{Var}(e) + \text{Var}(u))$ . The AUN approximation is good in average, but of course cannot handle the saturation effects.

Next, it is examined how the information and thus the CRLB depends on the quantization level. In Fig 4, the Fisher information  $J_m(\theta)$  is illustrated by plotting the lower bound  $J_m^{-1/2}(\theta)$  on the standard deviation for different quantization levels  $\Delta = 2/m$ . Here,  $q = \mathcal{Q}_m(\theta + e)$ ,  $e \in \mathcal{N}(0, \sigma^2)$  is used with  $\sigma = 0.1$ . Note that  $J_{100}^{-1/2}(\theta) \approx \sigma$  and that  $J_m$  converges to the AUN in (35) when  $m \rightarrow \infty$ , that is, when  $\Delta \rightarrow 0$ .

### 6.3 CRLB in the General Case

Consider now the original signal model where  $q_k = \mathcal{Q}_m(s_k(\theta) + e_k)$ , and  $\theta$  is vector valued. A linear regression model  $s_k(\theta) = H_k\theta$  is included as a special case.

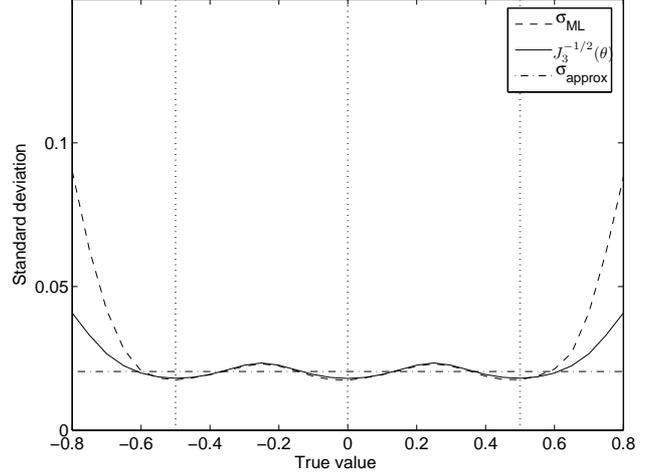


Fig 3. The CRLB  $J_3^{-1/2}(\theta)$  and ML standard deviation as a function of the true value compared to the additive noise approximation,  $\sigma_{\text{approx}}$ , when  $q = \mathcal{Q}_m(\theta + e)$  with  $m = 3$  levels are used.

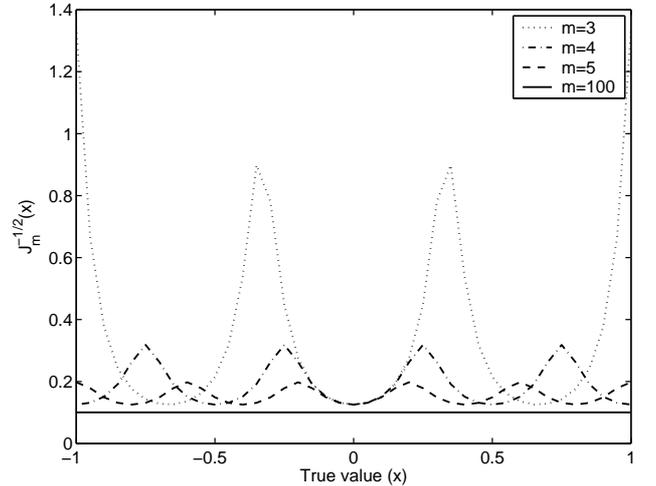


Fig 4. Fisher information for  $q = \mathcal{Q}_m(\theta + e)$ , used to compute the standard deviation lower bound  $J_m^{-1/2}(\theta)$  as a function of  $\theta$  for different quantization levels  $\Delta = 2/m$ .

The following theorem generalizes the results obtained above.

**Theorem 6** Consider the model  $q_k = \mathcal{Q}_m(s_k(\theta) + e_k)$  with  $e_k \in \mathcal{N}(0, \sigma_k^2)$ ,  $i = 1, \dots, N$ . Let  $s_k = H_k\theta$  and  $s_k = h_k(\theta)$  denote the signal part for linear and nonlinear regression, respectively. The Fisher information for quantized measurements is given by

$$J(\theta) = \sum_{k=1}^N H_k^T J_m(s_k(\theta)) H_k. \quad (41)$$

where

$$H_k = \frac{\partial}{\partial \theta} s_k(\theta) \quad (42)$$

and  $J_m(s_k(\theta))$  is given by (38) for binary quantization  $m = 1$  and (40) for multi-level quantization  $m > 1$ .

**PROOF.** Follows from the chain rule, the additivity of information and calculations according to Theorem 4 and Theorem 5, respectively.

Without quantization,  $J_m(s_i(\theta)) = \sigma^2$  independent of the signal value. Using the AUN assumption just increases the figure to  $J_m(s_i(\theta)) = \sigma^2 + \Delta^2/12$ . However, as Fig 4 illustrates, the information is highly dependent on the signal value  $s_k(\theta)$ .

As a final remark, dithering increases the CRLB as shown in (36), indicating that the attainable variance for the parameter estimate increases as well. One should here note that CRLB holds for unbiased estimators, and dithering actually removes the bias. That is, the total mean square error can decrease. This has to be examined on a case to case basis.

## 7 Conclusions

The implication of quantization on moment or likelihood based approaches to estimation and system identification has been studied. For all these approaches, a deep understanding is required of how quantization changes the statistics of data used in the estimator, in particular when the quantization level is large compared to the standard deviation of the measurement noise. The uncertainty in a quantized sensor measurement was split up in the following contributions: The measurement noise, the equivalent AUN noise, possible dithering noise and the remaining uncertainty which was interpreted as alias noise. The design of dithering noise was shown to be a trade-off between the uncertainty it adds in itself and how well it suppresses the alias noise. It was further explained how adding band-limited dithering noise can make moment and likelihood reconstruction easier. An open question is what and how much can be gained by dithering and how an optimal dithering noise should be designed.

Further, a detailed study on the Cramér-Rao lower bound was given. Several theoretical results and examples were presented to show that estimators utilizing knowledge of the quantization are superior to conventional estimators, where only the second order properties of the quantization is incorporated.

## A Proofs

**Proof of Theorem 4.** The probability function for  $q$  can be calculated using

$$p(q = -1|\theta) = \text{Prob}(\theta + e < 0) = \varrho(-\theta/\sigma). \quad (\text{A.1})$$

Similarly,

$$p(q = +1|\theta) = \text{Prob}(\theta + e \geq 0) = 1 - \varrho(-\theta/\sigma).$$

Hence, the discrete likelihood can be written as

$$p(q|\theta) = \varrho(-\theta/\sigma) \delta(q + 1) + (1 - \varrho(-\theta/\sigma)) \delta(q - 1),$$

where  $\delta(i)$  is the pulse function, which is one for  $i = 0$  and zero otherwise. To calculate the CRLB variance, apply (32b).

$$\begin{aligned} J(\theta) &= -\mathbb{E} \left( \frac{\frac{\partial^2 p(q|\theta)}{\partial \theta^2} p(q|\theta) - \left( \frac{\partial p(q|\theta)}{\partial \theta} \right)^2}{p^2(q|\theta)} \right) \\ &= - \sum_{j \in \{-1, 1\}} \frac{\partial^2 p(q = j|\theta)}{\partial \theta^2} - \frac{\left( \frac{\partial p(q = j|\theta)}{\partial \theta} \right)^2}{p(q = j|\theta)}. \end{aligned} \quad (\text{A.2})$$

This general expressions now used for the Gaussian case. From (A.1), we get

$$\frac{\partial p(q|\theta)}{\partial \theta} = \frac{e^{-\frac{\theta^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma} \times \begin{cases} 1, & q = -1 \\ -1, & q = 1 \end{cases} \quad (\text{A.3})$$

$$\frac{\partial^2 p(q|\theta)}{\partial \theta^2} = \frac{-\theta e^{-\frac{\theta^2}{2\sigma^2}}}{\sqrt{2\pi}\sigma^3} \times \begin{cases} 1, & q = -1 \\ -1, & q = 1 \end{cases} \quad (\text{A.4})$$

Inserting these equations into (A.2) gives

$$J(\theta) = \frac{e^{-\frac{\theta^2}{\sigma^2}}}{2\pi\sigma^2} \underbrace{\left( \frac{1}{(1 - \varrho(-\theta/\sigma))} + \frac{1}{\varrho(-\theta/\sigma)} \right)}_{\frac{1}{(1 - \varrho(-\theta/\sigma))\varrho(-\theta/\sigma)}}, \quad (\text{A.5})$$

which proves the theorem.

**Proof of Theorem 5.** Calculate the probability for each level  $j = -m + 1, \dots, m - 1$  as

$$\begin{aligned} p_j(\theta) &= \text{Prob}(j\Delta < \theta + e \leq (j + 1)\Delta) \\ &= \varrho\left(\frac{(j + 1)\Delta - \theta}{\sigma}\right) - \varrho\left(\frac{j\Delta - \theta}{\sigma}\right), \end{aligned} \quad (\text{A.6a})$$

where  $\varrho(\cdot)$  is defined in (A.1). The probability at the

end points are calculated as

$$p_{-m}(\theta) = \varrho \left( \frac{-m\Delta - \theta}{\sigma} \right), \quad (\text{A.6b})$$

$$p_{m-1}(\theta) = 1 - \varrho \left( \frac{m\Delta - \theta}{\sigma} \right). \quad (\text{A.6c})$$

Similar to the sign quantizer, the likelihood is given as

$$p(q|\theta) = \sum_{j=-m}^m p_j(\theta) \delta \left( q - j\Delta - \frac{\Delta}{2} \right). \quad (\text{A.7})$$

Proceeding in the same way as for the sign quantizer, the derivatives for  $j = -m + 1, \dots, m - 2$  are given by

$$\frac{\partial p_j(\theta)}{\partial \theta} = -\frac{1}{\sqrt{2\pi}\sigma} \left( e^{-\frac{1}{2} \left( \frac{(j+1)\Delta - \theta}{\sigma} \right)^2} - e^{-\frac{1}{2} \left( \frac{j\Delta - \theta}{\sigma} \right)^2} \right), \quad (\text{A.8})$$

$$\begin{aligned} \frac{\partial^2 p_j(\theta)}{\partial \theta^2} = & -\frac{(j+1)\Delta - \theta}{\sqrt{2\pi}\sigma^3} e^{-\frac{1}{2} \left( \frac{(j+1)\Delta - \theta}{\sigma} \right)^2} \\ & + \frac{j\Delta - \theta}{\sqrt{2\pi}\sigma^3} e^{-\frac{1}{2} \left( \frac{j\Delta - \theta}{\sigma} \right)^2}. \end{aligned} \quad (\text{A.9})$$

For  $j = -m$ , differentiating (A.6) yields

$$\frac{\partial p_{-m}(\theta)}{\partial \theta} = -\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2} \left( \frac{-m\Delta - \theta}{\sigma} \right)^2}, \quad (\text{A.10a})$$

$$\frac{\partial^2 p_{-m}(\theta)}{\partial \theta^2} = -\frac{-m\Delta - \theta}{\sqrt{2\pi}\sigma^3} e^{-\frac{1}{2} \left( \frac{-m\Delta - \theta}{\sigma} \right)^2}, \quad (\text{A.10b})$$

and similarly for  $j = m - 1$ . Note that the terms in (A.9) form a telescope sum, so together with (A.10b) it yields

$\sum_{j=-m}^{m-1} \frac{\partial^2 p_j}{\partial \theta^2} = 0$ . Hence, the Fisher information is

$$J(\theta) = \sum_{j=-m}^{m-1} \left( -\frac{\partial^2 p_j}{\partial \theta^2} + \frac{\left( \frac{\partial p_j(\theta)}{\partial \theta} \right)^2}{p_j} \right) = \sum_{j=-m}^{m-1} \frac{\left( \frac{\partial p_j(\theta)}{\partial \theta} \right)^2}{p_j}.$$

This, together with (A.6),(A.8) and (A.10a) proves the theorem.

## References

- Agüero, J.C., G.C. Goodwin and J.I. Yuz (2007). System identification using quantized data. In: *IEEE Conference on Decision and Control*. Vol. 46.
- Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton, NJ: Princeton University Press.
- Gut, A. (1995). *An Intermediate Course in Probability*. Springer-Verlag.
- Host-Madsen, A. and P. Händel (2000). Effects of sampling and quantization on single-tone frequency estimation. *IEEE Transactions on Signal Processing* **48**(3), 650–662.

- Karlsson, R. (2005a). Particle Filtering for Positioning and Tracking Applications. PhD thesis. Linköping University, Linköping, Sweden. Linköping Studies in Science and Technology. Dissertations No. 924.
- Karlsson, Rickard (2005b). Particle filtering for positioning and tracking applications. Dissertation no. 924. Linköping University, Sweden.
- Kay, S. (1993). *Fundamentals of Statistical Signal Processing*. Prentice Hall.
- Landes, R.L (2005). Statistical Methods for Application to Calibration Problems. PhD thesis. Iowa State University.
- Lehmann, E. L. (1983). *Theory of Point Estimation*. John Wiley and Sons.
- Lipshitz, S. P., R. A. Wannamker and J. Vanderkooy (1992). Quantization and dither: A theoretical survey. *Journal of Audio Eng. Soc* **40**(5), 355–375.
- Ljung, L. (1999). *System Identification, Theory for the User*. second ed.. Prentice Hall. Englewood Cliffs, New Jersey.
- Nandi, A.K., Ed. (1999). *Blind Estimation Using Higher-order Statistics*. Springer.
- Necchi, L., L. Lavagno, D. Pandini and L. Vanzago (2006). An ultra-low energy asynchronous processor for wireless sensor networks. In: *IEEE International Symposium on Asynchronous Circuits and Systems (ASYNC'06)*. p. 8 pp.
- Oppenheim, A. and R. Schaffer (1975). *Digital Signal Processing*. Prentice-Hall.
- Schuchman, L. (1964). Dither signals and their effect on quantization noise. *IEEE Transaction on Communication Technology* **12**, 162–165.
- Sripad, A.B. and D.L. Snyder (1977). A necessary and sufficient condition for quantization errors to be uniform and white. *IEEE Transactions on Signal Processing* **25**, 442–448.
- Stuart, A. and J. K. Ord (1994). *Kendall's Advanced Theory of Statistics*. Vol. 1. 6 ed.. London: Edward Arnold, New-York Wiley.
- Wang, L.Y. and G.G. Yin (2007). Asymptotically efficient parameter estimation using quantized output observations. *Automatica* **43**, 1178–1191.
- Wang, L.Y., J.F. Zhang and G.G. Yin (2003). System identification using binary sensors. *IEEE Transactions on Automatic Control* **48**(11), 1892–1906.
- Wannamaker, R.A., S.P. Lipshitz, J. Vanderkooy and J.N. Wright (2000). A theory for nonsubtractive dither. *IEEE Transactions on Signal Processing* **48**(2), 499–516.
- Widrow, B. and I. Kollar (2008). *Quantization Noise: Roundoff Error in Digital Computation, Signal Processing, Control, and Communications*. Cambridge University Press.
- Widrow, B., I. Kollar and M-C Liu (1996). Statistical theory of quantization. *IEEE Transactions on Signal Processing* pp. 353–361.