Deterministic Analysis of Oversampled A/D Conversion and $\Sigma\Delta$ Modulation, and Decoding Improvements using Consistent Estimates

Truong-Thao Nguyen

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY
1993
ABSTRACT

Deterministic Analysis of Oversampled A/D Conversion and $\Sigma\Delta$ Modulation, and Decoding Improvements using Consistent Estimates

Truong-Thao Nguyen

Analog-to-digital conversion (ADC) which consists in a double discretization of an analog signal in time and in amplitude is increasingly used in modern data acquisition. However, the conversion process always implies some loss of information due to amplitude quantization. Oversampling is the technique currently used to reduce this loss of accuracy. The error reduction can be performed by lowpass filtering the quantized signal, thus eliminating the high frequency components of the quantization error signal. This is the classical method used to reconstruct the analog signal from its oversampled and quantized version. This reconstruction scheme yields a mean squared error (MSE) inversely proportional to the oversampling ratio $R$.

The fundamental question pursued in this thesis is the following: how much information is available in the oversampled and quantized version of a bandlimited signal for its reconstruction? In order to identify this information, it is essential to go back to the original description of quantization which is typically deterministic. We show that a reconstruction scheme fully takes this information into account when it satisfies a certain condition, called consistency. We point out that the classical reconstruction method is not necessarily consistent and show, under certain conditions on the input signals, that the MSE of a consistent reconstruction scheme depends on $R$ in $\mathcal{O}(R^{-2})$ instead of $\mathcal{O}(R^{-1})$. This deterministic approach is also entirely applicable to $\Sigma\Delta$ modulation which is currently the most successful technique in oversampled ADC. When $n$ is the order of the modulator, consistent reconstruction yields an MSE in $\mathcal{O}(R^{-(2n+2)})$ instead of $\mathcal{O}(R^{-(2n+1)})$ in classical reconstruction. In general, consistent reconstruction yields an improvement of 3 dB per octave of oversampling over classical reconstruction, regardless of the order $n$.

The deterministic analysis of quantization also leads to methods of improvement of non-consistent reconstruction, and the description of optimal reconstruction. In the particular case of constant inputs, we show that $\mathcal{O}(R^{-(2n+2)})$ is also the lower bound to the optimal reconstruction MSE for multi-loop and multi-stage $\Sigma\Delta$ modulation.
# Contents

1 Introduction .......................................................... 1
   1.1 Introduction ..................................................... 1
   1.2 Overview of oversampled encoding .......................... 7
       1.2.1 Simple encoding .......................................... 7
       1.2.2 Predictive encoding and $\Delta$ modulation .......... 8
       1.2.3 Noise-shaping encoding and $\Sigma\Delta$ modulation .... 10
       1.2.4 Higher order noise-shaping encoding and multi-loop $\Sigma\Delta$ mod-
           ulation ...................................................... 12
       1.2.5 Multi-stage $\Sigma\Delta$ modulation .................... 15
   1.3 Basic idea of our approach and outline of the thesis .... 18
   1.4 Past work and contribution of this thesis ................ 19

2 Mathematical context and notations ............................... 22
   2.1 Continuous-time and discrete-time signals ............... 22
   2.2 Space $\mathbb{R}^N$ of discrete-time signals ............... 22
   2.3 Approximation of bandlimited signals .................... 24
   2.4 Space $\mathcal{V}$ of bandlimited signals ................... 24

3 Simple oversampled encoding ........................................ 27
   3.1 Deterministic description of quantization ............... 27
   3.2 The partitioning approach to simple encoding .......... 28
   3.3 Consistent estimates .......................................... 31
   3.4 Non-consistency of linear decoding ....................... 33
   3.5 Methods of improvement of non-consistent estimates .... 35
   3.6 Conceptual method for consistent reconstruction ...... 35
   3.7 Consistent decoding MSE upper bound .................... 36
   3.8 Numerical evaluation of consistent reconstruction .... 41
   3.9 Conclusion .................................................... 41

4 Predictive oversampled encoding .................................. 43
   4.1 Encoding with preprocessing ................................. 43
   4.2 Deterministic description of predictive encoding ....... 43
   4.3 The partitioning approach to predictive encoding ...... 46
   4.4 Consistent estimates and non-consistency of linear decoding .... 47
5 Single-loop and multi-loop ΣΔ modulation
5.1 Structure of multi-loop ΣΔ modulation
5.2 Deterministic description of multi-loop ΣΔ modulation encoding
5.3 The partitioning approach to noise-shaping encoding
5.4 Consistent estimates and non-consistency of linear decoding
5.5 Methods of improvement of non-consistent estimates
5.5.1 Case of single-loop ΣΔ modulation
5.5.2 Case of n-loop ΣΔ modulation with \( n \geq 2 \)
5.6 Numerical evaluation of consistent reconstruction
5.7 MSE upper bound in consistent decoding
5.8 Lower bound on optimal reconstruction with constant inputs
5.9 Unknown initial conditions
5.10 Conclusion

6 Multi-stage ΣΔ modulation
6.1 Introduction
6.2 Deterministic description of multi-stage ΣΔ modulation encoding
6.3 The partitioning approach to multi-stage ΣΔ modulation
6.4 Consistent estimates and non-consistency of linear decoding
6.5 Methods of improvements of non-consistent estimates
6.6 Numerical tests
6.7 Lower bound on optimal reconstruction with constant inputs
6.8 Unknown initial conditions
6.9 Conclusion

7 Conclusion, discussion and future work

A Appendices
A.1 Review of the z-transform
A.1.1 Discrete-time, linear and time invariant system
A.1.2 Definition of the z-transform and properties
A.1.3 Examples of description of LTI using the z-transform
A.1.4 Relation with the discrete Fourier transform
A.2 Lemma on the matrix \( \mathcal{W}(t_1, \ldots, t_W) \)
A.3 Proof of Lemma 4.7.5
A.4 Proof of Lemma 5.7.2
A.5 Step 2 of Algorithm 3' of Section 5.5.2
A.5.1 Reduction to a problem of minimization under equality constraints
A.5.2 Computation issues ........................................... 98
A.5.3 Recursive computation of $\Omega$ ............................... 100
A.5.4 Recursive computation of $\Omega_p$ and $\Psi_{p}$ ................... 102
A.5.5 Recursive construction of $K$ which satisfies condition (A.14) 105
A.5.6 Complete algorithm ........................................... 107
A.5.7 Algorithm C with limited buffer .............................. 107
A.6 Proof of Proposition A.5.4 ...................................... 108
A.7 Proof of Proposition A.5.17 ..................................... 110
A.8 Proof of Proposition A.7.3 ...................................... 113
A.9 Non-overload sufficient condition in multi-loop $\Sigma \Delta$ modulation ... 117
## List of Figures

1.1 Sampling schemes. (a) Nyquist rate. (b) Oversampling rate. .... 2
1.2 Description of amplitude quantization in the example of a two-bit quantizer. The quantized output $C(k)$ is the center of the quantization interval containing the input sample $X(k)$. The quantization intervals are separated by the quantization thresholds. The difference $E(k)$ between $X(k)$ and $C(k)$ is the quantizer error. .... 3
1.3 Simple encoding. (a) Amplitude quantization symbol. (b) Additive error source representation. ......................... 3
1.4 ADC schemes. (a) Nyquist rate. (b) Oversampling rate. .... 4
1.5 Typical power spectral density of encoded signal, under the white noise assumption of the quantizer error signal. (a) Simple encoding. (b) Single-loop $\Sigma\Delta$ modulation. .......................... 6
1.6 Block diagram of predictive encoding. (a) Encoding part. (b) Pre-decoding part. ................................. 8
1.7 Linearized analysis of the whole predictive system (encoding and predecoding) using the additive error source model of quantization. The whole system is equivalent to a simple encoder. .......... 9
1.8 Block diagram of a $\Delta$ modulator. ............................... 9
1.9 Block diagrams of the single-loop $\Sigma\Delta$ modulator. (a) Principle. (b) Actual implementation. ............................. 10
1.10 Linearized analysis of the single-loop $\Sigma\Delta$ modulator using the additive error source model of quantization. The whole system is itself equivalent to an additive source of error. .......... 11
1.11 Block diagrams of the double-loop $\Sigma\Delta$ modulator. (a) Derivation from the single-loop configuration. (b) Linearized analysis using the additive error source model of quantization. (c) Actual implementation. ................................. 13
1.12 Block diagrams of a $n^{th}$ order multi-loop $\Sigma\Delta$ modulator. (a) Actual implementation. (b) Equivalent block diagram. .............. 14
1.13 General block diagram of a noise-shaping encoder: $H$ and $G$ are linear filters satisfying the constraint (1.5) [5]. .............. 15
1.14 Block diagram of a two-stage $\Sigma\Delta$ modulator. (a) Encoding part. (b) Predecoding part. ................................. 16
1.15 Block diagram of a $p$-stage $\Sigma\Delta$ modulator. (a) Encoding part. (b) Predecoding part. ................................................. 17

3.1 Representation of quantization as many-to-one mapping. (a) Single sample quantization. (b) Discrete-time signal quantization. ............ 27

3.2 Geometric representation of the information “$X \in C(C) \cap V$” in simple encoding (uniform quantization). ............................ 29

3.3 Partition induced on $\mathbb{R}^N$ ($N = 2$) by a two-bit quantizer. The shaded lines represent the non-overload region. The dotted lines represent the case of infinite quantizer. .............................. 30

3.4 Partition on $V$ induced by a two-bit quantizer at oversampling ratio $R = 5$, where $V$ is the two-dimensional space of sinusoids of arbitrary phase and period $1$. .................................................. 31

3.5 Geometric representation of the improvement of a non-consistent estimate $\hat{X}$ in terms of distance, implied by a convex projection on $C(C) \cap V$. ................................................................. 32

3.6 Geometric representation of the non-consistency of the linear decoding estimate $P_V[C]$ in simple encoding. ............................... 33

3.7 Time representation of an example of non-consistency of a linear decoding estimate $P_V[C]$: the samples of $P_V[C]$ at time indices $k = 11, 12$ do not belong to the quantization intervals $q^{-1}[C(11)]$ and $q^{-1}[C(12)]$ respectively. ........................................ 34

3.8 QTC information provided by $C$ about any $X$ such that $Q[X] = C$, at two different sampling rates. ........................................... 37

3.9 Dependence of the MSE with the oversampling ratio $R$ for real linear decoding, “ideal” linear decoding and consistent decoding approached by alternating projections. ............................ 42

4.1 Structure of generalized encoder. ............................................. 43

4.2 Partition induced on $\mathbb{R}^N$ ($N = 2$) by two 2-bit predictive encoders of the type of Figure 4.3, including (a) one integrator (case of $\Delta$ modulation), (b) two integrators. The shaded lines represent the non-overload region. The dotted lines represent the case of infinite quantizer. .............................................. 45

4.3 Architecture of predictive encoding, as a high order generalization of $\Delta$ modulation [5]. ........................................................... 46

4.4 Representation of the solution $Y$ to Step 2 of Algorithm 2. .......... 48

4.5 Dependence of the MSE with the oversampling ratio $R$ for real linear decoding, “ideal” linear decoding and consistent decoding approached by alternating projections. ............................ 49

4.6 Representation of the different elements involved in the demonstration of Theorem 4.7,6 ................................................................. 53
5.1 Block diagrams of a multi-loop ΣΔ modulator with zero initial conditions ........................................ 56

5.2 Geometric representation of the information "X ∈ C(C) \cap V" and non-consistency of the linear decoding estimate P_v[C] in ΣΔ modulation. ................................................................. 58

5.3 Partition induced on \( \mathbf{R}^N \) \((N = 2)\) by 2-bit ΣΔ modulators. (a) Single-loop configuration. (b) Double-loop configuration. The shaded lines represent the non-overload region. The dotted lines represent the case of infinite quantizer. .......................... 59

5.4 Examples of partitions of the space V at the oversampling ratio R with single-bit ΣΔ modulators, in the case where V is of 2 dimensional space of zero-dc component sinusoids. The input signal region of operation is represented by the circle in dotted line. ............... 61

5.5 Representation of the solution Y to the "Thread algorithm". ................................. 63

5.6 Representation of the solution Y to Step 2 of Algorithm 3 for 2\textsuperscript{nd} order ΣΔ modulation. ................................................................. 65

5.7 Dependence of the decoding MSE with the oversampling ratio for single-bit single-loop, two-bit double-loop and three-bit triple-loop ΣΔ modulation. (a) Comparison between linear decoding and consistent decoding approached by alternating projections. (b)-(c)-(d) Decoding MSE normalized with the theoretical noise shaping equation (5.7) for linear decoding, consistent decoding approached by alternating projections and finite step alternating projections: (b) single-bit single-loop configuration; (c) two-bit double-loop configuration; (d) three-bit triple-loop configuration. ....................... 67

5.8 Example of partition of the space V at the oversampling ratio R = 4, in the case where V is of 2 dimensional space of zero-phase sinusoids. The partition of the one dimensional space of constant signals is simply the restriction of the 2 dimensional partition to the dc component axis (dashed line). ................................................................. 70

5.9 Relationship between the number of crossed cells and the coefficients \((\alpha_i)_{i=1,...,N}\) of decomposition in the lattice basis \((F_i)_{i=1,...,N}\). ...................... 72

6.1 Linear decoding in multi-stage ΣΔ modulation. ................................. 76

6.2 Equivalent block diagram of a p-stage ΣΔ modulator. ............................ 77

6.3 Equivalent mapping between the input and the p\textsuperscript{th} output quantized signal of a p-stage ΣΔ modulator. ................................................................. 78

6.4 Construction of the partition on \( \mathbf{R}^2 \) induced by a two-stage ΣΔ modulator composed of two single-bit single-loop modulators (MASH structure) in the non-overload region. (a) Partition of the first single-loop encoder (non-overload region). (b) Partition of the p\textsuperscript{th} encoder \((p = 2)\) in the equivalent block diagram of Figure 6.2 (non-overload region). (c) Nesting of partition (b) in the cells of partition (a). (d) Resulting partition of the two-stage modulator. ...................... 80
6.5 Dependence of the decoding MSE with the oversampling ratio in the case of single-bit single-loop single-stage, two-stage and three-stage $\Sigma \Delta$ modulation. (a) Comparison between linear decoding and consistent decoding approached by alternating projections. (b)-(c)-(d) Decoding MSE normalized with the theoretical noise shaping equation (5.7) for linear decoding, consistent decoding approached by alternating projections and finite step alternating projections: (b) single-stage configuration; (c) two-stage configuration; (d) three-stage configuration. ................................. 83

6.6 Division of the segment $\mathcal{D}$ in the partition induced by the $(p - 1)$-stage $\Sigma \Delta$ modulator in the example of a two-stage MASH configuration (see Figure 6.4 (c)). ................................................. 84

A.1 Examples of representation of LTIs with the $z$-transform notation. (a) Delay operator. (b) Discrete-time differentiator. (c) Discrete-time integrator. ......................................................... 91

A.2 Comparison between $X^{(-n)}(k)$ and $X^{(-n)}[\frac{k}{N}]$. ................................. 95

A.3 Example of construction of $K_p$, $Y_p$, (with $1 \leq p \leq 3$). The selected constraints are shown by dark arrows. ................................. 106

A.4 Block diagram derived from the single-loop $\Sigma \Delta$ modulator by replacing the quantizer by a built-in encoder. ................................. 118
Acknowledgements

I wish to express my gratitude to my advisor, Martin Vetterli, for the original discovery of the problem, and his support and collaboration in this work, thus introducing me to the world of research.

I would like to thank Murat Karabatur for his contribution to numerical experiments involved in this work. I would also like to thank Stephane Mallat from the Courant Institute of Mathematical Sciences for fruitful discussions in the field of signal reconstruction and harmonic analysis, Cormac Herley for his help in the publication of this work, Professor Chu from Columbia University for discussions about a conjecture used in the appendix and Professors Tsividis, Suyama and Vallancourt from Columbia University for showing their interest in our research.
Chapter 1

Introduction

1.1 Introduction

The digital communication and signal processing have known a tremendous development due to progress of the technology of digital circuits. However, most natural sources of information are typically analog whereas it is speech, image or signals for scientific measurements. Therefore, analog-to-digital conversion (ADC) is necessary to transform analog information into a digital format. Even in the ideal case of infinite digital processing resources, the ADC resolution, that is the description accuracy of the analog source achieved by the digital conversion, is limited by the analog precision of the front end circuits of the converter. Given the state of the art of the analog circuit technology, quite some effort is being spent to achieve the highest possible ADC resolution.

Analog-to-digital conversion is classically performed by discretizing the analog signal first in time, then in amplitude [1]. Thanks to Shannon’s well known sampling theorem, no information is lost in the time discretization if the input is bandlimited by some frequency $f_m$ and the sampling frequency $f_s$ is equal to (Figure 1.1(a)) or larger (Figure 1.1(b)) than the Nyquist rate $2f_m$. The bandlimited signal can always be uniquely recovered by sinc interpolation of its samples taken at the Nyquist rate (Figure 1.1(a)). To use Shannon’s sampling theorem, an analog lowpass filter is typically used to make the input signal bandlimited.

However, the second discretization in amplitude implies some irreversible loss of information. In other words, the encoding part of ADC is all localized in the second discretization step. Therefore, the discretization in amplitude will be called “amplitude encoding”, or simply “encoding”. The most basic encoding scheme is the amplitude quantization, as described in Figure 1.2. The figure shows that, by quantizing an input sample, an error is added to it. The quantization operator is symbolically represented in Figure 1.3(a) and its additive error source representation in Figure 1.3(b). When the sampling frequency is equal to the Nyquist rate (Figure 1.4(a)), the added error is entirely irreversible. In Nyquist rate ADC, the only way to limit this error is to reduce the quantization level spacing $q$, or
quantization step size, as much as possible. The drawback of this method is that it increases the number of quantization thresholds, and thus, the circuit complexity of the quantizer, while requiring analog circuits of higher precision.

Modern methods for high resolution ADC [2] are based on oversampling where the sampling frequency \( f_s \) is chosen larger than the Nyquist rate. Oversampling is unnecessary in the absence of amplitude encoding (Figure 1.1(b)). But, in ADC, it gives some redundancy that can be exploited to reduce the amount of information lost in the encoding process. Therefore, an extra digital circuit, called decoder (Figure 1.4(b)), is needed to perform this error reduction based on oversampling redundancy. Basically, with the oversampling method, high resolution ADC is achieved by compensating the complexity limitations of analog circuits by their speed performances. The compensation is performed by means of digital computation in the decoder.

In fact, the oversampling situation permits the design of various other amplitude encoding schemes, such as predictive encoding and noise-shaping encoding, where linear and discrete-time filters, and feedback loops are added to the basic quantizer [2]. Thus, very successful encoding architectures were designed achieving high ADC resolution after decoding, while having a simple analog circuit implementation.

Combined efforts invested in encoding optimization and decoding optimization will contribute to achieve the highest possible ADC resolution. The design possibilities for encoding are constrained by the characteristics of analog circuits of the current technology. Efforts are made to design implementable encoding struc-
Figure 1.2: Description of amplitude quantization in the example of a two-bit quantizer. The quantized output $C(k)$ is the center of the quantization interval containing the input sample $X(k)$. The quantization intervals are separated by the quantization thresholds. The difference $E(k)$ between $X(k)$ and $C(k)$ is the quantizer error.

Figure 1.3: Simple encoding. (a) Amplitude quantization symbol. (b) Additive error source representation.
Figure 1.4: ADC schemes. (a) Nyquist. (b) Oversampling rate.
tures which are robust to circuit inaccuracies and imperfections [2]. In the field of noise-shaping encoding, \( \Sigma \Delta \) modulation is currently the most successful encoding scheme, where, in basic versions, only a one-bit quantizer with a low accuracy requirement is necessary, while yielding high ADC resolution after decoding. Efforts invested for the decoding part, given a choice of encoding, consist in exploiting the oversampling redundancy to reduce as much as possible the amount of information lost in the encoding process. This type of research is different from the encoding improvement, since decoding is performed digitally and is a computational problem. The constraint is here the complexity feasibility of the decoding computation. However, compared to the analog circuit constraint, the digital constraint is conceptually more "flexible" in the sense that the complexity limitations depend on how much the user is willing to spend. In a related area of signal processing, one of the largest decoder ever built until 1991 was designed by JPL (Pasadena) to increase by 1 dB the decoding resolution of information sent by a satellite approaching Jupiter. This situation can be in general found in scientific areas where the best signal reconstruction from digital data acquired by remote sensors is desired.

In the present work, we study the question of how oversampling redundancy can be used in ADC to reduce the encoding loss of information as much as possible and what are the theoretical limits of decoding, given an encoding scheme and an oversampling ratio.

The classical approach to analyze the redundancy contained in an encoded signal is to consider the quantizer included in the encoder as an additive source of error which is a white noise independent of the input. This permits a linearized analysis of the various encoding schemes and leads to the conclusion that the encoded signal is the sum of the bandlimited input signal and an error signal which is not bandlimited and spreads out over the whole frequency range. Figure 1.5 shows the typical power spectral density of the encoded signal, under the white noise assumption, using a simple encoder (pure quantization) and a single-loop \( \Sigma \Delta \) modulator. The redundancy is then exploited by canceling the out-of-band energy of the encoded signal, using a linear lowpass filter. This is the classical decoding scheme that we call the linear decoding scheme. Although the white noise assumption is not theoretically justified [3], this methods leads to good performances which are well predicted by the linear model analysis.

Now, the question is whether the oversampling redundancy has been fully exploited using this method. In other words, while the out-of-band energy is obviously redundant, is the remaining in-band error (see Figure 1.5) irreversible? This would be the case if the error signal was really independent of the input signal. However, although a quantizer can always be seen as an additive source of error (Figure 1.3(b)), it is not true that the error signal is independent of the input [3]. If one wants to know the exact information that the oversampling redundancy can recover, it is important to analyze the quantizer starting from its basic definition which is deterministic.
Figure 1.5: Typical power spectral density of encoded signal, under the white noise assumption of the quantizer error signal. (a) Simple encoding. (b) Single-loop ΣΔ modulation.
By applying the deterministic approach to quantization, we show in this work that the linear decoding scheme does not make use of the full oversampling redundancy. In other words, the classical decoding technique does not use the full information about the analog signal available in the encoded signal. We show that the amount of information that is dropped by linear decoding increases asymptotically by 3 dB per octave of oversampling, regardless of the type of employed encoder (simple, predictive or noise-shaping encoder). This result is supported by numerical tests but also by some analytical derivations based on particular assumptions. We also propose methods for decoding improvement, which recover a part of the missing information, whose proportion depends on the computation complexity that the user is ready to invest. Finally, we ask whether this deterministic analysis of quantization leads to the theoretical limits of signal decoding. In multi-loop and multi-stage $\Sigma \Delta$ modulation and in the case of constant input signals, we prove that the 3 dB/octave improvement cannot be surpassed. This result shows the potential of the deterministic analysis to demonstrate the theoretical limits of signal reconstruction in oversampled ADC.

Since our decoding analysis is applied to existing encoding schemes, we give in Section 1.2 an overview of oversampled encoders. In Section 1.3, we introduce the basic idea of our approach and give the outline of the thesis. Finally, in Section 1.4, we situate our work in the context of past research done in the field and summarize the contribution of the thesis.

1.2 Overview of oversampled encoding

Although the concepts introduced in this thesis are quite basic (especially in the most simple case of amplitude quantization), their consequences are effective in the entire field of oversampled ADC which has known a great development since its origin. In particular, a large variety of encoders has been created and has contributed to progress in the amplitude resolution of A/D converters. Before presenting our contribution, it is important to have an overview of these existing encoding schemes.

1.2.1 Simple encoding

Simple encoding is the case where the amplitude discretization is a simple amplitude quantization (Figure 1.3). The quantization error signal was first analyzed by Bennett [4] in 1948. He showed that under certain conditions, this error signal can be satisfactorily approximated as a uniformly distributed white noise independent of the input. Among these conditions are the requirements of a large number of quantization levels and a small quantization step size compared to the input amplitude range. With this statistical approach, the typical variance of the quantization noise, which is naturally a function of the step size $q$, is equal to
Therefore, the natural way to minimize the quantization error is to increase the resolution of the quantizer, that is, increase the number of quantization levels and decrease the quantization step size. Adding one bit to the quantizer implies a quantization resolution multiplied by 2, and a quantization noise reduction of 6 dB. However, this method requires an increasing circuit complexity with increased analog component accuracy, and is sensitive to technological limitations.

In the oversampled situation, as long as the conditions mentioned by Bennett are maintained, the quantization noise spectrum remains flat and spreads out over the whole sampling frequency range, while keeping a constant total power equal to $\frac{q^2}{12}$ (see Figure 1.5(a)). Therefore, only a fraction $\frac{1}{R}$ of this power is located in the input signal bandwidth, where $R$ is the oversampling ratio, equal to $\frac{f_s}{2f_m}$. Then, using a digital lowpass filter with cut off frequency $f_m$ as decoder, the remaining mean square error (MSE) is equal to $\frac{q^2}{12R}$. This means that, whenever the resolution of time discretization is increased by 2 ($R$ is doubled), the quantization noise power is reduced by 3 dB. Therefore, the equivalence of a one-bit gain is achieved by increasing the time resolution by 4.

### 1.2.2 Predictive encoding and $\Delta$ modulation

The idea of predictive encoding is to minimize the amplitude range of the samples to be quantized so that low complexity quantizers can be used. The method is to quantize the difference between the input samples and a prediction of these samples (see Figure 1.6(a)) [5, 6]. The predictive operator, represented by $G$ in

![Figure 1.6: Block diagram of predictive encoding. (a) Encoding part. (b) Prede-coning part.](image)

the figure, uses the past digital values produced by the quantizer. The prediction is based on the features of the input signal, that is, in the case of oversampled ADC, its band limitation. Depending on the accuracy of the prediction, a quantizer with small amplitude range and of low complexity can be used.
An extra operation (we already consider as a part of the decoding process) has to be performed after the encoding to recover a quantized version of the input signal itself. This is shown in Figure 1.6(b). Figure 1.7 shows the equivalence of the whole conversion system using the additive error source representation of quantization. It shows that the error between the input signal $X(k)$ and the predecoded version $\hat{C}(k)$ is reduced to the error generated by the quantizer [6].

The simplest predictive encoder is the $\Delta$ modulator [7], where $G$ is a single discrete-time integrator and the quantizer is reduced to two levels (one-bit quantizer). The block diagram is shown in Figure 1.8 (for the $z$-transform notations, see Appendix A.1). Although in general the conditions of Bennett are no longer valid in predictive encoding [8], the quantization error is still often modeled as a white noise. Despite its limited validity, this approach leads to very good predictions of the system’s performance in practice. As in the simple encoding case, the white noise assumption implies that the in-band quantization noise power of signal $\hat{C}(k)$ is equal to $\frac{q^2}{12R}$. This implies that, for a fixed quantizer, whenever the oversampling ratio is doubled, the quantization noise power contained in the lowpass version $X(k)$ of $\hat{C}(k)$ is reduced by 3 dB as in simple encoding. The difference in predictive encoding is that the quality of prediction also improves.
with the oversampling ratio \([5]\). While keeping the same number of quantization levels, improved prediction allows a reduction of the quantization step size. In \(\Delta\) modulation, the step size can be reduced in proportion to the oversampling ratio. This implies that when the step size \(q\) is optimally chosen for each oversampling ratio, \(\frac{q^2}{12R}\) decreases by 9 dB per octave of \(R\) \([5]\).

### 1.2.3 Noise-shaping encoding and \(\Sigma\Delta\) modulation

Noise-shaping encoders are the second family of high resolution converters using low resolution quantizers. The simplest version is obtained by feeding an integrated version of the input signal to a \(\Delta\) modulator as shown in Figure 1.9(a). This encoder is called \(\Sigma\Delta\) modulator. In practice, the two integrators (input integrator and feedback integrator) are merged into a single integrator by linear combination. This leads to the classical single-loop configuration of \(\Sigma\Delta\) modulation, shown in Figure 1.9(b) \([9, 10, 11]\). This encoder was introduced in \([12]\) and became very successful due to its good performances, its simplicity of implementation and its robustness to circuit imperfections. One of its strong features is its stability property. The quantizer, although limited to two levels \(\pm \frac{q}{2}\), never overloads if the input signal is confined in the interval \([-\frac{q}{2}, \frac{q}{2}]\). Considering the quantizer as an additive source of error, the output can be linearly expressed as a function of the input and
the quantization error signal. Using $z$-transform notations (see Appendix A.1), it can be easily derived that [13]:

$$C(z) = X(z) + (1 - z^{-1})E(z),$$

(1.1)

where $E$ is the quantization error signal. This expression shows that the output is the sum of the input signal (without any linear distortion) and the error signal. Therefore, the encoder can be itself considered as an additive source of error as represented by Figure 1.10. The power spectral density of this error signal is equal to $4 \left( \sin \frac{\omega}{2} \right)^2 |E(\omega)|^2$ (see Appendix A.1.4), which is the spectral density of the quantizer error signal multiplied by the “shaping” function $4 \left( \sin \frac{\omega}{2} \right)^2$. This justifies the term of “noise-shaping” encoder. Although the conditions of Bennett are not valid, it is still often assumed that the quantizer behaves like a white noise source [10]. With this approximation, the power spectral density of the encoded output signal has the shape given in Figure 1.5(b). The shaping function has the effect to “push” away the quantization noise power from the low frequency region. Under the white noise assumption, it is shown in [13] that, for the oversampling ratio $R$, the in-band portion of the quantization noise is equal to $\frac{\pi^2}{3} \frac{q^2}{12R^3}$. This implies that by lowpass filtering the encoded output signal, the quantization noise power is reduced by $9$ dB per octave of $R$. This result is similar to $\Delta$ modulation with the difference that the quantization step size is chosen according to the input amplitude range, regardless of the oversampling ratio. The noise reduction is completely based on the higher order dependence of the noise power with $R$. Although the white noise assumption is not verified in theory [3], the $9$ dB/octave is observed in practice when lowpass filtering the encoded signal.

Figure 1.10: Linearized analysis of the single-loop $\Sigma\Delta$ modulator using the additive error source model of quantization. The whole system is itself equivalent to an additive source of error.
1.2.4 Higher order noise-shaping encoding and multi-loop \(\Sigma \Delta\) modulation

In general, a noise-shaping encoder is obtained when the output can be expressed in terms of the input and the quantization error signal as follows:

\[
C(z) = X(z) + S(z)E(z),
\]

where \(S(z)\) is some “shaping” function. To obtain high performance in-band noise reduction, the goal is to design an encoder such that \(S(z)\) has minimum energy at low frequencies. An idea is to replace the quantizer of Figure 1.9(b) by an encoder of higher performance, for example, a single-loop \(\Sigma \Delta\) modulator. From equation (1.1), a single-loop \(\Sigma \Delta\) modulator can be itself seen as an additive source of error equal to \((1 - z^{-1})E(z)\), where \(E(z)\) is the error signal of the built-in quantizer. Therefore, Figure 1.11(a) is equivalent to Figure 1.11(b). Thus, it can be easily derived that the relationship between the input, the output and \(E(z)\) is

\[
C(z) = X(z) + (1 - z^{-1}) \left( (1 - z^{-1}) E(z) \right) = X(z) + (1 - z^{-1})^2 E(z).
\]

The shaping function \(S(z) = (1 - z^{-1})^2\) has lower energy at low frequencies since its power spectral density is proportional to \((\sin \frac{\pi}{2})^4\). Assuming that the quantization error signal is still a white noise, it is shown in [5, 14] that the in-band noise power is proportional to \(R^{-5}\). The block diagram of Figure 1.11(a) leads to the equivalent implementation shown in Figure 1.11(c) and called double-loop \(\Sigma \Delta\) modulation. This architecture was first introduced in [12]. This configuration owns its success to the fact that the system remains stable although the quantizer is limited to two levels.

In a similar way, noise-shaping of higher order can be obtained by iterating the nesting of single-loop \(\Sigma \Delta\) modulators. This leads to the encoder of Figure 1.12(a) called \(n^{th}\) order multi-loop \(\Sigma \Delta\) modulator. This generalized architecture was first introduced in [15]. The relationship (1.2) has the form

\[
C(z) = X(z) + (1 - z^{-1})^n E(z),
\]

and the in-band noise power contained in the encoded signal is [5]

\[
\frac{\pi^{2n}}{2n + 1} \cdot \frac{q^2}{12R^{2n+1}}.
\]

However, the drawback of this configuration is that, when the quantizer is limited to two levels, the system becomes unstable as soon as \(n \geq 3\). Therefore, this requires the quantizer to be multi-bit.

In general, according to [5] a noise-shaping encoder can be thought of as the composition of some linear filter with a predictive encoder, as shown in Figure 1.13. This is already true in single-loop \(\Sigma \Delta\) modulation but can be seen for the multi-
Figure 1.11: Block diagrams of the double-loop ΣΔ modulator. (a) Derivation from the single-loop configuration. (b) Linearized analysis using the additive error source model of quantization. (c) Actual implementation.
Figure 1.12: Block diagrams of a \( n \)th order multi-loop \( \Sigma\Delta \) modulator. (a) Actual implementation. (b) Equivalent block diagram.
loop configuration as shown by the equivalent block diagram of Figure 1.12(b). The constraint is that equation (1.2) be respected. From Figure 1.13, it can be easily derived that

\[ C(z) = \frac{H(z)}{1 + G(z)} \cdot X(z) + \frac{1}{1 + G(z)} \cdot E(z). \]

Therefore, a noise-shaping encoder is in general an encoder whose block diagram can be equivalently described by Figure 1.13 with the constraint

\[ H(z) = 1 + G(z). \]  \hfill (1.5)

Then, the noise-shaping relation (1.2) will be satisfied with the particular shaping function

\[ S(z) = H^{-1}(z). \]  \hfill (1.6)

In the case of \( n \)th order multi-loop \( \Sigma \Delta \) modulation, \( H \) is an \( n \)th order integrator, as shown in Figure 1.12(a), which implies in \( z \)-transform notations that \( H(z) = \frac{1}{(1 - z^{-1})^n} \). This is another way to derive (1.3) using (1.6).

### 1.2.5 Multi-stage \( \Sigma \Delta \) modulation

Another way to improve noise-shaping is to reencode the quantization error signal resulting from a noise-shaping encoder. Suppose an input signal \( X \) is encoded through a first noise-shaping encoder and yields the output signal

\[ C_1(z) = X(z) + H_1^{-1}(z)E_1(z). \]

Suppose that the quantization error \( E_1 \) is reencoded through a second noise-shaping encoder giving the output

\[ C_2(z) = -E_1(z) + H_2^{-1}(z)E_2(z) \]
(we assume more precisely that $-E_1$ is reencoded). By taking the following linear combination $\hat{C}_2$ of $C_1$ and $C_2$:

$$\hat{C}_2(z) = C_1(z) + H_1^{-1}(z)C_2(z),$$

we cancel the error signal $E_1$ and find

$$\hat{C}_2(z) = X(z) + \left(H_1^{-1}(z)H_2^{-1}(z)\right)E_2(z).$$

The complete encoding scheme thus obtained is shown in Figure 1.14. By this

![Block diagram of a two-stage ΣΔ modulator](image)

Figure 1.14: Block diagram of a two-stage ΣΔ modulator. (a) Encoding part. (b) Predecoding part.

method, the noise included in $\hat{C}_2$ is generated by the second quantizer, but shaped with the more efficient function $H_1^{-1}(z)H_2^{-1}(z)$. For example, if the two encoders are single-loop ΣΔ modulators, the noise-shaping function of the complete system is $(1 - z^{-1})^2$. This gives the same function as in double-loop ΣΔ modulation. This configuration is usually called two-stage ΣΔ modulation and was introduced in [16]. It owns its success to the fact that single-bit quantizers can be used, avoiding the non-linearity problems of multi-bit quantizers due to circuit imperfections.

The general multi-stage configuration is obtained by iterating this cascading process as shown in Figure 1.15. It is easy to derive that the shaping function of such a system is

$$S(z) = H_1^{-1}(z)H_2^{-1}(z)...H_p^{-1}(z).$$

In the case of an $n$-stage ΣΔ modulator composed of single-loop modulators, we obtain

$$\hat{C}_n(z) = X(z) + (1 - z^{-1})^nE_n(z)$$

and leads to the expression (1.3) of in-band quantization noise power. The multi-stage configuration was introduced in [17] and called the MASH structure.
Figure 1.15: Block diagram of a $p$-stage $\Sigma\Delta$ modulator. (a) Encoding part. (b) Predecoding part.
1.3 Basic idea of our approach and outline of the thesis

The first hint that the linear decoding scheme is not optimal can be seen in the case of simple encoding. We saw in Section 1.2.1, that, to obtain an equivalent amplitude resolution increased by 2, it is necessary to increase the time resolution by 4. This asymmetry is disappointing, since a bandlimited signal with bounded amplitude has a limited slope. Thus, one expects that a variation along the amplitude dimension, or at least an upper bound, is linearly equivalent to a variation along the time dimension. This lack of symmetry is a hint that linear decoding might be suboptimal, not just by a fixed amount, but rather, by a factor dependent on $R$.

Using a deterministic analysis in the case of simple encoding (Chapter 3), we show that the information which is missing in linear decoding is the consistency constraint. We call an estimate of an analog input signal consistent when it reproduces the same quantized signal if it is to be requantized. To see this, we explain in Section 3.1 that a quantizer operates like a many-to-one mapping in the space of discrete-time signals. The set of consistent estimates is simply obtained by inverse image of this mapping. We point out in Section 3.2 that by collecting the sets of consistent estimates for every possible encoded signals, a partition of the space of input signals is defined and uniquely characterizes the encoder (partitioning approach). We show in Section 3.3 that any set of consistent estimate has the property to be convex. As a consequence, we show that any non-consistent estimate is not optimal and can be improved. We show in Section 3.4 that a linear decoding estimate is not necessarily consistent. We propose in Section 3.5 methods of improvements of non-consistent estimates based on convex projections. Moreover, consistent estimates can be approached by iterating these convex projections (alternating projection method). In Section 3.7 we derive analytically what improvement should be achieved by taking a consistent estimate. Under certain assumptions about the quantization threshold crossings of the input signal, we show that the MSE of a consistent estimate decreases at the speed $O(R^{-2})$ instead of $O(R^{-1})$ for linear decoding. Thus, we recover the expected symmetry between time and amplitude. Using our methods of improvement, we confirm this $O(R^{-2})$ result numerically in Section 3.8. This gives the first demonstration of the 3 dB/octave improvement.

In the next chapters (Chapters 4, 5 and 6), we apply the same approach to predictive, multi-loop and multi-stage ΣΔ modulation. Following the same reasoning as in Chapter 3, we consider any type of encoder as a many-to-one mapping in the space of discrete-time signals. The notion of consistent estimate and the partitioning approach are constructed in a similar way. The properties of convexity hold and the principles of non-consistent estimate improvement are still valid. The non-consistency of linear decoding appears also in a similar way.
In predictive encoding, methods of improvement of non-consistent estimates are similar to simple encoding, with the same computation complexity. In the case of Δ modulation, numerical tests show that consistent decoding improves linear decoding by, again, 3 dB/octave. This means that when the quantization step size \( q \) is optimized according to the oversampling ratio \( R \) (see Section 1.2.2), the consistent decoding MSE increases to 12 dB/octave, instead of 9 dB/octave. The theoretical analysis of this result is more difficult. However, starting with some model assumptions, we derive in Section 4.7 that the MSE dependence of consistent estimates is upper bounded by \( \mathcal{O}\left(\frac{q^2}{R^2}\right) \) instead of \( \mathcal{O}\left(\frac{q^2}{R}\right) \).

The deterministic analysis applied to noise-shaping encoders is studied in Chapter 5 with emphasis on the multi-loop configuration of ΣΔ modulation (including the popular single-loop configuration). Methods of improvements are based on the same principles but require algorithms of higher complexity. The single-loop configuration requires an original algorithm called the “Thread algorithm” (Section 5.5.1) and higher order configurations require more sophisticated algorithms fully described in Appendix A.5. Numerical tests show that, regardless of the order of the modulator, consistent decoding improves the linear decoding MSE from \( \mathcal{O}(R^{-(2n+1)}) \) to \( \mathcal{O}(R^{-(2n+2)}) \). The theoretical analysis is performed in Section 5.7 in a way similar to predictive encoding, starting with the same model assumptions. It confirms the \( \mathcal{O}(R^{-(2n+2)}) \) behavior. In Section 5.8, we prove the important result that the order \( \mathcal{O}(R^{-(2n+2)}) \) is also the lower bound in the case of constant input signals. The derivation is based on the partitioning properties of multi-loop ΣΔ modulators. With this result, we get to the goal of theoretical limits of signal reconstruction in oversampled ADC.

The final chapter deals with multi-stage ΣΔ modulators (Chapter 6). The more complex architecture of these encoders requires some preliminary analysis before the deterministic approach can be applied (Section 6.2). Then, Sections 6.3 to 6.6 show that the study of consistent decoding is similar to the case of multi-loop ΣΔ modulation: methods of improvement, algorithms, numerical results and analysis of MSE upper bound. The same lower bound \( \mathcal{O}(R^{-(2n+2)}) \) is derived in the case of constant input signals in Section 6.7.

### 1.4 Past work and contribution of this thesis

Historically, signal reconstruction in oversampled ADC is based on the white noise approach of the quantization error signal, and is performed by linear processing (linear decoding method). The first non-classical approach to signal reconstruction in oversampled ADC was introduced in [18, 19] in the particular field of ΣΔ modulation (single-loop and double-loop configurations) and the context of constant input signals. It was pointed out that a given output bit stream (or encoded signal) corresponds to a whole interval of input dc values. Therefore, in the con-
text of constant inputs, optimal decoding is achieved by taking the center of this interval. It was shown that this decoding scheme is not linear. Unfortunately, the performances obtained from this non-linear decoder showed the asymptotic dependence of the MSE in $\mathcal{O}(R^{-3})$ and $\mathcal{O}(R^{-5})$ for the single-loop and the double-loop $\Sigma\Delta$ modulators, which are already obtained with linear decoding. It was also proved in [19, 20] that this constant input non-linear decoder cannot yield better performances than $\mathcal{O}(R^{-3})$ and $\mathcal{O}(R^{-5})$ respectively.

The first non-linear study of signal reconstruction of time-varying bandlimited input signals was introduced in [21] in the general context of oversampled ADC. Although the term of consistency was not used, the set of consistent estimates was introduced and its convexity was proved. The non-consistency of linear decoding estimates was shown and their non-optimality was deduced as a consequence of the consistent set convexity. The alternating projection algorithm was introduced to achieve bandlimited consistent estimates and led to the MSE performance of the order of $\mathcal{O}(R^{-3})$, $\mathcal{O}(R^{-4})$ and $\mathcal{O}(R^{-6})$ in simple encoding, single-loop and double-loop $\Sigma\Delta$ modulation respectively, instead of $\mathcal{O}(R^{-1})$, $\mathcal{O}(R^{-3})$ and $\mathcal{O}(R^{-5})$ using linear decoding (provided some conditions on the input signals in the case of simple encoding). This was the first indication of asymptotic improvements achievable by non-linear decoding.

The idea of consistent set and alternating projection was later adopted in [22, 23, 24] with an approximated implementation of the convex projections. This permitted improved signal reconstruction for a $4^{th}$ order single-loop modulator introduced in [25]. Meanwhile, algorithms for the convex projections in the case of $n^{th}$ order multi-loop and multi-stage $\Sigma\Delta$ modulation were developed and described in [26]. Numerical results showed the general $\mathcal{O}(R^{-(2n+2)})$ MSE behavior obtained by the alternating projections on $\Sigma\Delta$ modulators of any order $n$, thus indicating an asymptotic improvement of 3 dB per octave over linear decoding, regardless of the type of encoder.

The question of analytical justification of this improvement was dealt at the same time in several papers. A qualitative justification of the consistent decoding MSE in simple encoding was given in [21] and the complete proof was detailed in [26, 27]. Starting from some model assumptions, the $\mathcal{O}(R^{-(2n+2)})$ MSE behavior of consistent decoding in $\Sigma\Delta$ modulation was derived in [26, 28]. The formalism of consistent reconstruction in relation with the deterministic description of quantization was developed in [29, 28]. First results of reconstruction theoretical limits were derived in [32].

This thesis formalizes all the results we originally obtained in [21, 26, 28, 30, 27, 29, 31, 32], and presents them as consequences of the fundamental deterministic definition of quantization. The contributions of this thesis are listed below.
(i) Study of oversampled ADC without the approximation of white quantization noise.
(ii) Description of quantization and amplitude encoding as signal mapping, instead
of additive source of error.

(iii) Development of mathematical tools to derive the mapping defined by encoders used in oversampled ADC (simple encoders, predictive encoders, $\Sigma\Delta$ modulation).

(iv) Presentation of the exact information contained in a quantized or encoded signal in oversampled ADC, as a signal set, called set of consistent estimates.

(v) Equivalent deterministic description of an encoder as inducing a signal partition.

(vi) Proof of convexity of the set of consistent estimates.

(vii) Consequence on consistency as a necessary condition for reconstruction optimality.

(viii) Geometric localization of quantized (or encoded) signal with respect to the set of consistent estimate (as the center of a hyper-parallelepiped).

(ix) Performance of consistent decoding MSE measured by numerical experiments showing an improvement of 3 dB per octave of oversampling over linear decoding regardless of the encoding scheme (simple, predictive encoding or $\Sigma\Delta$ modulation).

(x) Mathematical derivation of this improvement in simple encoding.

(xi) First analytical derivations of this improvement in predictive encoding and $\Sigma\Delta$ modulation, based on some model assumptions.

(xii) Methods of improvement of non-consistent estimates based on convex projections.

(xiii) Algorithms for these convex projections, for simple, predictive encoding, 1st order and $n^{th}$ order multi-loop and multi-stage $\Sigma\Delta$ modulation.

(xiv) Derivation of theoretical limit of reconstruction in multi-loop and multi-stage $\Sigma\Delta$ modulation in the case of constant inputs.
Chapter 2

Mathematical context and notations

The deterministic approach in oversampled ADC implies the description of quantization, and amplitude encoding in general, in a non-classical way and requires the introduction of a specific formalism. This chapter defines all the non-classical notations which will be essential to this work.

2.1 Continuous-time and discrete-time signals

Continuous-time signals, denoted by $X[t]$, are real signals and are assumed to be observed on a finite length time interval $[0, T_0]$. By a time rescaling, we can always assume that $T_0 = 1$. The error between two signals $X[t]$ and $X'[t]$ is measured by the mean squared error (MSE) equal to

$$MSE(X[t], X'[t]) = \int_0^1 |X'[t] - X[t]|^2 dt.$$ (2.1)

We assume that continuous-time signals are sampled $N$ times in the interval $[0, 1]$ at instants $\frac{k}{N}$, for $k = 1, \ldots, N$. The discrete-time signals are thus elements of $\mathbb{R}^N$ and denoted as $Y = (Y(k))_{k=1,\ldots,N}$. The sampled version of a continuous-time signal $X[t]$ is the discrete-time signal $X$ such that $X(k) = X(\frac{k}{N})$ for $k = 1, \ldots, N$.

2.2 Space $\mathbb{R}^N$ of discrete-time signals

Notations relative to $\mathbb{R}^N$ are defined here. Subsets of $\mathbb{R}^N$ are designated by calligraphic letters (e.g. $\mathcal{A}, \mathcal{B}, \mathcal{C}$). When $B_1, \ldots, B_N$ are $N$ subsets of $\mathbb{R}$, $\mathcal{B} = B_1 \times \cdots \times B_N$ is the subset of $\mathbb{R}^N$ defined by

$$\mathcal{B} = \{ X \in \mathbb{R}^N \mid \forall k = 1, \ldots, N, \ X(k) \in B_k \}.$$ (2.2)
If $\mathcal{C}$ is a subset of $\mathbb{R}^N$, $\mathcal{C} + X$ denotes the translated version of $\mathcal{C}$ by $X$, that is:

$$\mathcal{C} + X = \{ Y + X \mid Y \in \mathcal{C} \}.$$ 

A mapping of $\mathbb{R}^N$ is a function $H$ mapping any element of $\mathbb{R}^N$ into another unique element of $\mathbb{R}^N$ denoted $H[X]$. The value of the sequence $H[X]$ at instant $k$ is denoted by $H[X](k)$. The notation $I$ denotes the identity mapping. If $H$ is a one-to-one (or invertible) mapping, $H^{-1}$ designates the inverse mapping. If $H_1$ and $H_2$ are two mappings of $\mathbb{R}^N$, $H_2H_1$ is the mapping such that $H_2H_1[X] = H_2[H_1[X]]$. If $\mathcal{C}$ is a subset of $\mathbb{R}^N$, $H[\mathcal{C}]$ and $H^{-1}[\mathcal{C}]$ designate the forward and inverse images of $\mathcal{C}$ through $H$ respectively ($H$ need not be invertible). When $\mathcal{C}$ is reduced to a singleton $\mathcal{C} = \{C\}$, by abuse of notation $H^{-1}[\mathcal{C}]$ is denoted by $H^{-1}[C]$. With this notation, when $H$ is not invertible, $H^{-1}[\mathcal{C}]$ is a subset of $\mathbb{R}^N$, that is $H^{-1}[\mathcal{C}] = \{ X \in \mathbb{R}^N \mid H[X] = C \}$. For illustration purposes, we show here an example of use of these notations, which will be used later in this work.

If $\mathcal{C}$ is an element of $\mathbb{R}^N$ and $H$, $G$, $Q$ are three mappings of $\mathbb{R}^N$ where $H$ is invertible, then $H^{-1}[Q^{-1}[\mathcal{C}] + G[\mathcal{C}]]$ is a subset of $\mathbb{R}^N$ which is constructed as follows:

(i) $Q^{-1}[\mathcal{C}]$ is a subset of $\mathbb{R}^N$, which is the inverse image of $\mathcal{C}$ through $Q$,

(ii) $Q^{-1}[\mathcal{C}] + G[\mathcal{C}]$ is the subset $Q^{-1}[\mathcal{C}]$ translated by the fixed element $G[\mathcal{C}]$ of $\mathbb{R}^N$,

(iii) $H^{-1}[Q^{-1}[\mathcal{C}] + G[\mathcal{C}]]$ is the forward image of the subset $Q^{-1}[\mathcal{C}] + G[\mathcal{C}]$ through the mapping $H^{-1}$.

For $Y \in \mathbb{R}^N$, $Y^{(1)}$, $Y^{[1]}$ and $Y^{(-1)}$ are respectively the backward discrete-time derivative, the forward discrete-time derivative and the discrete-time integral of $Y$, defined as follows:

$$\begin{align*}
  \begin{cases} 
    Y^{(1)}(1) = Y(1), \\
    Y^{(1)}(k) = Y(k) - Y(k - 1), \quad \text{for } k = 2, \ldots, N.
  \end{cases} 
\end{align*}$$

$$\begin{align*}
  \begin{cases} 
    Y^{[1]}(k) = Y(k + 1) - Y(k), \quad \text{for } k = 1, \ldots, N - 1, \\
    Y^{[1]}(N) = -Y(N)
  \end{cases} \quad \text{(2.4)}
\end{align*}$$

$$Y^{(-1)}(k) = \sum_{j=1}^{k} Y(j), \quad \text{for } k = 1, \ldots, N. \quad \text{(2.5)}$$

For $Y \in \mathbb{R}^N$ and $n \geq 0$, $Y^{(n)}$, $Y^{[n]}$ and $Y^{(-n)}$ are respectively the $n^{th}$ order backward discrete-time derivative, forward discrete-time derivative and discrete-time integral of $Y$, recursively defined as follows:

$$Y^{(0)} = Y \quad \text{and for } n \geq 0, \quad \begin{align*}
  \begin{cases} 
    Y^{(n+1)} = \left( Y^{(n)} \right)^{(1)}, \\
    Y^{[n+1]} = \left( Y^{[n]} \right)^{[1]}, \\
    Y^{(-n-1)} = \left( Y^{(-n)} \right)^{(-1)}.
  \end{cases}
\end{align*} \quad \text{(2.6)}$$
Finally, for \( X, Y \in \mathbb{R}^N \), we define the inner product \( \langle X, Y \rangle = \frac{1}{N} \sum_{k=1}^{N} X(k)Y(k) \). We recall that \( (\mathbb{R}^N, \langle \cdot, \cdot \rangle) \) is a euclidean space. The associated norm is \( \|X\| = \langle X, X \rangle^{1/2} \).

### 2.3 Approximation of bandlimited signals

The considered continuous-time signals are assumed to be perfectly bandlimited with cut-off frequency \( f_m \). This implies that they have an infinite support, or, are non-zero on an interval of infinite length. However, as already said, they will be observed only on a finite interval, as is always the case in practice, such as \([0, 1]\) which can be obtained by time rescaling. Even in the case of oversampling, Shannon’s sampling theorem does not lead to unicity of reconstruction if the known samples are limited to the interval \([0, 1]\). We therefore introduce an approximation of bandlimited signals which will allow us to recover Shannon’s unicity. We assume that the energy outside \([0, 1]\) of the considered bandlimited signals is small enough and decays fast enough so that their restrictions to \([0, 1]\) are “almost equal” to the periodized versions, obtained by adding subsequently their translated versions by integer multiples. By “almost”, we mean that the error made by this approximation (or aliasing error) is at least small compared to other sources of errors existing in the ADC process, such as those due to quantization for example. In the frequency domain, this means that the Fourier transform of a bandlimited signal is approximated by its discrete frequency version, where the period of discretization is equal to \( f_0 = \frac{1}{T_0} = 1 \). Therefore, the approximation consists in saying that the restrictions of input signals to the interval \([0, 1]\) are elements of the space \( \mathcal{V} \) of all possible signals bandlimited to \( f_m \) and 1 periodic.

In this work, we thus propose to study the behavior of oversampled ADC when input signals are in general elements of \( \mathcal{V} \).

### 2.4 Space \( \mathcal{V} \) of bandlimited signals

We assume that \( f_m \) is an integer \( M \). The elements of \( \mathcal{V} \) have the form

\[
X[t] = \sum_{i=-M}^{M} X_i e^{j2\pi i t}, \quad \text{where} \quad X_{-i} = X_i^* \quad \text{for} \quad i = 0, \ldots, M. \tag{2.7}
\]

Therefore, \( \mathcal{V} \) is a real vector space of dimension

\[
W = 2M + 1. \tag{2.8}
\]

When \( X[t] \) is given by (2.7), we define the complex vector

\[
\vec{X} = [X_{-M} \cdots X_M] \in \mathbb{C}^W. \tag{2.9}
\]
There is a one-to-one mapping between \( X \in \mathbb{R}^N \) and its complex vector representation \( \tilde{X} \). Moreover, if we define the norm

\[
\| \tilde{X} \| = \left( \sum_{i=-M}^{M} |X_i|^2 \right)^{\frac{1}{2}},
\]

then from Parseval’s equality, we have

\[
\int_{0}^{1} |X[t]|^2 dt = \| \tilde{X} \|^2.
\]

In other words, there is an isometry relationship between \( X[t] \in \mathcal{V} \) and \( \tilde{X} \in \mathbb{C}^W \). Similarly, we know that, when \( N \geq W \), there is an isometric relationship between \( \tilde{X} \in \mathbb{C}^W \) and the sampled version \( X \in \mathbb{R}^N \) of \( X[t] \in \mathcal{V} \). We first have the following theorem:

**Theorem 2.4.1** When \( N \geq W \), there is a linear one-to-one mapping between \( \tilde{X} \in \mathbb{C}^W \) and the sampled version \( X \in \mathbb{R}^N \) of \( X[t] \in \mathcal{V} \).

**Proof:** Using (2.7), there is a trivial linear mapping from \( \tilde{X} \) to the sampled version \( X \) of \( X[t] \). Let us show this is a one-to-one mapping. For \( W \) instants \( t_1, \ldots, t_W \in [0, 1] \), let us define the Vandermonde matrix:

\[
\mathcal{W}(t_1, \ldots, t_W) = \begin{bmatrix} e^{i2\pi t_{1k}} \end{bmatrix}_{1 \leq k \leq W} \cdot e^{i2\pi t_{i'k}}
\end{bmatrix}_{-M \leq i \leq M}.
\]

We recall that this matrix is invertible when \( e^{i2\pi t_{ik}} \neq e^{i2\pi t_{i'k}} \) for any \( t_k \neq t_{i'} \). One can check that

\[
[X[t_1] \cdots X[t_W]]^T = \mathcal{W}(t_1, \ldots, t_W) \cdot \tilde{X}.
\]

Since \( N \geq W \), we can take \( t_k = \frac{k}{N} \in [0, 1] \). For this choice of \( t_k \), \( \mathcal{W}(t_1, \ldots, t_W) \) is invertible. Therefore \( \tilde{X} \) is uniquely defined by \( [X[t_1] \cdots X[t_W]]^T \) which is itself uniquely defined by \( X = (X(k))_{1 \leq k \leq N} = \left(X[k]\right)_{1 \leq k \leq N} \). □

The isometric relation

\[
\| \tilde{X} \| = \| X \|
\]

is obtained by applying the discrete version of Parseval’s equality on

\[
X(k) = X[k] = \sum_{i=-M}^{M} X_i e^{i2\pi \frac{ik}{N}}.
\]

Therefore, as a finite dimensional version of Shannon’s sampling theorem, we have shown that, when \( N \geq W \), \( X[t] \in \mathcal{V} \) is uniquely defined by its sampled version \( X \) (\( W \) is analogous to the bandwidth and \( N \) to the sampling frequency).
Moreover, when \( N \geq W \), if \( \mathbf{X} \) and \( \mathbf{X}' \) are the sampled versions of two signals \( X[t] \) and \( X'[t] \) of \( \mathcal{V} \), we have from (2.1) and (2.11)

\[
MSE(X[t], X'[t]) = \| \mathbf{X}' - \mathbf{X} \|^2. \tag{2.16}
\]

Because of these unicity and isometry properties, the bandlimited signals will be implicitly considered in their discrete-time version, and \( \mathcal{V} \) will be considered as a subspace of \( \mathbb{R}^N \). Oversampling occurs when \( N > W \). In this case, \( \mathcal{V} \) is a subspace of \( \mathbb{R}^N \) in the strict sense. The oversampling ratio

\[
R = \frac{N}{W} \tag{2.17}
\]

gives the ratio between the dimensions of \( \mathbb{R}^N \) and \( \mathcal{V} \).
Chapter 3

Simple oversampled encoding

The basic mechanisms used in a deterministic analysis of oversampled ADC can be most easily demonstrated in the case of simple encoding. Their detailed description is necessary in order to convey the basic concepts and prepare the framework for generalization to predictive, noise-shaping encoding and multi-stage encoding.

3.1 Deterministic description of quantization

In a deterministic approach, quantization is simply a many-to-one mapping of \( \mathbb{R}^N \). In the particular case \( N = 1 \), where signals are reduced to single sample values, quantization is a mapping of \( \mathbb{R} \), denoted by \( q \), such that whole intervals of \( \mathbb{R} \), called quantization intervals, are mapped into single discrete values, called quantization levels (see Figure 3.1(a)). The quantization interval corresponding

![Diagram](image)

Figure 3.1: Representation of quantization as many-to-one mapping. (a) Single sample quantization. (b) Discrete-time signal quantization.
to a quantization level \(c\) is denoted by \(q^{-1}[c]\). The second characteristic of the
quantization mapping is that it is a consistent mapping. By this, we mean that,
for any quantization level \(c\), \(q[c] = c\), or, \(c \in q^{-1}[c]\). In practice, \(c\) is typically
chosen to be the center of the \(q^{-1}[c]\). This will be assumed in this thesis. Actually,
a quantizer has a finite number of quantization levels and the two extreme quanti-
zation intervals are necessarily infinite. If \(c\) is one of the two extreme quantization
levels, it is typically chosen to be the center of \(q^{-1}[c] \cap B\), where \(B\) is the specified
bounded region of input samples, called non-overload region. When quantization
is uniform (this will not be necessarily assumed), the quantization intervals have
a common length, denoted by \(q\) and called quantization step size. For the extreme
quantization intervals, it is implicitly defined as \(q^{-1}[c] \cap B\) which has length \(q\).

In the general case where \(N \geq 1\), quantization is a mapping \(Q\) of \(R^N\) such that
\[
\forall X \in R^N, \quad C = Q[X] \iff \forall k = 1, \ldots, N, \quad C(k) = q[X(k)].
\]
We say that \(C\) is the encoded version of \(X\). Rigorously speaking, if an input signal
\(X \in R^N\) is only known by its encoded version \(C\), the full information available
about \(X\) is: \(\{ X \in \mathcal{C}(C) \}\) where \(\mathcal{C}(C)\) is the set of possible input signals with
encoded version \(C\). By definition, we have \(\mathcal{C}(C) = Q^{-1}[C]\), where
\[
Q^{-1}[C] = \{ Y \in R^N / \forall k = 1, \ldots, N, \; Y(k) \in q^{-1}[C(k)] \}.
\]
Therefore, when quantization is uniform, \(\mathcal{C}(C)\) is a hyper-cube of \(R^N\) (see Figure
3.1(b)) and \(C\) is its geometric center. In the more general case of non-uniform
quantization, \(\mathcal{C}(C)\) is a rectangular hyper-parallelepiped\(^1\).

Now, in oversampled ADC, we have the extra information about the input
signal that it belongs to \(\mathcal{V}\). We therefore have the following proposition:

**Proposition 3.1.1** In oversampled ADC, when a bandlimited signal \(X\) is only
known by the encoded signal \(C\) it produces through the encoder, the full information
available about \(X\) is:

\[
\{ X \in \mathcal{C}(C) \cap \mathcal{V} \}.
\]

This information is represented geometrically in Figure 3.2.

### 3.2 The partitioning approach to simple encoding

Another way to characterize a mapping is to consider its inverse transform. In
the case of quantization which is not a one-to-one mapping, describing its inverse

\(^1\)To be rigorous, \(Q^{-1}[C]\) may not be bounded. However, the picture of hyper-cube or hyper-
parallelepiped with \(C\) as geometric center holds when taking \(\mathcal{C}(C) = Q^{-1}[C] \cap B^N\), which simply
assumes that input signals \(X\) belong to the non-overload region, or, \(X(k) \in B\) for \(k = 1, \ldots, N\).
Figure 3.2: Geometric representation of the information “$X \in C(C) \cap \mathcal{V}$” in simple encoding (uniform quantization).

The transform is mapping any encoded signal $C$ into the set of all signals of $\mathbb{R}^N$ which produce $C$ through the quantizer, that is, the set $C(C)$. In other words, a quantizer can be equivalently characterized by the following partition of $\mathbb{R}^N$:

\[
\{ C(C) / C \text{ is an encoded signal} \}. \tag{3.1}
\]

We call it the partition of $\mathbb{R}^N$ induced by the quantizer. Figure 3.3 shows the partition induced by a two-bit quantizer in the space $\mathbb{R}^2$ ($N = 2$).

If the quantizer is moreover uniform and infinite, that is, has an infinite number of equally spaced quantization levels with step size $q$, the partition forms a hyper-cubic lattice in $\mathbb{R}^N$, of period $q$. This lattice is indicated by the dotted lines in Figure 3.3. It is generated by a vector basis $(E_i)_{i=1,...,N}$ which is simply the canonical basis of $\mathbb{R}^N$ multiplied by $q$ (see Figure 3.3). The definition of the lattice basis is more precisely:

\[
\forall i = 1, ..., N, \begin{cases} E_i(k) = q & \text{if } k = i, \\ E_i(k) = 0 & \text{otherwise}. \end{cases} \tag{3.2}
\]

In general, the partition induced by a finite quantizer can always be thought of as extracted from the lattice (see the solid lines in Figure 3.3). We will say that the partition is an extraction of the lattice.

In oversampled ADC, quantization can be regarded as a mapping from $\mathcal{V}$ and $\mathbb{R}^N$. Following the same reasoning, the quantizer can be equivalently characterized by
Figure 3.3: Partition induced on $\mathbb{R}^N \ (N = 2)$ by a two-bit quantizer. The shaded lines represent the non-overload region. The dotted lines represent the case of infinite quantizer.
the following partition of $\mathcal{V}$:

$$\{C(C) \cap \mathcal{V} \mid C \text{ is an encoded signal}\}^2.$$ \hfill (3.3)

We call it the partition of $\mathcal{V}$ induced by the quantizer. Figure 3.4 shows the partition when the quantizer is 2 bit, the oversampling ratio is $R = 5$ and $\mathcal{V}$ is the two dimensional space of sinusoids of arbitrary phase and period 1.

![Diagram](image)

Figure 3.4: Partition on $\mathcal{V}$ induced by a two-bit quantizer at oversampling ratio $R = 5$, where $\mathcal{V}$ is the two-dimensional space of sinusoids of arbitrary phase and period 1. As confirmed by this figure, there is no reason for the partition of $\mathcal{V}$ to have a lattice structure or to be a extraction of a lattice. However, one has to keep in mind that this partition is derived by intersection of $\mathcal{V}$ with a hyper-cubic lattice, or an extraction of it. In the case of Figure 3.4, the hyper-cubic lattice is of dimension 10.

### 3.3 Consistent estimates

The goal of signal decoding in oversampled ADC is to reconstruct an estimate $\hat{X}$ of an original input signal $X$ from its encoded version $C$. Since the knowledge of $C$ implies the information “$X \in C(C) \cap \mathcal{V}$”, it is tempting to pick as estimate of $X$ an element $\hat{X}$ of $C(C) \cap \mathcal{V}$. For reasons which will become clear later, we are

\[^2\text{Not every encoded signal can be produced by bandlimited signals. In this case, we simply have } C(C) \cap \mathcal{V} = \emptyset\]
particularly interested in the set of estimates \( \overline{C(C)} \cap \mathcal{V} \), where \( \overline{C(C)} \) is the closure of \( C(C) \). We propose the following definition:

**Definition 3.3.1** When \( C \) is the encoded signal produced by a bandlimited input signal \( X \), the elements of \( \overline{C(C)} \cap \mathcal{V} \) are called the consistent estimates of \( X \). If \( \hat{X} \) is a consistent estimate of \( X \in \mathcal{V} \), then \( \hat{X}[t] \) will be called a consistent estimate of \( X[t] \).

We show that when an estimate of \( X \) is not consistent, then, by necessity, it can be theoretically improved by a consistent estimate. This is based on the fact (easy to verify) that \( C(C) \) and \( \mathcal{V} \) are convex sets and the two following lemmas.

**Lemma 3.3.2** [33] Let \( Y \) be an element of \( \mathbb{R}^N \) and \( S \) a closed and convex set. There exists a unique element \( Y' \) of \( S \) such that for all \( Z \in S \), \( \| Y' - Y \| \leq \| Z - Y \| \). The transformation from \( Y \) to \( Y' \) is then a mapping of \( \mathbb{R}^N \) called the convex projection on \( S \) and denoted by \( P_S \).

**Lemma 3.3.3** [34] If \( Y' \) is the convex projection of \( Y \) on a closed and convex set \( S \), and \( Y \notin S \), then for all \( Z \in S \), \( \| Y' - Z \| < \| Y - Z \| \).

It is easy to verify that \( S = \overline{C(C)} \cap \mathcal{V} \) is closed and convex. Applying Lemma 3.3.3 to \( S = \overline{C(C)} \cap \mathcal{V} \) and using the fact that \( X \in \overline{C(C)} \cap \mathcal{V} \), we obtain the following property:

**Property 3.3.4** Let \( X \) be a bandlimited signal producing the encoded signal \( C \). Let \( \hat{X} \) be a non-consistent estimate of \( X \), that is \( \hat{X} \notin \overline{C(C)} \cap \mathcal{V} \). Then the convex projection \( \hat{X}' \) of \( \hat{X} \) on \( \overline{C(C)} \cap \mathcal{V} \) is a consistent estimate of \( X \) and \( \| \hat{X}' - X \| < \| \hat{X} - X \| \).

This property is illustrated by Figure 3.5 which shows that the projection \( \hat{X}' \) of

![Figure 3.5: Geometric representation of the improvement of a non-consistent estimate \( \hat{X} \) in terms of distance, implied by a convex projection on \( \overline{C(C)} \cap \mathcal{V} \).](image)

\( \hat{X} \) on \( \overline{C(C)} \cap \mathcal{V} \) is necessarily closer to \( X \) than \( \hat{X} \). Basically, this property implies that, when \( \hat{X} \) is a non-consistent estimate of \( X \), the knowledge of \( \hat{X} \) and \( C \) gives enough information to characterize a consistent estimate which is always better than \( \hat{X} \).
3.4 Non-consistency of linear decoding

We show in this section that linear decoding in oversampled ADC, does not necessarily provide consistent estimates. This can first be seen geometrically. Linear decoding consists in lowpass filtering the encoded signal $C$ at cut-off frequency $f_m$. In the space $\mathbb{R}^N$, this amounts to performing an orthogonal projection of $C$ on the subspace of bandlimited signals. In other words, the linear decoding scheme provides as estimate of $X$ the projection on $V$ of the center of the hyper-cube $\mathcal{C}(C)$. As indicated in Figure 3.6, unless the cube $\mathcal{C}(C)$ lies at a particular “angle”

![Diagram](image)

Figure 3.6: Geometric representation of the non-consistency of the linear decoding estimate $P_V[C]$ in simple encoding.

with $V$, there is no reason for this estimate to belong to $\mathcal{C}(C)$.

Figure 3.7 shows a concrete example of this non-consistency in the time domain. A bandlimited signal is generated numerically, sampled at the oversampling ratio $R = 4$ and encoded. The figure shows that the estimate $\hat{X}$ obtained from lowpass filtering the encoded signal $C$ does not belong to $\mathcal{C}(C)$ since, for example, the samples $\hat{X}(11)$ and $\hat{X}(12)$ do not belong to quantization intervals $q^{-1}[C(11)]$ and $q^{-1}[C(12)]$, respectively.
Figure 3.7: Time representation of an example of non-consistency of a linear decoding estimate $P_y[C]$: the samples of $P_y[C]$ at time indices $k = 11, 12$ do not belong to the quantization intervals $q^{-1}[C(11)]$ and $q^{-1}[C(12)]$ respectively.
3.5 Methods of improvement of non-consistent estimates

Property 3.3.4 gives the mathematical justification for possible improvement of non-consistent estimates, but does not really provide a method to achieve it. In fact, a method for at least partial improvement can be obtained by applying Lemma 3.3.3 on either $\mathcal{S} = \mathcal{V}$ or $\mathcal{S} = \mathcal{C}(\mathcal{C})$. If the non-consistent estimate $\hat{X}$ does not belong to $\mathcal{V}$, then according to Lemma 3.3.3, it will be necessarily improved by a projection on $\mathcal{V}$, that is a lowpass filtering at cut-off frequency $f_m$. Similarly, if $\hat{X} \notin \mathcal{C}(\mathcal{C})$, an improvement can be achieved by using a projection on $\mathcal{C}(\mathcal{C})$. This projection is performed by the following algorithm, which is similar to the algorithm introduced in [35] for two dimensional image reconstruction.

**Algorithm 1**: At every instant $k$,

(i) if $\hat{X}(k) \in q^{-1}[C(k)]$, take $X'(k) = \hat{X}(k)$,

(ii) else, take $X'(k)$ equal to the bound of the quantization interval $q^{-1}[C(k)]$ closest to $\hat{X}(k)$.

Qualitatively speaking, the algorithm consists in projecting each sample $\hat{X}(k)$ on the quantization interval $q^{-1}[C(k)]$ indicated by the encoded value $C(k)$ when $\hat{X}(k) \notin q^{-1}[C(k)]$. It is easy to verify that this leads to an estimate $\hat{X}'$ which is the projection of $\hat{X}$ on $\mathcal{C}(\mathcal{C})$. In particular, this projection algorithm can be used for immediate improvement of the linear decoding estimate, since this corresponds to the case $\hat{X}(k) \notin q^{-1}[C(k)]$. This improvement is illustrated in Figure 3.7 by the dark arrows.

In fact, as long as an estimate does not belong to $\mathcal{C}(\mathcal{C})$ or $\mathcal{V}$, a projection on either $\mathcal{C}(\mathcal{C})$ or $\mathcal{V}$ can be reiterated, thus implying further reductions of the distance between the current estimate and $X$. The obtained improvement will always be an increasing function of the number of iterations.

3.6 Conceptual method for consistent reconstruction

It was shown in [34] that alternating projections infinitely between two intersecting closed and convex sets necessarily converges to their intersection. This property became the basis of the algorithm of alternating projection classically used in signal processing, and first introduced by Youla [36] for image reconstruction. In our case, this property ensures that the limit of alternating projections between $\mathcal{C}(\mathcal{C})$ and $\mathcal{V}$ constitute a consistent estimate of $X$. Numerically speaking, this implies that a consistent estimate can at least be approached using finite step alternating projections.
3.7 Consistent decoding MSE upper bound

Starting the alternating projection scheme from the estimate provided by linear decoding is a way to find a consistent estimate which automatically improves this decoding scheme. The question is now to have an analytical evaluation of this improvement. Our approach is to find an upper bound to any consistent estimate of $X$ and compare it with the expected MSE of a classical estimate. In this section we show the following theorem:

**Theorem 3.7.1** Let $X[t] \in \mathcal{V}$ be a bandlimited signal which crosses the quantization thresholds at least $W$ times in the interval $[0, 1]$. Then, there exists a constant $\alpha_X > 0$ which does not depend on the oversampling ratio $R$ such that, for $R$ high enough, and any consistent estimate $\hat{X}[t]$ of $X[t]$ at oversampling ratio $R$,

$$MSE(X[t], \hat{X}[t]) \leq \frac{\alpha_X}{R^2}.$$ 

This theorem is based on

(i) a particular interpretation of the information contained in the quantized signal $C$ about the original signal $X[t]$;

(ii) the fact that bandlimited and bounded signals have a bounded slope.

Let us first describe the interpretation of (i). The signal $X[t]$ can be viewed as a two-dimensional graph where the two axes represent the time and the amplitude dimensions respectively (see Figure 3.8(a)). With this approach, the quantized signal $C$ gives the location of graph points, with absolute precision in time and “uncertainty” in amplitude. In Figure 3.8(a), the localized points are represented by crosses. The amplitude uncertainty is an interval specification represented by shaded vertical segments. The proof of Theorem 3.7.1 will be based on a dual interpretation which says that $C$ equivalently gives the location of a set of graph points with absolute precision in amplitude and uncertainty in time. This can be seen as follows. Suppose that the quantization intervals indicated by $C(k - 1)$ and $C(k)$ at the two consecutive time indices $k - 1$ and $k$ are located at two different levels and on each side of a quantization threshold $l$. An example can be seen in Figure 3.8(a) at $k = 4$ and $l = 4q$. Since $t \mapsto X[t]$ is a continuous graph, we know for sure that $X[t]$ crosses the level $l$ somewhere between $t = \frac{k-1}{N}$ and $t = \frac{k}{N}$. We will say that $X[t]$ has a quantization crossing (QTC) at level $l$ and in the time interval $t \in [\frac{k-1}{N}, \frac{k}{N}]$. This QTC is a graph point (represented in Figure 3.8(a) by a dot) whose location is known with absolute precision in amplitude and uncertainty in time. This location information is represented in the figure by dark horizontal segments.

If the oversampling ratio is high enough so that the sampling period is smaller than the minimum distance between the QTCs of $X[t]$, every QTC will be localized in a horizontal segment as shown in Figure 3.8(b). In other words, every single QTC is localized with a time uncertainty equal to the sampling period $T_s = \frac{1}{N}$.
Figure 3.8: QTC Information provided by $C$ about any $X$ such that $Q[X] = C$, at two different sampling rates.
(length of the horizontal segments). Let \( n \) be the number of QTCs and \((t_1, \ldots, t_n)\) their instants. If \( \hat{X}[t] \) is a consistent estimate of \( X[t] \), it has necessarily the same quantized version \( C \). As a consequence, \( \hat{X}[t] \) has \( n \) QTCs at the same levels as those of \( X[t] \) and at instants \((t'_1, \ldots, t'_n)\) such that

\[
\forall i = 1, \ldots, n, \quad |t'_i - t_i| \leq T_s = \frac{1}{N} \tag{3.4}
\]

(see Figure 3.8(b)). Note that, when the oversampling is increased infinitely, \((t'_1, \ldots, t'_n)\) tends to be equal to \((t_1, \ldots, t_n)\). Consequently, at infinite oversampling, any consistent estimate \( \hat{X}[t] \) of \( X[t] \) necessarily coincides in amplitude with \( X[t] \) at the \( n \) distinct instants \( t_1, \ldots, t_n \). Suppose that \( n \geq W \). We have that, at infinite oversampling, any consistent estimate \( \hat{X}[t] \) of \( X[t] \) coincides with \( X[t] \) at least at \( W \) distinct instants \( t_1, \ldots, t_W \). The invertibility of the Vandermonde matrix \( \mathbf{W}(t_1, \ldots, t_W) \) defined in (2.12) (Section 2.4) and found in equation (2.13) implies that any consistent estimate coincides with \( X[t] \). Qualitatively speaking, we conclude that when \( X[t] \) has more than \( W \) QTCs, perfect reconstruction is achieved by linear decoding at infinite oversampling.

Now, suppose that the oversampling is finite, but still high enough so that (3.4) holds. Although we may have \( t'_i \neq t_i \), we still have

\[
\forall i = 1, \ldots, n, \quad X[t_i] = \hat{X}[t'_i], \tag{3.5}
\]

since the QTCs of \( X[t] \) and \( \hat{X}[t] \) are respectively at the same levels. Moreover, from the property mentioned in (ii), \( \hat{X}[t] \) has a maximum slope \( S_{\text{max}} \), and its variations between \( t_i \) and \( t'_i \) is upper bounded as follows:

\[
\forall i = 1, \ldots, n, \quad |\hat{X}[t'_i] - X[t_i]| \leq S_{\text{max}} \cdot |t'_i - t_i| \tag{3.6}
\]

From (3.4), (3.5) and (3.6), we deduce:

\[
\forall i = 1, \ldots, n, \quad |\hat{X}[t_i] - X[t_i]| = |\hat{X}[t_i] - \hat{X}[t'_i]| \leq \frac{S_{\text{max}}}{N}
\]

Qualitatively speaking, this means that any consistent estimate \( \hat{X}[t] \) of \( X[t] \) coincides with \( X[t] \) at the instants \((t_1, \ldots, t_n)\), with an amplitude error in \( \mathcal{O}(N^{-1}) \). Assuming \( n \geq W \), the invertibility of the Vandermonde matrix in equation (2.13) implies that any consistent estimate \( \hat{X}[t] \) differs from \( X[t] \) uniformly in the interval \([0, 1]\) by an error proportional to \( N^{-1} \). In other words, because of the bounded slope, errors in amplitude are linearly equivalent to errors in time. In terms of MSE, the error between \( X[t] \) and \( \hat{X}[t] \) is of the order of \( \mathcal{O}(N^{-2}) = \mathcal{O}(R^{-2}) \).

We have thus given a qualitative justification of the result of Theorem 3.7.1. We now show the rigorous proof.

**Proof of Theorem 3.7.1:** Let us choose \( W \) QTCs of \( X[t] \), call \( 0 < t_1 < \ldots < t_W \leq T \) their instants and \((l_1, \ldots, l_W)\) their levels. We have \( X[t_i] = l_i \) for all
\( i = 1, \ldots, W \). Using the convention \( t_0 = 0 \), let \( \delta \) be the minimum distance between \( t_0, \ldots, t_W \), that is \( \delta = \min_{1 \leq i \leq W} (t_i - t_{i-1}) \) (see Figure 3.8). Let us choose \( R \) large enough so that the sampling period \( T_s = \frac{1}{N} \) is smaller than \( \frac{\delta}{3} \). With this choice of \( R \), these instants belong to distinct sampling intervals. We conclude that, for any consistent estimate \( \hat{X}' \) of \( \hat{X} \)

\[
\exists t_1', \ldots, t_W' \in [0, 1], \forall i = 1, \ldots, W, \quad X'[t_i] = l_i \text{ and } |t_i' - t_i| < T_s = \frac{1}{N}. \tag{3.7}
\]

From the condition \( T_s \leq \frac{\delta}{3} \), we have the following constraint:

\[
\forall i = 1, \ldots, W, \quad t_i' \geq t_{i-1}' + \frac{\delta}{3}, \text{ with the convention } t_0' = 0, \tag{3.8}
\]

since

\[
t_i' - t_{i-1}' = (t_i' - t_i) + (t_i - t_{i-1}) + (t_{i-1} - t_{i-1}') \geq t_i - t_{i-1} - |t_i' - t_i| - |t_{i-1} - t_{i-1}'| \geq \delta - T_s - T_s \geq \frac{\delta}{3}.
\]

Using the fact that \( X'[t_i'] = X[t_i] = l_i \) for every \( i = 1, \ldots, W \), and the definition of \( \mathcal{W}(t_1, \ldots, t_W) \) in (2.12), we find that

\[
\mathcal{W}(t_1', \ldots, t_W') \cdot \vec{X}' = \mathcal{W}(t_1, \ldots, t_W) \cdot \vec{X} = [l_1 \cdots l_W]^T, \tag{3.9}
\]

Subtracting \( \mathcal{W}(t_1, \ldots, t_W) \cdot \vec{X}' \) in the two first terms of this equation, we find

\[
(\mathcal{W}(t_1', \ldots, t_W') - \mathcal{W}(t_1, \ldots, t_W)) \cdot \vec{X}' = -\mathcal{W}(t_1, \ldots, t_W) \cdot (\vec{X}' - \vec{X}). \tag{3.10}
\]

Because of (3.7), \( |t_i' - t_i| \) is upper bounded by \( \frac{1}{N} \) and thus goes to zero when \( N \) goes to infinity. Therefore, in the limit of \( N \) going to \( \infty \), we have

\[
\mathcal{W}(t_1', \ldots, t_W') - \mathcal{W}(t_1, \ldots, t_W) \overset{N \to \infty}{\approx} \sum_{i=1}^{W} (t_i' - t_i) \frac{\partial \mathcal{W}}{\partial t_i}(t_1, \ldots, t_W) \tag{3.11}
\]

Because \( (t_1, \ldots, t_W) \) are distinct in \([0,1] \), \( \mathcal{W}(t_1, \ldots, t_W) \) is invertible (see proof of Theorem 2.4.1). Because of (3.8), \( (t_1', \ldots, t_W') \) are also distinct in \([0,1] \) and \( \mathcal{W}(t_1', \ldots, t_W') \) is also invertible. In fact, relation (3.8) implies the stronger fact that \( [\mathcal{W}(t_1', \ldots, t_W')]^{-1} \) is bounded. This is a consequence of Lemma A.2.1 shown in Chapter A.2. Therefore, this implies that \( \vec{X}' = [\mathcal{M}(t_1', \ldots, t_W')]^{-1} \cdot [l_1 \cdots l_W]^T \) is bounded (expression obtained from (3.9)). Using (3.10) and (3.11), we can thus write

\[
\vec{X}' - \vec{X} \overset{N \to \infty}{\approx} -[\mathcal{W}(t_1, \ldots, t_W)]^{-1} \sum_{i=1}^{W} (t_i' - t_i) \frac{\partial \mathcal{W}}{\partial t_i}(t_1, \ldots, t_W) \cdot \vec{X}' \tag{3.12}
\]

39
Since \( \tilde{X}' \) is bounded, the right hand side goes to zero when \( N \) goes to \( \infty \). Therefore, \( \tilde{X}' \) tends to \( \tilde{X} \), and (3.12) is still true when replacing \( \tilde{X}' \) by \( \tilde{X} \) in the right hand side. We obtain

\[
\tilde{X}' - \tilde{X} \approx N \rightarrow \infty \sum_{i=1}^{W} (t_i' - t_i) \vec{F}_i \quad \text{where} \quad \vec{F}_i = - [W(t_1, ..., t_W)]^{-1} \frac{\partial W}{\partial t_i}(t_1, ..., t_W) \cdot \tilde{X}.
\]

(3.13)

For each \( i = 1, ..., W \), \( \vec{F}_i \) is a vector which depends only on the signal \( X \). Using the fact that \( |t_i' - t_i| \) is upper bounded by \( \frac{1}{N} \), (3.13) implies that, for \( N \) large enough \( \|\tilde{X}' - \tilde{X}\| \leq \frac{1}{N} \sum_{i=1}^{W} \|\vec{F}_i\| \). Using relations (2.16) and (2.14), and the fact that \( N = RW \), the proof is completed by taking \( \alpha_x = \left( \frac{1}{W} \sum_{i=1}^{W} \|\vec{F}_i\| \right)^2 \quad \square \)

The above result is to be compared with the classical decoding MSE which is proportional to \( \frac{1}{R} \). This means that under the special condition on the quantization threshold crossings of Theorem 3.7.1, the consistent decoding MSE asymptotically decreases at least at the rate of 6dB per octave of \( R \), instead of 3dB per octave for the linear decoding MSE. Regardless of the value of \( \alpha_x \), this means that the speed of MSE reduction is improved by 3dB per octave of \( R \), when \( R \) is high enough.

**Remark 1:** The only assumption used in this theorem was the number of QTCs of the input signal. For example, this does not require quantization to be uniform. However, the constant \( \alpha_x \) obtained in the upper bound may depend on how the input signal crosses the thresholds. In particular, \( \alpha_x \) depends on \( \vec{F}_i \) which contains the term \( \frac{\partial W}{\partial t_i}(t_1, ..., t_W) \cdot \tilde{X} \). One can verify that this expression contains the slope of \( X \) at the \( i^{th} \) QTC.

**Remark 2:** If the number of QTCs is less than \( W \), the size of the set \( \mathcal{C}(\mathcal{C}) \cap \mathcal{V} \) tends to a non-zero value when the oversampling goes to infinity, and perfect reconstruction cannot be achieved. This condition for perfect reconstruction is similar to the Nyquist condition in Shannon’s sampling theorem. In the case of bandlimited and periodic signals, the Nyquist condition requires that the number of samples \( N \) be larger or equal to \( W \) (see Section 2.4). The number of samples can be interpreted as the number of crossings of the graph \( t \mapsto X[t] \) with the sampling thresholds, that is, the vertical lines located at each sampling instant (see Figure 3.8(a)).

\(^3^\text{This is valid provided that} \tilde{X} \neq 0. \text{This is indeed the case because the QTC requirement on} \ X \text{implies that} \ X \text{is a non-zero signal.}\)
3.8 Numerical evaluation of consistent reconstruction

Numerical tests were performed to evaluate the performance of consistent estimates obtained by alternating projections, with respect to the oversampling ratio $R$. For a given number $W$, bandlimited signals of the form (2.7) were randomly generated, with the constraint that they cross the quantization thresholds at least $W$ times. This is ensured by imposing a minimum to their peak-to-peak amplitude (PPA). For example, in uniform quantization with step size $q$, when $W = 3$, forcing the PPA to be equal to $2q$ where $q$ is the quantization step size, ensures that the input signals have at least 3 quantization threshold crossings. Then, for a fixed oversampling ratio $R$, the encoded version of each of these input signals was computed, as well as an approximately consistent estimate obtained by alternating projections. The linear decoding estimate was used as first estimate in the projection iteration, and the alternating process was stopped as soon as the estimate MSE decrement per iteration was less than 0.001dB. The resulting MSE is averaged over all randomly generated input signals. Although this MSE is not the MSE of rigorously consistent estimates, since the iteration of alternating projections is finite, one can be sure that the consistent estimates obtained from infinite iteration would necessarily yield an even smaller MSE.

The evolution of the averaged MSE versus the oversampling ratio is plotted in Figure 3.9, in the case where $W = 3$, quantization is uniform, PPA=2$q$ and $R$ varies between 20 and 170 approximately. The MSE is averaged over 1000 generated input signals. In the same figure we plot the MSE predicted by the classical decoding model equal to $\frac{q^2}{12R}$, whose slope is -3dB/octave. Note that with PPA=2$q$, the number of quantization level is limited to 3. With this very low quantization resolution it is known that the classical decoding MSE no longer decreases by 3dB/octave since the quantization error signal becomes very correlated to the input signal. The numerical results of Figure 3.9 show that oversampling does no longer reduce the real linear decoding MSE which stagnates. For consistent estimates, there is no such stagnation, and the slope of -6dB/octave is experimentally verified.

3.9 Conclusion

We have shown the deterministic approach of oversampled ADC in the most basic case of encoding which consists in simple quantization. A simple encoder defines a many-to-one mapping in the space of discrete-time signals $\mathbb{R}^N$. The image of a quantized signal $C$ through the inverse of this mapping is a whole set $\mathcal{C}(C)$, which is a rectangular hyper-parallelepiped of $\mathbb{R}^N$ (a hyper-cube if quantization is uniform) whose center is $C$ itself. In the context of oversampled ADC, simple encoding is a many-to-one mapping from the space $\mathcal{V}$ of bandlimited signals to $\mathbb{R}^N$. Therefore, the exact information about an input signal $X \in \mathcal{V}$ from its quantized
version $C$, is that $X$ belongs to the image of $C$ through the inverse of this mapping, which is the set $C(C) \cap \mathcal{V}$, called the set of consistent estimates. In fact, the encoder, as mapping from $\mathcal{V}$ to $\mathbb{R}^N$, can be equivalently described by the partition constituted by the family of sets $\{C(C) \cap \mathcal{V} / C\}$, called the partition induced by the encoder. When $X$ produces $C$, because $C(C) \cap \mathcal{V}$ is convex, any estimate of $X$ should be consistent, otherwise it is not optimal. We have shown that estimates obtained from the classical linear decoding are not necessarily consistent. Moreover, we show that, under certain conditions on the input quantization thresholds, the dependence of consistent reconstruction MSE with $R$ is at least in $O(R^{-2})$ instead of $O(R^{-1})$ in linear decoding. A method for immediate improvement of a non-consistent estimate consists in performing a convex projection of the estimate either on $\mathcal{C}(C)$ or $\mathcal{V}$. While the projection on $\mathcal{V}$ is a lowpass filtering, the algorithm for the projection on $\mathcal{C}(C)$ has a very straightforward implementation. These projections have the extra feature that they indeed converge to a consistent estimate when alternated and iterated infinitely. This method was used to approach consistent estimates numerically. Experiments show the $O(R^{-2})$ MSE behavior of consistent estimates (provided the same conditions on the input quantization threshold crossings). This represents an asymptotic improvement of the MSE of 3 dB per octave of oversampling.
Chapter 4

Predictive oversampled encoding

4.1 Encoding with preprocessing

Predictive encoders were described in Section 1.2.2 and their block diagram given in Figure 1.6(a). It includes the particular case of \( \Delta \) modulation, when \( \mathbf{G} \) is a discrete-time integrator. In the deterministic approach, \( \mathbf{G} \) is considered as a mapping, as was already done for \( \mathbf{Q} \). We will only assume that \( \mathbf{G} \) is a strictly causal mapping of \( \mathbb{R}^N \), that is, when \( \mathbf{D} = \mathbf{G}[\mathbf{C}] \), \( D(k) \) only depends on \( C(1), \ldots, C(k-1) \) for \( k = 1, \ldots, N \). In particular, \( D(1) \) is a constant \( G_1 \) independent of \( \mathbf{C} \).

We show in Section 4.2 that a predictive encoder is itself a many-to-one mapping of \( \mathbb{R}^N \) which has the particular structure of Figure 4.1, where the one-to-one mapping \( \mathbf{F} \) can be expressed in terms of \( \mathbf{G} \) and \( \mathbf{Q} \). Then, the resulting expression of \( \mathcal{C}(\mathbf{C}) \) will show that \( \mathcal{C}(\mathbf{C}) \) is still a rectangular hyper-parallelepiped of \( \mathbb{R}^N \), as in simple encoding. The notion of consistent estimation can be applied, and linear decoding non-consistency and methods of improvements will be studied as extensions of the simple encoding case.

![Diagram](image)

Figure 4.1: Structure of generalized encoder.

4.2 Deterministic description of predictive encoding

The equivalence of a predictive encoder with the block diagram of Figure 4.1 is based on the following lemma:
Lemma 4.2.1 \( I + GQ \) is a one-to-one mapping of \( \mathbb{R}^N \).

Proof: For a given \( X \in \mathbb{R}^N \), suppose that there exists \( A \in \mathbb{R}^N \) such that \( X = (I + GQ)[A] \). Let \( D = G[Q[A]] \). This implies that \( A = X - D \). From the assumption of strict causality of \( G \), we know that \( D(1) \) is a constant \( G_1 \), independent of the input of \( G \). Therefore \( A(1) \) is uniquely defined. Now, suppose that for a certain \( k = 1, ..., N - 1 \), we have proved that \( A(1), ..., A(k) \) are uniquely defined from the knowledge of \( X \). Then \( D(k+1) \) is uniquely defined from \( q[A(1)], ..., q[A(k)] \) because \( G \) is strictly causal. This uniquely defines \( A(k+1) = X(k+1) - D(k+1) \). We have therefore proved by induction that when \( A \) exists, it is uniquely defined. This induction actually shows an explicit construction of \( A \), and therefore gives at the same time the existence proof. Therefore \( I + GQ \) is an invertible (or one-to-one) mapping. \( \Box \)

As a consequence we have the following two propositions:

**Proposition 4.2.2** The predictive encoder of Figure 1.6(a) has the general structure of Figure 4.1 where the preprocessing one-to-one mapping is

\[
F = (I + GQ)^{-1}.
\]

Proof: The signal \( A \) defined in Figure 1.6(a) verifies \( A = X - G[Q[A]] \) which implies that \( (I + GQ)[A] = X \). We have from Lemma 4.2.1 that \( I + GQ \) is a one-to-one mapping. Therefore \( A = F[X] \) where \( F = (I + GQ)^{-1} \) \( \Box \)

**Proposition 4.2.3** The set of signals producing an encoded signal \( C \) through the predictive encoder of Figure 1.6(a) is

\[
C(C) = Q^{-1}[C] + G[C].
\]

Proof: Applying Proposition 4.2.2, we have

\[
C(C) = F^{-1}[Q^{-1}[C]] = (I + GQ)[Q^{-1}[C]] = I[Q^{-1}[C]] + GQ[Q^{-1}[C]] = Q^{-1}[C] + G[C] \quad \Box
\]

This means that, as in simple encoding, \( C(C) \) is a rectangular hyper-parallelepiped of \( \mathbb{R}^N \) (or a hyper-cube if quantization is uniform), since it is equal to the rectangular hyper-parallelepiped \( Q^{-1}[C] \) translated by the vector \( G[C] \) of \( \mathbb{R}^N \).

The main point is that Proposition 3.1.1 relative to the encoded information in the oversampling situation is still applicable in the case of predictive encoding and the geometric representation of this information is still that of Figure 3.2, since \( C(C) \) is a hyper-cube.
Figure 4.2: Partition induced on $\mathbb{R}^N$ ($N = 2$) by two 2-bit predictive encoders of the type of Figure 4.3, including (a) one integrator (case of $\Delta$ modulation), (b) two integrators. The shaded lines represent the non-overload region. The dotted lines represent the case of infinite quantizer.
4.3 The partitioning approach to predictive encoding

As in simple encoding, a predictive encoder can be equivalently characterized by the partition it defines on \( \mathbb{R}^N \), using the same definition (3.1) (in Section 3.2). As illustration, Figure 4.2(a) shows the partition of \( \mathbb{R}^2 \) induced by a two-bit \( \Delta \) modulator with uniform quantization. The partition induced by a predictive encoder can be easily derived from the partition induced by the quantizer. Indeed, according to (4.2.3), each cell \( C(C) \) is obtained by translating \( Q^{-1}[C] \), which is a cell of the partition induced by the quantizer, with the vector \( G[C] \). The partition which would be obtained if the quantizer was uniform and infinite, is represented in dotted lines. Note that, as in simple encoding, the resulting partition forms a cubic lattice of period \( q \). This is because the values of the translating vector \( D = G[C] \) are multiples of the quantization step \( q \), since this is the case for the encoded signal \( C \) and the operator \( G \) is limited to pure summations. To be more rigorous, the possible values of \( D(k) \) at a given instant \( k \) are multiples of \( q \) with a possible offset. However, this offset only depends on the instant \( k \) and can be called \( G_0(k) \). Then the translation of each cell \( Q^{-1}[C] \) by \( G[C] \) falls into a cubic lattice of period \( q \), with origin translated by the vector \( G_0 \) (see the case of Figure 4.2(a)). As in simple encoding, this lattice is generated by the canonical basis \( (E_i)_{i=1,...,N} \) defined in (3.2) (Section 3.2).

There exists a whole family of predictive encoders whose partition forms a cubic lattice. It is derived from \( \Delta \) modulation by increasing the number of integrators in the feedback (see Figure 4.3). This class of predictive encoders was studied in [5]. Figure 4.2(b) shows the partition in the case where the feedback includes two integrators and the quantizer is two-bit. The dotted lines show the cubic lattice structure of the partition when the quantizer is infinite. In general, the cubic lattice structure is maintained whenever the feedback operator \( G \) is limited to pure summations.

![Diagram](https://example.com/diagram.png)

Figure 4.3: Architecture of predictive encoding, as a high order generalization of \( \Delta \) modulation [5].
As in simple encoding, a predictive encoder can be characterized, in the oversampling situation, as a mapping from \( \mathcal{V} \) to \( \mathbb{R}^N \), or equivalently, by the partition of \( \mathcal{V} \) defined from (3.3). It is simply obtained by taking the restriction of \( \mathbb{R}^N \) to \( \mathcal{V} \).

### 4.4 Consistent estimates and non-consistency of linear decoding

We have the same notion of consistent estimates (Definition 3.3.1), and Property 3.3.4 is still valid. Let us show that the linear decoding scheme does not necessarily provide consistent estimates. We recall from Section 1.2.2 that the linear decoding scheme consists in lowpass filtering the signal \( \hat{C} \) obtained from \( C \) through the predecoding shown in Figure 1.6(b). We have the following property:

**Proposition 4.4.1** \( \hat{C} \) is the center of the set \( \tilde{\mathcal{C}}(C) \).

**Proof:** Using mapping notations, we have \( \hat{C} = (I + G)[C] \). Since \( C \) is the center of \( Q^{-1}[C] \), then \( \hat{C} = C + G[C] \) is the center of \( Q^{-1}[C] + G[C] \) equal to \( \hat{C}(C) \). \( \square \)

Therefore, as in simple encoding, linear decoding consists in performing an orthogonal projection of the center of \( \tilde{\mathcal{C}}(C) \) on \( \mathcal{V} \). This is still represented by Figure 3.6, where \( C \) has to be replaced by \( \hat{C} \). Therefore, linear decoding estimates are not necessarily consistent.

### 4.5 Methods of improvement of non-consistent estimates

As in simple encoding, the principle of projection, with the option of alternating projections, can be used to improve non-consistent estimates. The projection on \( \tilde{\mathcal{C}}(C) \) is slightly modified since the new expression of \( \tilde{\mathcal{C}}(C) \) derived from Proposition 4.2.3 is:

\[
\hat{\mathcal{C}}(C) = \left\{ Y \in \mathbb{R}^N \mid \forall k = 1, \ldots, N, \ Y(k) \in q^{-1}[C(k)] + D(k), \ \text{where} \ D = G[C] \right\}.
\]

In this expression, note that \( q^{-1}[C(k)] + D(k) \) is simply the quantization interval \( q^{-1}[C(k)] \) translated by the real value \( D(k) \). This leads to the following algorithm:

**Algorithm 2:**

Step 1: calculate the signal \( D = G[C] \).

Step 2: At every instant \( k \),

(i) if \( \hat{X}(k) \in q^{-1}[C(k)] + D(k) \), take \( \hat{X}'(k) = \hat{X}(k) \),

(ii) else, take \( X'(k) \) equal to the bound of the interval \( q^{-1}[C(k)] + D(k) \) closest to \( \hat{X}(k) \).
We propose an equivalent form to Algorithm 2, which may look more complicated, but whose principle will be used for future generalization to \( \Sigma \Delta \) modulation. This consists in performing a change of variable by taking \( \hat{X} \), the estimate to be projected, as the origin of the space \( \mathbf{R}^N \). Then, looking for the signal \( \hat{X}' \) which is the projection of \( \hat{X} \) on \( C(C) \) amounts to looking for the signal \( Y = \hat{X}' - \hat{X} \) which is the projection of the zero signal on \( C(C) - \hat{X} \). Note that \( C(C) - \hat{X} = Q^{-1}[C] + D \), where \( D = G[C] - \hat{X} \). This leads to the following algorithm:

Algorithm 2:\ '
Step 1: calculate the signal \( D = G[C] - \hat{X} \).
Step 2: At every instant \( k \),
   (i) if \( 0 \in q^{-1}[C(k)] + D(k) \), take \( Y(k) = 0 \),
   (ii) else, take \( Y(k) \) equal to the bound of the interval \( q^{-1}[C(k)] + D(k) \) closest to 0.
Step 3: calculate the signal \( \hat{X}' = Y + \hat{X} \).

The computation of the signal \( Y \) in Step 2 has the simple graphic representation shown in Figure 4.4.

![Diagram](image)

- \( Y(k) \) solution to the constrained minimization problem
- boundaries of the interval \( q^{-1}[C(k)] \) shifted by \( D(k) \)

Figure 4.4: Representation of the solution \( Y \) to Step 2 of Algorithm 2\'.

48
4.6 Numerical evaluation of consistent reconstruction

Experiments similar to the case of simple encoding were performed on a one-bit Δ modulator of quantization step size $q$. Input signals were chosen to have the bandwidth $W = 3$. For each choice of oversampling ratio, their PPA was adjusted so that the maximum variation between two consecutive samples is equal to $\frac{q}{2}$. This ensures that the quantizer does not overload. For the case $W = 3$, it can be shown that the constraint is to take

$$PPA = \frac{3}{2\pi} Rq.$$  \hfill (4.3)

The results are shown in Figure 4.5, in a way similar to the case of simple encoding. They show that, for the fixed quantization step $q$, the MSE decreases by 6 dB/octave in consistent decoding versus 3 dB/octave in linear decoding. In reality,
the PPA is the given parameter, and the quantization step size \( q \) is adjusted to it according to the relation (4.3) for each oversampling ratio \( R \). Therefore, every time \( R \) is multiplied by 2, \( q \) is divided by 2. With this adjustment of \( q \), we find that the MSE decreases by 12 dB/octave in consistent decoding versus 9 dB/octave in linear decoding. In general, we find that the order of the consistent decoding MSE is \( O\left(\frac{q^2}{R^2}\right) \), instead of \( O\left(\frac{q^2}{R}\right) \) in linear decoding.

### 4.7 MSE upper bound in consistent decoding

The derivation of a consistent decoding MSE upper bound in the general case is a difficult problem. In particular, a deterministic derivation of the upper bound would require an explicit description of the feedback behavior.

The goal of this section is to see if the results obtained experimentally at least derived with some conjectured model assumption about the system. This is often the approach in experimental physics, where real systems are too complex to be understood in detail. The goal is more to find some “asymptotic” laws whose consequences fit the experimental results. We propose the derivation of an upper bound to consistent reconstruction MSE of the order \( O\left(\frac{q^2}{R}\right) \) as a consequence of three assumptions. While the first assumption only specifies some conditions of operation, the two other ones constitute conjectured model assumptions about the predictive encoder and the consistent decoder.

**Assumption 4.7.1** The quantizer is uniform and the input signal \( X \) is a random vector of a bounded domain \( \mathcal{D} \) included in the non-overload region of \( \mathcal{V} \).

This assumption can be realized in practice. For example, in single-bit \( \Delta \) modulation, non-overload is ensured if the quantization step size is larger than the maximum variation of the input signal between two sampling instants. Since input signals are bandlimited, a domain \( \mathcal{D} \) can be defined by limiting the maximum amplitude of the signals.

This assumption implies a certain property on the quantization error signal obtained when \( X \in \mathcal{D} \) is input. This signal, denoted by \( E[X] \), is the difference between the input and the output of the quantizer when \( X \) is input. More precisely, \( E[X] = C - A = (Q - I)[A] = (Q - I)F[X] \). The operator \( E \) is called the quantization error operator and is equal to

\[
E = (Q - I)F. \tag{4.4}
\]

When there is no overloading, \( \forall k = 1, \ldots, N, |C(k) - A(k)| \leq \frac{q}{2} \). Therefore,

\[
\forall X \in \mathcal{D}, \quad E[X] \in \left[-\frac{q}{2}, \frac{q}{2}\right]^N. \tag{4.5}
\]
But one can verify that we also have

$$\forall X \in D, \forall \hat{X} \text{ consistent estimate of } X, \quad E[\hat{X}] \in \left[-\frac{q}{2}, \frac{q}{2}\right]^N. \quad (4.6)$$

**Assumption 4.7.2** For each random input signal $X \in D$, the chosen consistent estimate $\hat{X}$ is such that $\hat{X} \in D$ and the direction of the error vector $\hat{X} - X$ in $\mathcal{V}$ is independent of the random vector $X$.

This assumption is difficult to verify for a specific consistent decoding algorithm. We take it as a model assumption.

**Assumption 4.7.3** For the random vector $X \in D$, the quantization error $E[X]$ is a random vector of $[-\frac{q}{2}, \frac{q}{2}]^N$ which verifies the following property:

$$\exists c_0 > 0, \forall N \geq W, \quad \text{Prob} \{E[X] \in \mathcal{R}\} \leq c_0 \cdot \frac{\text{Vol}(\mathcal{R})}{q^N}, \quad (4.7)$$

for any subset $\mathcal{R} \subset [-\frac{q}{2}, \frac{q}{2}]^N$ such that $\mathcal{R} = I_1 \times \cdots \times I_N$ where, for $k = 1, \ldots, N$, $I_k$ is a subinterval of $[-\frac{q}{2}, \frac{q}{2}]$ of the type $[-\frac{q}{2}, a]$ or $[a, \frac{q}{2}]$.

Qualitatively speaking, $c$ is an upper bound to the density of $E[X]$ in $[-\frac{q}{2}, \frac{q}{2}]^N$, independent of $N$, when the density is evaluated in rectangular volumes of the specific type of Assumption 4.7.3. Note that this does not assume that $E[X]$ has a bounded probability distribution in $[-\frac{q}{2}, \frac{q}{2}]^N$ nor that it is uncorrelated with $X$. Again, this assumption is difficult to verify in practice, and we take it as a model assumption.

With these three assumptions, we show that the expectation of $\|\hat{X} - X\|^2$, denoted by $E \left(\|\hat{X} - X\|^2\right)$, has an upper bound of the order of $O \left(\frac{q^2}{R^2}\right)$. The proof is based on the following proposition and lemma:

**Proposition 4.7.4** If $X \in C(C)$, then $E[X] = (I + G)[C] - X$.

Proof: From Figure 1.6(a), $A = X - G[C]$. Therefore, $E[X] = C - A = C - X + G[C] = (I + G)[C] - X \quad \Box$

**Lemma 4.7.5** $\exists c_1 > 0$, for $N$ high enough, $\forall X \in \mathcal{V}$,

$$\sum_{k=1}^{N} |X(k)| \geq c_1 N \|X\|.$$
This is proved in Appendix A.3. This lemma which will be crucial for the derivation of the upper bound is obvious in the particular case where $\mathcal{V}$ is reduced to constant signals. However, as shown in Appendix A.3, the lemma holds for any choice of space $\mathcal{V}$ of bandlimited signals. An MSE upper bound, which confirms the numerical results of the previous section, is then obtained according to the following theorem:

**Theorem 4.7.6** Under Assumptions 4.7.1, 4.7.2 and 4.7.3, $\exists \alpha > 0$, for $R$ high enough,

$$E\left(\|\hat{X} - X\|^2\right) \leq \alpha \frac{q^2}{R^2}$$

Proof: Before starting the proof, let us first define some notations. The set of unitary vectors of $\mathcal{V}$ is denoted by $\mathcal{U}$ and is explicitly defined by $\mathcal{U} = \{U \in \mathcal{V} / \|U\| = 1\}$. When $U \in \mathcal{U}$, $dU$ designates the elementary solid angle centered around $U$. For $\lambda \geq 0$ and $U \in \mathcal{U}$, we write

$$Prob\left\{\hat{X} - X \in dU \text{ and } \lambda \leq \|\hat{X} - X\|\right\} = P(dU, \lambda). \quad (4.8)$$

Consistent estimates verifying this probability are illustrated in Figure 4.6. Then,

$$Prob\left\{\hat{X} - X \in dU \text{ and } \lambda \leq \|\hat{X} - X\| \leq \lambda + d\lambda\right\} = \frac{P(dU, \lambda)}{d\lambda} d\lambda. \quad (4.9)$$

Let us start the proof. We have

$$E\left(\|\hat{X} - X\|^2\right) = \int_{\lambda=0}^{+\infty} \lambda^2 P\left\{\lambda \leq \|\hat{X} - X\| \leq \lambda + d\lambda\right\}. \quad (4.10)$$

Using (4.9), we have

$$Prob\left\{\lambda \leq \|\hat{X} - X\| \leq \lambda + d\lambda\right\} = \int_{U \in \mathcal{U}} Prob\left\{\hat{X} - X \in dU \text{ and } \lambda \leq \|\hat{X} - X\| \leq \lambda + d\lambda\right\}$$

$$= \int_{U \in \mathcal{U}} \frac{P(dU, \lambda)}{d\lambda} d\lambda.$$  

Coming back to (4.9) and performing an integration by part, we find

$$E\left(\|\hat{X} - X\|^2\right) = \int_{U \in \mathcal{U}} \int_{\lambda=0}^{+\infty} \lambda^2 \left(\frac{P(dU, \lambda)}{d\lambda}\right) d\lambda = 2 \int_{U \in \mathcal{U}} \int_{\lambda=0}^{+\infty} \lambda P(dU, \lambda) d\lambda. \quad (4.11)$$

We have used the fact that $P(dU, +\infty) = 0$ since $\|\hat{X} - X\|$ is necessarily bounded. Let us find an upper bound to $P(dU, \lambda)$. When $(X, \hat{X})$ is such that $\hat{X} - X \in dU$ and $\lambda \leq \|\hat{X} - X\|$ we have the following properties:

(i) $\frac{\hat{X} - X}{\|\hat{X} - X\|} \simeq U$. 

52
Figure 4.6: Representation of the different elements involved in the demonstration of Theorem 4.7.6.
(ii) \( \hat{X} + \frac{\lambda}{\|\hat{X} - X\|} (\hat{X} - X) \) is a consistent estimate of \( X \).

Property (i) comes from the fact that \( \hat{X} - X \in dU \). Property (ii) uses the fact that the set of consistent estimates is convex (see Figure 4.6). Then, for any \( \mu \in [0,1] \), \( X + \mu(\hat{X} - X) = (1 - \mu)X + \mu\hat{X} \) is a consistent estimate of \( X \). Property (ii) is obtained for the particular choice \( \mu = \frac{\lambda}{\|\hat{X} - X\|} \).

Then, from properties (i) and (ii), we derive that \( X + \lambda U \) is a consistent estimate \( X \). Since \( X \) and \( X + \lambda U \) belong to a common set \( \mathcal{C}(C) \), it is easy to see from Proposition (4.7.4) that

\[
E[X] = E[X + \lambda U] + \lambda U.
\] (4.12)

Therefore, applying (4.6) on \( X + \lambda U \), we find that \( E[X] \in [-\frac{\lambda}{2}, \frac{\lambda}{2}]^N + \lambda U \). Since we also have \( E[X + \lambda U] \in [-\frac{\lambda}{2}, \frac{\lambda}{2}]^N \), then \( E[X] \in \mathcal{R}_U \) where, in general, \( \mathcal{R}_W \) is defined by \( \mathcal{R}_W = [-\frac{\lambda}{2}, \frac{\lambda}{2}]^N \cap \left( [-\frac{\lambda}{2}, \frac{\lambda}{2}]^N + W \right) \). We have just shown that

\( \hat{X} - X \in dU \) and \( \lambda \leq \|\hat{X} - X\| \implies \hat{X} - X \in dU \) and \( E[X] \in \mathcal{R}_U \).

Therefore,

\[
P(dU, \lambda) \leq \text{Prob}\{\hat{X} - X \in dU \text{ and } E[X] \in \mathcal{R}_U\}.
\]

Using Assumption 4.7.2, we have

\[
P(dU, \lambda) \leq \text{Prob}\{\hat{X} - X \in dU\} \cdot \text{Prob}\{E[X] \in \mathcal{R}_U\}.
\] (4.13)

Let us assume that \( \max_{1 \leq k \leq N} \lambda|U(k)| \leq q \). One can see that \( \mathcal{R}_U = I_1 \times \cdots \times I_N \) where for \( k = 1, \ldots, N \), \( I_k = [-\frac{\lambda}{2}, \frac{\lambda}{2}] \cap \left( [-\frac{\lambda}{2}, \frac{\lambda}{2}]^N + \lambda U(k) \right) \). Therefore \( \mathcal{R}_U \) is of the type of Assumption 4.7.3. Since the length of \( I_k \) is \( q - \lambda|U(k)| \), the volume of \( \mathcal{R}_U \) is \( \prod_{k=1}^N (q - \lambda U(k)) \). Applying Assumption 4.7.3, we find

\[
\text{Prob}\{E[X] \in \mathcal{R}_U\} \leq c_0 \cdot \prod_{k=1}^N \left( 1 - \frac{\lambda}{q}|U(k)| \right).
\] (4.14)

Using the inequality \( 1 - a \leq e^{-a} \), we find

\[
\text{Prob}\{E[X] \in \mathcal{R}_U\} \leq c_0 \cdot \exp \left( -\frac{\lambda}{q} \sum_{k=1}^N |U(k)| \right).
\]

Then applying Lemma 4.7.5, we find

\[
\text{Prob}\{E[X] \in \mathcal{R}_U\} \leq c_0 \cdot e^{-\lambda q_N}.
\] (4.15)

54
Gathering (4.11), (4.13) and (4.15), we find
\[
E \left( \| \hat{X} - X \|^2 \right) \leq 2 \int_{U \in \mathcal{U}} \text{Prob} \left\{ \hat{X} - X \in dU \right\} \cdot \int_{\lambda=0}^{+\infty} \lambda e^{-\frac{1}{\lambda^2} N} d\lambda. \tag{4.16}
\]

The first integral is equal to 1. By an integration by part, the second integral gives \(\frac{c_0 q^2}{c_1 N^2}\). Using the fact that \(R = \frac{N}{W}\), (4.16) leads to \(E \left( \| \hat{X} - X \|^2 \right) \leq \alpha \frac{q^2}{N^2}\) where \(\alpha = \frac{2c_0}{c_1 W^2} > 0\) \(\square\)

### 4.8 Conclusion

In this chapter, we have shown that the deterministic approach of simple encoding is entirely applicable to predictive encoding. The constant idea is to describe the encoder as a mapping of \(\mathbb{R}^N\). Although a predictive encoder includes a feedback loop, we show how the global mapping can be determined. As a result, we see that the set \(\mathcal{C}(C)\), image of \(C\) through the inverse of this mapping, is still a rectangular hyper-parallelepiped. Every notion presented in simple encoding, such as estimate consistency, the partitioning interpretation, the linear decoding non-consistency, the methods of improvements of non-consistent estimates and the alternating projection algorithm are applicable in predictive encoding. In the case of \(\Delta\) modulation, the consistent estimates, approached by alternating projections, yield an MSE in \(O\left(\frac{q^2}{R^2}\right)\) instead of \(O\left(\frac{q^2}{R}\right)\) in linear decoding (where \(q\) is the quantization step size). For a fixed step size \(q\), this confirms again the MSE improvement of 3 dB per octave of oversampling. Contrary to simple encoding, this result is obtained without any particular condition on the input signals. On the other hand, the analytical justification of this result is more difficult. However, the choice of two particular model assumptions (Assumptions 4.7.2 and 4.7.3) leads to the derivation of the result \(O\left(\frac{q^2}{R^2}\right)\) as an upper bound to consistent reconstruction MSE.

The deterministic analysis of predictive encoding will also serve as an intermediate stage towards the study of noise-shaping encoders in the next chapter.
Chapter 5

Single-loop and multi-loop \( \Sigma \Delta \) modulation

5.1 Structure of multi-loop \( \Sigma \Delta \) modulation

The structures of single-loop and multi-loop \( \Sigma \Delta \) modulators where presented in Sections 1.2.3 and 1.2.4 and shown in Figures 1.9(b) and 1.12(a). We recall that these encoders belong to the family of noise-shaping encoders whose block diagram is presented in Figure 1.13. In the deterministic approach, \( H \) and \( G \) are considered as mappings of \( \mathbb{R}^N \). In mapping notations, the constraint (1.5) between \( H \) and \( G \) becomes:

\[
H = I + G,
\]

(5.1)

where \( I \) is the identity mapping. Let us consider these two mappings in the case of a multi-loop \( \Sigma \Delta \) modulator. According to Figure 1.12(b), the mapping \( H \) is an \( n^{th} \) order discrete-time integrator. Using the notation (2.6) of Section 2.2, \( H \) is such that \( H[X] = X^{(-n)} \) (see notations of Section 2.2). For the mapping \( G \) to be completely defined, the initial conditions \( A_0^{(0)}, ..., A_0^{(n-1)}, C_0 \) need to be known. In most of this chapter, we will assume that these initial conditions are zero. We will therefore consider the block diagram shown in Figure 5.1. The case of non-

![Block diagram of a multi-loop \( \Sigma \Delta \) modulator with zero initial conditions.](image)

Figure 5.1: Block diagrams of a multi-loop \( \Sigma \Delta \) modulator with zero initial conditions.

56
5.2 Deterministic description of multi-loop $\Sigma\Delta$ modulation encoding

Using the properties of predictive encoders derived in Chapter 4, we immediately have the following propositions.

Proposition 5.2.1 The noise-shaping encoder of Figure 1.13 has the structure of Figure 4.1 where the preprocessing one-to-one mapping is

$$F = (I + GQ)^{-1} H.$$  \hfill (5.2)

Proposition 5.2.2 The set of signals producing an encoded signal $C$ through the noise-shaping encoder of Figure 1.13 is

$$C(C) = H^{-1} \left[ Q^{-1}[C] + G[C] \right].$$  \hfill (5.3)

Since $Q^{-1}[C] + G[C]$ is in general a rectangular hyper-parallelepiped of $\mathbb{R}^N$, and $H^{-1}$ a linear mapping, then $C(C)$, although not necessarily rectangular, is still a hyper-parallelepiped in $\mathbb{R}^N$. Proposition 3.1.1 relative to the encoded information in the oversampling situation is still applicable, but the representation of this information is modified as shown in Figure 5.2. The deformation of $C(C)$ is due to the mapping $H^{-1}$.

5.3 The partitioning approach to noise-shaping encoding

As in simple and predictive encoding, a noise-shaping encoder can be equivalently characterized by the partition it induces on $\mathbb{R}^N$ (definition (3.1) of Section 3.2). This partition is derived in the following way. Using Figure 1.13, one can first derive the partition induced by the predictive encoder included in the block diagram. We saw in Section 4.3 how this can be done. Then, by applying the linear transform $H^{-1}$, one obtains the partition of the whole noise-shaping encoder.

Let us study the case of a multi-loop $\Sigma\Delta$ modulator. As we already mentioned, the predictive encoder contained in this type of encoder has the structure Figure 4.3. According to Section 4.3, predictive encoders of this type induce a partition which forms a cubic lattice or is an an extraction of a cubic-lattice. This lattice is generated by the canonical basis $(E_i)_{i=1,...,N}$ defined in (3.2). By applying the
Figure 5.2: Geometric representation of the information “$X \in \mathcal{C}(\mathcal{C}) \cap \mathcal{V}$” and non-consistency of the linear decoding estimate $P_{\mathcal{V}}[\mathcal{C}]$ in $\Sigma\Delta$ modulation.
linear operator $H^{-1}$, we obtain a new lattice generated by the basis $(F_i)_{i=1,...,N}$ defined by

$$\forall i = 1,...,N, \quad F_i = H^{-1}[E_i]. \quad (5.4)$$

We show in Figure 5.3 the partition induced by a single-loop and a double-loop

![Figure 5.3: Partition induced on $R^N$ ($N = 2$) by 2-bit $\Sigma\Delta$ modulators. (a) Single-loop configuration. (b) Double-loop configuration. The shaded lines represent the non-overload region. The dotted lines represent the case of infinite quantizer.](image)

$\Sigma\Delta$ modulator using a two-bit quantizer. The inherent lattice structure is also represented (dotted lines) with the basis $(F_i)_{i=1,...,N}$. These two partitions are simply obtained by applying the operator $H^{-1}$ on the two partitions of Figure 4.2, where $H^{-1}$ is a first order and a second order integrator for the single-loop and the double-loop configurations respectively. For example, consider a double-loop modulator. The built-in predictive encoder has the structure of Figure 4.3 with two integrators. We showed its partition in Figure 4.2(b). The partition of the double-loop modulator is then obtained by applying the linear operator $H^{-1}$ which is a $2^{nd}$ order discrete-time integrator and we obtain the partition of Figure 5.3. The dotted lines show the lattice structure of this partition in the case of a uniform and infinite quantizer. The solid lines correspond to the case of a two-bit quantizer and show that the partition is an extraction of the lattice.
Also, in the situation of oversampling, a noise-shaping encoder can be equivalently characterized by the partition it induces in $\mathcal{V}$ (definition (3.3) of Section 3.2), as in simple and predictive encoding. Figure 5.4 shows various examples of partitions with the single and the double-loop configurations and various oversampling ratios, when $\mathcal{V}$ is the two-dimensional space of sinusoids of arbitrary phase and period 1. Again, these partitions do no longer have a lattice structure. However, one should keep in mind that they are obtained by intersecting a lattice (or an extraction of it) with $\mathcal{V}$.

### 5.4 Consistent estimates and non-consistency of linear decoding

We again use the notion of consistent estimates with its associated property (Property 3.3.4), and show that linear decoding is not necessarily consistent. We recall that, in noise-shaping encoding, linear decoding consists in lowpass filtering the encoded signal $C$. Once again, one would guess that $C$ is the center of $\mathcal{C}(C)$. We have indeed the following proposition:

**Proposition 5.4.1** In noise-shaping encoding, $C$ is the center of $\mathcal{C}(C)$.

**Proof:** An element $C'$ is the center of the parallelepiped $H^{-1}[Q^{-1}[C] + G[C]]$ if and only if $H[C']$ is the center of the parallelepiped $Q^{-1}[C] + G[C]$. This is because $H$ is linear. But the center of $Q^{-1}[C] + G[C]$ is simply $C + G[C]$. Then using equation (5.1) we have $C' = H^{-1}[C + G[C]] = H^{-1}(I + G)C = H^{-1}H(C) = C \, \square$

Then, as in simple and predictive encoding, linear decoding consists in performing the orthogonal projection of the center of $\mathcal{C}(C)$ on $\mathcal{V}$. Therefore, as can be seen in Figure 5.2, linear decoding estimates are not necessarily consistent.

### 5.5 Methods of improvement of non-consistent estimates

We first describe a general scheme for the design of algorithms for the projection on $\mathcal{C}(C)$. Then, we will separately consider the case of single-loop modulation and the case of higher order modulation. The formulation of the scheme starts with the same idea of change of variable proposed in Algorithm 2' for predictive encoding which will prove to be particularly convenient in the present case. The difference here is that we include the invertible mapping $H$ in the change of variable. From Lemma 3.3.2, the projection of $\hat{X}$ on $\mathcal{C}(C)$ is the signal $\hat{X}'$ which minimizes $\|\hat{X}' - \hat{X}\|$ subject to the constraint $\hat{X}' \in \mathcal{C}(C)$. By the change of variable $Y = H[\hat{X}' - \hat{X}]$, this amounts to finding the signal $Y$ which minimizes the functional $\|H^{-1}[Y]\|$ subject to the constraint $Y \in H[\mathcal{C}(C) - \hat{X}]$. Using the expression of
Figure 5.4: Examples of partitions of the space $\mathcal{V}$ at the oversampling ratio $R$ with single-bit $\Sigma\Delta$ modulators, in the case where $\mathcal{V}$ is of 2 dimensional space of zero-dc component sinusoids. The input signal region of operation is represented by the circle in dotted line.
\(C(C)\) from Proposition 5.2.2, we simply have \(H \left[ C(C) - \hat{X} \right] = Q^{-1}[C] + D\), where 
\(D = G[C] - H[\hat{X}]\). Also, minimizing \(\|H^{-1}[Y]\|\) is minimizing \(\phi(Y)\) where,
\[
\phi(Y) = \frac{1}{2} \|H^{-1}[Y]\|^2.
\]

The projection algorithm is therefore:

**Algorithm 3:**

Step 1: calculate the signal \(D = G[C] - H[\hat{X}]\).

Step 2: Find the minimum \(Y\) of \(\phi\) subject to \(Y \in Q^{-1}[C] + D\).

Step 3: calculate the signal \(\hat{X}' = H^{-1}[Y] + \hat{X}\).

Step 3 is essentially the computation of an \(n^{th}\) order discrete derivative. Step 2 is a problem of minimization of a quadratic functional under convex constraints. From the theory of convex analysis [37], there exists a characterization for the minimum in such a problem. Since the gradient of \(\phi\) is involved in this characterization, we introduce the following notation.

**Notation 5.5.1** \(\nabla_\phi\) denotes the mapping of \(R^N\) such that, for any \(Y \in R^N\), \nabla_\phi[Y](k) is the partial derivative of \(\phi(Y)\) with respect to \(Y(k)\), or \nabla_\phi[Y](k) = \frac{\partial \phi(Y)}{\partial Y(k)}.

Then, the characterization of the minimum is given by the following lemma:

**Lemma 5.5.2** [37] A quadratic functional \(\phi\) has a unique minimum in a convex set \(S\). It is the signal \(Y\) such that
\[
Y \in S \quad \text{and} \quad \forall Y' \in S, \left\langle \nabla_\phi[Y], Y' - Y \right\rangle \geq 0. \quad (5.5)
\]

### 5.5.1 Case of single-loop \(\Sigma\Delta\) modulation

In single-loop \(\Sigma\Delta\) modulation, \(H\) is a first order integrator. In this case, the expression of \(\nabla_\phi\) is given by the following property:

**Property 5.5.3** For any \(Y \in R^N\),
\[
\forall k = 1, \ldots, N, \quad \nabla_\phi[Y](k) = - \left\{ [Y(k + 1) - Y(k)] - [Y(k) - Y(k - 1)] \right\}, \quad (5.6)
\]

with the convention \(Y(0) = Y(N + 1) = 0\).

Proof: Since \(H^{-1}\) is the first order discrete-time differentiator, then \(\phi(Y) = \frac{1}{2} \|H^{-1}[Y]\|^2 = \frac{1}{2} \sum_{j=1}^{N} |Y(j) - Y(j - 1)|^2\). For \(k = 2, \ldots, N - 1\),
\[
\nabla_\phi[Y](k) = \frac{\partial \phi(Y)}{\partial Y(k)} = [Y(k) - Y(k - 1)] - [Y(k + 1) - Y(k)].
\]

62
One can check that $\nabla\phi[Y](1) = Y(1) - [Y(2) - Y(1)]$ and $\nabla\phi[Y](N) = [Y(N) - Y(N - 1)] - Y(N)$. All these cases are summarized in (5.6) \( \square \)

Then, $\nabla\phi[Y](k)$ is minus the change of slope of $Y$ about time index $k$. Applying Lemma 5.5.2 on $S = Q^{-1}[C] + D$ and using this property, we derive an algorithm for the computation of Step 2 of Algorithm 3. This algorithm was first introduced in [21]. We present it in a “physical” manner.

“Thread algorithm”:

(i) Represent graphically in the time domain the set of constraint $Q^{-1}[C] + D$ as sequence of the quantization intervals $q^{-1}[C(k)]$ translated by $D(k)$ (see Figure 5.5).

(ii) Attach a “thread” at node $(k = 0, Y(0) = 0)$, pull it taut with a horizontal force between the constraints defined by these intervals (arrows in Figure 5.5).

(iii) For $k = 1, \ldots, N$, take $Y(k), k \geq 1$ on the path of the resulting thread position (see Figure 5.5).

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{thread_algorithm}
\caption{Representation of the solution $Y$ to the “Thread algorithm”.

Let us show that the signal $Y$ thus obtained is the minimum of $\phi$ subject to $Q^{-1}[C] + D$. It is sufficient to prove that $Y$ satisfies the criterion (5.5). First, it is
obvious that $Y \in Q^{-1}[C] + D$. Then, whenever $\nabla_\theta [Y](k) < 0$, or a change of slope occurs about time $k$, the thread touches a constraint (in Figure 5.5, at $k = 5, 9, 12, 14$). But, the change of slope is always opposite to the direction of the constraint. For example, in Figure 5.5, the change of slope is negative only when the arrow of contact goes up ($k = 5, 9$) and vice-versa ($k = 12$). When an arrow of contact goes up, for example at $k = 5$, for any other admissible signal $Y' \in Q^{-1}[C] + D$, then we necessarily have $Y'(k) - Y(k) > 0$. Similarly, if an arrow of contact goes down at $k = 12$, then $Y'(k) - Y(k) < 0$. Since $\nabla_\theta [Y]$ gives the opposite of the slope’s variation, we have just shown that whenever $\nabla_\theta [Y](k) > 0$, then $\nabla_\theta [Y](k)$ and $Y'(k) - Y(k)$ have the same sign. This implies (5.5).

While presented in a physical manner, this algorithm can be easily translated in terms of computer operations.

5.5.2 Case of n-loop $\Sigma \Delta$ modulation with $n \geq 2$

The projection on $\mathcal{C}(C)$ in the general case of order $n \geq 2$ is quite involved. In this section we will only sketch the main ideas used for the design of the algorithm. More detailed explanations can be found in Appendix A.5.

With the single-loop configuration, the “thread algorithm” gave the signal $Y$ which meets the constraints and at the same time minimizes the mean squared “slope”. For an $n$-loop configuration in general, the functional $\phi$ to minimize is the $n^{th}$ order discrete derivative. For example, when $n = 2$, we will have to minimize the “curvature” of the signal. This is similar to the problem of constrained thin flexible beam in the field of mechanics. Figure 5.6 shows the solution minimizing the “curvature” under the same constraints as in figure 5.5.

Unfortunately, in the $n^{th}$ order case, performing the direct projection on $\mathcal{C}(C)$ in one step as in the single-loop case is not possible. Iterative algorithms from the theory of non-linear programming [38] could be used. But the aim is to find a single transformation based on the knowledge of $C$ which leads to a necessary improvement of the estimate. This would give us the freedom to use it once, or to alternate it with the projection on $\mathcal{Y}$.

The algorithm we designed for the $n^{th}$ order case is able to perform a projection on a convex set which does not contain the estimate $\hat{X}$, always includes $\mathcal{C}(C)$ and therefore contains $X$. Projecting $\hat{X}$ on such a convex set necessarily insures the improvement of $\hat{X}$ as estimate of $X$ according to Lemma 3.3.3. Instead of projecting $\hat{X}$ on $\mathcal{C}(C) = H^{-1} [Q^{-1}[C] + G[C]]$, the idea is to project $\hat{X}$ on the larger set $C_K(C) = H^{-1} \left[ Q^{-1} [S_K(C)] + G[C] \right]$, where $S_K(C)$ is the set of signals which coincide with $C$ on a certain subset of time indices $K \subset \{1, ..., N\}$. More precisely,

$$S_K(C) = \left\{ C' \in \mathbb{R}^N \; / \; \forall k \in K, \; C'(k) = C(k) \right\},$$
\* Y(k) solution to the constrained minimization problem

| boundaries of the interval q^{-1}[C(k)] shifted by D(k) |

| thin beam path |

Figure 5.6: Representation of the solution \( Y \) to Step 2 of Algorithm 3 for 2\(^{nd} \) order \( \Sigma\Delta \) modulation.
It is easy to show that $\mathcal{C}_K(C)$ is a convex set which includes $\mathcal{C}(C)$. This is true for any choice of $K \subseteq \{1, \ldots, N\}$. For a given $K$, Algorithm 3 is modified as follows:

**Algorithm 3':**
Step 1: calculate the signal $D = G[C] - H[\hat{X}]$.
Step 2: Find the minimum $Y$ of $\phi$ subject to $Y \in Q^{-1}[\mathcal{S}_K(C)] + D$.
Step 3: calculate the signal $\hat{X}' = H^{-1}[Y] + \hat{X}$.

In Appendix A.5, we design an algorithm which solves Step 2, where the subset $K$ is progressively constructed in time and such that $\hat{X} \notin \mathcal{C}_K(C)$. The improvement achieved by this algorithm has been numerically evaluated. The results are presented in the next section.

### 5.6 Numerical evaluation of consistent reconstruction

Experiments similar to the case of simple encoding were performed on single-bit single-loop, two-bit double-loop and three-bit triple-loop $\Sigma\Delta$ modulation. The quantizers were uniform with step size $q$. Input signals were chosen to have $2M + 1 = 7$ as bandwidth and were constrained to have $\frac{q}{2}$ as maximum amplitude. The oversampling ratio $R$ was chosen between 18 and 585 approximately. For each experiment, the decoding MSE was averaged over 600 randomly generated input signals. In Figure 5.7(a), we show the comparison between the linear decoding and alternating projection schemes. The projection iteration was stopped as soon as the MSE decrement per step was less than 0.1dB. This figure shows that the linear decoding scheme yields an MSE slope of $-(2n+1) \times 3$ dB/octave and that the alternating projection scheme increasingly improves this MSE reduction with the oversampling ratio, regardless of the encoder’s order.

Figures 5.7(b) to (d) show this improvement separately for the three types of encoders. To emphasize the MSE slope tendency, we plotted the difference (normalized MSE) between the measured MSE and the theoretical linear decoding MSE given from [5] by:

$$MSE(\text{linear decoding}) = \frac{\pi^{2n}}{2n+1} \cdot \frac{\sigma_q^2}{R^{2n+1}}, \quad \text{where } \sigma_q^2 = \frac{q^2}{12}. \quad (5.7)$$

One can see that when $R$ is high enough, the MSE obtained from the alternating projection scheme decreases with $R$ faster than the linear decoding MSE by approximately 3 dB/octave. This means that the global asymptotic MSE slope achieved by the alternating projection scheme is $-(2n+2) \times 3$ dB/octave. This confirms the improvement of the asymptotic behavior from $O(R^{-(2n+1)})$ to $O(R^{-(2n+2)})$ for consistent reconstruction, discussed in the previous section. Figure 5.7(b) to (d)
Figure 5.7: Dependence of the decoding MSE with the oversampling ratio for single-bit single-loop, two-bit double-loop and three-bit triple-loop ΣΔ modulation. (a) Comparison between linear decoding and consistent decoding approached by alternating projections. (b)-(c)-(d) Decoding MSE normalized with the theoretical noise shaping equation (5.7) for linear decoding, consistent decoding approached by alternating projections and finite step alternating projections: (b) single-bit single-loop configuration; (c) two-bit double-loop configuration; (d) three-bit triple-loop configuration.
also show that a substantial fraction of the total improvement obtained by “infinite” alternating projections is achieved with very limited numbers of iteration (one to three steps).

5.7 MSE upper bound in consistent decoding

As in predictive encoding, we derive a consistent MSE upper bound, starting from the same assumptions as that of Section 4.7 (Assumptions 4.7.1, 4.7.2 and 4.7.3). Again, Assumption 4.7.1 can be realized in practice. For example, in single-bit single-loop ΣΔ modulation, a signal \( X \in [-\frac{q}{2}, \frac{q}{2}]^N \) will never overload the quantizer, as mentioned in the introduction (Section 1.2.3). This can also be shown in the case of \( n \)-bit \( n \)-loop ΣΔ modulation for the same domain of inputs (see Appendix A.9). Under this assumption, properties (4.4), (4.5) and (4.6) are still true. Assumptions 4.7.2 and 4.7.3 are still considered as model assumptions. Proposition 4.7.4 and Lemma 4.7.5 have a new form.

**Proposition 5.7.1** If \( X \in C(C) \), then \( \mathbf{E}[X] = H[C - X] \).

Proof: From Figure 1.13, \( A = H[X] - G[C] \). Therefore, \( \mathbf{E}[X] = C - A = C - H[X] + G[C] = (I + G)[C] - H[X] = H[C - X] \). \( \square \)

**Lemma 5.7.2** \( \exists c_1 > 0 \), for \( N \) high enough, \( \forall X \in \mathcal{V} \),

\[
\sum_{k=1}^{N} |H[X](k)| \geq c_1 N^{n+1} \|X\|.
\]

This is proved in Appendix A.4. An MSE upper bound is then obtained according to the following theorem:

**Theorem 5.7.3** Under Assumptions 4.7.1, 4.7.2 and 4.7.3 (in Section 4.7), \( \exists \alpha > 0 \), for \( R \) high enough,

\[
\mathbf{E} \left( \| \hat{X} - X \|^2 \right) \leq \alpha \frac{q^2}{R^{2n+2}}.
\]  

(5.8)

Outline of the proof: This is similar to that of Theorem 4.7.6. Using the same notations, the equation

\[
\mathbf{E} \left( \| \hat{X} - X \|^2 \right) = 2 \int_{U \in \mathcal{U}} \int_{\lambda = 0}^{+\infty} \lambda P(dU, \lambda) d\lambda
\]  

(5.9)

obtained from (4.11) is still valid, where \( P(dU, \lambda) \) is defined in (4.8). Similarly, considering a couple \((X, \hat{X})\) such that \( \hat{X} - X \in dU \) and \( \lambda \leq \|X - X\| \), then \( X + \lambda U \) is a consistent estimate of \( X \). Using Proposition 5.7.1, we find that

\[
\mathbf{E}[X] = \mathbf{E}[X + \lambda U] + \lambda \mathbf{H}[U].
\]
This is similar to (4.12) where \( U \) is replaced by \( V = H[U] \). This leads to
\[
P(dU, \lambda) \leq \text{Prob}\{ \hat{X} - X \in dU \} \cdot \text{Prob}\{ E[X] \in \mathcal{R}_{AV} \},
\]
where the second factor can be upper bounded as
\[
\text{Prob}\{ E[X] \in \mathcal{R}_{AV} \} \leq c_0 \cdot \exp\left( -\frac{\lambda}{q} \sum_{k=1}^{N} |V(k)| \right).
\]
in a way similar to (4.14). Then, applying Lemma 5.7.2, we find
\[
\text{Prob}\{ E[X] \in \mathcal{R}_{AV} \} \leq c_0 \cdot \exp\left( -\frac{\lambda}{q} \right).
\]
Finally, gathering (5.9), (5.10) and (5.12) and performing an integration by part, we find
\[
E\left( \| \hat{X} - X \|^2 \right) \leq \alpha \frac{q^2}{N^{2n+2}} \text{ where } \alpha = \frac{2c_0}{c_1 W^{2N+2}} > 0 \quad \square
\]

5.8 Lower bound on optimal reconstruction with constant inputs

Under Assumptions 4.7.1, 4.7.2 and 4.7.3, we have shown that consistent reconstruction MSE has an asymptotic behavior of the order of or better than \( O(R^{-2n+2}) \). The numerical results of Section 5.6 show that the averaged MSE achieved by alternating projections is of the order of \( O(R^{-2n+2}) \). However, alternating projections only tend to a consistent estimate which belongs to the border of \( C(C) \cap \mathcal{V} \). One wonders whether this performance could be improved by taking as consistent estimate the center of gravity of \( C(C) \cap \mathcal{V} \). Assuming that input signals are uniformly distributed in the domain of operation \( D \subset \mathcal{V} \), this would represent the optimal reconstruction, since the expectation of the MSE would be minimized.

In this section, we show in the case where \( \mathcal{V} \) is reduced to constant input signals, that the optimal reconstruction MSE is lower bounded by \( O(R^{-2n+2}) \) for \( n \)th order multi-loop \( \Sigma \Delta \) modulators with uniform quantization. In this case \( \mathcal{V} \) is a one dimensional space \( (W = 1) \). Therefore the oversampling ratio \( R \) coincides with the total number of samples \( N \). More important, the partition \( \{ C(C) \cap \mathcal{V} \} / C \) defines a subdivision of the dc component axis into intervals. In Figure 5.8, we show the partition defined by a single-loop \( \Sigma \Delta \) modulator in the two dimensional space of zero-phase sinusoids with dc component. The partition of the space of constant input signals can be simply derived by taking the restriction of this two-dimensional partition to the dc component axis. The one-dimensional interval subdivision appears naturally. Optimal reconstruction consists in identifying in what interval of this subdivision the input signal belongs to, and providing as
Figure 5.8: Example of partition of the space $\mathcal{V}$ at the oversampling ratio $R = 4$, in the case where $\mathcal{V}$ is of 2 dimensional space of zero-phase sinusoids. The partition of the one dimensional space of constant signals is simply the restriction of the 2 dimensional partition to the dc component axis (dashed line).
estimate the midpoint of this interval. This description of optimal reconstruction of constant inputs was previously given in [18, 20].

Assuming that the constant input amplitudes are uniformly distributed in a certain segment \( \mathcal{D} \) of \( \mathcal{V} \), it was shown in [18] (Theorem 1) that the optimal reconstruction MSE expectation is necessarily lower bounded by

\[
MSE_{\text{min}} = \frac{d^2}{12N_d^2},
\]

(5.13)

where \( d \) is the length of the segment \( \mathcal{D} \) and \( N_d \) is the total number of subdivision intervals intersected by \( \mathcal{D} \). By finding an upper bound to \( N_d \), this result was used in [18] to derive a lower bound to \( MSE_{\text{min}} \) in single-loop \( \Sigma\Delta \) modulation. We use the same approach here to derive a lower bound for any \( n^{th} \) order multi-loop modulator.

The basic idea of the derivation is to see that \( N_d \) is also the number of cells of the partition defined by the encoder on \( \mathbb{R}^N \) that are intersected by \( \mathcal{D} \). This number is maximized when the quantizer is infinite. Since we are looking for an upper bound to \( N_d \), we can assume this is the case. We saw in Section 5.3 that the partition is a lattice generated by the basis \( (F_i)_{i=1,\ldots,N} \) defined in (5.4). Figure 5.9 shows a piece of this lattice extracted from the partition defined by a double-loop \( \Sigma\Delta \) modulator shown in Figure 5.3(b). Because of the lattice structure, there is an immediate way to derive an upper bound to \( N_d \). Any signal \( X \) of \( \mathcal{V} \) can be decomposed on the basis as:

\[
X = \alpha_1 F_1 + \cdots + \alpha_N F_N.
\]

This decomposition is represented in Figure 5.9 in two dimensions. Working on this figure, let us start with \( X = O \) and move it along the \( \mathcal{V} \) axis. One can see that as soon as \( X \) enters a new cell, the integer part of one of its coordinates \( (\alpha_1, \ldots, \alpha_N) \) increases by 1. As a consequence, when we stop at some signal \( X_0 = \alpha_{0,1} F_1 + \cdots + \alpha_{0,N} F_N \), the number of cells crossed cannot be larger than \( \alpha_{0,1} + \cdots + \alpha_{0,N} + 1 \). Since the length of the segment \( \mathcal{D} \) is \( d \), to derive an upper bound of \( N_d \), it is sufficient to find a decomposition of the signal \( X_d \in \mathcal{V} \) which is at the distance \( d \) from \( O \). Let

\[
X_d = \sum_{i=1}^{N} \alpha_{d,i} \cdot F_i
\]

(5.14)

be the decomposition of \( X_d \) in \( (F_i)_{i=1,\ldots,N} \). Then, we have

\[
N_d \leq 1 + \sum_{i=1}^{N} \alpha_{d,i}.
\]

(5.15)

Let us now derive the coefficients \( (\alpha_{d,i})_{i=1,\ldots,N} \). By definition, \( X_d \) is the constant sequence such that \( \|X_d\| = d \). This implies that

\[
\forall k = 1, \ldots, N, \quad X_d(k) = d.
\]

(5.16)
Figure 5.9: Relationship between the number of crossed cells and the coefficients $(\alpha_i)_{i=1,\ldots,N}$ of decomposition in the lattice basis $(F_i)_{i=1,\ldots,N}$. 
Therefore, applying the linear mapping \( H \) on equation (5.14) and using (5.4), we find:

\[
H[X_d] = \sum_{i=1}^{N} \alpha_{d,i} \cdot H[F_i] = \sum_{i=1}^{N} \alpha_{d,i} \cdot E_i.
\]

From the definition (3.2) of the basis \( (E_i)_{i=1,...,N} \), we derive that

\[
\forall k = 1, ..., N, \quad H[X_d](k) = \alpha_{d,k}.
\]  

But \( H[X_d] \) is the \( n^{th} \) order discrete-time integration of the constant signal \( X_d \) defined by (5.16). When \( n = 1 \), we simply have \( H[X_d](k) = d \cdot k \). For higher order \( n \), \( H[X_d](k) \) is of the order of \( d \cdot \frac{k^n}{n!} \). The exact expression is:

\[
\forall k = 1, ..., N, \quad H[X_d](k) = d \cdot S_n(k),
\]

where \( S_n(k) = \frac{1}{n!}k(k+1)...(k+n-1) \) as defined in (A.25) (Appendix A.5.4). Relation (5.18) is a consequence of Proposition A.8.1. Therefore, from the relations (5.15), (5.17) and (5.18), we conclude that

\[
N_d \leq 1 + d \cdot \sum_{i=1}^{N} S_n(i)
\]

\[
\leq 1 + d \cdot S_{n+1}(N),
\]

using Proposition A.8.1 a second time. Now, using the explicit expression of \( S_n(k) \), we have

\[
N_d \leq 1 + \frac{d}{(n+1)!} N(N+1) \cdots (N+n) \sim_{N \to \infty} \frac{d}{(n+1)!} N^{n+1}.
\]

From (5.13) we derive the asymptotic lower bound

\[
MSE_{\min} \geq \frac{(n+1)!^2}{12} \frac{1}{N^{2n+2}} = \mathcal{O}\left(R^{-(2n+2)}\right).
\]

Future work will be done to generalize this lower bound \( \mathcal{O}(R^{-(2n+2)}) \) to spaces \( \mathcal{V} \) of higher dimension \( W \).

## 5.9 Unknown initial conditions

In this section, we show that the control of the initial conditions of a multi-loop modulator is not essential for the deterministic approach of signal reconstruction. Suppose that the initial condition is no longer zero, but \( A_0^{(0)}, ..., A_0^{(n-1)}, C_0 \) as shown in Figure 1.12(a). Then, it can be verified that feeding this encoder with \( X \), is equivalent to feeding the zero initial condition encoder of Figure 5.1 with \( X + I \),
where \( I \) is the signal defined as follows: \( I \) is zero everywhere except at the \( n \) first instants \( k = 1, \ldots, n \) where
\[
I(k) = \sum_{i=1}^{n-k+1} (-1)^{k-1} \binom{n-i}{k-1} (A_0^{i-1} - C_0).
\]

We call \( I \) the initial condition signal. This means that, rigorously speaking, we no longer have \( X \in \mathcal{C}(C) \), but \( X + I \in \mathcal{C}(C) \) or \( X \in \mathcal{C}(C) - I \). In practice, \( n \) is almost zero compared to the total number of samples \( N \) (typically \( n \leq 4 \)). Therefore \( I \) is negligible, and practically \( X \in \mathcal{C}(C) \).

At first, one might think that unknown initial conditions make the use of the digital output sequence \( C \) impractical, since they imply digital errors, amplified by the feedback and the internal integrator. Indeed, the output sequence \( C \) may be digitally completely different from that obtained in the case of zero initial conditions. However, the important result is that the input \( X \) is close to \( \mathcal{C}(C) \) in the MSE sense.

### 5.10 Conclusion

This chapter shows that the deterministic approach of simple encoding, which was applicable to predictive encoding, can be generalized to noise-shaping encoding as well. The implications are the same: necessity of consistency for optimality, non-consistency of linear decoding, convex projections as a tool for automatic improvement of non-consistent estimates. However, certain particularities have to be noted. The set \( \mathcal{C}(C) \) is no longer rectangular, although it is still a hyper-parallelepiped with center \( C \). The convex projection on \( \mathcal{C}(C) \) requires more elaborate algorithms. We developed a specific algorithm, called the “Thread Algorithm” for the case of single-loop \( \Sigma \Delta \) modulation. We also proposed an algorithm for the general case of \( n \)th order multi-loop \( \Sigma \Delta \) modulation, which is the object of a long appendix (Appendix A.5). Numerical evaluation of consistent estimates approached by alternating projections show the general \( \mathcal{O}(R^{-2n+3}) \) MSE asymptotic behavior, instead of \( \mathcal{O}(R^{-2n+1}) \) in linear decoding. This implies an improvement of 3 dB per octave of oversampling, regardless of the order of the modulator. The same approach as in predictive encoding is used to analyze this result theoretically. The same model assumptions (Assumptions 4.7.2 and 4.7.3) lead to the derivation of \( \mathcal{O}(R^{-2n+2}) \) as an upper bound to consistent reconstruction MSE. Finally, the partitioning approach of encoding applied to multi-loop \( \Sigma \Delta \) modulators show that, in the case of constant inputs, \( \mathcal{O}(R^{-2n+2}) \) is the asymptotic lower bound to consistent reconstruction, including optimal reconstruction.

This analysis of noise-shaping encoders will be used for the study of multi-stage \( \Sigma \Delta \) modulators in the next chapter.
Chapter 6

Multi-stage $\Sigma\Delta$ modulation

6.1 Introduction

Multi-stage $\Sigma\Delta$ modulators where presented in Section 1.2.5 and their block diagram was shown in Figure 1.15(a). Using the mapping description of noise-shaping encoders from Chapter 5, each encoder is characterized by:

$$H_i = I + G_i.$$ 

In this chapter, we consider the case where each stage is a single-loop or multi-loop $\Sigma\Delta$ modulator. Therefore, for each $i = 1, \ldots, p$, $H_i$ is an integrator whose order is designated by $n_i$.

We recall from Section 1.2.5 that the linear decoding scheme consists in lowpass filtering a predecoded version of the encoded outputs $C_1, \ldots, C_p$ of the $p$-stages. The predecoding scheme is shown in Figure 1.15(b). For convenience, let us define the cumulative integrator mapping

$$H'_i = H_i \cdots H_2 H_1$$

(by convention $H'_0 = I$). The mapping $H'_i$ is simply an integrator of order $n'_i = n_1 + \ldots + n_i$. We call $n = n'_p = n_1 + \ldots + n_p$ the total order of the multi-stage modulator. Using these notations, the predecoded version $\hat{C}_p$ of $(C_1, \ldots, C_p)$ can be expressed as $\hat{C}_p = \sum_{j=1}^{p} H'^{-1}_{j-1} |C_j|$. In general, for $i = 1, \ldots, p$, we define the predecoded version $\hat{C}_i$ of $(C_1, \ldots, C_i)$ as:

$$\hat{C}_i = \sum_{j=1}^{i} H'^{-1}_{j-1} |C_j|.$$  \hspace{1cm} (6.2)

The whole linear decoding scheme is represented in Figure 6.1.
6.2 Deterministic description of multi-stage $\Sigma\Delta$ modulation encoding

The natural way to generalize the deterministic approach to multi-stage encoding is simply to say that, when an input $X$ produces simultaneously the $p$-tuple of encoded signals through the cascade structure, then the information available about $X$ is “ $X \in C(C_1, ..., C_p)$ ” where $C(C_1, ..., C_p)$ designates the set of all signals producing $(C_1, ..., C_p)$ through the $p$ encoders. We show in this section that $C(C_1, ..., C_p)$ is the intersection of $p$ parallelepipeds.

The derivation is performed recursively on the number of stage $p$. In general, for $i = 1, ..., p$, $C(C_1, ..., C_i)$ designates the set of all signals producing $(C_1, ..., C_i)$ through the first $i$ encoders. We show that a $p$-stage $\Sigma\Delta$ modulator is equivalent to the structure of Figure 6.2. This structure shows that the last encoded signal $C_p$ can be equivalently obtained by a virtual encoder whose input is $X - \hat{C}_{p-1}$, where $\hat{C}_{p-1}$ is the predecoded version of $C_1, ..., C_{p-1}$ output by the first $(p - 1)$ encoders. The virtual encoder itself is similar to the $p^{th}$ noise-shaping encoder of the original structure, where the mapping $H_p$ has to be replaced by $H'_p$. We recall that $H'_p$ is the cumulative integrator $H_p ... H_1$ of order $n'_p = n$.

This equivalence is derived from the dependence of the quantization error signal $E_i$ produced by the $i^{th}$ noise-shaping encoder with the input $X$ (see Figure 1.15(a)). To emphasize this dependence, we write $E_i = E_i[X]$. We have the following proposition:

**Proposition 6.2.1** In a $p$-stage $\Sigma\Delta$ modulator, for $i \in \{1, ..., p\}$, when $X \in C(C_1, ..., C_i)$, the quantization error signal of the $i^{th}$ modulator is

$$E_i[X] = H'_i[\hat{C}_i - X],$$

(6.3)
where $\hat{C}_i$ is the predecoded version of $(C_1, \ldots, C_i)$.

Proof: Assume $X \in C(C_1, \ldots, C_i)$. The output of the first encoder is then $C_1$. Since the resulting quantization error signal is denoted by $E_1[X]$, applying Proposition 5.7.1 we have

$$E_1[X] = H_1[C_1 - X].$$

We can also apply this proposition to the $j^{th}$ encoder where $2 \leq j \leq i$. The input of this encoder is $-E_{j-1}[X]$ (see Figure 1.15(a)), its output is $C_j$ and the resulting error signal is $E_j[X]$. Therefore,

$$E_j[X] = H_j[C_j + E_{j-1}[X]].$$

By induction, we can express $E_i[X]$ in terms of $X$. We find

$$E_i[X] = \sum_{j=1}^{i} H_i \cdots H_j[C_j] - H_i \cdots H_1[X].$$

Using (6.1) and 6.2, we find:

$$E_i[X] = \sum_{j=1}^{i} H_i H_{j-1}^{-1}[C_j] - H_i'[X] = H_i'[\hat{C}_i - X] \quad \Box$$

This proposition implies that the input of the last encoder is $-E_{p-1}[X] = H_{p-1}[X - \hat{C}_{p-1}]$. Therefore, when $X \in C(C_1, \ldots, C_{p-1})$, the output $C_p$ of the $p^{th}$ encoder is
equivalently obtained as shown in Figure 6.3. Using the fact that $H_p H'_{p-1} = H'_p$, we obtain the global equivalence of Figure 6.2.

To derive the set $\mathcal{C}(C_1, \ldots, C_p)$, we can now work on the equivalent structure. We say that $X \in \mathcal{C}(C_1, \ldots, C_p)$ if and only if $X \in \mathcal{C}(C_1, \ldots, C_{p-1})$ and $X - \hat{C}_{p-1}$ produces $C_p$ through the virtual encoder. Applying Proposition 5.2.2 to the virtual encoder, we find that

$$X \in \mathcal{C}(C_1, \ldots, C_p) \iff X \in \mathcal{C}(C_1, \ldots, C_{p-1}) \text{ and } X - \hat{C}_{p-1} \in H'_{p-1} \left[ Q^{-1}_p[C_p] + G_p[C_p] \right].$$

For $i = 1, \ldots, p$, $C$ an encoded signal and $\hat{C} \in \mathbb{R}^N$, let us define the parallelepiped $\mathcal{C}_i(C/\hat{C})$ as

$$\mathcal{C}_i(C/\hat{C}) = H_i^{-1} \left[ Q^{-1}_i[C] + G_i[C] \right] + \hat{C}. \quad (6.4)$$

We have proved that

$$X \in \mathcal{C}(C_1, \ldots, C_p) \iff X \in \mathcal{C}(C_1, \ldots, C_{p-1}) \text{ and } X \in \mathcal{C}_p(C_p/\hat{C}_{p-1}).$$

Therefore, we have:

**Proposition 6.2.2**

$$\mathcal{C}(C_1, \ldots, C_p) = \mathcal{C}(C_1, \ldots, C_{p-1}) \cap \mathcal{C}_p(C_p/\hat{C}_{p-1}).$$

This result can be used to derive $\mathcal{C}(C_1, \ldots, C_{p-1})$ and so forth. By induction, we find:

**Proposition 6.2.3**

$$\mathcal{C}(C_1, \ldots, C_p) = \bigcap_{i=1}^p \mathcal{C}_i(C_i/\hat{C}_{i-1}).$$

Therefore $\mathcal{C}(C_1, \ldots, C_p)$ is the intersection of $p$ parallelepipeds of the type (6.4).
6.3 The partitioning approach to multi-stage $\Sigma\Delta$ modulation

As in for the previous types of encoding, a multi-stage $\Sigma\Delta$ modulator defines a partition on $\mathbb{R}^N$ induced by

$$\{\mathcal{C}(C_1, \ldots, C_p) \mid C_1, \ldots, C_p \text{ are encoded signals}\}. \quad (6.5)$$

We saw in Sections 3.2 and 4.3 that this partition has a lattice structure in simple encoding and $\Delta$ modulation. For multi-loop $\Sigma\Delta$ modulation, we saw in section 5.3 that the partition loses the cubic (orthogonal) property but still has a lattice structure. Unfortunately, in the case of multi-stage $\Sigma\Delta$ modulation, the partition loses the lattice structure itself. However, it still has “locally” a lattice structure in a sense we are going to explain. This local property of the partition will find its importance when deriving lower bounds to optimal reconstruction MSE (Section 6.7).

Consider the equivalent block diagram of a $p$-stage $\Sigma\Delta$ modulator in Figure 6.2. Suppose we know the partition induced by the $(p-1)$-stage modulator, that is

$$\{\mathcal{C}(C_1, \ldots, C_{p-1}) \mid C_1, \ldots, C_{p-1} \text{ are encoded signals}\}. \quad (6.6)$$

For each cell $\mathcal{C}(C_1, \ldots, C_{p-1})$ of this partition, the following family

$$\{\mathcal{C}(C_1^0, \ldots, C_{p-1}^0, C_p) \mid C_p \text{ is an encoded signal}\}, \quad (6.6)$$
forms a subpartition of $\mathcal{C}(C_1^0, \ldots, C_{p-1}^0)$. We show that this partition has a lattice structure.

When input signals are confined to belong to $\mathcal{C}(C_1^0, \ldots, C_{p-1}^0)$, the input of the last encoder of Figure 6.2 is subtracted from $\dot{C}_{p-1}^0$ which is a fixed signal, since it is a function of $(C_1^0, \ldots, C_{p-1}^0)$. Therefore, the subpartition (6.6) will be locally the partition induced by the last encoder, translated by the vector $\dot{C}_{p-1}^0$. This partition has a lattice structure because this encoder has the structure of a multi-stage $\Sigma\Delta$ modulator.

The same reasoning can be applied to the $(p-1)$-stage modulator. On the whole, the partition induced by a multi-stage $\Sigma\Delta$ modulator has a multi-level structure of locally nested lattices. Figure 6.4 shows the example of the two-stage $\Sigma\Delta$ modulator ($p = 2$) composed of two single-bit single-loop $\Sigma\Delta$ modulators (MASH structure).

In the situation of oversampling, the partition induced on $\mathcal{V}$ by a multi-stage $\Sigma\Delta$ modulator is naturally defined by

$$\{\mathcal{C}(C_1, \ldots, C_p) \cap \mathcal{V} \mid C_1, \ldots, C_p \text{ are encoded signals}\}. \quad (6.7)$$
Figure 6.4: Construction of the partition on $\mathbb{R}^2$ induced by a two-stage $\Sigma\Delta$ modulator composed of two single-bit single-loop modulators (MASH structure) in the non-overload region. (a) Partition of the first single-loop encoder (non-overload region). (b) Partition of the $p^{th}$ encoder ($p = 2$) in the equivalent block diagram of Figure 6.2 (non-overload region). (c) Nesting of partition (b) in the cells of partition (a). (d) Resulting partition of the two-stage modulator.
6.4 Consistent estimates and non-consistency of linear decoding

The consistent estimates are naturally the elements of $\mathcal{C}(C_1, ..., C_p) \cap \mathcal{V}$. Property 3.3.4 still holds when replacing $\mathcal{C}(C)$ by $\mathcal{C}(C_1, ..., C_p)$ since $\mathcal{C}(C_1, ..., C_p)$ is a convex set as intersection of convex sets. We have

**Proposition 6.4.1** $\hat{C}_p$ is the center of the parallelepiped $\mathcal{C}_p(\hat{C}_p/\hat{C}_{p-1})$.

Proof: We have $\mathcal{C}_p(\hat{C}_p/\hat{C}_{p-1}) = H_p^{-1} \left[ Q_p^{-1}[C_p] + G_p[C_p] \right] + \hat{C}_{p-1}$. Applying Proposition 5.2.2 and Proposition 5.4.1 to the $p$th encoder, $\hat{C}_p$ is the center of $H_p^{-1} \left[ Q_p^{-1}[C_p] + G_p[C_p] \right]$. Therefore, the center of $\mathcal{C}_p(\hat{C}_p/\hat{C}_{p-1})$ is $H_p^{-1}[C_p] + \hat{C}_{p-1}$, which is equal to $\hat{C}_p$, from the general definition of $\hat{C}_i$ in (6.2) \(\square\)

Therefore, the linear decoding scheme also has the geometric interpretation of Figure 5.2 where the parallelepiped is $\mathcal{C}_p(\hat{C}_p/\hat{C}_{p-1})$. Thus, $P_\mathcal{V}[\hat{C}_p]$ is not necessarily consistent because it does not necessarily belong to $\mathcal{C}_p(\hat{C}_p/\hat{C}_{p-1})$.

6.5 Methods of improvements of non-consistent estimates

A non-consistent estimate can be improved by a projection on $\mathcal{V}$ or on one of the parallelepipeds $\mathcal{C}_i(\hat{C}_i/\hat{C}_{i-1})$. We have

$$\mathcal{C}_i(\hat{C}_i/\hat{C}_{i-1}) = H_i^{-1} \left[ Q_i^{-1}[C_i] + \left( G_i[C_i] + H_i'[\hat{C}_{i-1}] \right) \right].$$

Since $H_i^{-1}$ is the discrete-time differentiator of order $n'_i$, the expression of $\mathcal{C}_i(\hat{C}_i/\hat{C}_{i-1})$ is analogous to that of $\mathcal{C}(C)$ in an $n'_i$-loop $\Sigma\Delta$ modulator, after replacing $G[C]$ by $G_i[C_i] + H_i'[\hat{C}_{i-1}]$. The projection algorithm presented for multi-loop $\Sigma\Delta$ modulation can be directly used without any special modification.

As a generalization of single-stage encoding, consistent estimates can be obtained by iterating projections infinitely and periodically on the $p + 1$ convex sets $\mathcal{V}$, $\mathcal{C}_1(C_1)$, $\mathcal{C}_2(C_2/C_1)$, ..., $\mathcal{C}_p(C_p/C_{p-1})$, as justified in [34]. Another approach is to see that

$$\mathcal{C}(C_1, ..., C_p) \cap \mathcal{V} = \bigcap_{i=1}^{p} \left( \mathcal{C}_i(\hat{C}_i/\hat{C}_{i-1}) \cap \mathcal{V} \right).$$

Because $\mathcal{C}_i(\hat{C}_i/\hat{C}_{i-1}) \cap \mathcal{V}$ is similar to the set of consistent estimates of an $n'_i$th order multi-loop modulator, it is expected that estimates chosen in such a set will yield an MSE of the order of $R^{-(2n_i+2)}$. Since $n'_i$ is increasing with $i$, $\mathcal{C}_p(\hat{C}_p/\hat{C}_{p-1}) \cap \mathcal{V}$ is the set of “smallest” size. Therefore, the size of $\mathcal{C}(C_1, ..., C_p) \cap \mathcal{V}$ is of the same order as that of $\mathcal{C}_p(\hat{C}_p/\hat{C}_{p-1}) \cap \mathcal{V}$. Then it will be sufficient to perform alternating projections between $\mathcal{C}(C_1, ..., C_p)$ and $\mathcal{V}$ only, and the infinite iteration will be expected to yield an MSE of the order of $R^{-(2n+2)}$. This is the improvement method used in the next section for numerical tests.
6.6 Numerical tests

Numerical tests in the same conditions as in multi-loop ΣΔ modulation were performed on multi-stage ΣΔ modulators, where each encoder is a single-bit single-loop ΣΔ modulator. The number of stages varied between 1 and 3 (the single-stage configuration coincides with the single-loop case). The results are plotted in Figure 6.5 in the same format as in Figure 5.7. The asymptotic behavior \( O(R^{-2n}) \) is again confirmed, where \( n \) is the total order of the multi-stage encoder.

6.7 Lower bound on optimal reconstruction with constant inputs

The \( O(R^{-2n}) \) behavior was obtained numerically in Section 6.6 by alternating projections between \( \mathcal{V} \) and the last parallelepiped \( C_p/C_{p-1} \). Although this method does not lead to a consistent estimate, that is, an element of \( C(C_1, \ldots, C_p) \cap \mathcal{V} \), we explained qualitatively in Section 6.5 that \( O(R^{-2n}) \) is the expected order of consistent reconstruction. A way to verify the validity of this conjecture is to evaluate a lower bound on optimal reconstruction. In the case where we find the same order \( O(R^{-2n}) \), we will be certain that consistent reconstruction cannot be better than that. In this section, we study the case of constant inputs, as in section 5.8. We show that, in this case, the expectation of the optimal reconstruction MSE is indeed lower bounded by \( O(R^{-2n}) \).

We use the same method of derivation as in Section 5.8. Consider a certain segment of input signals \( \mathcal{D} \) included in the one dimensional space \( \mathcal{V} \) of constant signals. The expectation of the optimal reconstruction MSE is lower bounded by \( MSE_{\text{min}} = \frac{d^2}{12N_d^2} \) given in (5.13). We recall that \( d \) is the length of the segment \( \mathcal{D} \) and \( N_d \) is the total number of cells intersected by \( \mathcal{D} \) in the partition induced by the encoder on \( \mathbb{R}^N \). As in Section 5.8, the method is to evaluate an upper bound to \( N_d \).

We propose a derivation by induction on the number of stages \( p \). For any segment \( \mathcal{D} \subset \mathcal{V} \) of length \( d \), let us denote by \( N_d^{(p)} \) the number of cells intersected by \( \mathcal{D} \) in the partition induced by the \( p \)-stage modulator. Consider the equivalent block diagram of Figure 6.2. Suppose we know the number \( N_d^{(p-1)} \) of cells defined by the \( (p-1) \)-stage modulator and intersected by \( \mathcal{D} \). The segment \( \mathcal{D} \) is divided into \( N_d^{(p-1)} \) subsegments \( \mathcal{D}_i \) by intersection with these \( N_d^{(p-1)} \) cells (see Figure 6.6). If \( d_i \) is the length of \( \mathcal{D}_i \), we have of course:

\[
d = \sum_{i=1}^{N_d^{(p-1)}} d_i. \tag{6.8}
\]

But we also have:

\[
N_d^{(p)} = \sum_{i=1}^{d} N_d^{(p-1)} \tag{6.9}
\]
Figure 6.5: Dependence of the decoding MSE with the oversampling ratio in the case of single-bit single-loop single-stage, two-stage and three-stage \(\Sigma\Delta\) modulation. (a) Comparison between linear decoding and consistent decoding approached by alternating projections. (b)-(c)-(d) Decoding MSE normalized with the theoretical noise shaping equation (5.7) for linear decoding, consistent decoding approached by alternating projections and finite step alternating projections: (b) single-stage configuration; (c) two-stage configuration; (d) three-stage configuration.
Figure 6.6: Division of the segment $\mathcal{D}$ in the partition induced by the $(p-1)$-stage $\Sigma \Delta$ modulator in the example of a two-stage MASH configuration (see Figure 6.4(c)).
This relation simply says that, when considering the partition induced by the \( p \)-stage modulator, the number of cells intersected by \( D \) is obtained by counting this number for each segment \( D_i \) and add up the result over all indices \( i \).

For a fixed index \( i \), the segment \( D_i \) is entirely confined in a cell \( \mathcal{C}(C_1^i, ..., C_{p-1}^i) \) defined by the \((p-1)\) stage modulator. Therefore, in the partition induced by the \( p \)-stage modulator, the cells intersected by \( D_i \) necessarily have the form \( \mathcal{C}(C_1^i, ..., C_{p-1}^i, \bar{C}_p) \). From Section 6.3, we know that the subpartition

\[
\{ \mathcal{C}(C_1^0, ..., C_{p-1}^0, \bar{C}_p) \mid \bar{C}_p \text{ is an encoded signal} \},
\]

of the cell \( \mathcal{C}(C_1^i, ..., C_{p-1}^i) \) has the lattice structure of the partition induced by the last encoder. This encoder is a multi-loop modulator of order \( n'_p = n_1 + ... + n_p \). Using the upper bound result (5.19) applicable for multi-loop \( \Sigma \Delta \) modulators, we conclude that

\[
N_{d_n}^{(p)} \leq 1 + d_i \cdot S_{n'_p+1}(N). \tag{6.10}
\]

We recall the definition of \( S_n(N) \) given in (A.25) in Section A.5.4:

\[
S_n(N) = \frac{1}{n!} N(N+1) \cdot (N+n-1) \simeq \frac{N^n}{n!}. \tag{6.11}
\]

Applying successively (6.9) and (6.8), we find that

\[
N_{d_n}^{(p)} \leq \sum_{i=1}^{N_{d_n}^{(p-1)}} \left( 1 + d_i \cdot S_{n'_p+1}(N) \right)
\]

\[
\leq N_{d_n}^{(p-1)} + \sum_{i=1}^{N_{d_n}^{(p-1)}} d_i \cdot S_{n'_p+1}(N)
\]

\[
\leq N_{d_n}^{(p-1)} + d \cdot S_{n'_p+1}(N)
\]

This technique used to find an upper bound to \( N_{d_n}^{(p)} \) can be applied on \( N_{d_n}^{(p-1)} \) and so forth. By induction, we find

\[
N_{d_n}^{(p)} \leq d \cdot S_{n'_1+1}(N) + \cdots + d \cdot S_{n'_p+1}(N) \simeq d \cdot S_{n'_p+1}(N).
\]

The last equivalence results from the fact that \( n'_p > n'_{p-1} > ... > n'_1 \) which implies that \( n'_p \) is the dominant power. Then, applying (5.13) and (6.11), we find the asymptotic lower bound

\[
MSE_{\min} \geq \left( \frac{(n+1)!^2}{12} \frac{1}{N^{2n+2}} \right) \mathcal{O} \left( R^{-(2n+2)} \right),
\]

where \( n = n'_p \) is the total order of the multi-stage \( \Sigma \Delta \) modulator.
6.8 Unknown initial conditions

Assume that each encoder of the multi-stage modulator has non-zero initial conditions. This is equivalent to saying that the \(i^{th}\) encoder has an initial condition signal \(I_i\) (see Section 5.9). It can be shown that this has the effect to shift each parallelepiped \(\mathcal{C}_i(C_i/C_{i-1})\) by the fixed signal \(-\hat{I}_i\), where \(\hat{I}_i\) is the predecoded version of \((I_1, ..., I_i)\), that is \(\hat{I}_i = \sum_{j=1}^{i} H_{j-1}^{-1}[I_j]\). We know from Section 5.9, that \(I_i\) is non-zero only at the \(n_i\) first instants. It is easy to show that \(\hat{I}_i\) is non-zero only at the \(n'_i\) first instants. Then \(\hat{I}_i\) can be neglected for the deterministic analysis.

6.9 Conclusion

In a deterministic approach, a multi-stage \(\Sigma\Delta\) modulator is a mapping between an input signal \(X\) and a \(p\)-tuple of quantized signals \((C_1, ..., C_p)\). Although the analysis of this mapping is more complex than for the previous types of encoders, the reasoning remains the same. The exact information about an input signal \(X\) from its output \((C_1, ..., C_p)\), is that \(X\) belongs to the image of \((C_1, ..., C_p)\) through the inverse of this mapping, which is the set \(\mathcal{C}(C_1, ..., C_p)\). In the context of oversampled ADC, the information is that \(X\) belongs to \(\mathcal{C}(C_1, ..., C_p)\cap\mathcal{V}\), called the set of consistent estimates. We show that \(\mathcal{C}(C_1, ..., C_p)\) is the intersection of \(p\) hyper-parallelepiped, \(\mathcal{C}(C_i/C_{i-1})\) (with \(i = 1, ..., p\)) similar to those generated by single-stage modulators. The center of the last parallelepiped \(\mathcal{C}(C_p/C_{p-1})\) coincides with the signal \(\hat{C}_p\) used in linear decoding as the signal to be lowpass filtered. Although consistent estimates are rigorously defined by the set \(\mathcal{C}(C_1, ..., C_p)\cap\mathcal{V}\), we show that estimates chosen in the subset \(\mathcal{C}(C_p/C_{p-1})\cap\mathcal{V}\) should give results of the same order. The algorithm for the convex projection on \(\mathcal{C}(C_p/C_{p-1})\) is identical to that used for a multi-loop \(\Sigma\Delta\) modulator. Experimental results show that estimates of \(\mathcal{C}(C_p/C_{p-1})\cap\mathcal{V}\) approached by alternating projections yield an MSE of the type \(\mathcal{O}(R^{-(2n+2)})\) where \(n\) is the total order of the multi-stage \(\Sigma\Delta\) modulator, whereas linear decoding yields \(\mathcal{O}(R^{-(2n+1)})\).

As in multi-loop \(\Sigma\Delta\) modulation, we also prove using the partitioning approach, that consistent reconstruction (in the rigorous sense) is lower bounded by \(\mathcal{O}(R^{-(2n+2)})\) in the case of constant inputs.
Chapter 7

Conclusion, discussion and future work

Following Shannon’s sampling theorem, definite rules of signal reconstruction are available when a bandlimited signal is given by its sampled version. However, in the context of oversampling, this is not quite the case when sampling is followed by amplitude quantization or encoding. In the classical approach, the quantization error signal is considered as an independent white noise, and an estimate of the original signal is obtained by lowpass filtering the quantized or encoded signal. Although this method leads to a practical method for analog signal reconstruction with good performances, the theoretical question of best possible reconstruction from an encoded signal in oversampled ADC had not really been studied.

To identify the exact information about an analog signal available in its encoded version, a deterministic approach of oversampled ADC is necessary, since quantization is fundamentally a deterministic operation. In general, any encoding system, whether it is a simple quantizer, a predictive encoder or a ΣΔ modulator, can be described as a many-to-one mapping. The exact information about the original signal is its location in the set obtained by inverting the mapping in the space of bandlimited signals and called the set of consistent estimates. The convexity property of this set implies that consistency is a necessary condition for reconstruction optimality.

The deterministic approach then indicates that the reconstruction achieved by the classical method (linear decoding) is not complete, since it does not necessarily lead to consistent estimates. The non-consistency of linear decoding can be seen in the difference of performance between consistent decoding and linear decoding. Indeed, the MSE of consistent estimates decreases with the oversampling ratio $R$ by at least by 3 dB per octave faster than the MSE of linear decoding estimates. In simple encoding, this result is both observed numerically (Section 3.8) and derived analytically (Theorem 3.7.1), provided that some conditions on the input signal’s quantization threshold crossings holds. In Δ modulation and ΣΔ modulation, this result is observed numerically, without any particular condition on the input
signal (Sections 4.6, 5.6 and 6.6). It can also be derived analytically (Theorems 4.7.6 and 5.7.3) assuming non-overloading (Assumption 4.7.1) and starting from some theoretical model assumptions (Assumptions 4.7.2 and 4.7.3). Explicitly, the dependence with the oversampling ratio of the consistent estimate MSE is in \( O(R^{-2n+2}) \) instead of \( O(R^{-2n+1}) \) in linear decoding, where \( n \) is the order of the modulator.

Thanks to the partitioning interpretation of an encoder (Sections 3.2, 4.3, 5.3 and 6.3), we prove for multi-loop and multi-stage \( \Sigma \Delta \) modulators (Sections 5.8 and 6.7) that the performance \( O(R^{-2n+2}) \) is also a lower bound to optimal reconstruction in the case where input signals are confined to be constant. This gives first hints that, whereas linear decoding is not optimal, the theoretical limit of reconstruction in terms of the MSE asymptotic dependence with \( R \), might be achieved by the consistent estimates. This gives preliminary results for a future “oversampled quantization theorem”, which is the necessary complement to Shannon’s sampling theorem to establish rules of reconstruction in the presence of quantization.

While the deterministic analysis of oversampled ADC shows the non-consistency of linear decoding, and thus its non-optimality, it also gives methods and algorithms for improvements of non-consistent estimates. They are based on convex projections and are proposed for the various types of encoders, including simple encoding (Section 3.5), predictive encoding (Section 4.5), multi-loop \( \Sigma \Delta \) modulation (Section 5.5 and Appendix A.5) and multi-stage \( \Sigma \Delta \) modulation (Section 6.5). Consistent estimates can themselves be approached by iteration of the convex projections. But achieving a consistent estimate is not necessary to obtain some improvement. In general, it is sufficient to modify the estimate so that it gets closer to the set of consistent estimates. This can be achieved for example by finite step projections. Sections 5.6 and 6.6 showed that very substantial improvement is achieved with very limited numbers of steps. Qualitatively speaking, the set of consistent estimates does not represent the available information contained in the encoded signal, it also represents the “direction” towards which one should tend to improve non-consistent estimates. Methods of reconstruction more practical than convex projections but following this general criterion will be investigated in a future work.

An important part of this work was to formalize the signal transformation produced by A/D converters, going back to the original deterministic definition of quantization. We developed a complete and consistent formalism, based on the notion of signal mapping, which permits the deterministic description of a large variety of encoders. We saw the effective consequence of this approach in the context of oversampling, where input signals have some extra features related to bandlimitation. However, the deterministic analysis of quantization can also be used when input signals have extra features of a different nature. For example, this could be used when input signals have some power spectrum characteristics rather than bandlimitation characteristics.
As a part of current investigations, consistent reconstruction in oversampled ADC is studied in the case where input signals are not perfectly bandlimited. In fact, this corresponds to the real case of oversampled ADC where input signals are made bandlimited by a non-ideal analog lowpass filter. Although not perfectly bandlimited, power spectrum characteristics of the input signals can be deduced from the specifications of the non-ideal filter. Principle of reconstruction are currently investigated, where the deterministic information $X \in C(C)$ is combined with the minimization of a quadratic functional derived from the power spectrum features. This will lead to the classical reconstruction method of steepest descent with convex projection [39].

Signal reconstruction using the deterministic approach of quantization will also be studied in the presence of errors existing in the real operation of A/D converters (circuit imperfections, thermal noise). According to current research, we show that an imperfect encoder can be modeled as an ideal encoder provided that some fictive error signal $N$ is added to the input signal, where $N$ can be expressed as a function of the imperfections of the real encoder. In fact, this kind of modelization has been already introduced in this work in the case where the sources of errors are the initial conditions of the built-in integrators in $\Sigma \Delta$ modulation (Sections 5.9 and 6.8). In this situation, the information about the input signal $X$, given the encoded signal $C$, is that $X$ is bandlimited and $X + N \in C(C)$. Signal reconstruction will be investigated, based on the information $X \in C(C) - N$ which basically separates the errors due to the real factors ($N$) of statistical nature, and the errors due to quantization ($C(C)$) of deterministic nature. The question of theoretical limit of reconstruction in the presence of random factors will be investigated.

In conclusion, this work shows the power and potential of a deterministic analysis of quantization in oversampled ADC, demonstrating experimental and theoretical signal reconstruction improvements on known encoding schemes. This analysis implied the development of mathematical tools (concept of mapping) formalizing the basic mechanisms of A/D converters. The good results obtained in oversampled ADC indicate possible directions for future investigations based on this mathematical framework.
Appendix A

Appendices

A.1 Review of the $z$-transform

The $z$-transform is an algebraic notation used to describe discrete-time linear and time invariant systems.

A.1.1 Discrete-time, linear and time invariant system

A discrete-time, linear and time invariant system (LTI) is an operator $H$ which maps a sequence $(X(k))_{k \in \mathbb{Z}}$ into another sequence $(Y(k))_{k \in \mathbb{Z}} = H [(X(k))_{k \in \mathbb{Z}}]$, such that:

(i) $H$ is a linear mapping,
(ii) $\forall k_0 \in \mathbb{Z}$, $(Y(k - k_0))_{k \in \mathbb{Z}} = H [(X(k - k_0))_{k \in \mathbb{Z}}].$

Let $(\delta(k))_{k \in \mathbb{Z}}$ be the sequence defined by

$$
\begin{cases}
\delta(0) = 1, \\
\delta(k) = 0, \text{ for } k \neq 0
\end{cases}
$$

The sequence $(H(k))_{k \in \mathbb{Z}} = H [(\delta(k))_{k \in \mathbb{Z}}]$ is by definition the impulse response of $H$. It is shown [1] that, if $H$ is an LTI, $H$ is uniquely defined by its impulse response $(H(k))_{k \in \mathbb{Z}}$ and that

$$(Y(k))_{k \in \mathbb{Z}} = H [(X(k))_{k \in \mathbb{Z}}] \implies \forall k \in \mathbb{Z}, \ Y(k) = \sum_{j \in \mathbb{Z}} H(j) X(k - j). \quad (A.1)$$

The signal resulting from $(A.1)$ is denoted by

$$(Y(k))_{k \in \mathbb{Z}} = (H(k))_{k \in \mathbb{Z}} \ast (X(k))_{k \in \mathbb{Z}},$$

where $\ast$ is called the convolution operation.
A.1.2 Definition of the $z$-transform and properties

The $z$-transform is the transform which maps a discrete-time sequence $(X(k))_{k \in \mathbb{Z}}$ into the polynomial $P(z) = \sum_{k \in \mathbb{Z}} X(k) z^k$. This is a one-to-one mapping. By abuse of notation, the $z$-transform $P(z)$ of $(X(k))_{k \in \mathbb{Z}}$ is simply denoted by $X(z)$.

It is easy to show that, if $(Y(k))_{k \in \mathbb{Z}} = (H(k))_{k \in \mathbb{Z}} \ast (X(k))_{k \in \mathbb{Z}}$, then the $z$-transform $Y(z)$ is the product (in the sense of polynomials) of the $z$-transforms $H(z)$ and $X(z)$. We simply write $Y(z) = H(z)X(z)$. If $H(z)$ has an inverse in the set of polynomials, its inverse is denoted by $H^{-1}(z)$ (but also by $\frac{1}{H(z)}$) and $X(z) = H^{-1}(z)Y(z)$. We recall that the multiplication in the set of polynomials is associative and commutative.

A.1.3 Examples of description of LTI using the $z$-transform

A typical example of LTI is the delay operator $H$ defined by:

$$(Y(k))_{k \in \mathbb{Z}} = H[(X(k))_{k \in \mathbb{Z}}] \implies \forall k \in \mathbb{Z}, \ Y(k) = X(k - 1).$$

It is easy to show that $H$ is an LTI, with impulse response $(\delta(k - 1))_{k \in \mathbb{Z}}$ and $z$-transform $H(z) = z^{-1}$. This operator is represented as shown in Figure A.1(a). In the $z$-domain, the delay operation is written as $Y(z) = z^{-1}X(z)$. The differenti-
The integration operation is the inverse of the differentiation operation, is shown in Figure A.1(c) and is written in the $z$-domain as $Y(z) = (1 - z^{-1})^{-1} X(z) = \frac{1}{1 - z^{-1}} X(z)$.

### A.1.4 Relation with the discrete Fourier transform

The polynomial $X(z)$ taken at the complex value $z = e^{j\omega}$ is equal to $\sum_{k \in \mathbb{Z}} X(k) e^{j\omega k}$ and coincides with the discrete Fourier transform of $(X(k))_{k \in \mathbb{Z}}$. By abuse of notation, the value of $X(z)$ at $z = e^{j\omega}$ is denoted by $X(\omega)$. As a consequence, if $Y(z) = H(z) X(z)$, then $Y(\omega) = H(\omega) X(\omega)$. For example, if $H$ is the differentiation operator, $Y(z) = (1 - z^{-1})X(z)$. Therefore, $Y(\omega) = (1 - e^{-j\omega}) X(\omega) = \left(2j e^{-j\frac{\omega}{2}} \right) \sin(\frac{\omega}{2}) X(\omega)$.

### A.2 Lemma on the matrix $\mathcal{W}(t_1, ..., t_W)$

**Lemma A.2.1** For given a real number $\delta > 0$, $[\mathcal{W}(t_1, ..., t_W)]^{-1}$ is well defined and bounded on the set $S_\delta = \{(t_1, ..., t_w) \in [\delta, T]^W \mid \forall i = 2, ..., W, t_i \geq t_{i-1} + \delta\}$.

**Proof:** For $(t_1, ..., t_W) \in S_\delta$, $t_1, ..., t_W$ are necessarily distinct. Therefore $\mathcal{W}(t_1, ..., t_W)$ is invertible. Let us write $[\mathcal{W}(t_1, ..., t_W)]^{-1} = [M_{ij}(t_1, ..., t_W)]_{1 \leq i \leq j \leq W}$. Using the algebraic expression of the inverse of a matrix, it can be seen that $M_{ij}(t_1, ..., t_W)$ is a continuous function of $(t_1, ..., t_W)$ on the set $S_\delta$. Since $S_\delta$ is a closed and bounded set, it is compact. Its image through $M_{ij}(t_1, ..., t_W)$ is therefore compact and thus bounded. In other words $[M_{ij}(t_1, ..., t_W)]_{1 \leq i \leq j \leq W}$ is bounded on $S_\delta$. □

### A.3 Proof of Lemma 4.7.5

**Notation A.3.1** For a continuous function $X[t]$ on $[0, 1]$, 

$$\|X\|_\infty = \max_{t \in [0,1]} |X[t]| \quad \text{and} \quad \|X\|_1 = \int_0^1 |X[t]| dt.$$

**Notation A.3.2** For a function $X[t]$ differentiable on $[0, 1]$, $X^{(1)}[t]$ designates the derivative of $X[t]$ on $[0, 1]$.

**Lemma A.3.3** For a function $X[t]$ continuously differentiable on $[0, 1]$ and $N$ a positive integer, 

$$\left| \frac{1}{N} \sum_{k=1}^N |X(k)| - \|X\|_1 \right| \leq \frac{1}{N} \|X^{(1)}\|_\infty. \quad (A.2)$$
Proof: Let $X[t]$ be a function continuously differentiable on $[0, 1]$. We have

$$
\|X\|_1 = \int_0^1 |X[t]|dt = \sum_{k=1}^N \int_{\frac{k-1}{N}}^{\frac{k}{N}} |X[t]|dt.
$$

Since $|X[t]|$ is continuous on $[0, 1]$, according to the average theorem, we have

$$
\forall k = 1, \ldots, N, \exists t_k \in \left[ \frac{k-1}{N}, \frac{k}{N} \right], \quad \int_{\frac{k-1}{N}}^{\frac{k}{N}} X[t]dt = \frac{1}{N} X[t_k].
$$

Since $X[t]$ is continuously differentiable, using Lagrange inequality, we have

$$
\forall k = 1, \ldots, N, \left| \left| X\left[ \frac{k}{N} \right] \right| - \left| X[t_k] \right| \right| \leq \frac{1}{N} \max_{s \in \left[ \frac{k-1}{N}, \frac{k}{N} \right]} |X^{(1)}(s)| \leq \frac{1}{N} \|X^{(1)}\|_\infty.
$$

Therefore, using the inequality $|a| - |b| \leq |a - b|$, we find

$$
\left|\frac{1}{N} \sum_{k=1}^N \left| X\left[ \frac{k}{N} \right] \right| - \|X\|_1 \right| \leq \sum_{k=1}^N \frac{1}{N} \left| X\left[ \frac{k}{N} \right] \right| - \sum_{k=1}^N \int_{\frac{k-1}{N}}^{\frac{k}{N}} X[t]dt \leq \frac{1}{N} \sum_{k=1}^N \left| X\left[ \frac{k}{N} \right] - X[t_k] \right|
$$

$$
\leq \frac{1}{N} \sum_{k=1}^N \left| X\left[ \frac{k}{N} \right] - X[t_k] \right| \leq \frac{1}{N} \sum_{k=1}^N \frac{1}{N} \|X^{(1)}\|_\infty.
$$

The proof is then completed using the notation $X\left[ \frac{k}{N} \right] = X(k)$ defined in Section 2.1 \[\square\]

**Proof of Lemma 4.7.5**

Let $N$ be a positive integer and $X \in \mathcal{V}$. From the expression (2.7) we can see that the continuous version $X[t]$ of $X$ is continuously differentiable on $[0, 1]$. Therefore, from Lemma A.3.3, we have

$$
\frac{1}{N} \sum_{k=1}^N |X(k)| \geq \|X\|_1 - \frac{1}{N} \|X^{(1)}\|_\infty.
$$

To lower bound the right hand side, it is sufficient to lower bound $\|X\|_1$ and upper bound $\|X^{(1)}\|_\infty$. Since $\|X\|_1$ is a norm on $\mathcal{V}$ and $\mathcal{V}$ is of finite dimension, $\|X\|$ and $\|X\|_1$ are equivalent. Therefore, there exists a constant $B > 0$ (which depends on the definition of $\mathcal{V}$, but not on $N$) such that

$$
\forall X \in \mathcal{V}, \quad \|X\|_1 \geq B \|X\|.
$$

Unfortunately, $\|X^{(1)}\|_\infty$ is not a norm on $\mathcal{V}$. However, we only need an upper bound on $\|X^{(1)}\|_\infty$. This can be done as follows: choose a basis $\{U_i\}_{1 \leq i \leq W}$ of $\mathcal{V}$ which is orthonormal for the norm $\|\cdot\|$ in $\mathcal{V}$, and define $C = W \max_{1 \leq i \leq W} \|U_i^{(1)}\|_\infty$. If
\[ X = \sum_{i=1}^{W} \alpha_i U_i \] is the decomposition of \( X \) on \{\( U_i \)\}_{1 \leq i \leq W} \), and \( \|X\|^2 = \sum_{i=1}^{W} |\alpha_i|^2 \) which implies that \( |\alpha_i| \leq \|X\| \) for all \( i = 1, ..., W \). Therefore,

\[
\|X^{(t)}\|_\infty \leq \sum_{i=1}^{N} |\alpha_i|\|U_i^{(t)}\|_\infty \leq \max_{1 \leq i \leq W} \|U_i^{(t)}\|_\infty \leq C\|X\|.
\]

Then, taking \( c_1 = \frac{B}{2} > 0 \), we have \( \forall N \geq \frac{2c}{B} \), \( \forall X \in \mathcal{V} \),

\[
\frac{1}{N} \sum_{k=1}^{N} |X(k)| \geq B\|X\| - \frac{B}{2C} C\|X\| = c_1\|X\| \quad \square
\]

### A.4 Proof of Lemma 5.7.2

**Notation A.4.1** For a function \( X[t] \) integrable on \([0, 1]\), \( X^{(-1)}[t] \) designates the integral of \( X[t] \) on \([0, 1]\) defined as follows:

\[
\forall t \in [0, 1], \quad X^{(-1)}[t] = \int_{0}^{t} X[s]ds.
\]

**Notation A.4.2** For function \( X[t] \) integrable on \([0, 1]\) and \( n \geq 0 \), \( X^{(-n)}[t] \) designates the \( n^{th} \) order integral of \( X[t] \) recursively defined as follows:

\[
X^{(0)}[t] = X[t] \quad \text{and for } n \geq 0, \quad X^{(-n-1)}[t] \text{ is the integral of } X^{(-n)}[t].
\]

**Remark:** We recall from Section 2.1 that for a given integer \( N \), and a function \( X[t] \) defined on \([0, 1]\), \( X \) designates the sequence \((X(k))_{1 \leq k \leq N}\) such that for \( k = 1, ..., N \), \( X(k) = X[k/N] \). For \( n \geq 0 \), \( X^{(-n)} \) is the \( n^{th} \) order discrete-time integral of \( X \), defined in Section 2.2 (equation (2.6)), whose value at index \( k \) is denoted by \( X^{(-n)}(k) \). This should not be confused with \( X^{(n)}[k/N] \) which is the value of \( X^{(-n)}[t] \) (defined in Notation A.3.2) at \( t = k/N \). Qualitatively speaking, \( X^{(-n)}(k) \) is obtained by sampling \( X[t] \), performing an \( n^{th} \) order **discrete-time** integration and taking the \( k^{th} \) sample, whereas \( X^{(n)}[k/N] \) is obtained by performing an \( n^{th} \) order **continuous-time** integral of \( X[t] \), sampling the resulting signal and taking the \( k^{th} \) sample (see Figure A.2). The next lemma gives an analytical comparison between \( X^{(-n)}(k) \) and \( X^{(-n)}[k/N] \) when \( X[t] \) is continuously differentiable.

**Lemma A.4.3** For \( n \geq 0 \), \( X[t] \) a function continuously differentiable on \([0, 1]\) and \( N \) a positive integer,

\[
\forall k = 1, ..., N, \quad \left| \frac{1}{N^n} X^{(-n)}(k) - X^{(-n)}[k/N] \right| \leq \frac{1}{N} \sum_{m=-n+2}^{1} \|X^{(m)}\|_\infty. \tag{A.3}
\]

**Proof:** Let \( X[t] \) be a function continuously differentiable on \([0, 1]\) and \( N \) a positive integer. Let us show (A.3) by induction.
Figure A.2: Comparison between $X^{(-n)}(k)$ and $X^{(-n)}[\frac{k}{N}]$.

First, (A.3) is trivial for $n = 0$, since $\forall k = 1, \ldots, N$, $X^{(0)}(k) = X[\frac{k}{N}]$, $X^{(0)}[\frac{k}{N}] = X[\frac{k}{N}]$ and $\sum_{m=-n+3}^1 \|X^{(m)}\|_\infty = 0$. Next, suppose for a certain $n \geq 1$ that (A.3) is true at $n - 1$, that is:

$$\forall k = 1, \ldots, N, \quad \left| \frac{1}{N^{n-1}} X^{(-n+1)}(k) - X^{(-n+1)}[\frac{k}{N}] \right| \leq \frac{1}{N} \sum_{m=-n+3}^1 \|X^{(m)}\|_\infty. \quad (A.4)$$

Let $k$ be a fixed integer of $\{1, \ldots, N\}$. Using the fact that

$$\frac{1}{N^n} X^{(-n)}(k) = \frac{1}{N} \sum_{j=1}^k \frac{1}{N^{n-1}} X^{(-n+1)}(j)$$

and

$$X^{(-n)}[\frac{k}{N}] = \int_{0}^{\frac{k}{N}} X^{(-n+1)}[t] dt = \sum_{j=1}^k \int_{\frac{j-1}{N}}^{\frac{j}{N}} X^{(-n+1)}[t] dt,$$

we can write

$$\frac{1}{N^n} X^{(-n)}(k) - X^{(-n)}[\frac{k}{N}] = \frac{1}{N} \sum_{j=1}^k (C_j + D_j), \quad (A.5)$$

where for $j \in \{1, \ldots, k\}$,

$$C_j = \frac{1}{N^{n-1}} X^{(-n+1)}(j) - X^{(-n+1)}[\frac{j}{N}]$$

and

$$D_j = X^{(-n+1)}[\frac{j}{N}] - N \int_{\frac{j-1}{N}}^{\frac{j}{N}} X^{(-n+1)}[t] dt.$$  

Applying (A.4), we have $|C_j| \leq \frac{1}{N} \sum_{m=-n+3}^1 \|X^{(m)}\|_\infty$. Since $X^{(-n+1)}[t]$ is continuous on $[0,1]$, then from the average theorem,

$$\exists t_j \in \left[ \frac{j-1}{N}, \frac{j}{N} \right], \quad \int_{\frac{j-1}{N}}^{\frac{j}{N}} X^{(-n+1)}[t] dt = \frac{1}{N} X^{(-n+1)}[t_j]. \quad (A.6)$$

95
Since $X^{(-n+1)}[t]$ is continuously differentiable, then from Lagrange inequality,
\[
\left| X^{(-n+1)}[\frac{k}{N}] - X^{(-n+1)}[t_j] \right| \leq \left| \frac{k}{N} - t_j \right| \left\| (X^{(-n+1)})^{(1)} \right\|_{\infty} \leq \frac{1}{N} \left\| X^{(-n+2)} \right\|_{\infty} . \quad (A.7)
\]
From (A.6) and (A.7) we find that $|D_j| \leq \frac{1}{N} \left\| X^{(-n+2)} \right\|_{\infty}$. From (A.5) we have
\[
\left| \frac{1}{N^{n+1}} \sum_{k=1}^{N} X^{(-n)}(k) - X^{(-n)}[\frac{k}{N}] \right| \leq \frac{k}{N} \left( \frac{1}{N} \sum_{m=-n+3}^{1} \left\| X^{(m)} \right\|_{\infty} + \frac{1}{N} \left\| X^{(-n+2)} \right\|_{\infty} \right).
\]
\[
\leq \frac{1}{N} \sum_{m=-n+2}^{1} \left\| X^{(m)} \right\|_{\infty} , \quad (A.8)
\]
since $k \leq N$. Therefore (A.3) is true. The induction is completed \(\square\)

**Lemma A.4.4** For a function $X[t]$ continuously differentiable on $[0,1]$ and $N$ a positive integer,
\[
\left| \frac{1}{N^{n+1}} \sum_{k=1}^{N} X^{(-n)}(k) - \left\| X^{(-n)} \right\|_{1} \right| \leq \frac{1}{N} \sum_{m=-n+1}^{1} \left\| X^{(m)} \right\|_{\infty} . \quad (A.9)
\]

Proof: We can write
\[
\frac{1}{N^{n+1}} \sum_{k=1}^{N} X^{(-n)}(k) - \left\| X^{(-n)} \right\|_{1} = C + D, \quad (A.10)
\]
where
\[
C = \frac{1}{N} \sum_{k=1}^{N} \left( \frac{1}{N^n} \left| X^{(-n)}(k) - X^{(-n)}[\frac{k}{N}] \right| \right) \quad \text{and} \quad D = \frac{1}{N} \sum_{k=1}^{N} \left| X^{(-n)}[\frac{k}{N}] - \left\| X^{(-n)} \right\|_{1} \right| .
\]
Using the inequality $|a| - |b| \leq |a - b|$ and applying Lemma A.4.3, we have
\[
|C| \leq \frac{1}{N} \sum_{k=1}^{N} \left( \frac{1}{N^n} \left| X^{(-n)}(k) - X^{(-n)}[\frac{k}{N}] \right| \right) \leq \frac{1}{N} \sum_{m=-n+2}^{1} \left\| X^{(m)} \right\|_{\infty} . \quad (A.11)
\]
Applying Lemma A.3.3 on $X^{(-n)}[t]$ which is continuously differentiable on $[0,1]$, we have
\[
|D| \leq \frac{1}{N} \left\| (X^{(-n)})^{(1)} \right\|_{1} = \frac{1}{N} \left\| X^{(-n+1)} \right\|_{\infty} . \quad (A.12)
\]
Therefore, from (A.10), (A.11) and (A.12), we find
\[
\left| \frac{1}{N^{n+1}} \sum_{k=1}^{N} X^{(-n)}(k) - \left\| X^{(-n)} \right\|_{1} \right| \leq |C| + |D| \leq \frac{1}{N} \sum_{m=-n+1}^{1} \left\| X^{(m)} \right\|_{\infty} \quad \square
\]

**Proof of Lemma 5.7.2**
From Lemma A.4.4, for any positive integer $N$, we have
\[
\sum_{k=1}^{N} |X^{(n)}(k)| \geq N^{n+1} \left( \|X^{(n)}\|_1 - \frac{1}{N} \sum_{m=-n+1}^{1} \|X^{(m)}\|_\infty \right).
\]

It can be shown that there exist two numbers $B, C > 0$ such that
\[
\forall X \in \mathcal{V}, \quad \|X^{(n)}\|_1 \geq B \|X\| \quad \text{and} \quad \sum_{m=-n+1}^{1} \|X^{(m)}\|_\infty \leq C \|X\|.
\]

This is shown in a way similar to the proof of Lemma 4.7.5. For the first inequality, note that $X[t] \mapsto \|X^{(n)}\|_1$ is a norm on $\mathcal{V}$. For the second inequality, use the technique in the proof of Lemma 4.7.5. Therefore
\[
\sum_{k=1}^{N} |X^{(n)}(k)| \geq N^{n+1} \left( B - \frac{1}{N} C \right) \|X\|.
\]

Taking $c_1 = \frac{B}{2C}$, we have $\forall N \geq \frac{2C}{B}$, $\forall X \in \mathcal{V},$
\[
\sum_{k=1}^{N} |H[X](k)| = \sum_{k=1}^{N} |X^{(n)}(k)| \geq c_1 \|X\| N^{n+1} \quad \Box
\]

A.5 Step 2 of Algorithm 3' of Section 5.5.2

A.5.1 Reduction to a problem of minimization under equality constraints

For convenience, given sequences $C$ and $D$, we will simply write $Q^{-1} [S_K(C)] + D = \mathcal{I}(K)$. Also, for each $k = 1, ..., N$, we define $b_-(k)$ and $b_+(k)$ as the lower and upper bounds of the interval $q^{-1}[C(k)] + D(k)$ respectively (they can be equal to $-\infty$ and $+\infty$ respectively). With these notations, Step 2 of Algorithm 3' becomes:

Find $Y$ minimizing $\phi$ subject to the inequality constraints:
\[
\mathcal{I}(K) = \left\{ Y \in \mathbb{R}^N / \forall k \in K, \ b_-(k) \leq Y(k) \leq b_+(k) \right\}.
\]

Proposition A.5.1 Let $Y$ be the signal minimizing $\phi$ subject to the set of inequality constraints $\mathcal{I}(K)$. Suppose, we know the values of $Y$ at $k \in K$, $Y(k) = b(k)$. Then $Y$ is the signal minimizing $\phi$ under the set of equality constraints
\[
\mathcal{E}(K) = \left\{ Y \in \mathbb{R}^N / \forall k \in K, \ Y(k) = b(k) \right\}.
\]

(A.13)

Proof: This is immediate since $\mathcal{E}(K)$ is a subset of $\mathcal{I}(K) \quad \Box$

Conversely, we have following two propositions:
Proposition A.5.2 Let \((b(k))_{k \in K}\) a sequence of values defined for \(k \in K\), \(E(K)\) be the set of equality constraints defined in (A.13) and \(Y\) be the signal minimizing \(\phi\) subject to \(E(K)\). Then \(\forall k \notin K, \nabla_{\theta}[Y](k) = 0\).

Proof: Let \(l \notin K\) and \(\epsilon \in \{-1, 1\}\). Let \(Y_{\epsilon}\) be the signal equal to \(Y_{K}\) everywhere except at \(k = l\) where \(Y_{\epsilon}(l) = Y_{K}(l) + \epsilon\). By definition of \(E(K)\), one can see that \(Y_{\epsilon} \in E(K)\). Applying Lemma 5.5.2 on \(S = E(K)\), \(Y = Y_{K}\) and \(Y' = Y_{\epsilon}\), we find that \(\nabla_{\theta}[Y_{K}](l) \cdot \epsilon \geq 0\). Since this is true for \(\epsilon = \pm 1\), then \(\nabla_{\theta}[Y_{K}](l) = 0\). This is true for any \(l \notin K\) □

Proposition A.5.3 Let \((b(k))_{k \in K}\) a sequence of values defined for \(k \in K\) such that \(b_{-}(k) \leq b(k) \leq b_{+}(k)\), \(E(K)\) be the set of equality constraints defined in (A.13) and \(Y\) be the signal minimizing \(\phi\) subject to \(E(K)\). Suppose that \(Y\) verifies the following conditions:

\[
\forall k \in K, \begin{cases} 
\text{if } b(k) = b_{+}(k) & , & \nabla_{\theta}[Y](k) \leq 0 \\
\text{if } b(k) = b_{-}(k) & , & \nabla_{\theta}[Y](k) \geq 0 \\
\text{otherwise} & , & \nabla_{\theta}[Y](k) = 0. 
\end{cases} \tag{A.14}
\]

Then \(Y\) minimizes \(\phi\) subject to the set of inequality constraints \(I(K)\).

Proof: Let \(Y' \in I(K)\). We show that \(\langle \nabla_{\theta}[Y], Y' - Y \rangle \geq 0\). Let \(k \in \{1,...,N\}\) such that \(\nabla_{\theta}[Y] \neq 0\). By necessity \(k \in K\) (because of Proposition A.5.2) and \(b(k)\) is either equal to \(b_{+}(k)\) or \(b_{-}(k)\), according to (A.14). Suppose that \(b(k) = b_{+}(k)\). We have \(\nabla_{\theta}[Y](k) < 0\). Since \(Y' \in I(K)\), \(Y'(k) \leq b_{+}(k) = b(k) = Y(k)\). Therefore, \(\nabla_{\theta}[Y](k) \cdot (Y'(k) - Y(k)) \geq 0\). The proof is completed by applying Lemma 5.5.2 on \(S = I(K)\) □

In Section A.5.5, we will propose an algorithm which chooses \(K\) and a sequence \((b(k))_{k \in K}\) such that (A.14) is verified. With such a choice, the minimization problem is reduced to dealing with the set of equality constraints \(E(K)\). From Sections A.5.2 to A.5.4, we concentrate on solving such a problem.

### A.5.2 Computation issues

Let us call \(Y_{K}\) the signal minimizing \(\phi\) subject to the set of equality constraints \(E(K)\). We recall that Step 3 of Algorithm 3' requires the computation of the signal \(Z_{K} = H^{-1}[Y_{K}] = Y_{K}^{(n)}\). Computationally speaking, although \(Z_{K}\) is theoretically defined from \(Y_{K}\), it is not required to calculate \(Y_{K}\) entirely before calculating \(Z_{K}\). In fact, we will try to compute \(Z_{K}\) as directly as possible. The following propositions give some properties of \(Z_{K}\).

Proposition A.5.4 When \(H\) is the \(n^{th}\) order discrete-time integrator, then

\[
\forall Y \in R^{N}, \quad \nabla_{\theta}[Y] = (-1)^{n}Z^{[n]}, \text{ where } Z = H^{-1}[Y] = Y^{(n)}.
\]

98
This is shown in Section A.6. As a consequence of Lemma 5.5.2 and Proposition A.5.4 we have:

**Proposition A.5.5** $\forall p = 1, ..., m, \forall k \not\in K$, $Z_{K}^{[p]}(k) = 0$.

Proof: This is a consequence of Propositions A.5.2 and A.5.4.

Let us write $K = \{k_1, k_2, ..., k_r\}$ such that $1 \leq k_1 < k_2, ..., < k_r \leq N$. As a consequence of Proposition A.5.5, we have:

**Proposition A.5.6** $Z_{K}^{[n-2]}$ is piecewise linear on intervals $\{k_{i-1} + 1, ..., k_i + 1\}$ for $1 \leq i \leq p$, and on $\{k_r + 1, ..., N\}$.

Proof: For $1 \leq i \leq p$, $\forall k = k_{i-1} + 1, ..., k_i - 1$, $Z_{K}^{[i]}(k) = 0$, because of Proposition A.5.5. Since

$$Z_{K}^{[i]}(k) = (Z_{K}^{[n-2]}(k + 2) - Z_{K}^{[n-2]}(k + 1)) - (Z_{K}^{[n-2]}(k + 1) - Z_{K}^{[n-2]}(k)),$$

by definition of the forward derivative, we conclude that

$$\forall k = k_{i-1} + 1, ..., k_i - 1,$$

$$Z_{K}^{[n-2]}(k + 2) - Z_{K}^{[n-2]}(k + 1) = Z_{K}^{[n-2]}(k + 1) - Z_{K}^{[n-2]}(k)$$

Therefore, $Z^{[n-2]}$ is piecewise linear on intervals $\{k_{i-1} + 1, ..., k_i + 1\}$.

Let us define the following column vectors of size $(n - 1)$:

$$Z_i = \begin{bmatrix} Z_{K}^{[n-2]}(k_i + 1) \\ \vdots \\ Z_{K}^{[n]}(k_i + 1) \end{bmatrix}.$$  \hfill (A.15)

We have:

**Proposition A.5.7** $Z_K$ is uniquely defined by the sequence of vectors $(Z_i)_{0 \leq i \leq r}$.

Proof: $Z_{K}^{[n-2]}$ can be obtained by linear interpolation of the coefficients $(Z_{K}^{[n-2]}(k_i + 1))_{0 \leq i \leq r}$ given by the top components of the vectors $(Z_i)_{0 \leq i \leq r}$. Starting from the initial conditions $Z_{K}^{[1]}(0), ..., Z_{K}^{[n-1]}(1)$ given by the vector $Z_0$, $Z_K$ is then uniquely obtained by $(n - 2)^{th}$ order discrete-time integration of $Z_{K}^{[n-2]}$.

Theoretically, for $i \geq 1$, the $(n - 1)$ bottom components of $Z_i$ are not needed. However, in practice, errors are expected to be accumulated in time during the integration. The $(n - 1)$ bottom components of $Z_i$ can be used to readjust the integration constants at every $k = k_i + 1$.

In fact, the following proposition implies that $Z_r$ is equal to the null column vector $O_{n-1}$ of size $(n - 1)$.

**Proposition A.5.8** $\forall k = k_r + 1, ..., N$, $\forall q \geq 0$, $Z_{r}^{[q]}(k) = 0$.  

99
Proof: Suppose there exists $h > k_p$ such that $Z_p(h) \neq 0$. Let $Z_p'$ the sequence equal to $Z_p$ everywhere except at $k = h$ where $Z_p'(h) = 0$ and call $Y_p' = H[Z_p']$. We have $\|Z_p'\| < \|Z_p\|$ which implies $\phi(Y_p') < \phi(Y_p)$. Since $H$ is causal, $Y_p'$ will differ from $Y_p$ only for $k \geq h$. Therefore $Y_p' \in I(K_p)$ and $\phi(Y_p') \geq \phi(Y_p)$ which is impossible. We conclude that $\forall k > k_p$, $Z_p(k) = 0$. It will be also the case for its successive forward derivatives for $k > k_p$. The reasoning is the same for $Z_p'$. \hfill \square

Therefore, we have:

**Proposition A.5.9** $Z_K$ is uniquely defined by the column vector of size $r(n-1)$

$$
\Omega = \begin{bmatrix}
Z_0 \\
\vdots \\
Z_{r-1}
\end{bmatrix}.
$$

(A.16)

We propose an algorithm which computes $\Omega$ recursively in time, in the next section.

### A.5.3 Recursive computation of $\Omega$

Let us start with some definitions and properties. For $p = 1, \ldots, r$,

(i) $K_p = \{k_1, \ldots, k_p\}$ (by convention $K_0 = \emptyset$).

(ii) $\mathcal{E}(K) = \{Y_p \in \mathbb{R}^N \mid \forall i = 1, \ldots, p, \ Y_p(k_i) = b(k_i)\}$,

\[\mathcal{E}(K) = \{Y_p' \in \mathbb{R}^N \mid \forall i = 1, \ldots, p-1, \ Y_p'(k_i) = 0 \text{ and } Y_p'(k_p) = 1\}\].

(iii) $Y_p$ and $\overline{Y}_p$ are the signals minimizing $\phi$ subject to the sets of equality constraints $\mathcal{E}(K_p)$ and $\overline{\mathcal{E}}(K_p)$ respectively (it is obvious that $Y_0 = \mathcal{O}$).

(iv) $Z_p = Y_p^{(n)}$ and $\overline{Z}_p = \overline{Y}_p^{(n)}$.

(v) For $i = 1, \ldots, p$,

$$
Y_{p,i} = \begin{bmatrix}
Y_p^{(n-1)}(k_i) \\
\vdots \\
Y_p^{(1)}(k_i)
\end{bmatrix} \quad \text{and} \quad \overline{Y}_{p,i} = \begin{bmatrix}
\overline{Y}_p^{(n-1)}(k_i) \\
\vdots \\
\overline{Y}_p^{(1)}(k_i)
\end{bmatrix}.
$$

(A.17)

By convention $Y_{p,0} = \overline{Y}_{p,0} = \mathcal{O}_{n-1}$ ($(n-1)$ zero column vector).

(vi) For $i = 1, \ldots, p$,

$$
Z_{p,i} = \begin{bmatrix}
Z_p^{[n-2]}(k_i + 1) \\
\vdots \\
Z_p^{[0]}(k_i + 1)
\end{bmatrix} \quad \text{and} \quad \overline{Z}_{p,i} = \begin{bmatrix}
\overline{Z}_p^{[n-2]}(k_i + 1) \\
\vdots \\
\overline{Z}_p^{[0]}(k_i + 1)
\end{bmatrix}.
$$

(A.18)

By convention $Z_{p,N} = \overline{Z}_{p,N} = \mathcal{O}_{n-1}$ in the case where $k_r = N$.

(viii)

$$
\Omega_p = \begin{bmatrix}
Z_{p,0} \\
\vdots \\
Z_{p,p-1}
\end{bmatrix} \quad \text{and} \quad \overline{\Omega}_p = \begin{bmatrix}
\overline{Z}_{p,0} \\
\vdots \\
\overline{Z}_{p,p-1}
\end{bmatrix}.
$$

(A.19)
We propose an algorithm which computes \( \Omega_p \) recursively with \( p = 0, \ldots, r \). This will lead to the computation of \( \Omega = \Omega_r \). The recursion is obtained from the following propositions:

**Proposition A.5.10** \( \forall p = 1, \ldots, r, \ Y_p = Y_{p-1} + \delta_p Y_p \) where \( \delta_p = b(k_p) - Y_{p-1}(k_p) \).

Proof: Let us first consider the case \( \delta_p = 0 \). Then \( Y_{p-1} + \delta_p Y_p = Y_{p-1} \) which minimizes \( \phi \) subject to \( \mathcal{E}(K_{p-1}) \). But it also minimizes \( \phi \) subject to \( \mathcal{E}(K_p) \), since \( \mathcal{E}(K_p) \subseteq \mathcal{E}(K_{p-1}) \). Let us now consider the case \( \delta_p \neq 0 \). Let us write \( Y = Y_{p-1} + \delta_p Y_p \). By linearity of \( \phi \) we have \( \nabla_{\phi}[Y] = \nabla_{\phi}[Y_{p-1}] + \delta_p \nabla_{\phi}[Y_p] \). Let \( Y' \in \mathcal{E}(K_p) \). We have

\[
\left< \nabla_{\phi}[Y], Y' - Y \right>
= \left< \nabla_{\phi}[Y_{p-1}], (Y' - \delta_p Y_p) - Y_{p-1} \right> + \left< \delta_p \nabla_{\phi}[Y_p], (Y' - Y_{p-1}) - \delta_p Y_p \right>
\]

where \( A = Y' - \delta_p Y_p \) and \( B = \frac{\delta_p}{\delta_{p}}(Y' - Y_{p-1}) \). Let us show that \( A \in \mathcal{E}(K_{p-1}) \) and \( B \in \mathcal{E}(K_p) \). Since \( Y' \in \mathcal{E}(K_p) \) and \( Y_p \in \mathcal{E}(K_p) \), \( \forall i = 1, \ldots, p-1, Y'(k_i) = b(k_i) \) and \( Y_p(k_i) = 0 \), which implies that \( Y'(k_i) - \delta_p Y_p(k_i) = b(k_i) \). Therefore, \( A = Y' - \delta_p Y_p \). Since \( Y' \in \mathcal{E}(K_p) \) and \( Y_{p-1} \in \mathcal{E}(K_{p-1}) \), \( \forall i = 1, \ldots, p-1, Y'(k_i) = Y_{p-1}(k_i) = b(k_i) \) which implies that \( \frac{1}{\delta_p}(Y'(k_i) - Y_{p-1}(k_i)) = 0 \). But we also have \( Y'(k_p) = b(k_p) \) which implies that \( \frac{1}{\delta_p}(Y'(k_p) - Y_{p-1}(k_p)) = \frac{\delta_p}{\delta_p} = 1 \). Therefore, \( B \in \mathcal{E}(K_p) \). Applying Lemma 5.5.2 on \( S = \mathcal{E}(K_{p-1}) \) and \( S = \mathcal{E}(K_p) \), we derive that the first term and the second term of (A.20) are non-negative. This implies that \( \left< \nabla_{\phi}[Y], Y' - Y \right> \geq 0 \). Since this is true for any \( Y' \in \mathcal{E}(K_p) \) we derive that \( Y = Y_p \). \( \square \)

**Proposition A.5.11** \( \forall p = 1, \ldots, m, \forall k \notin K_p, Z_p^{[n]}(k) = \overline{Z}_p^{[n]}(k) = 0 \).

**Proposition A.5.12** \( \forall k = k_p + 1, \ldots, N, \forall q \geq 0, \ Z_p^{[q]}(k) = \overline{Z}_p^{[q]}(k) = 0 \).

**Proposition A.5.13** \( \forall p = 1, \ldots, m, Z_p^{[n-2]} \) and \( z_p^{[n-2]} \) are piecewise linear on intervals \( \{k_{i-1} + 1, \ldots, k_i + 1\} \) for \( 1 \leq i \leq p \), and on \( \{k_r + 1, \ldots, N\} \).

These three propositions are shown in a way similar to Propositions A.5.5, A.5.8 and A.5.6.

**Proposition A.5.14** \( \forall p = 1, \ldots, r, \)

\[
\Omega_p = \left[ \Omega_{p-1} \right]_{n-1} + \delta_p \overline{Y}_p \tag{A.21}
\]

\[
Y_{p,p} = Y_{p-1,p} + \delta_p \overline{Y}_{p,p} \tag{A.22}
\]

where \( \delta_p = b(k_p) - Y_{p-1}(k_p) \).
Proof: Equation (A.22) is an immediate consequence of Proposition A.5.10. As a second consequence, we have for \( p = 1, \ldots, r \), \( Z_p = Z_{p-1} + \delta_p Z_p \). Therefore,

\[
\forall i = 0, \ldots, p - 2, \quad Z_{p,i} = Z_{p-1,i} + \delta_p Z_{p,i}.
\] (A.23)

From Proposition A.5.12, we have \( Z_{p-1,p-1} = O_{n-1} \). Therefore, \( Z_{p,p-1} = O_{n-1} + \delta_p Z_{p,p-1} \). This is summarized in (A.21) \( \Box \)

Assume that we know how to calculate \( \Omega_p \) and \( Y_{p,p} \) for any \( p = 0, \ldots, r \). For Proposition to give a recursive algorithm, we need one more property.

**Proposition A.5.15** Suppose \( Y_{p,p} \) is known. For \( q = 0, \ldots, n - 1 \), the values of \( Y_p^{(q)}(k) \) for \( k = k_p + 1, \ldots, N \) are obtained \((n - q)^{th}\) order discrete-time integration of the zero signal starting from the initial conditions at \( k = k_p \) given by \([Y_p^{(n-1)}(k_p) \cdots Y_p^{(1)}(k_p)]^T = Y_{p,p} \) and \( Y_p^{(0)} = b(k_p) \).

Proof: This is simply based on the fact that \( Y_p^{(q)} \) is the \((n - q)^{th}\) order discrete-time integration of \( Z_p \), since \( Z_p = Y_p^{(n)} = (Y_p^{(q)})^{(n-q)} \) and that \( Z_p(k) = 0 \) for any \( k = k_p + 1, \ldots, N \) due to Proposition A.5.12 \( \Box \)

From this we conclude that \( Y_p(k_p) \) and \( Y_{p,p+1} \) can be obtained from \( Y_{p,p} \) and \( b(k_p) \) by discrete-time integration. This leads to the following algorithm for the computation of \( \Omega_p \):

**Algorithm A:**

(i) Initialization: \( Y_{0,0} = O_{n-1}, \Omega_0 = [ ] \) (empty matrix), \( p = 1 \).

(ii) Calculate \( \Omega_p \) and \( Y_{p,p} \).

(iii) From \( Y_{p-1,p-1} \), calculate \( Y_{p-1}(k_p) \) and \( Y_{p-1,p} \).

(iv) Calculate \( \delta_p = b(k_p) - Y_{p-1}(k_p) \).

(v) Calculate \( \Omega_p = [\Omega_{p-1} \ O_{n-1}] + \delta_p \Omega_p \) and \( Y_{p,p} = Y_{p-1,p} + \delta_p Y_{p,p} \).

(vi) If \( p < r \), increment \( p \) and go to (ii).

**A.5.4 Recursive computation of \( \Omega_p \) and \( Y_{p,p} \)**

The computation of \( \Omega_p \) and \( Y_{p,p} \) is based on the following proposition:

**Proposition A.5.16** The signals \( \overline{Y}_p \) and \( \overline{Z}_p \) verify the following system:

\[
\forall k = 1, \ldots, N, \quad \left\{ \begin{array}{ll}
(i) & \text{if } k = k_i, \quad \overline{Y}_p(k_i) = 0, \quad \text{for } i = 1, \ldots, p - 1 \\
(ii) & \text{if } k = k_p, \quad \overline{Y}_p(k_p) = 1 \\
(iii) & \text{if } k = k_p + 1, \quad \overline{Z}_p^q(k) = 0, \quad \text{for } q = 0, \ldots, n - 2 \\
(iv) & \text{for other } k, \quad \overline{Z}_p^q(k) = 0.
\end{array} \right.
\] (A.24)
Proof: Conditions (i) and (ii) come from the definition of \( \mathbf{Y}_p \), conditions (iii) and (iv) come from Propositions A.5.11 and A.5.12 respectively. □

The derivation of \( \mathbf{Q}_p \) and \( \mathbf{Y}_{p,p} \) is detailed in Section A.7. Their computation is based on the definition of four square matrices \( \mathbf{M}_1(l) \) to \( \mathbf{M}_4(l) \) of size \((n-1)\) and two column vectors \( \mathbf{N}_1(l) \) and \( \mathbf{N}_2(l) \) of size \((n-1)\), all functions of an integer \( l \), and constructed as follows:

\[
\forall p \geq 0, \quad \left\{ \begin{array}{ll}
S_p(k) = 0, & \text{for } k \leq 0 \\
S_p(k) = \frac{1}{p}k(k+1)...(k+p-1), & \text{for } k > 0
\end{array} \right.
\]  \hspace{1cm} (A.25)

\[
\mathbf{M}_{01}(l) = \begin{bmatrix}
S_0(l+1) & 0 & \ldots & 0 \\
S_1(l) & S_0(l+1) & 0 & \vdots \\
\vdots & \vdots & \ddots & \ddots \\
S_{n-2}(l-(n-3)) & \ldots & S_1(l) & S_0(l+1)
\end{bmatrix}
\]  \hspace{1cm} (A.26)

\[
\mathbf{M}_{02}(l) = \begin{bmatrix}
S_0(l) & 0 & \ldots & 0 \\
S_1(l) & S_0(l) & 0 & \vdots \\
\vdots & \vdots & \ddots & \ddots \\
S_{n-2}(l) & \ldots & S_1(l) & S_0(l)
\end{bmatrix}
\]  \hspace{1cm} (A.27)

\[
\mathbf{M}_{03}(l) = \begin{bmatrix}
S_{n-1}(l-(n-2)) & \ldots & S_2(l-1) & S_1(l) \\
S_{n-2}(l-(n-2)) & \ldots & S_3(l-1) & S_2(l) \\
\vdots & \vdots & \ddots & \ddots \\
S_{2n-3}(l-(n-2)) & \ldots & S_{n-1}(l-1) & S_{n-1}(l)
\end{bmatrix}
\]  \hspace{1cm} (A.28)

\[
\mathbf{M}_{04}(l) = \mathcal{O}^2_{n-1}
\]

\[
\mathbf{M}_0(l) = \begin{bmatrix}
\mathbf{M}_{01}(l) & \mathbf{M}_{04}(l) \\
\mathbf{M}_{03}(l) & \mathbf{M}_{02}(l)
\end{bmatrix}
\]  \hspace{1cm} (A.29)

\[
\mathbf{N}_{01}(l) = \begin{bmatrix}
S_1(l) \\
S_2(l-1) \\
\vdots \\
S_{n-1}(l-(n-2))
\end{bmatrix}
\]  \hspace{1cm} (A.30)

\[
\mathbf{N}_{02}(l) = \begin{bmatrix}
S_n(l-(n-1)) \\
S_{n+1}(l-(n-1)) \\
\vdots \\
S_{2n-2}(l-(n-1))
\end{bmatrix}
\]  \hspace{1cm} (A.31)

\[
\mathbf{N}_{03}(l) = \begin{bmatrix}
S_{2n-2}(l-(n-2)) \\
S_{n-1}(l-1) \\
\vdots \\
S_n(l)
\end{bmatrix}
\]  \hspace{1cm} (A.32)

\[
\mathbf{N}_{04}(l) = \begin{bmatrix}
S_{n-1}(l) \\
S_2(l) \\
\vdots \\
S_1(l)
\end{bmatrix}
\]  \hspace{1cm} (A.33)

\[
\begin{bmatrix}
\mathbf{N}_1(l) \\
\mathbf{N}_2(l)
\end{bmatrix} = \mathbf{N}(l) = \{S_{2n-1}(l-(n-1))\}^{-1} \begin{bmatrix}
\mathbf{N}_{01}(l) \\
\mathbf{N}_{02}(l)
\end{bmatrix}
\]  \hspace{1cm} (A.34)
In the particular case $n = 2$, $\mathcal{M}_j(l)$ and $\mathcal{N}_j(l)$ are simply scalars. Their direct expressions in terms of $l$ are:

\[
\begin{bmatrix}
\mathcal{M}_1(l) & \mathcal{M}_3(l) \\
\mathcal{M}_3(l) & \mathcal{M}_2(l)
\end{bmatrix}
= - \frac{1}{l^2 - 1} \begin{bmatrix}
2l^2 + 3l + 1 & 6l \\
\frac{1}{2}l(l^2 - 1) & 2l^2 - 3l + 1
\end{bmatrix}
\]  

(A.35)

and

\[
\begin{bmatrix}
\mathcal{N}_1(l) \\
\mathcal{N}_2(l)
\end{bmatrix}
= \frac{6}{l^2 - 1} \begin{bmatrix}
1 \\
\frac{1}{2}(l - 1)
\end{bmatrix}.
\]  

(A.36)

In the general case, we have the following proposition:

**Proposition A.5.17** Let $(\mathcal{R}_i)_{i \leq i \leq p}$ and $(\Lambda_i)_{i \leq i \leq p}$ be the matrix sequences recursively defined by

\[
\left\{ \begin{array}{l}
\mathcal{R}_0 = \mathcal{O}^2_{n-1} \\
\Lambda_0 = [ ]
\end{array} \right. \quad \text{and} \quad \forall i = 1, \ldots, p, \quad \left\{ \begin{array}{l}
\mathcal{T}_i = (\mathcal{M}_1(l_i) + \mathcal{M}_4(l_i)\mathcal{R}_{i-1})^{-1} \\
\mathcal{R}_i = (\mathcal{M}_3(l_i) + \mathcal{M}_2(l_i)\mathcal{R}_{i-1})\mathcal{T}_i \\
\Lambda_i = \left[ \begin{array}{c}
\Lambda_{i-1} \\
T^2_{i-1}
\end{array} \right] \mathcal{T}_i
\end{array} \right.
\]  

(A.37)

where the matrices $(\mathcal{M}_i(l))_{i \leq i \leq 4}$ are defined in (A.34) and $l_i = k_i - k_{i-1}$. Then $\overline{\Omega}_p$ and $\overline{\mathcal{Y}}_{p,p}$ can be expressed in terms of $\Lambda_p$ and $\mathcal{R}_p$ as follows:

\[
\overline{\Omega}_p = -\Lambda_p\mathcal{N}_1(l_p) \quad \text{and} \quad \overline{\mathcal{Y}}_{p,p} = -\mathcal{R}_p\mathcal{N}_1(l_p) + \mathcal{N}_2(l_p),
\]  

(A.38)

where the column vectors $\mathcal{N}_1(l)$ and $\mathcal{N}_2(l)$ are defined in (A.33).

This is proved in Section A.7. In the proof, we do not justify the invertibility of the matrix $\mathcal{M}_1(l_i) + \mathcal{M}_4(l_i)\mathcal{R}_{i-1}$. However, we observed numerically that this matrix is indeed invertible and that its inversion is well conditioned.

As an example, we show the details of the recursion in the case $n = 2$. The matrices $\mathcal{T}_p$, $\mathcal{R}_p$ and $\overline{\mathcal{Y}}_{p,p}$ are simply scalars and the matrices $\Lambda_p$ and $\Omega_p$ are column vectors of size $p$. Using the expressions of $\mathcal{M}_j(l)$ and $\mathcal{N}_j(l)$ from (A.35) and (A.36), we have:

\[
\left\{ \begin{array}{l}
\mathcal{R}_0 = 0 \\
\Lambda_0 = [ ]
\end{array} \right. \quad \text{and} \quad \forall i = 1, \ldots, p, \quad \left\{ \begin{array}{l}
\mathcal{T}_i = -(l_i^2 - 1)((2l_i^2 + 3l_i + 1) + (6l_i)\mathcal{R}_{i-1})^{-1} \\
\mathcal{R}_i = -(l_i - 1)\left(\frac{1}{2}l_i(l_i + 1) + (2l_i - 1)\mathcal{R}_{i-1}\right)\mathcal{T}_i \\
\Lambda_i = \left[ \begin{array}{c}
\Lambda_{i-1} \\
1
\end{array} \right] \mathcal{T}_i
\end{array} \right.
\]  

(A.39)

\[
\overline{\Omega}_p = -\frac{6}{l_p^2 - 1}\Lambda_p \quad \text{and} \quad \overline{\mathcal{Y}}_{p,p} = -\frac{6}{l_p^2 - 1}\left(\mathcal{R}_p - \frac{1}{2}(l_p - 1)\right).
\]  

(A.40)

104
A.5.5 Recursive construction of $K$ which satisfies condition (A.14)

Our recursive construction of $K$ is based on the following conjecture.

**Conjecture A.5.18** If $\forall i = 1, ..., r$, $k_i - k_{i-1} \geq n$, then for

$$\forall p = 1, ..., r, \forall i = 1, ..., p, \quad \text{sign} \left( Z_p^{[n]}(k_i) \right) = (-1)^{p-i}$$  \hspace{1cm} (A.41)

We discovered this property numerically for any order $n$. We have the proof for the case $n = 2$.

Proof in the case $n = 2$: From (A.39), since $l_i \geq 2$, one can verify that the scalars $\mathcal{T}_i$ and $\mathcal{R}_i$ are such that, for all $i = 1, ..., r$, $\mathcal{T}_i > 0$ and $\mathcal{R}_i < 0$. This implies that the coefficients of $\mathbf{A}_i$ have alternating signs with the constraint that the last coefficient be negative. From (A.40), this implies that the coefficients of $\mathbf{A}_p$ have alternating signs with the constraint that the last coefficient be positive.

According to (A.18) and (A.19) with $n = 2$, the coefficients of $\mathbf{Z}_p$ are the scalars $\mathbf{Z}_{p,i} = Z_p(k_i + 1)$ for $i = 1, ..., p - 1$. Therefore, $\text{sign}(Z_p(k_i + 1)) = (-1)^{p-i-1}$.

Since $n = 2$, Proposition A.5.13 implies that $\mathbf{Z}_p$ is piecewise linear on intervals $\{k_{i-1} + 1, ..., k_i + 1\}$ for $i = 1, ..., p$. Therefore, the sign of the slope of $\mathbf{Z}_p$ is $(-1)^{p-i-1}$. Since

$$Z_p^{[2]}(k) = (Z_p(k+2) - Z_p(k+1)) - (Z_p(k+1) - Z_p(k))$$

is the variation of slope of $\mathbf{Z}_p$ about $k+1$, we derive that $\forall i = 1, ..., p, \text{sign} \left( Z_p^{[2]}(k_i) \right) = (-1)^{p-i} \square$

For $\epsilon \in \{-1, 1\}$, $b_\epsilon(k)$ designates $b_\epsilon(k)$ when $\epsilon = 1$ and $b_\epsilon(k)$ when $\epsilon = -1$. As a consequence, we have the following proposition:

**Proposition A.5.19** Suppose that the following conditions are satisfied:

(i) $\forall i = 1, ..., r$, $k_i - k_{i-1} \geq n$.

(ii) $\exists \epsilon \in \{-1, 1\}, \forall p = 1, ..., r$, $b(k_p) = b_{(\epsilon-1)p}(k_p)$.

(iii) $\forall p = 1, ..., r$, $\text{sign}(\delta_p) = \epsilon(-1)^{p+1}$ where $\delta_p = b(k_p) - Y_{p-1}(k_p)$.

Then $Y_r$ verifies the condition (A.14) of Proposition A.5.3. Therefore, $Y_K = Y_r$ minimizes $\phi$ subject to the set of inequality constraints $\mathcal{I}(K)$.

Proof: From Proposition A.5.10 we have $Y_r = \sum_{p=1}^{r} \delta_p Y_p$ which implies $Z_r^{[n]} = \sum_{p=1}^{r} \delta_p Z_p^{[n]}$. Therefore $\bigvee_{p=1}^{r} Y_p(k_i) = Z_r^{[n]}(k_i) = \sum_{p=1}^{r} \delta_p Z_p^{[n]}(k_i)$, since $Z_r^{[n]}(k) = 0$ when $k \geq k_p + 1$. Each term of the summation have a sign equal to $\epsilon(-1)^{p+1}(-1)^{p-i} = \epsilon(-1)^{i+1}$. Since $b(k_i) = b_{(\epsilon-1)p}(k_i)$, one can check that condition (A.14) is satisfied.

The fact that $Y_r$ minimizes $\phi$ subject to $\mathcal{I}(K)$ is then a consequence of Proposition A.5.3 \square

We propose a recursive construction of $K_p$ so that the conditions (i), (ii) and (iii) of Proposition A.5.19 are satisfied.
Algorithm B:
(i) Initialization:

• \( k_0 = 0, \ K_0 = \emptyset, \ Y_0 = O, \ p = 1, \)
• choose \( k_1 \geq n \) such that \( 0 \notin [b_-(k_1), b_+(k_1)], \)
• take \( \epsilon = 1 \) if \( 0 \geq b_+(k_1) \) and \( \epsilon = -1 \) if \( 0 \leq b_-(k_1). \)
(ii) Define \( K_p = K_{p-1} \cup \{k_p\}, \ b(k_p) = b_0(k_p) \) and \( \delta_p = b(k_p) - Y_{p-1}(k_p). \)
(iii) Calculate the signal \( Y_p \) minimizing \( \phi \) subject to \( \mathcal{E}(K_p). \)
(iv) Change the sign of \( \epsilon. \)
(v) Choose \( k_{p+1} \geq k_p + n \) such that \( \text{sign}(b_0(k_{p+1}) - Y_p(k_{p+1})) = -\epsilon. \)
   Else \( r = p. \)

It can be easily verified that the elements \( K_p, \ b(k_p) \) and \( Y_p \) obtained from this algorithm satisfy the conditions (i), (iii) and (iii) of Proposition A.5.19. Therefore, the resulting signal \( Y_r \) necessarily minimizes \( \phi \) subject to the set of inequality constraints \( \mathcal{E}(K_r). \) Note that line (iv) amounts to:

(iv) Choose \( k_{p+1} \geq k_p + n \) such that \( Y_p(k_{p+1}) \geq b_+(k_{p+1}) \) if \( \epsilon = 1 \) and \( Y_p(k_{p+1}) \leq b_-(k_{p+1}) \) if \( \epsilon = -1. \)

An example of construction of \( K_p, b(k_p) \) and \( Y_p \) from the algorithm is shown in Figure A.3.

![Diagram](image)

Figure A.3: Example of construction of \( K_p, Y_p, \) (with \( 1 \leq p \leq 3 \)). The selected constraints are shown by dark arrows.

106
A.5.6 Complete algorithm

Combining Algorithm A, Algorithm B and the result of Proposition A.5.17, we have the following algorithm:

Algorithm C:
(i) Initialization:
   • \( k_0 = 0, K_0 = \emptyset, Y_0 = \mathcal{O}, \Omega_0 = [\ ], \mathcal{R}_0 = \mathcal{O}_{n-1}^2, \Lambda_0 = [\ ], p=1, \)
   • choose \( k_1 \geq n \) such that \( 0 \notin [b_-(k_1), b_+(k_1)] \),
   • take \( \epsilon = 1 \) if \( 0 \geq b_+(k_1) \) and \( \epsilon = -1 \) if \( 0 \leq b_-(k_1) \).
(ii) Define \( K_p = K_{p-1} \cup \{ k_p \} \) and \( l_p = k_p - k_{p-1} \).
(iii) Determine \( \mathcal{M}_1(l_p), \mathcal{M}_2(l_p), \mathcal{M}_3(l_p), \mathcal{M}_4(l_p), \mathcal{N}_1(l_p), \mathcal{N}_2(l_p) \), using (A.34) and (A.33).
(iv) Calculate
   \[
   \begin{align*}
   \mathcal{T}_p &= (\mathcal{M}_1(l_p) + \mathcal{M}_4(l_p)\mathcal{R}_{p-1})^{-1} \\
   \mathcal{R}_p &= (\mathcal{M}_3(l_p) + \mathcal{M}_2(l_p)\mathcal{R}_{p-1})\mathcal{T}_p \\
   \Lambda_p &= \left[ \begin{array}{c} \Lambda_{p-1} \\ \mathcal{T}_{n-1}^2 \end{array} \right] \mathcal{T}_p
   \end{align*}
   \]
   • \( \Omega_p = -\Lambda_p^T \mathcal{N}_1(l_p) \) and \( \mathcal{Y}_{p,p} = -\mathcal{R}_p \mathcal{N}_1(l_p) + \mathcal{N}_2(l_p) \).
(v) \( \delta_p = b_\epsilon(k_p) - Y_{p-1}(k_p) \).
(vi) From \( \mathcal{Y}_{p-1,p-1} \) calculate \( \mathcal{Y}_{p-1,p} \).
(vii) Calculate
   \[
   \begin{align*}
   \Omega_p &= \left[ \begin{array}{c} \Omega_{p-1} \\ \mathcal{O}_{n-1} \end{array} \right] + \delta_p \Omega_p \\
   \mathcal{Y}_{p,p} &= \mathcal{Y}_{p-1,p} + \delta_p \mathcal{Y}_{p,p}
   \end{align*}
   \]
(viii) Change the sign of \( \epsilon \).
(ix) From \( \mathcal{Y}_{p,p} \) calculate \( Y_p(k) \) for \( k > k_p \) (according to Proposition A.5.15).
(x) Choose \( k_{p+1} \geq k_p + n \) such that \( Y_p(k_{p+1}) \geq b_+(k_{p+1}) \) if \( \epsilon = 1 \) and \( Y_p(k_{p+1}) \leq b_-(k_{p+1}) \) if \( \epsilon = -1 \).
(xi) If \( k_{p+1} \) is found, increment \( p \) and go to (ii).
    Else \( r = p \).

This algorithm provides a set of indices \( K = K_r \) and a column vector \( \Omega = \Omega_r \), such that \( Y_K \) minimizes \( \phi \) subject to \( \mathcal{I}(K) \) and \( Z_K \) is obtained by integration of the elements of \( \Omega \) according to Propositions A.5.7 and A.5.9.

A.5.7 Algorithm C with limited buffer

The drawback of this algorithm is that it requires the memorization of matrices \( \Lambda_p \) and \( \Omega_p \) with a size growing with \( p \). However, for analytical reasons, we show that, in practice, this matrix can be truncated to a limited size, by keeping the bottom part.

From the relation (A.23) and the fact that \( \mathcal{Z}_{i,i} = \mathcal{O}_{n-1} \), it can be derived that for
\[ i = 0, \ldots, r - 1, \]
\[ \forall q = i + 1, \ldots, r, \quad \mathbf{Z}_{q,i} = \sum_{p=i+1}^{q} (\mathbf{Z}_{p,i} - \mathbf{Z}_{p-1,i}) + \mathbf{Z}_{r,i} = \sum_{p=i+1}^{q} \delta_p \mathbf{Z}_{p,i} \]

In particular, for the case \( q = r \), we have:
\[ \forall i = 0, \ldots, r - 1, \quad \mathbf{Z}_{r,i} = \sum_{p=i+1}^{r} \delta_p \mathbf{Z}_{p,i}. \quad (A.42) \]

We observed numerically that for a given \( i, \mathbf{Z}_{p,i} \) has an exponential decay when \( p \) increases. This can be easily proved in the case \( n = 2 \). From (A.39), (A.40) and the definition of \( \Omega_p \) in (A.19), one can derive that \( \mathbf{Z}_{p,i} = -\frac{\delta_p}{\delta_{p-1}} T_{i+1} \cdots T_p \). From (A.39), using the fact that \( \mathcal{R}_{i-1} > 0 \), it is easy to derive that \( |T_i| \leq \frac{1}{2} \). Since \( l_p \geq 2 \), we find that \( |\mathbf{Z}_{p,i}| \leq 2^{-(p-i+1)} \) which has an exponential decay.

As a consequence, it can be decided that a predetermined number \( i_0 \) of terms in the summation of (A.42) is enough to give a good approximation of \( \mathbf{Z}_{r,i} \)

\[ \mathbf{Z}_{r,i} \approx \sum_{p=i+1}^{i+i_0} \delta_p \mathbf{Z}_{p,i} = \mathbf{Z}_{i+i_0,i}. \quad (A.43) \]

In other words, every \( \mathbf{Z}_{p,i} \) included in the computation of \( \Omega_p \) (line (vii) of Algorithm C) should be output as soon as \( p \geq i + i_0 \) or \( i \leq p - i_0 \). This implies that the \((i_0 - 1)\) bottom subvectors \( \mathbf{Z}_{p,i} \) of \( \Omega_p \) only, that is \( [\mathbf{Z}_{p,(i_0-1)} \cdots \mathbf{Z}_{p,p-1}]^T \), will be needed and kept in memory for the next iteration step \( (p + 1) \). This requires a buffer of length \((i_0 - 1)(n - 1)\). Another consequence is that, at step \( p \), \( [\mathbf{Z}_{p,(i_0-1)} \cdots \mathbf{Z}_{p,p-1}]^T \) are the only values needed from line (vii) of Algorithm C. This means that, in line (iv), only the last \( i_0(n-1) \) rows of \( \mathbf{A}_p \) need to be known. Therefore, this requires the memorization of the last \((i_0 - 1)(n - 1)\) rows of \( \mathbf{A}_{p-1} \) from the previous step \( (p - 1) \) for the calculation of line (iv) at step \( p \). Since the rows of \( \mathbf{A}_{p-1} \) have length \((n - 1)\), this requires a buffer of length \((i_0 - 1)(n - 1) \times (n - 1)\). In total, we need a buffer of size \((i_0 - 1)n(n - 1)\).

Conversely, if we have a buffer with a size limited to \( s \), then the number of terms in the approximation (A.43) will be the integer part of \( \frac{s}{n(n - 1)} + 1 \).

### A.6 Proof of Proposition A.5.4

Let us define for \( k \in \{0, \ldots, n\} \), \( D_n(k) = (-1)^k \left( \begin{array}{c} n \\ k \end{array} \right) \) and for \( k \notin \{0, \ldots, n\} \), \( D_n(k) = 0 \). From the classical property \( \left( \begin{array}{c} n + 1 \\ k \end{array} \right) = \left( \begin{array}{c} n \\ k \end{array} \right) + \left( \begin{array}{c} n \\ k - 1 \end{array} \right) \) valid for \( k = 1, \ldots, n + 1 \), we derive that
\[ D_{n+1}(k) = D_n(k) - D_n(k - 1). \quad (A.44) \]
One can check that (A.44) is also valid for any $k \in \mathbf{Z}$. We show that $\forall \mathbf{Y}, \mathbf{Z} \in \mathbb{R}^N$,

$$\forall j = 1, \ldots, N, \quad Y^{(n)}(j) = \sum_{i=1}^{N} D_n(j - i)Y(i), \quad \text{(A.45)}$$

$$\forall k = 1, \ldots, N, \quad Z^{[n]}(k) = (-1)^n \sum_{i=1}^{N} D_n(i - k)Z(i). \quad \text{(A.46)}$$

In general, from (2.3) and (2.4), we have, $\forall n \geq 0$, $Y^{(n+1)}(1) = Y^{(n)}(1) = Z^{[n+1]}(N) = Z^{[n]}(N)$ which implies that $\forall n \geq 0$, $Y^{(n)}(1) = Y(1)$ and $Z^{[n]}(N) = Z(N)$. One can check that this already proves that (A.45) is true for $j = 1$ and (A.46) for $k = N$. Now let us prove (A.45) and (A.46) for the other time indices by induction on $n$. For $n = 0$, this is trivial. Suppose this has been proved for a certain $n \geq 0$. For $j = 2, \ldots, N$, we have

$$Y^{(n+1)}(j) = Y^{(n)}(j) - Y^{(n)}(j - 1)$$

$$= \sum_{i=0}^{N} (D_n(j - i) - D_n(j - 1 - i))Y(i)$$

$$= \sum_{i=1}^{N} D_n+1(j - i)Y(i).$$

For $k = 1, \ldots, N - 1$, we have

$$Z^{[n+1]}(k) = Z^{[n]}(k + 1) - Z^{[n]}(k)$$

$$= (-1)^n \sum_{i=0}^{N} (D_n(i - k - 1) - D_n(i - k))Z(i)$$

$$= (-1)^{n+1} \sum_{i=1}^{N} D_{n+1}(i - k)Z(i).$$

Therefore, (A.45) and (A.46) are true at $n + 1$. The induction is completed.

When $H$ is an $n^{th}$ order integrator, we have $\phi(\mathbf{Y}) = \frac{1}{2} \sum_{j=1}^{N} |Y^{(n)}(j)|^2$. Therefore, for $k = 1, \ldots, N$,

$$\nabla_\phi[\mathbf{Y}](k) = \frac{1}{2} \sum_{j=1}^{N} \frac{\partial |Y^{(n)}(j)|^2}{\partial Y(k)} = \sum_{j=1}^{N} \frac{\partial Y^{(n)}(j)}{\partial Y(k)} Y^{(n)}(j) \quad \text{(A.47)}$$

Using (A.45), we have $\frac{\partial Y^{(n)}(j)}{\partial Y(k)} = D_n(j - k)$. Consider $\mathbf{Z} = Y^{(n)}$. Then, (A.46) can be written as $\nabla_\phi[\mathbf{Y}](k) = \sum_{j=1}^{N} D_n(j - k)Z(j)$. Using (A.47), we find $\nabla_\phi[\mathbf{Y}](k) = (-1)^n Z^{[n]}(k) \quad \Box$

109
A.7 Proof of Proposition A.5.17

Notation A.7.1 \( \mathcal{P} \) and \( \mathcal{Q} \) are the two \((n-1) \times (2n-2)\) matrices \( \mathcal{P} = [I_{n-1}^2 \, O_{n-1}^2] \) and \( \mathcal{Q} = [O_{n-1}^2 \, I_{n-1}^2] \), where \( I_{n-1}^2 \) and \( O_{n-1}^2 \) are respectively the identity and null square matrices of size \((n-1)\).

Notation A.7.2 For any sequence \( y \in \mathbb{R}^N \), we define the column vector
\[
U_y(k) = \left[ z^{[n-2]}(k+1) \cdots z^{[0]}(k+1) \, y^{(n-1)}(k) \cdots y^{(1)}(k) \right]^T
\]
where \( k \in \mathbb{N} \) and \( z = y^{(n)} \) (by convention for any \( q \geq 0 \), \( y^{(q)}(0) = 0 \)).

Note that, when taking \( y = Y_p \),
\[
\forall i = 0, \ldots, r, \quad U_{Y_p}(i) = \begin{bmatrix} z_{p,i} \\ \overline{Y}_{p,i} \end{bmatrix}.
\]

Using Notation A.7.1, we can write
\[
\overline{Z}_{p,i} = \mathcal{P} U_{Y_p}(i)
\]
and
\[
\overline{Y}_{p,i} = \mathcal{Q} U_{Y_p}(i)
\]
but also
\[
U_{Y_p}(i) = \mathcal{P}^T \overline{Z}_{p,i} + \mathcal{Q}^T \overline{Y}_{p,i}.
\]

In the particular case \( i = 0 \) \((k_0 = 0)\), we have
\[
U_{Y_p}(0) = \mathcal{P}^T \overline{Z}_{p,0}.
\]

We have the following preliminary proposition:

Proposition A.7.3 Let \( k \geq 1, \, l \geq n \) be two integers, and \( y \in \mathbb{R}^N \) such that \( \forall h = 1, \ldots, l, \, z^{[n]}(k+h) = 0 \) where \( z = y^{(n)} \). Then \( U_y(k+l) \) is uniquely determined by \( U_y(k), \, y(k) \) and \( y(k+l) \) as follows:
\[
U_y(k+l) = \mathcal{M}(l) U_y(k) + \mathcal{N}(l)(y(k+l) - y(k)),
\]
where \( \mathcal{N}(l) \) and \( \mathcal{M}(l) \) are respectively defined in (A.33) and (A.34).

This is proved in Appendix A.8.

Proposition A.7.4 Let \((S_i')_{0 \leq i \leq p}\) and \((\Lambda'_i)_{0 \leq i \leq p}\) be the matrix sequences recursively defined by
\[
\begin{cases}
S'_0 = \mathcal{P}^T \\
\Lambda'_0 = [ ]
\end{cases}
\quad \text{and} \quad \forall i = 1, \ldots, p, \quad \begin{cases}
S'_i = \mathcal{M}(l_i) S'_{i-1} \\
\Lambda'_i = [ \Lambda'_{i-1} ]
\end{cases}
\]

Then \( \overline{\Omega}_p \) and \( \overline{Y}_{p,p} \) can be expressed in terms of \( \Lambda'_p \) and \( S'_p \) as
\[
\overline{\Omega}_p = -\Lambda'_p (\mathcal{P} S'_p)^{-1} \mathcal{P} \mathcal{N}(l_p) \quad \text{and} \quad \overline{Y}_{p,p} = - (\mathcal{Q} S'_p)(\mathcal{P} S'_p)^{-1} \mathcal{P} \mathcal{N}(l_p) + \mathcal{Q} \mathcal{N}(l_p).
\]
Proof: This is based on the properties (i) to (iv) of Proposition A.5.16 verified by \( \mathbf{Y}_p \) and \( \mathbf{Z}_p \). Because of (iv), we have for every \( i = 1, \ldots, p, \)
\[
\forall l = 1, \ldots, l - 1, \quad \mathbf{Z}_p^{[l]} (k_{l+1} + l) = 0
\]
Applying Proposition A.7.3 on \( y = \mathbf{Y}_p \) with \( k = k_{l+1}, l = l \), we find that
\[
\mathcal{U}_{\mathbf{Y}_p}(k_l) = \mathcal{M}(l) \mathcal{U}_{\mathbf{Y}_p}(k_{l-1}) + \mathcal{N}(l)(\mathbf{Y}_p(k_l) - \mathbf{Y}_p(k_{l-1}))
\]
since \( k_l = k_{l-1} + l \). Applying the properties (i) and (ii) of Proposition A.5.16, we have
\[
\begin{cases}
\mathcal{U}_{\mathbf{Y}_p}(k_l) = \mathcal{M}(l) \mathcal{U}_{\mathbf{Y}_p}(k_{l-1}) & \text{for } i = 1, \ldots, p - 1 \\
\mathcal{U}_{\mathbf{Y}_p}(k_p) = \mathcal{M}(l_p) \mathcal{U}_{\mathbf{Y}_p}(k_{p-1}) + \mathcal{N}(l_p)
\end{cases}
\]
Defining \( \mathcal{P}_i = \mathcal{M}(l_i) \mathcal{M}(l_{i-1}) \ldots \mathcal{M}(l_1) \) for \( i = 1, \ldots, p \) and \( \mathcal{P}_0 = \mathcal{I}^2_{n-1} \), we obtain
\[
\mathcal{U}_{\mathbf{Y}_p}(k_l) = \mathcal{P}_i \mathcal{U}_{\mathbf{Y}_p}(0) \quad \text{for } i = 0, \ldots, p - 1 \quad (A.55)
\]
\[
\mathcal{U}_{\mathbf{Y}_p}(k_p) = \mathcal{P}_p \mathcal{U}_{\mathbf{Y}_p}(0) + \mathcal{N}(l_p) \quad (A.56)
\]
Using (A.48) and (A.51), (A.55) and (A.56) become
\[
\mathbf{Z}_{p,i} = \mathcal{P}_i \mathcal{P}_p \mathcal{P}_p^T \mathbf{Z}_{p,0} \quad \text{for } i = 0, \ldots, p - 1 \quad (A.57)
\]
\[
\mathbf{Z}_{p,p} = \mathcal{P}_p \mathcal{P}_p^T \mathbf{Z}_{p,0} + \mathcal{P} \mathcal{N}(l_p). \quad (A.58)
\]
Using (A.49) and (A.51), (A.56) becomes
\[
\mathbf{Y}_{p,p} = \mathcal{Q} \mathcal{P}_p \mathcal{P}_p^T \mathbf{Z}_{p,0} + \mathcal{Q} \mathcal{N}(l_p) \quad (A.59)
\]
From (A.53), it is easy to see to show by induction that
\[
\mathbf{S}_p = \mathcal{P}_p \mathcal{P}_p^T \quad \text{and} \quad \mathbf{A}' = \begin{bmatrix} \mathcal{P} \mathcal{P}_0 \mathcal{P}_p^T \\ \vdots \\ \mathcal{P} \mathcal{P}_{p-1} \mathcal{P}_p^T \end{bmatrix}.
\]
Therefore, (A.57) and (A.59) lead to
\[
\overline{\mathbf{O}}_p = \begin{bmatrix} \mathbf{Z}_{p,0} \\ \vdots \\ \mathbf{Z}_{p,p-1} \end{bmatrix} = \mathbf{A}' \mathbf{Z}_{p,0} \quad \text{and} \quad \mathbf{Y}_{p,p} = \mathcal{Q} \mathbf{S}_p' \mathbf{Z}_{p,0} + \mathcal{Q} \mathcal{N}(l_p). \quad (A.60)
\]
Finally, property (iii) of Proposition A.5.16 implies that \( \mathbf{Z}_{p,p} = \mathcal{O}_{n-1} \). Therefore (A.58) implies that
\[
\mathbf{Z}_{p,0} = -(\mathcal{P} \mathcal{P}_p \mathcal{P}_p^T)^{-1} \mathcal{P} \mathcal{N}(l_p) = -(\mathcal{P} \mathbf{S}_p')^{-1} \mathcal{P} \mathcal{N}(l_p). \quad (A.61)
\]
The proof is completed by replacing this value of \( \mathbf{Z}_{p,0} \) in (A.60) \( \square \)
The inversion of \( \mathcal{P} \mathbf{S}_p' \) in relation (A.54) can lead to numerical difficulties as \( \mathbf{S}_p' \) is a matrix growing with \( p \) (see relation (A.53)). This problem can be solved with the following proposition:
Proposition A.7.5 Formula (A.54) is still true when replacing \((S_i)_{s \leq i \leq p}\) and \((\Lambda_i)_{s \leq i \leq p}\) by the matrix sequences \((S_i)_{s \leq i \leq p}\) and \((\Lambda_i)_{s \leq i \leq p}\) recursively defined by

\[
\begin{align*}
S_0 &= \mathcal{P}^T \\
\Lambda_0 &= [ \ ] \\
\forall i &= 1, \ldots, p, \quad \begin{cases} 
S_i = \mathcal{M}(l_i)S_{i-1}T_i \\
\Lambda_i = \begin{bmatrix} \Lambda_{i-1} \\
\mathcal{P}S_{i-1} \end{bmatrix} T_i
\end{cases}
\end{align*}
\]  

(A.62)

where \((T_i)_{s \leq i \leq p}\) is an arbitrary sequence of invertible matrices of size \((n - 1)\).

Proof: For every \(p \geq 0\), we define \(V_p = T_1 T_2 \ldots T_p\) \((V_0 = I_{n-1})\). Let us show by induction the proposition

\[
P(p) : \text{"} \Lambda_p = \Lambda'_p V_p \text{ and } S_p = S'_p V_p.\text{"}
\]

We have: \(\forall p \geq 0\), \(V_p T_{p+1} = V_{p+1}\).

\(P(0)\) is true because \(V_0 = I_{n-1}^2\), \(\Lambda'_0 = \Lambda_0 = [ \ ]\) and \(S'_0 = S_0 = \mathcal{P}^T\).

Now, suppose that \(P(p)\) has been proved for some \(p \geq 1\). Then

\[
\Lambda_{p+1} = \begin{bmatrix} \Lambda_p \\
\mathcal{P}S_p \end{bmatrix} T_{p+1} = \begin{bmatrix} \Lambda'_p V_p \\
\mathcal{P}S'_p V_p \end{bmatrix} T_{p+1} = \begin{bmatrix} \Lambda'_p \\
\mathcal{P}S'_p \end{bmatrix} V_{p+1} = \Lambda'_{p+1} V_{p+1}
\]

and

\[
S_{p+1} = \mathcal{M}(K_p)S_p T_{p+1} = \mathcal{M}(K_p)S'_p V_p T_{p+1} = \mathcal{M}(K_p)S'_p V_{p+1} = S'_{p+1} V_{p+1}
\]

This implies that \(P(p+1)\) is true. The induction is completed \(\square\)

To prevent \(S_p\) from growing, we propose to choose at step \(p\) the matrix \(T_p = (\mathcal{P} \mathcal{M}(l_p)S_{p-1})^{-1}\). This is how Proposition A.5.17 is derived. We prove it here in a formal way.

**Proof of Proposition A.5.17:**

Let us start with the matrix sequences \((R_i)_{s \leq i \leq p}\), \((\Lambda_i)_{s \leq i \leq p}\) and \((T_i)_{s \leq i \leq p}\) defined in (A.53). Let us define

\[
S_i = \begin{bmatrix} T_{n-1}^i \\
R_{i} \end{bmatrix} \quad \text{for } i = 0, \ldots, p.
\]  

(A.63)

Note that

\[
\mathcal{P}S_i = T_{n-1}^i \quad \text{and} \quad QS_i = R_i.
\]  

(A.64)

Let us show that \((S_i)_{s \leq i \leq p}\), \((\Lambda_i)_{s \leq i \leq p}\) and \((T_i)_{s \leq i \leq p}\) verify (A.62). We already have \(\Lambda_0 = [ \ ]\) and \(S_0 = \mathcal{P}^T\) since \(R_0 = O_{n-1}^2\). Let \(i \in \{1, \ldots, p\}\). Using (A.63) at \(i - 1\) we can see from (A.37) that

\[
T_{i}^{-1} = \begin{bmatrix} \mathcal{M}_1(l_i) \mathcal{M}_3(l_i) \end{bmatrix} S_{i-1} = \mathcal{P} \mathcal{M}(l_i) S_{i-1}
\]

\[
R_{i} = \begin{bmatrix} \mathcal{M}_3(l_i) \mathcal{M}_2(l_i) \end{bmatrix} S_{i-1} T_i = Q \mathcal{M}(l_i) S_{i-1} T_i.
\]
Therefore

\[ S_i = \left[ \begin{array}{c} \mathcal{I}^2_{n-1} \\ \mathcal{R}_i \end{array} \right] = \left[ \begin{array}{c} \mathcal{P}\mathcal{M}(l_i)S_{i-1}T_i \\ \mathcal{Q}\mathcal{M}(l_i)S_{i-1}T_i \end{array} \right] = \mathcal{M}(l_i)S_{i-1}T_i. \]

Also, because of (A.64), \( \Lambda_i = \left[ \begin{array}{c} \Lambda_{i-1} \\ \mathcal{I}^2_{n-1} \end{array} \right] T_i. \) Therefore (A.62) is verified. Proposition A.7.5 can then be applied. We find

\[ \overline{\Omega}_p = -\Lambda_p(\mathcal{P}\mathcal{S}_p)^{-1}\mathcal{P}\mathcal{N}(l_p) \quad \text{and} \quad \overline{\Psi}_{p,p} = -(\mathcal{Q}\mathcal{S}_p)(\mathcal{P}\mathcal{S}_p)^{-1}\mathcal{P}\mathcal{N}(l_p) + \mathcal{Q}\mathcal{N}(l_p). \]

Using (A.33) and (A.64), this gives

\[ \overline{\Omega}_p = -\Lambda_p\mathcal{N}_1(l_p) \quad \text{and} \quad \overline{\Psi}_{p,p} = -\mathcal{R}_p\mathcal{N}_1(l_p) + \mathcal{N}_2(l_p) \]

\[ \square \]

### A.8 Proof of Proposition A.7.3

Let \( k \geq 1, l \geq n \) be two integers, and \( y \in \mathbb{R}^N \) such that \( \forall h = 1, ..., l, \ z^{[n]}(k+h) = 0 \) where \( z = y^{(n)} \). The proof of Proposition A.7.3 is based on Propositions A8.1 to A8.4.

**Proposition A8.1**

\[ \forall p \geq 0 \ , \forall k \geq 1, \ S_{p+1}(k) = \sum_{l=1}^{k} S_p(l). \quad \text{(A.65)} \]

Proof: This is obtained by induction with \( p \), using the definition of \( S_p(k) \) in (A.25) \( \square \)

**Proposition A8.2** For \( p \geq 0 \), and \( h_1, h_2 \) two integers such that \( 0 \leq h_1 \leq h_2 \), then

\[ \sum_{r=1}^{h_2} S_p(r-h_1) = S_{p+1}(h_2-h_1) \]

Proof: With the change of variable \( r' = r - h_1 \), using the fact that \( 1 - h_1 \leq 1 \) and definition A8.1, we find

\[ \sum_{r=1}^{h_2} S_p(r-h_1) = \sum_{r'=1}^{h_2-h_1} S_p(r') = \sum_{r'=1}^{h_2-h_1} S_p(r') = S_{p+1}(h_2-h_1) \]

\[ \square \]

**Proposition A8.3** \( \forall q \in \{1, ..., n\}, \ \forall h = 1, ..., l + q - 1, \)

\[ z^{[n-q]}(k+h) = \sum_{j=0}^{q-1} S_j(h-j) z^{[n-q+j]}(k+1) \quad \text{(A.66)} \]
Proof: We will use the fact that
\[ 
\forall j \geq 0, \forall h \geq 1, \quad z^{[j]}(k + h) = z^{[j]}(k + 1) + \sum_{r=1}^{h-1} z^{[j+1]}(k + r) \tag{A.67} 
\]
which is a consequence of the relation \( z^{[j+1]}(h) = z^{[j]}(h + 1) - z^{[j]}(h) \). We are going to prove (A.66) by induction with \( q \in \{0, \ldots, n\} \).

By assumption, \( \forall h = 1, \ldots, l - 1, \quad z^{[n]}(k + h) = 0 \). Applying (A.67) with \( j = n - 1 \), we find that \( \forall h = 1, \ldots, l, \)
\[
 z^{[n-1]}(k + h) = z^{[n-1]}(k + 1) = S_0(h)z^{[n-1]}(k + 1)
\]
since \( S_0(h) = 1 \). This proves (A.66) for \( q = 1 \).

Now let us assume that (A.66) is true for a certain \( q \in \{1, \ldots, n - 1\} \). Let us consider \( h \in \{1, \ldots, l + q\} \). Applying (A.67) with \( j = n - q - 1 \) we find
\[
 z^{[n-q-1]}(k + h) = z^{[n-q-1]}(k + 1) + \sum_{r=1}^{h-1} z^{[n-q]}(k + r) \tag{A.68}
\]
Using the assumption that (A.66) is true at \( q \), we find that
\[
 \sum_{r=1}^{h-1} z^{[n-q]}(k + r) = \sum_{r=1}^{h-1} \left( \sum_{j=0}^{q-1} S_j(r - j)z^{[n-q+j]}(k + 1) \right)
 = \sum_{j=0}^{q-1} \left( \sum_{r=1}^{h-1} S_j(r - j) \right) z^{[n-q+j]}(k + 1)
\]
Using Proposition A.8.2 and the fact that \( S_0(h) = 1 \), (A.68) becomes
\[
 z^{[n-q-1]}(k + h) = S_0(h)z^{[n-q-1]}(k + 1) + \sum_{j=0}^{q-1} S_{j+1}(h - 1 - j)z^{[n-q+j]}(k + 1)
\]
With the change of variable \( j' = j + 1 \) and inserting the first term into the sum, we find that, for \( h = 1, \ldots, l + (q + 1) - 1 \)
\[
 z^{[n-(q+1)]}(k + h) = \sum_{j'=0}^{(q+1)-1} S_{j'}(h - j')z^{[n-(q+1)+j']}\tag{A.69}
\]
which proves (A.66) for \( (q + 1) \). The induction is completed \( \Box \)

**Proposition A.8.4** \( \forall q \in \{1, \ldots, n\}, \ \forall h = 1, \ldots, l, \)
\[
 y^{(n-q)}(k + h) = \sum_{j=0}^{q-1} S_j(h) y^{(n-q+j)}(k) + \sum_{j=0}^{n-1} S_q+j(h - j)z^{[j]}(k + 1) \tag{A.69}
\]

114
Proof: We will use the fact that
\[ \forall j \geq 0, h \geq 1, \quad y^{(j)}(k + h) = y^{(j)}(k) + \sum_{r=1}^{h} y^{(j+1)}(k + r) \quad (A.70) \]

which is consequence of the relation \( y^{[j+1]}(h) = y^{[j]}(h) - y^{[j]}(h - 1) \). Note that (A.70) is still true at \( k = 0 \), with the convention \( y^{(j)}(0) = 0 \). We are going to prove (A.69) by induction with \( q \in \{1, ..., n\} \).

Applying formula (A.66) of Proposition A.8.3 at \( q = n \), we have
\[ \forall h = 1, ..., l, \quad y(k + h) = \sum_{j=0}^{n-1} S_j(h - j)z^{[j]}(k + 1) \]

Using (A.70) with \( j = n - 1 \) and the fact that \( z = y^{(n)} \), we find that, for \( h \in \{1, ..., l\} \),
\[ y^{(n-1)}(k + h) = y^{(n-1)}(k) + \sum_{r=1}^{h} y(k + r) \]
\[ = y^{(n-1)}(k) + \sum_{r=1}^{h} \left( \sum_{j=0}^{n-1} S_j(r - j)z^{[j]}(k + 1) \right) \]
\[ = y^{(n-1)}(k) + \sum_{j=0}^{n-1} \left( \sum_{r=1}^{h} S_j(r - j) \right) z^{[j]}(k + 1) \quad (A.71) \]

Using Proposition A.8.2 and the fact that \( S_0(h) = 1 \) we have
\[ y^{(n-1)}(k + h) = S_0(h)y^{(n-1)}(k) + \sum_{j=0}^{n-1} S_{j+1}(h - j)z^{[j]}(k + 1) \]

This proves (A.69) for \( q = 1 \).

Now, let us suppose that (A.69) is true for a certain \( q \in \{1, ..., n - 1\} \). Let us consider \( h \in \{1, ..., l\} \). Applying (A.70) with \( j = n - q - 1 \), we find
\[ y^{(n-q-1)}(k + h) = y^{(n-q-1)}(k) + \sum_{r=1}^{h} y^{(n-q)}(k + r) \]

Using the assumption that Proposition A.8.4 is true at \( q \), we can write
\[ \sum_{r=1}^{h} y^{(n-q)}(k + r) = \sum_{r=1}^{h} \left( \sum_{j=0}^{n-q-1} S_j(r) y^{(n-q+j)}(k) + \sum_{j=0}^{n-1} S_{q+j}(r - j)z^{[j]}(k + 1) \right) \]
\[ = \sum_{j=0}^{n-1} \left( \sum_{r=1}^{h} S_j(r) \right) y^{(n-q+j)}(k) + \sum_{j=0}^{n-1} \left( \sum_{r=1}^{h} S_{q+j}(r - j) \right) z^{[j]}(k + 1) \]

115
Using Proposition A.8.2, performing the change of variable $j' = j + 1$ in the first summation, and using the fact that $S_0(h) = 1$, (A.71) becomes

\[ y^{(n-(q+1))}_0(k + h) = \sum_{j' = 0}^{(q+1)-1} S_{j'}(h) y^{(n-(q+1)+j')}_0(k) + \sum_{j = 0}^{n-1} S_{(q+1)+j}(h - j) z^{[l]}(k + 1) \]

This proves (A.69) for $(q + 1)$. The induction is completed $\square$

**Proof of Proposition A.7.3:**

Considering $q = 2, \ldots, n$, relation (A.66) of Proposition A.8.3 can be written in matrix notations:

\[ \forall h = 1, \ldots, l, \begin{bmatrix} z^{[n-2]}(k + h) \\ z^{[n-3]}(k + h) \\ \vdots \\ z^{[0]}(k + h) \end{bmatrix} = \mathbf{M}_{01}(h) \cdot \begin{bmatrix} z^{[n-2]}(k + 1) \\ z^{[n-3]}(k + 1) \\ \vdots \\ z^{[0]}(k + 1) \end{bmatrix} + \mathbf{N}_{01}(h) \cdot z^{[n-1]}(k + 1) \quad (A.72) \]

where $\mathbf{M}_{01}(h)$ is the $(n - 1)$ square matrix given in (A.26) and $\mathbf{N}_{01}(h)$ is the $(n - 1)$ column vector given in (A.31). Similarly, considering $q = 1, \ldots, n - 1$, relation (A.69) of Proposition A.8.4 can be written in matrix notations:

\[ \forall h = 1, \ldots, l, \begin{bmatrix} y^{(n-1)}(k + h) \\ y^{(n-2)}(k + h) \\ \vdots \\ y^{(1)}(k + h) \end{bmatrix} = \mathbf{M}_{02}(h) \cdot \begin{bmatrix} y^{(n-1)}(k) \\ y^{(n-2)}(k) \\ \vdots \\ y^{(1)}(k) \end{bmatrix} + \mathbf{M}_{03}(h) \cdot \begin{bmatrix} z^{[n-2]}(k + 1) \\ z^{[n-3]}(k + 1) \\ \vdots \\ z^{[0]}(k + 1) \end{bmatrix} + \mathbf{N}_{02}(h) \cdot z^{[n-1]}(k + 1) \quad (A.73) \]

where $\mathbf{M}_{02}(h)$ and $\mathbf{M}_{03}(h)$ are the $(n - 1)$ square matrices given in (A.27) and (A.28) respectively and $\mathbf{N}_{02}(h)$ is the $(n - 1)$ column vector given in (A.31). Using the notation $\mathbf{U}_y(k)$ (Notation A.7.1), the relations (A.72) and (A.73) can be summarized as:

\[ \forall h = 1, \ldots, l, \quad \mathbf{U}_y(k + h) = \mathbf{M}_0(h) \mathbf{U}_y(k) + \begin{bmatrix} \mathbf{N}_{01}(h) \\ \mathbf{N}_{02}(h) \end{bmatrix} z^{[n-1]}(k + 1) \quad (A.74) \]
where $\mathcal{M}_0(h)$ is the $(2n-2)$ square matrix given in (A.30). Relation (A.69) applied at $q = n$ can be written:
$$\forall h = 1, ..., l,$$
$$y(k+h) - [\mathcal{N}(h) \mathcal{N}_0(h)] U_y(k) + S_{2n-1}(h -(n-1)) z^{[n-1]}(k+1) \quad (A.75)$$
where $\mathcal{N}(h)$ and $\mathcal{N}_0(h)$ are the $(n-1)$ row vectors given in (A.32) and (A.32) respectively. When taking $h = l$, $S_{2n-1}(l -(n-1)) \neq 0$ since $l \geq n$. We can therefore eliminate $z^{[n-1]}(k+1)$ between (A.74) and (A.75) at $h = l$. We find
$$U_y(k + l) =$$
$$\left( \mathcal{M}_0(l) - \{S_{2n-1}(l -(n-1))\}^{-1} \begin{bmatrix} \mathcal{N}_{01}(l) \\ \mathcal{N}_{02}(l) \end{bmatrix} \begin{bmatrix} \mathcal{N}_0(l) \\ \mathcal{N}_0(l) \end{bmatrix} \right) U_y(k) +$$
$$\{S_{2n-1}(l -(n-1))\}^{-1} \begin{bmatrix} \mathcal{N}_{01}(l) \\ \mathcal{N}_{02}(l) \end{bmatrix} (y(k + l) - y(k)).$$

Using (A.34) and (A.33) this leads to (A.52) \hfill \square

### A.9 Non-overload sufficient condition in multi-loop $\Sigma \Delta$ modulation

We show that an $n$-bit $n$-loop $\Sigma \Delta$ modulator never overloads if the input $X$ belongs to the domain $[-\frac{q}{2}, \frac{q}{2}]^N$, where $q$ is the step size of the built-in quantizer. Let us first introduce the following notation:
$$\forall X \in \mathbb{R}^N, \quad |X| = \max_{1 \leq k \leq N} |X(k)|.$$ 

With this notation, we have:
$$|X| \leq B \iff \forall k = 1, ..., N, \quad |X(k)| \leq B \iff X \in [-B, B]^N.$$ 

The non-overload region of a quantizer was defined in Section 3.1. In the case of an $n$-bit uniform quantizer, the non-overload region is defined by $|X| \leq 2^{n-1} q$. Figure 1.2 shows the case $n = 2$. When $X$ belongs to the non-overload region of a uniform quantizer of size $q$, one can see that (Figure 1.2) that $|C - X| \leq \frac{q}{2}$, where $C$ is the quantized version of $X$. Note that the quantizer is an encoder which verifies $\exists B > 0, \exists \Delta > 0, \forall X \in \mathbb{R}^N,$
$$|X| \leq B \implies |C - X| \leq \Delta,$$ \hfill (A.76)

where $E = C - X$ is the quantization error signal. Explicitly, an $n$-bit uniform quantizer verifies (A.76) with $B = 2^{n-1} q$ and $\Delta = \frac{q}{2}$. We will see that more general...
encoders such as $\Sigma \Delta$ modulators verify a similar proposition but weaker: $\exists B > 0, \exists \delta > 0, \forall X \in \mathbb{R}^N$, 

$$\forall k_0 = 1, ..., N, \quad (\forall k = 1, ..., k_0, \ |X(k)| \leq B) \implies (\forall k = 1, ..., k_0, \ |C(k) - X(k)| \leq \Delta). \quad (A.77)$$

One can check that (A.76) implies (A.77). We will show that multi-loop $\Sigma \Delta$ modulators verify (A.77). The procedure is the following. One has to see that a multi-loop $\Sigma \Delta$ modulator has the structure of Figure A.4, similar to a single-loop modulator, where the quantizer is in general replaced by an encoder. This is

![Figure A.4: Block diagram derived from the single-loop $\Sigma \Delta$ modulator by replacing the quantizer by a built-in encoder.](image)

already true for a 1st order $\Sigma \Delta$ modulator. We saw in Section 1.2.4 that a double-loop $\Sigma \Delta$ modulator has the structure of Figure A.4 where the built-in encoder is itself a single-loop $\Sigma \Delta$ modulator. One can verify that, in general, an $m$-loop $\Sigma \Delta$ modulator has this structure where the built-in encoder is an $(m - 1)$-loop $\Sigma \Delta$ modulator. The sufficient condition for non-overloading will be derived from two propositions (Propositions A.9.1 and A.9.2) relative to the encoding structure of Figure A.4.

**Proposition A.9.1** In the structure of Figure A.4, assume that the built-in encoder verifies (A.77) with $B > \Delta > 0$. Then, the following proposition is true for all $k_0 \in \{1, ..., N\}$:

$$P(k_0): \ (\forall k = 1, ..., k_0, \ |X(k)| \leq B - \Delta) \implies (\forall k = 1, ..., k_0, \ |A(k)| \leq B),$$

where $A$ is the input of the built-in encoder.

Proof: Let us prove $P(k_0)$ by induction. Suppose that $|X(1)| \leq B - \Delta$. We have $|A(1)| = |X(1)| \leq B - \Delta \leq B$. Therefore $P(1)$ is true. Suppose now that $P(k_0)$ has been proved for some $k_0 \in \{1, ..., N - 1\}$. According to Figure A.4,

$$\forall k_0 = 1, ..., N, \quad A(k) = A(k - 1) - C(k - 1) + X(k), \quad (A.78)$$

118
with the convention \(A(0) = C(0) = 0\). Applying (A.78) at \(k = k_0 + 1\), we derive:

\[
|A(k_0 + 1)| \leq |C(k_0) - A(k_0)| + |X(k_0 + 1)|. \tag{A.79}
\]

Assume that \(\forall k = 1, \ldots, k_0 + 1, |X(k)| \leq B - \Delta\). Because \(P(k_0)\) is true, this implies that \(\forall k = 1, \ldots, k_0, |A(k)| \leq B\). Applying (A.77) to the built-in encoder, one consequence is that \(|C(k_0) - A(k_0)| \leq \Delta\). Since by assumption we also have \(|X(k_0 + 1)| \leq B - \Delta\), (A.79) implies that \(|A(k_0 + 1)| \leq B\). To summarize, we have \(\forall k = 1, \ldots, k_0 + 1, |A(k)| \leq B\). Thus, we have proved that \(P(k_0 + 1)\) is true. The induction is completed \(\Box\)

**Application 1:** An \(n\)-bit single-loop \(\Sigma\Delta\) modulator is the case where the built-in encoder is an \(n\)-bit uniform quantizer of step size \(q\). Applying Proposition A.9.1, \(|X| \leq 2^{n-1}q - \frac{q}{2}\) implies \(|A| \leq 2^{n-1}q\). Therefore, when \(|X| \leq 2^{n-1}q - \frac{q}{2}\), the quantizer never overloads.

**Application 2:** For the case of a single-loop \(\Sigma\Delta\) modulator, \(n = 1\) and \(|X| \leq \frac{q}{2}\) is a sufficient condition of non-overloading.

**Proposition A.9.2** Let be an encoder which has the structure of Figure A.4 such that the built-in encoder verifies (A.77) with the constants \(B\) and \(\Delta\). Then, the whole encoder also verifies (A.77) where the constants \(B\) and \(\Delta\) have to be replaced by \(B' = B - \Delta\) and \(\Delta' = 2\Delta\).

Proof: Using the signal notations of Figure A.4, we can always write:

\[
|C(k) - X(k)| \leq |C(k) - A(k)| + |A(k) - X(k)|.
\]

From (A.78) we have \(A(k) - C(k) = A(k - 1) - C(k - 1)\). Therefore,

\[
|C(k) - X(k)| \leq |C(k) - A(k)| + |C(k - 1) - A(k - 1)|. \tag{A.80}
\]

Let \(k_0 \in \{1, \ldots, N\}\). Suppose that \(\forall k = 1, \ldots, k_0, |X(k)| \leq B - \Delta\). According to \(P(k_0)\) from Proposition A.9.1, we have \(\forall k = 1, \ldots, k_0, |A(k)| \leq B\), which implies that \(\forall k = 1, \ldots, k_0, |C(k) - A(k)| \leq \Delta\), since the built-in encoder verifies (A.77). Using the convention \(A(0) = C(0) = 0\), (A.80) implies that \(\forall k = 1, \ldots, k_0, |C(k) - X(k)| \leq 2\Delta\) \(\Box\)

**Application 3:** An \(n\)-bit single-loop \(\Sigma\Delta\) modulator of step size \(q\) verifies (A.77) with \(B = 2^{n-1}q - \frac{q}{2}\) and \(\Delta = 2 \cdot \frac{q}{2}\) and \(\Delta = 2 \cdot \frac{q}{2}\).

Proof: The built-in encoder is an \(n\)-bit quantizer which verifies (A.76), and thus (A.77), with \(B = 2^{n-1}q\) and \(\Delta = \frac{q}{2}\). Then, apply Proposition A.9.1 \(\Box\)

**Application 4:** An \(n\)-bit \(m\)-loop \(\Sigma\Delta\) modulator of step size \(q\) with \(m \leq n\) verifies (A.77) with \(B = 2^{n-1}q - (2^m - 1)\frac{q}{2}\) and \(\Delta = 2^m \cdot \frac{q}{2}\).
Proof: Let us show this by induction on $m \in \{1, ..., n\}$. This is already true at $m = 1$, according to Application 3. Suppose this has been proved at some $m \in \{1, ..., n - 1\}$. To prove the case $m + 1$, consider an $n$-bit $(m + 1)$-loop $\Sigma \Delta$ modulator. It has the structure of Figure A.4 where the built-in encoder is an $n$-bit $m$-loop $\Sigma \Delta$ modulator. By assumption, this encoder verifies (A.77) with $B = 2^{n-1}q - (2^m - 1)\frac{q}{2}$ and $\Delta = 2^m \cdot \frac{q}{2}$. Applying Proposition A.9.2, we conclude that the $(m + 1)$-loop modulator verifies (A.77) with

$$B = \left(2^{n-1}q - (2^m - 1)\frac{q}{2}\right) - \left(2^m \cdot \frac{q}{2}\right) = 2^{n-1}q - (2^{m+1} - 1)\frac{q}{2} \text{ and } \Delta = 2^{m+1} \cdot \frac{q}{2}.$$  

The induction is completed  

**Application 5:** For an $n$-bit $m$-loop $\Sigma \Delta$ modulator of step size $q$ with $m \leq n$, $|X| \leq 2^{n-1}q - (2^m - 1)\frac{q}{2}$ is a sufficient condition of non-overloading.

Proof: Let us show this by induction on $m \in \{1, ..., n\}$. Application 1 already justifies the case $m = 1$. Suppose this has been proved for a certain $m \in \{1, ..., n - 1\}$. Let $X \in \mathbb{R}^n$ such that

$$|X| \leq 2^{n-1}q - (2^m - 1)\frac{q}{2}. \quad (A.81)$$

According to (A.78):

$$|A(k)| \leq |C(k - 1) - A(k - 1)| + |X(k)|. \quad (A.82)$$

Note that, at $k = 1$,

$$|A(1)| \leq |X(1)| \leq 2^{n-1}q - (2^m - 1)\frac{q}{2} \leq 2^{n-1}q - (2^m - 1)\frac{q}{2}.$$  

Suppose that for a certain $k_0 \in \{1, ..., n - 1\}$, we have proved that $\forall k = 1, ..., k_0$,

$$|A(k)| \leq 2^{n-1}q - (2^m - 1)\frac{q}{2}. \quad (A.83)$$

In the equivalent structure of Figure A.4 of an $(m + 1)$-loop $\Sigma \Delta$ modulator, $A$ is the input of an $m$-loop $\Sigma \Delta$ modulator. Applying Application 4, we derive that $\forall k = 1, ..., k_0$, $|C(k) - A(k)| \leq 2^m \cdot \frac{q}{2}$. Applying (A.82) at $k = k_0 + 1$, we obtain that

$$A(k_0 + 1) \leq 2^m \cdot \frac{q}{2} + 2^{n-1}q - (2^{m+1} - 1)\frac{q}{2} = 2^{n-1}q - (2^m - 1)\frac{q}{2}.$$  

This means that (A.83) is also true at $k = k_0 + 1$. By induction on $k_0$, this means that,

$$|A| \leq 2^{n-1}q - (2^m - 1)\frac{q}{2}. \quad (A.84)$$

120
According to the assumption that Application 5 is true at $m$, the relation (A.84) is a sufficient condition for the built-in encoder not to overload. This implies that (A.81) is a sufficient condition of non-overloading for the $(m + 1)$-loop $\Sigma\Delta$ modulator. Application 5 is therefore true at $m + 1$. The induction in $m$ is completed $\square$

**Application 6:** $X \in [-\frac{q}{2}, \frac{q}{2}]^N$ is a sufficient condition for non-overloading for an $n$-bit $n$-loop $\Sigma\Delta$ modulator.

Proof: Applying Application 5 to $m = n$, we find that a sufficient condition for non-overloading is $|X| \leq 2^{n-1}q - (2^n - 1)\frac{q}{2} = \frac{q}{2}$ $\square$
Bibliography


