UNIVERSIDAD DE BUENOS AIRES

Facultad de Ciencias Exactas y Naturales

Departamento de Matemática

Tesis de Licenciatura

CUANTIFICACIÓN SIGMA-DELTA ($\Sigma\Delta$)

Magali Anastasio

Director: Dr. Carlos Cabrelli

Marzo de 2007

Agradecimientos

Quiero agradecer a mi director, Carlos Cabrelli, por haberme guiado en este trabajo, por confiar en mí y tener que soportar mi lluvia de mails con actualizaciones.

A mi papá, a mi mamá y a mis hermanas, porque nunca me hicieron dudar de mis elecciones, porque siempre estuvieron para mí y sé que siempre van a estar.

A Damian, por entender mis locuras y conocerme mejor que nadie, por abrirme los ojos, por hacerme feliz.

A Santi y Benji, porque son hermosos y me encanta ser su tía. Al resto de mi familia: primos, cuñados, tíos y abuelos, por bancar mi ausencia en reuniones familiares en época de examen.

A mis amigas de toda la vida, Sole y Eli, por ser incondicionales, por escucharme y aconsejarme.

A todo el grupo de gente que conocí en la facu y que me brindaron su amistad: Pablis, Lau P., Edu, Lean L., JL, Martin, Ana, Caro, Eze, Lau B., Lean D., Pablo D. En especial a Vicky e Isa, porque me dieron fuerzas para hacer un montón de cosas y porque disfruté poder compartir todo con ellas.

Índice general

Capítulo 1

Introducción

Las señales digitales están definidas en un conjunto discreto y toman valores en un conjunto finito. Ejemplos de esta clase de señales son: los estados de un interruptor (encendido/apagado), la cantidad de individuos de una población a lo largo de los años, etc.
En la actualidad se maneja cada vez más frecuentemente información en formato digital, como por ejemplo cámaras de foto y video digitales, CDs, DVDs, teléfonos celulares, etc.

Sin embargo, la mayoría de la información generada por procesos naturales es analógica, es decir, que para su representación necesita un conjunto continuo, en tiempo o amplitud, de valores. Ejemplos de este tipo de información son las señales de audio, las señales de video, las mediciones de un sismógrafo o un termómetro, etc.

El manejo de datos en formato digital ofrece numerosas ventajas, como ser: la señal digital es más resistente al ruido y a las distorsiones, el procesamiento digital es menos costoso y más versátil que el analógico y a través del mismo se logra mayor precisión y rapidez en la transmisión y en el almacenamiento de la información, además en el formato digital se pueden implementar sistemas de corrección de errores.
La demanda de este tipo de representación de la información, hace que haya gran interés en métodos de conversión analógico a digital (notados como A/D).

El proceso de digitalización tiene básicamente dos pasos:
El primero es la búsqueda de la representación de la señal a través de un

conjunto discreto de coeficientes. Si la señal es s, se busca una familia de núcleos $\{\varphi_n\}_{n\in\Lambda}$, de forma que

$$s = \sum_{n\in\Lambda} s_n \varphi_n,$$

donde el conjunto Λ es a lo sumo numerable y $s_n \in \mathbb{R}$ ó $\mathbb{C}$. De esta manera s queda definida por la sucesión $\{s_n\}_{n\in\Lambda}$.
La descomposición de la señal se dice redundante si la elección de los coeficientes s_n no es única en la familia $\{\varphi_n\}_{n\in\Lambda}$, que se denomina *diccionario*.

Si bien en este primer paso se logra representar a s por un conjunto discreto de números, aún no se ha logrado la digitalización, ya que los valores de s_n están dentro de un conjunto continuo, pues son números reales o complejos.

El segundo paso en la digitalización es la *cuantificación* de los coeficientes s_n, es decir, se reduce el rango continuo de valores que pueden tomar estos coeficientes a un rango discreto (preferentemente finito) de valores prefijados. Básicamente se sustituye cada valor $s_n \in \mathbb{R}$ ó $\mathbb{C}$ por un valor $q_n \in \mathcal{A}$, donde $\mathcal{A}$ es un conjunto discreto de valores permitidos llamado *alfabeto*.
La cuantificación se denomina *uniforme* si se utilizan alfabetos cuyos elementos se encuentran distribuidos de forma equidistante.

Existen principalmente dos mecanismos de cuantificación uniforme:
La *cuantificación fina*: Se aproxima cada coeficiente s_n con gran precisión, es decir, se emplea un alfabeto con una numerosa cantidad de elementos.
Y la *cuantificación espaciada*: Se utiliza un alfabeto con pocos elementos, pero se explota la redundancia del diccionario.

Como resultado de los pasos mencionados anteriormente se obtiene una señal aproximante

$$\tilde{s} = \sum_{n\in\Lambda} q_n \varphi_n. \tag{1.1}$$

Este tipo de reconstrucción se denomina *lineal*, existen otras formas de construir la señal aproximante dados los coeficientes q_n, pero en este trabajo utilizaremos la expresión (1.1).

Una vez realizada la cuantificación, se procede a la *codificación*, que consiste en asignarle a cada elemento del alfabeto un código binario. En el caso en que el alfabeto tenga dos elementos, a uno de ellos se le asigna el valor 0 y al otro el 1. Esta cuantificación se denomina de 1-bit, ya que se necesita 1 bit de información para almacenar la aproximación de cada coeficiente s_n.

Si la cantidad de elementos del alfabeto es K, donde $K > 2$, la cuantificación es conocida con el nombre de multibit. En este caso la cantidad de bits que se necesitan por coeficiente es $\log_2(K)$.

La eficacia del sistema de digitalización se traduce en encontrar una adecuada descomposición de la señal s en coeficientes $\{s_n\}_{n\in\Lambda}$ y un mecanismo de construcción de la sucesión cuantificadora $\{q_n\}_{n\in\Lambda}$, de forma que se logre conjuntamente que el error al aproximar s por $\tilde{s}$ sea pequeño y que la aplicación de los procedimientos no resulte costosa.

En este trabajo estudiaremos los algoritmos de cuantificación *Sigma-Delta* (abreviado por $\Sigma\Delta$), que a partir de la cuantificación espaciada, logran una buena aproximación de la señal y son de simple implementación.

Los procesos de cuantificación $\Sigma\Delta$ fueron introducidos en los años 60 por Inose y Yasuda en su trabajo [IY]. Ellos usaron el nombre $Delta - Sigma$ para los esquemas, la denominación $Sigma - Delta$ comenzó a usarse en los años 70 por los ingenieros de AT&T.
En los años 80, Gray enunció resultados muy importantes respecto a estos cuantificadores en [GCW] y [Gr]. Pero fue a fines de los años noventa, que se le dió un marco teórico matemático a los algoritmos $Sigma - Delta$, gracias a Ingrid Daubechies y Ron DeVore [DDV].
Se comenzó utilizando estos esquemas para cuantificar funciones de banda limitada (funciones cuya transformada de Fourier tiene soporte compacto), que son utilizadas para modelar señales de audio. Luego se fue ampliando su campo de aplicación, hasta ser utilizados por ejemplo para cuantificar expansiones en marcos finitos [BPY]-[BPY06]. En la actualidad estos algoritmos despiertan gran interés en la comunidad matemática.

En el segundo capítulo de este trabajo, vamos a estudiar la digitalización de funciones de banda limitada. Para desarrollar el primer paso en la conversión A/D, comenzaremos con la teoría de muestreo y sobremuestreo. Para el segundo paso, estudiaremos los esquemas $\Sigma\Delta$ y desarrollaremos un método de cuantificación fina denominado Modulación por Impulsos Codificados (conocido por las siglas PCM, debido a su designación en inglés de Pulse Code Modulation), con el objetivo de comparar los distintos mecanismos de cuantificación.

En el tercer capítulo se aplicarán los algoritmos $\Sigma\Delta$ para la cuantificación de coeficientes en marcos finitos, ajustados y uniformes de vectores de $\mathbb{R}^d$. Allí

también estudiaremos el método de Modulación por Impulsos Codificados (PCM).

Capítulo 2

Cuantificación Sigma-Delta ($\Sigma\Delta$) de funciones de banda limitada

En este capítulo vamos a estudiar el proceso de digitalización, generado por los cuantificadores $\Sigma\Delta$, para funciones de banda limitada (las mismas se utilizan para modelar, por ejemplo, señales de audio).
Una función se dice de banda limitada si pertenece al espacio

$$B_\Omega = \left\{ f \in L^2(\mathbb{R}) \,/\, \operatorname{sop} \hat{f} \subseteq [-\Omega, \Omega] \right\},$$

donde $\Omega > 0$ y la trasformada de Fourier viene dada por:

$$\hat{f}(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(t) e^{-it\xi}\, dt.$$

El mecanismo de conversión analógico a digital se basa en dos pasos fundamentales: el muestreo y la cuantificación.
Cada una de esas etapas será estudiada con profundidad en este capítulo. Comenzaremos desarrollando algunas cuestiones referentes a la teoría de muestreo.

2.1. Muestreo

Debido a que las funciones de B_Ω están definidas en un conjunto continuo de valores (los números reales), para la digitalización de las mismas se busca

discretizar ese dominio mediante el *muestreo*.
Dada $f \in B_\Omega$, el procedimiento de muestreo consiste en evaluar f en una sucesión de elementos de su dominio, de forma que la función quede completamente representada por los valores que toma en esos puntos. Cada uno de los elementos de la sucesión de evaluaciones se denomina *muestra*.
El teorema clásico de muestreo establece que la función f queda completamente definida por su valor en las muestras $f(\frac{n\pi}{\Omega})$, ya que las mismas resultan ser (salvo constante multiplicativa) los coeficientes de f en una base ortonormal del espacio B_Ω. A continuación se incluye la demostración de dicho teorema.

Teorema 2.1.1. *Si $f \in B_\Omega$, entonces*

$$f(t) = \sum_{n\in\mathbb{Z}} f\Big(\frac{n\pi}{\Omega}\Big)\frac{sen(\Omega t - n\pi)}{(\Omega t - n\pi)},$$

con convergencia en $L^2(\mathbb{R})$ y uniforme sobre $\mathbb{R}$.

Demostración. Es sabido que cualquier función $g \in L^2[-\pi,\pi]$ puede ser representada por su serie de Fourier de la forma

$$g(x) = \sum_{n\in\mathbb{Z}} c_n e^{-inx}, \quad \text{con} \quad c_n = \frac{1}{2\pi}\int_{-\pi}^{\pi} g(y)e^{iny}\,dy.$$

Como $\hat{f}\big(\frac{x\Omega}{\pi}\big) \in L^2[-\pi,\pi]$, ya que $\hat{f}(x) \in L^2[-\Omega,\Omega]$, podemos decir que

$$\hat{f}\Big(\frac{x\Omega}{\pi}\Big) = \sum_{n\in\mathbb{Z}} c_n e^{-inx} \quad \text{para } |x| \le \pi, \tag{2.1}$$

con

$$\begin{aligned} c_n &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{f}\Big(\frac{y\Omega}{\pi}\Big)e^{iny}\,dy = \frac{1}{2\pi}\int_{-\Omega}^{\Omega} \hat{f}(z)e^{\frac{inz\pi}{\Omega}}\frac{\pi}{\Omega}\,dz \\ &= \frac{1}{2\Omega}\int_{-\Omega}^{\Omega} \hat{f}(z)e^{\frac{inz\pi}{\Omega}}\,dz. \end{aligned}$$

Si aplicamos la fórmula de inversión $\Big(f(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\hat{f}(t)e^{it\xi}\,dt\Big)$ en esta

última igualdad, queda que

$$c_n = \frac{1}{2\Omega}\int_{-\Omega}^{\Omega} \hat{f}(z)e^{\frac{inz\pi}{\Omega}}\,dz = \frac{1}{2\Omega}\int_{-\infty}^{+\infty} \hat{f}(z)e^{\frac{inz\pi}{\Omega}}\,dz$$
$$= \frac{1}{\Omega}\sqrt{\frac{\pi}{2}} f\Big(\frac{n\pi}{\Omega}\Big).$$

Tomando $\xi = \frac{x\Omega}{\pi}$ en (2.1), resulta que

$$\hat{f}(\xi) = \frac{1}{\Omega}\sqrt{\frac{\pi}{2}}\sum_{n\in\mathbb{Z}} f\Big(\frac{n\pi}{\Omega}\Big)e^{\frac{-in\xi\pi}{\Omega}} \quad \text{para } |\xi| \leq \Omega. \tag{2.2}$$

Como $\operatorname{sop}\hat{f} \subseteq [-\Omega, \Omega]$, vale que

$$\hat{f}(\xi) = \frac{1}{\Omega}\sqrt{\frac{\pi}{2}}\sum_{n\in\mathbb{Z}} f\Big(\frac{n\pi}{\Omega}\Big)e^{\frac{-in\xi\pi}{\Omega}}\chi_{[-\Omega,\Omega]}(\xi).$$

Aplicando la trasformada de Fourier inversa, y usando que la antitransformada de la función

$$\frac{1}{\Omega}\sqrt{\frac{\pi}{2}}\chi_{[-\Omega,\Omega]}(\xi)e^{\frac{-in\xi\pi}{\Omega}}$$

es

$$\frac{\operatorname{sen}(\Omega t - n\pi)}{(\Omega t - n\pi)},$$

se obtiene que

$$f(t) = \sum_{n\in\mathbb{Z}} f\Big(\frac{n\pi}{\Omega}\Big)\frac{\operatorname{sen}(\Omega t - n\pi)}{(\Omega t - n\pi)}.$$

□

Si definimos la función seno cardinal como

$$\operatorname{senc}(x) := \frac{\operatorname{sen}(x)}{x},$$

entonces el teorema anterior nos dice que

$$f(t) = \sum_{n\in\mathbb{Z}} f\Big(\frac{n\pi}{\Omega}\Big)\operatorname{senc}(\Omega t - n\pi). \tag{2.3}$$

Observación 2.1.1. *A partir del teorema clásico de muestreo se puede probar fácilmente que el conjunto* $\left\{\sqrt{\frac{\Omega}{\pi}}senc(\Omega\cdot -n\pi)\right\}_{n\in\mathbb{Z}}$ *es una base ortonormal de* B_Ω*, valiendo que* $f\left(\frac{n\pi}{\Omega}\right) = \frac{\Omega}{\pi}\langle f, senc(\Omega\cdot -n\pi)\rangle$.

La fórmula de reconstrucción (2.3) no es utilizada en la práctica debido a la poca estabilidad que ofrece.
Al muestrear la función f se cometen errores en el cálculo de los coeficientes $f(\frac{n\pi}{\Omega})$, si llamamos ε_n al error cometido en la muestra n-ésima, resulta que en la fórmula (2.3) los coeficientes $f(\frac{n\pi}{\Omega})$ son reemplazados por $\tilde{f}_n = f(\frac{n\pi}{\Omega}) + \varepsilon_n$. En ese caso, la nueva función quedaría

$$\tilde{f}(t) = \sum_{n\in\mathbb{Z}} \tilde{f}_n \,\mathrm{senc}(\Omega t - n\pi),$$

y el error cometido al utilizar $\tilde{f}$ en lugar de f sería

$$|f(t) - \tilde{f}(t)| = \Big|\sum_{n\in\mathbb{Z}} \varepsilon_n \,\mathrm{senc}(\Omega t - n\pi)\Big|. \tag{2.4}$$

La función $\mathrm{senc}(x)$ no tiene buen decaimiento, más aún

$$\sum_{n\in\mathbb{Z}} |\mathrm{senc}(\Omega t - n\pi)| = \infty,$$

entonces no sólo no se puede establecer una cota para el error (2.4), sino que ni siquiera se puede asegurar la convergencia del mismo.

Para resolver este problema de estabilidad en el muestreo, se usa el sobremuestreo. Es aquí donde entra en juego la importancia de la redundancia en el proceso de digitalización de una señal.

2.1.1. Sobremuestreo

En la sección anterior se descompuso a f en núcleos que tenían poco decaimiento, esto hizo que la introducción de errores en el cálculo de las muestras produzcan errores globales importantes.

En esta sección se busca reemplazar esos núcleos por otros que tengan buen decaimiento. Una forma de hacer esto es tomar una función cuya trasformada de Fourier sea además de localizada, suave.

Dado $\lambda > 1$, sea g_λ tal que:

$$\hat{g}_\lambda \in C^\infty \qquad \text{y} \qquad \hat{g}_\lambda(\xi) = \begin{cases} \frac{1}{\sqrt{2\pi}} & \text{si} \quad |\xi| \leq \pi \\ 0 & \text{si} \quad |\xi| > \lambda\pi. \end{cases} \tag{2.5}$$

Utilizaremos la notación $g := g_\lambda$ para simplificar la escritura.

Para probar que el seno cardinal se puede reemplazar por la función g en la descomposición de f de forma que la misma sea estable, debemos primero demostrar el siguiente lema.

Lema 2.1.1. *Sea $\lambda > 1$ y g una función derivable en $\mathbb{R}$, tal que $g \in L^1(\mathbb{R})$ y $g' \in L^1(\mathbb{R})$, entonces*

$$\frac{1}{\lambda} \sum_{n\in\mathbb{Z}} \Big| g\Big(\frac{\Omega}{\pi}t - \frac{n}{\lambda}\Big)\Big| \leq C_g \qquad \forall\ t \in \mathbb{R},$$

donde $C_g = \frac{1}{\lambda}\|g'\|_{L^1} + \|g\|_{L^1}$.

Demostración. Sea $\xi_{t,n}$ tal que

$$|g(\xi_{t,n})| = \text{mín}\left\{ |g(s)| \Big/ s \in \Big[\frac{\Omega t}{\pi} - \frac{n+1}{\lambda}, \frac{\Omega t}{\pi} - \frac{n}{\lambda}\Big]\right\}.$$

Como

$$g\Big(\frac{\Omega t}{\pi} - \frac{n}{\lambda}\Big) - g(\xi_{t,n}) = \int_{\xi_{t,n}}^{\frac{\Omega t}{\pi} - \frac{n}{\lambda}} g'(s)\, ds,$$

vale que

$$\begin{aligned} \Big| g\Big(\frac{\Omega t}{\pi} - \frac{n}{\lambda}\Big)\Big| &\leq \Big| g\Big(\frac{\Omega t}{\pi} - \frac{n}{\lambda}\Big) - g(\xi_{t,n})\Big| + |g(\xi_{t,n})| \\ &\leq \int_{\xi_{t,n}}^{\frac{\Omega t}{\pi} - \frac{n}{\lambda}} |g'(s)|\, ds + |g(\xi_{t,n})| \\ &\leq \int_{\frac{\Omega t}{\pi} - \frac{n+1}{\lambda}}^{\frac{\Omega t}{\pi} - \frac{n}{\lambda}} |g'(s)|\, ds + |g(\xi_{t,n})|. \end{aligned} \tag{2.6}$$

Por definición de $\xi_{t,n}$ se cumple que

$$|g(\xi_{t,n})| \leq |g(s)| \ \ \forall s \in \Big[\frac{\Omega t}{\pi} - \frac{n+1}{\lambda}, \frac{\Omega t}{\pi} - \frac{n}{\lambda}\Big],$$

entonces

$$\frac{1}{\lambda}|g(\xi_{t,n})| \leq \int_{\frac{\Omega t}{\pi}-\frac{n+1}{\lambda}}^{\frac{\Omega t}{\pi}-\frac{n}{\lambda}} |g(s)|\,ds. \tag{2.7}$$

Usando (2.6) y (2.7), queda que

$$\begin{aligned}\frac{1}{\lambda}\sum_{n\in\mathbb{Z}}\Big|g\Big(\frac{\Omega t}{\pi}-\frac{n}{\lambda}\Big)\Big| &\leq \frac{1}{\lambda}\sum_{n\in\mathbb{Z}}\int_{\frac{\Omega t}{\pi}-\frac{n+1}{\lambda}}^{\frac{\Omega t}{\pi}-\frac{n}{\lambda}} |g'(s)|\,ds + \frac{1}{\lambda}\sum_{n\in\mathbb{Z}}|g(\xi_{t,n})| \\ &\leq \frac{1}{\lambda}\int_{\mathbb{R}}|g'(s)|\,ds + \sum_{n\in\mathbb{Z}}\int_{\frac{\Omega t}{\pi}-\frac{n+1}{\lambda}}^{\frac{\Omega t}{\pi}-\frac{n}{\lambda}} |g(s)|\,ds \\ &\leq \frac{1}{\lambda}\int_{\mathbb{R}}|g'(s)|\,ds + \int_{\mathbb{R}}|g(s)|\,ds.\end{aligned}$$

□

Corolario 2.1.1. *Si g cumple (2.5) y $\lambda > 1$, entonces*

$$\frac{1}{\lambda}\sum_{n\in\mathbb{Z}}\Big|g\Big(\frac{\Omega}{\pi}t-\frac{n}{\lambda}\Big)\Big| \leq C_g \qquad \forall\ t\in\mathbb{R},$$

donde $C_g = \frac{1}{\lambda}\|g'\|_{L^1} + \|g\|_{L^1}$.

Demostración. Basta observar que si g cumple (2.5), entonces se encuentra en la clase de Schwartz, ya que $\hat{g}$ está en dicha clase. Esto hace que tanto g como g' tengan buen decaimiento, resultando ser ambas integrables.

□

Teorema 2.1.2. *Sea $f \in B_\Omega$, si g cumple (2.5) y $\lambda > 1$, entonces*

$$f(t) = \frac{1}{\lambda}\sum_{n\in\mathbb{Z}} f\Big(\frac{n\pi}{\lambda\Omega}\Big)g\Big(\frac{\Omega}{\pi}t-\frac{n}{\lambda}\Big),$$

con convergencia absoluta y uniforme en $\mathbb{R}$.

Demostración. Como $\operatorname{sop}\hat{f} \subseteq [-\Omega,\Omega]$ y $\lambda > 1$, $\operatorname{sop}\hat{f} \subseteq [-\lambda\Omega,\lambda\Omega]$. Aplicando la igualdad (2.2) obtenida en la demostración del teorema 2.1.1, pero ahora viendo a $f \in B_{\lambda\Omega}$, queda que

$$\hat{f}(\xi) = \frac{1}{\lambda\Omega}\sqrt{\frac{\pi}{2}}\sum_{n\in\mathbb{Z}} f\Big(\frac{n\pi}{\lambda\Omega}\Big)e^{\frac{-in\xi\pi}{\lambda\Omega}} \quad \text{para } |\xi| \leq \lambda\Omega. \tag{2.8}$$

Usando esta expresión para $\hat{f}$ podemos afirmar que

$$\hat{f}(\xi) = \frac{\pi}{\lambda\Omega} \sum_{n\in\mathbb{Z}} f\Big(\frac{n\pi}{\lambda\Omega}\Big) e^{\frac{-in\xi\pi}{\lambda\Omega}} \hat{g}\Big(\frac{\pi\xi}{\Omega}\Big) \quad \forall\, \xi \in \mathbb{R}, \tag{2.9}$$

ya que:

- Si $|\xi| \leq \Omega$, entonces $\hat{g}\big(\frac{\pi\xi}{\Omega}\big) = \frac{1}{\sqrt{2\pi}}$. Reemplazando en (2.8) a $\sqrt{\frac{\pi}{2}}$ por $\pi\hat{g}\big(\frac{\pi\xi}{\Omega}\big)$ queda (2.9).
- Si $\Omega < |\xi| \leq \lambda\Omega$, como $\hat{f}(\xi) = 0$ y la fórmula (2.8) es válida, resulta que

  $$\frac{1}{\lambda\Omega} \sum_{n\in\mathbb{Z}} f\Big(\frac{n\pi}{\lambda\Omega}\Big) e^{\frac{-in\xi\pi}{\lambda\Omega}} = 0,$$

  por lo que

  $$\pi\hat{g}\Big(\frac{\pi\xi}{\Omega}\Big) \frac{1}{\lambda\Omega} \sum_{n\in\mathbb{Z}} f\Big(\frac{n\pi}{\lambda\Omega}\Big) e^{\frac{-in\xi\pi}{\lambda\Omega}} = 0 = \hat{f}(\xi).$$

- Si $|\xi| \geq \lambda\Omega$, $\hat{f}(\xi) = 0$ y $\hat{g}\big(\frac{\pi\xi}{\Omega}\big) = 0$, ya que $\operatorname{sop}\hat{g} \subseteq [-\lambda\pi, \lambda\pi]$. Esto hace que

  $$\hat{g}\Big(\frac{\pi\xi}{\Omega}\Big) \frac{\pi}{\lambda\Omega} \sum_{n\in\mathbb{Z}} f\Big(\frac{n\pi}{\lambda\Omega}\Big) e^{\frac{-in\xi\pi}{\lambda\Omega}} = 0 = \hat{f}(\xi).$$

Si aplicamos la trasformada de Fourier inversa en (2.9) y usamos que la antitransformada de

$$e^{\frac{-in\xi\pi}{\lambda\Omega}} \hat{g}\Big(\frac{\pi\xi}{\Omega}\Big)$$

es

$$\frac{\Omega}{\pi} g\Big(\frac{\Omega}{\pi}t - \frac{n}{\lambda}\Big),$$

tenemos que

$$f(t) = \frac{1}{\lambda} \sum_{n\in\mathbb{Z}} f\Big(\frac{n\pi}{\lambda\Omega}\Big) g\Big(\frac{\Omega}{\pi}t - \frac{n}{\lambda}\Big) \qquad \forall\ t \in \mathbb{R}.$$

La convergencia absoluta de la serie se prueba usando el corolario 2.1.1:

$$\Big|\frac{1}{\lambda}\sum_{n\in\mathbb{Z}} f\Big(\frac{n\pi}{\lambda\Omega}\Big)g\Big(\frac{\Omega}{\pi}t-\frac{n}{\lambda}\Big)\Big| \leq \|f\|_{L^\infty(\mathbb{R})}\frac{1}{\lambda}\sum_{n\in\mathbb{Z}}\Big|g\Big(\frac{\Omega}{\pi}t-\frac{n}{\lambda}\Big)\Big| \leq \|f\|_{L^\infty(\mathbb{R})}C_g.$$

Para probar la convergencia uniforme de la serie, veremos que la sucesión de sumas parciales es una sucesión de Cauchy.
Para la prueba usaremos el resultado del teorema 2.1.1 aplicado al caso en que $f \in B_{\lambda\Omega}$, es decir que vale la descomposición

$$f(t) = \sum_{n\in\mathbb{Z}} f\Big(\frac{n\pi}{\lambda\Omega}\Big)\mathrm{senc}(\lambda\Omega t - n\pi). \tag{2.10}$$

Usando esta representación de f y que $\Big\{\sqrt{\frac{\lambda\Omega}{\pi}}\mathrm{senc}(\lambda\Omega t - n\pi)\Big\}_{n\in\mathbb{Z}}$ es una base ortonormal para $B_{\lambda\Omega}$, podemos decir que

$$\sum_{n\in\mathbb{Z}}\Big|f\Big(\frac{n\pi}{\lambda\Omega}\Big)\Big|^2 = \frac{\pi}{\lambda\Omega}\|f\|^2_{L^2(-\lambda\Omega,\lambda\Omega)} < \infty.$$

Esto nos asegura que dado $\varepsilon > 0$ existe n_0, tal que si $N, M \geq n_0$, con $N \leq M$, entonces

$$\Bigg(\sum_{N\leq|n|\leq M}\Big|f\Big(\frac{n\pi}{\lambda\Omega}\Big)\Big|^2\Bigg)^{1/2} < \varepsilon.$$

Usando Cauchy-Schwarz queda que

$$\begin{aligned}
\Big|\frac{1}{\lambda}\sum_{N\leq|n|\leq M} f\Big(\frac{n\pi}{\lambda\Omega}\Big)g\Big(\frac{\Omega}{\pi}t-\frac{n}{\lambda}\Big)\Big| &\leq \frac{1}{\lambda}\sum_{N\leq|n|\leq M}\Big|f\Big(\frac{n\pi}{\lambda\Omega}\Big)\Big|\Big|g\Big(\frac{\Omega}{\pi}t-\frac{n}{\lambda}\Big)\Big| \\
&\leq \Big(\sum_{n\leq|n|\leq M}\Big|f\Big(\frac{n\pi}{\lambda\Omega}\Big)\Big|^2\Big)^{1/2}\frac{1}{\lambda}\Big(\sum_{N\leq|n|\leq M}\Big|g\Big(\frac{\Omega}{\pi}t-\frac{n}{\lambda}\Big)\Big|^2\Big)^{1/2} \\
&\leq \varepsilon\frac{1}{\lambda}\Big(\sum_{n\in\mathbb{Z}}\Big|g\Big(\frac{\Omega}{\pi}t-\frac{n}{\lambda}\Big)\Big|^2\Big)^{1/2} \leq \varepsilon\Big(\frac{\|g\|_{L^\infty(\mathbb{R})}C_g}{\lambda}\Big)^{1/2}.
\end{aligned}$$

En la última desigualdad hemos usado el corolario 2.1.1, pues

$$\sum_{n\in\mathbb{Z}}\Big|g\Big(\frac{\Omega}{\pi}t-\frac{n}{\lambda}\Big)\Big|^2 \leq \|g\|_{L^\infty(\mathbb{R})}\sum_{n\in\mathbb{Z}}\Big|g\Big(\frac{\Omega}{\pi}t-\frac{n}{\lambda}\Big)\Big| \leq \|g\|_{L^\infty(\mathbb{R})}C_g\lambda.$$

□

Observación 2.1.2. *Bajo las hipótesis del teorema anterior vale que*

$$f\Big(\frac{n\pi}{\lambda\Omega}\Big) = \frac{\Omega}{\pi}\Big\langle f, g\Big(\frac{\Omega}{\pi}\cdot - \frac{n}{\lambda}\Big)\Big\rangle,$$

por lo que

$$f(t) = \frac{\Omega}{\lambda\pi}\sum_{n\in\mathbb{Z}}\Big\langle f, g\Big(\frac{\Omega}{\pi}\cdot - \frac{n}{\lambda}\Big)\Big\rangle g\Big(\frac{\Omega}{\pi}t - \frac{n}{\lambda}\Big).$$

Demostración. Usando Plancherel,

$$\begin{aligned}\Big\langle f, g\Big(\frac{\Omega}{\pi}\cdot - \frac{n}{\lambda}\Big)\Big\rangle &= \int_{\mathbb{R}} f(t)\overline{g\Big(\frac{\Omega}{\pi}t - \frac{n}{\lambda}\Big)}\,dt = \int_{-\Omega}^{\Omega}\hat{f}(\xi)\hat{g}\Big(\frac{\pi\xi}{\Omega}\Big)\frac{\pi}{\Omega}e^{\frac{in\xi\pi}{\lambda\Omega}}\,d\xi \\ &= \frac{1}{\sqrt{2\pi}}\frac{\pi}{\Omega}\int_{-\Omega}^{\Omega}\hat{f}(\xi)e^{\frac{in\xi\pi}{\lambda\Omega}}\,d\xi = \frac{\pi}{\Omega}f\Big(\frac{n\pi}{\lambda\Omega}\Big).\end{aligned}$$

□

Probaremos ahora que la descomposición

$$f(t) = \frac{1}{\lambda}\sum_{n\in\mathbb{Z}} f\Big(\frac{n\pi}{\lambda\Omega}\Big)g\Big(\frac{\Omega}{\pi}t - \frac{n}{\lambda}\Big)$$

resulta estable bajo la introducción de errores en las muestras.
Generalmente no se conocen con exactitud los errores ε_n cometidos al muestrear la función, sólo se sabe que $|\varepsilon_n| \leq \varepsilon$, con $\varepsilon > 0$ conocido.
Dada

$$\tilde{f}(t) = \frac{1}{\lambda}\sum_{n\in\mathbb{Z}} \tilde{f}_n\, g\Big(\frac{\Omega}{\pi}t - \frac{n}{\lambda}\Big),$$

con $\tilde{f}_n = f(\frac{n\pi}{\Omega\lambda}) + \varepsilon_n$, el error al utilizar $\tilde{f}$ en lugar de f resulta

$$|f(t) - \tilde{f}(t)| = \frac{1}{\lambda}\Big|\sum_{n\in\mathbb{Z}}\varepsilon_n\, g\Big(\frac{\Omega}{\pi}t - \frac{n}{\lambda}\Big)\Big| \leq \varepsilon\frac{1}{\lambda}\Big|\sum_{n\in\mathbb{Z}} g\Big(\frac{\Omega}{\pi}t - \frac{n}{\lambda}\Big)\Big| \leq \varepsilon C_g.$$

Esto nos dice que la descomposición es estable, ya que pequeños errores en el cálculo de las muestras, no producen diferencias considerables al reemplazar f por $\tilde{f}$.

Si las muestras de una función son tomadas periódicamente a un intervalo de tiempo T, se define la *frecuencia de muestreo* como $\frac{1}{T}$. Resulta entonces

que la cantidad de muestras que se toman en una unidad de tiempo es $\lfloor \frac{1}{T} \rfloor$, donde $\lfloor x \rfloor = \text{máx}\{n \in \mathbb{Z} \,/\, n \leq x\}$.
En la sección anterior, se utilizaban las muestras $f(\frac{n\pi}{\Omega})$, esto hacía que la frecuencia de muestreo fuera $\frac{\Omega}{\pi}$, que se conoce con el nombre de frecuencia Nyquist.
En esta sección usamos las muestras $f\big(\frac{n\pi}{\lambda\Omega}\big)$, por lo que la frecuencia de muestreo es de $\frac{\lambda\Omega}{\pi}$ y dado que $\lambda > 1$, resulta que la cantidad de muestras que se toman por unidad de tiempo es mayor. Es por esto que este procedimiento se denomina *sobremuestreo*.

2.2. Cuantificación

Hasta ahora logramos expresar a $f \in B_\Omega$ a partir de sus muestras de la siguiente forma:

$$f(t) = \frac{1}{\lambda} \sum_{n \in \mathbb{Z}} f\Big(\frac{n\pi}{\lambda\Omega}\Big) g\Big(\frac{\Omega}{\pi} t - \frac{n}{\lambda}\Big).$$

Como f queda representada por un conjunto discreto de números reales, que es la sucesión $\{f\big(\frac{n\pi}{\lambda\Omega}\big)\}_{n\in\mathbb{Z}}$, podemos decir que obtuvimos la discretización de f en función del tiempo.
Dado que los coeficientes $f\big(\frac{n\pi}{\lambda\Omega}\big)$ son números reales, para la digitalización de la señal, nos resta discretizar la amplitud de las muestras.

Si notamos $x_n^\lambda := f\big(\frac{n\pi}{\lambda\Omega}\big)$, el objetivo de la cuantificación es asignarle a cada x_n^λ un valor q_n^λ que pertenezca a un conjunto finito preestablecido llamado *alfabeto* y notado como $\mathcal{A}$.
La señal aproximante generada a partir de la sucesión cuantificadora $\{q_n^\lambda\}_{n\in\mathbb{Z}}$, será

$$\tilde{f}_\lambda(t) = \frac{1}{\lambda} \sum_{n \in \mathbb{Z}} q_n^\lambda g\Big(\frac{\Omega t}{\pi} - \frac{n}{\lambda}\Big).$$

Definiremos el *error de aproximación* como $\|f - \tilde{f}_\lambda\|_{L^\infty(\mathbb{R})}$.

En lo sucesivo, para simplificar los cálculos, vamos a suponer que $\Omega = \pi$ y que $\|f\|_{L^\infty} \leq 1$.
A partir de estas normalizaciones, podemos decir que en la sección anterior obtuvimos que una función f en la clase

$$\mathcal{C} := \big\{h \in L^2(\mathbb{R}) \,/\, \|h\|_{L^\infty} \leq 1, \operatorname{sop} \hat{h} \subseteq [-\pi, \pi]\big\},$$

se puede descomponer en base a sus muestras $f(\frac{n}{\lambda})$ de la forma

$$f(t) = \frac{1}{\lambda} \sum_{n \in \mathbb{Z}} f\Big(\frac{n}{\lambda}\Big) g\Big(t - \frac{n}{\lambda}\Big).$$

Con estas suposiciones, la función aproximante va a ser

$$\tilde{f}_\lambda(t) = \frac{1}{\lambda} \sum_{n \in \mathbb{Z}} q_n^\lambda g\Big(t - \frac{n}{\lambda}\Big).$$

Tal como lo mencionamos en la introducción, se utilizan generalmente dos métodos para construir la sucesión cuantificadora $\{q_n^\lambda\}_{n \in \mathbb{Z}}$: la *cuantificación espaciada* y la *cuantificación fina*.

La cuantificación espaciada combina una alta frecuencia de muestreo con una pequeña cantidad de elementos en el alfabeto cuantificador.
Recordemos que la frecuencia de muestreo al utilizar el sobremuestreo era de $\frac{\lambda\Omega}{\pi}$. Si λ aumenta su valor, como se incrementa la cantidad de muestras tomadas por unidad de tiempo, se tiene más información acerca de la señal y por consiguiente se logra que la descomposición de la misma tenga mayor precisión. El problema es que si se tiene asignado un presupuesto de cierta cantidad de bits para el total de datos requeridos para la digitalización, al tomar muestras a una mayor frecuencia se va a poder gastar menos bits en cada una de ellas, es por esto que el alfabeto necesariamente debe tener pocos elementos.
En este capítulo estudiaremos los algoritmos $\Sigma\Delta$ que hacen uso de este tipo de cuantificación y trabajaremos con el caso extremo en que q_n^λ sólo puede tomar dos valores, es decir, que el alfabeto $\mathcal{A}$ tendrá dos elementos.

La cuantificación fina, para minimizar el error de aproximación, fija la frecuencia de muestreo e incrementa la cantidad de elementos del alfabeto. El método de cuantificación denominado Modulación por Impulsos Codificados utiliza este principio para la reducción del error. Comenzaremos estudiando este algoritmo con el objetivo de establecer una comparación en el rendimiento y aplicabilidad de ambos métodos de cuantificación.

2.3. Algoritmo de Modulación por Impulsos Codificados

Si se tiene un presupuesto de N+1 bits por muestra, con $N \in \mathbb{N}$, el algoritmo de Modulación por Impulsos Codificados (conocido como PCM) le asigna a cada $x_n^\lambda = f(\frac{n}{\lambda})$ su desarrollo binario truncado en la cifra N.

Dado que $x_n^\lambda \in [-1, 1]$, pues $x_n^\lambda = f(\frac{n}{\lambda})$ y $\|f\|_\infty \leq 1$, su desarrollo binario es del tipo:

$$x_n^\lambda = -1 + \sum_{i=0}^{\infty} 2^{-i} b_i^n, \quad \text{con } b_i^n \in \{0, 1\}.$$

El algoritmo define al elemento cuantificador de la siguiente forma:

$$q_n^\lambda = -1 + \sum_{i=0}^{N} 2^{-i} b_i^n.$$

Para relacionar este algoritmo con la definición de cuantificación dada anteriormente, podemos pensar que se le asigna a cada x_n^λ el elemento menor más cercano del alfabeto que es una progresión aritmética de paso $\delta = 2^{-N}$ con 2^{N+1} elementos y cuyo primer término es -1. Es decir,

$$\mathcal{A} = \left\{ -1, -1 + 2^{-N}, \ldots, -2^{-N}, 0, 2^{-N}, \ldots \ldots, 1 - 2^{-N+1}, 1 - 2^{-N} \right\}$$

y

$$q_n^\lambda = \text{máx}\left\{ q \in \mathcal{A} : q \leq x_n^\lambda \right\}.$$

El error que se comete al cuantificar cada muestra es

$$|x_n^\lambda - q_n^\lambda| = \sum_{i=N+1}^{\infty} 2^{-i} b_i^n \leq \sum_{i=N+1}^{\infty} 2^{-i} = 2^{-N}.$$

El algortimo PCM es un ejemplo de cuantificación fina, ya que incrementando la cantidad de elementos del alfabeto (aumentando N), logra aproximar con exactitud cada muestra.

El error global de cuantficación resulta ser

$$\begin{aligned} |f(t) - \tilde{f}_\lambda(t)| &\leq \frac{1}{\lambda} \sum_{n \in \mathbb{Z}} \left| x_n^\lambda - q_n^\lambda \right| \left| g\Big(t - \frac{n}{\lambda}\Big) \right| \\ &\leq \frac{1}{\lambda} \sum_{n \in \mathbb{Z}} 2^{-N} \left| g\Big(t - \frac{n}{\lambda}\Big) \right| \leq 2^{-N} C_g. \end{aligned}$$

Podemos ver que la acotación es de tipo exponencial en N, por lo que si se incrementa la cantidad de bits asignados a cada muestra, se obtiene una buena aproximación de la función f.

Si bien este algoritmo ofrece un error con óptimo decaimiento, en el proceso de conversión analógico a digital (A/D), no se utiliza este mecanismo, ya que resulta muy costoso construir dispositivos analógicos que posean el grado de precisión requerido para este tipo de cuantificación. Veremos más adelante que el algoritmo PCM no es robusto, ya que ante la introducción de pequeños errores en el cálculo de los coeficientes cuantificados, se producen errores globales significativos.

2.4. Algoritmos de cuantificación Sigma-Delta ($\Sigma\Delta$)

En la sección anterior vimos que en el algoritmo de Modulación por Impulsos Codificados, la elección del elemento cuantificador q_n^λ dependía exclusivamente del valor de la muestra x_n^λ.
En esta sección estudiaremos los algoritmos $\Sigma\Delta$, que tienen en cuenta para la elección del valor aproximante, no sólo al coeficiente x_n^λ sino a los errores producidos al cuantificar las muestras en los pasos anteriores. Estos algoritmos no buscan que cada q_n^λ esté lo más cerca posible del x_n^λ, sino que aproximan los promedios de ambas sucesiones.

Los esquemas que vamos a estudiar producen cuantificaciones cuyo error disminuye, no al incrementar la cantidad de elementos del alfabeto, sino al aumentar la frecuencia de muestreo. En este capítulo, para simplificar el estudio del algoritmo $\Sigma\Delta$, trabajaremos con $\mathcal{A} = \{-1, 1\}$, es decir que utilizaremos la cuantificación de 1-bit. En el próximo capítulo estudiaremos la cuantificación multibit.

Para el cálculo de la sucesión cuantificadora, los esquemas $\Sigma\Delta$ introducen variables internas. El orden del algoritmo viene dado por la cantidad de variables utilizadas. Comenzaremos estudiando los de orden uno y luego veremos los de orden arbitrario.

2.4.1. Cuantificación Sigma-Delta (ΣΔ) de orden uno

Algoritmo Sigma-Delta (ΣΔ) de orden uno para sucesiones con subíndices naturales

Dado que el alfabeto elegido es $\mathcal{A} = \{-1, 1\}$, podríamos tomar $q_n^\lambda = \text{sign}(f(\frac{n}{\lambda}))$, pero esta elección no es apropiada, ya que cualesquiera dos funciones positivas tendrían la misma función aproximante.

Buscamos definir el algoritmo ΣΔ para una sucesión de datos $\{s_n\}_{n\in\mathbb{Z}}$ tal que $-1 \leq s_n \leq 1$ para todo $n \in \mathbb{Z}$, ya que buscamos cuantificar las muestras $f\left(\frac{n}{\lambda}\right)$ (recordemos que $\|f\|_{L^\infty} \leq 1$). Comenzaremos definiendo el algoritmo sólo para sucesiones con subíndices naturales, veremos luego cómo se extiende al caso general.

Para entender el origen de la fórmula y del nombre del algoritmo ΣΔ, vamos a definir a continuación los operadores *sigma* (Σ) y *delta* (Δ).

Definición 2.4.1. *i) Dada $s = \{s_n\}_{n\in\mathbb{N}_0}$ una sucesión de número reales, se define el operador integral discreto Σ de la siguiente manera:*

$$\begin{cases} (\Sigma s)_n = \sum_{i=1}^{n} s_i & \forall\ n \in \mathbb{N} \\ (\Sigma s)_0 = s_0. \end{cases}$$

ii) El operador diferencial discreto Δ se define por:

$$(\Delta s)_n = s_n - s_{n-1} \quad \forall\ n \in \mathbb{N}.$$

Observación 2.4.1. *A partir de las definiciones anteriores, si $s = \{s_n\}_{n\in\mathbb{N}}$, vale que*

$$\Delta\Sigma s = s.$$

Demostración. Si $n \geq 2$, entonces

$$(\Delta\Sigma s)_n = (\Sigma s)_n - (\Sigma s)_{n-1} = \sum_{i=1}^{n} s_i - \sum_{i=1}^{n-1} s_i = s_n.$$

Si $n = 1$, tomando $s_0 = 0$,

$$(\Delta\Sigma s)_1 = (\Sigma s)_1 - (\Sigma s)_0 = s_1 - s_0 = s_1.$$

□

El algoritmo $\Sigma\Delta$ busca que, dada una sucesión $\{s_n\}_{n\in\mathbb{N}}$ tal que $-1 \leq s_n \leq 1$ para todo $n \in \mathbb{N}$, se construya una sucesión $\{q_n\}_{n\in\mathbb{N}}$ tomando valores en $\{-1, 1\}$ de forma que exista $C > 0$ tal que

$$\Big| \sum_{1+j}^{N+j} s_n - \sum_{1+j}^{N+j} q_n \Big| \leq C, \quad \forall\ N \in \mathbb{N} \text{ y } j \in \mathbb{N}_0. \tag{2.11}$$

Pues se logra así que

$$\left| \frac{1}{N} \sum_{k=1+j}^{N+j} s_k - \frac{1}{N} \sum_{k=1+j}^{N+j} q_k \right| \leq \frac{C}{N} \quad \forall\ N \in \mathbb{N} \text{ y } j \in \mathbb{N}_0,$$

es decir, que el promedio de cualesquiera N coeficientes q_n consecutivos aproxima con error del orden de N^{-1} al promedio de los datos.

Si definimos la variable interna $u := \Sigma s - \Sigma q$, es decir

$$u_n = \sum_{i=1}^{n} s_i - \sum_{i=1}^{n} q_i \quad \forall\ n \in \mathbb{N}$$

y tomamos $u_0 \in \mathbb{R}$, la propiedad (2.11) se traduce en

$$|u_{N+j} - u_j| \leq C \quad \forall\ N \in \mathbb{N} \text{ y } j \in \mathbb{N}_0.$$

Para lograr esta condición, pediremos que la sucesión $\{u_n\}_{n\in\mathbb{N}_0}$ sea acotada, más aún que $-1 \leq u_n < 1$ para todo $n \in \mathbb{N}_0$.

Una vez establecida la variable interna u, se busca definir la sucesión cuantificadora $q = \{q_n\}_{n\in\mathbb{N}}$, con $q_n \in \{-1, 1\}$, de forma que $-1 \leq u_n < 1$ para todo $n \in \mathbb{N}_0$.
Como

$$u = \Sigma s - \Sigma q,$$

aplicando el operador Δ y usando la observación 2.4.1,

$$\Delta u = s - q,$$

es decir que

$$u_n = u_{n-1} + s_n - q_n.$$

Supongamos que $-1 \leq u_{n-1} < 1$, como $-1 \leq s_n \leq 1$, sucede que $-2 \leq u_{n-1} + s_n < 2$. Tenemos entonces dos casos:

Si $0 \leq u_{n-1} + s_n < 2$, dado que $u_n = u_{n-1} + s_n - q_n$ y $q_n \in \{-1, 1\}$, para que $-1 \leq u_n < 1$, debemos pedir que $q_n = 1$.

Si $-2 \leq u_{n-1} + s_n < 0$, para que $-1 \leq u_n < 1$ debemos pedir que $q_n = -1$.

Resulta entonces que

$$q_n = \begin{cases} 1 & \text{si } 0 \leq u_{n-1} + s_n < 2 \\ -1 & \text{si } -2 \leq u_{n-1} + s_n < 0. \end{cases}$$

Si definimos la función signo como

$$\operatorname{sign}(x) = \begin{cases} 1 & \text{si } x \geq 0 \\ -1 & \text{si } x < 0, \end{cases}$$

tenemos que $q_n = \operatorname{sign}(u_{n-1} + s_n)$.

Obtuvimos finalmente que, suponiendo $-1 \leq u_{n-1} < 1$, para que $-1 \leq u_n < 1$ debemos pedir que $q_n = \operatorname{sign}(u_{n-1} + s_n)$. Es por esto que dado $u_0 \in [-1, 1)$, y la sucesión $\{s_n\}_{n\in\mathbb{N}}$, con $-1 \leq s_n \leq 1$ para todo $n \in \mathbb{N}$, el algoritmo $\Sigma\Delta$ de orden uno se define recursivamente como

$$\begin{cases} u_n = u_{n-1} + s_n - q_n \\ q_n = \operatorname{sign}\big(u_{n-1} + s_n\big), \end{cases} \tag{2.12}$$

para $n \in \mathbb{N}$.

Logramos entonces definir la sucesión cuantificadora $\{q_n\}_{n\in\mathbb{N}}$, con $q_n \in \{-1, 1\}$, de forma que los promedios de la misma aproximen con precisión los promedios de la sucesión de datos. Observemos que esto se logró definiendo la sucesión de variables internas $\{u_n\}_{n\in\mathbb{N}_0}$, que cumple $-1 \leq u_n < 1$ para todo $n \in \mathbb{N}_0$. El esquema (2.12) se dice estable debido a la acotación uniforme de las variables internas.
La función $Q : \mathbb{R} \longrightarrow \{-1, 1\}$, definida por $Q(u) = \operatorname{sign}(u)$, se denomina *cuantificador escalar*.

El algoritmo (2.12) se utiliza para cuantificar una sucesión acotada por el valor uno a partir del alfabeto $\mathcal{A} = \{-1, 1\}$. Si se quiere emplear este algoritmo para una sucesión génerica de números reales acotada, y un alfabeto arbitrario, el cuantificador escalar Q cambiaría, y el esquema sería el siguiente:

Definición 2.4.2. *Dada $\{s_n\}_{n\in\mathbb{N}}$ una sucesión de números reales, y dado $u_0 \in \mathbb{R}$. Si $\mathcal{A}$ es el alfabeto establecido para la cuantificación y $Q : \mathbb{R} \longrightarrow \mathcal{A}$ es el cuantificador escalar, el algoritmo $\Sigma\Delta$ de orden uno se define a partir de la siguiente recursión:*

$$\begin{cases} u_n = u_{n-1} + s_n - q_n \\ q_n = Q(u_{n-1} + s_n), \end{cases} \tag{2.13}$$

donde $n \geq 1$.

Para implementar el algoritmo, dado el elemento u_0 se calcula primero $q_1 = Q(u_0 + s_1)$. Usando el coeficiente q_1 se obtiene $u_1 = u_0 + s_1 - q_1$. El siguiente paso es calcular q_2 utilizando u_1, luego se calcula u_2 y así sucesivamente. De esta manera se generan las sucesiones $\{u_n\}_{n\in\mathbb{N}_0}$ de variables internas y $\{q_n\}_{n\in\mathbb{N}}$ de elementos cuantificadores.

Para que el algoritmo (2.13) sea aplicable, el *cuantificador escalar* Q, debe ser elegido de forma tal que la sucesión $\{u_n\}_{n\in\mathbb{N}_0}$ resulte acotada. En ese caso el algoritmo se dice *estable*. Obviamente para lograr la estabilidad, se va a requerir que la sucesión de datos $\{s_n\}_{n\in\mathbb{N}}$ también sea acotada.

Una vez enunciada la generalización del algoritmo, volveremos al caso que nos interesa, la recursión (2.12). Nosotros vamos a utilizarla para la sucesión $s_n^\lambda = f(\frac{n}{\lambda})$, es decir, trabajaremos con el siguiente esquema:

Definición 2.4.3. *Dados $\lambda > 1$, $f \in \mathcal{C}$ y $u_0^\lambda \in [-1, 1)$, el algoritmo $\Sigma\Delta$ de orden uno se define por*

$$\begin{cases} u_n^\lambda = u_{n-1}^\lambda + f(\frac{n}{\lambda}) - q_n^\lambda \\ q_n^\lambda = sign\big(u_{n-1}^\lambda + f(\frac{n}{\lambda})\big), \end{cases} \tag{2.14}$$

para $n \in \mathbb{N}$.

Recordemos que este esquema sólo sirve para cuantificar las muestras con subíndices naturales. En la sección de filtros finitos veremos que en la

práctica sólo se utilizan las cuantificaciones con esos subíndices, pero que por fines teóricos se define el algoritmo para todos los $n \in \mathbb{Z}$. Antes de definir el esquema para $n \leq 0$, veremos que el sistema (2.14) hereda las propiedades del esquema (2.12), es decir, la estabilidad (la acotación de las variables internas) y la aproximación de los promedios.

Estabilidad

Proposición 2.4.1. *Sea $f \in \mathcal{C}$ y $\lambda > 1$, si $-1 \leq u_0^\lambda < 1$, entonces $-1 \leq u_n^\lambda < 1$ para todo $n \geq 0$, donde $\{u_n^\lambda\}_{n\in\mathbb{N}}$ es la sucesión definida en el esquema (2.14).*

Demostración. Se prueba por inducción. La prueba es obvia debido a que la sucesión $\{q_n^\lambda\}_{n\in\mathbb{N}}$ fue definida de forma tal que la sucesión $\{u_n^\lambda\}_{n\in\mathbb{N}_0}$ cumpla que $-1 \leq u_n^\lambda < 1$ para todo $n \in \mathbb{N}_0$.

□

Dado que las variables internas están acotadas independientemente de λ, es decir, si $-1 \leq u_0^\lambda < 1$, entonces

$$|u_n^\lambda| \leq 1 \text{ para todo } n \in \mathbb{N}, \lambda > 1,$$

usaremos la notación $u_n := u_n^\lambda$.

Observación 2.4.2. *La variable interna u_n es la suma de los errores cometidos al aproximar $f\left(\frac{k}{\lambda}\right)$ por q_k^λ con $1 \leq k \leq n$.*

Demostración. Por el esquema (2.14),

$$u_n - u_{n-1} = f\left(\frac{n}{\lambda}\right) - q_n^\lambda,$$

entonces

$$u_N - u_0 = \sum_{k=1}^{N} \left(u_k - u_{k-1}\right) = \sum_{k=1}^{N} \left(f\left(\frac{k}{\lambda}\right) - q_k^\lambda\right).$$

Resulta finalmente que

$$u_N = u_0 + \sum_{k=1}^{N} \left(f\left(\frac{k}{\lambda}\right) - q_k^\lambda\right).$$

El error cometido al cuantificar la muestra n-ésima es $E_n := f\left(\frac{n}{\lambda}\right) - q_n^\lambda$. Si definimos $E_0 := u_0$, resulta que

$$u_N = \sum_{n=0}^{N} E_n.$$

□

Observación 2.4.3. *Recordemos que la estabilidad del sistema, es decir, la acotación de las variables internas, hace que el promedio de cualesquiera N coeficientes q_n^λ consecutivos aproxime con error del orden de N^{-1} al promedio de las muestras, pues*

$$\left|\frac{1}{N}\sum_{k=1+j}^{N+j} f\left(\frac{k}{\lambda}\right) - \frac{1}{N}\sum_{k=1+j}^{N+j} q_k^\lambda\right| = \left|\frac{1}{N}(u_{N+j} - u_j)\right| \leq \frac{2}{N}, \quad \forall\ N \in \mathbb{N}\ y\ j \in \mathbb{N}_0.$$

A diferencia del esquema de Modulación por Impulsos Codificados, el algoritmo $\Sigma\Delta$ no logra generalmente que el error al aproximar individualmente cada muestra sea pequeño. De hecho sólo se puede afirmar que

$$\left|f\left(\frac{n}{\lambda}\right) - q_n^\lambda\right| = |u_n - u_{n-1}| \leq 2.$$

Pero este algoritmo logra una buena aproximación de los promedios de las sucesiones.

Finalmente podemos observar que

$$q_N^\lambda = \operatorname{sign}\left(u_{N-1} + f\left(\frac{N}{\lambda}\right)\right) = \operatorname{sign}\left(\sum_{n=0}^{N-1} E_n + f\left(\frac{N}{\lambda}\right)\right).$$

Esto muestra que el algoritmo $\Sigma\Delta$ para cuantificar la muestra N-ésima, tiene en cuenta los errores cometidos anteriormente.

Veremos a continuación cómo definir el algoritmo $\Sigma\Delta$ para subíndices negativos.

Algoritmo Sigma-Delta (ΣΔ) de orden uno para sucesiones con subíndices enteros

Para dar una relación recursiva en el caso de subíndices negativos necesitamos definir u_{n-1} en función de u_n.
De (2.14) sabemos que para $n \geq 1$

$$u_{n-1} = u_n - f\Big(\frac{n}{\lambda}\Big) + \text{sign}\Big(u_{n-1} + f\Big(\frac{n}{\lambda}\Big)\Big).$$

Además vale que $\text{sign}\big(u_{n-1} + f\big(\frac{n}{\lambda}\big)\big) = -\text{sign}\big(u_n - f\big(\frac{n}{\lambda}\big)\big)$ ya que,

$$\begin{aligned} u_{n-1} + f\Big(\frac{n}{\lambda}\Big) \geq 0 \implies & \text{sign}\Big(u_{n-1} + f\Big(\frac{n}{\lambda}\Big)\Big) = 1 \\ \implies & u_n = u_{n-1} + f\Big(\frac{n}{\lambda}\Big) - 1 \\ \implies & u_n - f\Big(\frac{n}{\lambda}\Big) = u_{n-1} - 1 < 0. \end{aligned}$$

$$\begin{aligned} u_{n-1} + f\Big(\frac{n}{\lambda}\Big) < 0 \implies & \text{sign}\Big(u_{n-1} + f\Big(\frac{n}{\lambda}\Big)\Big) = -1 \\ \implies & u_n = u_{n-1} + f\Big(\frac{n}{\lambda}\Big) + 1 \\ \implies & u_n - f\Big(\frac{n}{\lambda}\Big) = u_{n-1} + 1 \geq 0. \end{aligned}$$

(En las últimas implicaciones usamos que $-1 \leq u_k < 1$ para todo $k \geq 0$).

Entonces la fórmula en el caso en que $n \geq 1$ podemos pensarla como:

$$u_{n-1} = u_n - f\Big(\frac{n}{\lambda}\Big) - \text{sign}\Big(u_n - f\Big(\frac{n}{\lambda}\Big)\Big).$$

Si copiamos este esquema para $n \leq 0$, tenemos lo siguiente:

Definición 2.4.4. *Dados $\lambda > 1$, $f \in \mathcal{C}$ y $u_0 \in [-1, 1)$, el algoritmo ΣΔ para $n \leq 0$ se define:*

$$\begin{cases} u_{n-1} = u_n - f\big(\frac{n}{\lambda}\big) + q_n^\lambda \\ q_n^\lambda = -sign\big(u_n - f\big(\frac{n}{\lambda}\big)\big). \end{cases} \tag{2.15}$$

La siguiente proposición enuncia la estabilidad para el caso $n \in \mathbb{Z}$.

Proposición 2.4.2. *Si $f \in \mathcal{C}$ y $-1 \leq u_0 < 1$, entonces $-1 \leq u_n < 1$ para todo $n \in \mathbb{Z}$, donde los u_n son los definidos en los esquemas (2.14) y (2.15).*

Demostración. La demostración es análoga que para el caso $n \in \mathbb{N}$. □

Observación 2.4.4. *Se puede demostrar por inducción que si en la proposición anterior en lugar de requerir que $-1 \leq u_0 < 1$, se hubiese pedido que $|u_0| \leq 1$, se habría obtenido que $|u_n| \leq 1$ para todo $n \in \mathbb{Z}$.*

Hasta ahora, los resultados obtenidos acerca del algoritmo $\Sigma\Delta$, es decir la estabilidad y la aproximación de los promedios, son válidos para cualquier tipo de sucesión de números reales acotada por el valor uno. A continuación veremos que, para el caso en que la sucesión son las muestras de una función, la estabilidad del sistema produce un error de aproximación que depende de la frecuencia de muestreo.

Error de aproximación

Vimos ya que los algoritmos $\Sigma\Delta$ no buscan que el error local de aproximación sea pequeño, sino que los promedios de la sucesión cuantificadora aproximen con precisión a los promedios de la sucesión original. Vamos a estudiar ahora el error global de aproximación, representado por la diferencia en norma infinito entre la función original f y la función aproximante:

$$\tilde{f}_\lambda(t) = \frac{1}{\lambda} \sum_{n \in \mathbb{Z}} q_n^\lambda g\Big(t - \frac{n}{\lambda}\Big), \tag{2.16}$$

donde $\{q_n^\lambda\}_{n \in \mathbb{Z}}$ es la sucesión generada por los esquemas (2.14) y (2.15).

Teorema 2.4.1. *Sea $f \in \mathcal{C}$, $\lambda > 1$, $u_0 \in [-1, 1)$ y g que cumple (2.5). Si definimos la sucesión $\{q_n^\lambda\}_{n \in \mathbb{Z}}$ como en (2.14) y (2.15), y $\tilde{f}_\lambda$ como en (2.16), entonces*

$$|f(t) - \tilde{f}_\lambda(t)| \leq \frac{1}{\lambda} \|g'\|_{L^1} \quad \forall\, t \in \mathbb{R}.$$

Demostración. Usando la fórmula de sumación por partes:

$$\sum_{n \in \mathbb{Z}} (w_n - w_{n-1}) v_n = \sum_{n \in \mathbb{Z}} w_n (v_n - v_{n+1}),$$

y que $|u_n| \leq 1$ para todo $n \in \mathbb{Z}$, tenemos que

$$\begin{aligned}
|f(t) - \tilde{f}_\lambda(t)| &= \Big|\frac{1}{\lambda}\sum_{n\in\mathbb{Z}} f\Big(\frac{n}{\lambda}\Big)g\Big(t-\frac{n}{\lambda}\Big) - \frac{1}{\lambda}\sum_{n\in\mathbb{Z}} q_n^\lambda g\Big(t-\frac{n}{\lambda}\Big)\Big| \\
&= \frac{1}{\lambda}\Big|\sum_{n\in\mathbb{Z}}\Big(f\Big(\frac{n}{\lambda}\Big) - q_n^\lambda\Big)g\Big(t-\frac{n}{\lambda}\Big)\Big| \\
&= \frac{1}{\lambda}\Big|\sum_{n\in\mathbb{Z}}(u_n - u_{n-1})g\Big(t-\frac{n}{\lambda}\Big)\Big| \\
&= \frac{1}{\lambda}\Big|\sum_{n\in\mathbb{Z}} u_n\Big(g\Big(t-\frac{n}{\lambda}\Big) - g\Big(t-\frac{n+1}{\lambda}\Big)\Big)\Big| \\
&\leq \frac{1}{\lambda}\sum_{n\in\mathbb{Z}}\Big|g\Big(t-\frac{n}{\lambda}\Big) - g\Big(t-\frac{n+1}{\lambda}\Big)\Big| \\
&= \frac{1}{\lambda}\sum_{n\in\mathbb{Z}}\Big|\int_{t-\frac{n+1}{\lambda}}^{t-\frac{n}{\lambda}} g'(s)\,ds\Big| \leq \\
&\leq \frac{1}{\lambda}\sum_{n\in\mathbb{Z}}\int_{t-\frac{n+1}{\lambda}}^{t-\frac{n}{\lambda}} |g'(s)|\,ds = \frac{1}{\lambda}\|g'\|_{L^1}.
\end{aligned}$$

□

Observemos que si bien la función g depende de λ, pues $\operatorname{sop}\hat{g} \subseteq [-\lambda\pi, \lambda\pi]$, si fijamos $\lambda_0 > 1$, entonces la función g_{λ_0} cumple (2.5) para todo $\lambda \geq \lambda_0$. Entonces a partir del teorema anterior podemos decir que

$$\|f - \tilde{f}_\lambda\|_{L^\infty} \leq \frac{1}{\lambda}\|g'_{\lambda_0}\|_{L^1} \quad \forall\ \lambda \geq \lambda_0. \tag{2.17}$$

Con lo cual el error tiende a cero si λ tiende a infinito, es decir, si la frecuencia de muestreo tiende a infinito.

Esto es lo que comentamos en la introducción. El algoritmo $\Sigma\Delta$ tiene error pequeño si la cantidad de bits (1 en este caso) usados por muestra está fija y se incrementa la redundancia del muestreo.

La acotación (2.17) se dice de tipo $O(\lambda^{-1})$, ya que existe una constante $C > 0$ que no depende de λ, tal que $\|f - \tilde{f}_\lambda\|_{L^\infty} \leq \frac{C}{\lambda}$ para todo $\lambda \geq \lambda_0$.

Algoritmos Sigma-Delta ($\Sigma\Delta$) de orden uno con cuantificadores imperfectos

En la práctica, debido a que los esquemas de cuantificación deben implementarse con dispositivos analógicos que introducen errores, en los sistemas (2.14) y (2.15) que definen el algoritmo $\Sigma\Delta$, la función $Q(x) = \text{sign}(x)$ puede llegar a ser reemplazada por $Q(x) = \text{sign}(x+\delta)$. Donde de δ sólo se sabe que $|\delta| < \tau$ con $\tau > 0$.
Incluso el error δ puede variar de un paso a otro, haciendo variar el cuantificador, entonces podemos decir que $Q_n(x) = \text{sign}(x + \delta_n)$ con $|\delta_n| < \tau$ para todo $n \in \mathbb{Z}$.
Observemos que:

- $x \geq \tau > 0 \Longrightarrow x + \delta_n \geq \tau + \delta_n > 0 \Longrightarrow Q_n(x) = 1 = \text{sign}(x)$.
- $x \leq -\tau < 0 \Longrightarrow x + \delta_n \leq -\tau + \delta_n < 0 \Longrightarrow Q_n(x) = -1 = \text{sign}(x)$.
- Para $-\tau < x < \tau$, el comportamiento de Q_n depende de δ_n que es desconocido, por lo que, al ser Q_n una traslación de la función signo, sólo podemos decir que $|Q_n(x)| \leq 1$.

Obtenemos finalmente que:

$$\begin{cases} Q_n(x) = \text{sign}(x) & \text{para } |x| \geq \tau \\ |Q_n(x)| \leq 1 & \text{para } |x| < \tau. \end{cases} \tag{2.18}$$

Definición 2.4.5. *Si Q_n cumple (2.18), el esquema $\Sigma\Delta$ para $n \geq 1$ con estos nuevos cuantificadores es*

$$\begin{cases} u_n = u_{n-1} + f\left(\frac{n}{\lambda}\right) - q_n^\lambda \\ q_n^\lambda = Q_n\left(u_{n-1} + f\left(\frac{n}{\lambda}\right)\right). \end{cases} \tag{2.19}$$

El algoritmo se dice robusto, si sigue siendo estable aunque el cuantificador Q implemente pequeños errores de los cuales sólo conocemos la cota τ. Veamos que efectivamente los esquemas $\Sigma\Delta$ son robustos.

Proposición 2.4.3. *Sea $f \in \mathcal{C}$, $\tau > 0$ y $\{u_n\}_{n\in\mathbb{N}}$ definida como en (2.19). Si $|u_0| \leq 1 + \tau$, entonces $|u_n| \leq 1 + \tau$ para todo $n \geq 0$.*

Demostración. Usamos inducción.
Para el caso $n = 0$ es trivial.
Supongamos que $|u_{n-1}| \leq 1 + \tau$, entonces $\left|u_{n-1} + f\left(\frac{n}{\lambda}\right)\right| \leq \tau + 2$.
Separemos en casos,

- Si $\tau \leq u_{n-1} + f\left(\frac{n}{\lambda}\right) \leq \tau + 2$, entonces

$$q_n = Q_n\Big(u_{n-1} + f\Big(\frac{n}{\lambda}\Big)\Big) = \operatorname{sign}\Big(u_{n-1} + f\Big(\frac{n}{\lambda}\Big)\Big) = 1,$$

  por lo que

$$-\tau - 1 < \tau - 1 \leq u_n = u_{n-1} + f\Big(\frac{n}{\lambda}\Big) - 1 \leq \tau + 1,$$

  lo que probaría que $|u_n| \leq \tau + 1$.

- Si $-\tau - 2 \leq u_{n-1} + f\left(\frac{n}{\lambda}\right) \leq -\tau$, entonces

$$q_n = Q_n\Big(u_{n-1} + f\Big(\frac{n}{\lambda}\Big)\Big) = \operatorname{sign}\Big(u_{n-1} + f\Big(\frac{n}{\lambda}\Big)\Big) = -1,$$

  por lo que

$$-\tau - 1 \leq u_n = u_{n-1} + f\Big(\frac{n}{\lambda}\Big) + 1 \leq -\tau + 1 \leq \tau + 1,$$

  lo que probaría que $|u_n| \leq \tau + 1$.

- Si $-\tau < u_{n-1} + f\left(\frac{n}{\lambda}\right) < \tau$, entonces

$$\Big|Q_n\Big(u_{n-1} + f\Big(\frac{n}{\lambda}\Big)\Big)\Big| \leq 1,$$

  por lo que

$$-\tau - 1 \leq u_n = u_{n-1} + f\Big(\frac{n}{\lambda}\Big) - Q_n\Big(u_{n-1} + f\Big(\frac{n}{\lambda}\Big)\Big) \leq \tau + 1.$$

□

El algoritmo para subíndices negativos lo definimos en función a la deducción hecha en el caso en que $Q(x) = \operatorname{sign}(x)$.

Definición 2.4.6. *Si Q_n cumple (2.18), el algoritmo $\Sigma\Delta$ para $n < 0$ con estos cuantificadores, se define*

$$\begin{cases} u_n = u_{n+1} - f(\frac{n+1}{\lambda}) + q_{n+1}^{\lambda} \\ q_{n+1}^{\lambda} = -Q_n\big(u_{n+1} - f(\frac{n+1}{\lambda})\big). \end{cases} \tag{2.20}$$

Proposición 2.4.4. *Sea $f \in \mathcal{C}$, $\{u_n\}_{n\in\mathbb{Z}}$ definida en (2.18) y (2.20). Si $|u_0| \leq 1 + \tau$, entonces $|u_n| \leq 1 + \tau$ para todo $n \in \mathbb{Z}$.*

Demostración. Es análoga a la anterior.

□

Veamos ahora el efecto de los cuantificadores imperfectos en la acotación del error.

Teorema 2.4.2. *Sea $f \in \mathcal{C}$, $\lambda > 1$ y g que cumple (2.5). Si definimos la sucesión $\{q_n^{\lambda}\}_{n\in\mathbb{Z}}$ como en (2.18) y (2.20), con $|u_0| \leq 1 + \tau$, entonces*

$$|f(t) - \tilde{f}_{\lambda}(t)| \leq \frac{1+\tau}{\lambda}\|g'\|_{L^1} \quad \forall\, t \in \mathbb{R}.$$

Demostración. La demostración es análoga a la del teorema para el caso $Q(x) = \text{sign}(x)$.

□

El teorema anterior enuncia una de las propiedades más sobresalientes del algoritmo $\Sigma\Delta$ que es la robustez. Al introducir cuantificadores imperfectos, si bien la cota del error aumentó en $1 + \tau$, ésta sigue siendo del tipo $O(\lambda^{-1})$. El hecho de que no haya ningún tipo de restricción sobre τ (que es el margen de error de los cuantificadores Q_n), hace que estos esquemas sean los preferidos para la conversión A/D de señales de audio.

Filtros finitos

Hemos mencionado al definir los esquemas $\Sigma\Delta$ de orden uno, que los coeficientes q_n con $n \leq 0$ se definen sólo por cuestiones teóricas. Esto se debe a que en la práctica, en la fórmula de la función aproximante $\tilde{f}_{\lambda}$, en lugar de utilizar las funciones g que cumplen (2.5), se usan núcleos G de soporte

compacto.
Obviamente al pedir que el soporte de G sea acotado, se pierde la condición de que $\hat{G}$ sea de soporte compacto, sin embargo veremos que sigue valiendo la acotación del orden de λ^{-1} en el error de aproximación.

Supongamos que $\operatorname{sop} G \subseteq [-R, R]$ y $G \in C^\infty$, entonces

$$\begin{aligned}\Big|f(t) - \frac{1}{\lambda}\sum_{n\in\mathbb{Z}} q_n^\lambda G\Big(t - \frac{n}{\lambda}\Big)\Big| \leq \quad & \Big|f(t) - \tfrac{1}{\lambda}\textstyle\sum_{n\in\mathbb{Z}} f\big(\frac{n}{\lambda}\big)G\big(t - \frac{n}{\lambda}\big)\Big| \\ & + \tfrac{1}{\lambda}\Big|\textstyle\sum_{n\in\mathbb{Z}}\big(f\big(\frac{n}{\lambda}\big) - q_n^\lambda\big)G\big(t - \frac{n}{\lambda}\big)\Big|.\end{aligned}$$

El segundo sumando puede ser acotado, usando los mismos argumentos que en la prueba del teorema 2.4.1, por

$$\frac{1}{\lambda}\Big|\sum_{n\in\mathbb{Z}}\Big(f\Big(\frac{n}{\lambda}\Big) - q_n^\lambda\Big)G\Big(t - \frac{n}{\lambda}\Big)\Big| \leq \frac{1}{\lambda}\|G'\|_{L^1(\mathbb{R})}.$$

Para el primer sumando podemos usar el resultado del lema 2.1.1 para la función $G - g$, donde g es tal que cumple (2.5). La acotación resulta entonces

$$\begin{aligned}\Big|f(t) - \frac{1}{\lambda}\sum_{n\in\mathbb{Z}} f\Big(\frac{n}{\lambda}\Big)G\Big(t - \frac{n}{\lambda}\Big)\Big| &= \frac{1}{\lambda}\Big|\sum_{n\in\mathbb{Z}} f\Big(\frac{n}{\lambda}\Big)\Big(g\Big(t - \frac{n}{\lambda}\Big) - G\Big(t - \frac{n}{\lambda}\Big)\Big)\Big| \\ &\leq \frac{1}{\lambda}\sum_{n\in\mathbb{Z}}\Big|g\Big(t - \frac{n}{\lambda}\Big) - G\Big(t - \frac{n}{\lambda}\Big)\Big| \\ &\leq \|G - g\|_{L^1(\mathbb{R})} + \frac{1}{\lambda}\|G' - g'\|_{L^1(\mathbb{R})}.\end{aligned}$$

Finalmente obtenemos que

$$\begin{aligned}&\Big|f(t) - \frac{1}{\lambda}\sum_{n\in\mathbb{Z}} q_n^\lambda G\Big(t - \frac{n}{\lambda}\Big)\Big| \leq \frac{1}{\lambda}\|G'\|_{L^1(\mathbb{R})} + \|G - g\|_{L^1(\mathbb{R})} + \frac{1}{\lambda}\|G' - g'\|_{L^1(\mathbb{R})} \\ &\leq \frac{1}{\lambda}\|G' - g'\|_{L^1(\mathbb{R})} + \frac{1}{\lambda}\|g'\|_{L^1(\mathbb{R})} + \|G - g\|_{L^1(\mathbb{R})} + \frac{1}{\lambda}\|G' - g'\|_{L^1(\mathbb{R})}.\end{aligned}$$

Fijada g que cumple (2.5), sea G_λ tal que $\|G_\lambda - g\|_{L^1(\mathbb{R})} \leq \frac{1}{\lambda}$ y $\|G_\lambda' - g'\|_{L^1(\mathbb{R})} \leq 1$, entonces el error al aproximar f por la reconstrucción

$$\tilde{f}_\lambda(t) = \frac{1}{\lambda}\sum_{n\in\mathbb{Z}} q_n^\lambda G_\lambda\Big(t - \frac{n}{\lambda}\Big) \tag{2.21}$$

va a ser del tipo $O(\lambda^{-1})$.

Como el soporte de G está contenido en $[-R, R]$, sólo se utilizarán en (2.21) los coeficientes q_n^λ con $|t - \frac{n}{\lambda}| \leq R$. Dado que en la práctica generalmente se busca aproximar los valores de $f(t)$ para $t > t_0$, resulta que sólo se considerarán los coeficientes con subíndice n tales que $n \geq t_0 - R$.

Este análisis justifica el hecho de que se le otorgue mayor relevancia a las recursiones del algoritmo $\Sigma\Delta$ definidas para $n \geq 1$.

2.4.2. Comparación entre los esquemas Sigma-Delta ($\Sigma\Delta$) y los de Modulación por Impulsos Codificados

En la sección 2.3 estudiamos el algoritmo de Modulación por Impulsos Codificados para funciones de banda limitada. Obtuvimos que el error global era acotado por

$$\|f - \tilde{f}_\lambda\|_{L^\infty} \leq 2^{-N} C_g,$$

donde $C_g = \lambda^{-1}\|g'\|_{L^1} + \|g\|_{L^1}$ y $N + 1$ es la cantidad de bits asignados a cada muestra.

En la sección correspondiente al algoritmo $\Sigma\Delta$ de orden uno, vimos que el error al cuantificar con esos esquemas es:

$$\|f - \tilde{f}_\lambda\|_\infty \leq \lambda^{-1}\|g'\|_{L^1}.$$

Comparando estas dos cotas vemos que con el algoritmo de Modulación por Impulsos Codificados obtenemos una aproximación de orden exponencial en N, mientras que con los $\Sigma\Delta$ es polinómica en λ.
Claramente, el grado de precisión logrado con el primer algoritmo es mucho mayor que con el segundo. Sin embargo, la eficacia del algoritmo $\Sigma\Delta$ radica en su robustez, tal como lo explicamos anteriormente.

Veremos a continuación que el costo de implementación del algoritmo PCM no es alto, ya que para la aplicación del mismo se puede utilizar un esquema muy similar al del $\Sigma\Delta$. Pero demostraremos que la gran desventaja del PCM es su falta de robustez.

Recordemos que el PCM utiliza como sucesión aproximante, aquella que proviene del desarrollo binario de las muestras hasta el coeficiente N-ésimo.

Presentamos a continuación un esquema introducido en [DDVGV], similar a los de $\Sigma\Delta$, que permite calcular el desarrollo binario de un número.

Dado $r \in [-1, 1]$, como $|r| \in [0, 1]$, su expansión binaria es de la forma

$$|r| = \sum_{i=1}^{\infty} b_i 2^{-i}, \text{ con } b_i \in \{0, 1\}.$$

Podemos decir entonces que

$$r = b_0 \sum_{i=1}^{\infty} b_i 2^{-i}, \text{ donde } b_0 = \text{sign}(r). \tag{2.22}$$

Sean

$$Q_0(x) := \text{sign}(x) = \begin{cases} -1 & x < 0 \\ 1 & x \geq 0 \end{cases}$$

y

$$Q_1(x) := \begin{cases} 0 & x < 1 \\ 1 & x \geq 1. \end{cases}$$

Buscamos definir recursivamente a partir de estos cuantificadores, a los coeficientes b_i, para $i \geq 0$, de la descomposición (2.22).
Resulta entonces que $b_0 = Q_0(r)$.
Para calcular b_1, recordemos que como es el primer coeficiente en el desarrollo binario de $|r|$, vale que

$$b_1 = \begin{cases} 0 & |r| < \frac{1}{2} \\ 1 & |r| \geq \frac{1}{2}, \end{cases}$$

por lo que podemos decir que $b_1 = Q_1(u_1)$ si $u_1 = 2|r|$.
En el caso de b_2 sucede que

$$b_2 = \begin{cases} 0 & |r| - b_1 2^{-1} < \frac{1}{4} \\ 1 & |r| - b_1 2^{-1} \geq \frac{1}{4}, \end{cases}$$

dicho de otra manera

$$b_2 = \begin{cases} 0 & 2(2|r| - b_1) < 1 \\ 1 & 2(2|r| - b_1) \geq 1, \end{cases}$$

con lo cual $b_2 = Q_1(u_2)$, donde $u_2 = 2(u_1 - b_1)$.
Si se continúa con estos razonamientos, se tiene el siguiente algoritmo que calcula los coeficientes b_{i+1} para $i \geq 1$:

$$\begin{cases} u_{i+1} = 2(u_i - b_i) \\ b_{i+1} = Q_1(u_{i+1}), \end{cases} \tag{2.23}$$

donde $u_1 = 2|r|$ y $b_1 = Q_1(u_1)$.

Veamos que el algoritmo es estable.

Proposición 2.4.5. *Dado $r \in [-1, 1]$, si definimos la sucesión $\{u_i\}_{i\in\mathbb{N}}$ a partir del esquema (2.23), entonces $u_i \in [0, 2]$ para todo $i \geq 1$.*

Demostración. Lo probamos por inducción.
Para $i = 1$ vale pues $|r| \in [0, 1]$ y $u_1 = 2|r|$.
Supongamos vale para i, queremos ver que vale para $i+1$. Tenemos dos casos:

Si $1 \leq u_i \leq 2$, entonces $b_i = Q_1(u_i) = 1$, por lo que $0 \leq u_{i+1} = 2(u_i - 1) \leq 2$.

Si $0 \leq u_i < 1$, entonces $b_i = Q_1(u_i) = 0$, por lo que $0 \leq u_{i+1} = 2u_i \leq 2$. □

Si definimos la sucesión $\{z_n\}_{n\in\mathbb{N}_0}$ como $z_0 = |r|$ y $z_i = 0$ para todo $i \geq 1$, usando (2.23), el algoritmo para calcular la expansión binaria de $|r|$ resulta ser

$$\begin{cases} u_{i+1} = 2(u_i + z_i - b_i) \\ b_{i+1} = Q_1(u_{i+1}), \end{cases}$$

para todo $i \geq 0$, tomando a $u_0 = 0$ y $b_0 = 0$.
Esto nos muestra que el esquema anterior es muy parecido al $\Sigma\Delta$ (salvo el factor multiplicativo 2), por lo que el costo de implementación de ambos métodos es el mismo.

Aplicando el algoritmo (2.23) hasta el paso N, obtenemos la expansión binaria de $|r|$ hasta la cifra N-ésima. Dado que la expansión binaria de r es

$$r = b_0 \sum_{i=1}^{\infty} b_i 2^{-i},$$

el elemento aproximante obtenido a partir del esquema (2.23), tomando $b_0 = Q_0(r)$, será

$$\tilde{r}_N = b_0 \sum_{i=1}^{N} b_i 2^{-i}.$$

Si definimos el error de aproximación como $E_N = |r - \tilde{r}_N|$, entonces $E_N \leq 2^{-N}$.

El esquema (2.23) presenta dos grandes ventajas: tiene bajo costo de implementación y logra un error de aproximación de tipo exponencial. La gran desventaja de este esquema, que lo hace inaplicable a los fines prácticos, es su falta de robustez.

Como la señal original es analógica, el cuantificador Q_1 va a ser generalmente analógico, por lo que es probable que al hacer el cambio rotundo del valor 1 al 0, introduzca pequeños errores.
Supongamos que el cuantificador Q_1 se desvía un poco de su definición original, presentándose como el siguiente cuantificador con error δ:

$$Q_{1,\delta} = \begin{cases} 0 & x < 1 + \delta \\ 1 & x \geq 1 + \delta, \end{cases}$$

sabiendo sólo que $|\delta| < \tau$, con $\tau > 0$ conocido.
Nuestro esquema sería para $i \geq 1$:

$$\begin{cases} u_{i+1} = 2(u_i - b_i) \\ \widetilde{b}_{i+1} = Q_{1,\delta}(u_{i+1}), \end{cases} \tag{2.24}$$

con $u_1 = 2|r|$, $\widetilde{b}_1 = Q_{1,\delta}(u_1)$ y $\widetilde{b}_0 = \text{sign}(r)$.
Si $|r| \in (\frac{1}{2}, \frac{1+\delta}{2})$, entonces $2|r| < 1 + \delta$, por lo que $\widetilde{b}_1 = Q_{1,\delta}(2|r|) = 0$. Pero como $|r| > 1/2$ debería ser $b_1 = 1$.
El elemento aproximante con este nuevo esquema (2.24) quedaría acotado de la siguiente forma

$$|\tilde{r}_N| = \sum_{i=1}^{N} \widetilde{b}_i 2^{-i} = \sum_{i=2}^{N} \widetilde{b}_i 2^{-i} \leq \sum_{i=2}^{N} 2^{-i} = \frac{1}{2} - 2^{-N},$$

entonces

$$|r| - |\tilde{r}_N| \geq |r| - \frac{1}{2} + 2^{-N} > |r| - \frac{1}{2} > 0.$$

Este pequeño error en el cuantificador Q_1, lleva a la siguiente cota mínima para el error $\widetilde{E}_N$:

$$\widetilde{E}_N = |r - \tilde{r}_N| = |\widetilde{b}_0||r| - |\tilde{r}_N|| = ||r| - |\tilde{r}_N|| > |r| - \frac{1}{2}. \tag{2.25}$$

Esto nos dice que aunque incrementemos la cantidad de bits N asignados a cada muestra, el error no va a poder achicarse. De hecho éste puede llegar a ser de hasta $\frac{\delta}{2}$ (pues $|r| \in (\frac{1}{2}, \frac{1+\delta}{2})$), y por consiguiente $\frac{\tau}{2}$. Entonces para pedir que el error sea pequeño se debe pedir que τ lo sea, es decir, que el cuantificador Q_1 sea preciso, lo que acarrea aumento de costos.

Supongamos ahora que el cuantificador Q_0 introduce errores, presentándose como

$$Q_{0,\delta} = \begin{cases} -1 & x < \delta \\ 1 & x \geq \delta, \end{cases}$$

y supongamos que $r \in [0, \delta)$. Resultará entonces que $\widetilde{b}_0 = Q_{0,\delta}(r) = -1$, cuando en realidad debería valer uno, pues r es mayor o igual que 0.
En este caso el elemento aproximante resulta

$$\tilde{r}_N = -\sum_{i=1}^{N} b_i 2^{-i},$$

por lo que

$$\widetilde{E}_N = |r - \tilde{r}_N| = \sum_{i=1}^{\infty} b_i 2^{-i} + \sum_{i=1}^{N} b_i 2^{-i} = r + |\tilde{r}_N| \geq r. \tag{2.26}$$

Lo que prueba nuevamente que el error no podrá achicarse aunque N tienda a infinito.

Con estos dos ejemplos hemos probado que el algoritmo PCM, a diferencia del $\Sigma\Delta$, no es robusto frente a la introducción de errores en los cuantificadores.

Otra cuestión que surge al analizar los ejemplos, es que la cota mínima del error E_N cambia según cual sea el coeficiente b_i en el que se introduce el error. Esto nos muestra que el algoritmo PCM no le da un trato igualitario a todos los elementos de la sucesión cuantificadora. Veamos que esto no sucede con los esquemas $\Sigma\Delta$.

Si suponemos que la sucesión a cuantificar $\{s_n\}_{n\in\mathbb{N}}$, toma constantemente el valor $r \in \mathbb{R}$, con $-1 \leq r \leq 1$, en la observación 2.4.3 vimos que

$$\Big|r - \frac{1}{N}\sum_{k=1}^{N} q_k\Big| = \Big|\frac{1}{N}\sum_{k=1}^{N} r - \frac{1}{N}\sum_{k=1}^{N} q_k\Big| = \Big|\frac{1}{N}\sum_{k=1}^{N} s_k - \frac{1}{N}\sum_{k=1}^{N} q_k\Big| \leq \frac{2}{N}.$$

Si se cometiese un error ε_{n_0} en el cálculo del coeficiente q_{n_0}, obtendríamos una nueva sucesión cuantificadora $\{\tilde{q}_n\}_{n\in\mathbb{N}}$, definida como $\tilde{q}_n = q_n$ si $n \neq n_0$ y $\tilde{q}_{n_0} = \varepsilon_{n_0} + q_{n_0}$. Sucedería entonces que

$$\Big|r - \frac{1}{N}\sum_{k=1}^{N} \tilde{q}_k\Big| = \Big|r - \frac{1}{N}\sum_{k=1}^{N} q_k\Big| + \Big|\frac{1}{N}(q_{n_0} - \tilde{q}_{n_0})\Big| \leq \frac{2}{N} + \frac{\varepsilon_{n_0}}{N}.$$

Los promedios seguirían aproximando al valor r con orden de N^{-1}, independientemente del coeficiente en el que se introdujo el error.

Este trato igualitario de todos los q_n en los algoritmos $\Sigma\Delta$ es estudiado en [CD] como el concepto de cuantificación "democrática". El algoritmo PCM es considerado no "democrático", ya que a partir de (2.25) y (2.26) podemos decir que la cota mínima del error depende de si el bit modificado es b_1 ó b_0.

2.4.3. Cuantificación Sigma-Delta ($\Sigma\Delta$) de órdenes mayores

Con los esquemas $\Sigma\Delta$ de orden uno obtuvimos una cota para el error global del tipo $O(\lambda^{-1})$. Los esquemas de mayor orden nos van a permitir mejorar esa cota.

Recordemos que el algoritmo de orden uno para una sucesión $\{x_n\}_{n\in\mathbb{N}}$ de números reales, tal que $|x_n| \leq 1$ para todo $n \in \mathbb{N}$, era

$$\begin{cases} u_n = u_{n-1} + x_n - q_n \\ q_n = \text{sign}(u_{n-1} + x_n). \end{cases}$$

Al estudiar este esquema, introdujimos al operador diferencial discreto Δ, que en la sucesión $s = \{s_n\}_{n\in\mathbb{Z}}$ actuaba de la siguiente manera:

$$\Delta_n s = s_n - s_{n-1} \quad \forall\ n \in \mathbb{Z}.$$

En realidad, habíamos definido Δ para subíndices naturales, pero se puede extender su aplicación a todos los enteros.

Definición 2.4.7. *Dado $k \geq 1$, el operador diferencial discreto de orden k en la sucesión $s = \{s_n\}_{n\in\mathbb{Z}}$, se define de la siguiente forma:*

$$\Delta_n^k s := \Delta_n^1\big(\Delta^{k-1}s\big), \quad con$$
$$\Delta_n^1 s := s_n - s_{n-1} \ y \ \Delta_n^0 s := s_n.$$

Observación 2.4.5. *A partir de la definición 2.4.7 se puede probar que*

$$\Delta_n^k s = \sum_{j=0}^{k}(-1)^j \binom{k}{j} s_{n-j}, \quad \forall\ k \geq 0 \ y\ n \in \mathbb{Z}.$$

Utilizando el operador diferencial discreto, el sistema de orden uno para subíndices naturales, puede ser interpretado como una ecuación diferencial discreta de la siguiente manera:

$$\begin{cases} \Delta_n^1 u & = x_n - q_n \\ q_n & = \operatorname{sign}\big(F\big(\Delta_{n-1}^0 u, x_n\big)\big), \end{cases}$$

donde $F(x, y) = x + y$.
El algoritmo $\Sigma\Delta$ de orden k es una generalización del algoritmo anterior, en lugar de utilizar los operador diferencial discreto de orden uno, se usa el de orden k.

Definición 2.4.8. *Dados $k \geq 1$ y $\lambda > 1$. Si $\{x_n\}_{n\in\mathbb{N}}$ es tal que $|x_n| \leq 1$ para todo $n \in \mathbb{N}$, el algoritmo $\Sigma\Delta$ de orden k con el cuantificador escalar $Q(x) = sign(x)$, se define para $n \geq 1$, como:*

$$\begin{cases} \Delta_n^k u & = x_n - q_n \\ q_n & = sign\big(F\big(\Delta_{n-1}^0 u, \Delta_{n-1}^1 u, \ldots, \Delta_{n-1}^{k-1} u, x_n\big)\big), \end{cases}$$

donde $u_0, u_{-1}, \ldots, u_{-k+1}$ son dados como dato y $F : \mathbb{R}^{k+1} \longrightarrow \mathbb{R}$ es una función elegida de forma tal que el sistema resulte estable, es decir, que la sucesión $\{u_n\}_{n\in\mathbb{N}}$ sea acotada.

Observemos que la definición anterior corresponde a una ecuación diferencial discreta de orden k.

Nosotros utilizaremos el esquema anterior aplicado al caso en que $x_n^\lambda = f\big(\frac{n}{\lambda}\big)$.

La prueba de la estabilidad y la robustez para los sistemas de orden uno era sencilla. Para órdenes mayores resulta bastante complicada, de hecho se necesitan nociones de diferentes campos de la matemática, pues se utilizan resultados de análisis armónico, teoría de números y sistemas dinámicos.

En su trabajo [DDV], Ron DeVore e Ingrid Daubechies fueron los primeros en construir una familia de esquemas $\Sigma\Delta$ estable de orden arbitrario. Incluso lograron que el algoritmo tenga la característica de ser robusto, es decir, al modificar la función $\text{sign}(x)$ por $\text{sign}(x+\tau)$ y las constantes multiplicativas M en la función F por $M+\varepsilon$, el sistema seguía teniendo variables internas uniformemente acotadas.
Antes de especificar la familia con la que se trabajó en [DDV], veamos cómo resulta la cota del error en el caso en que el algoritmo de orden k sea estable.

Error de aproximación para algoritmos $\Sigma\Delta$ estables de orden arbitrario

Para la prueba del teorema de acotación del error necesitaremos una definición y algunos lemas que se encuentran en [Ch].

Definición 2.4.9. *Se define la B-spline de orden k como:*

$$\varphi_k := \underbrace{\chi_{[0,1]} * \ldots * \chi_{[0,1]}}_{k \text{ veces}},$$

donde $\chi_{[0,1]}$ *es la función característica del intervalo* $[0,1]$.

Proposición 2.4.6. *La B-spline de orden k tiene las siguientes propiedades:*

i) $\varphi_k \geq 0$.

ii) $sop\,\varphi_k \subseteq [0,k]$.

Demostración. i) Es trivial, ya que φ_k es la convolución de funciones no negativas.

ii) Vale pues,

$$\begin{aligned} \text{sop}(\chi_{[0,1]} * \ldots * \chi_{[0,1]}) &\subseteq \text{sop}\chi_{[0,1]} + \ldots + \text{sop}\chi_{[0,1]} \\ &\subseteq [0,1] + \ldots + [0,1] = [0,k]. \end{aligned}$$

□

Lema 2.4.1. *Si φ_k es la B-spline de orden k, entonces:*

a) Dada $f \in L^1(\mathbb{R})$, vale que

$$\int_{\mathbb{R}} f(x)\varphi_k(x)\,dx = \int_0^1 \dots \int_0^1 f(x_1 + \dots + x_k)\,dx_1 \dots dx_k.$$

b) Si $g \in C^k(\mathbb{R})$ y $g^{(k)} \in L^1(\mathbb{R})$, entonces

$$\int_{\mathbb{R}} g^{(k)}(x)\varphi_k(x)\,dx = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} g(j).$$

Demostración. a) Se prueba por inducción.
Para $k = 1$,

$$\int_{\mathbb{R}} f(x)\varphi_1(x)\,dx = \int_{\mathbb{R}} f(x)\chi_{[0,1]}(x)\,dx = \int_0^1 f(x)dx.$$

Supongamos que vale para k, para probar la propiedad en $k+1$ usamos que $\varphi_{k+1} = \varphi_k * \chi_{[0,1]}$.

$$\begin{aligned}
\int_{\mathbb{R}} f(x)\varphi_{k+1}(x)\,dx &= \int_{\mathbb{R}} f(x) \int_{\mathbb{R}} \varphi_k(x-t)\chi_{[0,1]}(t)\,dtdx \\
&= \int_{\mathbb{R}} f(x) \int_0^1 \varphi_k(x-t)\,dtdx \\
&= \int_{\mathbb{R}} \int_0^1 f(x+t)\varphi_k(x)\,dtdx \\
&= \int_{\mathbb{R}} \left(\int_0^1 f(x+t)\,dt \right) \varphi_k(x)\,dx \\
&= \int_0^1 \dots \int_0^1 \int_0^1 f(x_1 + \dots + x_k + t)\,dtdx_1 \dots dx_k.
\end{aligned}$$

En la última igualdad utilizamos la hipótesis inductiva.

b) Usando el item a) vale que

$$\int_{\mathbb{R}} g^{(k)}(x)\varphi_k(x)\,dx = \int_0^1 \dots \int_0^1 g^{(k)}(x_1 + \dots + x_k)\,dx_1 \dots dx_k.$$

Resta probar que

$$\int_0^1 \dots \int_0^1 g^{(k)}(x_1 + \dots + x_k)\, dx_1 \dots dx_k = \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} g(j).$$

Esto se demuestra por inducción.
Para $k = 1$

$$\int_0^1 g'(x)\, dx = g(1) - g(0).$$

Suponiendo que vale para k,

$$\begin{aligned}
&\int_0^1 \dots \int_0^1 g^{(k+1)}(x_1 + \dots + x_{k+1})\, dx_1 \dots dx_{k+1} \\
&= \int_0^1 \dots \int_0^1 \Big(g^{(k)}(x_1 + \dots + x_k + 1) - g^{(k)}(x_1 + \dots + x_k) \Big)\, dx_1 \dots dx_k \\
&= \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} g(j+1) - \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} g(j) \\
&= \sum_{j=1}^{k+1} (-1)^{k-j+1} \binom{k}{j-1} g(j) - \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} g(j) \\
&= -\sum_{j=1}^{k} (-1)^{k-j} \left(\binom{k}{j-1} + \binom{k}{j} \right) g(j) + g(k+1) - (-1)^k g(0) \\
&= \sum_{j=1}^{k} (-1)^{k+1-j} \binom{k+1}{j} g(j) + g(k+1) + (-1)^{k+1} g(0) \\
&= \sum_{j=0}^{k+1} (-1)^{k+1-j} \binom{k+1}{j} g(j).
\end{aligned}$$

□

Lema 2.4.2. *Si φ_k es la B-spline de orden k, entonces*

$$\sum_{m \in \mathbb{Z}} \varphi_k(y + m) = 1 \qquad \forall\ y \in \mathbb{R}.$$

Demostración. Usando el item a) del lema anterior, y que $\sum_{m\in\mathbb{Z}} \chi_{[-m,1-m)}(z) = 1$ para todo $z \in \mathbb{R}$, tenemos que

$$\begin{aligned}
\sum_{m\in\mathbb{Z}} \varphi_k(y+m) &= \sum_{m\in\mathbb{Z}} \int_{\mathbb{R}} \varphi_{k-1}(t)\chi_{[0,1]}(y+m-t)\,dt \\
&= \sum_{m\in\mathbb{Z}} \int_{\mathbb{R}} \varphi_{k-1}(t)\chi_{[0,1)}(y+m-t)\,dt \\
&= \sum_{m\in\mathbb{Z}} \int_0^1 \dots \int_0^1 \chi_{[0,1)}\big(y+m-(x_1+\dots+x_{k-1})\big)\,dx_1\dots dx_{k-1} \\
&= \int_0^1 \dots \int_0^1 \sum_{m\in\mathbb{Z}} \chi_{[-m,1-m)}\big(y-(x_1+\dots+x_{k-1})\big)\,dx_1\dots dx_{k-1} \\
&= \int_0^1 \dots \int_0^1 1\,dx_1\dots dx_{k-1} = 1.
\end{aligned}$$

□

Lema 2.4.3. *Si $g \in C^k(\mathbb{R})$, $g^{(k)} \in L^1(\mathbb{R})$ y $\varphi_k(x)$ es la B-spline de orden k, entonces*

$$\sum_{j=0}^{k}(-1)^j \binom{k}{j} g\Big(t-\frac{n+j}{\lambda}\Big) = \frac{1}{\lambda^{k-1}} \int_0^{\frac{k}{\lambda}} g^{(k)}\left(t-\frac{n+k}{\lambda}+s\right)\varphi_k(\lambda s)\,ds.$$

Demostración. Usando el item ii) de la proposición 2.4.6,

$$\begin{aligned}
&\frac{1}{\lambda^{k-1}} \int_0^{\frac{k}{\lambda}} g^{(k)}\left(t-\frac{n+k}{\lambda}+s\right)\varphi_k(\lambda s)\,ds \\
=&\frac{1}{\lambda^k} \int_0^{k} g^{(k)}\left(t-\frac{n+k}{\lambda}+\frac{u}{\lambda}\right)\varphi_k(u)\,du \\
=&\frac{1}{\lambda^k} \int_{\mathbb{R}} g^{(k)}\left(t-\frac{n+k}{\lambda}+\frac{u}{\lambda}\right)\varphi_k(u)\,du.
\end{aligned}$$

Sea $h(u) := g\big(t-\frac{n+k}{\lambda}+\frac{u}{\lambda}\big)$, entonces $h^{(k)}(u) = \frac{1}{\lambda^k} g^{(k)}\big(t-\frac{n+k}{\lambda}+\frac{u}{\lambda}\big)$, por lo

que usando el item b) del lema 2.4.1, tenemos que

$$\begin{aligned}
&\frac{1}{\lambda^{k-1}}\int_0^{\frac{k}{\lambda}} g^{(k)}\Big(t-\frac{n+k}{\lambda}+s\Big)\varphi_k(\lambda s)\,ds\\
&=\int_{\mathbb{R}} h^{(k)}(u)\varphi_k(u)\,du=\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}h(j)\\
&=\sum_{j=0}^{k}(-1)^{k-j}\binom{k}{j}g\Big(t-\frac{n+(k-j)}{\lambda}\Big).\\
&=\sum_{j=0}^{k}(-1)^{j}\binom{k}{j}g\Big(t-\frac{n+j}{\lambda}\Big).
\end{aligned}$$

□

Teorema 2.4.3. *Sea $f\in\mathcal{C}$, $\lambda>1$ y g que satisface (2.5). Si la sucesión $\{q_n^\lambda\}_{n\in\mathbb{Z}}$ cumple que $\Delta_n^k u^\lambda = f\big(\frac{n}{\lambda}\big)-q_n^\lambda$ para todo $n\in\mathbb{Z}$, donde $\{u_n^\lambda\}_{n\in\mathbb{Z}}$ es una sucesión acotada. Entonces vale que*

$$|f(t)-\tilde{f}_\lambda(t)|\le\frac{1}{\lambda^k}\|u^\lambda\|_{l^\infty}\|g^{(k)}\|_{L^1}\quad \textit{para todo } t\in\mathbb{R}.$$

Demostración. Utilizando la observación 2.4.5 obtenemos que

$$\begin{aligned}
|f(t)-\tilde{f}_\lambda(t)| &= \frac{1}{\lambda}\Big|\sum_{n\in\mathbb{Z}}\Big(f\Big(\frac{n}{\lambda}\Big)-q_n^\lambda\Big)g\Big(t-\frac{n}{\lambda}\Big)\Big|\\
&= \frac{1}{\lambda}\Big|\sum_{n\in\mathbb{Z}}\Delta_n^k(u^\lambda)g\Big(t-\frac{n}{\lambda}\Big)\Big|\\
&= \frac{1}{\lambda}\Big|\sum_{n\in\mathbb{Z}}\Big(\sum_{j=0}^{k}(-1)^j\binom{k}{j}u_{n-j}^\lambda\Big)g\Big(t-\frac{n}{\lambda}\Big)\Big|\\
&= \frac{1}{\lambda}\Big|\sum_{n\in\mathbb{Z}}u_n^\lambda\Big(\sum_{j=0}^{k}(-1)^j\binom{k}{j}g\Big(t-\frac{n+j}{\lambda}\Big)\Big)\Big|. \qquad (2.27)
\end{aligned}$$

Esto último vale por la sumación por partes, es decir

$$\begin{aligned}\sum_{n\in\mathbb{Z}}\Big(\sum_{j=0}^{k}(-1)^j\binom{k}{j}u^\lambda_{n-j}\Big)g\Big(t-\frac{n}{\lambda}\Big) &= \sum_{j=0}^{k}(-1)^j\binom{k}{j}\Big(\sum_{n\in\mathbb{Z}}u^\lambda_{n-j}\,g\Big(t-\frac{n}{\lambda}\Big)\Big)\\ &= \sum_{j=0}^{k}(-1)^j\binom{k}{j}\Big(\sum_{n\in\mathbb{Z}}u^\lambda_{n}g\Big(t-\frac{n+j}{\lambda}\Big)\Big)\\ &= \sum_{n\in\mathbb{Z}}u^\lambda_n\Big(\sum_{j=0}^{k}(-1)^j\binom{k}{j}g\Big(t-\frac{n+j}{\lambda}\Big)\Big).\end{aligned}$$

Usando el lema 2.4.3, el item i) y ii) de la proposición 2.4.6 y el lema 2.4.2, a partir de (2.27) tenemos que:

$$\begin{aligned}|f(t)-\tilde{f}_\lambda(t)| &= \frac{1}{\lambda}\Big|\sum_{n\in\mathbb{Z}}u^\lambda_n\Big(\sum_{j=0}^{k}(-1)^j\binom{k}{j}g\Big(t-\frac{n+j}{\lambda}\Big)\Big)\Big|\\ &= \frac{1}{\lambda}\Big|\sum_{n\in\mathbb{Z}}u^\lambda_n\Big(\frac{1}{\lambda^{k-1}}\int_0^{\frac{k}{\lambda}}g^{(k)}\Big(t-\frac{n+k}{\lambda}+s\Big)\varphi_k(\lambda s)\,ds\Big)\Big|\\ &= \frac{1}{\lambda}\Big|\sum_{n\in\mathbb{Z}}u^\lambda_n\Big(\frac{1}{\lambda^{k-1}}\int_{-\infty}^{+\infty}g^{(k)}\Big(t-\frac{n+k}{\lambda}+s\Big)\varphi_k(\lambda s)\,ds\Big)\Big|\\ &\le \frac{1}{\lambda^k}\sum_{n\in\mathbb{Z}}|u^\lambda_n|\Big(\int_{-\infty}^{+\infty}\Big|g^{(k)}\Big(t-\frac{n+k}{\lambda}+s\Big)\Big|\varphi_k(\lambda s)\Big)\,ds\\ &= \frac{1}{\lambda^k}\sum_{n\in\mathbb{Z}}|u^\lambda_n|\Big(\int_{-\infty}^{+\infty}|g^{(k)}(y)|\varphi_k(\lambda y-\lambda t+n+k)\Big)\,dy\\ &\le \frac{1}{\lambda^k}\|u^\lambda\|_{l^\infty}\int_{-\infty}^{+\infty}|g^{(k)}(y)|\Big(\sum_{n\in\mathbb{Z}}\varphi_k(\lambda y-\lambda t+n+k)\Big)\,dy\\ &= \frac{1}{\lambda^k}\|u^\lambda\|_{l^\infty}\int_{-\infty}^{+\infty}|g^{(k)}(y)|\,dy=\frac{1}{\lambda^k}\|u^\lambda\|_{l^\infty}\|g^{(k)}\|_{L^1}.\end{aligned}$$

□

Recordemos que la función g dependía de λ, pero que habíamos acordado notar $g = g_\lambda$. Suponiendo que elegimos $g_\lambda = g_{\lambda_0}$ para todo $\lambda \ge \lambda_0$, el teorema nos dice que, si la sucesión $\{u^\lambda_n\}_{n\in\mathbb{Z}}$ está uniformemente acotada por

una constante independiente de λ, con el algoritmo de orden k, el error es del tipo $O(\lambda^{-k})$. Esto prueba que incrementando el orden en un sistema estable, se logra mayor precisión en la cuantificación.

Estabilidad y robustez

En esta sección trabajaremos con esquemas definidos para una sucesión $\{x_n\}_{n\in\mathbb{N}}$ que cumple $|x_n| \leq 1$ para todo $n \in \mathbb{N}$. Tomando $x_n^\lambda = f\left(\frac{n}{\lambda}\right)$, se aplican los resultados para la cuantificación de las muestras.

La principal dificultad en órdenes mayores que uno es la prueba de la estabilidad y de la robustez del sistema. Estudiaremos los esquemas de orden dos para entender los aspectos que resultan conflictivos.

La definición 2.4.8 establece que el algoritmo de orden dos para $n \geq 1$ es:

$$\begin{cases} \Delta_n^2 u = x_n - q_n \\ q_n = \operatorname{sign}\big(F\big(\Delta_{n-1}^0 u, \Delta_{n-1}^1 u, x_n\big)\big). \end{cases}$$

Copiando los métodos utilizados para ecuaciones diferenciales en variables continuas, se define una nueva variable $v_n := \Delta_n^1 u = u_n - u_{n-1}$. A partir de esto, el esquema anterior puede ser reescrito como:

$$\begin{cases} \Delta_n^1 v = x_n - q_n \\ \Delta_n^1 u = v_n \\ q_n = \operatorname{sign}(F(u_{n-1}, v_{n-1}, x_n)), \end{cases}$$

es decir,

$$\begin{cases} v_n = v_{n-1} + x_n - q_n \\ u_n = u_{n-1} + v_n \\ q_n = \operatorname{sign}(F(u_{n-1}, v_{n-1}, x_n)). \end{cases} \tag{2.28}$$

Como dijimos anteriormente la estabilidad del sistema va a depender de la elección de F. Se ha estudiado el comportamiento del algoritmo (2.28) con distintas funciones, por ejemplo:
Yilmaz trabajó en [Yi] con

$$F(u, v, x) = v + \gamma u,$$

Thao utilizó en [Th]:

$$F(u,v,x) = \frac{6x - 7\text{sign}(x)}{3} + \Big(v + \frac{x + 3 + \text{sign}(x)}{2}\Big)^2 + 2(1 - |x|)u.$$

Para probar la estabilidad, ambos demostraron que existe un conjunto $A_\alpha \subseteq \mathbb{R}^2$ acotado, tal que si $|x_n| \leq \alpha < 1$ para todo n y $(u_0, v_0) \in A_\alpha$, entonces $(u_n, v_n) \in A_\alpha$ para todo n.
Veremos en el capítulo correspondiente a la cuantificación de vectores de $\mathbb{R}^d$, una idea general de la prueba utilizada por Yilmaz, para la demostración de lo anteriormente dicho.

Ron DeVore e Ingrid Daubechies utilizaron en [DDV] un método distinto para probar la estabilidad de su sistema, que a diferencia del de Yilmaz, se extiende a mayores órdenes.
Ellos trabajaron con

$$F(u,v,x) = v + x + M\,\text{sign}(u),$$

donde $M > 1$.
No sólo probaron que su algoritmo es estable, sino que modificando la función $\text{sign}(x)$ por $\text{sign}(x + \tau)$, y la constante M por $M + \varepsilon$, probaron la robustez del sistema.

Definición 2.4.10. *Dado $0 < \mu < 1$, el esquema utilizado en [DDV] para $n \geq 1$ es:*

$$\begin{cases} v_n = v_{n-1} + x_n - q_n \\ u_n = u_{n-1} + v_n \\ q_n = Q_n^1[v_{n-1} + x_n + M(1 + \varepsilon_n)Q_n^2(u_{n-1})], \end{cases} \tag{2.29}$$

donde Q_n^1 y Q_n^2 cumplen (2.18), $|\varepsilon_n| \leq \mu$ para todo $n \geq 1$ y u_0, v_0 son dados.

Para deducir el algoritmo en el caso de subíndices negativos, observemos que para $n \geq 1$, en los esquemas de orden dos se cumple que

$$\Delta_n^2 u = u_n - 2u_{n-1} + u_{n-2} = x_n - q_n,$$

por lo que podemos decir que

$$\Delta_{n+2}^2 u = u_{n+2} - 2u_{n+1} + u_n = x_{n+2} - q_{n+2}.$$

Esto nos permite definir el algoritmo para $n < -2$ como

$$\begin{cases} u_{n+2} - 2u_{n+1} + u_n &= x_{n+2} - q_{n+2} \\ q_{n+2} &= Q_n^1[v_{n+1} + x_{n+2} + M(1+\varepsilon_n)Q_n^2(u_{n+1})]. \end{cases}$$

Tomando $v_n = u_n - u_{n+1}$, resulta que el algoritmo es:

Definición 2.4.11. *Sean $0 < \mu < 1$, u_0, v_0 los definidos en (2.29). Si $v_{-1} = -v_0$ y $u_{-1} = u_0 - v_0$, se define el algoritmo para $n \leq -2$ como:*

$$\begin{cases} v_n &= v_{n+1} + x_{n+2} - q_{n+2} \\ u_n &= u_{n+1} + v_n \\ q_{n+2} &= Q_n^1[v_{n+1} + x_{n+2} + M(1+\varepsilon_n)Q_n^2(u_{n+1})], \end{cases} \tag{2.30}$$

donde Q_n^1, Q_n^2 cumplen (2.18) y $|\varepsilon_n| \leq \mu$ para todo $n \leq -2$.

A partir de la definición de los esquemas (2.29) y (2.30), se puede decir que

$$\Delta_n^2 u = x_n - q_n \quad \forall\ n \in \mathbb{Z}.$$

Con lo cual, para poder aplicar el resultado obtenido en el teorema 2.4.3, debemos probar la acotación uniforme de $\{u_n\}_{n\in\mathbb{Z}}$.

La siguiente proposición demostrada en [DDV] afirma la estabilidad del sistema de orden dos presentado anteriormente.

Proposición 2.4.7. *Supongamos que $|x_n| \leq \alpha < 1$ para todo $n \in \mathbb{Z}$, sean $\{u_n\}_{n\in\mathbb{Z}}$, $\{v_n\}_{n\in\mathbb{Z}}$, y $\{q_n\}_{n\in\mathbb{Z}}$ definidos a partir de (2.29) y (2.30), con $M \geq \frac{2\alpha+\tau+1}{1-\mu}$. Entonces, si $|v_0| \leq M(1+\mu) + 1 + \tau$, resulta que $|v_n| \leq M(1+\mu) + 1 + \tau$ para todo $n \in \mathbb{Z}$. Más aún, si $|u_0|, |v_0| \leq \frac{\tau}{2}$, vale que $|u_n| \leq \tau + \frac{[M(1+\mu)+\tau+3/2-\alpha/2]^2}{2(1-\alpha)}$ para todo $n \in \mathbb{Z}$.*

Para probar esta proposición se usan dos lemas, realizaremos la demostración del primero, la otra demostración no fue incluida debido a que los argumentos para su prueba son similares a los utilizados en el primer lema.

Lema 2.4.4. *Si $|v_0| \leq M(1+\mu) + 1 + \tau$, entonces $|v_n| \leq M(1+\mu) + 1 + \tau$ para todo $n \in \mathbb{Z}$.*

Demostración. Probaremos el lema para el caso $n \in \mathbb{N}$ por inducción, para los subíndices negativos se procede de manera análoga.
Para $n = 0$ vale por hipótesis.
Supongamos $|v_{n-1}| \leq M(1+\mu) + 1 + \tau$.
Si $|v_{n-1} + x_n| > M(1+\varepsilon_n) + \tau$, entonces usando que $|Q_n^2| \leq 1$,

$$\begin{aligned} |v_{n-1} + x_n + M(1+\varepsilon_n)Q_n^2(u_{n-1})| &\geq |v_{n-1} + x_n| - M(1+\varepsilon_n)|Q_n^2(u_{n-1})| \\ &\geq M(1+\varepsilon_n) + \tau - M(1+\varepsilon_n) = \tau. \end{aligned}$$

Como Q_n^1 cumple (2.18),

$$Q_n^1(v_{n-1} + x_n + M(1+\varepsilon_n)Q_n^2(u_{n-1})) = \operatorname{sign}(v_{n-1} + x_n + M(1+\varepsilon_n)Q_n^2(u_{n-1})).$$

Si $v_{n-1} + x_n > 0$, entonces $v_{n-1} + x_n > M(1+\varepsilon_n) + \tau$, por lo que

$$\begin{aligned} v_{n-1} + x_n + M(1+\varepsilon_n)Q_n^2(u_{n-1}) &> M(1+\varepsilon_n) + \tau + M(1+\varepsilon_n)Q_n^2(u_{n-1}) \\ &> M(1+\varepsilon_n) + \tau + M(1+\varepsilon_n)(-1) = \tau > 0. \end{aligned}$$

Entonces $q_n = 1$, y usando la acotación sobre M, resulta que

$$\begin{aligned} v_n = v_{n-1} + x_n - 1 = |v_{n-1} + x_n| - 1 > M(1+\varepsilon_n) + \tau - 1 \\ > M(1-\mu) + \tau - 1 > 2\alpha + \tau + 1 + \tau - 1 > 2\alpha + 2\tau > 0. \end{aligned}$$

Finalmente queda que

$$|v_n| = v_n = |v_{n-1} + x_n| - 1 \leq |v_{n-1}| + \alpha - 1 < M(1+\mu) + 1 + \tau.$$

Si $v_{n-1} + x_n < 0$, entonces $v_{n-1} + x_n = -|v_{n-1} + x_n| < -M(1+\varepsilon_n) - \tau$, por lo que

$$v_{n-1} + x_n + M(1+\varepsilon_n)Q_n^2(u_{n-1}) < -M(1+\varepsilon_n) - \tau + M(1+\varepsilon_n) = -\tau.$$

Entonces $q_n = -1$, y resulta que

$$v_n = v_{n-1} + x_n + 1 = -|v_{n-1} + x_n| + 1 < -M(1+\varepsilon_n) - \tau + 1 < 0.$$

Finalmente queda que

$$|v_n| = |v_{n-1} + x_n| - 1 \leq |v_{n-1}| + \alpha - 1 < M(1+\mu) + 1 + \tau.$$

En el caso en que $|v_{n-1} + x_n| \leq M(1+\varepsilon_n) + \tau$, sucede que

$$|v_n| \leq |v_{n-1} + x_n| + 1 \leq M(1+\varepsilon_n) + \tau + 1 \leq M(1+\mu) + 1 + \tau.$$

□

Lema 2.4.5. *Sea $u_k \leq \tau$, y $u_{k+1}, u_{k+2}, \ldots, u_{k+L} > \tau$. Si existe $j \in \{1, \ldots, L\}$ tal que $v_{k+j} + x_{k+j+1} < -M(1-\mu) + 1 + \alpha + \tau$, sea*

$$j_0 = \text{mín}\left\{j \in \{1, \ldots, L\} \,/\, v_{k+j} + x_{k+j+1} < -M(1-\mu) + 1 + \alpha + \tau\right\},$$

entonces $v_{k+i} - v_{k+i+1} \geq 1 - \alpha$ para todo $1 \leq i < j_0$, y $v_{k+i} < 0$ para todo $j_0 \leq i \leq L$.
Si $v_{k+i} + x_{k+i+1} \geq -M(1-\mu) + 1 + \alpha + \tau$ para todo $i \in \{1, \ldots, L\}$, entonces $v_{k+i} < 0$ para todo $1 \leq i \leq L$.
Lo mismo vale para el caso en que $u_k \geq -\tau$, y $u_{k+1}, u_{k+2}, \ldots, u_{k+L} < -\tau$, si se invierten todos los signos.

Este lema se prueba usando la definición de las variables internas y de la sucesión $\{q_n\}_{n\in\mathbb{Z}}$, además se usa la acotación de M. Su demostración se encuentra en [DDV].

Veamos cómo estos lemas prueban la prosición 2.4.7 que establece la acotación uniforme de la sucesión $\{u_n\}_{n\in\mathbb{Z}}$.

Demostración. La primer parte de la proposición fue probada en el lema 2.4.4. Demostraremos la segunda parte para $n > 0$, el caso $n < 0$ es análogo. Si $u_n \leq \tau$ para todo $n \in \mathbb{N}$, entonces vale la acotación. Si no sucede esto, como $u_0 \leq \frac{\tau}{2}$, existen k y L tales que $u_k \leq \tau$ y $u_{k+1}, \ldots, u_{k+L} > \tau$.
Si existe un $j \in \{1, \ldots, L\}$ tal que $v_{k+j} + x_{k+j+1} < -M(1-\mu) + 1 + \alpha + \tau$, por el lema 2.4.5, si $m \in \{1, \ldots, L\}$, tenemos que

$$u_{k+m} = u_k + \sum_{i=1}^{m} v_{k+i} \leq \tau + \sum_{i=1}^{j_0-1} v_{k+i}.$$

Además dados $j, i \in \{1, \ldots, j_0 - 1\}$,

$$v_{k+j} - v_{k+j+1} \geq 1 - \alpha,$$

por lo que

$$v_{k+1} - v_{k+i} = \sum_{j=1}^{i-1} v_{k+j} - v_{k+j+1} \geq (1-\alpha)(i-1),$$

resultando que

$$v_{k+1} - (1-\alpha)(i-1) \geq v_{k+i}.$$

A partir de esto obtenemos que

$$\begin{aligned} u_{k+m} &\leq \tau + \sum_{i=1}^{j_0-1}(v_{k+1} - (1-\alpha)(i-1)) \\ &\leq \tau + \max_{n\geq 1}\sum_{i=1}^{n}(v_{k+1} - (1-\alpha)(i-1)) \\ &= \tau + \max_{n\geq 1}\Big(v_{k+1}\frac{n(n+1)}{2} - (1-\alpha)\frac{(n-1)n}{2}\Big) \\ &\leq \tau + \max_{n\geq 1}\Big(\frac{M(1+\mu)+\tau+\alpha}{2}n^2 + \frac{M(1+\mu)+2+\tau-\alpha}{2}n\Big) \\ &\leq \tau + \frac{[M(1+\mu)+\tau+3/2-\alpha/2]^2}{2(1-\alpha)}. \end{aligned}$$

Si $v_{k+i} + x_{k+i+1} \geq -M(1-\mu) + 1 + \alpha + \tau$ para todo $i \in \{1,\ldots,L\}$, por el lema 2.4.5, sucede que si $m \in \{1,\ldots,L\}$, entonces

$$u_{k+m} = u_k + \sum_{i=1}^{m} v_{k+i} \leq \tau.$$

Probamos así que $u_n \leq \tau + \frac{[M(1+\mu)+\tau+3/2-\alpha/2]^2}{2(1-\alpha)}$ para todo $n \in \mathbb{N}$.
La acotación $u_n \geq -\tau - \frac{[M(1+\mu)+\tau+3/2-\alpha/2]^2}{2(1-\alpha)}$ para todo $n \in \mathbb{N}$ se prueba de la misma forma.

□

Recordemos que los esquemas (2.29) y (2.30) fueron construidos de forma que

$$\Delta_n^2 u = x_n - q_n \quad \forall\ n \in \mathbb{Z}.$$

Si en dichos esquemas utilizamos $x_n^\lambda = f\big(\frac{n}{\lambda}\big)$, resulta que

$$\Delta_n^2 u^\lambda = f\Big(\frac{n}{\lambda}\Big) - q_n^\lambda \quad \forall\ n \in \mathbb{Z}.$$

Es por esto que si $\|f\|_\infty \leq \alpha < 1$ y $|u_0|, |v_0| \leq \frac{\tau}{2}$, entonces

$$|u_n^\lambda| \leq \tau + \frac{[M(1+\mu)+\tau+3/2-\alpha/2]^2}{2(1-\alpha)} \quad \text{para todo } n \in \mathbb{Z}, \lambda > 1.$$

Debido a que la sucesión de variables internas resulta uniformemente acotada independientemente de λ, a partir del teorema 2.4.3 se puede afirmar que el error de aproximación global es del tipo $O(\lambda^{-2})$.

En [DDV] se presentan esquemas $\Sigma\Delta$ de orden k general y se demuestra la proposición que enuncia la estabilidad de los mismos. La prueba de la proposición es una generalización de lo realizado para el caso de orden dos. Dado que en esos esquemas se cumple que $\Delta_n^k(u) = x_n - q_n$ para todo $n \in \mathbb{Z}$, donde $\{u_n\}_{n\in\mathbb{Z}}$ es acotada, a partir del teorema 2.4.3 se obtiene que el error de aproximación es del tipo $O(\lambda^{-k})$.

Capítulo 3

Cuantificación Sigma-Delta ($\Sigma\Delta$) de expansiones en marcos finitos

En este capítulo estudiaremos la digitalización, a partir de los esquemas $\Sigma\Delta$, de vectores de $\mathbb{R}^d$. El proceso constará de dos pasos: la descomposición del vector en un marco finito, ajustado y uniforme, y la cuantificación de los coeficientes del vector en el marco.

Para comenzar con el desarrollo del primer paso, recordaremos la definición y algunas propiedades de los marcos finitos.

3.1. Teoría de marcos finitos

En las siguientes definiciones J será un conjunto a lo sumo numerable.

Definición 3.1.1. *Una sucesión de elementos $\{e_j\}_{j\in J}$ en un espacio de Hilbert separable H se dice un marco de H si existen*

$$0 < A \leq B < \infty$$

tales que

$$\forall\ x \in H, \quad A\|x\|^2 \leq \sum_{j\in J} |\langle x, e_j\rangle|^2 \leq B\|x\|^2.$$

Las constantes A y B se denominan cotas del marco.
El marco se dice ajustado si $A = B$, finito si J es finito y $H = \mathbb{R}^d$ ó $H = \mathbb{C}^d$, y uniforme si $\|e_j\| = 1$ para todo $j \in J$.

Observación 3.1.1. *$\{e_j\}_{j\in J}$ es un marco ajustado de H con constante A si y sólo si*

$$\forall\ x \in H, \quad A\|x\|^2 = \sum_{j\in J} |\langle x, e_j\rangle|^2.$$

Proposición 3.1.1. *Si $\{e_j\}_{j\in J}$ es un marco ajustado uniforme con constante A en un espacio de Hilbert separable H,*

$$A = 1 \Longleftrightarrow \{e_j\}_{j\in J} \text{ es una base ortonormal.}$$

Demostración. $\Longrightarrow$) Si $A = 1$, por la observación 3.1.1 vale que

$$\|x\|^2 = \sum_{j\in J} |\langle x, e_j\rangle|^2.$$

Si $\langle x, e_j\rangle = 0$ para todo $j \in J$, entonces $x = 0$, por lo que la sucesión $\{e_j\}_{j\in J}$ genera H.
Falta ver que son ortogonales,

$$\|e_n\|^2 = \sum_{j\in J} |\langle e_n, e_j\rangle|^2 = \|e_n\|^4 + \sum_{j\neq n} |\langle e_n, e_j\rangle|^2.$$

Como el marco es uniforme, entonces

$$1 = 1 + \sum_{j\neq n} |\langle e_n, e_j\rangle|^2 \Longrightarrow \sum_{j\neq n} |\langle e_n, e_j\rangle|^2 = 0 \Longrightarrow \langle e_n, e_j\rangle = 0 \quad \forall j \neq n.$$

$\Longleftarrow$) Dado $x \in H$, como $\{e_j\}_{j\in J}$ es una base ortonormal, vale que

$$x = \sum_{j\in J} \langle x, e_j\rangle e_j,$$

y

$$\|x\|^2 = \sum_{j\in J} |\langle x, e_j\rangle|^2.$$

Entonces resulta por la observación 3.1.1 que $\{e_j\}_{j\in J}$ es un marco ajustado de constante uno.

□

Definición 3.1.2. *Sea $\{e_j\}_{j\in J}$ un conjunto de vectores de H.*
Se define el operador de análisis

$$L : H \longrightarrow l^2(J)$$

como:

$$(Lx)_k = \langle x, e_k\rangle.$$

El operador de síntesis

$$L^* : l^2(J) \longrightarrow H$$

es el adjunto de L, definido como

$$L^*(\{c_j\}_{j\in J}) = \sum_{j\in J} c_j e_j.$$

El operador de marco

$$S : H \longrightarrow H,$$

es

$$S = L^*L.$$

Proposición 3.1.2. *$\{e_j\}_{j\in J}$ es un marco de H con constantes A y B si y sólo si el operador de marco S satisface*

$$AI \leq S \leq BI,$$

donde I es el operador identidad en H.

Demostración. Por definición de S, dado $x \in H$, vale que

$$Sx = L^*Lx = L^*(\{\langle x, e_j\rangle\}_{j\in J}) = \sum_{j\in J}\langle x, e_j\rangle e_j.$$

Entonces

$$\langle Sx, x\rangle = \sum_{j\in J} |\langle x, e_j\rangle|^2,$$

por lo que resulta que

$$A\|x\|^2 \leq \sum_{j\in J} |\langle x, e_j\rangle|^2 \leq B\|x\|^2 \Longleftrightarrow A\|x\|^2 \leq \langle Sx, x\rangle \leq B\|x\|^2.$$

□

Corolario 3.1.1. *$\{e_j\}_{j\in J}$ es un marco ajustado de H con constante A si y sólo si $S = AI$.*

Teorema 3.1.1. *Sea $\{e_j\}_{j\in J}$ un marco de H con constantes A y B, y sea S el operador de marco. Entonces $\{S^{-1}e_j\}_{j\in J}$ es un marco de H con constantes A^{-1} y B^{-1}, denominado marco dual canónico de $\{e_j\}_{j\in J}$. Más aún, para todo $x \in H$ vale*

$$x = \sum_{j\in J} \langle x, e_j\rangle (S^{-1}e_j) = \sum_{j\in J} \langle x, S^{-1}e_j\rangle e_j$$

con convergencia incondicional en ambas sumas.

La prueba de este teorema puede encontrarse en [Da].

Corolario 3.1.2. *Si $\{e_j\}_{j\in J}$ es un marco ajustado de H con constante A, entonces*

$$\forall\ x \in H, \quad x = A^{-1} \sum_{j\in J} \langle x, e_j\rangle e_j.$$

Dado que en la descomposición de la señal utilizaremos marcos finitos, ajustados y uniformes de $\mathbb{R}^d$, enunciaremos dos proposiciones referidas a ese tipo de marcos. Notaremos $\|\cdot\|$ como la norma dos euclídea de $\mathbb{R}^d$ ó $\mathbb{C}^d$.

Proposición 3.1.3. *Un conjunto de vectores $\{e_n\}_{n=1}^N$ en $H = \mathbb{R}^d$ ó $\mathbb{C}^d$ es un marco ajustado con constante A si y sólo si la matriz asociada $L \in \mathbb{C}^{N\times d}$, cuya n-ésima fila es $\overline{e_n}$, satisface $S = L^*L = AI_d$, donde L^* es la matriz transpuesta conjugada de L e I_d es la matriz identidad.*

Demostración. Como las filas de L son los vectores $\overline{e_n}$, entonces L es el operador de análisis, ya que

$$(Lx)_k = \langle x, e_k\rangle.$$

La demostración es trivial debido al corolario 3.1.1.

□

Proposición 3.1.4. *Si $\{e_n\}_{n=1}^N$ en $H = \mathbb{R}^d$ ó $\mathbb{C}^d$ es un marco finito ajustado uniforme con constante A, entonces $A = \frac{N}{d}$.*

Demostración. Como $\{e_n\}_{n=1}^N$ es un marco ajustado, por la proposición 3.1.3, vale que $L^*L = AI_d$, entonces $\text{tr}(L^*L) = \text{tr}(AI_d) = dA$.
Usando la definición de L, tenemos que

$$\text{tr}(L^*L) = \text{tr}(LL^*) = \sum_{n=1}^{N} \overline{e_n} e_n^t = \sum_{n=1}^{N} \|e_n\|^2 = \sum_{n=1}^{N} 1 = N.$$

Resulta entonces que $N = \text{tr}(L^*L) = dA$, por lo que $A = \frac{N}{d}$.

□

A continuación veremos que construir marcos finitos en $\mathbb{R}^d$ es muy sencillo, ya que cualquier conjunto de generadores del espacio cumple con la condición de marco.

Observación 3.1.2. *$\{e_n\}_{n=1}^N$ es un marco finito de $\mathbb{R}^d$ si y sólo si es un conjunto de generadores de ese espacio.*

Demostración. $\Longrightarrow$) Sea $x \in \mathbb{R}^d$ tal que $\langle x, e_n\rangle = 0$ para todo $1 \leq n \leq N$, queremos ver que $x = 0$.
Debido a que el conjunto $\{e_n\}_{n=1}^N$ cumple la condición de marco, existe $A > 0$ tal que:

$$A\|x\|^2 \leq \sum_{j=1}^{N} |\langle x, e_j\rangle|^2 = 0,$$

por lo que $x = 0$.

$\Longleftarrow$) Si definimos $B := \sum_{n=1}^N \|e_n\|^2$, usando Cauchy-Schwarz, vale que

$$\sum_{n=1}^{N} |\langle x, e_n\rangle|^2 \leq \|x\|^2 \sum_{n=1}^{N} \|e_n\|^2 = \|x\|^2 B.$$

Para lograr la cota inferior de la condición de marco, tomamos

$$A := \Big(\min_{\|x\|=1} \|L(x)\| \Big)^2,$$

donde L es el operador de análisis, es decir

$$L : \mathbb{R}^d \longrightarrow \mathbb{R}^N, \qquad L(x) = \Big(\langle x, e_1\rangle, \ldots, \langle x, e_N\rangle \Big).$$

Observemos que $A > 0$, ya que si A=0, entonces existe $x \in \mathbb{R}^d$ tal que $\|x\| = 1$ y $L(x) = 0$. Por la definición del operador L, esto implicaría que para ese x vale que $\langle x, e_j \rangle = 0$ para todo $1 \leq j \leq N$, pero dado que $\{e_n\}_{n=1}^N$ genera $\mathbb{R}^d$, debería suceder que $x = 0$, y esto resulta absurdo ya que $\|x\| = 1$. La constante A verifica la desigualdad de marco, ya que

$$\|L(x)\| \geq \Big(\min_{\|x\|=1} \|L(x)\| \Big) \|x\| \quad \forall\ x \in \mathbb{R}^d,$$

implica que

$$\|L(x)\|^2 = \sum_{n=1}^{N} |\langle x, e_n \rangle|^2 \geq A\|x\|^2 \quad \forall\ x \in \mathbb{R}^d.$$

□

Para construir un marco finito, ajustado y uniforme en $\mathbb{R}^d$ no basta con que el conjunto de vectores genere el espacio, ya que se deben chequear las proposiciones 3.1.3 y 3.1.4.
Veamos dos ejemplos típicos de este tipo de marcos que tienen una cantidad arbitraria de elementos.

Ejemplo 3.1.1. ***Raíces de la unidad***
Consideremos las raíces N-ésimas de la unidad $R_N = \{e_n^N\}_{n=1}^N$, *con* $e_n^N = \big(cos\big(\frac{2\pi n}{N}\big), sen\big(\frac{2\pi n}{N}\big)\big)$, *donde* $N \geq 2$.
La matriz asociada al operador de análisis es

$$L = \begin{pmatrix} cos\big(\frac{2\pi}{N}\big) & sen\big(\frac{2\pi}{N}\big) \\ \vdots & \vdots \\ cos(2\pi) & sen(2\pi) \end{pmatrix}$$

La matriz del operador de marco es

$$L^*L = \begin{pmatrix} \sum_{n=1}^N cos^2\big(\frac{2\pi n}{N}\big) & \sum_{n=1}^N cos\big(\frac{2\pi n}{N}\big) sen\big(\frac{2\pi n}{N}\big) \\ \sum_{n=1}^N cos\big(\frac{2\pi n}{N}\big) sen\big(\frac{2\pi n}{N}\big) & \sum_{n=1}^N sen^2\big(\frac{2\pi n}{N}\big) \end{pmatrix}$$

Observemos que

$$\sum_{n=1}^{N} cos\Big(\frac{2\pi n}{N}\Big) sen\Big(\frac{2\pi n}{N}\Big) = \frac{1}{2} \sum_{n=1}^{N} sen\Big(\frac{4\pi n}{N}\Big) = \frac{1}{2} Im\Big(\sum_{n=1}^{N} e^{\frac{4\pi n i}{N}} \Big) = 0,$$

$$\sum_{n=1}^{N}\Big(cos^2\Big(\frac{2\pi n}{N}\Big)-sen^2\Big(\frac{2\pi n}{N}\Big)\Big)=\sum_{n=1}^{N}cos\Big(\frac{4\pi n}{N}\Big)=Re\Big(\sum_{n=1}^{N}e^{\frac{4\pi ni}{N}}\Big)=0 \tag{3.1}$$

y

$$\sum_{n=1}^{N}\Big(cos^2\Big(\frac{2\pi n}{N}\Big)+sen^2\Big(\frac{2\pi n}{N}\Big)\Big)=N. \tag{3.2}$$

Sumando las ecuaciones (3.1) y (3.2) queda que

$$2\sum_{n=1}^{N}cos^2\Big(\frac{2\pi n}{N}\Big)=N\Longrightarrow\sum_{n=1}^{N}cos^2\Big(\frac{2\pi n}{N}\Big)=\frac{N}{2},$$

restándolas

$$2\sum_{n=1}^{N}sen^2\Big(\frac{2\pi n}{N}\Big)=N\Longrightarrow\sum_{n=1}^{N}sen^2\Big(\frac{2\pi n}{N}\Big)=\frac{N}{2}.$$

Entonces

$$L^*L=\begin{pmatrix}\frac{N}{2} & 0\\ 0 & \frac{N}{2}\end{pmatrix},$$

por lo que, usando la proposición 3.1.3, probamos que para cada $N\geq 2$, R_N es un marco de $\mathbb{R}^2$ finito y ajustado de longitud N. Además resulta uniforme, pues los vectores e_n son raíces de la unidad.

Ejemplo 3.1.2. ***Marcos armónicos***
El ejemplo anterior sólo sirve para $\mathbb{R}^2$, veremos un ejemplo que funciona para $\mathbb{R}^d$ con $d\geq 2$.
Dado $N\geq d$, consideramos la familia $H_N^d=\{e_j^N\}_{j=0}^{N-1}$ definida de la siguiente forma:
Si d es par

$$e_j^N=\sqrt{\frac{2}{d}}\Big[\cos\Big(\frac{2\pi j}{N}\Big),sen\Big(\frac{2\pi j}{N}\Big),cos\Big(\frac{2\pi 2j}{N}\Big),sen\Big(\frac{2\pi 2j}{N}\Big),\dots$$
$$\dots,cos\Big(\frac{2\pi\frac{d}{2}j}{N}\Big),sen\Big(\frac{2\pi\frac{d}{2}j}{N}\Big)\Big].$$

Si d es impar

$$e_j^N = \sqrt{\frac{2}{d}}\Big[\frac{1}{\sqrt{2}}, cos\Big(\frac{2\pi j}{N}\Big), sen\Big(\frac{2\pi j}{N}\Big), cos\Big(\frac{2\pi 2j}{N}\Big), sen\Big(\frac{2\pi 2j}{N}\Big), \dots \\ \dots, cos\Big(\frac{2\pi \frac{d-1}{2} j}{N}\Big), sen\Big(\frac{2\pi \frac{d-1}{2} j}{N}\Big)\Big].$$

Observemos que para $d = 2$ los marcos armónicos son las raíces de la unidad reordenadas.

Tal como en el ejemplo anterior se prueba, usando la proposición 3.1.3, que H_N^d es un marco ajustado en $\mathbb{R}^d$.
Además observemos que el marco fue construido de forma que los vectores tengan norma uno, con lo cual resultan marcos finitos, ajustados y uniformes para todo $d \geq 2$ y $N \geq d$.

Usando el corolario 3.1.2 y la proposición 3.1.4, podemos decir que si $\{e_n\}_{n=1}^N$ es un marco finito, ajustado y uniforme de $\mathbb{R}^d$, entonces

$$\forall\ x \in \mathbb{R}^d, \quad x = \frac{d}{N}\sum_{n=1}^{N} x_n e_n, \quad \text{con} \quad x_n = \langle x, e_n\rangle. \tag{3.3}$$

Logramos así concretar el primer paso en la digitalización, que es conseguir una descomposición de la señal $x \in \mathbb{R}^d$ en base a los coeficientes $x_n = \langle x, e_n\rangle$ en el marco $\{e_n\}_{n=1}^N$.

Observemos que si $\{e_n\}_{n=1}^N$ es un marco finito, ajustado y uniforme de $\mathbb{R}^d$, por la observación 3.1.2, resultará ser un conjunto de generadores de ese espacio y por consiguiente sucederá que $N \geq d$.
Dado que la constante de marco es $A = \frac{N}{d} \geq 1$, si $A > 1$, entonces $\{e_n\}_{n=1}^N$ es un conjunto linealmente dependiente, por lo que la descomposición (3.3) no será única. Podemos afirmar entonces que cuanto mayor sea A, mayor será N y por lo tanto más redundante será el marco.

Si utilizamos valores grandes de N, como la descomposición (3.3) incrementa su redundancia, la pérdida o modificación de algunos coeficientes x_n no acarrea graves consecuencias en la señal digitalizada. Es por esto que la representación en marcos de la señal permite compensar errores introducidos por ciertos hardwares durante la digitalización, además asegura la estabilidad númerica del proceso y minimiza los efectos del ruido. Cuanto mayor es

la cantidad de elementos del marco, mayor es la robustez en los procesos de cuantificación.

Antes de continuar con la cuantificación para marcos finitos, recordemos que para funciones de banda limitada se obtuvo una descomposición similar a (3.3), ya que en la observación 2.1.2 del capítulo anterior vimos que

$$\forall\ f \in B_\pi, \quad f(t) = \frac{1}{\lambda}\sum_{n\in\mathbb{Z}} f\Big(\frac{n}{\lambda}\Big)g\Big(t-\frac{n}{\lambda}\Big),\ \text{con}\ f\Big(\frac{n}{\lambda}\Big) = \Big\langle f, g\Big(\,\cdot -\frac{n}{\lambda}\Big)\Big\rangle.$$

En este caso la constante $A := \lambda$ también mide la redundancia de la descomposición, ya que cuanto mayor es la misma, mayor es la frecuencia de muestreo.
Una diferencia que se puede marcar entre la expresión obtenida para funciones de banda limitada y la de marcos finitos, es que en la primera, los coeficientes de la descomposición son muestras de la función.

3.2. Cuantificación

El segundo paso en la digitalización de señales representadas por vectores es la cuantificación, es decir, la asignación a cada coeficiente x_n del marco de un elemento q_n que se encuentra dentro de un alfabeto fijo.

A diferencia de lo hecho para funciones de banda limitada, para el algoritmo $\Sigma\Delta$ de orden uno, vamos a usar alfabetos que pueden tener una cantidad arbitraria de elementos. Es decir que trabajaremos con la cuantificación multibit.

Dados $\delta > 0$ y $K \in \mathbb{N}$, el alfabeto formará una progresión aritmética de paso δ con $2K$ elementos definida por

$$\mathcal{A}_K^\delta := \Big\{\Big(-K+\frac{1}{2}\Big)\delta, \Big(-K+\frac{3}{2}\Big)\delta, \ldots, -\frac{1}{2}\delta, \frac{1}{2}\delta, \ldots, \Big(K-\frac{3}{2}\Big)\delta, \Big(K-\frac{1}{2}\Big)\delta\Big\}. \tag{3.4}$$

Al cambiar el alfabeto también se modificará el cuantificador escalar, ya no será la función signo, sino que estará definido por

$$Q : \mathbb{R} \longrightarrow \mathcal{A}_K^\delta, \quad Q(u) = \arg\Big\{\min_{q\in\mathcal{A}_K^\delta} |u-q|\Big\}. \tag{3.5}$$

Es decir, que el cuantificador le asignará a cada u el elemento más cercano en el alfabeto. En el caso en que dos elementos del alfabeto estén a igual distancia, se toma $Q(u)$ como el mayor de ellos.
A partir de la definición (3.5), se puede probar que

$$Q(u) = \begin{cases} \left\lfloor \frac{u}{\delta} \right\rfloor \delta + \frac{\delta}{2} & \text{si } |u| < K\delta \\ \text{sign}(u)(K - 1/2)\delta & \text{si } |u| \geq K\delta. \end{cases}$$

Vimos en la sección anterior que la señal x queda representada en un marco ajustado uniforme $\{e_n\}_{n=1}^N$ de la siguiente forma:

$$x = \frac{d}{N} \sum_{n=1}^{N} x_n e_n, \quad \text{con} \quad x_n = \langle x, e_n \rangle.$$

Una vez cuantificados los coeficientes x_n, la señal aproximante será

$$\tilde{x} = \frac{d}{N} \sum_{n=1}^{N} q_n e_n. \tag{3.6}$$

Definiremos el *error de aproximación* como $\|x - \tilde{x}\|$, donde $\|\cdot\|$ es la norma dos euclídea de $\mathbb{R}^d$.

Tal como lo hicimos para las funciones de banda limitada, vamos a estudiar primero el algoritmo PCM, luego vamos a presentar los esquemas de cuantificación $\Sigma\Delta$ de orden uno y haremos una breve comparación de ambos algoritmos. Por último analizaremos los esquemas $\Sigma\Delta$ de orden dos.

3.3. Algoritmo de Modulación por Impulsos Codificados

En el caso de funciones de banda limitada, el PCM asignaba a cada x_n su desarrollo binario hasta las cifra N. Vimos que en realidad podíamos pensarlo como que el alfabeto era

$$\mathcal{A} = \{-1, -1 + 2^{-N}, \ldots, -2^{-N}, 0, 2^{-N}, \ldots, 1 - 2^{-N+1}, 1 - 2^{-N}\}$$

y se le asignaba a cada coeficiente el elemento menor más cercano en el alfabeto.
Dicho de otra forma,

$$q_n = Q_B(x_n),$$

con

$$Q_B(u) := \text{máx}\big\{q \in \mathcal{A} \,/\, q \leq u\big\}.$$

En el caso en que la señal es un vector de $\mathbb{R}^d$, para compatibilizar con la definición que daremos del algoritmo $\Sigma\Delta$ de orden uno, vamos a cambiar el cuantificador Q y el alfabeto. Definiremos al PCM como el proceso que le asigna a cada x_n el elemento más cercano del alfabeto $\mathcal{A}_K^\delta$ con $K \in \mathbb{N}$ y $\delta > 0$ cualesquiera.
Formalmente,

$$q_n = Q(x_n),$$

con Q y $\mathcal{A}_K^\delta$ definidos como en (3.5) y (3.4) respectivamente.

Observemos que por la forma en que se define q_n y $\mathcal{A}_K^\delta$, como los elementos del alfabeto distan δ uno de otro, resulta que:

$$|x_n| \leq K\delta \Longrightarrow |x_n - q_n| \leq \frac{\delta}{2}. \tag{3.7}$$

Veamos cual es el error de aproximación al utilizar el PCM.

Proposición 3.3.1. *Dado $\{e_n\}_{n=1}^N$ un marco ajustado uniforme de $\mathbb{R}^d$, $\delta > 0$ y $K \in \mathbb{N}$. Si $\|x\| \leq K\delta$ y $\tilde{x}$ es definido como en (3.6) a partir de los coeficientes $\{q_n\}_{n=1}^N$ producidos por el algoritmo PCM, entonces*

$$\|x - \tilde{x}\| \leq \frac{d}{2}\delta.$$

Demostración. Como $x_n = \langle x, e_n\rangle$ y $\|e_n\| = 1 \quad \forall\ 1 \leq n \leq N$, vale que

$$|x_n| = |\langle x, e_n\rangle| \leq \|x\|\|e_n\| = \|x\| \leq K\delta,$$

entonces podemos afirmar por (3.7) que $|x_n - q_n| \leq \frac{\delta}{2}$. Usando esta acotación, el error queda

$$\begin{aligned}\|x - \tilde{x}\| = \frac{d}{N}\Big\|\sum_{n=1}^N (x_n - q_n)e_n\Big\| &\leq \frac{d}{N}\sum_{n=1}^N |x_n - q_n|\|e_n\| \\ &\leq \frac{d}{N}\frac{\delta}{2}\sum_{n=1}^N \|e_n\| = \frac{d}{N}\frac{\delta}{2}N = \frac{d}{2}\delta.\end{aligned}$$

□

La cota obtenida en la proposición 3.3.1 sólo depende de δ, por lo que se podría conjeturar que no siempre el incremento de la redundancia del marco implica la disminución del error producido por el PCM. De hecho veremos un ejemplo que muestra que aunque aumentemos la cantidad de elementos del marco, si se mantiene fijo el paso δ del alfabeto, el error en el procedimiento PCM no se puede reducir.

A continuación se enuncia una propiedad necesaria para demostrar una desigualdad en el ejemplo que presentaremos.

Proposición 3.3.2.

$$\|e^{i\alpha} - e^{i\beta}\| \leq |\alpha - \beta| \quad \forall\ \alpha, \beta \in \mathbb{R}.$$

Demostración. Como

$$\|e^{i\alpha} - e^{i\beta}\| = \|e^{i\beta}\|\|e^{i(\alpha-\beta)} - 1\| = \|e^{i(\alpha-\beta)} - 1\|,$$

basta probar que

$$\|e^{i\alpha} - 1\| \leq |\alpha| \quad \forall\ \alpha \in \mathbb{R}.$$

Resulta entonces que

$$\begin{aligned}
\|\mathrm{e}^{i\alpha} - 1\| &= \|\cos\alpha + i\mathrm{sen}\alpha - 1\| \\
&= \Big\| \cos^2\Big(\frac{\alpha}{2}\Big) - \mathrm{sen}^2\Big(\frac{\alpha}{2}\Big) + i2\cos\Big(\frac{\alpha}{2}\Big)\mathrm{sen}\Big(\frac{\alpha}{2}\Big) - 1\Big\| \\
&= \Big\| -2\mathrm{sen}^2\Big(\frac{\alpha}{2}\Big) + 2i\cos\Big(\frac{\alpha}{2}\Big)\mathrm{sen}\Big(\frac{\alpha}{2}\Big)\Big\| \\
&= 2\Big|\mathrm{sen}\Big(\frac{\alpha}{2}\Big)\Big|\Big\|i\cos\Big(\frac{\alpha}{2}\Big) - \mathrm{sen}\Big(\frac{\alpha}{2}\Big)\Big\| \\
&\leq 2\Big|\frac{\alpha}{2}\Big| = |\alpha|.
\end{aligned}$$

□

Ejemplo 3.3.1. *En la sección de teoría de marcos vimos que* $R_N = \{e_n^N\}_{n=1}^N$, *las raíces N-ésimas de la unidad, forman un marco ajustado uniforme de* $\mathbb{R}^2$. *Si consideramos* $x = (0, b)$ *con* $0 < b < 1$, *los coeficientes de* x *en el marco resultan*

$$x_n^N = \Big\langle (0,b), \Big(cos\Big(\frac{2\pi n}{N}\Big), sen\Big(\frac{2\pi n}{N}\Big)\Big)\Big\rangle = bsen\Big(\frac{2\pi n}{N}\Big).$$

Supongamos que N es par, entonces para $1 \leq n \leq \frac{N}{2}$ y $n = N$, $x_n^N \geq 0$, y para $\frac{N}{2} < n < N$, $x_n^N < 0$.
Fijemos el alfabeto $\mathcal{A}_K^\delta$ con $\delta = 1$ y $K = 2$, es decir, $\mathcal{A}_2^1 = \{-1/2, 1/2\}$. Entonces si aplicamos el método PCM para cuantificar los coeficientes en el marco, obtenemos:

$$q_n^N = \begin{cases} 1/2 & si \;\; 1 \leq n \leq N/2,\, n = N \\ -1/2 & si \;\; N/2 < n < N. \end{cases}$$

Usando esto y la proposición 3.3.2 para $\alpha = \frac{2\pi}{N}$ y $\beta = 0$, si identificamos $e_n^N = e^{\frac{2\pi i n}{N}}$ y suponemos $N \geq 4$, resulta que

$$\begin{aligned}
\|\tilde{x}_N\| &= \Big\| \frac{2}{N} \sum_{n=1}^{N} q_n^N e_n^N \Big\| = \frac{2}{N} \Big\| \sum_{n=1}^{N/2} \frac{1}{2} e_n^N + \frac{1}{2} e_N^N - \sum_{n=N/2+1}^{N-1} \frac{1}{2} e_n^N \Big\| \\
&= \frac{2}{N} \Big\| \sum_{n=1}^{N/2} \frac{1}{2} e^{\frac{2\pi i n}{N}} + \frac{1}{2} - \sum_{n=1}^{N/2-1} \frac{1}{2} e^{\frac{2\pi i (n+N/2)}{N}} \Big\| \\
&= \frac{2}{N} \Big\| \sum_{n=1}^{N/2} \frac{1}{2} e^{\frac{2\pi i n}{N}} + \frac{1}{2} + \sum_{n=1}^{N/2-1} \frac{1}{2} e^{\frac{2\pi i n}{N}} \Big\| \\
&= \frac{2}{N} \Big\| \sum_{n=1}^{N/2-1} e^{\frac{2\pi i n}{N}} + \frac{1}{2} (e^{\frac{2\pi i}{N}})^{N/2} + \frac{1}{2} \Big\| = \frac{2}{N} \Big\| \sum_{n=1}^{N/2-1} e^{\frac{2\pi i n}{N}} \Big\| \\
&= \frac{2}{N} \Big\| \frac{-2}{e^{\frac{2\pi i}{N}} - 1} - 1 \Big\| = \frac{2}{N} \Big\| \frac{e^{\frac{2\pi i}{N}} + 1}{e^{\frac{2\pi i}{N}} - 1} \Big\| \geq \frac{2}{N} \frac{|Re(e^{\frac{2\pi i}{N}} + 1)|}{\|e^{\frac{2\pi i}{N}} - 1\|} \\
&\geq \frac{2}{N} \frac{|\cos(\frac{2\pi}{N}) + 1|}{\frac{2\pi}{N}} \geq \frac{1}{\pi}
\end{aligned}$$

Se puede probar que en el caso en que N sea impar, se obtiene la misma cota para $\tilde{x}_N$. Resulta entonces que si b es pequeño, como $\|x\| = b$ y $\|\tilde{x}_N\| \geq \frac{1}{\pi}$, vale que

$$\|x - \tilde{x}_N\| \geq \frac{1}{\pi} - b > 0.$$

Esto prueba que aunque N tienda a infinito, el error siempre será mayor que una constante positiva.

El ejemplo anterior nos muestra un caso en el que el PCM no hace uso de la redundancia del marco para lograr disminuir el error de aproximación.

Teniendo en cuenta el resultado de la proposición 3.3.1, podemos decir que para reducir el error obtenido al utilizar el PCM, se podría pedir que δ tienda a cero. Pero si esto pasa, como la acotación vale para los x tales que $\|x\| \leq K\delta$, si dejáramos el valor K fijo, sólo podríamos cuantificar x cerca del origen. Entonces generalmente se trabaja con $K = \lceil \frac{1}{\delta} \rceil$, donde $\lceil x \rceil = \text{mín}\{n \in \mathbb{Z} \,/\, x \leq n\}$. Con esta elección de K, logramos que el conjunto de los valores que pueden ser cuantificados, contenga a la bola de radio 1 alrededor del origen.

Si δ tiende a cero, $K = \lceil \frac{1}{\delta} \rceil$ tiende a infinito, por lo que se achica el paso del alfabeto $\mathcal{A}_K^\delta$, pero se incrementa la cantidad de elementos. Esto hace que el cuantificador Q deba realizar muchos saltos entre valores que distan muy poco, entonces para que el algoritmo PCM resulte robusto, se debe extremar la precisión de Q, lo que produce que la implementación del mecanismo sea muy costosa.

En la práctica, para demostrar que el PCM utiliza la redundancia en promedio, se usaron suposiciones probabilísticas sobre $x_n - q_n$, como por ejemplo: la *suposición del ruido blanco de Bennett*. La misma consiste en asumir que los errores en los canales individuales $\eta_n := x_n - q_n$ se comportan como *ruido blanco*, es decir, que $\{\eta_n\}_{n=1}^N$ son variables aleatorias independientes e idénticamente distribuidas con media 0 y varianza $\frac{\delta^2}{12}$.

El *error cuadrático medio* se define como

$$ECM := E(\|x - \tilde{x}\|^2),$$

donde E es la función esperanza.
Se prueba en [GK] el siguiente teorema.

Teorema 3.3.1. *Bajo la suposición del ruido blanco de Bennett, el error cuadrático medio al utilizar el algortimo PCM resulta:*

$$ECM_{PCM} = \frac{d^2\delta^2}{12N}.$$

Demostración. Dado que $\{\eta_i\}_{i=1}^N$ son variables aleatorias independientes e idénticamente distribuidas con media 0 y varianza $\frac{\delta^2}{12}$, sucede que

$$E(\eta_i) = 0 \quad \text{y} \quad E(\eta_i\eta_k) = \delta_{ik}\frac{\delta^2}{12} \quad \text{para todo } i, k \in \{1, \ldots, N\}.$$

Tenemos entonces que

$$E(\|x-\tilde{x}\|^2) = E\left(\left\|\frac{d}{N}\sum_{i=1}^{N}\eta_i e_i\right\|^2\right) = \frac{d^2}{N^2}E\left(\sum_{i=1}^{N}\sum_{k=1}^{N}\eta_i\eta_k\langle e_i,e_k\rangle\right)$$
$$= \frac{d^2}{N^2}\sum_{i=1}^{N}\sum_{k=1}^{N}\delta_{ik}\frac{\delta^2}{12}\langle e_i,e_k\rangle = \frac{d^2}{N^2}\frac{\delta^2}{12}\sum_{i=1}^{N}\|e_i\|^2 = \frac{d^2\delta^2}{12N}.$$

En la última igualdad hemos usado que $\|e_i\| = 1$ para todo $1 \leq i \leq N$,

□

A partir de las hipótesis de Bennett, se obtuvo un error cuadrático que depende de N, por lo que se puede decir que el PCM usa en promedio la redundancia del marco. Sin embargo se presenta el problema de que lo asumido respecto a la independencia de las variable η_n no se cumple en general, ya que como los coeficientes x_n corresponden a una descomposición de x redundante, están relacionados entre sí.

3.4. Algoritmos de cuantificación Sigma-Delta (ΣΔ)

3.4.1. Cuantificación Sigma-Delta (ΣΔ) de orden uno

A diferencia del caso de funciones de banda limitada donde $\mathcal{A} = \{-1, 1\}$, utilizaremos en esta sección el alfabeto $\mathcal{A}_K^\delta$ definido en (3.4) que tiene $2K$ elementos.
En el capítulo anterior podríamos haber usado este tipo de alfabetos, pero para simplificar las cuentas decidimos trabajar con la cuantificación de 1-bit.

Dado que el algoritmo ΣΔ utiliza para el cálculo del coeficiente q_n a los x_m con $m \leq n$, el orden en que se cuantifica la sucesión $\{x_n\}_{n=1}^N$ adquiere relevancia. Es por esto que en la definición del esquema que daremos a continuación, va a aparecer una permutación p de los N elementos del marco, que marca dicho orden. Explicaremos despúes la razón por la cual no se tenía en cuenta esta cuestión para las funciones de banda limitada.

Definición 3.4.1. *Dados $K \in \mathbb{N}$ y $\delta > 0$, consideremos el alfabeto $\mathcal{A}_K^\delta$ y el cuantificador Q definidos en (3.4) y (3.5) respectivamente. Sea $\{x_n\}_{n=1}^N$ la sucesión de coeficientes del vector $x \in \mathbb{R}^d$ en el marco ajustado uniforme $\{e_n\}_{n=1}^N$ y p una permutación de $\{1, 2, \ldots, N\}$ que marca el orden de cuantificación de los coeficientes.*
El esquema de cuantificación $\Sigma\Delta$ de primer orden se define por la siguiente iteración:

$$\begin{cases} u_n = u_{n-1} + x_{p(n)} - q_n \\ q_n = Q(u_{n-1} + x_{p(n)}), \end{cases} \tag{3.8}$$

donde $n = 1, 2, \ldots, N$ y $u_0 = 0$.
El algoritmo produce una sucesión $\{u_n\}_{n=1}^N$ de variables internas y una sucesión de coeficientes cuantificados $\{q_n\}_{n=1}^N$.

Observemos que si tomamos $\delta = 2$ y $K = 1$, el alfabeto resulta ser $\mathcal{A}_1^2 = \{-1, 1\}$, y el cuantificador Q es la función signo. Por lo que podemos decir que estamos trabajando con un alfabeto y un cuantificador que son una generalización de los utilizados para funciones de banda limitada.

Tal como lo hicimos en el capítulo anterior, veremos primero que el algoritmo es estable, es decir, que las variables internas están uniformemente acotadas si la sucesión de coeficientes $\{x_n\}_{n=1}^N$ también lo está. Luego haremos estimaciones sobre la cota del error de aproximación.

Estabilidad

Proposición 3.4.1. *Sea $\{u_n\}_{n=1}^N$ la sucesión de variables internas producidas por el esquema (3.8), si*

$$|x_n| \leq (K - 1/2)\delta \qquad \forall\, n \in \{1, 2, \ldots, N\},$$

entonces

$$|u_n| \leq \delta/2 \qquad \forall\, n \in \{0, 1, \ldots, N\}.$$

Demostración. Se prueba por inducción.
Para $n = 0$ vale pues $u_0 = 0$.
Supongamos que $|u_{n-1}| \leq \delta/2$. Como $|x_{p(n)}| \leq (K - 1/2)\delta$, entonces $|u_{n-1} + x_{p(n)}| \leq K\delta$. Usando (3.7)

$$|u_{n-1} + x_{p(n)}| \leq K\delta \Longrightarrow |u_n| = |u_{n-1} + x_{p(n)} - Q(u_{n-1} + x_{p(n)})| \leq \delta/2.$$

□

Podemos notar que la prueba de la estabilidad es similar a la de funciones de banda limitada, de hecho lo podemos ver como una generalización de la misma.

Una vez probada la estabilidad, nos gustaría obtener una expresión de la cota del error de cuantificación.

Error de aproximación

Observemos en (3.8) que el coeficiente q_n cuantifica a $x_{p(n)}$. Si bien nuestra señal la habíamos representado por

$$x = \frac{d}{N} \sum_{n=1}^{N} x_n e_n,$$

podemos permutar el orden de la suma y resulta que

$$x = \frac{d}{N} \sum_{n=1}^{N} x_{p(n)} e_{p(n)}.$$

A partir de la definición de los q_n, la señal aproximante obtenida por reconstrucción lineal será:

$$\tilde{x} = \frac{d}{N} \sum_{n=1}^{N} q_n e_{p(n)}. \tag{3.9}$$

Para la acotación del error necesitaremos introducir la noción de *variación del marco*, que es un coeficiente que relaciona el orden en que se eligen los coeficientes para cuantificar, es decir, la permutación p, con los elementos del marco.

Definición 3.4.2. *Sea $F = \{e_n\}_{n=1}^{N}$ un marco finito, ajustado, uniforme en $\mathbb{R}^d$, y sea p una permutación de $\{1, \ldots, N\}$. Se define la variación del marco F con respecto a p como*

$$\sigma(F, p) := \sum_{n=1}^{N-1} \|e_{p(n)} - e_{p(n+1)}\|.$$

Calculemos ahora el error al aproximar x por $\tilde{x}$ usando el algoritmo $\Sigma\Delta$ de orden uno. La demostración del teorema es similar a la de funciones de banda limitada, se usa la fórmula de sumación por partes.

Lema 3.4.1. ***Fórmula de sumación por partes***
Dados $\{w_n\}_{n=0}^N$ *y* $\{v_n\}_{n=1}^N$, *entonces*

$$\sum_{n=1}^{N}(w_n - w_{n-1})v_n = w_N v_N - w_0 v_1 + \sum_{n=1}^{N-1} w_n(v_n - v_{n+1}).$$

Teorema 3.4.1. *Dado el esquema* $\Sigma\Delta$ *definido por (3.8), sea* $F = \{e_n\}_{n=1}^N$ *un marco finito, ajustado, uniforme en* $\mathbb{R}^d$, *y sea* p *una permutación de* $\{1,\dots,N\}$. *Si* $x \in \mathbb{R}^d$ *satisface* $\|x\| \leq (K-1/2)\delta$ *y* $\tilde{x}$ *es definido como en (3.9), entonces*

$$\|x - \tilde{x}\| \leq \frac{d}{N}\Big(\sigma(F,p)\frac{\delta}{2} + |u_N|\Big).$$

Demostración. Usando el lema 3.4.1, obtenemos que

$$\begin{aligned}
\|x - \tilde{x}\| &= \frac{d}{N}\Big\| \sum_{n=1}^{N}(x_{p(n)} - q_n)e_{p(n)}\Big\| = \frac{d}{N}\Big\| \sum_{n=1}^{N}(u_n - u_{n-1})e_{p(n)}\Big\| \\
&= \frac{d}{N}\Big\| \sum_{n=1}^{N-1} u_n(e_{p(n)} - e_{p(n+1)}) + u_N e_{p(N)} - u_0 e_{p(1)}\Big\| \\
&\leq \frac{d}{N}\Big(\sum_{n=1}^{N-1} |u_n|\|e_{p(n)} - e_{p(n+1)}\| + |u_N|\|e_{p(N)}\| + |u_0|\|e_{p(1)}\|\Big).
\end{aligned}$$

Para acotar la última expresión, como $\|x\| \leq (K-1/2)\delta$ implica que

$$|x_n| = |\langle x, e_n\rangle| \leq \|x\|\|e_n\| \leq (K-1/2)\delta \qquad \forall\ n \in \{1,\dots,N\},$$

aplicando la proposición 3.4.1, tenemos que

$$|u_n| \leq \delta/2 \qquad \forall\ n \in \{1,\dots,N\}.$$

Si además usamos que $u_0 = 0$ y $\|e_n\| = 1$ para todo n, resulta que

$$\begin{aligned}
\|x - \tilde{x}\| &\leq \frac{d}{N}\Big(\frac{\delta}{2}\sum_{n=1}^{N-1}\|e_{p(n)} - e_{p(n+1)}\| + |u_N|\Big) \\
&= \frac{d}{N}\Big(\frac{\delta}{2}\sigma(F,p) + |u_N|\Big).
\end{aligned}$$

□

La cota del error de aproximación al utilizar el algoritmo $\Sigma\Delta$ de orden uno para vectores de $\mathbb{R}^d$ es

$$\|x - \tilde{x}\| \leq \frac{d}{N}\Big(\sigma(F,p)\frac{\delta}{2} + |u_N|\Big)$$

y la obtenida en el caso de funciones de banda limitada era

$$\|f - \tilde{f}\|_\infty \leq \frac{1}{\lambda}\|g'\|_{L^1}.$$

La forma en que se demuestran los teoremas de acotación de ambos errores es la misma: se basa en la sumación por partes. Sin embargo, en la expresión de $\mathbb{R}^d$ aparece fuera de la sumatoria principal el término $|u_N|$, que es la suma de los errores cometidos hasta el paso N. Esto se debe a que la aplicación del algoritmo $\Sigma\Delta$ a sucesiones finitas, produce la aparición de errores no compensados.

Si aplicamos la cota obtenida en la proposición 3.4.1 al término u_N de la expresión del teorema 3.4.1, tenemos el siguiente corolario.

Corolario 3.4.1. *Dado el esquema $\Sigma\Delta$ definido por (3.8), sea $F = \{e_n\}_{n=1}^N$ un marco finito, ajustado, uniforme en $\mathbb{R}^d$, y sea p una permutación de $\{1,\ldots,N\}$. Si $x \in \mathbb{R}^d$ satisface $\|x\| \leq (K - 1/2)\delta$ y $\tilde{x}$ es definido como en (3.9), entonces*

$$\|x - \tilde{x}\| \leq \frac{d}{N}\frac{\delta}{2}(\sigma(F,p) + 1).$$

Demostración. La prueba es inmediata usando el teorema 3.4.1 y que $|u_N| \leq \frac{\delta}{2}$.

□

Observemos que, a diferencia de lo ocurrido en funciones de banda limitada, al utilizar alfabetos de paso general δ, en la acotación del error aparece el factor $\frac{\delta}{2}$, es decir, que si usamos alfabetos con paso pequeño podemos obtener menores cotas. Recordemos que generalmente se elige $K = \lceil \frac{1}{\delta} \rceil$, por lo que disminuir δ es equivalente a incrementar la cantidad de elementos del alfabeto.

A continuación enunciaremos un lema que nos va a permitir probar que, en el caso en que los elementos del marco sumen cero, si la extensión del marco es par, la constante de acotación del error es menor que en el caso en que sea impar.

Lema 3.4.2. *Dado el esquema $\Sigma\Delta$ definido por (3.8), sea $F = \{e_n\}_{n=1}^N$ un marco finito, ajustado, uniforme en $\mathbb{R}^d$, y sea p una permutación de $\{1, \ldots, N\}$. Si $x \in \mathbb{R}^d$ satisface $\|x\| \leq (K - 1/2)\delta$ y F cumple la condición de suma cero, es decir,*

$$\sum_{n=1}^{N} e_n = 0,$$

entonces

$$|u_N| = \begin{cases} 0 & \text{si } N \text{ es par} \\ \delta/2 & \text{si } N \text{ es impar.} \end{cases}$$

Demostración. Por definición del esquema (3.8), tal como lo probamos en la sección de funciones de banda limitada, las variables internas son la suma de los errores, es decir,

$$u_N = u_0 + \sum_{n=1}^{N} x_n - \sum_{n=1}^{N} q_n = \sum_{n=1}^{N} x_n - \sum_{n=1}^{N} q_n.$$

Como F satisface la condición de suma cero,

$$\sum_{n=1}^{N} x_n = \sum_{n=1}^{N} \langle x, e_n \rangle = \Big\langle x, \sum_{n=1}^{N} e_n \Big\rangle = 0,$$

entonces resulta que

$$u_N = -\sum_{n=1}^{N} q_n.$$

Dado que $q_n \in \mathcal{A}_K^\delta = \{(-K+1/2)\delta, (-K+3/2)\delta, \ldots, (K-3/2)\delta, (K-1/2)\delta\}$, podemos decir que $q_n = M_n \frac{\delta}{2}$ con M_n un entero impar, por lo que

$$u_N = -\Big(\sum_{n=1}^{N} M_n \Big) \frac{\delta}{2}.$$

Si N es par, como la suma par de enteros impares es par, entonces

$$u_N = \frac{-(\sum_{n=1}^{N} M_n)}{2} \delta = A_N \delta$$

con $A_N \in \mathbb{Z}$. Por la proposición 3.4.1, dado que $|u_N| \leq \frac{\delta}{2}$, resulta que $|A_N| \leq \frac{1}{2}$, pero al pertenecer éste a los enteros, sucede que $A_N = 0$, provocando que $u_N = 0$.
Si N es impar, como la suma impar de enteros impares es impar,

$$|u_N| = A_N \frac{\delta}{2}$$

con A_N entero impar. El hecho de que $|u_N| \leq \frac{\delta}{2}$, implica que $|A_N| \leq 1$. Como A_N es impar, necesariamente debe suceder que $|A_N| = 1$, por lo que $|u_N| = \frac{\delta}{2}$.

□

Corolario 3.4.2. *Dado el esquema $\Sigma\Delta$ definido por (3.8), sea $F = \{e_n\}_{n=1}^N$ un marco finito, ajustado, uniforme en $\mathbb{R}^d$, y sea p una permutación de $\{1, \ldots, N\}$. Si $x \in \mathbb{R}^d$ satisface $\|x\| \leq (K - 1/2)\delta$, $\tilde{x}$ es definido como en (3.9) y F cumple la condición de suma cero, entonces*

$$\|x - \tilde{x}\| \leq \begin{cases} \frac{d}{N}\frac{\delta}{2}\sigma(F,p) & \text{si } N \text{ es par} \\ \frac{d}{N}\frac{\delta}{2}(\sigma(F,p)+1) & \text{si } N \text{ es impar.} \end{cases}$$

Demostración. Se aplica el teorema 3.4.1 y el lema 3.4.2.

□

Si bien el corolario anterior no marca gran diferencia entre las cotas para marcos con extensión par y las de impar, veremos más adelante que para los algoritmos de orden dos, esta variación del término u_N en función de la paridad de N, ocasiona diferencias sustanciales en la acotación del error.

Por el corolario 3.4.1, sabemos que para marcos ajustados, finitos y uniformes de $\mathbb{R}^d$, sucede que

$$\|x - \tilde{x}\| \leq \frac{d}{N}\frac{\delta}{2}(\sigma(F,p)+1),$$

con $\sigma(F,p) := \sum_{n=1}^{N-1} \|e_{p(n)} - e_{p(n+1)}\|$.

Para disminuir el error, podríamos fijar N y el marco, y hacer tender δ a cero (achicar cada vez más la distancia entre los elementos del alfabeto). Pero tal como explicamos en la sección del PCM, esto es costoso para ser implementado. En cambio, así como lo hicimos en el caso de funciones de

banda limitada, buscamos fijar δ (fijar el alfabeto) y hacer tender N a infinito logrando que el marco sea más redundante.

En el caso en que la señal era $f \in B_\pi$, como los núcleos g dependían de λ, concluimos que para hacer chico el error, debíamos tomar $g_\lambda = g_{\lambda_0}$ para todo $\lambda \geq \lambda_0$. Con esto logramos una cota del tipo $O(\lambda^{-1})$.

En el caso de marcos finitos, como $\sigma(F, p)$ depende de N, para lograr una disminución del error incrementando la redundancia, debemos buscar marcos que, para alguna permutación p, tengan la variación acotada independientemente de N. Si conseguimos esto, la acotación va a ser del orden de N^{-1}.

Vamos a ver a continuación que tanto las raíces de la unidad como los marcos armónicos tienen la variación uniformemente acotada si $p = id$.

Ejemplo 3.4.1. ***Raíces de la unidad***

Si $R_N = \{e_n^N\}_{n=1}^N$, con $e_n^N = \big(cos\big(\frac{2\pi n}{N}\big), sen\big(\frac{2\pi n}{N}\big)\big)$, usando la proposición 3.3.2, y haciendo la identificación $e_n^N = e^{\frac{2\pi n}{N} i}$, podemos decir que

$$\|e_n^N - e_{n+1}^N\| \leq \Big|\frac{2\pi n}{N} - \frac{2\pi(n+1)}{N}\Big| = \frac{2\pi}{N}.$$

Si tomamos $p = id$, entonces

$$\sigma(R_N, p) = \sum_{n=1}^{N} \|e_n^N - e_{n+1}^N\| \leq \frac{2\pi}{N} N = 2\pi \qquad \forall\ N.$$

Con lo cual la variación queda acotada independientemente del valor N. Observemos además que estos marcos tienen la condición de suma cero, ya que

$$\begin{aligned}\sum_{n=1}^{N} e_n^N &= \Big(\sum_{n=1}^{N} cos\Big(\frac{2\pi n}{N}\Big), \sum_{n=1}^{N} sen\Big(\frac{2\pi n}{N}\Big)\Big) \\ &= \Big(Re\Big(\sum_{n=1}^{N} e^{\frac{2\pi i n}{N}}\Big), Im\Big(\sum_{n=1}^{N} e^{\frac{2\pi i n}{N}}\Big)\Big) = (0,0).\end{aligned}$$

Si $x \in \mathbb{R}^d$ es tal que $\|x\| \leq (K - 1/2)\delta$, usando el corolario 3.4.2,

$$\|x - \tilde{x}_N\| \leq \begin{cases} \frac{\delta\pi}{N} & \text{si } N \text{ es par} \\ \frac{\delta}{2N}(2\pi + 1) & \text{si } N \text{ es impar.} \end{cases}$$

Ejemplo 3.4.2. ***Marcos armónicos***
Recordemos que dado $N \geq d$ ($d \geq 2$), la familia $H_N^d = \{e_j^N\}_{j=0}^{N-1}$ se definía de la siguiente forma:
Si d es par

$$e_j^N = \sqrt{\frac{2}{d}}\Big[cos\Big(\frac{2\pi j}{N}\Big), sen\Big(\frac{2\pi j}{N}\Big), cos\Big(\frac{2\pi 2j}{N}\Big), sen\Big(\frac{2\pi 2j}{N}\Big), \dots \\ \dots, cos\Big(\frac{2\pi \frac{d}{2} j}{N}\Big), sen\Big(\frac{2\pi \frac{d}{2} j}{N}\Big)\Big].$$

Si d es impar

$$e_j^N = \sqrt{\frac{2}{d}}\Big[\frac{1}{\sqrt{2}}, cos\Big(\frac{2\pi j}{N}\Big), sen\Big(\frac{2\pi j}{N}\Big), cos\Big(\frac{2\pi 2j}{N}\Big), sen\Big(\frac{2\pi 2j}{N}\Big), \dots \\ \dots, cos\Big(\frac{2\pi \frac{d-1}{2} j}{N}\Big), sen\Big(\frac{2\pi \frac{d-1}{2} j}{N}\Big)\Big].$$

La condición de suma cero se cumple sólo en el caso en que d sea par. En el caso impar vale que

$$\sum_{j=0}^{N-1} e_j^N = \Big(\frac{N}{\sqrt{d}}, 0, \dots, 0\Big).$$

Veamos que esta familia tiene la variación acotada independientemente de N.
Supongamos d par, $p = id$, si usamos la propiedad de la proposición 3.3.2

obtenemos que

$$\begin{aligned}
\sqrt{\frac{d}{2}}\sigma(H_N^d, id) &= \sqrt{\frac{d}{2}}\sum_{j=0}^{N-2}\|e_j^N - e_{j+1}^N\| \\
&\leq \sum_{j=0}^{N-2}\Bigg[\sum_{k=1}^{d/2}\Big(cos\frac{2\pi kj}{N} - cos\frac{2\pi k(j+1)}{N}\Big)^2 + \\
&\quad \Big(sen\frac{2\pi kj}{N} - sen\frac{2\pi k(j+1)}{N}\Big)^2\Bigg]^{1/2} \\
&= \sum_{j=0}^{N-2}\Bigg[\sum_{k=1}^{d/2}\Big\|e^{\frac{2\pi kj}{N}i} - e^{\frac{2\pi k(j+1)}{N}i}\Big\|^2\Bigg]^{1/2} \leq \sum_{j=0}^{N-2}\Bigg[\sum_{k=1}^{d/2}\Big(\frac{2\pi k}{N}\Big)^2\Bigg]^{1/2} \\
&\leq 2\pi\Bigg[\sum_{k=1}^{d/2}k^2\Bigg]^{1/2} = 2\pi\Bigg[\frac{d(d/2+1)(d+1)}{12}\Bigg]^{1/2} \\
&\leq 2\pi\sqrt{\frac{d}{12}}(d+1).
\end{aligned}$$

Si d es impar, utilizando las mismas acotaciones, resulta que

$$\sqrt{\frac{d}{2}}\sigma(H_n^d, id) \leq 2\pi\Bigg[\sum_{k=1}^{(d-1)/2}k^2\Bigg]^{1/2} \leq 2\pi\sqrt{\frac{d}{12}}(d+1).$$

Concluimos entonces que tanto para d par como para d impar vale que

$$\sigma(H_N^d, p) \leq \frac{2\pi(d+1)}{\sqrt{6}} \qquad \forall\ N.$$

La cota para el error, tomando $x \in \mathbb{R}^d$ tal que $\|x\| \leq (K-1/2)\delta$, y usando que estos marcos tienen la condición de suma cero si d es par, resulta ser

$$\|x - \tilde{x}_N\| \leq \begin{cases} \frac{\delta d}{2N}\frac{2\pi(d+1)}{\sqrt{6}} & \text{si } d \text{ es par y } N \text{ es par} \\ \frac{\delta d}{2N}\big[\frac{2\pi(d+1)}{\sqrt{6}} + 1\big] & \text{en otro caso.} \end{cases}$$

El siguiente teorema establece que en $\mathbb{R}^2$ siempre existe una permutación p de forma que la variación del marco sea uniformemente acotada, por lo que la cota del error al aplicar el algoritmo $\Sigma\Delta$ en $\mathbb{R}^2$ resultará siempre del tipo $O(N^{-1})$.

Teorema 3.4.2. *Sea $F_N = \{e_n\}_{n=1}^N$ un marco uniforme ajustado de $\mathbb{R}^2$, con $e_n = (cos(\alpha_n), sen(\alpha_n))$, y $0 \leq \alpha_n < 2\pi$. Si p es una permutación de $\{1, \ldots, N\}$ tal que $\alpha_{p(n)} \leq \alpha_{p(n+1)}$ para todo $n \in \{1, \ldots, N-1\}$, entonces $\sigma(F_N, p) \leq 2\pi$.*

Demostración. Identificando $e_n = e^{\alpha_n i}$ y aplicando la proposición 3.3.2, tenemos que

$$\|e_{p(n)} - e_{p(n+1)}\| \leq |\alpha_{p(n)} - \alpha_{p(n+1)}| = \alpha_{p(n+1)} - \alpha_{p(n)},$$

entonces

$$\begin{aligned}\sigma(F_N, p) &= \sum_{n=1}^{N-1} \|e_{p(n)} - e_{p(n+1)}\| \leq \sum_{n=1}^{N-1} (\alpha_{p(n+1)} - \alpha_{p(n)}) \\ &= \alpha_{p(N)} - \alpha_{p(1)} \leq \alpha_{p(N)} \leq 2\pi.\end{aligned}$$

□

A partir del último teorema y de los dos ejemplos mencionados anteriormente, podemos explicar la aparición de la permutación p en el algoritmo $\Sigma\Delta$ para vectores de $\mathbb{R}^d$. El hecho de poder elegir el orden de cuantificación, nos permite lograr que la variación del marco no dependa de N.
En el caso de funciones de banda limitada no hacía falta esto, ya que la variación:

$$\sum_{n\in\mathbb{Z}} \left| g\Big(t - \frac{n}{\lambda}\Big) - g\Big(t - \frac{n+1}{\lambda}\Big) \right|$$

era acotada por $\|g'\|_{L^1}$, y tal como lo hemos mencionado en otra oportunidad, tomando $g_\lambda = g_{\lambda_0}$ para todo $\lambda \geq \lambda_0$, se obtiene una constante independiente de λ.

3.4.2. Comparación entre los esquemas Sigma-Delta ($\Sigma\Delta$) y los de Modulación por Impulsos Codificados

En la sección correspondiente al algoritmo PCM probamos que el error cuadrático medio alcanzado bajo la suposición de ruido blanco de Bennett era

$$ECM_{PCM} = \frac{d^2\delta^2}{12N}.$$

Pero luego vimos que esa suposición podía fallar en ciertos casos, y que se basaba en hechos probabilísticos.

Al considerar el algoritmo $\Sigma\Delta$ de orden uno obtuvimos la cota para el error

$$\|x - \tilde{x}\| \leq \frac{\delta d}{2N}(\sigma(F,p) + 1). \tag{3.10}$$

Veamos cual sería el error cuadrático correspondiente a este algoritmo.

Teorema 3.4.3. *Si $B \subseteq \{x \in \mathbb{R}^d : \|x\| \leq (K - 1/2)\delta\}$, definimos el error cuadrático medio del esquema $\Sigma\Delta$ sobre B como*

$$ECM_{\Sigma\Delta} = \int_B \|x - \tilde{x}\|^2 d\mu(x),$$

donde μ es una medida de probabilidad sobre B. Entonces vale que

$$ECM_{\Sigma\Delta} \leq \frac{\delta^2 d^2}{4N^2}(\sigma(F,p) + 1)^2.$$

Demostración. Es trivial debido a (3.10). □

Si tenemos un marco con $\sigma(F,p)$ acotado uniformemente, podemos decir a partir del teorema anterior que

$$ECM_{\Sigma\Delta} \lesssim \frac{1}{N^2}.$$

(La notación $A \lesssim B$ implica que existe una constante $C > 0$ tal que $A \leq CB$).

La acotación obtenida para el error cuadrático medio en los algoritmos $\Sigma\Delta$ es mucho mejor que la que teníamos para el esquema PCM (pues en ese caso $ECM_{PCM} \backsim \frac{1}{N}$) y es válida aún sin hacer suposiciones probabilísticas.

Debemos aclarar que el algoritmo $\Sigma\Delta$ ofrece errores más pequeños que el algoritmo PCM cuando la redundancia del marco es grande. Si N es chico, puede suceder que el PCM supere en optimicidad al $\Sigma\Delta$. De hecho, si $N = d$ la constante del marco es igual a uno, entonces, tal como vimos en la sección de marcos finitos, el marco es una base ortonormal. En ese caso sucede que

$$\|x - \tilde{x}\|^2 = \Big\| \sum_{n=1}^{N} (x_n - q_n) e_n \Big\|^2 = \sum_{n=1}^{N} |x_n - q_n|^2.$$

Esta expresión va a minimizarse cuando cada diferencia $x_n - q_n$ sea lo más pequeña posible y esto se logra asignando el q_n más próximo a x_n en el alfabeto, es decir, aplicando el algoritmo PCM.

3.4.3. Cuantificación Sigma-Delta (ΣΔ) de órdenes mayores

El algoritmo ΣΔ de orden mayor que uno para vectores de $\mathbb{R}^d$, se define de manera análoga al de funciones de banda limitada, sólo que vamos a tener en cuenta, al igual que en el caso de orden uno, a la permutación p que marca el orden en que se cuantifican los coeficientes.
Por una cuestión de conveniencia en la notación, llamaremos v a la variable interna en lugar de u.

Definición 3.4.3. *Dados $K \in \mathbb{N}$, $\delta > 0$ y el alfabeto $\mathcal{A}_K^\delta = \{(-K + 1/2)\delta, (-K + 3/2)\delta, \ldots, (K - 3/2)\delta, (K - 1/2)\delta\}$, junto con el cuantificador $Q(u) = arg\{\text{mín}_{q\in\mathcal{A}_K^\delta} |u - q|\}$. Sea $\{x_n\}_{n=1}^N$ la sucesión de coeficientes del vector $x \in \mathbb{R}^d$ en el marco $\{e_n\}_{n=1}^N$, que es finito, ajustado y uniforme, y sea p una permutación de $\{1, 2, \ldots, N\}$.*
El esquema de cuantificación ΣΔ de orden k se define, para $1 \leq n \leq N$, por la siguiente iteración:

$$\begin{cases} \Delta_n^k v & = x_{p(n)} - q_n \\ q_n & = Q\big(F\big(\Delta_{n-1}^{k-1} v, \Delta_{n-1}^{k-2} v, \ldots, \Delta_{n-1}^0 v, x_{p(n)}\big)\big), \end{cases} \tag{3.11}$$

donde $F : \mathbb{R}^{k+1} \longrightarrow \mathbb{R}$ es una función que permite la estabilidad del sistema, y $v_i = 0$ para todo $i \in \{-k + 1, \ldots, -1, 0\}$.

En el capítulo correspondiente a funciones de banda limitada vimos que la estabilidad para sistemas de orden mayor que uno no se prueba fácilmente. Hablamos del enfoque utilizado en [DDV] y probamos que para cualquier sistema estable de orden k, el error de aproximación es del tipo de $O(\lambda^{-k})$.

Para marcos finitos la prueba de la estabilidad tampoco resultará sencilla. Estudiaremos los resultados obtenidos por J.Benedetto, A.Powell y O.Yilmaz en [BPY], quienes utilizan la teoría de conjuntos invariantes para probar la estabilidad de ciertos algoritmos de orden dos. Debido a que las ideas expuestas en [BPY] no pueden generalizarse a mayores órdenes, en esta sección sólo trabajaremos con esquemas de segundo orden.

Con respecto al error de aproximación, veremos que para marcos finitos, a diferencia de lo que sucede para funciones de banda limitada, no siempre se va a lograr un error del tipo $O(N^{-2})$. De hecho veremos que para las raíces N-ésimas de la unidad, en el caso de N impar, el error se comporta asintóticamente como $\frac{1}{N}$.

Para definir el esquema $\Sigma\Delta$ de orden dos, primero tomaremos $u_n := \Delta_n^1 v = v_n - v_{n-1}$ en (3.11), por lo que obtendremos el algoritmo:

$$\begin{cases} u_n = u_{n-1} + x_{p(n)} - q_n \\ v_n = u_n + v_{n-1} \\ q_n = Q\big(F\big(u_{n-1}, v_{n-1}, x_{p(n)}\big)\big), \end{cases}$$

con $u_0 = v_0 = 0$.
Para simplificar el estudio del algoritmo, utilizaremos el alfabeto de dos elementos y paso δ, es decir, $\mathcal{A}_1^\delta = \{-\frac{\delta}{2}, \frac{\delta}{2}\}$. El cuantificador Q con este alfabeto resulta ser

$$Q(x) = \frac{\delta}{2}\operatorname{sign}(x) = \begin{cases} \frac{\delta}{2} & \text{si } x \geq 0 \\ \frac{-\delta}{2} & \text{si } x < 0. \end{cases}$$

En el capítulo de funciones de banda limitada habíamos hablado de distintas elecciones de la función F del esquema, se trabajará en esta sección con $F(u, v, x) = u + \gamma v$, para $\gamma > 0$ fijo.

El esquema $\Sigma\Delta$ de orden 2 que utilizaremos será:

Definición 3.4.4.

$$\begin{cases} u_n = u_{n-1} + x_{p(n)} - q_n \\ v_n = v_{n-1} + u_n \\ q_n = \frac{\delta}{2}\, sign(u_{n-1} + \gamma v_{n-1}), \end{cases} \tag{3.12}$$

donde $n \in \{1, \ldots, N\}$, $u_0 = v_0 = 0$, p *es una permutación de* $\{1, \ldots, N\}$, *y* $\gamma > 0$.

Tal como lo hicimos para el algoritmo de orden uno, primero estudiaremos la estabilidad y luego el error de aproximación.

Estabilidad de los esquemas Sigma-Delta (ΣΔ) de orden dos

Para simplificar los cálculos supondremos que $p = id$ y $\delta = 2$ en la fórmula (3.12), veremos luego cómo extenderlo al caso general.

El esquema con estas suposiciones resulta

$$\begin{cases} u_n = u_{n-1} + x_n - q_n \\ v_n = v_{n-1} + u_n \\ q_n = \text{sign}(u_{n-1} + \gamma v_{n-1}), \end{cases} \tag{3.13}$$

donde $n \in \{1, \ldots, N\}$, $u_0 = v_0 = 0$ y $\gamma > 0$.

Para demostrar la estabilidad del sistema probaremos que existe un conjunto $R \subseteq \mathbb{R}^2$ acotado, tal que para ciertos $0 < \alpha < 1$ y $\gamma > 0$,

$$|x_n| \leq \alpha \quad \forall\, n \in \{1, \ldots, N\} \Longrightarrow (u_n, v_n) \in R \quad \forall\ 0 \leq n \leq N.$$

Esto probaría la estabilidad, ya que las variables internas se encontrarían dentro de un conjunto acotado.

Para hallar la región R, definiremos una función S tal que

$$(u_n, v_n) = S(u_{n-1}, v_{n-1}, x_n) \quad \forall\ 1 \leq n \leq N, \tag{3.14}$$

y probaremos que

$$\forall\, |x| \leq \alpha, \quad (u, v) \in R \Longrightarrow S(u, v, x) \in R.$$

El primer paso hacia la prueba de la estabilidad consistirá entonces en encontrar la función S de forma que se cumpla (3.14). Trabajaremos con coeficientes que cumplan $|x_n| \leq \alpha$, con $0 < \alpha < 1$ (las condiciones sobre α las estableceremos más adelante).

Dado $\delta_n := |x_n - q_n|$, vale que

$$u_n + \gamma v_n \geq 0 \Longrightarrow q_n = 1 \Longrightarrow q_n > \alpha > x_n \Longrightarrow \delta_n = q_n - x_n,$$

y

$$u_n + \gamma v_n < 0 \Longrightarrow q_n = -1 \Longrightarrow q_n < -\alpha < x_n \Longrightarrow \delta_n = x_n - q_n.$$

Usando el esquema (3.13),

$$u_n + \gamma v_n \geq 0 \Longrightarrow u_n = u_{n-1} - \delta_n \ \text{ y } \ v_n = v_{n-1} + u_{n-1} - \delta_n,$$

y

$$u_n + \gamma v_n < 0 \Longrightarrow u_n = u_{n-1} + \delta_n \quad \text{y} \quad v_n = v_{n-1} + u_{n-1} + \delta_n.$$

Si definimos

$$S_l^\delta(u, v) := (u - \delta, u + v - \delta)$$

y

$$S_r^\delta(u, v) := (u + \delta, u + v + \delta),$$

entonces

$$(u_n, v_n) = \begin{cases} S_l^{\delta_n}(u_{n-1}, v_{n-1}) & \text{si } u_n + \gamma v_n \geq 0 \\ S_r^{\delta_n}(u_{n-1}, v_{n-1}) & \text{si } u_n + \gamma v_n < 0. \end{cases}$$

Finalmente, podemos decir que

$$(u_n, v_n) = S(u_{n-1}, v_{n-1}, \delta_n),$$

donde

$$S(u, v, \delta) := \begin{cases} S_l^\delta(u, v) & \text{si } u + \gamma v \geq 0 \\ S_r^\delta(u, v) & \text{si } u + \gamma v < 0. \end{cases} \tag{3.15}$$

Una vez definida la función S, para probar la estabilidad veremos que para ciertos $0 \leq \alpha < 1$ y $\gamma > 0$, existe un conjunto $R \subseteq \mathbb{R}^2$ tal que si $\delta \in [1 - \alpha, 1 + \alpha]$ entonces

$$(u, v) \in R \Longrightarrow S(u, v, \delta) \in R.$$

Es decir que, si notamos $S_\delta := S(\cdot, \cdot, \delta) : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$, estamos buscando un conjunto $R \subseteq \mathbb{R}^2$ tal que

$$S_\delta(R) \subseteq R \quad \forall\ \delta \in [1 - \alpha, 1 + \alpha].$$

Definición 3.4.5. *Un conjunto $R \subseteq \mathbb{R}^d$ es invariante para el mapa $T : \mathbb{R}^d \longrightarrow \mathbb{R}^d$ si $T(R) \subseteq R$.*

En base a esta definición, podemos decir que estamos buscando un conjunto $R \subseteq \mathbb{R}^2$ que resulte invariante para el mapa $S_\delta := S(\cdot, \cdot, \delta) : \mathbb{R}^2 \longrightarrow \mathbb{R}^2$, para todo $\delta \in [1 - \alpha, 1 + \alpha]$.

Observemos que el mapa S_δ es afín a trozos, ya que si definimos

$$U_+ := \{(u, v) \in \mathbb{R}^2 \,/\, u + \gamma v \geq 0\} \text{ y } U_- := \{(u, v) \in \mathbb{R}^2 \,/\, u + \gamma v < 0\},$$

entonces

$$\forall\ (u,v) \in U_+, \quad S_\delta(u,v) = S_l^\delta(u,v) = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} + (-\delta) \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

y

$$\forall\ (u,v) \in U_-, \quad S_\delta(u,v) = S_r^\delta(u,v) = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} + \delta \begin{pmatrix} 1 \\ 1 \end{pmatrix}.$$

Dado que el esquema (3.13) puede ser visto como un sistema dinámico determinado por una regla que es afín a trozos, para hallar la región R buscada, se puede aplicar la teoría de conjuntos invariantes desarrollada para esta clase particular de sistemas. A continuación veremos la región definida por Benedetto, Powell y Yilmaz en [Yi] y [BPY].

Dados $0 \leq \alpha < 1$, $\delta_+ := 1+\alpha$, $\delta_- := 1-\alpha$ y $C > 0$ (que será determinado más adelante), se definen las funciones $B_1, B_2 : \mathbb{R} \longrightarrow \mathbb{R}$ de la siguiente forma:

$$B_1(u) = \begin{cases} -\frac{1}{2\delta_-}(u - \frac{\delta_-}{2})^2 + \frac{\delta_-}{8} + C & \text{si } u \geq 0 \\ -\frac{1}{2\delta_+}(u - \frac{\delta_+}{2})^2 + \frac{\delta_+}{8} + C & \text{si } u < 0 \end{cases} \tag{3.16}$$

y

$$B_2(u) = \begin{cases} \frac{1}{2\delta_+}(u + \frac{\delta_+}{2})^2 - \frac{\delta_+}{8} - C & \text{si } u \geq 0 \\ \frac{1}{2\delta_-}(u + \frac{\delta_-}{2})^2 - \frac{\delta_-}{8} - C & \text{si } u < 0. \end{cases} \tag{3.17}$$

Si Γ_{B_1} y Γ_{B_2} son los gráficos de las funciones B_1 y B_2 respectivamente, definimos el conjunto R como la región encerrada por ambas curvas. Es decir

$$R = \{(u,v) : B_2(u) \leq v \leq B_1(u)\}.$$

Dado que R depende de α, γ y C, para probar que la región es invariante para $S(\cdot,\cdot,\delta)$ con $\delta \in [\delta_-, \delta_+]$, se necesita imponer condiciones sobre esos parámetros.
Para enunciar el teorema de estabilidad demostrado en [BPY], primero se definen subregiones de R y coeficientes que involucran a α, γ y C.

Como la definición (3.15) de la función S depende de si $v \geq -\frac{1}{\gamma}u$ ó si $v < -\frac{1}{\gamma}u$, dada $l(u) := -\frac{1}{\gamma}u$, vamos a dividir la región R en dos subregiones:

$$R_1 = \{(u,v) : B_2(u) \leq v \leq B_1(u), v \geq l(u)\} = R \cap U_+$$

y

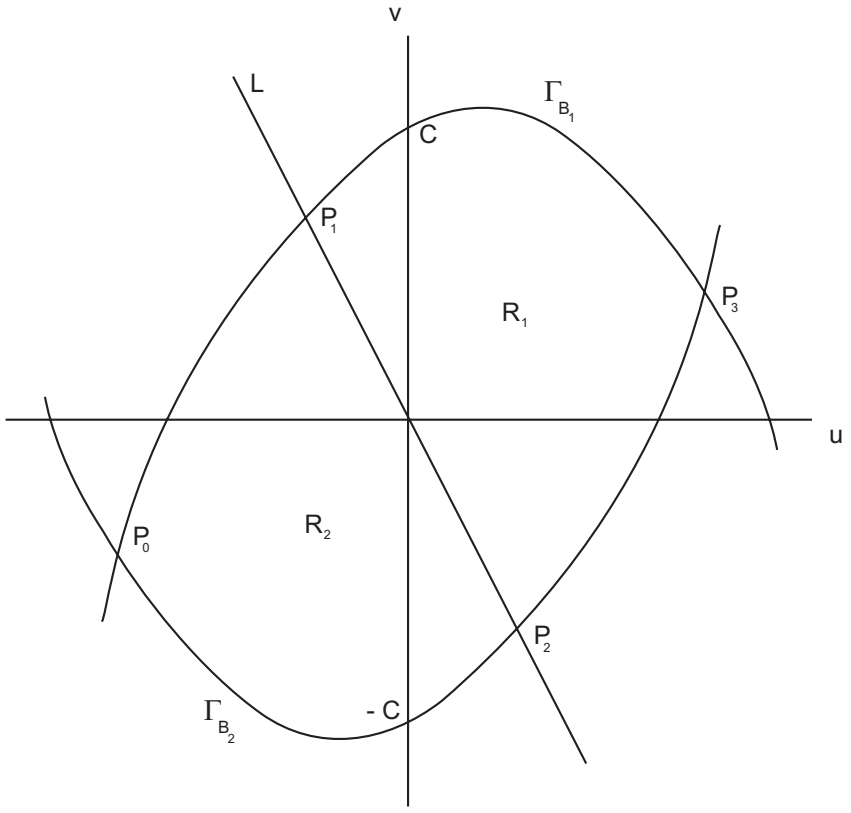
$$R_2 = \{(u,v) : B_2(u) \leq v \leq B_1(u), v < l(u)\} = R \cap U_-.$$

Si L es la gráfica de la recta l, definimos

$$P_0 = (a_0, b_0) := \Gamma_{B_1} \cap \Gamma_{B_2} \cap \{(u,v) \in \mathbb{R}^2 \,/\, u < 0\},$$

$$P_1 = (a_1, b_1) := L \cap \Gamma_{B_1}, \quad P_2 = (a_2, b_2) := L \cap \Gamma_{B_2}, \quad \text{y}$$
$$P_3 = (a_3, b_3) := \Gamma_{B_1} \cap \Gamma_{B_2} \cap \{(u,v) \in \mathbb{R}^2 \,/\, u > 0\} = (-a_0, -b_0).$$

Un esquema ilustrativo de las regiones y los puntos definidos anteriormente sería:



Usando la definición de las funciones B_1, B_2 y l, se tiene que

$$a_0 = -[2C(1-\alpha^2)]^{1/2}, \quad b_0 = B_1(a_0), \quad (a_2, b_2) = (-a_1, -b_1),$$

$$a_1 = \frac{(1+\alpha)(1+2/\gamma) - \sqrt{(1+\alpha)^2(1+2/\gamma)^2 + 8(1+\alpha)C}}{2} \quad \text{y} \quad b_1 = -\frac{1}{\gamma}a_1.$$

Enunciamos a continuación el teorema demostrado en [BPY] que sirve para probar que la región R es invariante para S_δ, con $\delta \in [\delta_-, \delta_+]$.

Teorema 3.4.4. *Sean P_0, P_1 y P_2 los puntos definidos anteriormente. Si se cumple que*

$$a_0 + \delta_+ \leq a_1, \tag{3.18}$$

$$a_2 + b_2 - \delta \geq B_2(a_2 - \delta) \qquad \textit{para} \quad \delta = \delta_+, \delta_- \tag{3.19}$$

y

$$a_1 + b_1 + \delta \leq B_1(a_1 + \delta) \qquad \textit{para} \quad \delta = \delta_+, \delta_-, \tag{3.20}$$

entonces $S_l^\delta(R_1) \subseteq R$ y $S_r^\delta(R_2) \subseteq R$ para todo $\delta \in [\delta_-, \delta_+]$.

La demostración del teorema no la incluiremos, se basa principalmente en la convexidad de la región R y la afinidad de S_l^δ y S_r^δ. La misma se puede encontrar en [BPY] y [Yi].
Veamos cómo a partir del teorema 3.4.4 se deducen corolarios que prueban la estabilidad de los algoritmos $\Sigma\Delta$ de orden dos para vectores de $\mathbb{R}^d$.

Corolario 3.4.3. *Si α, γ y C cumplen las condiciones (3.18), (3.19) y (3.20), entonces R es un conjunto invariante de $S(\cdot, \cdot, \delta)$ para todo $\delta \in [\delta_-, \delta_+]$. Es decir, $S(u, v, \delta) \in R$ para todo $(u, v) \in R$ y $\delta \in [\delta_-, \delta_+]$.*

Demostración. Dado $\delta \in [\delta_-, \delta_+]$ y $(u, v) \in R$ queremos probar que $S(u, v, \delta) \in R$.
Como $R = R_1 \cup R_2$, $(u, v) \in R_1$ ó $(u, v) \in R_2$.
Si $(u, v) \in R_1$, por cómo se define esa región, sucede que $u + \gamma v \geq 0$, entonces $S(u, v, \delta) = S_l^\delta(u, v)$. Por el teorema 3.4.4, sabemos que $S_l^\delta(R_1) \subseteq R$, resultando entonces que $S(u, v, \delta) \in R$.
Si $(u, v) \in R_2$, por cómo se define esa región, sucede que $u + \gamma v < 0$, entonces $S(u, v, \delta) = S_r^\delta(u, v)$, usando nuevamente el teorema 3.4.4, se prueba lo pedido.

□

Con este resultado podemos probar el siguiente teorema.

Teorema 3.4.5. *Dados α, γ y C que cumplen las condiciones (3.18), (3.19) y (3.20). Si $|x_n| \leq \alpha < 1$ para todo $1 \leq n \leq N$ y $\{u_n\}_{n=1}^N, \{v_n\}_{n=1}^N$ son las variables internas definidas por (3.13), con $u_0 = v_0 = 0$, entonces*

$$(u_n, v_n) \in R \qquad \forall\ n \in \{0, \ldots, N\}.$$

Demostración. Se prueba por inducción.
Como $(u_0, v_0) = (0, 0)$, $B_1(0) = C$ y $B_2(0) = -C$, entonces $B_2(0) \leq 0 \leq B_1(0)$, por lo que $(u_0, v_0) \in R$.
Para continuar observemos que $\delta_n \in [\delta_-, \delta_+] = [1 - \alpha, 1 + \alpha] \quad \forall\ 1 \leq n \leq N$, ya que $\delta_n = |x_n - q_n|$ y $|x_n| \leq \alpha$.
Supongamos que $(u_{n-1}, v_{n-1}) \in R$, como $(u_n, v_n) = S(u_{n-1}, v_{n-1}, \delta_n)$, usando el corolario anterior se deduce inmediatamente que $(u_n, v_n) \in R$.

□

Con el teorema 3.4.5 se concluye la estabilidad del sistema (3.13), ya que definiendo $C' := \frac{\delta_-}{8} + C$, se cumple que

$$R \subseteq [a_0, -a_0] \times [-C', C']$$

(pues C' representa el valor máximo de la función B_1 y $-C'$ el mínimo de B_2). Dado que las variables internas se mantienen dentro del conjunto R, esto nos dice que están acotadas.

Como el resultado anterior vale para el caso en que el esquema ΣΔ usa para cuantificar la función sign(x) (utiliza $\delta = 2$) y $p = id$, nos resta generalizar la prueba para cualesquiera $\delta > 0$ y p permutación de $\{1, \ldots, N\}$.

Teorema 3.4.6. *Si $|x_n| \leq \frac{\delta}{2}\alpha < 1$ para todo $1 \leq n \leq N$ y $\{u_n\}_{n=1}^N, \{v_n\}_{n=1}^N$ son las variables internas definidas por (3.12), con $u_0 = v_0 = 0$. Dados α, γ y C que cumplen las condiciones (3.18), (3.19) y (3.20), entonces*

$$(u_n, v_n) \in \frac{\delta}{2} R \qquad \forall\ n \in \{0, \ldots, N\}.$$

Demostración. Definimos la sucesión $\widetilde{x}_n := \frac{2}{\delta} x_{p(n)}$ y las variables $\widetilde{u}_n := \frac{2}{\delta} u_n$ y $\widetilde{v}_n := \frac{2}{\delta} v_n$. Como $\{u_n\}_{n=1}^N, \{v_n\}_{n=1}^N$ cumplen el esquema (3.12), resulta que $\{\widetilde{u}_n\}_{n=1}^N, \{\widetilde{v}_n\}_{n=1}^N$ chequean la definición (3.13). Además $|\widetilde{x}_n| \leq \alpha$, entonces vale el teorema 3.4.5, por lo que

$$(\widetilde{u}_n, \widetilde{v}_n) \in R \qquad \forall\ n \in \{1, \ldots, N\}.$$

El resultado de este teorema sigue de observar que $(\widetilde{u}_n, \widetilde{v}_n) = \left(\frac{2}{\delta}u_n, \frac{2}{\delta}v_n\right)$.

□

Observación 3.4.1. *Como*

$$R \subseteq [a_0, -a_0] \times [-C', C'],$$

el teorema 3.4.6 nos dice que

$$(u_n, v_n) \in \left[\frac{\delta}{2}a_0, -\frac{\delta}{2}a_0\right] \times \left[-\frac{\delta}{2}C', \frac{\delta}{2}C'\right] \qquad \forall\ n \in \{1, \ldots, N\},$$

es decir que

$$|u_n| \leq -\frac{\delta}{2}a_0 \qquad y \qquad |v_n| \leq \frac{\delta}{2}C' \qquad \forall\ n \in \{1, \ldots, N\}.$$

Estimación del error de aproximación producido al cuantificar con el algoritmo Sigma-Delta (ΣΔ) de orden dos

Una vez finalizada la prueba de la estabilidad para el esquema (3.12), comenzaremos con el análisis del error de aproximación. Tal como lo hicimos para el algoritmo de orden uno, vamos a definir un coeficiente que mide la variación del marco en relación al orden en que se cuantifican los coeficientes.

Definición 3.4.6. *Sea $F = \{e_n\}_{n=1}^N$ un marco finito en $\mathbb{R}^d$ y sea p una permutación de $\{1, \ldots, N\}$, se define la variación de orden k del marco F respecto a p como*

$$\sigma_k(F, p) = \sum_{n=1}^{N-k} \|\Delta^k e_{p(n+k)}\|,$$

donde Δ^k es el operador diferencial discreto de orden k.

Para los esquemas de orden dos vamos a usar la variación de orden dos:

$$\sigma_2(F, p) = \sum_{n=1}^{N-2} \| \ e_{p(n+2)} - 2e_{p(n+1)} + e_{p(n)}\|.$$

Teorema 3.4.7. *Sea $F = \{e_n\}_{n=1}^N$ un marco ajustado uniforme de $\mathbb{R}^d$ y sea p una permutación de $\{1, \dots, N\}$. Dado $x \in \mathbb{R}^d$, si $\{q_n\}_{n=1}^N$ son los coeficientes producidos por la iteración (3.12) y $\tilde{x} = \frac{d}{N}\sum_{n=1}^N q_n e_{p(n)}$, entonces*

$$\|x - \tilde{x}\| \leq \frac{d}{N}\big(\|v\|_\infty \sigma_2(F,p) + |v_{N-1}|\|e_{p(N-1)} - e_{p(N)}\| + |u_N|\big).$$

Demostración. Usando el lema 3.4.1 de sumación por partes y definiendo $f_n := e_{p(n)} - e_{p(n+1)}$, obtenemos que

$$\begin{aligned}
x - \tilde{x} &= \frac{d}{N}\sum_{n=1}^N (x_{p(n)} - q_n)e_{p(n)} = \frac{d}{N}\sum_{n=1}^N (u_n - u_{n-1})e_{p(n)} \\
&= \frac{d}{N}\Big(\sum_{n=1}^{N-1} u_n(e_{p(n)} - e_{p(n+1)}) - u_0 e_{p(1)} + u_N e_{p(N)}\Big) \\
&= \frac{d}{N}\Big(\sum_{n=1}^{N-1} (v_n - v_{n-1})f_n + u_N e_{p(N)}\Big) \\
&= \frac{d}{N}\Big(\sum_{n=1}^{N-2} v_n(f_n - f_{n+1}) + v_{N-1}f_{N-1} - v_0 f_1 + u_N e_{p(N)}\Big) \\
&= \frac{d}{N}\Big(\sum_{n=1}^{N-2} v_n(f_n - f_{n+1}) + v_{N-1}f_{N-1} + u_N e_{p(N)}\Big) \\
&= \frac{d}{N}\Big(\sum_{n=1}^{N-2} v_n(e_{p(n)} - 2e_{p(n+1)} + e_{p(n+2)}) + v_{N-1}(e_{p(N-1)} - e_{p(N)}) + u_N e_{p(N)}\Big).
\end{aligned} \tag{3.21}$$

Entonces

$$\begin{aligned}
\|x - \tilde{x}\| &\leq \frac{d}{N}\Big(\|v\|_\infty \sum_{n=1}^{N-2} \|\Delta^2 e_{p(n+2)}\| + |v_{N-1}|\|(e_{p(N-1)} - e_{p(N)})\| + |u_N|\Big) \\
&\leq \frac{d}{N}\big(\|v\|_\infty \sigma_2(F,p) + |v_{N-1}|\|e_{p(N-1)} - e_{p(N)}\| + |u_N|\big).
\end{aligned}$$

□

Tal como en el caso de orden uno, tenemos un lema que establece el valor del término u_N.

Lema 3.4.3. *Sea $F = \{e_n\}_{n=1}^N$ un marco finito, ajustado y uniforme de $\mathbb{R}^d$, y sea $S = \sum_{n=1}^N e_n$. Si $x \in \mathbb{R}^d$, entonces*

$$u_N \in \begin{cases} \langle x, S\rangle + \delta\mathbb{Z} & \textit{si } N \textit{ es par} \\ \langle x, S\rangle + \delta(\mathbb{Z} + \frac{1}{2}) & \textit{si } N \textit{ es impar.} \end{cases}$$

Demostración.

$$u_N = u_0 + \sum_{j=1}^N x_j - \sum_{j=1}^N q_j = \sum_{j=1}^N \langle x, e_j\rangle - \sum_{j=1}^N q_j = \langle x, S\rangle - \sum_{j=1}^N q_j.$$

Como $q_n \in \{-\delta/2, \delta/2\}$, con un análisis similar al realizado para los esquemas de orden uno, podemos decir que

$$\sum_{j=1}^N q_j \in \begin{cases} \delta\mathbb{Z} & \text{si N es par} \\ \delta(\mathbb{Z} + 1/2) & \text{si N es impar.} \end{cases}$$

□

Recordemos que para los esquemas de orden uno, si el marco F cumple la condición de suma cero, entonces

$$|u_N| = \begin{cases} 0 & \text{si N es par} \\ \frac{\delta}{2} & \text{si N es impar.} \end{cases}$$

La demostración de esta propiedad se basaba en el hecho de que $|u_N| \leq \frac{\delta}{2}$. Incluso, si se tiene en cuenta la prueba del lema 3.4.2, basta usar que $|u_N| < \delta$. Si queremos obtener una propiedad análoga para orden dos, utilizando el lema 3.4.3, deberíamos probar que bajo ciertos requisitos $|u_N| < \delta$.

A partir de la observación 3.4.1 sabemos que pidiendo que α, γ y C cumplan las condiciones (3.18), (3.19) y (3.20), entonces $|u_N| \leq -\frac{\delta}{2}a_0$. Si además pedimos que $a_0 > -2$, entonces $|u_N| < \delta$.

En el caso en que α, γ y C verifiquen (3.18), (3.19), (3.20) y $a_0 > -2$, diremos que α, γ y C cumplen la condición (A).

Para enunciar el siguiente teorema, recordemos que la notación $A \lesssim B$ es utilizada para expresar que existe una constante $\beta > 0$ tal que $A \leq \beta B$.

Corolario 3.4.4. *Sea $F = \{e_n\}_{n=1}^N$ un marco finito ajustado uniforme en $\mathbb{R}^d$ que cumple la condición de suma cero, p una permutación de $\{1,\dots,N\}$ y $x \in \mathbb{R}^d$ tal que $\|x\| \leq \frac{\delta}{2}\alpha$, con los parámetros α, γ y C que satisfacen la condición (A). Si el marco cumple que*

$$\sigma_2(F,p) \lesssim \frac{1}{N} \qquad y \qquad \|\Delta e_n\| \lesssim \frac{1}{N} \quad \forall\, n \in \{1,\dots,N\},$$

entonces si N es par, vale que

$$\|x - \tilde{x}\| \lesssim \frac{d\delta}{N^2},$$

Si N es impar y $N \geq 3$, vale que

$$\frac{d\delta}{N} \lesssim \|x - \tilde{x}\| \lesssim \frac{d\delta}{N}.$$

Las constantes que intervienen en la acotación no dependen de x ni de N.

Demostración. Como α, γ, C cumplen la condición (A), por lo observado anteriormente, vale que

$$(u_n, v_n) \in \frac{\delta}{2}R \quad \forall\, n \in \{0,\dots,N\} \qquad \text{y} \qquad R \subseteq (-2,2) \times [-C', C'],$$

entonces

$$|u_n| < \delta \tag{3.22}$$

y

$$|v_n| \leq C'\frac{\delta}{2} \quad \forall\, n \in \{1,\dots,N\}. \tag{3.23}$$

Si N es par, usando el lema 3.4.3 y que $S = \sum_{n=1}^N e_n = 0$ (condición de suma cero), sucede que $u_N \in \delta\mathbb{Z}$. Por (3.22) podemos decir que $u_N = 0$, por lo que la cota obtenida en el teorema 2.4.1 resulta

$$\begin{aligned}\|x - \tilde{x}\| &\leq \frac{d}{N}\big(\|v\|_\infty \sigma_2(F,p) + |v_{N-1}|\|e_{p(N-1)} - e_{p(N)}\| + |u_N|\big) \\ &\lesssim \frac{d}{N}\Big(\frac{C'\delta}{2}\frac{1}{N} + \frac{C'\delta}{2}\frac{1}{N}\Big) \lesssim \frac{d\delta}{N^2}.\end{aligned}$$

Si N es impar, por el lema 3.4.3, $u_N \in \delta(\mathbb{Z}+\frac{1}{2})$. Usando (3.22), vale que $|u_N| = \frac{\delta}{2}$, entonces

$$\begin{aligned}\|x-\tilde{x}\| &\leq \frac{d}{N}\big(\|v\|_\infty \sigma_2(F,p) + |v_{N-1}|\|e_{p(N-1)} - e_{p(N)}\| + |u_N|\big)\\ &\lesssim \frac{d}{N}\Big(\frac{C'\delta}{2}\frac{1}{N} + \frac{C'\delta}{2}\frac{1}{N} + \frac{\delta}{2}\Big) \lesssim \frac{d\delta}{N}.\end{aligned}$$

Para la acotación inferior usaremos la igualdad obtenida en (3.21) en el teorema 3.4.7:

$$x-\tilde{x} = \frac{d}{N}\Big(\sum_{n=1}^{N-2} v_n(e_{p(n)}-2e_{p(n+1)}+e_{p(n+2)}) + v_{N-1}(e_{p(N-1)}-e_{p(N)}) + u_N e_{p(N)}\Big).$$

A partir de esta expresión podemos decir que

$$\frac{d}{N}(u_N e_{p(N)}) = x-\tilde{x}-\frac{d}{N}\Big(\sum_{n=1}^{N-2} v_n(e_{p(n)}-2e_{p(n+2)}+e_{p(n+1)}) + v_{N-1}(e_{p(N-1)}-e_{p(N)})\Big).$$

Aplicando la norma euclidea queda que

$$\begin{aligned}\frac{d}{N}\frac{\delta}{2} = \frac{d}{N}|u_N|\|e_{p(N)}\| &\leq \|x-\tilde{x}\| + \frac{d}{N}(\|v\|_\infty \sigma_2(F,p) + |v_{N-1}|\|\Delta e_n\|)\\ &\lesssim \|x-\tilde{x}\| + \frac{d}{N}\Big(\frac{C'\delta}{2}\frac{1}{N} + \frac{C'\delta}{2}\frac{1}{N}\Big) \lesssim \|x-\tilde{x}\| + \frac{d\delta}{N^2},\end{aligned}$$

entonces

$$\|x-\tilde{x}\| \gtrsim \frac{d}{N}\frac{\delta}{2} - \frac{d\delta}{N^2} = \frac{d\delta}{N}\Big(\frac{1}{2}-\frac{1}{N}\Big).$$

Si $N \geq 3$, podemos afirmar que

$$\|x-\tilde{x}\| \gtrsim \frac{d\delta}{N}.$$

□

Como conclusión del corolario 3.4.4, podemos decir, que al aproximar un vector de $\mathbb{R}^d$ a partir de la cuantificación realizada con los esquemas $\Sigma\Delta$ de orden dos, no siempre se va a obtener una cota del tipo $O(N^{-2})$. Esto es muy diferente a lo que pasa en el caso de funciones de banda limitada, en donde siempre se podía lograr una estimación del tipo $O(\lambda^{-2})$. La razón de esta

diferencia de comportamiento en los errores de los distintos tipos de señales, se debe a la aparición en la expresión de la suma por partes de términos acotados no nulos.

A continuación veremos que los marcos armónicos y las raíces de la unidad cumplen con las condiciones del corolario 3.4.4.

Ejemplo 3.4.3. ***Marcos armónicos***

Si $H_N^d = \{e_n\}_{n=0}^{N-1}$ son los marcos armónicos, vimos que para d par se cumple la condición de suma cero. Supongamos que $p = id$, entonces a partir de la proposición 3.3.2 se demuestra que

$$\|\Delta e_n\| \leq \frac{\pi d}{N}.$$

Además usando el teorema del valor medio, si $\alpha_j^k = \frac{2\pi kj}{N}$, tenemos que

$$\begin{aligned} cos(\alpha_j^k) - 2cos(\alpha_{j+1}^k) + cos(\alpha_{j+2}^k) &= sen(\beta_j^k)\frac{2\pi k}{N} - sen(\beta_{j+1}^k)\frac{2\pi k}{N} \\ &= -\big(sen(\beta_{j+1}^k) - sen(\beta_j^k)\big)\frac{2\pi k}{N} \\ &= -cos(\gamma_j^k)(\beta_{j+1}^k - \beta_j^k)\frac{2\pi k}{N}, \end{aligned}$$

donde $\beta_j^k \in \big(\alpha_j^k, \alpha_{j+1}^k\big)$ y $\gamma_j^k \in \big(\beta_j^k, \beta_{j+1}^k\big)$.
Como

$$0 < \beta_{j+1}^k - \beta_j^k < \alpha_{j+2}^k - \alpha_j^k = 2\frac{2\pi k}{N},$$

resulta que

$$\begin{aligned} &\Big(cos(\alpha_j^k) - 2cos(\alpha_{j+1}^k) + cos(\alpha_{j+2}^k)\Big)^2 = \big(cos(\gamma_j^k)\big)^2\big(\beta_{j+1}^k - \beta_j^k\big)^2\Big(\frac{2\pi k}{N}\Big)^2 \\ &\leq \Big(2\frac{2\pi k}{N}\Big)^2\Big(\frac{2\pi k}{N}\Big)^2 = 4\Big(\frac{2\pi k}{N}\Big)^4. \end{aligned}$$

De forma análoga se puede probar que

$$\Big(sen(\alpha_j^k) - 2sen(\alpha_{j+1}^k) + sen(\alpha_{j+2}^k)\Big)^2 \leq 4\Big(\frac{2\pi k}{N}\Big)^4.$$

Finalmente obtenemos que

$$\begin{aligned}\sqrt{\frac{d}{2}}\sigma_2(H_N^d, id) &= \sqrt{\frac{d}{2}}\sum_{j=0}^{N-3}\|e_j^N - 2e_{j+1}^N + e_{j+2}^N\| \\ &\leq \sum_{j=0}^{N-3}\Big[\sum_{k=1}^{d/2}\Big(cos\frac{2\pi kj}{N} - 2cos\frac{2\pi k(j+1)}{N} + cos\frac{2\pi k(j+2)}{N}\Big)^2 + \\ &\quad \Big(sen\frac{2\pi kj}{N} - 2sen\frac{2\pi k(j+1)}{N} + sen\frac{2\pi k(j+2)}{N}\Big)^2\Big]^{1/2} \\ &\leq \sum_{j=0}^{N-3}\Big[8\sum_{k=1}^{d/2}\Big(\frac{2\pi k}{N}\Big)^4\Big]^{1/2} \leq \frac{4\pi^2}{N}\sqrt{8}\Big[\sum_{k=1}^{d/2}k^4\Big]^{1/2} \\ &\leq \frac{4\pi^2}{N}\sqrt{8}\Big[\sum_{k=1}^{d/2}\Big(\frac{d}{2}\Big)^4\Big]^{1/2} = \frac{2\pi^2 d^{\frac{5}{2}}}{N}.\end{aligned}$$

Por lo que

$$\sigma_2(H_N^d, p) \leq \frac{2^{\frac{3}{2}}\pi^2 d^2}{N}.$$

Usando el corolario 3.4.4 podemos afirmar que si α, γ, C cumplen (A) y d es par, entonces $\|x - \tilde{x}\|$ va a tener una acotación del tipo $O(N^{-2})$ si N es par, y se va a comportar asintóticamente como $\frac{d\delta}{N}$ si N es impar.

Ejemplo 3.4.4. ***Raíces de la unidad***

Dado que el marco $R_N = \{e_n\}_{n=1}^N$ corresponde a un reordenamiento de H_N^2, se puede probar que si $p = id$,

$$\sigma_2(R_N, p) \leq \frac{8\sqrt{2}\pi^2}{N} \ y \ \|\Delta e_n\| \leq \frac{2\pi}{N}.$$

Como estos marcos cumplen la condición de suma cero, entonces si α, γ, C verifican (A), valen las acotaciones obtenidas en el corolario 3.4.4.

Bibliografía

[BPY] J.Benedetto, A.Powell, Ö.Yilmaz, *Second-order Sigma-Delta* ($\Sigma\Delta$) *quantization of finite frame expansions*, Applied and Computational Harmonic Analysis, **20**, (2006), 126-148.

[BPY06] J.Benedetto, A.Powell, Ö.Yilmaz, *Sigma-Delta* ($\Sigma\Delta$) *Quantization and Finite Frames*, IEEE Transactions on Information Theory, **52 (5)**, (2006), 1990-2005.

[CD] A.R.Calderbank, I. Daubechies, *The pros and cons of democracy*, IEEE Transactions on Information Theory, **48 (6)**, (2002), 1721-1725.

[Ch] C.K.Chui, *An introduction to wavelets*, serie Wavelet analysis and its applications ; v.1, Academic Press, San Diego, 1992.

[Da] I.Daubechies, *Ten lectures on wavelets*, SIAM, Philadelphia, 1992.

[DDV] I.Daubechies, R. DeVore, *Approximating a bandlimited function using very coarsely quantized data: A family of stable sigma-delta modulators of arbritary order*, Annals of Mathematics, **158 (2)**, (2003), 679-710.

[DDVGV] I.Daubechies, R. DeVore, C.S.Gunturk, V.Vaishampayan *A/D Conversion with imperfect quantizers*, IEEE Transactions on Information Theory, **52 (3)**, (2006), 874-885.

[GCW] R.M.Gray, W.Chou, P.W.Wong, *Quantization noise in single-loop sigma-delta modulation with sinusoidal inputs*, IEEE Transactions on Communication, **37 (9)**, (1989), 956-968.

[GK] V.Goyal, J.Kovačević, *Quantized Frame Expansions with Erasures*, Applied and Computational Harmonic Analysis, **10**, (1998), 203-233.

[Gr] R.M.Gray, *Quantization noise espectra*, IEEE Transactions on Information Theory, **36 (6)**, (1990), 1220-1244.

[IY] H.Inose, Y.Yasuda, *A unity bit coding method by negative feedback*, Proceedings of the IEEE, **51 (11)**, (1963), 1524-1535.

[Th] N.T.Thao, *MSE behavior and centroid function of mth order asymptotic* $\Sigma\Delta$ *modulators*, IEEE Transactions on Information Theory, **50 (5)**, (2004), 839-860.

[Yi] Ö.Yilmaz, *Mathematical Properties of Coarse Quantization Schemes in Signal Analysis with New Applications*, Ph.D Thesis, PACM, Princeton University, 2002.