Scalar Quantization with Noisy Partitions and its Application to Flash ADC Design

Da Wang, Yury Polyanskiy and Gregory Wornell
EECS Dept., MIT
Cambridge, MA, USA
{dawang, yp, gww}@mit.edu

***Abstract*—Motivated by recent circuit designs for Flash ADCs with imperfect comparators, we investigate the problem of scalar quantization with noisy partition points, where the partition point locations are perturbed from the designated values by noise during the placement process. For this problem setting, we derive a high resolution approximation for mean square error, and analyze the optimal partition point density accordingly. Our results indicate that it is necessary to take the effect of noise into account in the design process. In particular, we derive the optimal partition point density when the input distribution is Gaussian or uniform, and show when noise variance exceeds a certain threshold, a peculiar phase transition occurs and the optimal point density degenerates into a delta function at the origin. These theoretical results allow to optimize the design of flash ADCs and gain 1 bit in resolution over existing designs.**

I. Introduction

In this paper we investigate the problem of scalar quantization with noisy partition points, where partition points are perturbed from the designated values by noise during the placement process. This problem is motivated by Analog-to-Digital Converter (ADC) design with imperfect comparators, where the comparator reference voltages are subject to stochastic manufacture variations, which becomes increasing salient as modern CMOS design approaches the physical limits of scaling. A series of work in circuit systems [1]–[3] have employed redundancy and/or reconfigurability to tackle this issue, in the context of Flash ADC design. On the other hand,

While there exists some theoretical investigation of imperfect scalar quantizer and ADC (e.g., see [4] and the references therein), they treat the ADC design as given and aim to improve the quantization (estimation) performance via post processing. When the fabrication variation is low, these types of techniques are useful in improving system performance. However, when the fabrication variation is high enough, it is simply impossible for the traditional designs to meet the performance specifications, and in this case, new ADC designs are called for, and our investigation aims to provide perspectives on these new designs, by taking the statistical property of fabrication variation into account.

Analogous to the development in classical scalar quantization theory, we adopt high resolution analysis for this new setting, and analyze the optimal partition point density. Finally, we discuss the implications of our results for Flash ADC design, in terms of both technology scaling and specific partition point density design.

II. Notation

We use lower-case letters (e.g. x) to denote a particular value of the corresponding random variable denoted in upper-case letters (e.g. X). We denote the support of a probability density function (p.d.f.) f_X by $\mathrm{Supp}(f_X) \triangleq \{x \in \mathbb{R} : f_X(x) > 0\}$. We use $x_i^j, j > i$ to denote a sequence of values $x_i, x_{i+1}, \ldots, x_j$, and x^n as a shorthand for x_1^n. We use the bold font to represent a vector, i.e., $\mathbf{x} \triangleq [x_1, x_2, \ldots, x_n]$. We denote the indicator function by $\mathbb{1}\{A\}$, where

$$\mathbb{1}\{A\} = \begin{cases} 1 & \text{clause } A \text{ is true,} \\ 0 & \text{otherwise.} \end{cases}$$

We let $\mathbb{R}$ denote the real line, and $\langle f, g \rangle$ denote the inner product of two functions on $\mathbb{R}$, i.e., for $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$, $\langle f, g \rangle \triangleq \int f(x) g(x) dx$.

Given a sequence c^n, we denote the number of points in c^n that falls in an interval $[a, b]$ by $N(a, b; c^n)$, i.e.,

$$N(a, b; c^n) \triangleq \sum_{i=1}^{n} \mathbb{1}\{a \le c_i \le b\}. \qquad (1)$$

We say $a_n \simeq b_n$ if $a_n = b_n(1 + \varepsilon_n)$ for some $\varepsilon_n \to 0$ as $n \to \infty$.

III. Background

In this section we describe the problem of Flash ADC design with imperfect comparators, whose abstraction leads to the problem formulation in Section IV.

Flash ADC is a popular high-speed ADC architecture, with a comparator bank being its key building block. As shown in Section III, an imperfect comparator bank consists of n imperfect comparators, where the block diagram of each comparator is shown in Section III. As the diagram indicates, an imperfect comparator has non-idealities on both its input voltage and reference voltage, and the input-output relationship of the comparator satisfies $Y_{\text{out}} = \mathbb{1}\{V_{\text{in}} + Z_{\text{in}} \ge V_{\text{ref}} + Z_{\text{ref}}\}$. Let $Z = Z_{\text{in}} - Z_{\text{ref}}$, then the output satisfies $Y_{\text{out}} = \mathbb{1}\{V_{\text{in}} + Z \ge V_{\text{ref}}\}$, where $Z \sim \mathsf{N}(0, \sigma^2)$ is the effective fabrication noise and σ^2 is the effective fabrication variance, as both Z_{in} and Z_{ref} can be modeled as independent zero-mean Gaussian random variables [5], [6]. We emphasize that the fabrication variations are *static* in the sense that they are determined at the time of fabrication and does not change during the quantization process.

The above model leads to the ADC block diagram in Fig. 2, where the fabricated reference voltages and design reference voltages are related by $\tilde{V}_i = v_i + Z_i, i = 1, 2, \ldots, n$ and the noisy comparator outputs satisfy $Y_i = \mathbb{1}\left\{X \ge \tilde{V}_i\right\}, i = 1, 2, \ldots, n$, where $Z_i \overset{i.i.d.}{\sim} \mathsf{N}(0, \sigma^2)$. In

This work was supported, in part, by AFOSR under Grant No. FA9550-11-1-0183, and by NSF under Grant No. CCF-1017772.

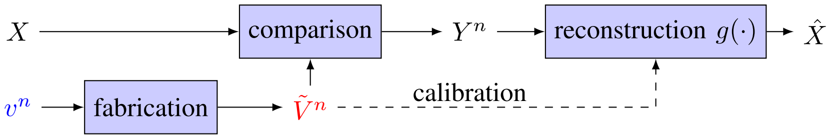
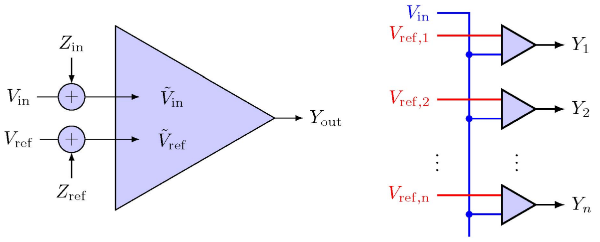


Fig. 2: System block diagram of a scalar quantizer with imperfect comparators. X is the input signal, v^n are the designed reference voltages and the $\tilde{V}^n$ are the fabricated reference voltages, which is a noisy version of v^n. A comparison of X and $\tilde{V}^n$ leads to the comparator outputs Y^n. The fabricated reference voltages are provided to the decoder via a calibration process. The encoder $g(\cdot,\cdot)$ takes both Y^n and $\tilde{V}^n$ to reproduce $\hat{X}$, an estimate of X.



(a) Imperfect comparator (b) Comparator bank

Fig. 1: Building blocks in a Flash ADC

this paper we study the case that the reconstructor $g(\cdot)$ has accurate knowledge about the fabricated reference voltages $\tilde{V}^n$, as calibration can be done with high accuracy with the help of extra calibration logic, as shown in [2].

IV. Problem formulation

In this section we abstract the problem of designing flash ADC with imperfect comparators, as described in Section III, as the problem of scalar quantization with noisy partition points. Before proposing the mathematical model and corresponding performance metrics in Section IV-B, we first review the classical scalar quantization problem in Section IV-A as our development can be seen as a generalization.

A. Classical scalar quantization problem

In the classical scalar quantization problem, an m-point scalar quantizer Q_m is a mapping $Q_m : \mathbb{R} \to \mathcal{C}$, where $\mathcal{C} = \{c_1, \dots, c_{m-1}, c_m\} \subset \mathbb{R}$ is the set of *reproduction values.* A quantizer $Q_m(x; v^n, \mathcal{C})$ is uniquely determined by its reproduction values $\mathcal{C}$ and its *partition points* v^n, where an input x is mapped to a value $c_j \in \mathcal{C}$ based on the quantization cell $(v_{i-1}, v_i]$ that x falls into. Given an input X with p.d.f. f_X, the MSE of the quantizer is

$$\begin{aligned} D(v^n, \mathcal{C}) &\triangleq \mathbb{E}_X\left[d(X; v^n, \mathcal{C})\right] \\ &\triangleq \sum_{i=1}^{n+1} \int_{v_{i-1}}^{v_i} f_X(x)(x - c_i)^2\, dx, \end{aligned}$$

where $d(\cdot;\cdot,\cdot)$ is the square error function $d(x; v^n, \mathcal{C}) \triangleq (x - Q_m(x; v^n, \mathcal{C}))^2$. Scalar quantization theory indicates that the optimal $\mathcal{C}$ in the MSE sense satisfies the *centroid condition* [7], i.e., $c_i = \mathbb{E}\left[X \middle| X \in (v_{i-1}, v_i]\right]$. Therefore, when discussing the MSE performance, we restrict our attention to the centroid reconstruction, and a scalar quantizer is uniquely determined by its partition points v^n and we denote the corresponding MSE by $D(v^n)$.

B. Scalar quantization with noisy partition points

In this section we introduce the problem of scalar quantization when the partition points are subject to random variations. More specifically, each partition point $\tilde{v}_i$ in an m-point scalar quantization is drawn independently from a distribution $F_{\tilde{V}_i}$, and we denote the quantizer as $Q_m(\cdot; \tilde{v}^n, \mathcal{C})$.

In this setting, a quantizer is randomly generated and all performance metrics become random variables. We take expectation over the random partition points $\tilde{V}^n$ when calculating the MSE, which leads to

$$\text{MSE} \triangleq \mathbb{E}_{\tilde{V}^n}\left[D\left(\tilde{V}^n\right)\right]. \tag{2}$$

Remark 1. *In (2) we average over both different realizations of quantizers and multiple uses of the same quantizer.*

For the problem of scalar quantization with noisy partition points, we investigate how the set of distributions $\left\{F_{\tilde{V}_i}\right\}$ impacts the MSE defined in (2). To investigate this, we develop a high resolution analysis that is analogous to the high resolution analysis of the classical scalar quantization problem in Section V.

More specifically, for the ADC with imperfect comparators model in Section III, if we design the reference voltages to be v^n, then $\tilde{V}_i \overset{\text{indep.}}{\sim} \mathsf{N}\left(v_i, \sigma^2\right), 1 \le i \le n$. Given σ^2, we utilize the theory in Section V-A to investigate how the choice of v^n impacts performance metrics, and present the detailed investigation in Section V-B.

Remark 2. *The problem of quantization with random uniformly distributed partition points for the uniform input distribution has been investigated in [8], under a different motivation, and turns out to be a useful building block in our analysis.*

V. High resolution analysis

For the classical quantization problem in Section IV-A, high resolution analysis leads to mathematical tractable performance results and yields useful approximate results for quantizer design. In this section we develop the analogous version of high resolution analysis for the problem in Section IV. We postpone most proofs and derivations to Section VII.

One key idea in high resolution approximation is that for a sequence of values that are dense enough, we can approximate it by a point density function.

Definition 1 (Point density function)**.** *A sequence of values v^n is said to have* point density function $\lambda(x)$ *if for a small enough interval dx,*

$$\lambda(x)\, dx = \lim_{n\to\infty} \frac{N(x, x + dx; v^n)}{n}, \quad x \in \mathbb{R}. \tag{3}$$

A. High resolution approximation of MSE

In this section we develop an analogous result to the high-resolution approximation of MSE for non-uniform quantization, as Bennett [9] did for classical quantization theory.

High resolution approximation: given input X with p.d.f. f_X and n independent random variables W^n, each with p.d.f. f_{W_i}, if the set of densities $\{f_{W_i}, 1 \le i \le n\}$ are all smooth and there exists $f_{\bar{W}}(\cdot)$ satisfies

$$f_{\bar{W}}(x) = \lim_{\delta\to 0}\lim_{n\to\infty}\frac{1}{n\delta}\mathbb{E}_{W^n}\left[N(x, x+\delta; W^n)\right], \tag{4}$$

then in the high resolution regime,

$$\mathbb{E}_{W^n}\left[D\left(X, W^n\right)\right] \simeq \frac{1}{2n^2}\int f_X(x) f_{\bar{W}}^{-2}(x)\,dx, \tag{5}$$

provided that the limit in (4) exists and the integral in (5) is finite (in particular, Supp$(f_{\bar{W}}) \supset$ Supp(f_X)).

The related derivations are presented in Section VII-A.

Remark 3. *When $W_i \overset{i.i.d.}{\sim} f_W$, $f_{\bar{W}} = f_W$ in (4).*

Remark 4. *For partition points v^n that have point density function $f_{\bar{W}}(\cdot)$, Bennett [9] shows the high-resolution approximation of MSE satisfies*

$$\text{MSE} \simeq \frac{1}{12n^2}\int f_X(x) f_{\bar{W}}^{-2}(x)\,dx, \tag{6}$$

which is exactly $1/6$ of (5). Therefore, in the high resolution regime, the random variation in placing partition points always lead to a 6-fold increase in MSE!

B. Application to Flash ADC design

We specialize (4) to the problem in Section IV and obtain (7).

Let $\phi(\cdot)$ be the density for Gaussian distribution $\mathsf{N}\left(0, \sigma^2\right)$, where $\sigma > 0$, and let the point density function of v^n be $\tau(x)$, then

$$\mathbb{E}_{\tilde{V}^n}\left[D\left(X, \tilde{V}^n\right)\right] \simeq \frac{1}{2n^2}\int f_X(x)\lambda^{-2}(x)\,dx \tag{7}$$

where λ is the convolution of two densities τ and ϕ:

$$\lambda(x) = (\tau * \phi)(x).$$

This result follows immediately from (5) and Lemma 1.

Lemma 1.

$$\frac{1}{n}\mathbb{E}\left[N\left(x, x+dx; \tilde{V}^n\right)\right] = (\tau * \phi)(x)dx.$$

Remark 5. *Due to the smoothness and support conditions, the smaller σ, the larger n we need to achieve the high resolution approximation, as shown by the Monte-Carlo simulation results in Fig. 3.*

And it is not hard to see that for a fixed n, taking $\sigma \to 0$ leads to high rate approximation in (6) rather than (5).

C. Optimal partition point density analysis

As shown in Section V-A, the integral

$$R(\tau) = \int f_X(x)(\tau * \phi)^{-2}(x)\,dx \tag{8}$$

is the key quantity in MSE calculation, and in this section we characterize τ that minimizes $R(\tau)$ in a variety of scenarios of interest.

Theorem 2. *τ minimizes $R(\tau)$ if and only if*

$$\sup_{x\in\mathcal{A}}\left[\frac{f_X}{(\tau*\phi)^3} * \phi\right](x) \le \left\langle f_X, \frac{1}{(\tau*\phi)^2}\right\rangle. \tag{9}$$

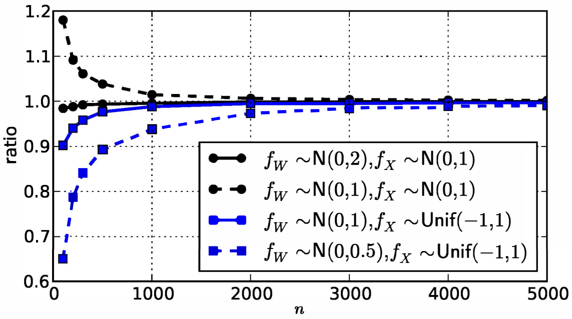


Fig. 3: The ratio of $\mathbb{E}_{W^n}\left[D(X, W^n)\right]$ obtained from Monte-Carlo simulation and numerical calculation of the integral in (5) when f_X is uniform or Gaussian.

In particular, if there exists τ such that

$$\tau * \phi \propto f_X^{1/3}, \tag{10}$$

then τ^* minimizes $R(\tau)$ and

$$R(\tau^*) = \left(\int f_X^{1/3}(x)dx\right)^3. \tag{11}$$

Based on Theorem 2, we can derive the optimal τ when the input distribution is Gaussian or uniformly distributed.

Theorem 3 (Gaussian input distribution)**.** *When $X \sim \mathsf{N}(0, \sigma_X^2)$,*

$$\tau^* \sim \begin{cases}\mathsf{N}(0, 3\sigma_X^2 - \sigma^2) & \text{when } 3\sigma_X^2 > \sigma^2\\ \delta(x) & \text{when } 3\sigma_X^2 \le \sigma^2\end{cases},$$

and

$$R(\tau^*) = \begin{cases}6\sqrt{3}\pi\sigma_X^2 & \text{when } 3\sigma_X^2 > \sigma^2\\ 2\pi\sigma^3/\sqrt{\sigma^2 - 2\sigma_X^2} & \text{when } 3\sigma_X^2 \le \sigma^2\end{cases}.$$

Theorem 4 (Uniform input distribution)**.** *When $X \sim \mathsf{Unif}([-1, 1])$ and $\sigma \ge \sigma_0 \approx 0.7228$, $\tau^*(x) = \delta(x)$ and*

$$R(\tau^*) = 2\pi\sigma^2\int_0^1 \exp\left(-\frac{x^2}{2\sigma^2}\right)dx.$$

Remark 6. *For both Gaussian and uniform input distributions, when σ large enough, $\tau^*(x) = \delta(x)$. In this case, simply aiming to place all partition points at $x = 0$ and letting the noisy placement process spread them out naturally is optimal, which is somewhat surprising.*

When the input is uniform and $\sigma < \sigma_0$, we obtain τ^* numerically. In particular, we approximate τ^* by a discrete distribution $\hat{\tau}$, i.e., $\hat{\tau}(x; \mathbf{p}, \mathbf{a}) = \sum_{i=0}^k p_i(\delta(x - a_i) + \delta(x + a_i))$, where $a_i \ge 0$ and the symmetry of $\hat{\tau}^*$ follows from the symmetry of f_X. Without loss of generality, we assume $a_0 = 0$.

We develop the following iterative optimization procedure to find the best $\mathbf{p}$ and $\mathbf{a}$, with some examples of the $\hat{\tau}$ in Fig. 4.

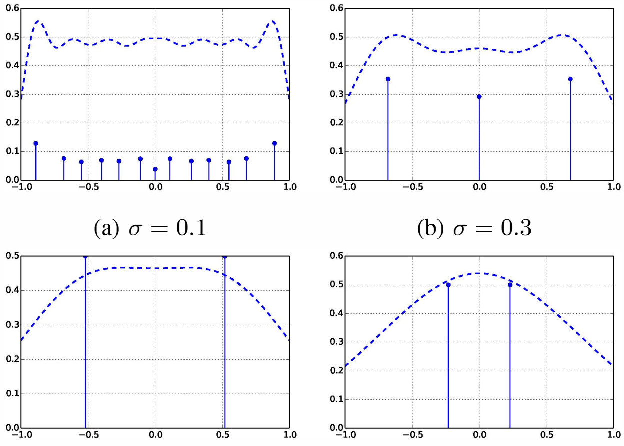
Remark 7. *Since the optimization problem is non-convex, Algorithm 1 only guarantees that it converges to local optimum. We use multiple randomly perturbed initial solutions to increase the probability of reaching global optimum.*

VI. Flash ADC design implications

We discuss two implications of our results on Flash ADC design with imperfect comparators.

Algorithm 1 Iterative optimization for $\hat{\tau}$.

```
$p_i^{(1)} = 1/(2k+1)$ for $0 \le i \le k$
$a_i^{(1)} = i/(k-1)$ for $1 \le i \le k$
$E_0 = 0$, $E_1 = R\left(\hat{\tau}\left(\cdot; \mathbf{p}^{(1)}, \mathbf{a}^{(1)}\right)\right)$, $t = 1$
while $|E_t - E_{t-1}| \ge \varepsilon$ do
    $\mathbf{p}^{(t+1)} = \arg\min_{\mathbf{p}} \hat{\tau}\left(x; \mathbf{p}, \mathbf{a}^{(t)}\right)$
    $\mathbf{a}^{(t+1)} = \arg\min_{\mathbf{a}} \hat{\tau}\left(x; \mathbf{p}^{(t+1)}, \mathbf{a}\right)$
    $E_{t+1} = R\left(\hat{\tau}\left(\cdot; \mathbf{p}^{(t+1)}, \mathbf{a}^{(t+1)}\right)\right)$
    $t = t + 1$
end while
```



(a) $\sigma = 0.1$

(b) $\sigma = 0.3$

(c) $\sigma = 0.5$

(d) $\sigma = 0.7$

Fig. 4: Algorithm 1 output for uniform input distribution over $[-1, 1]$ with $k = 7$ for all σ values. The stems indicate $\hat{\tau}(x)$ and the dashed curves indicate $(\hat{\tau} * \phi)(x)$.

A. Technology scaling

Section IV-A shows that in classical quantization, MSE D scales with the number of quantization points n as $1/n^2$.

However, in imperfect comparator fabrication, σ increases as the component size shrinks, with the relationship [5], [6] $\sigma^2 \propto 1/\text{Area}$. Taking only the component area into account, i.e., ignoring the wiring overhead etc.., $n \propto 1/\text{Area} \propto \sigma^2$. Therefore, for Gaussian input distribution, when $\sigma^2 \ge 3\sigma_X^2$, $D \propto \sigma^2/n^2 = \Theta(1/n)$. For uniform input distribution, when $\sigma \ge \sigma_0$, $D \propto \sigma^2/n^2 = \Theta(1/n)$. In conclusion, building more imperfect comparators is beneficial for reducing MSE. While in classical setting MSE $\propto 1/n^2$, with noisy fabrication, MSE $\propto 1/n$ when σ is large enough.

B. Comparison with stochastic ADC

In circuit system research, [3] presents a design that explores the idea of high resolution quantization. Assuming uniform input over $[-\sigma, \sigma]$, their design corresponds to n in the range of 1000 to 2000, and $\tau(x) = \delta(x - 1.078\sigma)/2 + \delta(x + 1.078\sigma)/2$, with the rationale of making the resulting density $\lambda = \tau * \phi$ as uniform as possible in the signal range $[-\sigma, \sigma]$. However, as we showed in Theorem 4, the optimal MSE solution is $\tau^*(x) = \delta(x)$. As Fig. 5 shows, assuming $\sigma = 1$, while $\lambda_{\text{stochastic}}$ is approximately flat in the input range $[-1, 1]$, many partition points are wasted as they are out of the input range. Calculation shows $\text{MSE}_{\text{stochastic}}/\text{MSE}^* \approx 2.15$, which corresponds to slightly more than 1 effective number of bit (ENOB) difference. This is significant for the design in [3] with ENOB in the range of 5 to 6 bits.

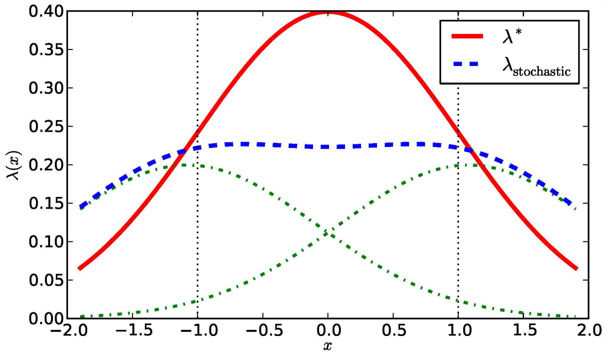


Fig. 5: Comparison of the optimal λ^* with the stochastic ADC density $\lambda_{\text{stochastic}}$. The two dash-dotted lines show the noisy partition point densities corresponding to $\delta(x - 1.078)/2$ and $\delta(x + 1.078)/2$, which are $\{\phi(x \pm 1.078)/2\}$ and sum to $\lambda_{\text{stochastic}}$.

VII. Derivations for high resolution analysis

A. High resolution analysis of MSE

In this section we first show a result for the MSE for a quantizer with random uniformly distributed partition points Lemma 5, and its extension in Lemma 6, then proceed to show the high resolution approximation result in (5) by showing the increase in MSE (comparing to (6)) is due to the random interval sizes resulting from the random partitioning, rather than the random number of partition points in an interval.

Lemma 5 (Theorem 1 in [8]). *Given $X \sim$ Unif$([0, \Delta])$ and $W_i \overset{i.i.d.}{\sim}$ Unif$([0, \Delta])$, $1 \le i \le n$, then*

$$\mathbb{E}_{X, W^n}\left[d\left(X, W^n\right)\right] = \frac{\Delta^2}{2(n+2)(n+3)}$$

Proof: See the proof in [8]. ■

Lemma 6. *Given $X \sim$ Unif$([0, \Delta])$, and for $1 \le i \le n$,*

$$\begin{aligned} &W_i \sim \textit{Unif}([0, \Delta]) && \textit{w.p. } p_i \\ &W_i \notin [0, \Delta] && \textit{w.p. } 1 - p_i, \end{aligned}$$

and let $k_n = \sum_{i=1}^n p_i$, then if for some $\varepsilon > 0$, $\lim_{n\to\infty} k_n / \left(n^{1/2+\varepsilon}\right) = c > 0$, then

$$\lim_{n\to\infty} k_n^2\, \mathbb{E}_{X, W^n}\left[d(X, W^n)\right] = \frac{\Delta^2}{2}.$$

Proof sketch: Define $U_i \triangleq \mathbb{1}\left\{W_i \sim \text{Unif}\left([0, \Delta]\right)\right\}$, then $U_i \sim \text{Bern}\left(p_i\right)$. Let $K \triangleq \sum_{i=1}^n U_i$, Lemma 5 indicates that

$$\mathbb{E}_{X, W^n}\left[d\left(X, W^n\right) \middle| K = k\right] = \frac{\Delta^2}{2(k+2)(k+3)}.$$

Noting K is the sum of n independent Bernoulli random variables, by Hoeffding's inequality,

$$\mathbb{P}\left[\left|K - \mathbb{E}\left[K\right]\right| > t\right] \le 2\exp\left(-\frac{2t^2}{n}\right).$$

Let $t_n = n^{1/2+\varepsilon/2}$, then $\mathbb{P}\left[\left|K - \mathbb{E}\left[K\right]\right| > t_n\right] \le 2\exp\left(-2n^{\varepsilon}\right)$. Then let $\mathcal{K} \triangleq \{k : k_n - t_n \le k \le k_n + t_n\}$, we can obtain $\lim_{n\to\infty} k_n^2\, \mathbb{E}_{X, W^n}\left[d\left(X, W^n\right)\right] \le \Delta^2/2$. by calculating $\mathbb{E}_{X, W^n}\left[d\left(X, W^n\right) \middle| K = k\right]$ for the case $k \in \mathcal{K}$ and $k \notin \mathcal{K}$. Similarly, we can show $\lim_{n\to\infty} k_n^2\, \mathbb{E}_{X, W^n}\left[d\left(X, W^n\right)\right] \ge \Delta^2/2$ and complete the proof. ■

Derivations for (5): We partition $\mathrm{Supp}(f_X)$ by m points $x_1, x_2, \ldots, x_m$ and let x_0 and x_{m+1} be the two ends points of $\mathrm{Supp}(f_X)$, which could be $-\infty$ and $+\infty$ when $\mathrm{Supp}(f_X)$ is unbounded. We assume m is large enough such that 1) each interval $\mathcal{R}_j \triangleq (x_{j-1}, x_j], 1 \leq j \leq m+1$ is small enough so that the densities $(f_X, \phi,\ f_{W_i})$ can be seen as constant over $\mathcal{R}_j$; 2) the expected number of partition points that fall into each region $\mathcal{R}_j$ satisfies $\mathbb{E}_{W^n}\left[N\left(x_{j-1}, x_j; W^n\right)\right] = \Omega\left(n^{1/2}\right)$. Then

$$\mathbb{E}_{W^n}\left[D\left(W^n\right)\right] = \sum_{j=1}^{m+1} \mathbb{E}\left[d\left(X, W^n\right) \middle| X \in \mathcal{R}_j\right] \mathbb{P}\left[X \in \mathcal{R}_j\right].$$

For each interval $\mathcal{R}_j, 1 \leq j \leq m+1$, based on the first assumption above, $\mathbb{P}\left[W_i \in \mathcal{R}_j\right] = f_{W_i}\left(x_j\right)\left|\mathcal{R}_j\right|$, and the conditional density given that $W_i \in \mathcal{R}_j$ is uniform over $\mathcal{R}_j$. Therefore, let $p_{ij} = f_{W_i}\left(x_j\right)\left|\mathcal{R}_j\right|$, then

$$\begin{aligned} W_i &\sim \mathsf{Unif}\left(\left[x_{j-1}, x_j\right]\right) && w.p.\ p_{ij} \\ W_i &\notin \left[x_{j-1}, x_j\right] && w.p.\ 1 - p_{ij}, \end{aligned}$$

and by the second assumption and Lemma 6, $\mathbb{E}_{X,W^n}\left[d\left(X, W^n\right) \middle| X \in \mathcal{R}_j\right] \simeq \left|\mathcal{R}_j\right|^2 / \left(2n_j^2\right)$, where $n_j \triangleq \sum_{i=1}^n p_{ij}$. By (4), $n_j = n f_{\bar{W}}\left(x_j\right)\left|\mathcal{R}_j\right|$. Therefore,

$$\begin{aligned} &\mathbb{E}_{X,W^n}\left[d\left(X, W^n\right)\right] \\ &= \sum_{j=1}^{m+1} \mathbb{E}_{X,W^n}\left[d\left(X, W^n\right) \middle| X \in \mathcal{R}_j\right] \mathbb{P}\left[X \in \mathcal{R}_j\right] \\ &\simeq \sum_{j=1}^{m+1} \frac{\left|\mathcal{R}_j\right|^3}{2\left(n f_{\bar{W}}\left(x_j\right)\right)^2} f_X\left(x_j\right)\left|\mathcal{R}_j\right| \simeq \frac{1}{2n^2} \int \frac{f_X(x)}{f_{\bar{W}}^2(x)}\, dx. \end{aligned}$$

■

B. Application to Flash ADC design

We derive the density function for the problem in Section IV in Lemma 1, leading to (7).

Proof for Lemma 1:

$$\begin{aligned} &\frac{1}{n}\mathbb{E}\left[N\left(x, x+dx; \tilde{V}^n\right)\right] \\ &= \int \phi(z) \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}\left[V_i \in [x-z, x-z+dx]\right] dz \\ &= \int \phi(z) \frac{1}{n} N\left(x-z, x-z+dx; V^n\right) dz \\ &= \int \phi(z) \tau(x-z) dx dz = (\tau * \phi)(x) dx. \end{aligned}$$

■

C. Optimal partition point density analysis

In this section we first prove the optimal conditions in Theorem 2. Following that we specialize Theorem 2 to $\tau^*(\cdot) = \delta(\cdot)$ in Lemma 7, and derive the corresponding conditions for Gaussian and uniform input distributions respectively in Lemmas 8 and 9.

Proof for Theorem 2: When the existence condition in (10) is satisfied, then (11) follows from the Panter and Dite formula [10].

In general, given the optimal τ^*, for any distribution h such that $\int h = 1$, $R((1-\epsilon)\tau^* + \epsilon h) \geq R(\tau^*)$. Therefore,

$$\lim_{\varepsilon \to 0} \frac{R((1-\epsilon)\tau^* + \epsilon h) - R(\tau^*)}{\varepsilon} \geq 0,$$

which leads to $\left\langle f_X, 1/(\tau^* * \phi)^2 \right\rangle \geq \left\langle h * \phi, f_X/(\tau^* * \phi)^3 \right\rangle = \left\langle h, f_X/(\tau^* * \phi)^3 * \phi \right\rangle$. Since the above holds for any h that satisfies Section VII-C, we have

$$\sup_x \left(\frac{f_X}{(\tau^* * \phi)^3} * \phi\right)(x) \leq \left\langle f_X, \frac{1}{(\tau^* * \phi)^2} \right\rangle. \qquad (12)$$

■

Lemma 7 (Condition for $\tau^*(x) = \delta(x)$)**.** *Define*

$$g(x) \triangleq \left(\frac{f_X}{\phi^3} * \phi\right)(x), \qquad (13)$$

then if for any $x \in \mathcal{A}$, $g'(x) \leq 0$, $\tau^*(x) = \delta(x)$.

Proof: Substitute $\tau^*(x) = \delta(x)$ in (12), we have

$$\sup_x \left(\left(f_X/\phi^3\right) * \phi\right)(x) \leq \left\langle f_X, 1/\phi^2 \right\rangle. \qquad (14)$$

Since f_X is symmetric and smooth on $\mathcal{A}$, $g(x)$ is an even function on $\mathcal{A}$ and is smooth, therefore, $g'(0) = 0$. Since $\sup_x g(x) \geq g(0) = \left\langle f_X, 1/\phi^2 \right\rangle$, we know if for any $x \in \mathcal{A}$, $g'(x) \leq 0$ then $x = 0$ maximizes $g(x)$, thus (14) is satisfied and hence $\delta(x)$ is indeed the optimal solution. ■

Below we show that for both Gaussian and uniform input distributions, $\tau^*(x) = \delta(x)$ when σ is large enough.

Lemma 8. *When $X \sim \mathsf{N}\left(0, \sigma_X^2\right)$, $\tau^*(x) = \delta(x)$ if and only if $\sigma^2 \geq 3\sigma_X^2$.*

Proof: When $\sigma^2 \geq 3\sigma_X^2$, straightforward algebra shows

$$g'(x) \propto \frac{x\sigma_X^2}{\sigma^2 - 2\sigma_X^2} - x = \frac{x\left(3\sigma_X^2 - \sigma^2\right)}{\sigma^2 - 2\sigma_X^2} \leq 0.$$

When $\sigma^2 < 3\sigma_X^2$, $\tau^*(x) \neq \delta(x)$ by (11) in Theorem 2. ■

Lemma 9. *When $X \sim \mathsf{Unif}([-1, 1])$ $\tau^*(x) = \delta(x)$ if and only if $\sigma \geq \sigma_0 \approx 0.7228$.*

Proof: For $\mathsf{Unif}\left([-1, 1]\right)$, algebra shows

$$g'(x) \propto \int_{-1}^{1} (t - x) \exp\left(\frac{t + x/2}{\sigma^2}\right)^2 dt.$$

Numerically solution indicates if $\sigma \geq \sigma_0 \approx 0.7228$, $g'(x) \leq 0$ for any x, and if $\sigma < \sigma_0$, $g''(0) > 0$, and (9) is violated when $\tau(x) = \delta(x)$. ■

ACKNOWLEDGEMENT

The authors thank Frank Yaul and Anantha Chandrakasan for helpful discussions.

References

[1] M. Flynn, C. Donovan, and L. Sattler, "Digital calibration incorporating redundancy of flash ADCs," *Circuits and Systems II: Analog and Digital Signal Processing, IEEE Transactions on*, vol. 50, no. 5, pp. 205–213, 2003.

[2] D. Daly and A. Chandrakasan, "A 6-bit, 0.2 v to 0.9 v highly digital flash ADC with comparator redundancy," *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 11, pp. 3030–3038, 2009.

[3] S. Weaver, B. Hershberg, P. Kurahashi, D. Knierim, and U. Moon, "Stochastic flash Analog-to-Digital conversion," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 57, no. 11, pp. 2825–2833, 2010.

[4] H. Lundin, "Characterization and correction of Analog-to-Digital converters," dissertation, KTH, 2005.

[5] P. Kinget, "Device mismatch and tradeoffs in the design of analog circuits," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 6, pp. 1212–1224, 2005.

[6] P. Nuzzo, F. D. Bernardinis, P. Terreni, and G. V. der Plas, "Noise analysis of regenerative comparators for reconfigurable ADC architectures," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 55, no. 6, pp. 1441–1454, 2008.

[7] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Boston: Kluwer Academic Publishers, 1992.

[8] V. Goyal, "Scalar quantization with random thresholds," *IEEE Signal Processing Letters*, vol. 18, no. 9, pp. 525–528, 2011.

[9] W. R. Bennett, "Spectra of quantized signals," *Bell System Technical Journal*, vol. 27, no. 3, pp. 446–472, 1948.

[10] P. F. Panter and W. Dite, "Quantization distortion in Pulse-Count modulation with nonuniform spacing of levels," *Proceedings of the IRE*, vol. 39, no. 1, pp. 44–48, 1951.