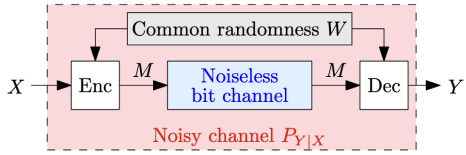
Subtractively Dithered Quantization and Universal Quantization

Nam Nguyen

School of Electrical Engineering and Computer Science
Oregon State University

September 18, 2025

Channel Simulation Review



Setup:

- Given a *noiseless channel*.
- Goal: simulate a noisy channel $P_{Y|X}$.
- Encoder observes input X and sends a *description* M noiselessly.
- Decoder produces output Y, s.t.

$$Y \sim P_{Y|X}(\cdot|X).$$

- Encoder/decoder may also share *common randomness* W.

Non-Dithered Quantization

Scheme:

$$Y = Q(X) = \Delta \left\lfloor \frac{X}{\Delta} + \frac{1}{2} \right\rfloor, \quad X \in \mathbb{R}$$

- Encoder sends $M = \lfloor X/\Delta + 1/2 \rfloor \in \mathbb{Z}$
  $\Rightarrow$ Using Huffman code to encode into a sequence of bits.
- Decoder reconstructs $Y = \Delta M$.
- Limitation: $P_{\tilde{Y}|X} \neq P_{Y|X}$.

Dithered Scalar Quantization

Introduce **randomness** with common dither $W \sim \mathrm{Unif}(-\frac{1}{2}, \frac{1}{2})$:

$$Y = \Delta\Big(\Big\lfloor \tfrac{X}{\Delta} + W + \tfrac{1}{2} \Big\rfloor - W\Big).$$

- Encoder: $K = \lfloor X/\Delta + W + 1/2 \rfloor$.
- Decoder: $Y = \Delta(K - W)$.
- Quantization error:

$$Y - X \sim \mathrm{Unif}\big(-\tfrac{\Delta}{2}, \tfrac{\Delta}{2}\big), \quad \textbf{independent of } X.$$

Example

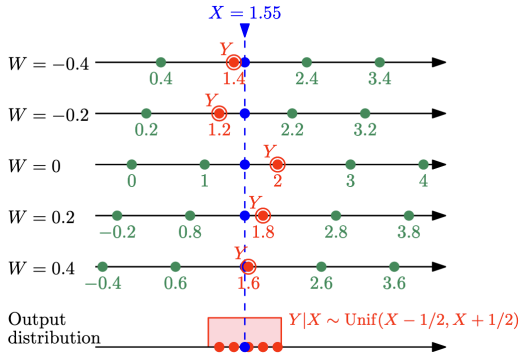


Figure: Subtractive dithering with $\Delta = 1$, $X = 1.55$. We first generate the dither signal $W \sim \mathrm{Unif}\left(-\frac{1}{2}, \frac{1}{2}\right)$, and then find the reconstruction level among $\{\ldots, -2 - W, -1 - W, -W, 1 - W, \ldots\}$ that is closest to X, and output it as Y.

Additive Noise Simulation View

- Goal: simulate
$$\mathbf{Y} = \mathbf{X} + \mathbf{Z}, \quad Z_i \sim \text{i.i.d. } \mathrm{Unif}\left(-\tfrac{\Delta}{2}, \tfrac{\Delta}{2}\right).$$
- Apply dithered quantization entrywise with common randomness
$$\mathbf{W} = (W_1, \ldots, W_n), \quad W_i \sim \text{i.i.d. } \mathrm{Unif}\left(-\tfrac{1}{2}, \tfrac{1}{2}\right).$$

Proposition

For any distribution P_X over $\mathbb{R}^n$,

$$H(\mathbf{Y} \mid \mathbf{W}) = I(\mathbf{X}; \mathbf{Y}) = h(\mathbf{Y}) - n \log_2 \Delta.$$

- Achieve the *minimum conditional entropy* for simulating the additive noise channel $P_{Y|X}$
- If encode $\mathbf{Y}$ using Huffman code, then the expected length is at most *1 bit* away from the minimal expected length L^*.

Universal Quantization Property

Theorem

With step size $\Delta = 2\sqrt{3D}$ (so $\mathbb{E}[\|\mathbf{X} - \mathbf{Y}\|^2]/n = D$), $D > 0$

$$H(\mathbf{Y} \mid \mathbf{W}) \;\le\; R(D) + \tfrac{n}{2}\log_2 \tfrac{\pi e}{3} \;\le\; R(D) + 0.755n.$$

where

$$R(D) := \inf_{P_{\hat{\mathbf{X}}|\mathbf{X}}:\, n^{-1}\mathbb{E}[\|\mathbf{X}-\hat{\mathbf{X}}\|^2] \le D} I(\mathbf{X}; \hat{\mathbf{X}})$$

- Dithered quantization is **universally near-optimal**.
- Gap to $R(D) \le 0.755$ bits/dimension.
- Works for *any* source distribution P_X.

Layered Randomized Quantization

- Add noise $\mathrm{Unif}(-\Delta/2, \Delta/2) \Rightarrow$ simulated noise can be **uniform** over *any interval*.
- **Idea:** extend this to simulate **any unimodal noise distribution** f [1].
- Approach: express f as a mixture of uniform distributions over intervals

$$f(x) = \int_0^\infty \mathrm{Vol}(L_s^+(f)) \cdot \mathrm{Unif}(x; L_s^+(f))\, ds,$$

where $L_s^+(f) := \{x \in \mathbb{R} : f(x) \geq s\}$ is the superlevel set of f;
$\mathrm{Vol}(L_s^+(f)) = \sup L_s^+(f) - \inf L_s^+(f)$. And

$$\mathrm{Unif}(x; L_s^+(f)) := \frac{\mathbf{1}\{x \in L_s^+(f)\}}{\mathrm{Vol}(L_s^+(f))} \text{ is pdf of } \mathrm{Unif}(L_s^+(f)).$$

$\Rightarrow$ simulate the noise distribution f, by randomly selecting s and applying **subtractive dithering** to produce a noise distribution $\mathrm{Unif}(L_s^+(f))$.

[1] A pdf $f : \mathbb{R} \to \mathbb{R}$ is unimodal if there exists $c \in \mathbb{R}$ such that $f(x)$ is nondecreasing over $x \in (-\infty, c]$, and nonincreasing over $x \in [c, \infty)$.

Scheme

1. Generate common randomness (S, W):

$$S \sim f_S(s) = \mathsf{Vol}(L_s^+(f)), \quad W \sim \mathsf{Unif}(-\tfrac{1}{2}, \tfrac{1}{2}).$$

2. Encoder (given X, S, W):

$$\Delta = f_S(S), \quad K = \left\lfloor \tfrac{X}{\Delta} + W + \tfrac{1}{2} \right\rfloor.$$

3. Decoder (given S, W, K):

$$B = \frac{\sup L_S^+(f) + \inf L_S^+(f)}{2}, \quad Y = \Delta \cdot (K - W) + B.$$

Key Properties:

- Conditional on $S = s$, the scheme reduces to **subtractive dithering**.
- Randomizing over S makes the overall noise law *exactly* f.
- Provides a channel simulation method for **any unimodal noise distribution**.

Asymptotic Optimality

- Not universally optimal, but becomes **asymptotically optimal in the high-SNR limit**.
- Setup: $X \sim \mathsf{Unif}(0,t)$, $Z \sim f$, $t \to \infty$.

$$I(X; X+Z) \;\approx\; \log_2 t + O(1).$$

- Hence, the optimal conditional entropy also grows like

$$H^*_{f,t} \;\approx\; \log_2 t + O(1).$$

Asymptotic Result

For unimodal f with finite mean,

$$H^*_{f,t} = \log_2 t - h_L(f) + O\left(\tfrac{1}{t}\right), \quad t \to \infty,$$

where

$$h_L(f) := \int_0^\infty \mathsf{Vol}(L_s^+(f)) \, \log_2 \mathsf{Vol}(L_s^+(f)) \, ds$$

is the *layered entropy* of f.

Nonasymptotic Bound

Nonasymptotic Result

For an unimodal f with finite mean, and any source distribution P_X,

$$H(Y \mid S, W) \;\leq\; I(X;Y) + h(f) - h_L(f),$$

where

$$h(f) = \text{differential entropy of } f, \quad h_L(f) = \text{layered entropy of } f.$$

- Expected description length is bounded by

$$I(X;Y) + h(f) - h_L(f) + 1.$$

Exact Fixed-Length Channel Simulation

- Description $M \in [N]$ has a **fixed size** N.
- Goal:

$$\begin{aligned} N^* &= \min\ N \\ &\text{s.t. } P_{Y|X} = P_{\hat{Y}|X}. \end{aligned}$$

- Only certain schemes (e.g., subtractive dithering, layered randomized quantizers) can be adapted to *exact fixed-length* form.
- Unlike variable-length, exact fixed-length simulation incurs an extra penalty.

Optimal Description Length

Optimal Description Size

For finite discrete $\mathcal{X}, \mathcal{Y}$ with common randomness, the exact fixed-length description size is

$$N^* = \min\left\{k : \mathbf{P}_{Y|X} \in \mathsf{conv}\big(\{\mathbf{Q}_{Y|X} : \|\mathbf{1}^T\mathbf{Q}_{Y|X}\|_0 \le k\}\big)\right\}.$$

- $\mathbf{P}_{Y|X}$: conditional probability matrices wit entries $(\mathbf{P}_{Y|X})_{x,y} = P_{Y|X}(y|x)$.
- $\mathbf{Q}_{Y|X}$: conditional probability matrices with at most k nonzero columns.
- $\mathsf{conv}(\cdot)$: convex hull of such matrices.

Penalty:

- Unlike variable-length (cost $\approx I(X;Y)$ bits), fixed-length may need much larger N^*.
- Hence N^* can be significantly larger than L^* (optimum variable length).
- The exact fixed-length channel simulation is *inefficient*.