EE269
Signal Processing and Quantization for Machine Learning

Dithering and Stochastic Rounding

Instructor : Mert Pilanci

Stanford University

Outline

- Why finite-bit quantization creates *structured artifacts*
- **Dither:** add noise then quantize (what "dithering" usually means)
- Fundamental analysis: *linearization*
- **Subtractive dither (variant):** exact uniform error
- **Stochastic rounding:** unbiased randomized rounding
- Quantization in GPUs
- Practical takeaways

Recall (Lecture 1): Bennett's theorem is asymptotic

- For a uniform quantizer with step Δ, the additive model assumes

$$X \mapsto Q(X) = X - \varepsilon, \qquad \varepsilon \sim \mathcal{U}(-\Delta/2, \Delta/2), \qquad \varepsilon \perp X.$$

- Bennett's theorem justifies this *as* $\Delta \to 0$ (many levels, smooth pdf, no overload).
- At finite bit-depth, $\varepsilon(x)$ is deterministic and correlated.

Finite rate artifacts: why randomness helps

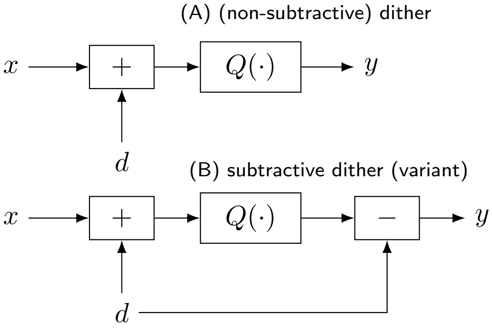
- Deterministic quantization maps smooth ramps to a **staircase** $\Rightarrow$ contouring/banding in images.
- Quantized sinusoids produce **spurious tones** (harmonics), not white noise.
- In feedback systems, quantization can create **limit cycles**.

Finite rate artifacts: why randomness helps

- Deterministic quantization maps smooth ramps to a **staircase** $\Rightarrow$ contouring/banding in images.
- Quantized sinusoids produce **spurious tones** (harmonics), not white noise.
- In feedback systems, quantization can create **limit cycles**.
- **Dithering idea:** add small noise before quantization to *randomize the staircase*.
  - trades structured distortion for “noise-like” distortion
  - often preferred perceptually (weak grain vs hard bands)

Outline

Additive dither: two canonical architectures



- In this lecture we start with (A): add noise then quantize.
- (B) is a useful *variant* when the same d is known at encoder/decoder.

Uniform rounding quantizer

We use a uniform rounding quantizer with step Δ:

$$Q_\Delta(z) \triangleq \Delta \cdot \mathrm{round}\left(\frac{z}{\Delta}\right).$$

Quantization error:

$$e(z) \triangleq Q_\Delta(z) - z \in \left[-\frac{\Delta}{2}, \frac{\Delta}{2}\right].$$

- Without randomization, $e(x)$ is a periodic sawtooth function of x.

Non-subtractive dither (definition)

Let d be a random dither, independent of the signal x.

Non-subtractive dithering:

$$\boxed{y = Q_\Delta(x + d)}$$

The reconstruction error is

$$\varepsilon \triangleq y - x = d + e(x + d).$$

- Benefit: breaks correlation between the signal and the quantization staircase.
- Cost: d appears directly in the output error.

Outline

Proof of linearization

Fix $x \in \mathbb{R}$ and write

$$x = k\Delta + r, \qquad r \in \left[-\frac{\Delta}{2}, \frac{\Delta}{2}\right).$$

With $d \sim \mathcal{U}[-\Delta/2, \Delta/2]$,

$$x + d \sim \mathcal{U}\left(k\Delta + r - \frac{\Delta}{2},\ k\Delta + r + \frac{\Delta}{2}\right),$$

an interval of length Δ, so

$$Q_\Delta(x + d) \in \{k\Delta, (k+1)\Delta\}.$$

Proof of linearization

Fix $x \in \mathbb{R}$ and write

$$x = k\Delta + r, \qquad r \in \left[-\frac{\Delta}{2}, \frac{\Delta}{2}\right).$$

With $d \sim \mathcal{U}[-\Delta/2, \Delta/2]$,

$$x + d \sim \mathcal{U}\left(k\Delta + r - \frac{\Delta}{2},\ k\Delta + r + \frac{\Delta}{2}\right),$$

an interval of length Δ, so

$$Q_\Delta(x + d) \in \{k\Delta, (k+1)\Delta\}.$$

Let

$$p = \mathbb{P}(Q_\Delta(x + d) = (k+1)\Delta \mid x) = \mathbb{P}(x + d > k\Delta + \Delta/2)$$

Since the interval is uniform and centered at x,

$$p = \frac{r}{\Delta}.$$

Proof of linearization

Fix $x \in \mathbb{R}$ and write

$$x = k\Delta + r, \qquad r \in \left[-\frac{\Delta}{2}, \frac{\Delta}{2}\right).$$

With $d \sim \mathcal{U}[-\Delta/2, \Delta/2]$,

$$x + d \sim \mathcal{U}\left(k\Delta + r - \frac{\Delta}{2},\ k\Delta + r + \frac{\Delta}{2}\right),$$

an interval of length Δ, so

$$Q_\Delta(x + d) \in \{k\Delta, (k+1)\Delta\}.$$

Let

$$p = \mathbb{P}(Q_\Delta(x + d) = (k+1)\Delta \mid x) = \mathbb{P}(x + d > k\Delta + \Delta/2)$$

Since the interval is uniform and centered at x,

$$p = \frac{r}{\Delta}.$$

Thus

$$\mathbb{E}[Q_\Delta(x+d) \mid x] = p(k+1)\Delta + (1-p)k\Delta = k\Delta + \Delta p = k\Delta + r = x.$$

Numerical example (linearization)

Quantize an 8-bit pixel using a 4-bit uniform quantizer.

- Dynamic range: 0 to 255.
- Step size: $\Delta = 16$ (levels at $\{\ldots, 96, 112, 128, \ldots\}$).
- Choose pixel value $x = 100$.

Write $x = 96 + 4$, so $k\Delta = 96$ and $r = 4$.

Let $d \sim \mathcal{U}(-8, 8)$. Then $x + d \sim \mathcal{U}(92, 108)$. The boundary between 96 and 112 is at 104.

$$\mathbb{P}(Q_\Delta(x+d) = 112) = \frac{r}{\Delta} = \frac{4}{16} = \frac{1}{4}$$

$$\mathbb{P}(Q_\Delta(x+d) = 96) = 1 - \frac{1}{4} = \frac{3}{4}.$$

Therefore, on average we get

$$\mathbb{E}[Q_\Delta(x+d) \mid x] = 112\,\frac{1}{4} + 96\,\frac{3}{4} = 100.$$

Dithering is stochastic rounding

Let Q_Δ be a uniform scalar quantizer and

$$d \sim \mathcal{U}\left[-\frac{\Delta}{2}, \frac{\Delta}{2}\right].$$

Assumption (no overload): $x \pm \frac{\Delta}{2}$ lies strictly inside the quantizer range.

For fixed x, $x + d$ is uniform over an interval of length Δ. Hence $Q_\Delta(x + d)$ can take only the two neighboring quantization levels.

Let $x \in [t, t + \Delta)$ and define

$$p \triangleq \frac{x - t}{\Delta} = \frac{\text{'distance from x to } t\text{'}}{\text{'distance from } t \text{ to } t + \Delta\text{'}}.$$

Then

$$Q_\Delta(x + d) = \begin{cases} t & \text{with probability } 1 - p, \\ t + \Delta & \text{with probability } p. \end{cases}$$

Dithering is stochastic rounding

Let Q_Δ be a uniform scalar quantizer and

$$d \sim \mathcal{U}\left[-\frac{\Delta}{2}, \frac{\Delta}{2}\right].$$

Assumption (no overload): $x \pm \frac{\Delta}{2}$ lies strictly inside the quantizer range.

For fixed x, $x + d$ is uniform over an interval of length Δ. Hence $Q_\Delta(x+d)$ can take only the two neighboring quantization levels.

Let $x \in [t, t+\Delta)$ and define

$$p \triangleq \frac{x-t}{\Delta} = \frac{\text{'distance from x to } t\text{'}}{\text{'distance from } t \text{ to } t+\Delta\text{'}}.$$

Then

$$Q_\Delta(x+d) = \begin{cases} t & \text{with probability } 1-p, \\ t+\Delta & \text{with probability } p. \end{cases}$$

Therefore, $\quad \mathbb{E}[Q_\Delta(x+d) \mid x] = (1-p)t + p(t+\Delta) = x.$

Stochastic rounding in NVFP4

- NVIDIA's NVFP4 data format was announced in March 2024 as a key feature of its Blackwell GPU architecture
- Its representable values are *not uniformly spaced*
- Each block is first scaled before quantization
- Popular in LLM inference and training

Stochastic rounding: Instead of always rounding to the nearest value, we randomly round up or down so that the average equals the original number. **How stochastic rounding works in NVFP4.**

- After scaling, a real number lies between two neighboring NVFP4 values
- Call them v_{low} and v_{high}
- We round to either one at random

$$\text{round to } v_{\text{high}} \text{ with probability } \frac{x - v_{\text{low}}}{v_{\text{high}} - v_{\text{low}}}, \qquad \text{otherwise round to } v_{\text{low}}.$$

Outline

How much dither should we add?

A standard choice is **one-LSB uniform dither**:

$$\boxed{d \sim \mathcal{U}\left(-\frac{\Delta}{2}, \frac{\Delta}{2}\right)}$$

- large enough to randomize the fractional part modulo Δ
- minimal support that yields the exact linearization result $\mathbb{E}[Q_\Delta(x+d) \mid x] = x$

How much dither should we add?

A standard choice is **one-LSB uniform dither**:

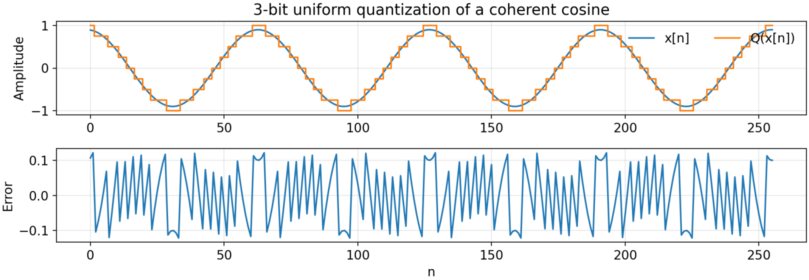
$$\boxed{d \sim \mathcal{U}\left(-\frac{\Delta}{2}, \frac{\Delta}{2}\right)}$$

- large enough to randomize the fractional part modulo Δ
- minimal support that yields the exact linearization result $\mathbb{E}[Q_\Delta(x+d) \mid x] = x$

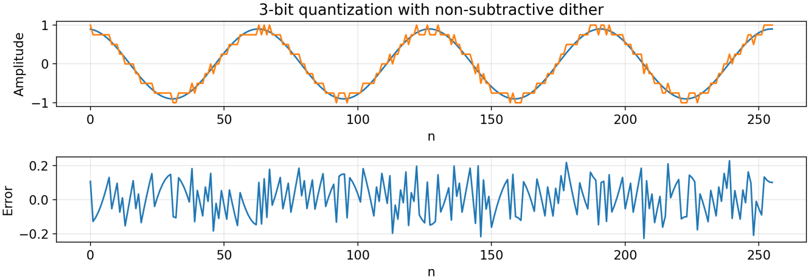
If the dither amplitude is much smaller than Δ, artifacts remain; if much larger, you add unnecessary noise.

- Examples
  - dithering in 1D signals
  - dithering in images
  - dithering in audio
  - spectrum (later)

Quantization of a pure sine wave

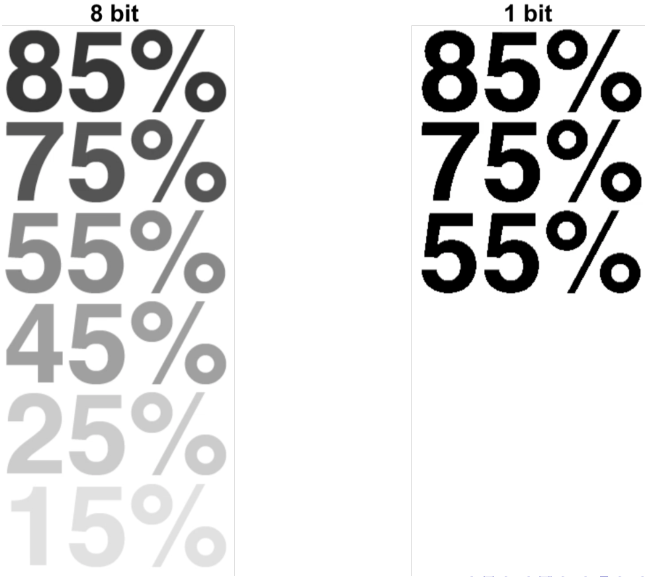


Dithered quantization of a pure sine wave

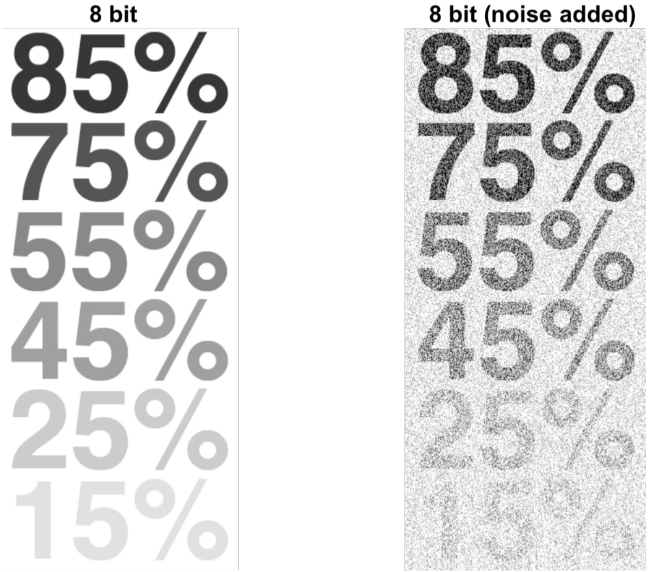


Uniform quantization of an image

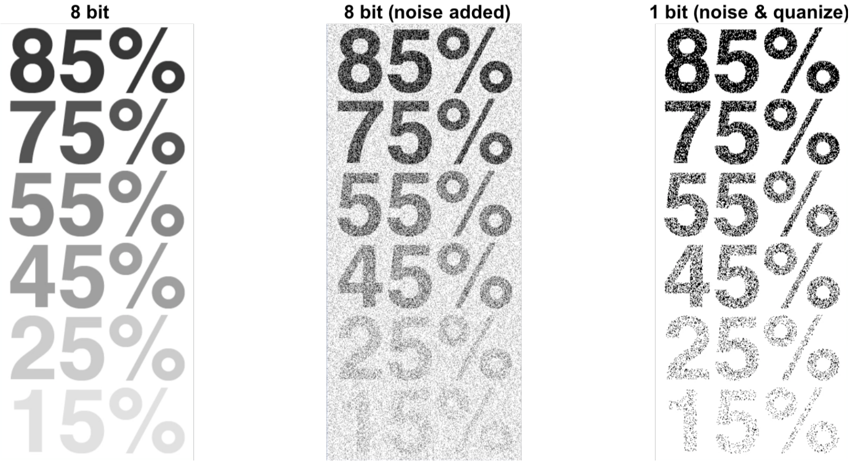
- values in $[0, 100]$ and threshold at 50



Adding noise

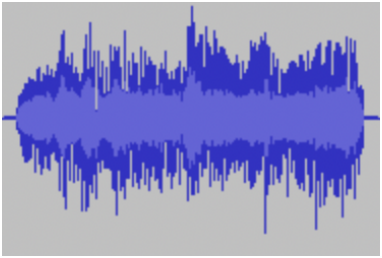


Dithering: Add noise and quantize

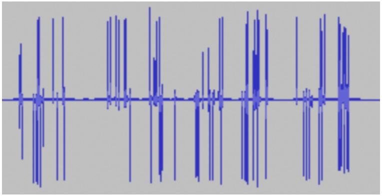


Dithering in music

- original audio



- 1-bit quantization



- 1-bit quantization with dithering

Outline

Subtractive dither: the key simplification

In subtractive dither,

$$y = Q_\Delta(x+d) - d \qquad \Rightarrow \qquad \varepsilon \triangleq y - x = Q_\Delta(x+d) - (x+d) = e(x+d).$$

- The reconstruction error equals the quantization error of the *dithered input*.
- If $(x+d) \bmod \Delta$ is uniform, then ε becomes uniform.

Theorem (uniform subtractive dither)

Claim. Let $d \sim \mathcal{U}(-\Delta/2, \Delta/2)$ be independent of x. In subtractive dithering:

$$y = Q_\Delta(x + d) - d$$

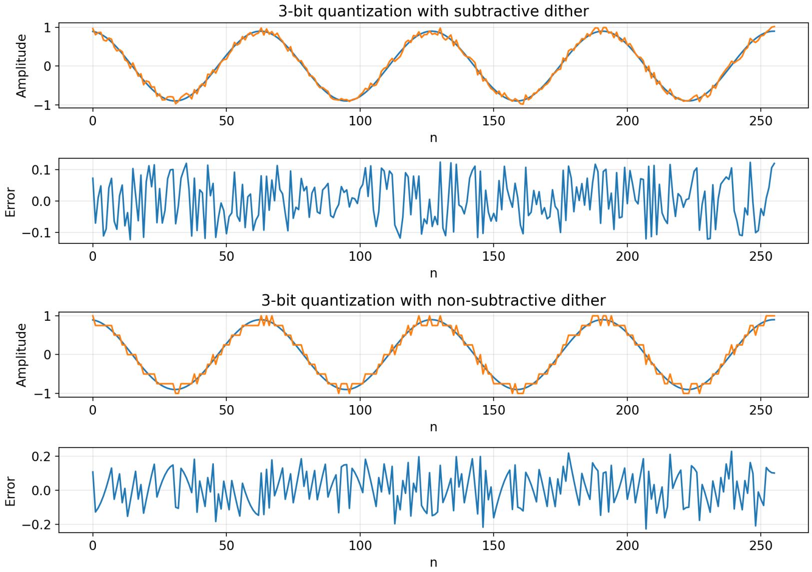
the error $\varepsilon = y - x$ satisfies

$$\boxed{\varepsilon \sim \mathcal{U}\left(-\frac{\Delta}{2}, \frac{\Delta}{2}\right), \qquad \varepsilon \perp x}$$

Quantization error and the signal are independent (see appendix for the proof). Consequently,

$$\mathbb{E}[\varepsilon \mid x] = 0, \qquad \mathbf{Var}(\varepsilon) = \frac{\Delta^2}{12}.$$

Subtractive vs non-subtractive dithering of a sine wave



Summary

- Bennett's theorem explains when quantization error *looks* uniform in the high-rate regime.
- Dithering is a way to *engineer* this behavior at finite rate.
  - non-subtractive dither: linearizes in expectation (good for images)
  - subtractive dither: gives uniform, input-independent error (great for analysis) but need to make the noise sequence available at the receiver
- Stochastic rounding is an unbiased randomized quantizer; tightly related to dither.
- NVFP4 utilizes stochastic rounding for unbiasedness.

Appendix: Why subtractive dither gives signal-independent noise

Let $d \sim \mathcal{U}(-\Delta/2, \Delta/2)$ and define

$$z = x + d, \qquad \hat{x} = Q_\Delta(z) - d.$$

Quantization depends only on the position of z within a cell. Write $x = k\Delta + r$ with $r \in [-\Delta/2, \Delta/2)$, so

$$z = k\Delta + (r + d).$$

Since d is uniform over one full quantization step, the random variable $(r + d)$ is uniform over an interval of length Δ, independent of r.
Wrapping any length-Δ interval into one cell produces a uniform variable on $[-\Delta/2, \Delta/2)$, hence

$$e \triangleq \hat{x} - x \sim \mathcal{U}(-\Delta/2, \Delta/2),$$

with a distribution that does not depend on x.

Conclusion: Subtractive dither yields uniform, signal-independent noise.