A Direct Upper Bound on Optimal Scalar Quantization Error

Sina Baghal

Last updated: March 27, 2026

Abstract

This note provides a self-contained proof of the scalar quantization upper bound used in the TurboQuant paper [1]. For any log-concave density f_X with compact support $[a, b]$, any $k \geq 1$, and boundary vanishing exponent $\beta \geq 0$, we give an explicit finite-k upper bound:

$$\mathcal{C}(f_X, k) \;\leq\; \frac{C^3}{12k^2} + \frac{M}{6}\left(\frac{C(3+\beta)}{3\ell^{1/3}}\right)^{9/(3+\beta)} k^{-9/(3+\beta)},$$

where $C = \int f_X^{1/3}$, $M = \max f_X$, and $\ell > 0$, $\beta \geq 0$ are the constants in the boundary lower bound $f_X(x) \geq \ell(x-a)^\beta$ near $x = a$ (and $f_X(x) \geq \ell(b-x)^\beta$ near $x = b$). The proof combines a variance bound for log-concave densities [2] with a direct analysis of the equal $f_X^{1/3}$-mass partition.

1 Background

A *scalar quantizer* with k levels for a real-valued source $X \sim f_X$ consists of an encoder $Q : \mathbb{R} \to [k]$ and a decoder $Q^{-1} : [k] \to \mathbb{R}$ producing reconstruction levels (centroids) $c_1 < \cdots < c_k$. The optimal encoder assigns each input to its nearest centroid, which in 1D partitions $\mathbb{R}$ into k contiguous *Voronoi cells* $[b_{i-1}, b_i]$ with boundaries $b_i = (c_i + c_{i+1})/2$. The resulting mean-squared error (MSE) distortion decomposes as

$$D = \sum_{i=1}^{k} \int_{b_{i-1}}^{b_i} |x - c_i|^2 \, f_X(x)\, dx. \tag{1}$$

The *Lloyd-Max problem* is to minimize (1) jointly over the k centroids. Because the boundaries are determined by the centroids via the nearest-neighbor rule, the optimization reads

$$\mathcal{C}(f_X, k) := \min_{c_1 \leq \cdots \leq c_k} \sum_{i=1}^{k} \int_{\frac{c_{i-1}+c_i}{2}}^{\frac{c_i+c_{i+1}}{2}} |x - c_i|^2 \, f_X(x)\, dx, \tag{2}$$

where $c_0 = -\infty$ and $c_{k+1} = +\infty$. This is the **continuous 1D k-means problem**: minimize the expected squared distance from a continuously distributed random variable X to its nearest centroid. In digital quantization, $k = 2^b$ where b is the bit-width, so $\mathcal{C}(f_X, b)$ denotes the optimal MSE at bit-width b.

2 A Direct Upper Bound

Theorem 1. *Let f_X be a log-concave probability density with compact support $[a, b]$, and let $C := \int_a^b f_X(x)^{1/3}\,dx$, $M = \max f_X$. Suppose f_X satisfies (12) with constants $\ell > 0$ and $\beta \geq 0$. Then for all $k \geq 1$:*

$$\mathcal{C}(f_X, k) \;\leq\; \frac{C^3}{12k^2} + \frac{M}{6}\left(\frac{C(3+\beta)}{3\,\ell^{1/3}}\right)^{9/(3+\beta)} k^{-9/(3+\beta)}. \tag{3}$$

The proof builds on the following proposition.

Proposition 1. *Under the same assumptions (with f_X log-concave), define the equal $f_X^{1/3}$-mass partition $b_0 < b_1 < \cdots < b_k$ by $\int_{b_{i-1}}^{b_i} f_X^{1/3}\,dx = C/k$, and let $\Delta_i = b_i - b_{i-1}$, $p_i = \int_{b_{i-1}}^{b_i} f_X\,dx$. Then for any $k \geq 1$:*

$$\mathcal{C}(f_X, k) \;\leq\; \frac{1}{12}\sum_{i=1}^{k} \Delta_i^2\, p_i. \tag{4}$$

The proof of Proposition 1 proceeds in two exact steps.

Step 1: Construct a feasible quantizer. Since $\mathcal{C}(f_X, k)$ is the *minimum* distortion over all k-level quantizers, it is bounded above by the distortion of any specific quantizer.

We build one tailored to the structure of f_X: Choose boundaries $b_0 < b_1 < \cdots < b_k$ so that each cell carries equal $f_X^{1/3}$-mass:

$$\int_{b_{i-1}}^{b_i} f_X(x)^{1/3}\,dx = \frac{C}{k}, \qquad i = 1, \ldots, k. \tag{5}$$

Such boundaries exist and are unique because $x \mapsto \int_{-\infty}^{x} f_X^{1/3}$ is continuous and strictly increasing. Set each centroid to the conditional mean: $c_i = \mathbb{E}[X \mid X \in [b_{i-1}, b_i]]$. Moreover, since $f_X^{1/3}$ is large where f_X is large, condition (5) concentrates many narrow cells in high-density regions and places wide cells in the tails, balancing the per-cell contribution to distortion.

Let $\Delta_i = b_i - b_{i-1}$ and $p_i = \int_{b_{i-1}}^{b_i} f_X\,dx$. The distortion of our specific quantizer Q decomposes cell by cell from (1):

$$D_Q = \sum_{i=1}^{k} \int_{b_{i-1}}^{b_i} (x - c_i)^2\, f_X(x)\,dx.$$

Multiply and divide each term by $p_i > 0$:

$$\int_{b_{i-1}}^{b_i} (x - c_i)^2\, f_X(x)\,dx = p_i \cdot \underbrace{\frac{\displaystyle\int_{b_{i-1}}^{b_i} (x - c_i)^2\, f_X(x)\,dx}{\displaystyle\int_{b_{i-1}}^{b_i} f_X(x)\,dx}}_{=\,\mathbb{E}[(X-c_i)^2 \mid X \in [b_{i-1}, b_i]]}.$$

Since $c_i = \mathbb{E}[X \mid X \in [b_{i-1}, b_i]]$ is the conditional mean, the conditional expectation of the squared deviation equals the conditional variance: $\mathbb{E}[(X - c_i)^2 \mid X \in [b_{i-1}, b_i]] = \mathrm{Var}(X \mid X \in [b_{i-1}, b_i])$. Therefore:

$$\mathcal{C}(f_X, k) \;\le\; D_Q = \sum_{i=1}^{k} \mathrm{Var}(X \mid X \in [b_{i-1}, b_i]) \cdot p_i. \tag{6}$$

Step 2: Bound $\mathrm{Var}(X \mid X \in [b_{i-1}, b_i])$**.** This step follows from the lemma below.

Lemma 1. *Let Z be a random variable with a log-concave density supported on an interval of width Δ. Then*

$$\mathrm{Var}(Z) \le \frac{\Delta^2}{12}, \tag{7}$$

with equality if and only if Z is uniform on the interval.[1]

Proof. By rescaling, $\Delta = 1$ on $[0, 1]$. Write $f = e^{\phi}$ with ϕ concave (log-concavity of f). The reference family is the truncated exponentials $g_\lambda(x) \propto e^{\lambda x}$, $\lambda \in \mathbb{R}$ ($g_0 \equiv 1$ is uniform with variance $1/12$). Choose λ so that $E_f[Z] = E_{g_\lambda}[Z]$ (such λ exists by the intermediate value theorem: $E_{g_\lambda}[Z] \to 0$ as $\lambda \to -\infty$ and $E_{g_\lambda}[Z] \to 1$ as $\lambda \to +\infty$). The ratio $f/g_\lambda = c\, e^{\phi(x) - \lambda x}$ is log-concave (as $\phi(x) - \lambda x$ is concave), so the densities f and g_λ satisfy the single-crossing condition. By the Karlin–Novikoff theorem (Lemma B in [2]), this yields convex-order domination $Z_f \prec_{cx} Z_{g_\lambda}$: for every convex ψ, $E_f[\psi(Z)] \le E_{g_\lambda}[\psi(Z)]$. Taking $\psi(z) = (z - c)^2$ and minimizing over c:

$$\mathrm{Var}(Z_f) \;\le\; \mathrm{Var}(Z_{g_\lambda}) \;\le\; \sup_{\lambda \in \mathbb{R}} \mathrm{Var}(Z_{g_\lambda}) \;=\; \tfrac{1}{12},$$

where the supremum equals $1/12$ because $\mathrm{Var}(Z_{g_\lambda}) = 1/\lambda^2 - e^\lambda/(e^\lambda - 1)^2$ (by direct calculation), which is continuous, equals $1/12$ at $\lambda = 0$, tends to 0 as $|\lambda| \to \infty$, and satisfies $1/\lambda^2 - e^\lambda/(e^\lambda - 1)^2 \le 1/12$ for all $\lambda \ne 0$. Equality forces ϕ affine, i.e. $f = g_\lambda$ for some λ; then $\mathrm{Var}(Z_f) = \mathrm{Var}(Z_{g_\lambda}) \le 1/12$, with equality only for $\lambda = 0$. □

We now proceed to use Lemma 1 to show Proposition 1. Log-concavity is preserved under conditioning to an interval and Lemma 1 therefore gives $\mathrm{Var}(X \mid X \in [b_{i-1}, b_i]) \le \Delta_i^2/12$. Substituting into (6):

$$\mathcal{C}(f_X, k) \;\le\; \frac{1}{12} \sum_{i=1}^{k} \Delta_i^2\, p_i. \tag{8}$$

We are now ready to prove Theorem 1. By the mean-value theorem applied to each cell, there exist $\xi_i^{(1)}, \xi_i^{(2)} \in [b_{i-1}, b_i]$ such that

$$\frac{C}{k} = f_X\left(\xi_i^{(1)}\right)^{1/3} \Delta_i, p_i = f_X\left(\xi_i^{(2)}\right) \Delta_i.$$

Thus,

$$\Delta_i = \frac{C}{k\, f_X(\xi_i^{(1)})^{1/3}},$$

[1] Without the log-concavity assumption, the best bound for a density supported on an interval of width Δ is $\mathrm{Var}(Z) \le \Delta^2/4$ (Hoeffding), which is three times weaker.

Substituting Δ_i and p_i directly:

$$\Delta_i^2\, p_i = \frac{C^2}{k^2\, f_X\left(\xi_i^{(1)}\right)^{2/3}} \cdot \frac{C\, f_X\left(\xi_i^{(2)}\right)}{k\, f_X\left(\xi_i^{(1)}\right)^{1/3}} = \frac{C^3}{k^3} \cdot \frac{f_X\left(\xi_i^{(2)}\right)}{f_X\left(\xi_i^{(1)}\right)}. \tag{9}$$

Write $r_i := f_X(\xi_i^{(2)})/f_X(\xi_i^{(1)})$, so from (9):

$$\sum_{i=1}^{k} \Delta_i^2\, p_i \;=\; \frac{C^3}{k^3} \sum_{i=1}^{k} r_i. \tag{10}$$

We now proceed to upper bound $\sum r_i$. Let $V_i = \int_{b_{i-1}}^{b_i} |f_X'|\, dx$ be the total variation of f_X on cell i. Since $|f_X(\xi_i^{(2)}) - f_X(\xi_i^{(1)})| \le V_i$:

$$r_i - 1 = \frac{f_X(\xi_i^{(2)}) - f_X(\xi_i^{(1)})}{f_X(\xi_i^{(1)})} \le \frac{V_i}{f_X(\xi_i^{(1)})} = \frac{V_i\, k^3\, \Delta_i^3}{C^3}.$$

Summing, then bounding $\sum_i V_i \Delta_i^3 \le \Delta_{\max}^3 \sum_i V_i = \Delta_{\max}^3 \cdot \mathrm{TV}(f_X)$:

$$\sum_{i=1}^{k} \Delta_i^2\, p_i \;\le\; \frac{C^3}{k^2} + \mathrm{TV}(f_X) \cdot \Delta_{\max}^3. \tag{11}$$

A log-concave density is unimodal with maximum M at m (since $\log f_X$ is concave, hence unimodal, and $e^{\cdot}$ is monotone). Thus f_X is non-decreasing on $[a, m]$ and non-increasing on $[m, b]$, so by the fundamental theorem of calculus:

$$\mathrm{TV}(f_X) = \int_a^b |f_X'| = \int_a^m f_X'\, dx + \int_m^b (-f_X')\, dx = \big(M - f_X(a)\big) + \big(M - f_X(b)\big) \le 2M.$$

We are left to show the upper bound on $\Delta_{\max}$. Let $m = \arg\max_{x\in[a,b]} f_X(x)$ and $M = f_X(m) > 0$. Suppose f_X satisfies the boundary lower bound

$$f_X(x) \;\ge\; \ell\,(x-a)^\beta \;\text{ for } x \in [a, m], \qquad f_X(x) \;\ge\; \ell\,(b-x)^\beta \;\text{ for } x \in [m, b], \tag{12}$$

for some $\ell > 0$ and $\beta \ge 0$. Since f_X is log-concave (hence unimodal), $f_X^{1/3}$ is smallest near the endpoints, so the boundary cells are widest: $\Delta_{\max} = \max(\Delta_1, \Delta_k)$. By symmetry assume it is Δ_1, the cell abutting a. Then $f_X^{1/3}(x) \ge \ell^{1/3}(x-a)^{\beta/3}$ on that cell, so:

$$\frac{C}{k} = \int_{b_0}^{b_1} f_X^{1/3}\, dx \;\ge\; \ell^{1/3} \int_{b_0}^{b_1} (x - b_0)^{\beta/3}\, dx = \frac{3\,\ell^{1/3}}{3+\beta}\, \Delta_{\max}^{(3+\beta)/3}.$$

Solving:

$$\Delta_{\max} \;\le\; \left(\frac{C(3+\beta)}{3\,\ell^{1/3}\,k}\right)^{3/(3+\beta)}. \tag{13}$$

Substituting into (11), using $\mathrm{TV}(f_X) \le 2M$, and dividing by 12 via (8):

$$\mathcal{C}(f_X, k) \;\le\; \frac{C^3}{12k^2} + \frac{M}{6}\left(\frac{C(3+\beta)}{3\,\ell^{1/3}}\right)^{9/(3+\beta)} k^{-9/(3+\beta)}. \tag{14}$$

This completes the proof of Theorem 1. □

3 Application to Beta density

We finish by computing the parameters in the main bound for the Beta distribution. For $d \geq 3$, we have $f_X(x) = K(1-x^2)^{(d-3)/2}$ on $[-1,1]$, where $K = \Gamma(d/2)/(\sqrt{\pi}\,\Gamma((d-1)/2))$ is the normalization constant and $M = \max f_X = K = f_X(0)$. The parameters in Theorem 1 are computed as follows.

Endpoint	**Quotient**	**Bound**
$x = -1$, on $[-1,0]$	$\dfrac{f_X(x)}{(x+1)^{(d-3)/2}} = K(1-x)^{(d-3)/2}$	$\geq K$
$x = 1$, on $[0,1]$	$\dfrac{f_X(x)}{(1-x)^{(d-3)/2}} = K(1+x)^{(d-3)/2}$	$\geq K$

Thus $\ell = M = K$ and $\beta = (d-3)/2$ satisfy both boundary conditions in (12). Substituting into (3) gives the explicit finite-k bound:

$$\mathcal{C}(f_X, k) \;\leq\; \frac{C^3}{12k^2} + \frac{M}{6}\left(\frac{C(d+3)/2}{3\,M^{1/3}}\right)^{18/(d+3)} k^{-18/(d+3)},$$

where $C = \int_{-1}^{1} f_X^{1/3}\,dx$ and $M = \Gamma(d/2)/(\sqrt{\pi}\,\Gamma((d-1)/2))$.

References

[1] A. Zandieh, M. Daliri, M. Hadian, and V. Mirrokni, TurboQuant: Online vector quantization with near-optimal distortion rate, *arXiv preprint arXiv:2504.19874*, 2025.

[2] S. Karlin and A. Novikoff, Generalized convex inequalities, *Pacific Journal of Mathematics*, 13(4):1251–1279, 1963.