Unified manuscript

Distribution-Controlled Selective Quantization

Distortion, allocation, and margin-survival accuracy curves

J. R. Landers

May 2026

Abstract. Worst-case quantization bounds price every coordinate as if every perturbation mattered equally. A distribution-controlled view does something sharper: it spends low precision where the model is insensitive on the data it actually sees. For linear readouts, blockwise ReLU networks, and coefficient-feature classifiers, the same first-order product governs distortion,

$$\text{selection probability} \times \text{error scale} \times \text{sensitivity}.$$

This manuscript unifies the selective-quantization note with its margin-survival extension. The product gives a threshold rule under a linear first-order bound, becomes a logistic gate under entropy smoothing, is modified by signed cancellation under deterministic rounding, and supports constrained memory–loss distribution shaping. Margins then convert cumulative output distortion into training-accuracy curves: sorted least-sensitive-first quantization accumulates distortion slowly, and accuracy survives until that cumulative distortion reaches the body of the margin distribution. On handwritten digits, a closed-form lognormal-plus-Weibull margin-survival model fits the least-sensitive-first coefficient-feature sweep with an RMSE of 0.159 percentage points. A separate ReLU bit-allocation experiment reaches full-precision test accuracy at 4.01 average bits and exceeds uniform 4-bit accuracy at 3.43 average bits.

1 The question

Quantization buys cheaper storage and arithmetic by replacing high-precision parameters with lower-precision approximations. The blunt analysis treats the resulting error as if it can hit every coordinate at once with equal consequence. That is safe, but it misses the geometry of a trained model: some directions move the output a great deal, while others barely register on the data distribution.

Selective quantization starts from the opposite end. It asks not simply "how large is the rounding error?" but "where does that error land?" The right object is therefore anisotropic: low precision should be allocated heavily to directions with small sensitivity, lightly to directions with large sensitivity, and smoothly or abruptly depending on the objective used to choose the allocation.

The extension in this manuscript adds one more layer. Output distortion is not yet accuracy. A classifier is correct as long as its margin survives the perturbation. The full story is therefore a composition:

$$\text{quantization schedule} \longrightarrow \text{cumulative distortion} \longrightarrow \text{margin survival} \longrightarrow \text{accuracy curve}.$$

2 Distribution-controlled gates

Consider a linear readout and its quantized counterpart,

$$f_w(x) = w^\top x + b, \qquad f_{\widetilde{w}}(x) = \widetilde{w}^\top x + \widetilde{b}, \qquad x, w \in \mathbb{R}^d.$$

Write the coordinatewise perturbation as

$$\widetilde{w}_i = w_i + M_i \xi_i.$$

The gate $M_i \in \{0, 1\}$ records whether coordinate i is placed in low precision, and ξ_i is the induced error. Assume

$$M_i \sim \text{Bernoulli}(q_i), \qquad |\xi_i| \le \varepsilon_i, \qquad |\widetilde{b} - b| \le \varepsilon_b.$$

The probabilities q_i are design variables. They are driven by a sensitivity score read from the data distribution $X \sim \mathcal{D}$, for example

$$s_i = \mathbb{E}_{\mathcal{D}}|X_i| \qquad \text{or} \qquad s_i = \mathbb{E}_{\mathcal{D}}X_i^2.$$

A simple selective family is

$$q_i \propto (s_i + \lambda)^{-\alpha}, \qquad \alpha > 0, \quad \lambda > 0,$$

so low-sensitivity coordinates are quantized more often. The exponent α is the selectivity dial; the floor λ prevents the smallest scores from absorbing all the mass.

The same rule can be read at population scale. Let $z = \log s$ and suppose the sensitivities have lognormal shape, $z \sim \mathcal{N}(m, \sigma^2)$. With $\lambda \to 0$, multiplying by $q(s) \propto s^{-\alpha} = e^{-\alpha z}$ shifts the log-density exactly:

$$z \mid \text{quantized} \sim \mathcal{N}(m - \alpha\sigma^2, \sigma^2).$$

Selectivity does not merely lower an average error; it moves the whole quantization mass away from the high-sensitivity tail.

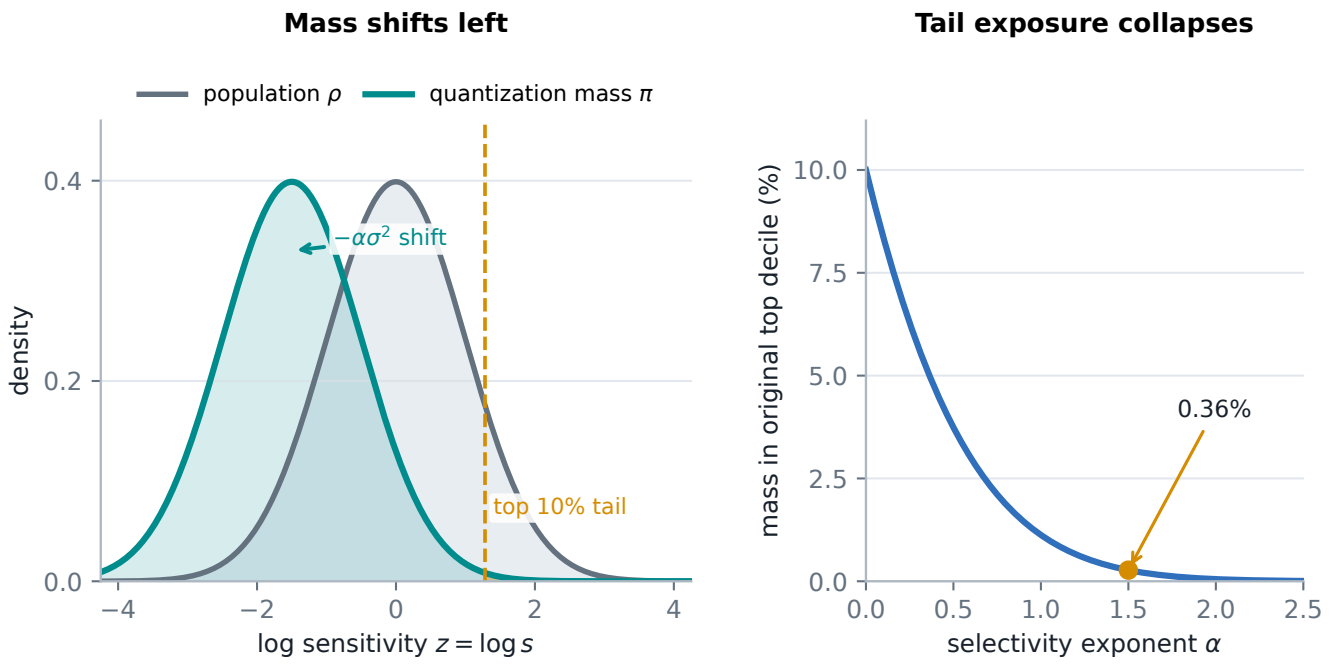

Figure 1: A lognormal sensitivity population reweighted by $q(s) \propto s^{-\alpha}$. The quantization mass shifts left by $\alpha\sigma^2$ in log-sensitivity. In the displayed case $m = 0$, $\sigma = 1$, and $\alpha = 1.5$, so the original top decile receives only about 0.36% of the quantization mass.

3 First-order distortion

Proposition 1 (Selective quantization bound). *For every fixed input $x \in \mathbb{R}^d$,*

$$\mathbb{E}_Q[|f_{\widetilde{w}}(x) - f_w(x)|] \le \sum_{i=1}^{d} q_i \varepsilon_i |x_i| + \varepsilon_b,$$

where the expectation is over the quantization randomness.

Proof. The output difference is

$$f_{\widetilde{w}}(x) - f_w(x) = \sum_i M_i \xi_i x_i + (\widetilde{b} - b).$$

The triangle inequality gives

$$|f_{\widetilde{w}}(x) - f_w(x)| \le \sum_i M_i |\xi_i| |x_i| + |\widetilde{b} - b|.$$

Taking expectations, using $|\xi_i| \le \varepsilon_i$, $|\widetilde{b} - b| \le \varepsilon_b$, and $\mathbb{E}M_i = q_i$, yields the claim. □

Averaging over the data and writing $s_i = \mathbb{E}|X_i|$ gives the central first-order object:

$$\boxed{\mathbb{E}_{X,Q}[|f_{\widetilde{w}}(X) - f_w(X)|] \le \sum_{i=1}^{d} q_i\, \varepsilon_i\, s_i + \varepsilon_b.}$$

The product $q_i \varepsilon_i s_i$ is the price of placing low precision at coordinate i. Uniform quantization sets $q_i = q$ and $\varepsilon_i = \varepsilon$, giving the coarser $q\varepsilon \sum_i s_i + \varepsilon_b$. Selectivity improves the same bound by moving probability mass to coordinates where the product is cheap.

3.1 The threshold rule

Suppose the error scale is common, $\varepsilon_i = \varepsilon$, and the expected number of quantized coordinates is fixed:

$$\sum_i q_i = K, \qquad 0 \le q_i \le 1.$$

The first-order problem is a linear program:

$$\min_q \ \varepsilon \sum_i q_i s_i + \varepsilon_b.$$

Because the objective is linear over a box intersected with one budget hyperplane, an optimum occurs at a vertex.

Proposition 2 (Budget-optimal selection)**.** *Let $s_{(1)} \le \cdots \le s_{(d)}$. If K is an integer, an optimizer is*

$$q_{(i)} = 1 \quad (i \le K), \qquad q_{(i)} = 0 \quad (i > K).$$

For noninteger K, the same sorting rule holds with one fractional coordinate.

Thus the familiar heuristic "quantize the unimportant coordinates" is not only intuitive. Under the first-order distortion bound it is the exact variational solution. The limitation is also clear: the rule is abrupt. A coordinate just below the threshold is fully exposed to low precision, and a coordinate just above it is fully spared.

4 Smoothing the threshold

Abrupt rules are sometimes undesirable. They are hard to tune, brittle under noisy sensitivity estimates, and inconvenient when the selection mechanism is trained by gradient methods. Add a convex entropy penalty to soften the vertex solution:

$$\min_{q \in [0,1]^d} \sum_i q_i c_i + \mu \sum_i [q_i \log q_i + (1 - q_i) \log(1 - q_i)] \quad \text{s.t.} \quad \sum_i q_i = K,$$

where $c_i = \varepsilon_i s_i$ and $\mu > 0$.

Proposition 3 (Smooth allocation)**.** *The unique minimizer is logistic in the coordinate cost:*

$$q_i^\star = \frac{1}{1 + \exp((c_i + \nu)/\mu)},$$

where ν is chosen so that $\sum_i q_i^\star = K$. As $\mu \to 0$, the solution steepens into the threshold of Proposition 2.

Proof. The Lagrangian stationarity condition is

$$c_i + \mu \log \frac{q_i}{1 - q_i} + \nu = 0.$$

Solving for q_i gives the displayed logistic form. The entropy term is strictly convex on $(0, 1)$ and its gradient diverges at the endpoints, so the minimizer is interior and unique. The remaining scalar ν is fixed by the monotone budget equation. □

The temperature μ is the bridge between selection and relaxation. At zero temperature the solution is a sorted subset. At positive temperature it is a graded allocation that still prefers low c_i, but no longer has to make a discontinuous decision.

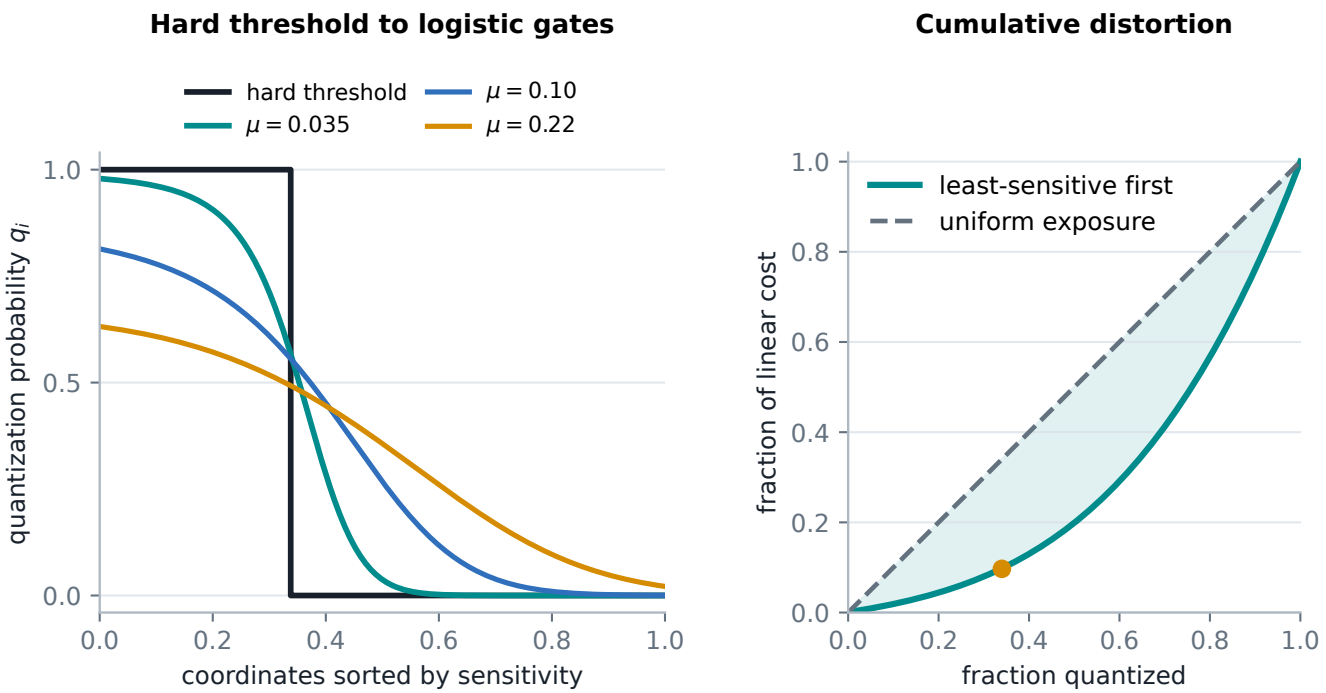

Figure 2: The first-order bound sorts coordinates by sensitivity and quantizes the cheapest ones first. Entropy smoothing rounds the threshold into a logistic gate. The right panel shows why the sorted rule helps: spending a fixed budget on low-sensitivity coordinates accumulates much less linear distortion than uniform exposure.

Remark 1. Replacing the Bernoulli gate with a continuous bit allocation b_i, with a scale such as $\varepsilon_i^2 \sim 2^{-2b_i}$, gives the classical reverse-water-filling shape from rate-distortion theory [7]. The entropy relaxation above is the Bernoulli-gate analogue: a discrete allocation is bent into a graded one.

5 When signs matter

The preceding bounds add magnitudes and ignore sign. If quantization errors are independent, bounded, and mean zero, as in idealized stochastic rounding [4], signed terms cancel and concentration gives a sharper tail. Define

$$Z(x) = f_{\widetilde{w}}(x) - f_w(x) - (\widetilde{b} - b) = \sum_i M_i \xi_i x_i.$$

Then $\mathbb{E}Z(x) = 0$ and

$$\mathrm{Var}(Z(x)) \le V(x) := \sum_i q_i \varepsilon_i^2 x_i^2.$$

With $a(x) = \max_i \varepsilon_i |x_i|$, Bernstein's inequality gives, with probability at least $1 - \delta$,

$$|f_{\widetilde{w}}(x) - f_w(x)| \le \sqrt{2V(x) \log \frac{2}{\delta}} + \frac{2}{3} a(x) \log \frac{2}{\delta} + \varepsilon_b.$$

This is sharper than the first-order magnitude bound, but its dependence on q_i is still linear inside $V(x)$. It changes the scale of the error, not the qualitative allocation pressure.

Deterministic rounding is different. If round-to-nearest commits the same signed error every time coordinate i is selected, write $\xi_i = \beta_i$ with $|\beta_i| \le \varepsilon_i$, and let the only randomness be the gate.

Proposition 4 (Selection variance and tail). *Let $u_i = \beta_i x_i$ and $Z(x) = \sum_i M_i u_i$, with $M_i \sim \mathrm{Bernoulli}(q_i)$. Then*

$$m(x) := \mathbb{E}Z(x) = \sum_i q_i u_i, \qquad \mathrm{Var}(Z(x)) = \sum_i q_i (1 - q_i) u_i^2.$$

With $a = \max_i |u_i| \le \max_i \varepsilon_i |x_i|$, with probability at least $1 - \delta$,

$$|Z(x)| \le \left| \sum_i q_i u_i \right| + \sqrt{2 \log \frac{2}{\delta}} \sqrt{\sum_i q_i (1 - q_i) u_i^2} + \frac{2}{3} a \log \frac{2}{\delta}.$$

Now the shape has changed. The bias term is signed, so positive and negative contributions can cancel. The fluctuation term contains $q_i(1-q_i)$, which is concave and vanishes at $q_i = 0$ and $q_i = 1$. Entropy smoothing pulls gates toward the interior; deterministic-rounding fluctuation pushes them back to the corners.

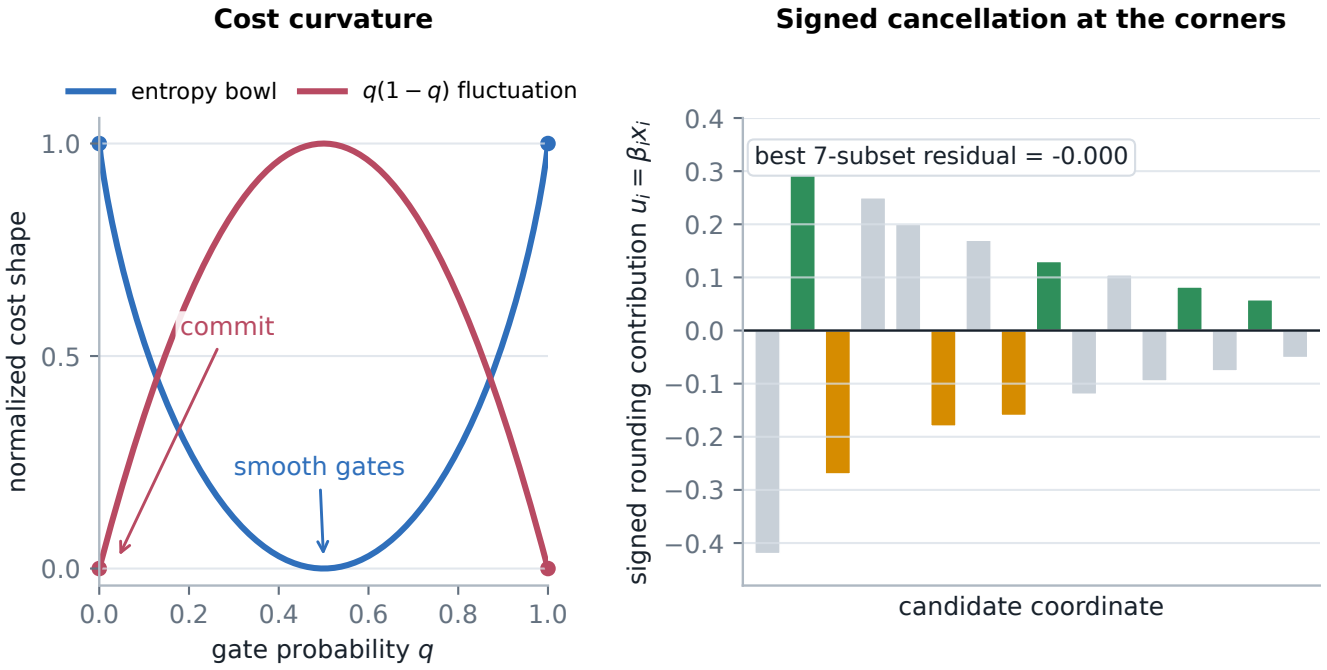

Figure 3: Curvature decides whether the allocation hedges or commits. The entropy penalty has a convex bowl and favors smooth gates. The deterministic-rounding variance $q(1-q)$ is concave and vanishes at the corners, so once signs are fixed the remaining task becomes choosing a subset whose signed contributions cancel.

Remark 2 (Commit, then cancel). At a deterministic vertex $q_i \in \{0, 1\}$, the gate variance disappears. With a budget $|S| = K$, the residual deterministic-rounding objective becomes

$$\min_{|S|=K} \left| \sum_{i \in S} \beta_i x_i \right|.$$

This is a signed subset-sum problem. The rule is no longer simply "choose the smallest nonnegative costs"; it uses the signs of the rounding errors, coupling coordinates that a coordinatewise sensitivity score would treat independently.

6 What survives under depth

The same anatomy lifts from coordinates to parameter blocks. Let f_θ be a ReLU network with blocks $\theta = (\theta_1, \ldots, \theta_m)$, and quantize block j by

$$\widetilde{\theta}_j = \theta_j + M_j \Delta_j, \qquad M_j \sim \text{Bernoulli}(q_j), \qquad \|\Delta_j\| \le \varepsilon_j.$$

For an input region $\mathcal{X}$, define the block sensitivity

$$G_j = \sup_{x \in \mathcal{X}} \left\| \frac{\partial f_\theta(x)}{\partial \theta_j} \right\|.$$

For ReLU networks this derivative exists inside each activation region; at boundaries G_j can be read as a local Lipschitz sensitivity over the visited region. A mean-value bound gives, to first order,

$$\|f_{\widetilde{\theta}}(x) - f_\theta(x)\| \lesssim \sum_{j=1}^{m} M_j \varepsilon_j G_j.$$

Taking expectation over the gates,

$$\boxed{\mathbb{E}_Q \left[\|f_{\widetilde{\theta}}(x) - f_\theta(x)\| \right] \lesssim \sum_{j=1}^{m} q_j \varepsilon_j G_j.}$$

The coordinate sensitivity s_i has become a block sensitivity G_j, but the product is unchanged. ReLU itself is 1-Lipschitz; amplification comes from the surrounding linear maps and the sensitivity of downstream computation.

7 Distribution shaping as constrained optimization

The preceding sections describe the geometry of a fixed allocation. The natural next step is to choose the allocation itself. There is no single point that simultaneously minimizes memory and loss except at trivial extremes; the object is a Pareto frontier. One convenient side is

$$\min_{\phi} \ \mathcal{M}(\phi) \qquad \text{s.t.} \qquad \mathcal{D}(\phi) \le \tau,$$

where τ is the allowed distortion. Let $z = \log G$ be log-sensitivity and let a policy ϕ produce a quantization probability $q_\phi(z)$ and a bit width $b_\phi(z) \in [b_{\min}, b_{\max}]$. With population density $\rho(z)$,

$$\mathcal{M}(\phi) = \int \rho(z)\Big[(1 - q_\phi(z))b_{\text{full}} + q_\phi(z)b_\phi(z)\Big]\, dz$$

and the first-order distortion proxy is

$$\mathcal{D}(\phi) = \int \rho(z)\, q_\phi(z)\, \varepsilon(b_\phi(z))\, e^z\, dz.$$

The empirical version replaces the integrals by sums over weights or blocks. A differentiable constrained optimizer then minimizes the augmented objective

$$\mathcal{J}(\phi, \lambda) = \mathcal{M}(\phi) + \lambda\, [\mathcal{D}(\phi) - \tau]_+^2 + \eta\, \mathcal{R}(\phi).$$

The regularizer $\mathcal{R}$ controls the shape of the quantization distribution

$$\pi_\phi(z) = \frac{\rho(z) q_\phi(z)}{\int \rho(u) q_\phi(u)\, du}.$$

For example, $\mathcal{R} = \mathrm{KL}(\pi_\phi \| \pi_{\text{target}})$ can pull the allocation toward a uniform law, a shifted lognormal law, or a learned mixture between them. Without this term the optimizer is free to find whatever distribution the memory–loss constraint demands.

Proposition 5 (Continuous-bit water filling). *Consider the deterministic bit-allocation problem*

$$\min_{b_{\min} \le b_j \le b_{\max}} \sum_j m_j b_j \quad \text{s.t.} \quad \sum_j a_j 2^{-b_j} \le \tau,$$

where $a_j = C_j G_j$ combines sensitivity and quantizer scale, and m_j is the memory price of block j. At any interior optimum,

$$b_j^\star = \log_2\left(\frac{\lambda a_j \log 2}{m_j}\right),$$

clipped to $[b_{\min}, b_{\max}]$, with λ chosen so the constraint is active.

Proof. The Lagrangian is

$$\sum_j m_j b_j + \lambda\left(\sum_j a_j 2^{-b_j} - \tau\right).$$

Stationarity gives $m_j - \lambda a_j (\log 2) 2^{-b_j} = 0$, which rearranges to the displayed expression. The box constraints clip the result at the endpoints. □

Thus the optimizer does not need to assume that the final allocation is uniform, lognormal, or thresholded. Those are possible shapes. The constrained problem chooses the shape implied by sensitivity, quantizer scale, and the allowed loss.

8 From distortion to margins

The preceding theory controls score movement. A classifier's accuracy is governed by whether that movement crosses decision margins. For a multiclass classifier with scores $f_c(x)$, define the clean training margin of example (x_i, y_i) by

$$m_i = f_{y_i}(x_i) - \max_{c \neq y_i} f_c(x_i).$$

The example is correctly classified when $m_i > 0$. After quantizing a cumulative set S_k, write the score perturbation as

$$\delta_c^{(k)}(x_i) = f_c^{(k)}(x_i) - f_c(x_i).$$

The quantized margin obeys

$$\begin{aligned} m_i(k) &= f_{y_i}(x_i) + \delta_{y_i}^{(k)}(x_i) - \max_{c \neq y_i} \left\{ f_c(x_i) + \delta_c^{(k)}(x_i) \right\} \\ &\geq m_i - \left| \delta_{y_i}^{(k)}(x_i) \right| - \max_{c \neq y_i} \left| \delta_c^{(k)}(x_i) \right|. \end{aligned}$$

Therefore the example is guaranteed to remain correct when

$$m_i > \left| \delta_{y_i}^{(k)}(x_i) \right| + \max_{c \neq y_i} \left| \delta_c^{(k)}(x_i) \right|.$$

For a linear score model, quantizing coefficient feature j changes class c's score by

$$\delta_c^{(k)}(x_i) = \sum_{j \in S_k} \Delta W_{cj} x_{ij}, \qquad \Delta W_{cj} = Q(W_{cj}) - W_{cj}.$$

A schedule-level proxy replaces the sample-specific perturbation by cumulative first-order exposure

$$D(k) = \sum_{j \in S_k} c_j, \qquad c_j = \varepsilon_j G_j.$$

This is the same cost appearing in the selective-quantization formalism. It controls output movement; margins convert output movement into accuracy loss.

Proposition 6 (Accuracy as margin survival)**.** *Let S_k be a cumulative quantization schedule and let $D(k) = \sum_{j \in S_k} c_j$. Suppose that, on the training set, the induced margin perturbation is well approximated by a common scale factor times cumulative distortion:*

$$|\Delta m_i(k)| \approx \gamma D(k).$$

Let F_M be the empirical CDF of clean positive margins among initially correct training examples. Ignoring rare repair events in which quantization fixes an initially incorrect example,

$$A(k) \approx A_0 \left[1 - F_M(\gamma D(k)) \right],$$

where A_0 is the full-precision training accuracy. With endpoint normalization to the fully quantized accuracy $A_\infty = A(d)$,

$$A(k) \approx A_0 - (A_0 - A_\infty) \frac{F_M(\gamma D(k)) - F_M(0)}{F_M(\gamma D(d)) - F_M(0)}.$$

Proof. For an initially correct example, the clean margin is positive. The quantized classifier keeps the example correct whenever the margin perturbation is smaller than the clean margin. Under the common-scale approximation, this event is $m_i > \gamma D(k)$. Averaging the indicator of this event over initially correct training examples gives the survival function $1 - F_M(\gamma D(k))$, and multiplying by A_0 gives the first approximation. The endpoint-normalized form rescales the cumulative margin-failure fraction so the model agrees with the observed fully quantized endpoint. □

accuracy curve ≈ margin survival ∘ cumulative quantization distortion.

The selective-quantization product supplies the distortion path $D(k)$. The margin distribution turns that path into discrete classification errors.

9 Closed-form least-sensitive-first accuracy

Specialize to the least-sensitive-first schedule. Let $\phi = k/d$ be the fraction of quantized coordinates and suppose coordinate costs are approximately lognormal:

$$C = \varepsilon G, \qquad \log C \sim \mathcal{N}(\mu, \sigma_s^2).$$

The least-sensitive-first cumulative distortion fraction is

$$R(\phi) = \frac{\mathbb{E}\left[C\,\mathbf{1}\{C \le Q_\phi\}\right]}{\mathbb{E}C},$$

where Q_ϕ is the ϕ-quantile of C. Since $C = \exp Z$ with $Z \sim \mathcal{N}(\mu, \sigma_s^2)$,

$$R(\phi) = \Phi\Big(\Phi^{-1}(\phi) - \sigma_s\Big).$$

This is the plateau generator. A large fraction of coordinates can be quantized while accumulating only a small fraction of the total sensitivity-weighted distortion.

Approximate the empirical margin CDF by a Weibull-type failure law,

$$F_M(t) \approx 1 - \exp\left[-\left(\frac{t}{\tau}\right)^\beta\right].$$

Absorbing the margin scale γ into τ, composition gives the explicit least-sensitive-first accuracy curve

$$A_{\text{blue}}(\phi) = A_0 - (A_0 - A_\infty)\frac{1 - \exp\left[-(R(\phi)/\tau)^\beta\right]}{1 - \exp\left[-(1/\tau)^\beta\right]}, \qquad R(\phi) = \Phi\Big(\Phi^{-1}(\phi) - \sigma_s\Big).$$

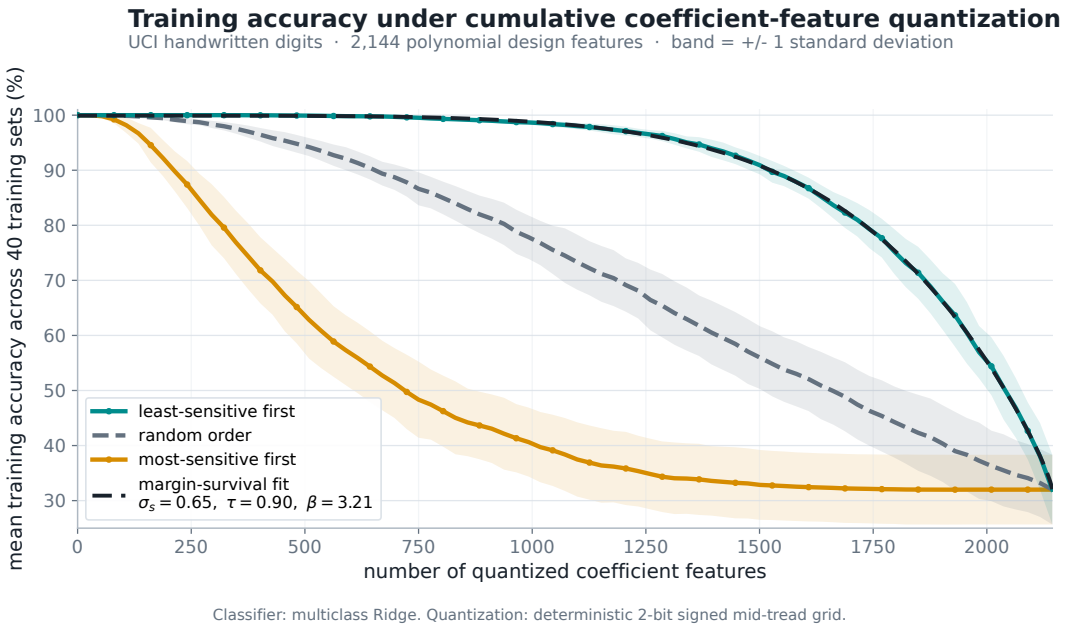

Figure 4: Cumulative coefficient-feature quantization on UCI/scikit-learn handwritten digits. The 64 pixel coordinates are expanded to 2,144 polynomial design features and a multiclass Ridge classifier is quantized cumulatively by coefficient-feature column using deterministic 2-bit signed mid-tread quantization. Shaded bands show ±1 standard deviation across 40 stratified training splits. The dashed black curve is the margin-survival fit. The least-sensitive-first schedule stays near perfect training accuracy until cumulative exposure enters the sensitive tail; random order degrades gradually; most-sensitive-first collapses early.

Fitting the closed-form curve to the least-sensitive-first training-accuracy curve in Figure 4 gives:

Quantity	Symbol	Fitted value
Full-precision training accuracy	A_0	99.923%
Fully quantized training accuracy	A_∞	32.065%
Log-cost spread	σ_s	0.651
Margin scale	τ	0.905
Margin-shape exponent	β	3.214
Fit RMSE	–	0.159 percentage points

The fit is not meant as a benchmark claim. It is a structural check. The original formalism predicts that least-sensitive-first quantization accumulates distortion slowly. The margin formalism predicts that accuracy remains stable until cumulative distortion enters the body of the margin distribution. Together they explain the plateau-then-collapse shape.

Remark 3 (Why the most-sensitive-first curve collapses early). For the most-sensitive-first schedule, $D(k)$ accumulates from the right tail of the cost distribution rather than from the left tail. In the lognormal approximation, the exposure of the largest ϕ fraction is

$$R_{\text{high}}(\phi) = 1 - \Phi\Big(\Phi^{-1}(1-\phi) - \sigma_s\Big),$$

so the margin distribution is hit immediately. The random schedule lies between the two extremes because, in expectation, it samples sensitivity mass roughly in proportion to coordinate count rather than in sorted order.

10 Experiment: a ReLU classifier

As a small check beyond the linear and Ridge-classifier settings, we trained a two-layer ReLU classifier $64 \to 96 \to 10$ on the scikit-learn handwritten digits data [8]. After training, each parameter received a sensitivity score

$$s_i = |\theta_i|\sqrt{\widehat{F}_i} + 10^{-12},$$

where $\widehat{F}_i$ is an empirical diagonal Fisher estimate on a held-out calibration split. We then solved the constrained bit-allocation problem with $b_i \in \{2, \dots, 8\}$. The continuous relaxation optimized average bits under the proxy constraint $\sum_i s_i \varepsilon_i(b_i) \le \tau$; a discrete repair step rounded bits while preserving the same proxy budget. Uniform k-bit quantization gives the baselines and supplies comparable distortion budgets.

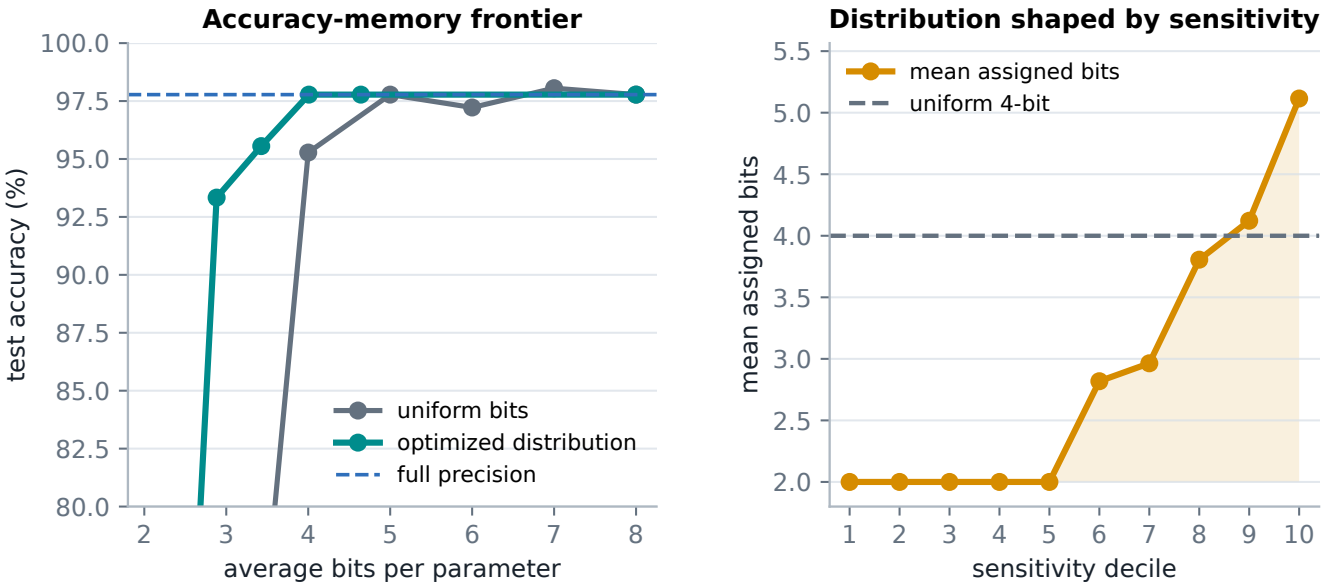

Figure 5: A small ReLU network follows the distributional prediction. *Left:* the optimized bit distribution reaches the full-precision test accuracy of 97.8% at 4.01 average bits, while uniform 4-bit quantization reaches 95.3%. At 3.43 average bits the optimized policy reaches 95.6%, slightly above uniform 4-bit at lower memory. *Right:* for the policy matched to the uniform 4-bit distortion budget, low-sensitivity deciles are assigned the minimum 2 bits, while high-sensitivity deciles receive more precision.

The experiment is deliberately small: it is not a benchmark claim. Its role is to verify that the product $q\varepsilon G$ and the resulting memory–loss tradeoff remain useful when the index set is a trained ReLU network rather than a single linear readout. The optimized distribution moves memory toward the sensitive tail and removes it from the flat directions, matching the theory.

11 One product, five geometries

$$\sum_j q_j \, \varepsilon_j \, G_j \qquad \text{selection probability} \times \text{error scale} \times \text{sensitivity}.$$

This product is the invariant. The allocation rule and the observed accuracy curve are determined by the curvature, constraints, and margin distribution built around it.

Regime	Shape of objective or readout	Behavior
First-order magnitude	Linear in q_i	Vertex solution; quantize least-sensitive coordinates first.
Entropy-relaxed selection	Convex smoothing term	Interior logistic gates; threshold recovered as $\mu \to 0$.
Deterministic rounding	Concave fluctuation $q_i(1-q_i)$ plus signed bias	Commit to corners, then choose subsets whose signed errors cancel.
Distribution control	Memory with loss constraint	Shape the quantization distribution along the Pareto frontier.
Margin survival	Thresholding by clean classification margins	Accuracy remains stable until cumulative distortion reaches the margin distribution.

The practical message is stable across the variants: spend numerical imprecision where sensitivity is small. What changes is the mode of spending. Linear costs commit by a threshold, convex relaxations hedge smoothly, deterministic signed errors return the problem to combinatorial commitment, constrained optimization turns allocation into direct distribution control, and margins turn cumulative distortion into the observed accuracy curve. Selective quantization is therefore not one algorithm so much as a geometry: a way of aligning compression with the distributional directions a model can afford to lose.

References

[1] Y. LeCun, J. S. Denker, and S. A. Solla. Optimal Brain Damage. *Advances in Neural Information Processing Systems*, 1989.

[2] Z. Dong, Z. Yao, Y. Cai, D. Arfeen, A. Gholami, M. W. Mahoney, and K. Keutzer. HAWQ-V2: Hessian Aware trace-Weighted Quantization of Neural Networks. *Advances in Neural Information Processing Systems*, 2020.

[3] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. *International Conference on Learning Representations*, 2023.

[4] S. Gupta, A. Agrawal, K. Gopalakrishnan, and P. Narayanan. Deep Learning with Limited Numerical Precision. *International Conference on Machine Learning*, 2015.

[5] C. Louizos, M. Welling, and D. P. Kingma. Learning Sparse Neural Networks through L_0 Regularization. *International Conference on Learning Representations*, 2018.

[6] E. Jang, S. Gu, and B. Poole. Categorical Reparameterization with Gumbel-Softmax. *International Conference on Learning Representations*, 2017.

[7] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 2011.