

CHAPTER 6

Compression

September 9, 2026

Compression must turn a small set of transform coefficients into a short binary description without introducing excessive reconstruction error. This requires controlling both the loss from quantization and the cost of recording coefficient values and locations. We derive error and bit-rate bounds for transform coding, explain how probability models reduce coding costs, and describe the wavelet and entropy-coding stages of JPEG-2000.

6.1 Transform Coding

6.1.1 Coding

Choose an orthonormal basis, such as a wavelet basis, and write the coefficients as

$$a_m = \langle f, \psi_m \rangle \in \mathbb{R}.$$

Quantization maps each coefficient to an integer using a step size $T > 0$:

$$q_m = Q_T(a_m) \in \mathbb{Z} \quad \text{where} \quad Q_T(x) = \text{sign}(x) \left\lfloor \frac{|x|}{T} \right\rfloor.$$

This quantizer has a zero bin of width $2T$: coefficients in $(-T, T)$ are set to zero.

Like hard thresholding (5.4), quantization discards coefficients smaller than T in magnitude. It also rounds the retained coefficients, introducing an additional error; see Figure 6.1.

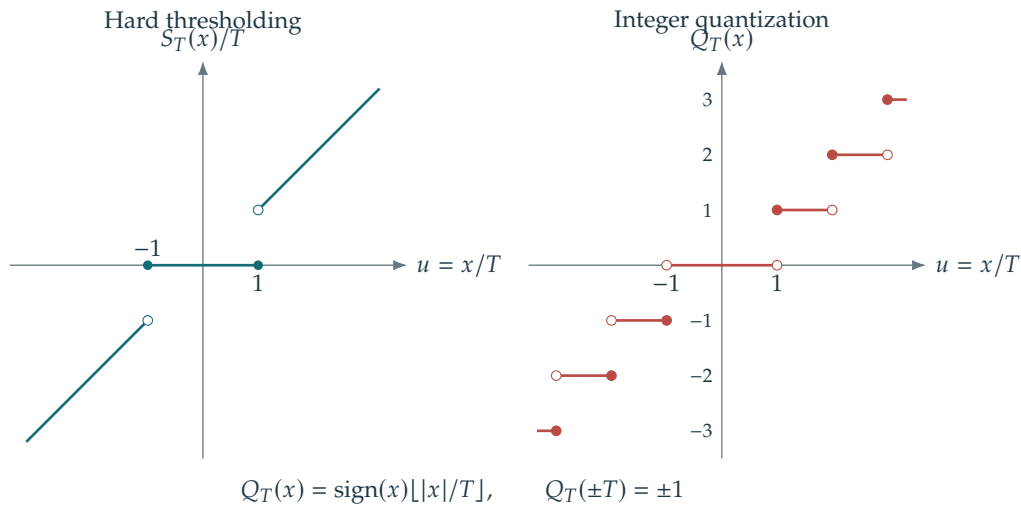


Figure 6.1. Hard thresholding and quantization as functions of the input coefficient.

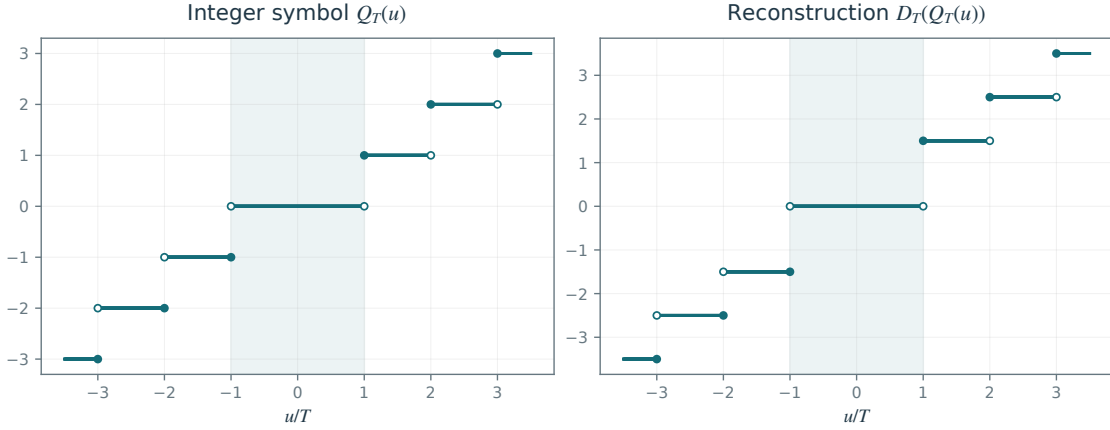
The integer coefficients $(q_m)_m$ are encoded in a binary stream of length R bits. Sections 6.1.3 and 6.2 describe two approaches to encoding these integers. The goal is to minimize R for an acceptable distortion.

6.1.2 Decoding

The decoder retrieves the integers q_m from the binary file and reconstructs coefficient values using

$$\tilde{a}_m = \text{sign}(q_m) \left(|q_m| + \frac{1}{2} \right) T. \quad (6.1)$$

Each nonzero symbol is reconstructed at the center of its quantization bin; zero is reconstructed as zero:



The decoded signal or image is reconstructed as

$$Q_T(f) \stackrel{\text{def}}{=} \sum_{m \in I_T} \tilde{a}_m \psi_m = \sum_m D_T(Q_T(\langle f, \psi_m \rangle)) \psi_m, \quad I_T = \{m : |a_m| \geq T\}, \quad (6.2)$$

where $D_T(q) = T \text{sign}(q)(|q| + 1/2)$ and $\text{sign}(0) = 0$. This produces a reconstruction error $\|f - Q_T(f)\|$.

Let $M = \#I_T$ and let f_M retain the coefficients indexed by I_T without quantizing them. This is the approximation (5.3), with equality at the threshold retained here. Orthogonality gives $\|f - f_M\| \leq \|f - Q_T(f)\|$. The following bound separates the approximation error from the additional error caused by rounding.

Proposition 6.1. *With this choice of f_M ,*

$$\|f - Q_T(f)\|^2 \leq \|f - f_M\|^2 + MT^2/4 \quad \text{where} \quad M = \#\{m : \tilde{a}_m \neq 0\}. \quad (6.3)$$

Proof. The reconstruction rule (6.1) gives, for $|a_m| \geq T$,

$$|a_m - \tilde{a}_m| \leq \frac{T}{2}.$$

By orthogonality,

$$\|f - Q_T(f)\|^2 = \sum_m (a_m - \tilde{a}_m)^2 \leq \sum_{|a_m| < T} |a_m|^2 + \sum_{|a_m| \geq T} \left(\frac{T}{2}\right)^2,$$

The first sum is $\|f - f_M\|^2$, and the second contains M terms, proving the bound. $\square$

6.1.3 Support Coding

To express the error bound (6.3) in terms of a bit budget, we must relate M and T to the cost of encoding the coefficients.

At high compression ratios, only a small number M of the N coefficients remain nonzero. Their support

$$I_M = \{m : \tilde{a}_m \neq 0\}$$

can be encoded separately from the nonzero values $q_m \neq 0$ for $m \in I_M$.

The following theorem gives a rate bound for this support-and-value coding strategy.

Theorem 6.2. Assume that $f \in L^2$ has best M -term approximation error $\|f - f_M\|^2 \leq CM^{-\alpha}$ for some $\alpha > 0$. Suppose that, for each sufficiently small $T > 0$, all coefficients of magnitude at least T occur among a known first $N(T)$ basis elements, with $N(T) \leq C_0 T^{-\gamma}$ for some $\gamma > 0$. Then $\mathbf{Q}_T(f)$ admits a binary description of length $R(T)$ such that, as $R(T) \rightarrow \infty$,

$$\|f - \mathbf{Q}_T(f)\|^2 = O\left(\left(\frac{\log R(T)}{R(T)}\right)^\alpha\right).$$

The decoder is assumed to know the basis and quantization step; describing these parameters adds any required finite header.

Proof. **Coefficient and quantization bounds.** By Proposition 5.1, the ordered magnitudes satisfy

$$d_m \leq C_1 m^{-(\alpha+1)/2}. \quad (6.4)$$

Put $K = T^{-2/(\alpha+1)}$. At most $M = O(K)$ coefficients have magnitude at least T . Summing the squared errors of the zero and nonzero quantization cells gives

$$\|f - \mathbf{Q}_T(f)\|^2 \leq \sum_{m \geq 1} \min(d_m^2, T^2) \leq [K]T^2 + \sum_{m > [K]} d_m^2 = O(K^{-\alpha}). \quad (6.5)$$

An upper approximation bound is enough for this estimate.

Finite computation. The finite-access assumption is

$$\forall m \geq N(T), \quad |\langle f, \psi_m \rangle| < T. \quad (6.6)$$

For a bounded signal on the unit cube with periodic boundary conditions, using periodized compactly supported d -dimensional wavelets, the coefficient bound is $O(2^{jd/2})$ at fine scales $j \rightarrow -\infty$. There are $O(2^{-jd})$ coefficients at scales coarser than J . Retaining scales down to J with $2^{Jd/2}$ proportional to T therefore suffices, and one may choose

$$N(T) = O(T^{-2}). \quad (6.7)$$

More generally, the hypothesis gives

$$N(T) = O(K^{\gamma(\alpha+1)/2}). \quad (6.8)$$

The ordering here is a fixed computable ordering of basis elements, not the unknown ordering by coefficient magnitude.

Support and value coding. Encode each of the M nonzero indices with $\lceil \log_2 N \rceil$ bits:

$$R_{\text{ind}} = M \lceil \log_2 N \rceil = O(K \log K). \quad (6.9)$$

Since $|a_m| \leq \|f\|_2$, every quantized value lies in $[-A, A] \cap \mathbb{Z}$ with $A \leq \|f\|_2/T$. Thus

$$R_{\text{val}} \leq M \lceil \log_2(2A + 1) \rceil = O(K \log K). \quad (6.10)$$

A header encoding M , N , and the integer amplitude bound A has logarithmic size and is absorbed in this bound. Consequently,

$$R = O(K \log K). \quad (6.11)$$

Because the inverse of $x \mapsto x \log x$ is asymptotic to $r/\log r$, this implies

$$K \geq cR/\log R \quad (6.12)$$

for large R . Combining this with (6.5) proves the claim. $\square$

The theorem links compression rates to nonlinear approximation rates and motivates choosing a basis adapted to the signal model Θ . A practical compression algorithm operates on a discrete signal or image of size N . It computes N transform coefficients $\{\langle f, \psi_m \rangle\}_{0 \leq m < N}$ from these data. Section 4.4 describes the discrete wavelet transform and gives a compatibility condition (4.5) under which these discrete coefficients coincide with the continuous inner products.

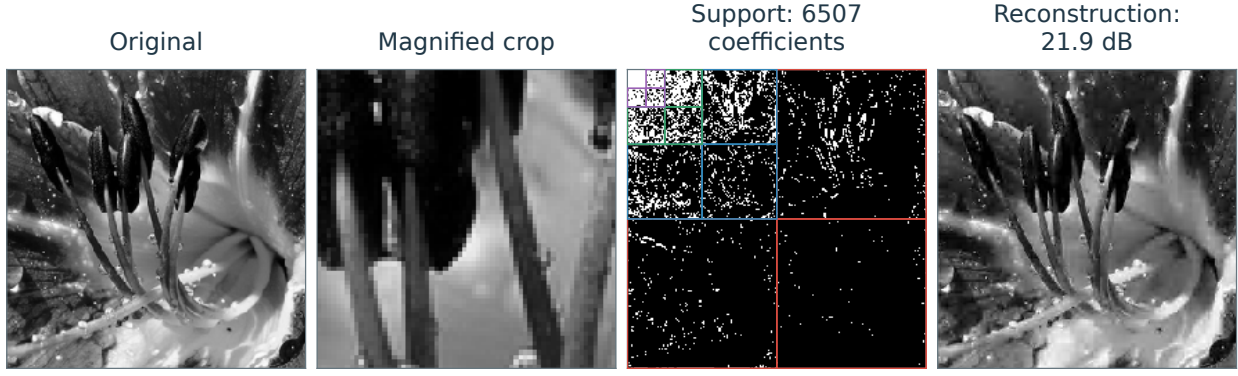


Figure 6.2. Image compression using wavelet support coding.

6.2 Entropy Coding

Entropy coding exploits the uneven distribution of the quantized coefficients to reduce the expected file size: zeros are frequent, while large values are rare. Shannon’s coding theory [1] quantifies the resulting savings.

We refer to Section 3.1 for the theoretical foundations of the methods described below.

Probabilistic modeling. The quantized coefficients $q_m \in \{-A, \dots, A\}$ take values in an alphabet of $Q = 2A + 1$ elements. A code maps the coefficient sequence to a binary string:

$$\{q_m\}_m \mapsto \{0, 1, 1, \dots, 0, 1\} \in \{0, 1\}^R.$$

We first model the q_m as independent draws from a known probability distribution:

$$\mathbb{P}(q_m = i) = p_i \in [0, 1].$$

Huffman code. A Huffman code assigns a variable-length binary string to each symbol:

$$q_m = i \in \{-A, \dots, A\} \mapsto c_i \in \{0, 1\}^{|c_i|}$$

where $|c_i|$ is the codeword length. Codeword lengths are assigned so that a less probable symbol receives no shorter a codeword than a more probable symbol. Huffman coding minimizes the expected length among binary prefix codes. Shannon’s construction gives lengths $\lceil -\log_2 p_i \rceil$ for positive-probability symbols, so optimality of Huffman’s code implies

$$\mathbb{E}R = N \sum_i p_i |c_i| \leq N(\mathcal{E}(p) + 1),$$

where $\mathcal{E}$ is the entropy of the distribution, defined as

$$\mathcal{E}(p) = - \sum_i p_i \log_2(p_i).$$

Figure 6.3 illustrates how entropy depends on the distribution. Concentrated distributions have low entropy, as is often the case for quantized wavelet coefficients because many are zero.

Symbol-by-symbol Huffman coding uses an integer number of bits per symbol. This constraint can cause a substantial gap above entropy when one symbol is very probable. Arithmetic coding encodes whole sequences and can achieve an expected length close to $N\mathcal{E}(p)$, with a small overhead, under the assumed probability model.

6.3 JPEG-2000

JPEG-2000 is a family of still-image compression standards based on wavelet transform coding and adaptive entropy coding. Its irreversible transform uses the biorthogonal CDF 9/7 filters; its reversible transform uses

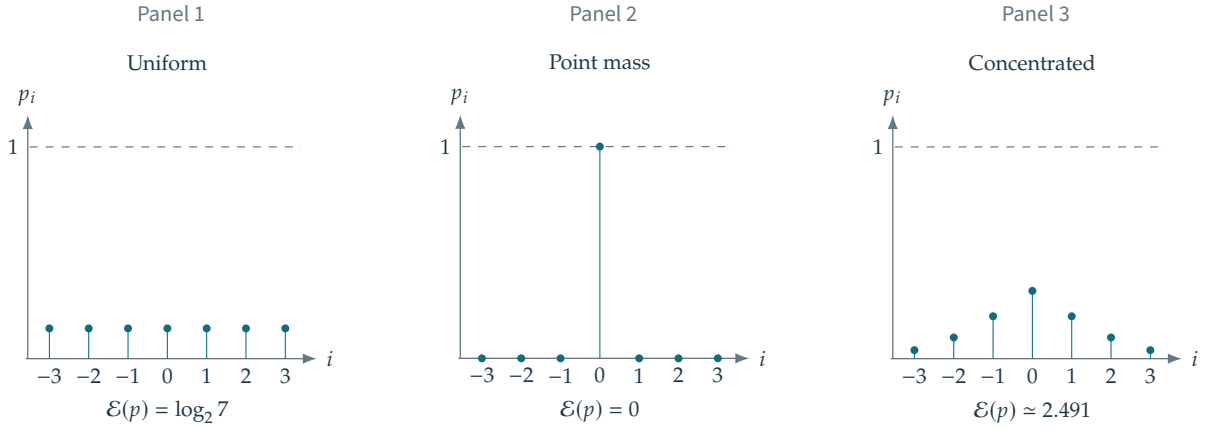


Figure 6.3. Three different probability distributions on the same alphabet of $Q = 7$ symbols. Entropy is measured in bits.

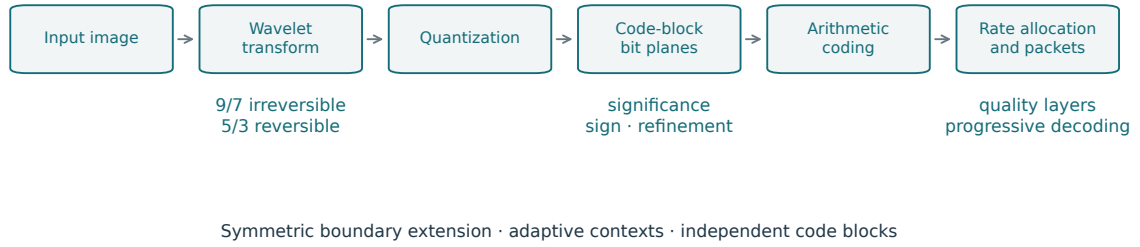


Figure 6.4. JPEG-2000 coding architecture.

integer-to-integer 5/3 lifting for lossless coding. Symmetric boundary extension limits edge artifacts. Because the transform is biorthogonal, the orthonormal energy identities above are replaced by bounds involving the synthesis operator.

Figure 6.4 outlines the JPEG-2000 architecture, and Figure 6.5 compares its reconstructions with those of JPEG. JPEG uses a blockwise DCT and can exhibit visible block boundaries at low bit rates. JPEG-2000 uses a wavelet transform and supports features such as regions of interest, which allow selected parts of an image to be encoded more accurately.

Progressive coefficient refinement. Bit-plane coding progressively reveals the binary digits of the quantized coefficient magnitudes. We describe this refinement schematically using thresholds $T_i = 2^{-i}T_0$: each new bit plane resolves coefficient values at a finer scale.

Rate allocation and quality layers. The coder processes coefficient blocks S_k independently, producing embedded streams c_i^k . The encoder selects coding-pass truncation points within these streams to balance rate against estimated distortion, then organizes the retained data into quality layers. Decoding additional layers improves image quality. This construction supports progressive transmission, with truncation at valid codestream boundaries. The allocation uses an additive distortion model; it does not guarantee the smallest reconstruction error for every possible bit budget.

Bit-plane coding. At threshold T_i , consider a coefficient $d_j^\omega[n]$ at scale j , orientation $\omega \in \{V, H, D\}$, and position $n \in S_k$. Three types of bits describe its significance, sign, and magnitude. We suppress (j, ω) in the notation for these bits.

- If $|d_j^\omega[n]| < T_{i-1}$, the coefficient was not significant at bit-plane $i - 1$. The encoder writes a significance bit $b_i^1[n]$ indicating whether $|d_j^\omega[n]| \geq T_i$.

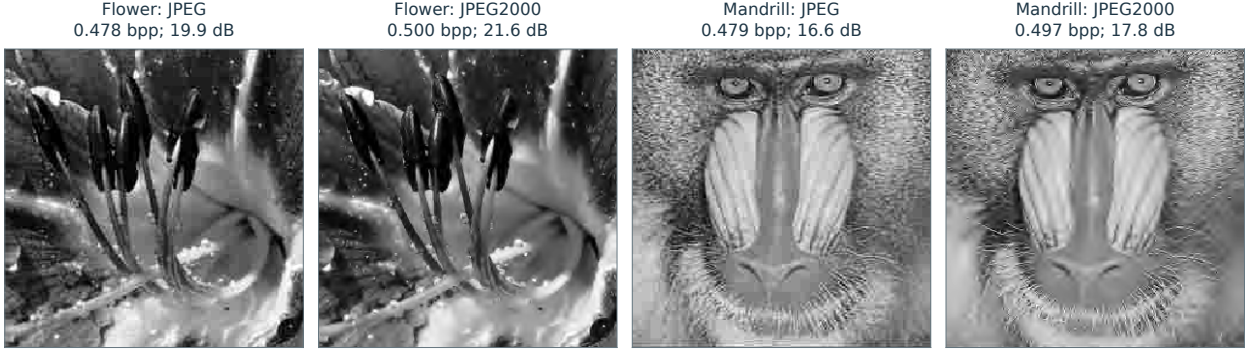


Figure 6.5. JPEG and JPEG-2000 reconstructions of the flower (first pair) and mandrill (second pair), at matched target bit rates.

- If $b_i^1[n] = 1$, the coefficient has just become significant, and the encoder writes its sign as a bit $b_i^2[n]$.
- For every position n that was previously significant, meaning $|d_j^\omega[n]| \geq T_{i-1}$, the encoder writes a refinement bit $b_i^3[n]$ giving the next binary digit of the coefficient magnitude, namely $\lfloor |d_j^\omega[n]|/T_i \rfloor \bmod 2$.

Context modeling. The bits $\{b_i^s[n]\}_{s=1}^3$ for $n \in S_k$ are compressed using context-adaptive arithmetic coding to form the streams c_i^k . Contexts exploit spatial dependence, especially near edges, where large wavelet coefficients tend to cluster. The scan proceeds through stripes four samples high, visiting each stripe column by column.

For a coefficient at position $n \in S_k$, the context index $v_i^s[n]$ summarizes its coding state and those of coefficients in its 3×3 neighborhood

$$w_n = \{(n_1 + \varepsilon_1, n_2 + \varepsilon_2) : \varepsilon_1, \varepsilon_2 \in \{-1, 0, 1\}\}.$$

The relevant states include whether neighboring coefficients have become significant, their known signs, and the coefficient's refinement history. Only information already available to the decoder may enter the context. The precise context rules differ for significance, sign, and refinement bits.

An adaptive arithmetic coder estimates the conditional probability $\mathbb{P}(b_i^s[n] | v_i^s[n])$ and uses it to encode $b_i^s[n]$. Informative contexts can reduce the conditional entropy and hence the expected number of bits.

References

- [1] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.