Mathematical Foundations of Data Sciences

Gabriel Peyré
CNRS & DMA
École Normale Supérieure
gabriel.peyre@ens.fr
https://mathematical-tours.github.io
www.numerical-tours.com

November 18, 2020

Chapter 5

Compression

5.1 Transform Coding

5.1.1 Coding

State of the art compression schemes correspond to transform coders, that code quantized coefficients in an ortho-basis. They first computes the coefficients of the decomposition of the signal into an well-chosen basis (for instance wavelets)

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

Quantization corresponds to rounding the coefficients to an integer using a step size $T > 0$

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

We note that this quantizer has a twice larger zero bin, so that coefficients in $[-T, T]$ are set to zero.

This quantizer nonlinearity should be compared to the hard thresholding nonlinearity (4.4) used to perform non-linear approximation. The quantizer not only set to zero small coefficients that are smaller than T in magnitude, it also modifies larger coefficients by rounding, see Figure 5.1.

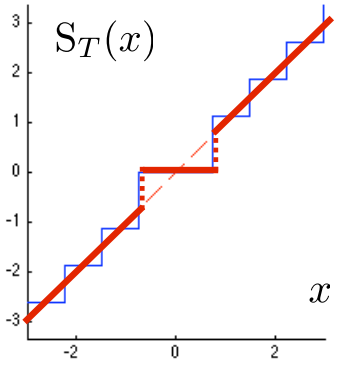


Figure 5.1: Thresholding and quantization non-linearity mappings.

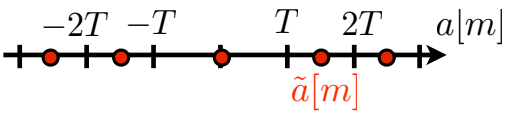
The resulting integer values $(q_m)_m$ are stored into a binary file of length R, which corresponds to a number of bits. Sections 5.1.3 and 5.2 detail two different approach to perform this transformation from integer to bits. The goal is to reduce as much as possible the number R of bits.

5.1.2 De-coding

The decoder retrieves the quantized coefficients q_m from the binary file, and dequantizes the coefficients using

$$\tilde{a}_m = \mathrm{sign}(q_m)\left(|q_m| + \frac{1}{2}\right)T. \tag{5.1}$$

This corresponds to retrieving the value from quantization at the center of quantization bins:



The compressed-decompressed image is then reconstructed as

$$\mathcal{Q}_T(f) \overset{\text{def.}}{=} \sum_{m \in I_T} \tilde{a}_m \psi_m = \sum_{m \in I_T} Q_T(\langle f, \psi_m \rangle)\psi_m,$$

thus producing a decompression error $\|f - \mathcal{Q}_T(f)\|$.

This decompression reconstruction (5.1.2) should be compared with the non-linear approximation formula (4.3). One necessarily has $\|f - f_M\| \leqslant \|f - \mathcal{Q}_T(f)\|$, but in practice these two errors have comparable magnitudes.

Proposition 20. *One has*

$$\|f - \mathcal{Q}_T(f)\|^2 \leqslant \|f - f_M\|^2 + MT^2/4 \quad \textit{where} \quad M = \#\left\{m \,;\, \tilde{a}_m \neq 0\right\}. \tag{5.2}$$

Proof. Indeed, the de-quantization formula (5.1) implies that for $|a_m| > T$,

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

One thus has

$$\|f - \mathcal{Q}_T(f)\|^2 = \sum_m (a_m - \tilde{a}_m)^2 \leqslant \sum_{|a_m| < T} |a_m|^2 + \sum_{|a_m| \geqslant T} \left(\frac{T}{2}\right)^2,$$

which leads to the desired bound. □

5.1.3 Support Coding

To measure how large is the additional error term $MT^2/4$ in (5.2), one needs to choose a method to store the quantized coefficients q_m into a file.

For aggressive compression scenario, where R and M are small with respect to the size N of the image, the support

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

is highly sparse. It thus make sense to code first this support and then to code the actual value $q_m \neq 0$ for $m \in I_M$.

The remaining of this section proves the following theorem.

Theorem 7. *We assume $\|f - f_M\|^2 \sim M^{-\alpha}$ where f_M is the M-term non-linear approximation of f in $\{\psi_m\}_m$. We also assume that the required number of discrete samples N used can be bounded polynomially with N (see (5.7) below for more details). Then for all T there exists a coding strategy of $\mathcal{Q}_T(f)$ using $R = R(T)$ bits such that*

$$\|f - \mathcal{Q}_T(f)\|^2 = O(\log^\alpha(R) R^{-\alpha}).$$

Proof. This proof is split in several parts.

Signals constraints. First, let us notice that, thanks to Proposition 16 the error decay hypothesis $\|f - f_M\|^2 \sim M^{-\alpha}$ is equivalent to imposing a fast decay of the ordered coefficients d_m defined in (4.1) satisfies

$$d_m \sim m^{-\frac{\alpha+1}{2}} \quad \Longrightarrow \quad T \sim M^{-\frac{\alpha+1}{2}}. \tag{5.3}$$

Thanks to Proposition 20, one thus has

$$\|f - f_M\|^2 \leqslant \|f - \mathcal{Q}_T(f)\|^2 \leqslant \|f - f_M\|^2 + MT^2/4 \sim M^{-\alpha} \sim \|f - f_M\|^2, \tag{5.4}$$

which shows that the compression error is comparable with the approximation error.

Discrete computation and scaling of N. For the compression from the discrete signals to be the same as a compression of a continuous signal, we impose that N is large enough so that

$$\forall\, m \geqslant N, \quad |\langle f, \psi_m \rangle| < T \tag{5.5}$$

so that the coefficients not quantized to 0 are contained within the set $\{\langle f, \psi_m \rangle\}_{0 \leqslant m < N}$ of the N computed coefficients. For instance, if f is bounded and one considers a wavelet basis, (19) ensures that $|\langle f, \psi_m \rangle| = O(2^{-jd/2})$, and thus choosing $T \sim 2^{-j_{\max} d/2}$ ensures (5.5) where $N = 2^{-j_{\max} d}$, i.e.

$$N = O(1/T^2). \tag{5.6}$$

The other hypothesis beyond (5.3) of Theorem 7 is that the number N of required discrete samples is not too large, and in particular, that there is a polynomial grows of N with respect to the number M of coefficients to code

$$N = O(M^\beta) \tag{5.7}$$

for some $\beta > 0$. For instance, under the decay condition (5.3) and the worse case scaling (5.6) for bounded function in a wavelet basis,

$$N = O(1/T^2) \quad \text{and} \quad T \sim M^{-\frac{\alpha+1}{2}} \quad \Longrightarrow \quad N = O(M^{\alpha+1}).$$

Support coding. One can simply encode the M indexes of each element $m \in I_M \subset \{1, \ldots, N\}$ using $\log_2(N)$ bits, so that the total number of bits for the support I_M is

$$R_{\text{ind}} = M \log_2(N) = O(M \log_2(M)) \tag{5.8}$$

where we used (5.3).

Values coding. The quantized values satisfy $q_m \in \{-A, \ldots, A\}$, with

$$A \leqslant \frac{1}{T} \max_m |\langle f, \psi_m \rangle| = O(T^{-1}),$$

so one can code them using a number of bits

$$R_{\text{val}} = O(M |\log_2(T)|) = O(M \log_2(M)) \tag{5.9}$$

where we have used hypothesis (5.3) that implies $|\log_2(T)| \sim \log_2(M)$.

Total number of bits. Putting (5.8) and (5.9) together, the total number of bits for this support coding approach is thus

$$R = R_{\text{ind}} + R_{\text{val}} = O(M \log_2(M)). \tag{5.10}$$

The function $\varphi(m) = m \log_2(m)$ is strictly increasing, one can invert it, and we now show that $\varphi^{-1}(r) = r/\log(r) + o(1)$ for large r. Indeed, writing $r = \varphi(m)$

$$m = \frac{r}{\log_2(m)} = \frac{r}{\log_2(r) - \log_2 \log_2(m)} \sim \frac{r}{\log_2(r)}$$

where we used the fact that since $m \leqslant r$, $\log_2 \log_2(m) = o(\log_2(r))$. Inverting this relationship (5.10) thus proves that

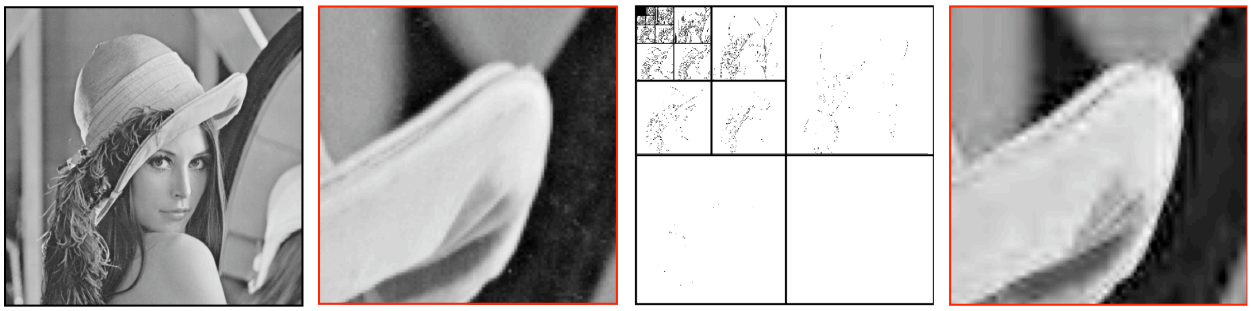
$$M \geqslant C \frac{R}{\log_2(R)} \tag{5.11}$$

for some constant C. Using (5.4) and (5.11), one thus finally arrives to

$$\|f - \mathcal{Q}_T(f)\|^2 \sim M^{-\alpha} = O(\log^{\alpha}(R) R^{-\alpha}).$$

□

This theorem shows the importance of the study of non-linear approximation, and in particular the design of bases that are efficient for approximation of a given signal model Θ. Let us also insists on the fact that a practical compression algorithm is only capable of dealing with discrete signals of size N. We thus considers that the algorithm has access to N inner products $\{\langle f, \psi_m \rangle\}_{1 \leqslant m < N}$ that are computed using a decomposition algorithm from a discretized signal or image of size N. For instance, Section **??** details a discrete wavelet transform, and introduces a compatibility condition (**??**) on the sampling operator for this inner product computation to be possible from discrete data.



Original f | Zoom | Wavelet support | Compressed $\mathcal{Q}_T(f)$

Figure 5.2: Image compression using wavelet support codding.

5.2 Entropic Coding

To further reduce the file size R (in bits), one can use an entropic coder to transform the integer values q_m into bits. Such coding scheme makes use of the statistical redundancy of the quantized values, that have a large number of zero entries and very few large entries. The theory of coding was formalized by Shannon [7].

We refer to Section 1.3 for the theoretical foundation associated to what we now describe.

Probabilistic modeling. The quantized coefficients $q_m \in \{-A, \ldots, A\}$ are assumed to take values in an alphabet of $Q = 2A + 1$ elements. A coding scheme performs the transformation

$$\{q_m\}_m \longmapsto \{0, 1, 1, \ldots, 0, 1\} \in \{0, 1\}^R.$$

To reduce the average value of R, one makes use of a statistical model, which assumes that the q_m are drawn independently at random from a known probability distribution

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

Huffman code. A Huffman code is a code with variable length, since it perform a mapping from symbols to binary strings

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

where $|c_i|$ is the length of the binary code word c_i, that should be larger if p_i is small. A Huffman tree algorithm is able to build a code such that

$$|c_i| \leqslant \lceil \log_2(p_i) \rceil$$

so that

$$R \leqslant (\mathcal{E}(p) + 1)N$$

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

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

Figure 5.3 shows different probability distribution. The entropy is small for highly sparse distribution. Wavelet coefficients of natural images tend to have a low entropy because many coefficients are small.

The Huffman scheme codes symbols independently, leading to a sub-optimal code if some of the p_i are large, which is usually the case for wavelet coefficients. One usually prefers arithmetic coding schemes, that codes groups of symbols, and are able to get close to the entropy bound $R \approx \mathcal{E}(p)N$ for large N.

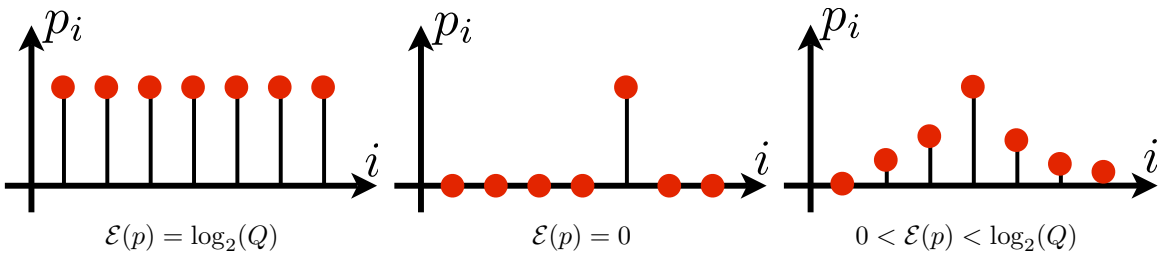


Figure 5.3: Three different probability distributions.

5.3 JPEG-2000

JPEG-2000 is the latest still image compression standard. It corresponds to a wavelet transform coder that performs a clever adaptive entropy coding that makes use of the statistical redundancy of wavelet coefficients of natural images. The wavelet transform is not orthogonal, it is a symmetric 7/9 biorthogonal, with symmetric boundary condition and a lifting implementation. This transform is however close to orthogonality, so that the previous discussion about orthogonal approximation and coding is still relevant.

Figure 5.4 shows an overview of JPEG-2000 architecture. Figure 5.5 shows a comparison between JPEG and JPEG-2000 compressors. JPEG is based on a local DCT transform, and suffers from blocking artifacts at low bit rates, which is not the case of JPEG-2000. This new standard also comes with several important features, such as regions of interest, which allows to refine the coding in some specific parts of the image.

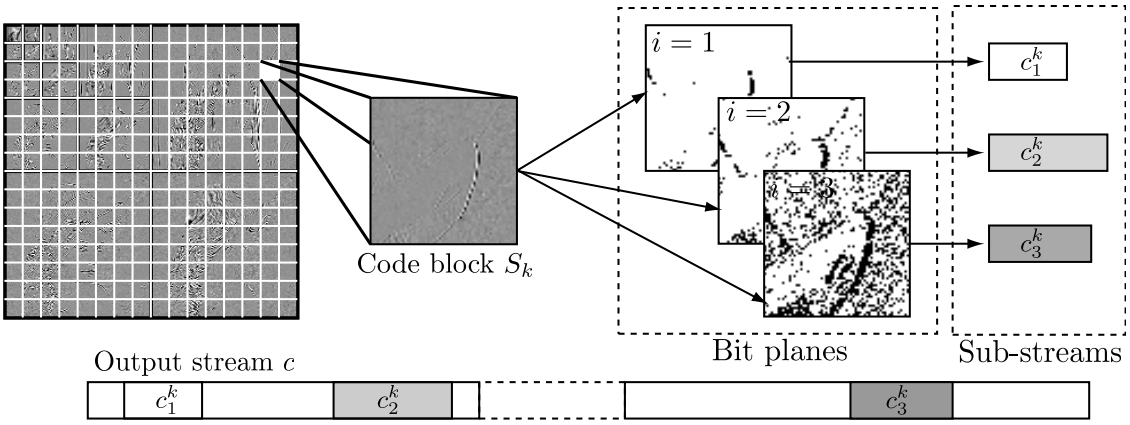


Figure 5.4: JPEG-2000 coding architecture.

Dyadic quantization. The wavelet coefficients are quantized with a varying quantization step $T_i = 2^{-i}T_0$. This allows one to progressively increase the precision of the coded coefficients. For each i, a bit plane coding pass produces new bits to refine the value of the coefficients when i increases.

Steam packing. The bits obtained using the bit plane pass with quantizer $T_i = 2^{-i}T_0$ are entropy coded using a contextual coder. This coder processes square blocks S_k of coefficients. This local coding enhances the parallelization of the method. This local block coding produces a bit stream c_i^k, and these streams are optimally packed into the final coded file to reduce the distortion $\|f - \mathcal{Q}_T(f)\|$ for almost every possible number R of bits. This stream packing ensures scalability of the output bit stream. It means that one can receive only the R first bits of a large coded file and get a low resolution decoded image $\mathcal{Q}_T(f)$ that has an almost minimal distortion $\|f - \mathcal{Q}_T(f)\|$.

Bit plane coding pass. For each threshold T_i, for each scale and orientation $(j, \omega \in \{V, H, D\})$, for each coefficient location $n \in S_k$, JPEG-2000 coder encodes several bit reflecting the value of the wavelet coefficient $d_j^\omega[n]$. In the following we drop the dependancy on (j, ω) for simplicity.

– If $d_j^\omega[n] < T_{i-1}$, the coefficient was not significant at bit-plane $i-1$. It thus encodes a significance bit $b_i^1[n]$ to tell wether $d_j^\omega[n] \geqslant T_i$ or not.

– If $b_i^1[n] = 1$, meaning that the coefficient has became significant, it codes it sign as a bit $b_i^2[n]$.

– For every position n that was previously significant, meaning $d_j^\omega[n] \geqslant T_{i-1}$, it codes a value refinement bit $b_i^3[n]$ to tell wether $d_j^\omega[n] \geqslant T_i$ or not.

Contextual coder. The final bits streams c_i^k are computed from the produced bits $\{b_i^s[n]\}_{s=1}^3$ for $n \in S_k$ using a contextual coder. The contextual coding makes use of spacial redundancies in wavelet coefficients, especially near edges and geometric singularities that create clusters of large coefficients. The coefficients $n \in S_k$ are traversed in zig-zag order as shown on Figure 5.6.

For each coefficient location $n \in S_k$, the context value $v_i^s[n]$ of the bit $b_i^s[n]$ to code at position x is an integer computed over a 3×3 window

$$w_n = \{(n_1 + \varepsilon_1, n_2 + \varepsilon_2)\}_{\varepsilon_i = \pm 1}.$$

This local context $v_i^s[n]$ integrates in a complicated way the previous bit plane values $\{b_{i-1}^s[\tilde{n}]\}_{\tilde{n} \in w_n}$, and neighboring bits at plane $\{b_i^s[\tilde{n}]\}_{\tilde{n} \in w_n, \tilde{n} \text{ coded}}$ that have already been coded.

The bit value $b_i^s[n]$ is then coded with an arithmetic coding by making use of the conditional probability distribution $\mathbb{P}(b_i^s[n] | v_i^s[n])$. The choice made for the computation $v_i^s[n]$ allows to reduce significantly the

Figure 5.5: Comparison of JPEG (left) and JPEG-2000 (right) coding.

entropy of this conditional probability condition with respect to the original distribution $\mathbb{P}(b_i^s[n])$, thus reducing the overall number of bits.

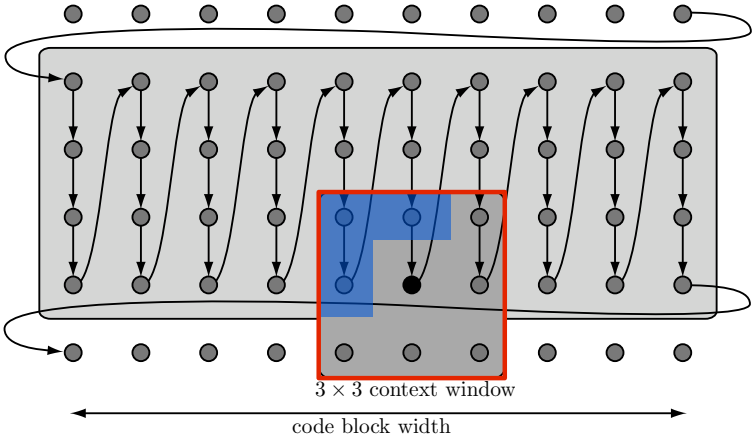


Figure 5.6: Coding order and context for JPEG-2000 coding.

Bibliography

[1] E. Candès and D. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise C^2 singularities. *Commun. on Pure and Appl. Math.*, 57(2):219–266, 2004.

[2] E. J. Candès, L. Demanet, D. L. Donoho, and L. Ying. Fast discrete curvelet transforms. *SIAM Multiscale Modeling and Simulation*, 5:861–899, 2005.

[3] Stephane Mallat. *A wavelet tour of signal processing: the sparse way.* Academic press, 2008.

[4] D. Mumford and J. Shah. Optimal approximation by piecewise smooth functions and associated variational problems. *Commun. on Pure and Appl. Math.*, 42:577–685, 1989.

[5] Gabriel Peyré. *L'algèbre discrète de la transformée de Fourier.* Ellipses, 2004.

[6] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, 1992.

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