Lossy Compression

Chapter 1

Ergodic Processes

1.1 Ergodic Processes

Note: In this class, as in many scientific communities, we define ergodicity of a process in the context of a stationary process.

Definition 1.1.1. *A finite-alphabet stationary process* $\mathbf{X}$ *is ergodic if for all* k *and every* $f : \mathcal{X}^{2k+1} \to \mathbb{R}$*, with* $E|f(X_k^k)| < \infty$*,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(X_{i-k}^{i+k}) \to \mathbb{E} f(X_{-k}^{k}) \qquad w.p.1 \tag{1.1}$$

Intuitivly, ergodic processes are "reliable" in that one process realization (L.H.S. of above equation) reveals all process statistics (R.H.S. of above equation).

Exercise 1.1.2. *Ergodic Examples [1, 4, 9]*

(a) Show X_i *i.i.d.* $\Rightarrow \mathbf{X}$ *is ergodic*

(b) Show that $\mathbf{X} = \left\{ \begin{array}{ll} \ldots 000000000 \ldots & w.p.\frac{1}{2} \\ \ldots 111111111 \ldots & w.p.\frac{1}{2} \end{array} \right\}$ *is not ergodic*

(c) Show that $\mathbf{X} = \left\{ \begin{array}{ll} \ldots 010101010 \ldots & w.p.\frac{1}{2} \\ \ldots 101010101 \ldots & w.p.\frac{1}{2} \end{array} \right\}$ *is ergodic.*

Theorem 1.1.3 (The "Ergodic Decomposition" Theorem)**.** *Every stationary process is a mixture of ergodic processes. That is,* $\forall \mathbf{P}$ *stationary,* $\exists$ *a family of ergodic processes* $\{\mathbf{P}_\theta\}_{\theta\in\Theta}$ *and a probability measure* $\mu(\theta)$ *on* Θ *such that* $\mathbf{P} = \int_\Theta \mathbf{P}_\theta d\mu(\theta)$*.*

Conversely, if $\mathbf{P}$ *is not ergodic, then it can be expressed as a non-trivial mixture of different stationary processes.*

For $\mathbf{X}$, let $\mathbf{X}^{(k)}$ denote the k-th order super source (i.e., $X_1^{(k)} = X_1^k$, $X_2^{(k)} = X_{k+1}^{2k}$).

Definition 1.1.4. $\mathbf{X}$ *is "totally ergodic" if* $\mathbf{X}^{(k)}$ *is ergodic* $\forall k$

Example 1.1.5. *Consider the source*

$$\mathbf{X} = \left\{ \begin{array}{ll} \ldots 01010101 \ldots & w.p. \frac{1}{2} \\ \ldots 10101010 \ldots & w.p. \frac{1}{2} \end{array} \right\}. \tag{1.2}$$

Then, we have

$$\mathbf{X}^{(2)} = \left\{ \begin{array}{ll} \ldots (01)(01)(01)(01) \ldots & w.p. \frac{1}{2} \\ \ldots (10)(10)(10)(10) \ldots & w.p. \frac{1}{2} \end{array} \right\}. \tag{1.3}$$

It is clear that $\mathbf{X}^{(2)}$ *is not ergodic as it is equivalent to the single-letter process*

$$\mathbf{Y} = \left\{ \begin{array}{ll} \ldots 000000000 \ldots & w.p. \frac{1}{2} \\ \ldots 111111111 \ldots & w.p. \frac{1}{2} \end{array} \right\} \tag{1.4}$$

which is not ergodic. Thus, $\mathbf{X}$ *is not totally ergodic.*

Exercise 1.1.6. *Let* $\mathbf{X}$ *be a "B-process" defined as* $X_i = f(Y_{i-k}^{i+k})$ *for i.i.d. process* $\mathbf{Y}$*. Show that* $\mathbf{X}$ *is totally ergodic.*

1.2 Ergodicity: General Setup

We consider a more general setup for ergodicity on the probability space $(\Omega, \mathfrak{F}, \Pr)$. A mapping $\varphi : \Omega \to \Omega$ is *measure-preserving* if $\Pr(\varphi^{-1}(A)) = \Pr(A)$ for all $A \in \mathcal{F}$. A random variable $X = X(\omega)$ is a measurable mapping from Ω to $\mathbb{R}$. An event $A \in \mathcal{F}$ is said to be *invariant* if $\Pr(\varphi^{-1}(A) \triangle A) = 0$, where the notation $A \triangle B = (A \setminus B) \cup (B \setminus A)$ is the symmetric difference of sets A and B.

Exercise 1.2.1. *Show that the class of invariant events* $\mathcal{I}$ *is a* σ*-field.*

Definition 1.2.2. *The mapping* φ *is "ergodic" if* $\mathcal{I}$ *is trivial, that is,* $\Pr(A) \in \{0, 1\}$ *for all* $A \in \mathcal{I}$*.*

Theorem 1.2.3 (The ergodic theorem/Birkhoff's ergodic theorem/Pointwise ergodic theorem[2])**.** *For any* $X \in L_1$*,*

$$\frac{1}{n} \sum_{m=0}^{n-1} X(\varphi^m \omega) \stackrel{n \to \infty}{\longrightarrow} \mathbb{E}[X|\mathcal{I}] \text{ a.s. and in } L_1. \tag{1.5}$$

Since $\mathbb{E}[X|\mathcal{F}] = \mathbb{E}[X]$ *a.s. for all* $X \in L_1$ *if and only if* $\mathcal{F}$ *is trivial, by the ergodic theorem and Definition 1.2.2, the following statements are equivalent.*

(a) For any $X \in L_1$, $(1/n)\sum_{m=0}^{n-1} X(\varphi^m \omega) \overset{n\to\infty}{\longrightarrow} \mathbb{E}[X]$ *w.p.1 and in* L_1.

(b) $\mathcal{I}$ *is trivial.*

(c) φ *is ergodic.*

To define ergodicity for random processes, we consider the probability space, where $\Omega = \mathcal{X}^\infty$, the probability measure Pr is the distribution of the random process, and $\varphi : \Omega \to \Omega$ is defined as a shift in time, namely $\varphi(\omega)_i = \omega_{i+1}$. Stationarity of the process is equivalent here to φ is measure preserving. Then, the above equivalence implies that ergodicity of a stationary process, as per Definition 1.1.1, is essentially equivalent to ergodicity, in the sense of Definition 1.2.2, of the time shift transformation.

Exercise 1.2.4. *For the examples given in Exercise 1.1.2, verify ergodicity or lack thereof using Definition 1.2.2.*

Definition 1.2.2 brings some insight into the ergodic decomposition theorem which states that a process $\mathbf{X}$ is ergodic if and only if it cannot be represented as a mixture of different stationary processes. By definition, if $\mathbf{X}$ is not ergodic then $\mathcal{I}$ defined by φ is not trivial. It follows that there exists an event A such that $\Pr(A) > 0$, $\Pr(A^c) > 0$, $\varphi(A) = A$, and $\varphi(A^c) = A^c$. Essentially, we can divide the sample space Ω into A and A^c, and interpret the random process $\mathcal{X}$ as a mixture of two processes in each subset with probability $\Pr(A)$ and $\Pr(A^c)$. For details we refer to Chapter 6 of [2].

Although there are multiple ways to formally define decaying memory, when a process is "mixing" or has decaying memory, even the weakest definition of a "mixing" process implies ergodicity. The converse is far from true, as demonstrated by part (c) of Exercise 1.1.2.

Chapter 2

Rate Distortion

2.1 Allowable Schemes

Let $\mathbf{X}$ be a stationary ergodic process with a finite alphabet. In lossy compression (also known as lossy source coding), our goal is to encode a source sequence of block length n, X^n, using only nR bits, in order to minimize a given distortion metric between the original source sequence and the reconstruction sequence, Y^n, chosen by the decoder. We assume that our given distortion function $d : (\mathcal{X}, \mathcal{Y}) \to \mathbb{R}_+$ operates bitwise and that the distortion D of a given reconstruction sequence Y^n is given by $D = \mathbb{E}(d(X^n, Y^n)) = \mathbb{E}\frac{1}{n}\sum_{i=1}^{n} d(X_i, Y_i)$.

Definition 2.1.1. *A fixed-rate rate-distortion scheme at rate R for block length n consists of:*

(a) An encoder, $f_n : \mathcal{X}^n \to \{1, \ldots, 2^{nR}\}$

(b) A decoder, $g_n : \{1, \ldots, 2^{nR}\} \to \mathcal{Y}^n$

(c) A reconstruction sequence, $Y^n = g_n(f_n(X^n))$

Figure 2.1: A variable-rate lossy compression scheme

Now we consider a more general class of allowable rate-distortion schemes analogous to the lossless compression setting. Consider the variable-rate lossy compression scheme shown in Figure 2.3. The encoder maps the n-block source X^n with $X_i \in \mathcal{X}$ into bit stream $C_n(X^n)$, which may have different lengths for different source sequences, that is, $C_n : \mathcal{X}^n \to \{0,1\}^*$. The decoder is a mapping $D_n : \{0,1\}^* \to \mathcal{Y}^n$ and it reconstructs the n-block sequence $Y^n = D_n(C_n(X^n))$, where $Y_i \in \mathcal{Y}$. A distortion matrix $d(\cdot,\cdot) : \mathcal{X} \times \mathcal{Y} \to [0, \infty)$ is given. Since the encoder can also reconstruct the sequence Y^n, the above rate-distortion scheme is equivalent to the scheme where the encoder first maps the source sequence

into a reconstruction sequence and then describes the reconstruction sequence losslessly. In the following we formally define the equivalent scheme.

Definition 2.1.2. *A "variable-rate" code for n-blocks is a triple* (B_n, ϕ_n, C_n), *where*

- $B_n \subseteq \mathcal{Y}^n$ *is a codebook consisting of all possible reconstruction sequences,*
- $\phi_n : \mathcal{X}^n \to B_n$, *and*
- $C_n : B_n \to \{0,1\}^*$ *is a U.D. code on* B_n *which describes the reconstruction sequences losslessly.*

The reconstruction sequence is $Y^n = \phi_n(X^n)$. *The number of bits expended is* $l_n(X^n)$, *which is the length of* $C_n(Y^n) = C_n(\phi_n(X^n))$. *The (expected) distortion is* $D = \mathbb{E}\left[d(X^n, Y^n)\right] = \mathbb{E}\left[(1/n)\sum_{i=1}^n d(X_i, Y_i)\right]$, *and the (expected) rate is* $R = \mathbb{E}\left[(1/n) l_n(X^n)\right]$.

Note that fixed-rate schemes are special cases of variable-rate schemes with $C_n : B_n \to \{0,1\}^{nR}$. Now we define achievability and the rate-distortion function.

2.2 The Rate-Distortion Function

Definition 2.2.1. *The pair (R,D) is called achievable if* $\forall \varepsilon > 0$, $\exists n$ *and a rate-distortion scheme at rate* $\leq R + \varepsilon$ *and (expected) distortion* $\leq D + \varepsilon$.

Definition 2.2.2. *The rate-distortion function is defined as* $R(\mathbf{X}, D) = R(D) = \inf\{R' : (R', D)$ is achievable$\}$. *Similarly, we define the "distortion-rate" function as* $D(R) = \inf\{D' : (R, D')$ is achievable$\}$

Exercise 2.2.3. *Show that for* $\mathbf{X}$ *stationary:*

(a)

$$R(D) = \lim_{n\to\infty} \min_{\mathbb{E}(d(X^n, Y^n)) \leq D} \frac{1}{n} H(Y^n) \tag{2.1}$$

(b)

$$R(D) = \inf\{\mathbb{H}(\mathbf{Y}) : \mathbb{E}(d(X_0, Y_0)) \leq D, (\mathbf{X}, \mathbf{Y}) \text{ jointly stationary}\} \tag{2.2}$$

Although the definition of the rate-distortion function above is in terms of a convex optimization problem which is in theory solvable, a much more useful expression for rate-distortion involves the mutual information between the source and output distributions, not only the output distribution's entropy.

Definition 2.2.4. *The informational rate-distortion function $R^{(I)}(D)$ is defined as*

$$R^{(I)}(D) = \lim_{k\to\infty} R_k^{(I)}(D) \tag{2.3}$$

where

$$R_k^{(I)}(D) = \min_{\mathbb{E}(d(X^k,Y^k))\leq D} \frac{1}{k} I(X^k, Y^k). \tag{2.4}$$

Exercise 2.2.5. *Equivalence of rate-distortion and informational rate-distortion functions*

(a) Show that the limit

$$R^{(I)}(D) = \lim_{k\to\infty} R_k^{(I)}(D) \tag{2.5}$$

exists and is equal to $\inf_{k\geq 1} R_k^{(I)}(D)$, $\forall \mathbf{X}$ *stationary*

(b) For $(\mathbf{X}, \mathbf{Y})$ *jointly stationary, define the mutual information rate as*

$$\mathbb{I}(\mathbf{X}, \mathbf{Y}) \triangleq \lim_{n\to\infty} \frac{1}{n} I(X^n, Y^n). \tag{2.6}$$

Show that this limit exists and that

$$R^{(I)}(D) = \inf\{\mathbb{I}(\mathbf{X}, \mathbf{Y}) : \mathbb{E}(d(X_0, Y_0)) \leq D, (\mathbf{X}, \mathbf{Y}) \text{ jointly stationary}\} \tag{2.7}$$

2.3 Rate-Distortion Theory

The main theorem of rate distortion theory for random processes can be state as:

Theorem 2.3.1. *The operational rate distortion function for a stationary ergodic process $\mathcal{X}$ with bounded distortion function $d(x, \hat{x})$ is equal to the associated information rate distortion function. Thus*

$$R(D) = R^{(I)}(D) \tag{2.8}$$

is the minimum achievable rate at distortion D.

Proof *Converse:* We need to show that $R(D) \geq R^{(I)}(D)$. For any scheme with $\mathbb{E}d(X^n, Y^n) \leq D$, let the expected number of bits that this variable length

scheme achieves be nR. Then, since we need to losslessly describe Y^n,

$$\begin{aligned} nR(D) &\geq H(Y^n) && (2.9)\\ &\geq I(X^n;Y^n) && (2.10)\\ &\geq \min_{\mathbb{E}d(X^n,\tilde{Y}^n)\leq D} I(X^n;\tilde{Y}^n) && (2.11)\\ &\triangleq nR_n^{(I)}(D) && (2.12)\\ &\geq n \inf_{k\geq 1} R_k^{(I)}(D) && (2.13)\\ &\triangleq nR^{(I)}(D). && (2.14) \end{aligned}$$

□

Exercise 2.3.2. Achievability: *We need to show that $R(D) \leq R^{(I)}(D)$. This is left as an exercise, but the outline of the proof is given here in three steps.*

(a) Prove $R(D) \leq R_1^{(I)}(D)$. We can use the ideas from rate distortion theory for a memoryless source. If we generate the codebook according to $Q(Y)$ which achieves the minimum in the information rate distortion function, then, with high probability, there is a codeword which is jointly typical with x^n for all typical x^n. The crucial idea is that if X^n is memoryless then it will be a typical sequence with high probability. Now for an ergodic random process $\mathbf{X}$, show that ergodicity implies that X^n is typical with respect to the first order marginal distribution with high probability.

(b) Assuming $\mathbf{X}$ is totally ergodic, prove $R(D) \leq R^{(I)}(D)$.
Hint: *Since $R_k^{(I)}(D)$ for $\mathbf{X}$ is $R_1^{(I)}(D)$ for $\mathbf{X}^{(k)}$, (a) implies (b).*

(c) For an ergodic random process $\mathbf{X}$, prove $R(D) \leq R^{(I)}(D)$.
Hint: *Either prove or look up the following lemma.*

Lemma 2.3.3. *If $\mathbf{X}$ is ergodic then $\mathbf{X}^{(k)}$ can be represented as a mixture of k equiprobable ergodic processes.*

Using this lemma and concavity of the mutual information prove the desired claim. Note that for a set of probability distributions $\{\Pr_{X^k}^{(a)} : a \in \{1,\ldots,k\}\}$, the following is always true.

$$\frac{1}{k}\sum_{a=1}^{k} I(\Pr_{X^k}^{(a)};\Pr_{Y^k|X^k}) \leq I(\frac{1}{k}\sum_{a=1}^{k}\Pr_{X^k}^{(a)};\Pr_{Y^k|X^k}). \quad (2.15)$$

2.4 Restating Rate-distortion

$$\begin{aligned} R(D) &= \inf\{R' \,:\, (R',D) \text{ is achievable } \}\,, && (2.16)\\ D(R) &= \inf\{D' \,:\, (R,D') \text{ is achievable } \}\,. && (2.17) \end{aligned}$$

The rate distortion function $R(D)$ is related to the question: if we are not allowing more than distortion D, what is the minimum rate we can get? For the distortion rate function $D(R)$, the related question is: If we want rate no more than R, what is the minimum distortion we can get? Naturally overall cost function we want to minimize is the linear combination of the rate and the distortion.

$$\min_{\text{all schemes}} \mathbb{E}\Big[\frac{1}{n} l_n(X^n) + \alpha\, d(X^n, Y^n))\Big] \,, \tag{2.18}$$

for some $\alpha \geq 0$, and minimization is over all schemes in the world, lossy or lossless.

Exercise 2.4.1. *Show that for any stationary source*

$$\min_{all\ schemes} \mathbb{E}\Big[\frac{1}{n} l_n(X^n) + \alpha\, d(X^n, Y^n))\Big] \stackrel{n\to\infty}{\longrightarrow} \min_{D\geq 0}\{R(D) + \alpha D\} \tag{2.19}$$

Hint : *First prove that*

$$\mathbb{E}[(\frac{1}{n}) l_n(X^n) + \alpha\, d(X^n, Y^n))] \geq R(\mathbb{E}d(X^n, Y^n)) + \alpha \mathbb{E}d(X^n, Y^n) \tag{2.20}$$

for all n. Then prove the upper bound.

If $R(D)$ is smooth, then the minimum is achieved at $R'(D) + \alpha = 0$, which is called the optimal "fixed-slope" performance. The reason is that the optimal solution achieves the rate distortion performance whose slope is $-\alpha$. In the case when the rate distortion function $R(D)$ is not smooth, there may exist multiple rate distortion pairs (R, D) such that the minimum is achieved.

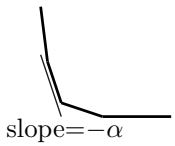

Figure 2.2: Non-smooth $R(D)$

Recall that the operational rate distortion function is defined as

$$R_n(D) = \min_{\mathbb{E}d(X^n, Y^n)\leq D} \frac{1}{n} H(Y^n) \,, \tag{2.21}$$

$$R(D) = \lim_{n\to\infty} R_n(D) \,, \tag{2.22}$$

and the informational rate distortion function is

$$R_k^{(I)}(D) \triangleq \min_{\mathbb{E}d(X^k, Y^k)\leq D} \frac{1}{k} I(X^k; Y^k) \,, \tag{2.23}$$

$$R^{(I)}(D) \triangleq \lim_{k\to\infty} R_k^{(I)}(D) \,. \tag{2.24}$$

Theorem 2.4.2. *Rate distortion theorem:* $R(D) = R^{(I)}(D)$.

What might be confusing is that $R_n(D)$ is what you can achieve with block of length n, while $R_k^{(I)}$ is not something you can achieve with blocklength k. In the proof of Theorem 2.4.2, to achieve $R_k^{(I)}$, we take kth order super symbols and work with a long block code of super symbols. The target rate distortion pair (R, D) is then achieved in the limit. Hence,

(a) $R_n(D)$ is achievable with blocklength n. $R_k^{(I)}$ is in general *not* achievable with blocklength k.

(b) For all k, both $R_k(D)$ and $R_k^{(I)}$ upper bound $R(D)$.

The above theorem states that to compute the operational rate distortion function, we can use the informational rate distortion function. A valid question to ask here is: why is informational rate distortion function any simpler or useful than the original operational rate distortion function? On the face of it, the informational rate distortion function also involves infinite dimensional optimization problem involving X^k for $k \to \infty$. One good answer we have already given in 376A is that for memoryless sources, informational RD collapses into a simple optimization problem:

Exercise 2.4.3. *Show that for a memoryless source, for all* k $R_k^{(I)}(D) = R_1^{(I)}$, *and conlcude that* $R(D) = R_1^{(I)}(D)$.

The informational representation yields explicit characterizations of the rate distortion function way beyond memoryless sources, as we can see in the next section.

2.5 Shannon Lower Bound

For simplicity assume that $\mathcal{X} = \mathcal{Y} = \{0, 1, \ldots, M-1\}$ and that the distortion measure is a *difference distortion measure*, namely $d(x, y) = d(x - y)$. And for aesthetics, assume that $d(\cdot) \geq 0$, and $d(a) = 0$ iff $a = 0$. Also, the additions and subtractions are modulo M. In some cases, we can add symmetric condition $d(-a) = d(a)$, but we won't be needing it in the following. Define the *maximum entropy function* $\Phi_d : \mathbb{R}^+ \to \mathbb{R}^+$:

$$\Phi_d(D) \triangleq \max_{\mathbb{E}d(V) \leq D} H(V) \,, \tag{2.25}$$

where the maximization is over all random variables V taking values in $\{0, \ldots, M-1\}$. Denote the achiever of the maximum entropy function $\Phi_d(D)$ by V_D. Note that the strict concavity of the entropy function guarantees that the achiever is unique (why?).

Exercise 2.5.1. *(a) For the binary case, $\mathcal{X} = \mathcal{Y} = \{0,1\}$, and hamming distortion*

$$d(a) = \begin{cases} 0 & \text{if } a = 0 \\ 1 & \text{otherwise} \end{cases}, \tag{2.26}$$

show that

$$\Phi_d(D) = \begin{cases} h(D) & \text{if } 0 \le D \le 1/2 \\ 1 & \text{if } D > 1/2\,. \end{cases} \tag{2.27}$$

(b) Generalize (a) to $M > 2$ and hamming distortion.

Exercise 2.5.2. *Prove the following properties of $\Phi_d(\cdot)$:*

(a) $\Phi_d(0) = 0$, $\Phi_d(D) = \log(M)$ if $D \ge (1/M)\sum_{i=1}^{M-1} d(i)$.

(b) Monotonic increasing.

(c) Concave.

(d) Strictly increasing in $[0, (1/M)\sum_{i=1}^{M-1} d(i)]$.

(e) In the interval $\left[0, (1/M)\sum_{i=1}^{M-1} d(i)\right]$, the maximum is attained uniquely by V_D with

$$\Pr(V_D = v) = \frac{1}{\sum_{v'=0}^{M-1} e^{-\beta d(v')}} e^{-\beta d(v)}\,, \tag{2.28}$$

where β is the value such that $\mathbb{E}d(V_D) = D$.
Hint*: Use Kullback-Leibler divergence function and the fact it is non-negative and zero if and only if the two distributions are equal.*

Assume now some $P_{Y^n|X^n}$ such that $\mathbb{E}d(X^n, Y^n) \le D$. Letting $X^n - Y^n$ denote the n-tuple $(X_1 - Y_1, \ldots, X_n - Y_n)$, $P_{Y^n|X^n}$

$$\begin{aligned} I(X^n;Y^n) &= H(X^n) - H(X^n|Y^n) && (2.29)\\ &= H(X^n) - H(X^n - Y^n|Y^n) && (2.30)\\ &\ge H(X^n) - H(X^n - Y^n) && (2.31)\\ &\ge H(X^n) - \sum_{i=1}^{n} H(X_i - Y_i) && (2.32)\\ &\ge H(X^n) - \sum_{i=1}^{n} \Phi_d(\mathbb{E}d(X_i - Y_i)) && (2.33)\\ &\ge H(X^n) - n\Phi_d\Big(\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}d(X_i - Y_i)\Big) && (2.34)\\ &\ge H(X^n) - n\Phi_d(D)\,, && (2.35) \end{aligned}$$

where (2.34) follows from concavity of $\Phi_d(\cdot)$ and (2.35) from monotonicity of $\Phi_d(\cdot)$ and the fact that $\mathbb{E}d(X^n, Y^n)$ is upperbounded by D. Hence,

$$R_n^{(I)}(D) \geq \frac{1}{n} H(X^n) - \Phi_d(D) \,, \tag{2.36}$$

which is known as the Shannon Lower Bound for $R_n^{(I)}(D)$. Taking the limit $n \to \infty$ gives

$$R(D) \geq \mathrm{H}(\mathbf{X}) - \Phi_d(D) \,, \tag{2.37}$$

which is the Shannon Lower Bound on $R(D)$. The following exercise explores the condition for the Shannon Lower bound to hold with equality.

Exercise 2.5.3. *Assuming $0 \leq D \leq (1/M)\sum d(i)$, show that*

(a) Equality holds in $R_n^{(I)}(D) \geq (1/n)H(X^n) - \Phi_d(D)$ if and only if there exists Y^n such that

$$N_i \text{ i.i.d.} \sim V_D \tag{2.38}$$

$$\downarrow \tag{2.39}$$

$$Y^n \longrightarrow \bigoplus \longrightarrow X^n \tag{2.40}$$

and in this case (X^n, Y^n) achieves the minimum in the definition of $R_n^{(I)}(D)$.

(b) (X^n, Y^n) uniquely achieve $R^{(I)}(D)$ if the Toeplitz matrix Π induced by $(\Pr(V_D = 0), \ldots, \Pr(V_D = M-1))$ is invertible.

(c) Conclude that, under the same condition on Π, equality in $R(D) \geq \mathrm{H}(\mathbf{X}) - \Phi_d(D)$ holds if and only if there exists a stationary process $\mathbf{Y}$ such that

$$N_i \text{ i.i.d.} \sim V_D \tag{2.41}$$

$$\downarrow \tag{2.42}$$

$$\mathbf{Y} \longrightarrow \bigoplus \longrightarrow \mathbf{X} \tag{2.43}$$

2.6 Geometric Interpretation of the Shannon Lower Bound

The distortion "ball" of radius D is denoted by

$$B_n(y^n, D) \triangleq \{x^n : d(x^n - y^n) \leq D\} \tag{2.44}$$

$$= \{x^n : \frac{1}{n}\sum_{i=1}^{n} d(x_i - y_i) \leq D\} \tag{2.45}$$

$$= \{x^n : x^n - y^n \in B_n(D)\} \,, \tag{2.46}$$

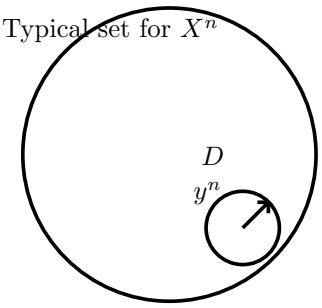

Figure 2.3: Typical set for X^n and distortion ball $B_n(y^n, D)$

for $B_n(D) = \{e_n : (1/n)\sum_{i=1}^n d(e_i) \leq D\}$. Hence,

$$|B_n(y^n, D)| = |B_n(D)| \doteq 2^{n \max_{\mathbb{E}d(V) \leq D} H(V)} = 2^{n\Phi_d(D)} \ . \tag{2.47}$$

Since the union of the distance D ball for all the codewords should at least cover an exponentially non-negligible fraction of the typical set, we have

$$\text{The number of codewords } \geq 2^{n\mathbf{H}(\mathbf{X})}/2^{n\Phi_d(D)} \ , \tag{2.48}$$

which implies that

$$R(D) \geq \mathrm{H}(\mathbf{X}) - \Phi_d(D) \ . \tag{2.49}$$

Exercise 2.6.1. *Assume* $\mathbf{X}$ *is ergodic and* $|\mathcal{X}| < \infty$.

(a) Let $B_n \subseteq \mathcal{X}^n$ *be a sequence of sets with*

$$\limsup_{n\to\infty} \frac{1}{n} \log |B_n| < \mathbf{H}(\mathbf{X}) \ . \tag{2.50}$$

Show that $\Pr(X^n \in B_n \ i.o.\ ) = 0$.
Hint*: Show that* $\Pr(X^n \in B_n)$ *is exponentially small, and apply the Borel-Cantelli lemma.*

(b) Assuming $0 \leq D \leq (1/M)\sum_{i=1}^{M-1} d(i)$*, show that any sequence of schemes of fixed rate* $R \leq \mathbf{H}(\mathbf{X}) - \Phi_d(D)$ *must satisfy*

$$\liminf_{n\to\infty} d(X^n, Y^n) \geq D \ \textit{with probability } 1. \tag{2.51}$$

Hint*: Use the geometric argument above to conclude that the set of source sequences covered to within distortion* $D - \epsilon$ *is exponentially smaller in size than* $2^{n\mathrm{H}(\mathbf{X})}$*, and invoke part (a).*

2.7 A Bit More on the Shannon Lower Bound

Assume distortion measure $d(\cdot)$ and the maximum entropy function $\Phi_d(D) = \max_{\mathbb{E}d(V)\leq D} H(V)$. The Shannon Lower Bound (SLB) states that, for any ergodic process $\mathbf{X}$:

$$R(D) \geq H(\mathbf{X}) - \Phi_d(D). \tag{2.52}$$

We have seen this from a geometric argument. The equality holds if and only if there exists $\mathbf{Y}$, such that the relationship in Fig. 2.4 is satisfied, where N_i is *i.i.d.* additive noise with distribution V_D. In that case, the infimum in

$$R(D) = \inf \{\mathrm{I}(\mathbf{X};\mathbf{Y}) : (\mathbf{X},\mathbf{Y}) \text{ joint stationary and } \mathbb{E}d(X_0,Y_0) \geq D\} \tag{2.53}$$

is achieved by the process pair described in Figure 2.4. But it does not imply that the reconstruction sequence associated with a code that operates close to the rate distortion curve resembles the process $\mathbf{Y}$ that satisfies the relationship in Figure 2.4. To see this clearly, consider the case where $\mathbf{X}$ is a memoryless binary process, X_i is *i.i.d.* $\sim \mathrm{Ber}(p)$. As we know well form EE376A, $R(D) = h(p) - h(D) = H(\mathbf{X}) - \Phi_d(D)$, where $D \leq p \leq 1/2$. I.e., the SLB is satisfied with equality for this source. Indeed, the *i.i.d.* process $\mathbf{Y}$ in Figure 2.5 yields $\mathbf{X}$ when corrupted by a $\mathrm{BSC}(D)$, which is equivalent to modulo 2 addition with the binary random variable achieving the maximum in the definition of $\Phi_d(D)$. But if we take a code such that $\mathbf{Y}$ is *i.i.d.*, then we would have $1/n\, H(Y^n) = H(Y_i)$, whereas for a good code $1/n\, H(Y^n) = I(X_i;Y_i)$!

Figure 2.4: Signal model that achieves equality in (2.52).

Figure 2.5: Signal model for a binary memoryless source achieving $R(D)$.

A non-trivial example for a case where the SLB is tight is the following. Let $\mathbf{X}$ be binary symmetric Markov source, illustrated in Figure 2.6. The SLB gives $R(D) \geq h(p) - h(D)$, where $D \leq p \leq 1/2$. As it turns out, there exists $\mathbf{Y}$ satisfying the relationship in Figure 2.4 if and only if $D \leq D^* \triangleq \left(1 - \sqrt{1 - (\frac{p}{1-p})^2}\right)/2$, but $\mathbf{Y}$ is not of simple form. See [3].

Figure 2.6: Binary symmetric Markov source.

Chapter 3

Universal Lossy Compression

3.1 Yang-Kieffer (Y-K) Codes

Let us first look at a universal lossy compression scheme due to Yang and Kieffer [12]. The codebook is generated by selecting codewords that have Lempel-Ziv (LZ) description length less than nR, *i.e.*, $B_n = \{y^n : l_{\mathrm{LZ}}(y^n) \leq nR\}$. The encoding involves two steps. In the first step, the source sequence X^n is mapped to the reconstruction sequence Y^n using minimum distortion criterion:

$$Y^n = \phi_n(X^n) = \arg \min_{y^n \in B_n} d(X^n, y^n). \tag{3.1}$$

In the second step, Y^n is mapped to its LZ description:

$$C_n\left(\phi_n(X^n)\right) = \text{LZ description of } \phi_n(X^n). \tag{3.2}$$

Note that the codebook is independent not only of any source statistics, but also of the distortion measure! Encoding, of course, depends on the particular distortion criterion used [cf. (3.1)]. By construction, the scheme described has rate $\leq R$, so, on any source, the distortion it achieves satisfies

$$\liminf_{n\to\infty} \mathbb{E}d(X^n, Y^n) \geq D_{\mathbf{X}}(R). \tag{3.3}$$

On the other hand, the following result shows that it universally attains this lower bound.

Theorem 3.1.1 (Yang&Kieffer [12])**.** *For all stationary ergodic source* $\mathbf{X}$*:*

$$\lim_{n\to\infty} \mathbb{E}d(X^n, Y^n) = D_{\mathbf{X}}(R). \tag{3.4}$$

We will prove universality of the fixed-slope version of the Yang-Kieffer scheme, to which we now turn.

3.2 Fixed Slope Version of Y-K Codes

In the fixed slope version of Y-K (YKFS) codes, we select the reconstruction sequence according to

$$\hat{X}^n(x^n) = \arg\min_{y^n} \left(\frac{1}{n} l_{LZ}(y^n) + \alpha \cdot d(x^n, y^n) \right). \tag{3.5}$$

The encoder then describes $\hat{X}^n$ using its LZ description. Denote $l_{\mathrm{YKFS}}(x^n) \triangleq l_{\mathrm{LZ}}\left(\hat{X}^n(x^n)\right)$.

Theorem 3.2.1. *For all stationary ergodic process* $\mathbf{X}$*:*

$$\mathbb{E}\left[\frac{1}{n} l_{\mathrm{YKFS}}(X^n) + \alpha \cdot d(X^n, \hat{X}^n(X^n))\right] \overset{n\to\infty}{\longrightarrow} \min_{D\geq 0}[R(D) + \alpha D]. \tag{3.6}$$

Proof Fix an arbitrary distortion level, an arbitrary $\delta > 0$, and recall

$$R(D) = \inf\left\{\mathrm{H}(\mathbf{Y}) : (\mathbf{X}, \mathbf{Y}) \text{ joint stationary and } \mathbb{E}d(X_0, Y_0) \geq D\right\}. \tag{3.7}$$

Let $\mathbf{Y}^\delta$ be a δ-achiever, *i.e.*, $\mathrm{H}(\mathbf{Y}^\delta) \leq R(D) + \delta$, where $(\mathbf{X}, \mathbf{Y}^\delta)$ are joint stationary and $\mathbb{E}d(X_0, Y_0^\delta) \leq D$. For any k:

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{n} l_{\mathrm{YKFS}}(X^n) + \alpha \cdot d(X^n, \hat{X}^n)\right] \\
&= \mathbb{E}\left[\min_{y^n} \frac{1}{n} l_{\mathrm{LZ}}(y^n) + \alpha \cdot d(X^n, y^n)\right] && (3.8)\\
&\leq \mathbb{E}\left[\frac{1}{n} l_{\mathrm{LZ}}(Y^{n,\delta}) + \alpha \cdot d(X^n, Y^{n,\delta})\right] && (3.9)\\
&\leq \mathbb{E}\left[H_k\left(Y_{-(k-1)}^{n,\delta}\right) + \varepsilon_n^{(k)}\right] + \alpha \cdot \mathbb{E}d(X_0, Y_0^\delta) && (3.10)\\
&\leq H(Y_0^\delta | Y_{-k}^{-1,\delta}) + \varepsilon_n^{(k)} + \alpha \cdot D && (3.11)
\end{aligned}$$

where (3.10) is due to Theorem **??**, and (3.11) is due to part (b) of Exercise **??**. This implies that

$$\limsup_{n\to\infty} \mathbb{E}\left[\frac{1}{n} l_{\mathrm{YKFS}}(X^n) + \alpha \cdot d(X^n, \hat{X}^n)\right] \leq H(Y_0^\delta | Y_{-k}^{-1,\delta}) + \alpha \cdot D. \tag{3.12}$$

Due to the arbitrariness of k, we have

$$\begin{aligned}
\limsup_{n\to\infty} \mathbb{E}\left[\frac{1}{n} l_{\mathrm{YKFS}}(X^n) + \alpha \cdot d(X^n, \hat{X}^n)\right] &\leq \mathrm{H}(\mathbf{Y}^\delta) + \alpha \cdot D && (3.13)\\
&\leq R(D) + \delta + \alpha \cdot D. && (3.14)
\end{aligned}$$

Since equation (3.14) holds for arbitrary $\delta > 0$, we have

$$\limsup_{n\to\infty} \mathbb{E}\left[\frac{1}{n} l_{\mathrm{YKFS}}(X^n) + \alpha \cdot d(X^n, \hat{X}^n)\right] \leq R(D) + \alpha \cdot D \tag{3.15}$$

Finally, because of arbitrariness of D,

$$\limsup_{n\to\infty} \mathbb{E}\left[\frac{1}{n} l_{\mathrm{YKFS}}(X^n) + \alpha \cdot d(X^n, \hat{X}^n)\right] \leq \min_{D\geq 0}[R(D) + \alpha \cdot D] \tag{3.16}$$

which completes the proof when combined with the obvious

$$\liminf_{n\to\infty} \mathbb{E}\left[\frac{1}{n} l_{\mathrm{YKFS}}(X^n) + \alpha \cdot d(X^n, \hat{X}^n)\right] \geq \min_{D\geq 0}[R(D) + \alpha \cdot D] \tag{3.17}$$

□

Remark 3.2.2. *(a) $\mathbf{Y}^\delta$ is introduced since the infimum of (3.7) is not necessarily achieved (and, in fact, can be shown to be achieved only for trivial processes and/or distortion measures and levels).*

(b) This scheme is not practical since (3.5) is hard to compute. Practicality of schemes, and practical schemes, will be discussed in the next section.

3.3 "Practical" Universal Lossy Compression

3.3.1 An Open Question that's Not Quite Open

We start by posing the following question:

Question 3.3.1. *Given a function $f(n)$ satisfying $\lim_{n\to\infty} f(n) = \infty$, and given a rate R, does there exist a (sequence) of scheme(s) with complexity $O(nf(n))$ and rate less than or equal to R that satisfies*

$$\lim_{n\to\infty} \mathbb{E}d(X^n, Y^n) = D(\mathbf{X}, R) \tag{3.18}$$

for every stationary ergodic source $\mathbf{X}$?

Although this question is widely thought to be open, we answer it in the affirmative by constructing such a scheme. See [5] for more on this.

3.3.2 Constructing a "Practical" Universal Lossy Compressor

Let us divide a block of n symbols from $\mathbf{X}$ into length-k sub-blocks:

$$\underbrace{X_1, \ldots, X_k}_{k}, \underbrace{X_{k+1}, \ldots, X_{2k}}_{k}, \ldots, \underbrace{X_{n-k-1}, \ldots, X_n}_{k}. \tag{3.19}$$

Let $\mathcal{N}_k(R)$ represent the set of all rate-R codebooks for blocklength k. The size of this set is then given by

$$|\mathcal{N}_k(R)| = \binom{|\mathcal{Y}|^k}{2^{kR}} \tag{3.20}$$

where we denote the output alphabet by $\mathcal{Y}$.

The encoding scheme we suggest is a three-step algorithm.

(a) Find the codebook $C \in \mathcal{N}_k(R)$ that minimizes the distortion to X^n when applied separately to each of the k-blocks.

(b) Describe C to the decoder.

(c) Use C on each of the sub-blocks and send the codeword indices to the decoder.

3.3.3 Evaluating the Performance

The number of bits expended in step two is $\log(|\mathcal{N}_k(R)|)$. In step three, kR bits are required for each of the n/k subblocks. This results in a rate of

$$R_{\text{scheme}} = \frac{1}{n}\log(|\mathcal{N}_k(R)|) + R. \tag{3.21}$$

The number of operations is dominated by step one. For every candidate codebook C in $\mathcal{N}_k(R)$, each of the n/k subblocks must be encoded in order to determine the distortion. Each of these encodings requires a comparison to 2^{kR} codewords. Therefore, the number of operations scales as

$$\text{ops} \leq |\mathcal{N}_k(R)|\,\frac{n}{k}2^{kR} \tag{3.22}$$

and the number of operations per source symbol scales as

$$\frac{\text{ops}}{n} \leq \frac{1}{k}\,|\mathcal{N}_k(R)|\,2^{kR}. \tag{3.23}$$

The right hand side of this equation is a function only of k, and so we denote it $a(k)$. Similarly, it can be shown that size of memory required per source symbol is bounded by a function $b(k)$.

The sub-block length k must now be chosen as a function of the blocklength n, similarly as we did for the $DUDE$. We choose $k = k_n$ such that the following conditions are satisfied:

(a) $k_n \to \infty$. This is required to approach the rate-distortion function.

(b) $\frac{1}{n}\log\left(|\mathcal{N}_{k_n}(R)|\right) \to 0$. This ensures that the overhead in describing the codebook is asymptotically negligible.

(c) $\max(a(k_n), b(k_n)) = o(f(n))$. This constrains the complexity. Note that we have full control over the rate of growth of $a(k_n)$ and $b(k_n)$, just as we have full control over the rate of growth of $|\mathcal{N}_{k_n}(R)|$.

For such an encoding scheme:

(a) The complexity per source symbol is $o(f(n))$.

(b) For any stationary ergodic source $\mathbf{X}$,

$$\lim_{n\to\infty} \mathbb{E}d(X^n, Y^n) = D(\mathbf{X}, R). \tag{3.24}$$

Exercise 3.3.2. *Prove the second of the above claims.*

3.3.4 Interpretation

The above analysis indicates that the encoding scheme we have constructed achieves the rate-distortion function asymptotically, has a complexity in n arbitrarily close to linear, and works for any stationary ergodic source. However, there are obvious flaws with this construction. Not only is the convergence to $R(D)$ slow, but the algorithm in question is severely lacking in elegance. Compare, for instance, with the elegance of Lempel-Ziv, which seems to compensate for its poor rates of convergence.

The moral of the story? Low-complexity (in blocklength), universal, $R(D)$-achieving lossy codes are *not* the holy grail of rate-distortion theory. Rate of convergence and "cuteness" are also important considerations.

To emphasize ease of implementation, we pose the following (open) question.

Question 3.3.3. *Suppose the source is drawn from a known memoryless distribution, and suppose that an arbitrarily slow rate of convergence is permissible. Under these circumstances, does(do) there exist a (sequence of) scheme(s) achieving $R(D)$ with linear complexity?*

Chapter 4

Universal Compression via MCMC

4.1 Lossy Compression via MCMC

Consider the encoder that, for a given input sequence x^n, assigns the LZ description of $\hat{X}^n$ where

$$\hat{X}^n(x^n) = \arg\min_{y^n} \left[H_k(y^n) + \alpha d(x^n, y^n) \right]. \tag{4.1}$$

Then the number of bits expended is given by $l_{LZ}(\hat{X}^n)$.

After computing this reconstruction, the encoder returns its LZ description. The decoder is simply the LZ decoder. For this scheme, the number of bits expended is the length of the LZ description, $l_{LZ}(\hat{X}^n)$. This scheme differs from the Yang-Kieffer code in that the Yang-Kieffer code optimizes directly for l_{LZ} in the formulation of $\hat{X}$. The following exercise identifies the k that must be chosen so that the resulting scheme is universal.

Exercise 4.1.1. *Identify $k = k_n$ such that*

$$\lim_{n\to\infty} E\left[\frac{1}{n} l_{LZ}(\hat{X}^n) + \alpha d(X^n, \hat{X}^n) \right] = \min_{D>0} \left[R(D) + \alpha D \right] \tag{4.2}$$

for all stationary ergodic sources $\mathbf{X}$.
Hint: by Ziv's inequality, for all k, there exists some positive number $\epsilon_n^{(k)}$ independent of the source sequence, such that

$$\max_{y^n} \left[\frac{1}{n} l_{LZ}(y^n) - H_k(y^n) \right] \leq \epsilon_n^{(k)}. \tag{4.3}$$

Thus,

$$\limsup_{n\to\infty} \max_{y^n} \left[\frac{1}{n} l_{LZ}(y^n) - H_k(y^n) \right] \leq 0. \tag{4.4}$$

Verify that the previous equation holds when we replace k with $k_n = o(\log n)$.

Evaluating $\hat{X}^n$ from (4.1) is, however, computationally hard. We try the following approach: Define an energy function:

$$\Psi(y^n) = H_k(y^n) + \alpha d(x^n, y^n) \tag{4.5}$$

Then the Boltzmann distribution on the sequences y^n is given by

$$P_\beta(y^n) = \frac{1}{Z_\beta} \exp\left(-\beta \Psi(y^n)\right) \tag{4.6}$$

where $Z_\beta = \sum_{y^n} \exp\left(-\beta \Psi(y^n)\right)$ is the normalization factor.

Exercise 4.1.2. *Show that*

$$\lim_{\beta \to \infty} P_\beta \left(\Psi(Y^n) = \min_{y^n} \Psi(y^n) \right) = 1. \tag{4.7}$$

We can now attempt to sample from the distribution P_β using the MCMC algorithm to approximate the optimization (4.1). A Markov Chain Monte Carlo (MCMC) method for sampling from the distribution P_{X^n} can be described as follows:

- Choose an arbitrary starting point x^n.
- Choose $i \sim \text{Unif}\{1, 2, \dots n\}$ and replace x_i by a sample randomly drawn according to $P_{X_i|x^{n\setminus i}}$.
- Iterate the above step.

Exercise 4.1.3. *Assuming a benign condition $P_{X_i|x^{n\setminus i}}(x_i) > 0, \forall x^n, i$, and denoting by $X^{n,(j)}$ the sequence after j iterations of the MCMC algorithm described above, show that*

$$\lim_{j \to \infty} P_{X^{n,(j)}} = P_{X^n}. \tag{4.8}$$

The condition on $P_{X_i|X^{n\setminus i}}$ above can be relaxed. This algorithm is commonly referred to as *"Heat-bath"* algorithm. Refer to [8] for more details.

Evidently, to apply the MCMC procedure for sampling from the Boltzmann distribution, we need to compute

$$P_{\beta, Y_i|Y^{n\setminus i}}(y_i|y^{n\setminus i}) = \frac{Z_\beta^{-1} e^{-\beta \Psi(y^n)}}{\sum_y Z_\beta^{-1} e^{-\beta \Psi(y^{i-1}, y, y_{i+1}^n)}}. \tag{4.9}$$

The Boltzmann distribution is often used in physics since the energy function defined there is commonly separable into a sum of (potential) functions, each involving only a small collection of y_is (whose indices are neighbors on a grid or a graph). Then, the above conditional distribution is easy to compute since it involves only the variables neighboring the ith one. In our case, the energy

function has a non-separable term $H_k(y^n)$, which cannot be decomposed into a sum of functions involving local terms. However, as we shall show below, it is still easy to calculate.

We introduce some notation. Define

$$c_k(y^n, u^k)[u] = \left|\left\{k+1 \le i \le n : y_{i-k}^{k-1} = u^k, y_i = u\right\}\right|. \tag{4.10}$$

We then have

$$H_k(y^k) = \frac{1}{n-k} \sum_{u^k} \left\|c_k(y^n, u^k)\right\|_1 \mathcal{H}(c_k(y^n, u^k)), \tag{4.11}$$

where $\mathcal{H}(v) = H(W)$, and $P_W = \frac{v}{\|v\|_1}$.

Consider a sequence $y_1, y_2, \ldots, y_{i-k}, \ldots, y_{i-1}, y_i, y_{i+1}, \ldots$. We change y_i to y. Evidently, by doing this, we can affect a number of contexts:

- $\{y_{i-k}, \ldots, y_{i-1}\}$: we can affect $c_k(y^n, \{y_{i-k}, \ldots, y_{i-1}\})$.
- $\{y_{i-k+1}, \ldots, y_i\}$: we can affect $c_k(y^n, \{y_{i-k+1}, \ldots, y_i\})$ and $c_k(y^n, \{y_{i-k+1}, \ldots, y\})$.

  $\vdots$

- $\{y_i, \ldots, y_{i+k-1}\}$: we can affect $c_k(y^n, \{y_i, \ldots, y_{i+k-1}\})$ and $c_k(y^n, \{y, \ldots, y_{i+k-1}\})$.

No more than $(2k+1)$ contexts can be affected by changing one y_i. Thus, it is linear in k (and independent of n) to evaluate the denominator $\Psi(y^{i-1}, y, y_{i+1}^n)$ in (4.9), given the numerator $\Psi(y^n)$. Hence, for each iteration of MCMC, the evaluation of $P_{\beta, Y_i|Y^{n\setminus i}}(\cdot)$ is linear in k.

Now let $\hat{X}^{n,\beta,l}(x^n)$ denote the result of MCMC for target distribution P_β after l iterations with starting point x^n. Let $\hat{X}^{n,\beta}$ denote a perfect sample from P_β. Consider:

$$\lim_{n\to\infty} \lim_{\beta\to\infty} \lim_{l\to\infty} \mathrm{E}\left[\frac{1}{n} l_{LZ}(\hat{X}^{n,\beta,l}(X^n)) + \alpha d(X^n, \hat{X}^{n,\beta,l})\right] \tag{4.12}$$

$$= \lim_{n\to\infty} \lim_{\beta\to\infty} \mathrm{E}\left[\frac{1}{n} l_{LZ}(\hat{X}^{n,\beta}(X^n)) + \alpha d(X^n, \hat{X}^{n,\beta})\right] \tag{4.13}$$

$$= \lim_{n\to\infty} \mathrm{E}\left[\frac{1}{n} l_{LZ}(\hat{X}^{n}(X^n)) + \alpha d(X^n, \hat{X}^{n})\right] \tag{4.14}$$

$$= \min_D (R(D) + \alpha D). \tag{4.15}$$

where (4.13) is due to Exercise 4.1.3, and (4.14) is due to Exercise 4.1.2, and (4.15) is because of Exercise 4.1.1.

One can also show [6] that if $\beta = \beta_l$ is constant, that is, letting β increase along with l sufficiently slowly, then

$$\lim_{n\to\infty} \lim_{l\to\infty} \mathrm{E}\left[\frac{1}{n} l_{LZ}(\hat{X}^{n,\beta_l,l}(X^n)) + \alpha d(X^n, \hat{X}^{n,\beta_l,l})\right] = \min_D (R(D) + \alpha D). \tag{4.16}$$

This is known as "simulated annealing".

4.2 MCMC for Continuous Sources

The scheme of the previous section, as is its exhaustive origin, is suitable for a finite reconstruction alphabet. To see how disastrously it might perform when the size of the alphabet is not small relative to the size of the data, consider the following:

Example 4.2.1. *For y^n such that $i \neq j$ implies $y_i \neq y_j$, we have $H_0(y^n) = \log n$, and $H_1(y^n) = 0$.*

The moral of this example is that the reconstruction alphabet needs to be small relative to the size of the data. This, however, does not limit the applicability of the approach as severely as one might initially suspect. There are many continuous sources whose optimal reconstruction alphabet is finite and, in fact, small. For example, in lossy compression of a memoryless source under squared error loss, we have the following [10]:

Example 4.2.2. *Consider a memoryless source $\mathbf{X}$ and squared error distortion $d(x^n, y^n) = \|x^n - y^n\|_2^2$. The Shannon Lower Bound (SLB) for this case assumes the form*

$$R(D) \geq h(X) - \frac{1}{2}\log(2\pi e D) \tag{4.17}$$

where h denotes differential entropy, and $\frac{1}{2}\log(2\pi e D)$ is the maximum differential entropy function with respect to a squared error distortion constraint. Equality is achieved iff there exists Y such that $X = Y + N$, where $N \sim \mathcal{N}(0, D)$ is independent of Y. When the SLB is not tight, the Y that achieves the minimum in $\min_{\mathbb{E}[(\mathrm{X}-\mathrm{Y})^2]\leq \mathrm{D}} I(X;Y)$ is discrete and finite. See Figure 4.1

Thus, for sources that do not satisfy the SLB with equality, which are most sources, the reconstruction alphabet is finite and, at high distortion, can be very small. The MCMC approach is applicable to such sources.

4.3 A Brief Recap of Universal Lossy Compression

The goal of universal lossy compression is to find Y^n such that the energy function

$$\Psi(y^n) = H_k(y^n) + \alpha d(x^n, y^n) \tag{4.18}$$

is minimized. We discussed, in particular, the MCMC approach from sampling a Boltzmann distribution. For each iteration, we consider the computation complexity in updating the energy function, i.e.,

$$\Psi(y^n) \longrightarrow \Psi(y^{i-1}, y, y_{i+1}^n), \tag{4.19}$$

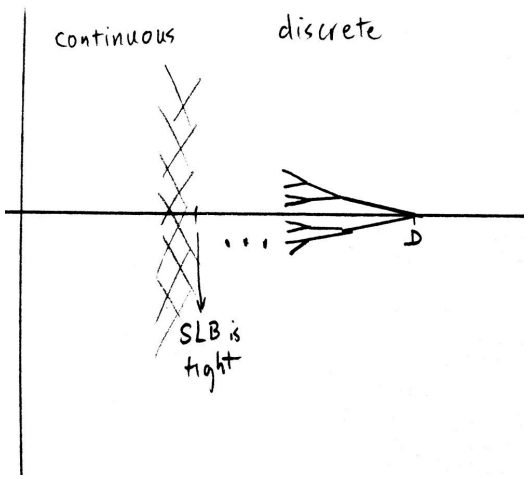

Figure 4.1: Discreteness of $\mathbf{Y}$.

or equivalently,

$$H_k(y^n) \longrightarrow H_k(y^{i-1}, y, y_{i+1}^n), \tag{4.20}$$

is linear in k and independent of n. Thus, to compute the Boltzmann distributions for all iterations, the cost is $O(kl)$, where l is the number of iteration.

Why do we use the energy function instead of the fixed-slope Yang-Kieffer code? The YK code replaces the $H_k(y^n)$ term in the energy function with the length of the LZ description of the reconstruction, $l_{LZ}(y^n)$. As changing one symbol in the reconstruction sequence could produce a global change in its LZ description, the complexity of calculating the YK cost function is $O(n)$. Over all iterations, the cost is $O(n^2)$. Thus it is much more economical to use the energy function.

A different approach, which approximates the exhaustive approach given Ψ, is given in [7].

Chapter 5

The Empirical Distribution of Rate-Distortion Codes

5.1 The Empirical Distribution of Rate-Distortion Codes

For a comprehensive discussion on this topic, see [11]. We first specify the setting for the problem we will discuss. We assume that the source and reconstruction alphabets, $\mathcal{X}$ and $\mathcal{Y}$, are finite. For ease of notation we define

$$I(P_{X,Y}) = I(P_X; P_{Y|X}) = I(X;Y) \tag{5.1}$$

$$I(P_{X^k,Y^k}) = I(P_{X^k}; P_{Y^k|X^k}) = I(X^k;Y^k) \tag{5.2}$$

and without loss of generality, assume that $d(i,j) \geq 0$ for all i, j.

Theorem 5.1.1. *Let* $\mathbf{X}$ *be a stationary random process, and let* $Y^n = Y^n(X^n)$ *be the (fixed block length) reconstruction under some scheme at rate* $\leq R$*. Let* $J \sim \text{Unif}(1, \ldots, n)$*, independent of* $\mathbf{X}$*. Then:*

(a) $I(X_J, Y_J) \leq R + H(X_1) - \frac{1}{n}H(X^n)$.
Notes:
- *For a memoryless source,* $I(X_J, Y_J) \leq R$.
- $P_{X_J,Y_J}(x,y) = \sum_{i=1}^n P_J(i) P_{X_i,Y_i}(x,y) = \frac{1}{n}\sum_{i=1}^n P_{X_i,Y_i}(x,y)$.
- *For* R *large, we can take* $Y = X$.

(b) Append an arbitrary $k-1$ *tuple to* Y^n*, i.e.* $[Y_1, ..., Y_n, y_{n+1}, ..., y_{n+k-1}]$.
$\frac{1}{k}I(X_J^{J+k-1}; Y_J^{J+k-1}) \leq \frac{n}{n-k}R + \frac{1}{k}H(X^k) - \frac{1}{n}H(X^n)$.
Note: $P_{X_J^{J+k-1},Y_J^{J+K-1}} = \frac{1}{n}\sum_{i=1}^n P_{X_i^{i+k-1},Y_i^{i+k-1}}$.

Proof

(a) We have

$$\begin{aligned}
nR &\geq H(Y^n) && (5.3)\\
&\geq I(X^n;Y^n) && (5.4)\\
&= H(X^n) - H(X^n|Y^n) && (5.5)\\
&= H(X^n) - nH(X_1) + \sum_{i=1}^{n} \left[H(X_i) - H(X_i|X^{i-1},Y^n)\right] && (5.6)\\
&\geq H(X^n) - nH(X_1) + \sum_{i=1}^{n} \left[H(X_i) - H(X_i|Y_i)\right] && (5.7)\\
&\geq H(X^n) - nH(X_1) + \sum_{i=1}^{n} I(X_i;Y_i) && (5.8)\\
&= H(X^n) - nH(X_1) + n\sum_{i=1}^{n} \frac{1}{n} I(X_i;Y_i) && (5.9)\\
&\geq H(X^n) - nH(X_1) + nI(P_{X_J};P_{Y_J|X_J}), && (5.10)
\end{aligned}$$

where (5.7) follows since removing conditioning increases entropy, and (5.10) follows since $P_{X_J} = P_{X_i}$ for all $1 \leq i \leq n$, and

$$P_{Y_J|X_J} = \frac{1}{n}\sum_{i=1}^{n} P_{Y_i|X_i}, \tag{5.11}$$

thus via the convexity of $I(X;Y)$ in $P_{Y|X}$, $\sum_{i=1}^{n} \frac{1}{n} I(X_i;Y_i) \geq I(P_{X_J};P_{Y_J|X_J})$.

(b) We provide an outline for this proof. For all $0 \leq j \leq k-1$, define $S_j = \{1 \leq i \leq n : i = j \bmod k\}$. Let $P^{(j)}_{X^k,Y^k} = \frac{1}{|S_j|}\sum_{i\in S_j} P_{X_i^{i+k-1},Y_i^{i+k-1}}$. Using part (a), with the following association:

$$\begin{aligned}
n &\longrightarrow |S_j| && (5.12)\\
X_i, Y_i &\longrightarrow X_i^{i+k-1}, Y_i^{i+k-1} && (5.13)\\
R &\longrightarrow \frac{nR}{|S_j|} && (5.14)
\end{aligned}$$

Now verify that

$$I(P^{(j)}_{X^k,Y^k}) \leq \frac{k}{n-k} nR + H(X^k) - \frac{k}{n} H(X^n). \tag{5.15}$$

Then, via the convexity of mutual information,

$$I(P_{X_J^{J+k-1}};P_{Y_J^{J+k-1}|X_J^{J+k-1}}) \leq \frac{k}{n-k} nR + H(X^k) - \frac{k}{n} H(X^n). \tag{5.16}$$

□

5.2 The Empirical Distribution of Good Codes

Definition 5.2.1. *Let* $\mathbf{X}$ *be ergodic. The code sequence* $\{Y^n(\cdot)\}_{n\geq 1}$ *and rates* $\{R_n\}_{n\geq 1}$ *is* good *for* $\mathbf{X}$ *and rate* R *if*

$$\limsup_{n\to\infty} R_n \leq R, \tag{5.17}$$

$$\limsup_{n\to\infty} \mathbb{E}\left[d(X^n, Y^n(X^n)\right] \leq D(\mathbf{X}, R). \tag{5.18}$$

Exercise 5.2.2. *Show that for* R *in the range where* $D(\mathbf{X}, R)$ *is strictly decreasing, the* lim sup *in the definition can be replaced by* lim.

Theorem 5.2.3 (Empirical Distribution Induced by A Good Code)**.** *Let* $\{Y^n\}$ *be a good code sequence for a stationary ergodic source* $\mathbf{X}$ *at distortion* D*. Let* $J_n \sim \text{Unif}(1, \ldots, n)$*, independent of* $\mathbf{X}$*. Fix some* k*. Assume the following is true:*

(a) $R_k^{(I)}(\mathbf{X}, D) = R(D) + \frac{1}{k}H(X^k) - \mathbb{H}(\mathbf{X})$

(b) $R_k^{(I)}(\mathbf{X}, D)$ *is achieved uniquely by* $(X^k, \tilde{Y}^k)$ *whose joint distribution ahcieves* $R_k^{(I)}(\mathbf{X}, D)$*.*

Then, for fixed k*, as* n *goes to infinity,*

$$\left(X_{J_n}^{J_n+k-1}, Y_{J_n}^{J_n+k-1}\right) \overset{n\to\infty}{\longrightarrow} \left(X^k, \tilde{Y}^k\right) \text{ in distribution.} \tag{5.19}$$

Notes that assumption (a) holds when $\mathbf{X}$ is memoryless or a source that satisfies SLB with equality, i.e.:

$$R_k^{(I)}(D) = \frac{1}{k}H(X^k) - \Phi(D) \tag{5.20}$$

$$= (\mathbb{H}(\mathbf{X}) - \Phi(D)) + \frac{1}{k}H(X^k) - \mathbb{H}(\mathbf{X}) \tag{5.21}$$

$$= R(D) + \frac{1}{k}H(X^k) - \mathbb{H}(\mathbf{X}). \tag{5.22}$$

Also, for a memoryless source where the components of $\tilde{Y}^k$ are i.i.d., we have

$$\frac{1}{k}H(Y_J^{J+k-1}) \overset{n\to\infty}{\longrightarrow} \frac{1}{k}H(\tilde{Y}^k) = H(\tilde{Y}_1), \tag{5.23}$$

while

$$\frac{1}{n}H(Y^n) \overset{n\to\infty}{\longrightarrow} I(X_1; \tilde{Y}_1) = R_1^{(I)}(D) = R(D). \tag{5.24}$$

The proof of Theorem 5.2.3 is as follows. First, note that

$$\lim_{n\to\infty} \mathbb{E}d(X_{J_n}^{J_n+k-1}, Y_{J_n}^{J_n+k-1}) = \lim_{n\to\infty} \mathbb{E}d(X^n, Y^n) = D. \tag{5.25}$$

where we leave the proof of (5.25) as an exercise.

Exercise 5.2.4. *Prove that*

$$\lim_{n\to\infty} \mathbb{E}d(X_{J_n}^{J_n+k-1}, Y_{J_n}^{J_n+k-1}) = \lim_{n\to\infty} \mathbb{E}d(X^n, Y^n) = D. \tag{5.26}$$

Hint: For $k = 1$, it is easy to show that the equality holds for any n, from our per-symbol distortion measure. Next for $k > 1$, show that for any n, $\mathbb{E}d(X_{J_n}^{J_n+k-1}, Y_{J_n}^{J_n+k-1})$ is within $\pm\frac{(k-1)^2}{nk}d_{max}$ from $\mathbb{E}d(X^n, Y^n)$ where $d_{max} = \max_{x\in\mathcal{X},y\in\mathcal{Y}} d(x, y)$.

On the other hand, from Theorem 5.1.1 (b), we know

$$\frac{1}{k}I(X_{J_n}^{J_n+k-1}; Y_{J_n}^{J_n+k-1}) \le \frac{n}{n-k}R_n + \frac{1}{k}H(X^k) - \frac{1}{n}H(X^n), \tag{5.27}$$

where R_n is a rate of the code for the source sequence of block length n, X^n. Then, by the "goodness" of the code,

$$\limsup_{n\to\infty} \frac{1}{k}I(X_{J_n}^{J_n+k-1}; Y_{J_n}^{J_n+k-1}) = R(D) + \frac{1}{k}H(X^k) - \mathbb{H}(\mathbf{X}) \tag{5.28}$$

$$= R_k^{(I)}(D), \tag{5.29}$$

where (5.29) is from Assumption (a). In the meantime, from the definition of the informational rate-distortion function, we know

$$\frac{1}{k}I(X_{J_n}^{J_n+k-1}; Y_{J_n}^{J_n+k-1}) \ge R_k^{(I)}(Ed(X_{J_n}^{J_n+k-1}; Y_{J_n}^{J_n+k-1})), \tag{5.30}$$

which in turn, implies that

$$\liminf_{n\to\infty} \frac{1}{k}I(X_{J_n}^{J_n+k-1}; Y_{J_n}^{J_n+k-1}) \ge \liminf_{n\to\infty} R_k^{(I)}(Ed(X_{J_n}^{J_n+k-1}; Y_{J_n}^{J_n+k-1})) \tag{5.31}$$

$$= R_k^{(I)}(D), \tag{5.32}$$

where the last equality is from (5.25) combined with the continuity of the rate-distortion function $R_k^{(I)}(D)$. Then, by combining (5.29) and (5.32), we can conclude that

$$\lim_{n\to\infty} \frac{1}{k}I(X_{J_n}^{J_n+k-1}; Y_{J_n}^{J_n+k-1}) = R_k^{(I)}(D). \tag{5.33}$$

Then, the rest of the proof will be complete with the following Lemma.

Lemma 5.2.5. *Let $P_{Y|X}$ uniquely attain the minimum in*

$$f(D) = \min_{P_{Y|X}: Ed(X,Y)\le D} I(X; Y) \tag{5.34}$$

and further let $\{P_{Y|X}^{(n)}\}$ satisfy

- $I(P_X; P_{Y|X}^{(n)}) \overset{n\to\infty}{\to} f(D)$

- $E_{P_X P^{(n)}_{Y|X}} d(X,Y) \overset{n\to\infty}{\to} D$,

then $P_X P^{(n)}_{Y|X} \overset{n\to\infty}{\to} P_X P_{Y|X}$

Exercise 5.2.6. *Prove Lemma 5.2.5*

Note that the lemma proves Theorem 5.2.3 by (5.25), (5.33), and the following associations:

- $X^k, \tilde{Y}^k \iff X, Y$
- $P_{X_{J_n}^{J_n+k-1}}, P_{Y_{J_n}^{J_n+k-1}} \iff P^{(n)}_{X,Y}$
- $R^{(I)}_k \iff f(D)$

Denote

$$Q^k_{emp}[X^n, Y^n](x^k, y^k) = \frac{1}{n-k+1} |\{0 \le i \le n-k : X_{i+1}^{i+k} = x^k, Y_{i+1}^{i+k} = y^k\}|, \tag{5.35}$$

the k-th order empirical distribution of (X^n, Y^n). Then under the same assumptions as Theorem 5.2.3,

$$\|Q^k_{emp}[X^n, Y^n] - P_{X^k,\tilde{Y}^k}\| \overset{n\to\infty}{\to} 0 \quad w.p.1 \tag{5.36}$$

See [11] for a proof.

5.3 Applications to Denoising

Assume the following setup:

- Noisefree source: $\mathbf{X}$ stationary, ergodic
- Additive white noise: $N_i \sim N$, i.i.d and independent of $\mathbf{X}$
- Noisy source: $\mathbf{Z}$, $Z_i = X_i \oplus_M N_i$, where we assume that the components of $\mathbf{X}, \mathbf{N}$, and $\mathbf{Z}$ all take values in $\{0, 1, \ldots, M-1\}$

Exercise 5.3.1. *Show that* $\mathbf{Z}$ *is ergodic*

Further assume that all the components of P_N are positive and that the Toeplitz matrix it induces is invertible.

Exercise 5.3.2. *Show that* $\max_V \{H(V) : Ed^{(N)}(V) \le H(N)\}$ *is uniquely attained by* N*, where* $d^{(N)}(a) = \log(\frac{1}{P_N(a)})$ *and the maximization is over all random variables* V *taking values in* $\{0, 1, \ldots, M\text{-}1\}$

Theorem 5.3.3. *Under* $d^{(N)}$*, we have* $R^{(I)}_k(\mathbf{Z}, H(N)) = \frac{1}{k} H(Z^k) - H(N)$*, and is achieved uniquely by* (Z^k, X^k)*.*

Proof Recall the setting of the Shannon lower bound in Section 2.5. Exercise 2.5.2 has shown us that N, the noise, is the unique achiever (in distribution) of $\phi_d(D)$, when d is the distortion measure $d^{(N)}$ and $D = H(N)$. Theorem 5.3.3 now follows directly from parts (a) and (b) in Exercise 2.5.3, with (Z^k, X^k) here playing the role of (X^k, Y^k) in that exercise. □

Thus, the combination of Theorem 5.2.3 with Theorem 5.3.3 implies that by using a good rate-distortion code, our reconstruction is, in effect, a 'sample from the posterior'. More precisely, for any fixed k, the kth-order distribution of the noisy and reconstruction sequences is converging to that of the noisy and *noise-free* sequences.

Bibliography

[1] T. Berger, "Rate-Distortion Theory", Prentice-Hall, 1971

[2] R. Durrett, "Probability: Theory and Examples", 3rd ed., Duxbury Press, 2004.

[3] R. M. Gray, "Information Rates of Autoregressive Processes," *IEEE Trans. Info. Theory*, vol. 16, no. 4, pp. 412–421, 1970.

[4] R. M. Gray, "Entropy and Information Theory", Springer-Verlag, 1990

[5] A. Gupta and S. Verdu and T. Weissman, "Rate-distortion in near-linear time", *IEEE International Symposium on Information Theory*, pp.847-851, 6-11 July 2008.

[6] S. Jalali and T. Weissman, "Rate Distortion Coding of Discrete Sources via Markov Chain Monte Carlo," *Proc. Int. Symp. Info. Theory*, Toronto, Ontario, Canada, July 2008.

[7] S. Jalali, A. Montanari and T. Weissman, "An Iterative Scheme for Near Optimal and Universal Lossy Compression," *Info. Theory Workshop*, Volos, Greece, 2009.

[8] A. Montanari and M. Mezard, "Information, Physics and Computation", http://www.stanford.edu/∼montanar/BOOK/book.html

[9] K. Petersen, "Ergodic theory", Cambridge University Press, 1983

[10] K. Rose. "A mapping approach to rate distortion," *IEEE Trans. Info. Theory*, vol. 40, no. 6, pp. 1939–1952, November 1994.

[11] T. Weissman and E. Ordentlich, "The empirical distribution of rate-constrained codes," *IEEE Trans. Inform. Theory*, vol. 51, no. 11, pp. 3718-3733, November 2005.

[12] E. Yang and J. C. Kieffer, "Simple universal lossy data compression schemes derived from the Lempel-Ziv algorithm," *IEEE Trans. Info. Theory*, vol. IT-42, pp. 239 - 245, January 1996.