CHAPTER 11: UNIVERSAL QUANTUM SOURCE COMPRESSION

§ 11.1 Classical sources and entropy

Consider a classical random variable (RV) X that emits symbols $x \in \{1, \ldots, d\}$ with probability p_x.

Ex.: A (biased) coin gives H with probability $p \in [0,1]$ and T with probability $1-p$.

We assume that we receive a sequence of symbols from X:

$$x^n = (x_1, x_2, \ldots, x_n),$$

where $x_i \in [d]$ for $i = 1, \ldots, n$.

Common assumption: source is independent and identically distributed (iid) or memoryless,

$$\Pr(x^n) = \prod_{i=1}^{n} p_{x_i}$$

Central question in information theory:

How much information do we gain when we learn $x^n = (x_1, \ldots, x_n)$?

Two extreme examples:

a) For a deterministic source with $p_{\hat{x}} = 1$ for some fixed $\hat{x} \in [d]$ and $p_y = 0$ for $y \neq \hat{x}$, we learn nothing new when receiving $x^n = (\hat{x}, \dots, \hat{x})$.

b) For a uniformly random source with $p_x = 1/d$ for all $x \in [d]$, all output sequences x^n are equally probable $(\Pr = \frac{1}{d^n})$, and hence a specific observed sequence x^n conveys a lot of information.

One of Shannon's many contributions:

make these observations quantitative using the concept of entropy.

$\boxed{\text{Def}}$ For a random variable $X \sim p_X$, the surprisal of an event $x \in [d]$ is defined as

$$I(x) := \log \frac{1}{p_x} = -\log p_x .$$

Intuition: the less likely an event, the more information we gain.

The expected surprisal of a source is defined as the (Shannon) entropy of the source:

$\boxed{\text{Def}}$ (Shannon entropy)

The Shannon entropy $H(X)$ of a source $X \sim p_x$ is given by the expected value of surprisals:

$$H(X) = \sum_x p_x I(x)$$

$$= - \sum_x p_x \log p_x .$$

Note that we use the convention $0 \log 0 = 0$ (since $x \log x \to 0$ as $x \to 0$). Hence, if $p_x = 0$ then x has infinite surprisal, but receives no weight in $H(X)$.

Some simple properties of Shannon entropy:

i) $0 \leq H(X) \leq \log d$ for any RV X taking values in $[d]$.

The bounds are saturated by the examples a), b) above: a source X is deterministic iff $H(X) = 0$, and uniform iff $H(X) = \log d$.

ii) Concavity: Let $X_1 \sim p_X$ and $X_2 \sim q_X$ be RV's on the same alphabet, and for $\lambda \in [0,1]$ define an RV $Z = \lambda X_1 + (1-\lambda) X_2 \sim \lambda p_X + (1-\lambda) q_X$.

Then $H(Z) \geq \lambda H(X_1) + (1-\lambda) H(X_2)$.

§ 11.2 Compressing a classical source

Task: Compress the signals $x^n = (x_1, \dots, x_n)$ of an iid. source $X \sim p_X$ without losing information (asymptotically, as $n \to \infty$).

§ 11.1 suggests that information content of a source X is quantified by Shannon entropy $H(X)$.

Shannon's theorem (1948): $H(X)$ is the optimal compression rate!

Idea of source compression:

some output signals of the source occur more frequently (determined by iid prob. dist p^{X^n}), and hence there is redundancy in the information.

Two ways of compression: variable-length and fixed-length variable-length coding: more frequent signals are assigned shorter code words (e.g. Huffman coding)

fixed-length coding: same code word length assigned to all signals (easier to decode)

We focus on fixed-length coding. How do we characterize the "frequent" output signals of a source?

$\boxed{\text{Def}}$ (Typical sequences)

Let $(X_i)_{i\in\mathbb{N}}$ be iid. RV's each taking values in $[d]$ and with common prob. mass function p_x, $x \in [d]$.

For $x^n = (x_1, \dots, x_n) \in [d]^n$ let $p(x^n) := \prod_{i=1}^{n} p_{x_i}$.

Fixing $\varepsilon > 0$, the ε-typical set $T_\varepsilon^{(n)}$ consists of those sequences $x^n \in [d]^n$ for which

$$2^{-n(H(X)+\varepsilon)} \leq p(x^n) \leq 2^{-n(H(X)-\varepsilon)},$$

where $X \sim p_x$.

This captures a notion of typicality:

Assume that each letter $x \in [d]$ appears roughly $n p_x$ times in a "typical" sequence x^n. Then,

$$p(x^n) \simeq \prod_{x\in[d]} p_x^{\,n p_x} = \prod_{x\in[d]} 2^{n p_x \log p_x}$$

$$= 2^{n \sum_{x\in[d]} p_x \log p_x}$$

$$= 2^{-nH(X)}.$$

$\boxed{\text{Prop}}$ (Properties of typical sequences)

Fix $\varepsilon > 0$. For any $\delta > 0$ there is $n_0 \in \mathbb{N}$ s.t. the following statements hold for all $n \geq n_0$:

i) $H(X) - \varepsilon \leq -\frac{1}{n} \log p(x^n) \leq H(X) + \varepsilon$ for all $x^n \in T_\varepsilon^{(n)}$.

ii) $\Pr(T_\varepsilon^{(n)}) \geq 1 - \delta$.

iii) $|T_\varepsilon^{(n)}| \leq 2^{n(H(X)+\varepsilon)}$

iv) $|T_\varepsilon^{(n)}| > (1-\delta)\, 2^{n(H(X)-\varepsilon)}$

Proof: See Ch. 14 in M. Wilde's book. $\square$

Shannon's compression thm for an iid. source $X \sim p_X$:

Fix a rate $R > H(X)$ and choose $\varepsilon > 0$ s.t. $H(X) + \varepsilon < R$.

For any $\delta > 0$ there is n_0 s.t., for $n \geq n_0$, there are at most $|T_\varepsilon^{(n)}| \leq 2^{n(H(X)+\varepsilon)} < 2^{nR}$ typical sequences. Now:

a) Index elements in $T_\varepsilon^{(n)}$ in some way using no more than $b = \lceil nR \rceil$ bits

b) Encoding:

For a received signal x^n, decide if $x^n \in T_\varepsilon^{(n)}$:

YES: assign index from a), prefix with symbol 1.

NO: assign fixed sequence of length b, prefix with 0.

c) Decoding:

On receiving sequence 1...., output respective typical sequence. For sequence 0...., declare error. The latter only occurs with probability δ.

In the limit $n \to \infty$ this defines a code with rate

$$r = \lim_{n \to \infty} \frac{1}{n}(nR + 1) = R$$ and error $e \to 0$.

Conversely, any code with rate $R < H(X)$ necessarily has $e \not\to 0$ as $n \to \infty$ (proof uses typicality again).

$\Rightarrow$ Shannon entropy $H(X)$ is the optimal compression rate.

§ 11.3 Strong typicality and universal compression

Last section: source compression based on typicality

Advantage: easy proof using law of large numbers

Disadvantage: encoding/decoding depend on source statistics

Goal: devise code that only depends on entropy of the source (= optimal source compression rate)

Requires stronger notion of typicality:

For a sequence $x^n = (x_1, \dots, x_n) \in [d]^n$ and $x \in [d]$, let

$$N(x|x^n) = \left| \{ i : x_i = x \} \right|$$

denote the number of occurences of x in x^n.

$\boxed{\text{Def}}$ (Type)

The type t_{x^n} of a sequence x^n is a probability distribution on $[d]$ defined as $t_{x^n}(x) = \frac{1}{n} N(x|x^n)$

Ex: let $d=3$, $n=5$, $x^n=(0,1,0,2,2)$

Then x^n has type $t_{x^n}=(2/5, 1/5, 2/5)$

Since $N(x|x^n)$ can only take $n+1$ possible values, there are at most $(n+1)^d$ different types.

This is only polynomial in n = sequence length.

Let $T_P \subseteq [d]^n$ denote the set of sequences x^n of type $t_{x^n}=P$. Then,

$$(n+1)^{-d}\, 2^{nH(P)} \leq |T_P| \leq 2^{nH(P)}$$

(Proof: textbook by Csiszár & Körner)

$\boxed{\text{Def}}$ (Strongly typical sequences)

Let $X \sim p_X$ be a source on $[d]$, and fix $\varepsilon > 0$.

A sequence x^n is called ε-strongly typical if

$$\left| \frac{1}{n} N(x|x^n) - p_x \right| \leq \varepsilon$$

for all $x \in [d]$ s.t. $p_x > 0$, and $N(x|x^n)=0$ if $p_x = 0$.

The set of all ε-strongly typical sequences is denoted $T^{(n)}_{X,\varepsilon}$.

Properties of strongly typical sequences:

i) For all $\delta > 0$ we have $\Pr\left(T_{X,\varepsilon}^{(n)}\right) \geq 1-\delta$

for sufficiently large n.

ii) $\left| \frac{1}{n} \log \left|T_{X,\varepsilon}^{(n)}\right| - H(X) \right| \leq c\varepsilon$ for some $c > 0$

and sufficiently large n.

iii) $2^{-n(H(X)+c\varepsilon)} \leq \Pr(x^n) \leq 2^{-n(H(X)-c\varepsilon)}$

for some constant $c > 0$.

Proof: See M. Wilde's book, Sec. 14.7.

Prop. iii) says that strong typicality implies typicality as defined in § 11.2 (often called weak typicality).

Prop i) + ii) give rise to a source compression protocol that only depends on $H(X)$:

For fixed $R > H(X)$, define

$$A^{(n)} := \bigcup_{P:\, H(P) < R} T_P ,$$

the set of all sequences of type P s.t. $H(P) < R$.

Then we have (Csiszár, Körner)

(*) $\quad |A^{(n)}| \leq (n+1)^d \, 2^{nR}$

because $|T_P| \leq 2^{nH(P)}$ and $\#(\text{types}) \leq (n+1)^d$, and

(**) $\quad \Pr(x^n \notin A^{(n)}) \leq (n+1)^d \exp\left[-n \min_{Q: H(Q) \geq R} D(Q \| P_X)\right]$

The protocol consists of only keeping sequences in $A^{(n)}$, for which by (*) we need at most

$$\frac{1}{n} \log\left[(n+1)^d \, 2^{nR}\right] = d \, \frac{\log(n+1)}{n} + R \xrightarrow{n\to\infty} R \text{ bits},$$

with the error decaying exponentially in n by (**).

§ 11.4 Quantum source compression

A quantum source emits quantum states with certain probabilities.

We restrict to pure state sources:

Let $(p_x, |\psi_x\rangle)_{x \in [d]}$ be a quantum state ensemble,

where $|\psi_x\rangle \in \mathcal{H}$ are pure states on a D-dim. Hilbert space.

Signal $|\psi_x\rangle$ is emitted with probability p_x.

iid. assumption: source emits sequences of states

$$|\psi_{x^n}\rangle := |\psi_{x_1}\rangle \otimes |\psi_{x_2}\rangle \otimes \dots \otimes |\psi_{x_n}\rangle$$

$\left(x^n = (x_1, \dots, x_n) \in [d]^n \text{ as before}\right)$ with probability

$$\Pr(\psi_{x^n}) = \prod_{i=1}^{n} p_{x_i}.$$

Let $\rho = \sum_{x \in [d]} p_x |\psi_x\rangle\langle\psi_x|$ be the ensemble average density operator. Then the average density operator after the source has emitted n signals is given by $\rho^{\otimes n}$.

A source compression protocol consists of:

i) an encoding or compression map

$$\mathcal{E} : \mathcal{L}(\mathcal{H}^{\otimes n}) \rightarrow \mathcal{L}(\tilde{\mathcal{H}}_n)$$

with $\dim \tilde{\mathcal{H}}_n < \dim \mathcal{H}^{\otimes n} = D^n$.

ii) a decoding operation $\mathcal{D} : \mathcal{L}(\tilde{\mathcal{H}}_n) \rightarrow \mathcal{L}(\mathcal{H}^{\otimes n})$.

Define the error $\varepsilon_n = 1 - \underbrace{\sum_{x^n} p_{x^n} F(\psi_{x^n}, \mathcal{D} \circ \mathcal{E}(\psi_{x^n}))}_{\text{average fidelity}}$

If $\varepsilon_n \to 0$ for $n \to \infty$, we call

$$R = \lim_{n\to\infty} \frac{1}{n} \log \dim \tilde{\mathcal{X}}_n$$

an achievable compression rate, and

$$R^* = \inf \{ R \text{ achievable} \}$$

the optimal rate of compression.

What is the equivalent entropy quantity here?

Def (von Neumann entropy)

The von Neumann entropy $S(\rho)$ of a density operator ρ with eigenvalues $\lambda = (\lambda_i)_{i=1,\dots,D}$ is defined as

$$S(\rho) = H(\lambda) = -\sum_i \lambda_i \log \lambda_i .$$

If $\rho = \sum_i \lambda_i |e_i\rangle\langle e_i|$ is a spectral decomposition, we can define the matrix logarithm

$$\log \rho = \sum_{i:\lambda_i > 0} \log \lambda_i \, |e_i\rangle\langle e_i| .$$

Then, $S(\rho) = -\operatorname{tr} \rho \log \rho$.

Properties of von Neumann entropy:

i) $0 \leq S(\rho) \leq \log D$ where $D = \dim \mathcal{H}$ $(\rho \in \mathcal{L}(\mathcal{H}))$

$S(\rho) = 0$ iff $\rho = |\psi\rangle\langle\psi|$ is pure.

$S(\rho) = \log D$ iff $\rho = \frac{1}{D} \mathbb{1}_{\mathcal{H}}$ is completely mixed.

ii) $S(\rho) = S(U \rho U^{\dagger})$ for every unitary $U \in \mathcal{U}(\mathcal{H})$

iii) $S(\lambda \rho_1 + (1-\lambda)\rho_2) \geq \lambda S(\rho_1) + (1-\lambda) S(\rho_2)$, $\lambda \in [0,1]$

iv) For any pure state $|\psi\rangle_{AB}$, we have $S(\psi_A) = S(\psi_B)$

(because of Schmidt decomposition).

v) A pure state $|\psi\rangle_{AB}$ is entangled iff $S(\psi_A) > 0$.

How can we achieve quantum source compression at a rate equal to the von Neumann entropy of the source?

Schumacher '95: use a quantum version of typicality

Let $\rho = \sum_x p_x |x\rangle\langle x|$ be a spectral decomposition of a density operator $\rho \in \mathcal{H}$. Consider the state $\rho^{\otimes n}$ with spectral decomposition $\rho^{\otimes n} = \sum_{x^n} p_{x^n} |x^n\rangle\langle x^n|$,

where $p_{x^n} := \prod_{i=1}^{n} p_{x_i}$ and $|x^n\rangle := \bigotimes_{i=1}^{n} |x_i\rangle$ (iid)

$\boxed{\text{Def}}$ (Typical subspace)

For $\varepsilon > 0$, the typical subspace $T_\varepsilon^{(n)}$ of a source $\rho = \sum_x p_x |x\rangle\langle x|$

is defined as

$$T_\varepsilon^{(n)} := \operatorname{span}\{ |x^n\rangle : x^n \text{ is } \varepsilon\text{-typical} \} \subseteq \mathcal{H}^{\otimes n}.$$

The projector onto $T_\varepsilon^{(n)}$ is given by $\Pi_\varepsilon^{(n)} := \sum_{x^n \in T_\varepsilon^{(n)}} |x^n\rangle\langle x^n|$

(we abuse notation and denote both the set of ε-typical sequences and the ε-typical subspace by the same symbol $T_\varepsilon^{(n)}$.)

Properties of the typical subspace:

i) For all $\delta > 0$ and n sufficiently large,

$$\operatorname{tr}(\Pi_\varepsilon^{(n)} \rho^{\otimes n}) \geq 1 - \delta.$$

ii) Let $S = S(\rho)$. Then for some constant $c > 0$,

$$\dim T_\varepsilon^{(n)} = \operatorname{tr} \Pi_\varepsilon^{(n)} \leq 2^{n(S + c\varepsilon)}.$$

iii) The operator $\tilde{\rho}_n := \Pi_\varepsilon^{(n)} \rho^{\otimes n} \Pi_\varepsilon^{(n)}$ is the 'typical' part of $\rho^{\otimes n}$ and satisfies $\tilde{\rho}_n \approx 2^{-nS} \Pi_\varepsilon^{(n)}$. Furthermore $\tilde{\rho}_n \approx \rho^{\otimes n}$ when n is large.

Schumacher's quantum source compression protocol:

i) Perform the typical subspace measurement to project the source signals to the typical subspace.

ii) Using some enumeration of the typical sequences in $T_\varepsilon^{(n)}$, construct a map $U_f = \sum_{x^n \in T_\varepsilon^{(n)}} |f(x^n)\rangle_W \langle x^n|_{A^n}$,

where $A^n \leftrightarrow \mathcal{H}^{\otimes n}$ and $W \leftrightarrow T_\varepsilon^{(n)}$ (typical subspace).

$T_\varepsilon^{(n)}$ has dimension at most $2^{n(S(\rho)+\varepsilon)}$.

U_f is the inverse of an isometry (i.e., $U_f U_f^\dagger = \mathbb{1}_W$).

iii) Decoding: essentially apply U_f^{-1}.

This achieves compression at a rate

$$R = \lim_{n\to\infty} \frac{1}{n} \log \dim \mathcal{H}_W = S(\rho),$$

with error $\varepsilon_n \to 0$ as $n \to \infty$.

We can again show that no asymptotically faithful compression protocol can achieve rates below the entropy $S(\rho)$.

$\Rightarrow$ $S(\rho)$ is the optimal rate of quantum source compression.

§ 11.5 Universal compression based on Schur-Weyl duality

Schumacher protocol achieves optimal compression rate, but is defined in terms of spectral decomposition of source state.

Want: compression protocol that only depends on $S(\rho)$.

How? Symmetries and Schur-Weyl duality!

Symmetries of quantum source compression:

i) Permutation symmetry: $Q_\pi \rho^{\otimes n} Q_\pi^\dagger = \rho^{\otimes n} \quad \forall \pi \in S_n$

ii) Unitary symmetry: $S(\rho) = S(U\rho U^\dagger) \quad \forall U \in \mathcal{U}(\mathcal{H})$

(entropy only depends on spectrum of ρ.)

$\Rightarrow$ use Schur-Weyl decomposition

$$\mathcal{H}^{\otimes n} \cong \bigoplus_{\lambda \vdash_D n} V_\lambda \otimes W_\lambda$$

Let P_λ be the projector onto $V_\lambda \otimes W_\lambda$.

For $\lambda \vdash_D n$ define $\bar{\lambda} = \frac{1}{n}\lambda$ (recall spectrum estimation).

Now fix $R > S(\rho)$ and define

$$\Pi_R := \sum_{\lambda : H(\bar{\lambda}) \leq R} P_\lambda .$$

This is a quantum version of the universal classical source compression code of § II.3!

Using Π_R as the projector in a source compression protocol, we can show (see Hayashi, arXiv: quant-ph/0202002):

i) With $\tilde{\mathcal{H}}_n := \Pi_R \mathcal{H}^{\otimes n} = \bigoplus_{\lambda : H(\bar{\lambda}) \leq R} V_\lambda \otimes W_\lambda$,

$$\dim \tilde{\mathcal{H}}_n = \operatorname{tr} \Pi_R \leq \operatorname{poly}(n)\, 2^{nR}$$

=> corresponding protocol has rate

$$\lim_{n \to \infty} \frac{1}{n} \log \dim \tilde{\mathcal{H}}_n \leq R.$$

ii) Exponential decay of decoding error:

$$\varepsilon_n \leq 2(n+D)^{4D} \exp\left(-n \min_{H(b) \geq R} D(b \| \lambda)\right),$$

where λ are the eigenvalues of the source ρ ($S(\rho) = H(\lambda)$), and as before $D(b \| \lambda) = \sum_x b_x \log \frac{b_x}{\lambda_x}$ is the relative entropy.

Since $S(\rho) = H(\lambda) < R \leq H(b)$, we have $b \neq \lambda$ for all b in the above optimization, and hence $\min_{H(b) \geq R} D(b \| \lambda) > 0$.

=> exponential decay of error ε_n for any rate $R > S(\rho)$.