COMP2610 – Information Theory

Lecture 12: The Source Coding Theorem

Mark Reid and Aditya Menon

Research School of Computer Science
The Australian National University

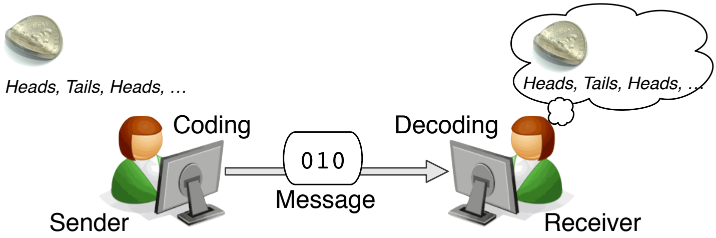
THE AUSTRALIAN NATIONAL UNIVERSITY

August 27th, 2014

A General Communication Game

Data compression is the process of replacing a message with a smaller message which can be reliably converted back to the original.

- Sender & Receiver agree on code for each outcome ahead of time (e.g., 0 for *Heads*; 1 for *Tails*)
- Sender observes outcomes then codes and sends message
- Receiver decodes message and recovers outcome sequence
- Want small messages on average when outcomes are from a fixed, known, but uncertain source (e.g., coin flips with known bias)



Definitions

Source Code

Given an ensemble X, the function $c : \mathcal{A}_X \to \mathcal{B}$ is a **source code** for X.
The number of symbols in $c(x)$ is the **length** $l(x)$ of the codeword for x.
The **extension** of c is defined by $c(x_1 \ldots x_n) = c(x_1) \ldots c(x_n)$

Smallest δ-sufficient subset

Let X be an ensemble and for $\delta \geq 0$ define S_δ to be the smallest subset of $\mathcal{A}_X$ such that

$$P(x \in S_\delta) \geq 1 - \delta$$

Essential Bit Content

Let X be an ensemble then for $\delta \geq 0$ the **essential bit content** of X is

$$H_\delta(X) \stackrel{\text{def}}{=} \log_2 |S_\delta|$$

A Note on Lossy Codes & Missing Codewords

When talking about a uniform lossy code c for $\mathcal{A}_X = \{\mathtt{a}, \mathtt{b}, \mathtt{c}\}$ we write

$$c(\mathtt{a}) = \mathtt{0} \quad c(\mathtt{b}) = \mathtt{1} \quad c(\mathtt{c}) = \mathtt{-}$$

where the symbol - means "no codeword". This is shorthand for "the receiver will decode this codeword incorrectly".

For the purposes of these lectures, this is equivalent to the code

$$c(\mathtt{a}) = \mathtt{0} \quad c(\mathtt{b}) = \mathtt{1} \quad c(\mathtt{c}) = \mathtt{1}$$

and the sender and receiver agreeing that the codeword 1 should always be decoded as b.

Assigning the outcome $\mathtt{a}_i$ the missing codeword "–" just means "it is not possible to send $\mathtt{a}_i$ reliably".

The Source Coding Theorem
(Theorem 4.1 in MacKay)

Our aim this week is to understand this:

The Source Coding Theorem

Let X be an ensemble with entropy $H = H(X)$ bits. Given $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 such that for all $N > N_0$

$$\left| \frac{1}{N} H_\delta \left(X^N \right) - H \right| < \epsilon.$$

The Source Coding Theorem

(Theorem 4.1 in MacKay)

Our aim this week is to understand this:

The Source Coding Theorem

Let X be an ensemble with entropy $H = H(X)$ bits. Given $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 such that for all $N > N_0$

$$\left| \frac{1}{N} H_\delta \left(X^N \right) - H \right| < \epsilon.$$

In English:

- Given outcomes drawn from X ...

The Source Coding Theorem

(Theorem 4.1 in MacKay)

Our aim this week is to understand this:

The Source Coding Theorem

Let X be an ensemble with entropy $H = H(X)$ bits. Given $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 such that for all $N > N_0$

$$\left| \frac{1}{N} H_\delta \left(X^N \right) - H \right| < \epsilon.$$

In English:

- Given outcomes drawn from X ...
- ... no matter what *reliability* $1 - \delta$ and *tolerance* ϵ you choose ...

The Source Coding Theorem

(Theorem 4.1 in MacKay)

Our aim this week is to understand this:

The Source Coding Theorem

Let X be an ensemble with entropy $H = H(X)$ bits. Given $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 such that for all $N > N_0$

$$\left| \frac{1}{N} H_\delta \left(X^N \right) - H \right| < \epsilon.$$

In English:

- Given outcomes drawn from X ...
- ... no matter what *reliability* $1 - \delta$ and *tolerance* ϵ you choose ...
- ... there is always a length N_0 so sequences X^N longer than this ...

The Source Coding Theorem
(Theorem 4.1 in MacKay)

Our aim this week is to understand this:

The Source Coding Theorem

Let X be an ensemble with entropy $H = H(X)$ bits. Given $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 such that for all $N > N_0$

$$\left| \frac{1}{N} H_\delta \left(X^N \right) - H \right| < \epsilon.$$

In English:

- Given outcomes drawn from X ...
- ... no matter what *reliability* $1 - \delta$ and *tolerance* ϵ you choose ...
- ... there is always a length N_0 so sequences X^N longer than this ...
- ... have an average essential bit content $\frac{1}{N} H_\delta(X^N)$ within ϵ of $H(X)$

The Source Coding Theorem

(Theorem 4.1 in MacKay)

Our aim this week is to understand this:

The Source Coding Theorem

Let X be an ensemble with entropy $H = H(X)$ bits. Given $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 such that for all $N > N_0$

$$\left| \frac{1}{N} H_\delta \left(X^N \right) - H \right| < \epsilon.$$

In English:

- Given outcomes drawn from X ...
- ... no matter what *reliability* $1 - \delta$ and *tolerance* ϵ you choose ...
- ... there is always a length N_0 so sequences X^N longer than this ...
- ... have an average essential bit content $\frac{1}{N} H_\delta(X^N)$ within ϵ of $H(X)$

$H_\delta(X^N)$ measures the *fewest* number of bits needed to uniformly code *smallest* set of N-outcome sequence S_δ with $P(x \in S_\delta) \geq 1 - \delta$.

Extended Ensembles (Review)

Instead of coding single outcomes, we now consider coding blocks and sequences of blocks

Example (Coin Flips):

$$
\begin{aligned}
\texttt{hhhhthhththh} &\rightarrow \texttt{hh hh th ht ht hh} && (6 \times 2 \text{ outcome blocks}) \\
&\rightarrow \texttt{hhh hth hth thh} && (4 \times 3 \text{ outcome blocks}) \\
&\rightarrow \texttt{hhhh thht hthh} && (3 \times 4 \text{ outcome blocks})
\end{aligned}
$$

Extended Ensembles (Review)

Instead of coding single outcomes, we now consider coding blocks and sequences of blocks

Example (Coin Flips):

$$\begin{aligned}
\texttt{hhhhthhththh} &\rightarrow \texttt{hh hh th ht ht hh} && (6 \times 2 \text{ outcome blocks})\\
&\rightarrow \texttt{hhh hth hth thh} && (4 \times 3 \text{ outcome blocks})\\
&\rightarrow \texttt{hhhh thht hthh} && (3 \times 4 \text{ outcome blocks})
\end{aligned}$$

Extended Ensemble

The **extended ensemble** of blocks of size N is denoted X^N. Outcomes from X^N are denoted $\mathbf{x} = (x_1, x_2, \ldots, x_N)$. The **probability** of $\mathbf{x}$ is defined to be $P(\mathbf{x}) = P(x_1)P(x_2)\ldots P(x_N)$.

Extended Ensembles (Review)

Instead of coding single outcomes, we now consider coding blocks and sequences of blocks

Example (Coin Flips):

$$
\begin{aligned}
\texttt{hhhhthhththh} &\rightarrow \texttt{hh hh th ht ht hh} && (6 \times 2 \text{ outcome blocks}) \\
&\rightarrow \texttt{hhh hth hth thh} && (4 \times 3 \text{ outcome blocks}) \\
&\rightarrow \texttt{hhhh thht hthh} && (3 \times 4 \text{ outcome blocks})
\end{aligned}
$$

Extended Ensemble

The **extended ensemble** of blocks of size N is denoted X^N. Outcomes from X^N are denoted $\mathbf{x} = (x_1, x_2, \ldots, x_N)$. The **probability** of $\mathbf{x}$ is defined to be $P(\mathbf{x}) = P(x_1)P(x_2)\ldots P(x_N)$.

What is the entropy of X^N?

Extended Ensembles (Review)

Example: Bent Coin

Let X be an ensemble with outcomes $\mathcal{A}_X = \{\mathtt{h}, \mathtt{t}\}$ with $p_\mathtt{h} = 0.9$ and $p_\mathtt{t} = 0.1$.

Consider X^4 – i.e., 4 flips of the coin.

$\mathcal{A}_{X^4} = \{\mathtt{hhhh}, \mathtt{hhht}, \mathtt{hhth}, \ldots, \mathtt{tttt}\}$

Extended Ensembles (Review)

Example: Bent Coin

Let X be an ensemble with outcomes $\mathcal{A}_X = \{\mathtt{h}, \mathtt{t}\}$ with $p_\mathtt{h} = 0.9$ and $p_\mathtt{t} = 0.1$.

Consider X^4 – i.e., 4 flips of the coin.

$$\mathcal{A}_{X^4} = \{\mathtt{hhhh}, \mathtt{hhht}, \mathtt{hhth}, \ldots, \mathtt{tttt}\}$$

What is the probability of

- Four heads? $P(\mathtt{hhhh}) = (0.9)^4 \approx 0.656$
- Four tails? $P(\mathtt{tttt}) = (0.1)^4 = 0.0001$

Let X be an ensemble with outcomes $\mathcal{A}_X = \{\mathtt{h}, \mathtt{t}\}$ with $p_{\mathtt{h}} = 0.9$ and $p_{\mathtt{t}} = 0.1$.

Consider X^4 – i.e., 4 flips of the coin.

$$\mathcal{A}_{X^4} = \{\mathtt{hhhh}, \mathtt{hhht}, \mathtt{hhth}, \ldots, \mathtt{tttt}\}$$

What is the probability of

- Four heads? $P(\mathtt{hhhh}) = (0.9)^4 \approx 0.656$
- Four tails? $P(\mathtt{tttt}) = (0.1)^4 = 0.0001$

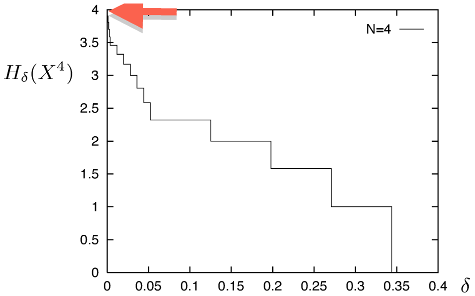
What is the entropy and raw bit content of X^4?

- The outcome set size is $|\mathcal{A}_{X^4}| = |\{0000, 0001, 0010, \ldots, 1111\}| = 16$
- Raw bit content: $H_0(X^4) = \log_2 |\mathcal{A}_{X^4}| = 4$
- Entropy: $H(X^4) = 4H(X) = 4.\,(-0.9 \log_2 0.9 - 0.1 \log_2 0.1) = 1.88$

Essential Bit Content of Extended Ensembles

What if we use a lossy uniform code on the extended ensemble?

x	$P(\mathbf{x})$	x	$P(\mathbf{x})$
hhhh	0.656	thht	0.008
hhht	0.073	thth	0.008
hhth	0.073	tthh	0.008
hthh	0.073	httt	0.001
thhh	0.073	thtt	0.001
htht	0.008	ttht	0.001
htth	0.008	ttth	0.001
hhtt	0.008	tttt	0.000

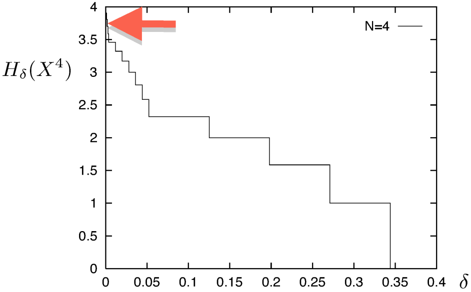


$$\delta = 0 \text{ gives } H_\delta\left(X^4\right) = \log_2 16 = 4$$

Essential Bit Content of Extended Ensembles

What if we use a lossy uniform code on the extended ensemble?

x	$P(\mathbf{x})$	**x**	$P(\mathbf{x})$
hhhh	0.656	thht	0.008
hhht	0.073	thth	0.008
hhth	0.073	tthh	0.008
hthh	0.073	httt	0.001
thhh	0.073	thtt	0.001
htht	0.008	ttht	0.001
htth	0.008	ttth	0.001
hhtt	0.008		

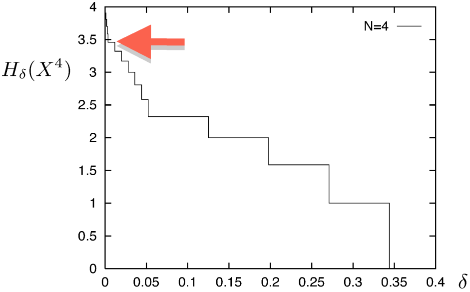


$$\delta = 0.0001 \text{ gives } H_\delta\left(X^4\right) = \log_2 15 = 3.91$$

Essential Bit Content of Extended Ensembles

What if we use a lossy uniform code on the extended ensemble?

$\mathbf{x}$	$P(\mathbf{x})$	$\mathbf{x}$	$P(\mathbf{x})$
hhhh	0.656	thht	0.008
hhht	0.073	thth	0.008
hhth	0.073	tthh	0.008
hthh	0.073		
thhh	0.073		
htht	0.008		
htth	0.008		
hhtt	0.008		

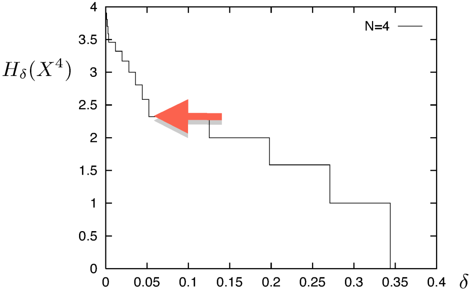


$$\delta = 0.005 \text{ gives } H_\delta\left(X^4\right) = \log_2 11 = 3.46$$

Essential Bit Content of Extended Ensembles

What if we use a lossy uniform code on the extended ensemble?

x	$P(\mathbf{x})$	**x**	$P(\mathbf{x})$
hhhh	0.656		
hhht	0.073		
hhth	0.073		
hthh	0.073		
thhh	0.073		

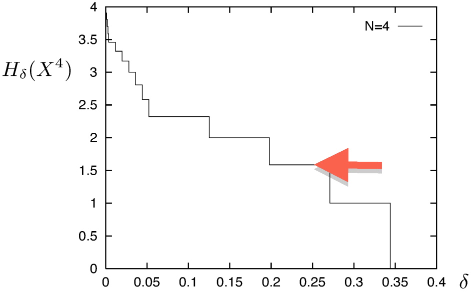


$$\delta = 0.05 \text{ gives } H_\delta\left(X^4\right) = \log_2 5 = 2.32$$

Essential Bit Content of Extended Ensembles

What if we use a lossy uniform code on the extended ensemble?

$\mathbf{x}$	$P(\mathbf{x})$	$\mathbf{x}$	$P(\mathbf{x})$
hhhh	0.656		
hhht	0.073		
hhth	0.073		

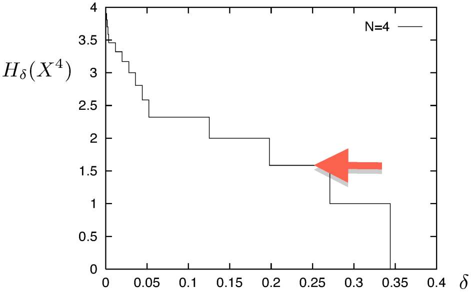


$$\delta = 0.25 \text{ gives } H_\delta\left(X^4\right) = \log_2 3 = 1.6$$

Essential Bit Content of Extended Ensembles

What if we use a lossy uniform code on the extended ensemble?

x	$P(\mathbf{x})$	**x**	$P(\mathbf{x})$
hhhh	0.656		
hhht	0.073		
hhth	0.073		

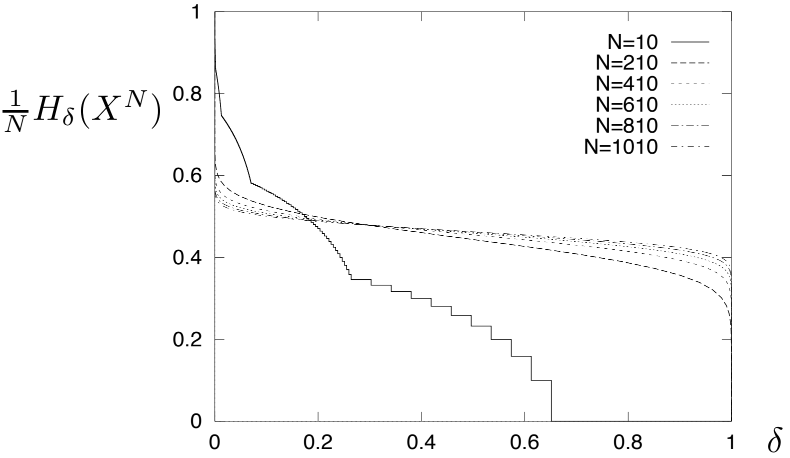


$$\delta = 0.25 \text{ gives } H_\delta\left(X^4\right) = \log_2 3 = 1.6$$

Unlike entropy, $H_\delta(X^4) \neq 4H_\delta(X) = 0$

Essential Bit Content of Extended Ensembles

What happens as N increases?



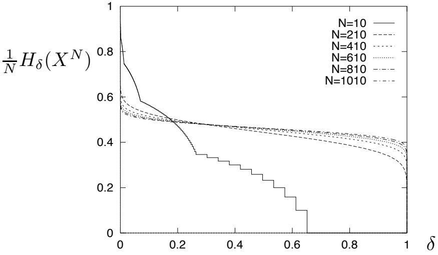
Recall that the entropy of a single coin flip with $p_{\mathrm{h}} = 0.9$ is $H(X) \approx 0.47$

The Source Coding Theorem

The Source Coding Theorem

Let X be an ensemble with entropy $H = H(X)$ bits. Given $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 such that for all $N > N_0$

$$\left| \frac{1}{N} H_\delta \left(X^N \right) - H \right| < \epsilon.$$



- Given a tiny probability of error δ, the average bits per outcome can be made as close to H as required.
- Even if we allow a large probability of error we cannot compress more than H bits per outcome for large sequences.

Typical Sets and the AEP (Review)

Typical Set

For "closeness" $\beta > 0$ the typical set $T_{N\beta}$ for X^N is

$$T_{N\beta} \stackrel{\text{def}}{=} \left\{ \mathbf{x} : \left| -\frac{1}{N} \log_2 P(\mathbf{x}) - H(X) \right| < \beta \right\}$$

The name "typical" is used since $\mathbf{x} \in T_{N\beta}$ will have roughly $p_1 N$ occurences of symbol a_1, $p_2 N$ of a_2, ..., $p_K N$ of a_K.

Typical Sets and the AEP (Review)

Typical Set

For "closeness" $\beta > 0$ the typical set $T_{N\beta}$ for X^N is

$$T_{N\beta} \stackrel{\text{def}}{=} \left\{ \mathbf{x} : \left| -\frac{1}{N} \log_2 P(\mathbf{x}) - H(X) \right| < \beta \right\}$$

The name "typical" is used since $\mathbf{x} \in T_{N\beta}$ will have roughly $p_1 N$ occurences of symbol a_1, $p_2 N$ of a_2, ..., $p_K N$ of a_K.

Asymptotic Equipartition Property (Informal)

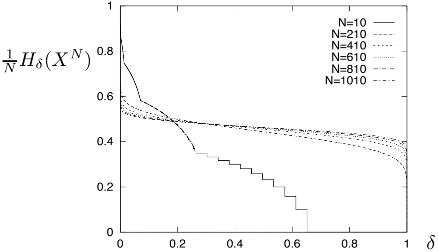
As $N \to \infty$, $\log_2 P(x_1, \ldots, x_N)$ is close to $-NH(X)$ with high probability.

For large block sizes "almost all sequences are typical" (i.e., in $T_{N\beta}$). This means $T_{N\beta}$ can be made to "look like" S_δ for any δ by choosing N large enough. This is useful since $T_{N\beta}$ is easy to count (size $\approx 2^{NH(X)}$) while S_δ is not (size varies with distribution)

The Source Coding Theorem

Let X be an ensemble with entropy $H = H(X)$ bits. Given $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 such that for all $N > N_0$

$$\left| \frac{1}{N} H_\delta \left(X^N \right) - H \right| < \epsilon.$$



Proof Idea: As N increases

- $T_{N\beta}$ has $\sim 2^{NH(X)}$ elements
- almost all **x** are in $T_{N\beta}$
- S_δ and $T_{N\beta}$ increasingly overlap
- so $\log_2 |S_\delta| \sim NH$

Proof of the SCT

The absolute value of a difference being bounded (e.g., $|x - y| \leq \epsilon$) says two things:

1. When $x - y$ is positive, it says $x - y < \epsilon$ which means $x < y + \epsilon$
2. When $x - y$ is negative, it says $-(x - y) < \epsilon$ which means $x < y - \epsilon$

$$|x - y| < \epsilon \quad \text{is equivalent to} \quad y - \epsilon < x < y + \epsilon$$

Proof of the SCT

The absolute value of a difference being bounded (e.g., $|x - y| \leq \epsilon$) says two things:

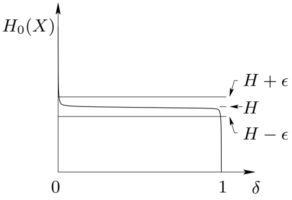
1. When $x - y$ is positive, it says $x - y < \epsilon$ which means $x < y + \epsilon$
2. When $x - y$ is negative, it says $-(x - y) < \epsilon$ which means $x < y - \epsilon$

$$|x - y| < \epsilon \quad \text{is equivalent to} \quad y - \epsilon < x < y + \epsilon$$

Using this, we break down the claim of the SCT into two parts: showing that for any ϵ and δ we can find N large enough so that:

Part 1: $\frac{1}{N} H_\delta(X^N) < H + \epsilon$

Part 2: $\frac{1}{N} H_\delta(X^N) > H - \epsilon$



Proof of the SCT (Part 1)

For $\epsilon > 0$ and $\delta > 0$, want N large enough so $\frac{1}{N} H_\delta(X^N) < H(X) + \epsilon$.

Proof of the SCT (Part 1)

For $\epsilon > 0$ and $\delta > 0$, want N large enough so $\frac{1}{N} H_\delta(X^N) < H(X) + \epsilon$.

Recall (see Lecture 10) for the *typical set* $T_{N\beta}$ we have for any N, β that

$$|T_{N\beta}| \leq 2^{N(H(X)+\beta)} \tag{1}$$

and, by the AEP, for any β as $N \to \infty$ we have $P(x \in T_{N\beta}) \to 1$.
So for any $\delta > 0$ we can always find an N such that $P(x \in T_{N\beta}) \geq 1 - \delta$.

Proof of the SCT (Part 1)

For $\epsilon > 0$ and $\delta > 0$, want N large enough so $\frac{1}{N} H_\delta(X^N) < H(X) + \epsilon$.

Recall (see Lecture 10) for the *typical set* $T_{N\beta}$ we have for any N, β that

$$|T_{N\beta}| \leq 2^{N(H(X)+\beta)} \tag{1}$$

and, by the AEP, for any β as $N \to \infty$ we have $P(x \in T_{N\beta}) \to 1$.
So for any $\delta > 0$ we can always find an N such that $P(x \in T_{N\beta}) \geq 1 - \delta$.

Now recall the definition of the *smallest δ-sufficient subset* S_δ: it is the smallest subset of outcomes such that $P(x \in S_\delta) \geq 1 - \delta$ so $|S_\delta| \leq |T_{N\beta}|$.

Proof of the SCT (Part 1)

For $\epsilon > 0$ and $\delta > 0$, want N large enough so $\frac{1}{N} H_\delta(X^N) < H(X) + \epsilon$.

Recall (see Lecture 10) for the *typical set* $T_{N\beta}$ we have for any N, β that

$$|T_{N\beta}| \leq 2^{N(H(X)+\beta)} \tag{1}$$

and, by the AEP, for any β as $N \to \infty$ we have $P(x \in T_{N\beta}) \to 1$.
So for any $\delta > 0$ we can always find an N such that $P(x \in T_{N\beta}) \geq 1 - \delta$.

Now recall the definition of the *smallest δ-sufficient subset* S_δ: it is the smallest subset of outcomes such that $P(x \in S_\delta) \geq 1 - \delta$ so $|S_\delta| \leq |T_{N\beta}|$.

So, given any δ and β we can find an N large enough so that, by (1)

$$|S_\delta| \leq |T_{N\beta}| \leq 2^{N(H(X)+\beta)}$$

Proof of the SCT (Part 1)

For $\epsilon > 0$ and $\delta > 0$, want N large enough so $\frac{1}{N}H_\delta(X^N) < H(X) + \epsilon$.

Recall (see Lecture 10) for the *typical set* $T_{N\beta}$ we have for any N, β that

$$|T_{N\beta}| \leq 2^{N(H(X)+\beta)} \tag{1}$$

and, by the AEP, for any β as $N \to \infty$ we have $P(x \in T_{N\beta}) \to 1$.
So for any $\delta > 0$ we can always find an N such that $P(x \in T_{N\beta}) \geq 1 - \delta$.

Now recall the definition of the *smallest δ-sufficient subset* S_δ: it is the smallest subset of outcomes such that $P(x \in S_\delta) \geq 1 - \delta$ so $|S_\delta| \leq |T_{N\beta}|$.

So, given any δ and β we can find an N large enough so that, by (1)

$$\log_2 |S_\delta| \leq \log_2 |T_{N\beta}| \leq N(H(X) + \beta)$$

Proof of the SCT (Part 1)

For $\epsilon > 0$ and $\delta > 0$, want N large enough so $\frac{1}{N} H_\delta(X^N) < H(X) + \epsilon$.

Recall (see Lecture 10) for the *typical set* $T_{N\beta}$ we have for any N, β that

$$|T_{N\beta}| \leq 2^{N(H(X)+\beta)} \tag{1}$$

and, by the AEP, for any β as $N \to \infty$ we have $P(x \in T_{N\beta}) \to 1$.
So for any $\delta > 0$ we can always find an N such that $P(x \in T_{N\beta}) \geq 1 - \delta$.

Now recall the definition of the *smallest δ-sufficient subset* S_δ: it is the smallest subset of outcomes such that $P(x \in S_\delta) \geq 1 - \delta$ so $|S_\delta| \leq |T_{N\beta}|$.

So, given any δ and β we can find an N large enough so that, by (1)

$$H_\delta(X^N) = \log_2 |S_\delta| \leq \log_2 |T_{N\beta}| \leq N(H(X) + \beta)$$

Setting $\beta = \epsilon$ and dividing through by N gives result.

Proof of the SCT (Part 2)

For $\epsilon > 0$ and $\delta > 0$, want N large enough so $\frac{1}{N}H_\delta(X^N) > H(X) - \epsilon$.

Suppose this was not the case – that is, for every N we have

$$\frac{1}{N}H_\delta(X^N) \leq H(X) - \epsilon \iff |S_\delta| \leq 2^{N(H(X)-\epsilon)}$$

Proof of the SCT (Part 2)

For $\epsilon > 0$ and $\delta > 0$, want N large enough so $\frac{1}{N}H_\delta(X^N) > H(X) - \epsilon$.

Suppose this was not the case – that is, for every N we have

$$\frac{1}{N}H_\delta(X^N) \leq H(X) - \epsilon \iff |S_\delta| \leq 2^{N(H(X)-\epsilon)}$$

Let's look at what this says about $P(x \in S_\delta)$ by writing

$$\begin{aligned} P(x \in S_\delta) &= P(x \in S_\delta \cap T_{N\beta}) + P(x \in S_\delta \cap \overline{T_{N\beta}}) \\ &\leq |S_\delta| 2^{-N(H-\beta)} + P(x \in \overline{T_{N\beta}}) \end{aligned}$$

since every $x \in T_{N\beta}$ has $P(x) \leq 2^{-N(H-\beta)}$ and $S_\delta \cap \overline{T_{N\beta}} \subset \overline{T_{N\beta}}$.

Proof of the SCT (Part 2)

For $\epsilon > 0$ and $\delta > 0$, want N large enough so $\frac{1}{N} H_\delta(X^N) > H(X) - \epsilon$.

Suppose this was not the case – that is, for every N we have

$$\frac{1}{N} H_\delta(X^N) \leq H(X) - \epsilon \iff |S_\delta| \leq 2^{N(H(X)-\epsilon)}$$

Let's look at what this says about $P(x \in S_\delta)$ by writing

$$\begin{aligned} P(x \in S_\delta) &= P(x \in S_\delta \cap T_{N\beta}) + P(x \in S_\delta \cap \overline{T_{N\beta}}) \\ &\leq |S_\delta| 2^{-N(H-\beta)} + P(x \in \overline{T_{N\beta}}) \end{aligned}$$

since every $x \in T_{N\beta}$ has $P(x) \leq 2^{-N(H-\beta)}$ and $S_\delta \cap \overline{T_{N\beta}} \subset \overline{T_{N\beta}}$.

So

$$P(x \in S_\delta) \leq 2^{N(H-\epsilon)} 2^{-N(H-\beta)} + P(x \in \overline{T_{N\beta}})$$

Proof of the SCT (Part 2)

For $\epsilon > 0$ and $\delta > 0$, want N large enough so $\frac{1}{N} H_\delta(X^N) > H(X) - \epsilon$.

Suppose this was not the case – that is, for every N we have

$$\frac{1}{N} H_\delta(X^N) \leq H(X) - \epsilon \iff |S_\delta| \leq 2^{N(H(X)-\epsilon)}$$

Let's look at what this says about $P(x \in S_\delta)$ by writing

$$\begin{aligned} P(x \in S_\delta) &= P(x \in S_\delta \cap T_{N\beta}) + P(x \in S_\delta \cap \overline{T_{N\beta}}) \\ &\leq |S_\delta| 2^{-N(H-\beta)} + P(x \in \overline{T_{N\beta}}) \end{aligned}$$

since every $x \in T_{N\beta}$ has $P(x) \leq 2^{-N(H-\beta)}$ and $S_\delta \cap \overline{T_{N\beta}} \subset \overline{T_{N\beta}}$.

So

$$P(x \in S_\delta) \leq 2^{-N(H-H+\epsilon-\beta)} + P(x \in \overline{T_{N\beta}})$$

Proof of the SCT (Part 2)

For $\epsilon > 0$ and $\delta > 0$, want N large enough so $\frac{1}{N} H_\delta(X^N) > H(X) - \epsilon$.

Suppose this was not the case – that is, for every N we have

$$\frac{1}{N} H_\delta(X^N) \leq H(X) - \epsilon \iff |S_\delta| \leq 2^{N(H(X)-\epsilon)}$$

Let's look at what this says about $P(x \in S_\delta)$ by writing

$$\begin{aligned} P(x \in S_\delta) &= P(x \in S_\delta \cap T_{N\beta}) + P(x \in S_\delta \cap \overline{T_{N\beta}}) \\ &\leq |S_\delta| 2^{-N(H-\beta)} + P(x \in \overline{T_{N\beta}}) \end{aligned}$$

since every $x \in T_{N\beta}$ has $P(x) \leq 2^{-N(H-\beta)}$ and $S_\delta \cap \overline{T_{N\beta}} \subset \overline{T_{N\beta}}$.

So

$$P(x \in S_\delta) \leq 2^{-N(\epsilon-\beta)} + P(x \in \overline{T_{N\beta}}) \to 0 \text{ as } N \to \infty$$

since $P(x \in T_{N\beta}) \to 1$.

Proof of the SCT (Part 2)

For $\epsilon > 0$ and $\delta > 0$, want N large enough so $\frac{1}{N}H_\delta(X^N) > H(X) - \epsilon$.

Suppose this was not the case – that is, for every N we have

$$\frac{1}{N}H_\delta(X^N) \le H(X) - \epsilon \iff |S_\delta| \le 2^{N(H(X)-\epsilon)}$$

Let's look at what this says about $P(x \in S_\delta)$ by writing

$$\begin{aligned} P(x \in S_\delta) &= P(x \in S_\delta \cap T_{N\beta}) + P(x \in S_\delta \cap \overline{T_{N\beta}}) \\ &\le |S_\delta| 2^{-N(H-\beta)} + P(x \in \overline{T_{N\beta}}) \end{aligned}$$

since every $x \in T_{N\beta}$ has $P(x) \le 2^{-N(H-\beta)}$ and $S_\delta \cap \overline{T_{N\beta}} \subset \overline{T_{N\beta}}$.

So

$$P(x \in S_\delta) \le 2^{-N(\epsilon-\beta)} + P(x \in \overline{T_{N\beta}}) \to 0 \text{ as } N \to \infty$$

since $P(x \in T_{N\beta}) \to 1$. But $P(x \in S_\delta) \ge 1 - \delta$, by defn. Contradiction

Interpretation of the SCT

The Source Coding Theorem

Let X be an ensemble with entropy $H = H(X)$ bits. Given $\epsilon > 0$ and $0 < \delta < 1$, there exists a positive integer N_0 such that for all $N > N_0$

$$\left| \frac{1}{N} H_\delta \left(X^N \right) - H \right| < \epsilon.$$

If you want to uniformly code blocks of N symbols drawn i.i.d. from X

- If you use more than $NH(X)$ bits per block you can do so without almost no loss of information as $N \to \infty$
- If you use less than $NH(X)$ bits per block you will almost certainly lose information as $N \to \infty$