Notes on Information Theory

Karl Stratos

(Work in progress)

1 Source Coding Theorem

We want to encode $\mathcal{X}^N$ into $\{0,1\}^B$ where $\mathcal{X}$ is a finite set of symbols. By the pigeonhole principle, we need $B \geq N \log |\mathcal{X}|$ to guarantee a lossless encoding of $|\mathcal{X}|^N$ possible sequences.[1] But suppose the sequence is a random variable $X \sim \mathbf{pop}^N$ where $\mathbf{pop}$ is a distribution over $\mathcal{X}$. Can we achieve an "almost lossless" encoding using fewer bits? By the usual interpretation of entropy, $B = H(\mathbf{pop}^N) = NH(\mathbf{pop})$ bits should be sufficient.

More formally, we consider a probabilistic compression of $\mathcal{X}^N$. Let $S_\delta(N)$ denote a subset of $\mathcal{X}^N$ such that

$$\Pr(X \in S_\delta(N)) \geq 1 - \delta \tag{1}$$

As $\delta \to 0^+$, it contains all "practical" sequences and $\log |S_\delta(N)|$ measures how many bits we need to encode $X \sim \mathbf{pop}^N$ without much loss of information.

1.1 Asymptotic Equipartition Principle

How small can $S_\delta(N)$ be? To answer this question, we first characterize the most *typical* realizations of X, because they will be size-efficient in capturing X. The crucial observation is that for a sequence $x \in \mathcal{X}^N$ drawn according to the generative process (i.e., $x_1 \ldots x_N \in \mathcal{X}$ are iid samples of $\mathbf{pop}$),

$$\lim_{N \to \infty} \left(-\frac{1}{N} \log \Pr(X = x) \right) = H(\mathbf{pop})$$

Thus as N gets bigger, *more* sequences $x \in \mathcal{X}^N$ will have a normalized negative log probability close to $H(\mathbf{pop})$. This motivates defining a typical set as

$$T_c(N) := \left\{ x \in \mathcal{X}^N : \left| -\frac{1}{N} \log \Pr(X = x) - H(\mathbf{pop}) \right| < c \right\} \tag{2}$$

for some $c > 0$. It follows from the weak law of large numbers (Tool B.3)

$$\Pr\left(X \in T_c(N)\right) \geq 1 - \frac{\sigma^2}{Nc^2} \tag{3}$$

where $\sigma^2 = \mathrm{Var}\left(-\log \mathbf{pop}(X_i)\right)$. At the same time, the definition (2) implies that any $x \in T_c(N)$ has a probability bounded as

$$2^{-N(H(\mathbf{pop})+c)} < \Pr\left(X = x\right) < 2^{-N(H(\mathbf{pop})-c)} \tag{4}$$

This is not surprising: typical sequences should be similarly probable, and no single sequence should hoard too much probability mass. (4) further implies that $T_c(N)$ cannot be too large. Specifically, since $|T_c(N)| \, 2^{-N(H(\mathbf{pop})+c)} < 1$, we must have

$$|T_c(N)| < 2^{N(H(\mathbf{pop})+c)} \tag{5}$$

The fact that, asymptotically in $N \to \infty$, $T_c(N)$ captures $X \in \mathcal{X}^N$ with only $2^{NH(\mathbf{pop})}$ sequences roughly having the same probability $2^{-NH(\mathbf{pop})}$ is referred to as the **asymptotic equipartition principle**.

[1] One such lossless encoding is

$$x \in \mathcal{X}^N \;\mapsto\; (\underbrace{b_1^{(1)} \ldots b_{\log|\mathcal{X}|}^{(1)}}_{\text{identify } x_1}, \;\; \underbrace{b_1^{(2)} \ldots b_{\log|\mathcal{X}|}^{(2)}}_{\text{identify } x_2}, \;\; \ldots, \;\; \underbrace{b_1^{(N)} \ldots b_{\log|\mathcal{X}|}^{(N)}}_{\text{identify } x_N}) \in \{0,1\}^{N \log |\mathcal{X}|}$$

1.2 Optimal Compression

We can now answer how small can $S_\delta(N)$ be. Let $S^\star_\delta(N)$ denote a smallest $S_\delta(N)$. Since we can choose $c_\delta(N) = \sigma(\delta N)^{-1/2}$ to have by (3)

$$\Pr\left(X \in T_{c_\delta(N)}\right) \geq 1 - \delta \tag{6}$$

whatever $S^\star_\delta(N)$ is, it has to be at least as small as $T_{c_\delta(N)}(N)$. Furthermore,

$$\begin{aligned} |S^\star_\delta(N)| &\leq \left|T_{c_\delta(N)}(N)\right| \\ &< 2^{N(H(\mathbf{pop})+c_\delta(N))} && \text{(by (5))} \\ &< 2^{N(H(\mathbf{pop})+\epsilon)} && \text{(for any } \epsilon > 0\text{, as long as } N \text{ is sufficiently large to drive } c_\delta(N) < \epsilon) \end{aligned}$$

We have proved the following lemma.

Lemma 1.1. Pick any $\epsilon > 0$ and $0 < \delta < 1$. There is some $N_0 \in \mathbb{N}$ such that for all $N > N_0$

$$|S^\star_\delta(N)| < 2^{N(H(\mathbf{pop})+\epsilon)} \tag{7}$$

By picking $\epsilon \to 0^+$ and $\delta \to 0^+$, we have that if N is sufficeintly large, choosing $2^{NH(\mathbf{pop})}$ (typical) sequences is sufficient to practically *guarantee* capturing $X \sim \mathbf{pop}^N$.

1.3 Any Compression

We can also show how big *any* $S_\delta(N)$ needs to be. Pick any $\epsilon > 0$ and $0 < \delta < 1$. For all sufficiently large N

$$|S_\delta(N)| > 2^{N(H(\mathbf{pop})-\epsilon)} \tag{8}$$

This happens mainly because

- For a large N, most sequences in $S_\delta(N)$ must also be in $T_c(N)$ by (3).
- But the probability of any $x \in T_c(N)$ is at most $2^{-N(H(\mathbf{pop})-c)}$ by (4).
- So $S_\delta(N)$ needs at least $O(2^{NH(\mathbf{pop})})$ elements to fulfill $\Pr(X \in S_\delta(N)) \geq 1 - \delta$.

See the proof of Lemma C.3 for details. By picking $\epsilon \to 0^+$ and $\delta \to 1^-$, we have that if N is sufficiently large, we can *never* capture any $X \sim \mathbf{pop}^N$ using fewer than $2^{NH(\mathbf{pop})}$ sequences.

1.4 A Combined Statement

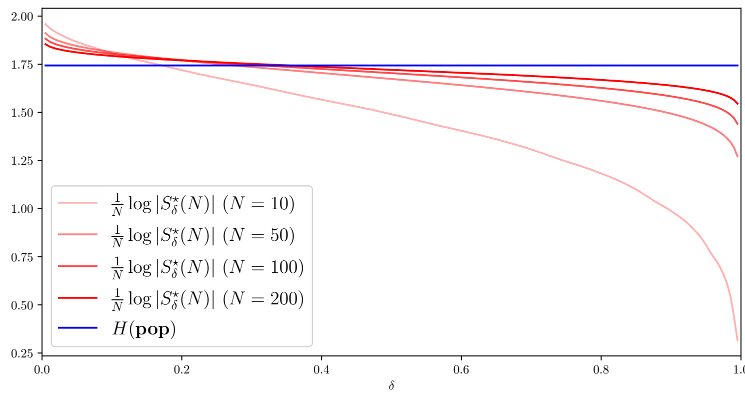
We can combine (7) and (8) as: for any $\epsilon > 0$ and $0 < \delta < 1$, for all sufficiently large N (Theorem C.4)

$$\left|\frac{1}{N}\log|S^\star_\delta(N)| - H(\mathbf{pop})\right| < \epsilon \tag{9}$$

In particular, pick $\epsilon \to 0^+$.[2] Then (9) holds for some large N and the "code rate" $\frac{1}{N}\log|S^\star_\delta(N)| \approx H(\mathbf{pop})$ is *constant* in δ. Thus it does not matter what δ is in the limit $N \to \infty$. Even if we are willing to lose almost all the information (i.e., δ is close to 1), we need the code rate of at least $H(\mathbf{pop})$ when N is sufficiently large. On the positive side, if we want to preserve almost all the information (i.e., δ is close to 0), we still need the code rate of only $H(\mathbf{pop})$ when N is sufficiently large.

[2] It is interesting to note that $\epsilon = 0$ is not allowed. But this simply reflects the fact that we must lose some information as long as we do not use all $|\mathcal{X}|^N$ sequences.

Visual proof. We set **pop** to be a random distribution over $\mathcal{X} = \{1, 2, 3, 4\}$. Given any N and δ, we can compute the size of $S^\star_\delta(N)$ by including most likely sequences $x \in \mathcal{X}^N$ (i.e., has the highest $\prod_{i=1}^N \mathbf{pop}(x_i)$) until $S^\star_\delta(N) \geq 1 - \delta$. The following plots the code rate as a function of $0 < \delta < 1$ for different values of N, as illustrated also in MacKay (2003).



References

Cover, T. M. and Thomas, J. A. (2006). *Elements of information theory*. John Wiley & Sons.

MacKay, D. J. (2003). *Information theory, inference and learning algorithms*. Cambridge university press.

A Binomial Coefficient

Analyzing an error-correcting code frequently involves the **binomial coefficient**: for $0 \leq k \leq n$,

$$\binom{n}{k} = \frac{n!}{(n-k)!k!}$$

is the number of ways to select k out of n items (unordered). It is also the coefficient of $x^{n-k}y^k$ in $(x+y)^n$ by the binomial theorem:

$$(x+y)^n = \sum_{k=1}^{n} \binom{n}{k} x^{n-k} y^k$$

Pascal's triangle states that, arranging $n = 0, 1, 2, \ldots$ as rows and $k = 0, \ldots, n$ as elements of the n-th row, we have

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}$$

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1

with the base case $\binom{0}{0} = 1$ (and 0 for all entries with $k < 0$). From the recurrence it is clear that

$$k^\star = \operatorname*{arg\,max}_{k \in \{0,\ldots,n\}} \binom{n}{k} \in \left\{ \left\lfloor \frac{n}{2} \right\rfloor, \left\lceil \frac{n}{2} \right\rceil \right\}$$

The ratio between $\binom{n}{k^\star}$ and the next maximum $\binom{n}{k^\star \pm 1}$ tends to 1 as $n \to \infty$,

$$\frac{\binom{n}{k^\star}}{\binom{n}{k^\star \pm 1}} = \frac{\left(\frac{n}{2}+1\right)!\left(\frac{n}{2}-1\right)!}{\left(\frac{n}{2}\right)!\left(\frac{n}{2}\right)!} = 1 + \frac{2}{n}$$

A.1 Information Theoretic Approximation

Using the fact that the binomial distribution $B(n, \frac{k}{n})$ (which involves the binomial coefficient) is normalized, we can show (Lemma C.1):

$$\frac{1}{n+1} 2^{nH_2\left(\frac{k}{n}\right)} \leq \binom{n}{k} \leq 2^{nH_2\left(\frac{k}{n}\right)} \tag{10}$$

where $H_2(p) := H(\text{Ber}(p))$ for any $p \in [0,1]$. A sharper bound exists for $0 < k < n$:

$$\sqrt{\frac{n}{8k(n-k)}} 2^{nH_2\left(\frac{k}{n}\right)} \leq \binom{n}{k} \leq \sqrt{\frac{n}{\pi k(n-k)}} 2^{nH_2\left(\frac{k}{n}\right)} \tag{11}$$

which follows from (a non-asympototic version of) Stirling's approximation; see Lemma 17.5.1 in Cover and Thomas (2006) for a proof. Thus we may approximate

$$\boxed{\binom{n}{k} \approx 2^{nH_2\left(\frac{k}{n}\right)}} \tag{12}$$

By (11), their ratio satisfies

$$\frac{\binom{n}{k}}{2^{nH_2\left(\frac{k}{n}\right)}} = \Theta\left(\sqrt{\frac{n}{k(n-k)}}\right)$$

In particular, choosing $k = \frac{n}{2}$ (assuming n is even) and using the fact that $H_2(\frac{1}{2}) = 1$,

$$\frac{\binom{n}{\frac{n}{2}}}{2^n} = \Theta\left(\sqrt{\frac{1}{n}}\right) \tag{13}$$

B Analytical Tools

Tool B.1 (Chebyshev's inequality 1)**.** For a nonnegative random variable $X \geq 0$ and a positive constant $c > 0$:

$$\Pr(X \geq c) \leq \frac{\mathbf{E}[X]}{c} \tag{14}$$

Proof. It is derived directly from the definition of $\mathbf{E}[X]$. □

Tool B.2 (Chebyshev's inequality 2)**.** For a random variable $X \in \mathbb{R}$ and a positive constant $c > 0$:

$$\Pr\left((X - \mathbf{E}[X])^2 \geq c\right) \leq \frac{\mathrm{Var}(X)}{c} \tag{15}$$

Proof. It is a corollary of Tool B.1 with $Y = (X - \mathbf{E}[X])^2 \geq 0$ as the nonnegative random variable satisfying $\mathbf{E}[Y] = \mathrm{Var}(X)$. □

Tool B.3 (Weak law of large numbers[3])**.** Let $X_1 \ldots X_N \in \mathbb{R}$ be iid random variables with a mean $\mu \in \mathbb{R}$ and a variance $\sigma^2 > 0$. For any positive constant $c > 0$:

$$\Pr\left(\left(\frac{1}{N}\sum_{i=1}^{N} X_i - \mu\right)^2 \geq c\right) \leq \frac{\sigma^2}{Nc} \tag{16}$$

Proof. It is a corollary of Tool B.2 with $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$ as the random variable satisfying $\mathbf{E}\left[\bar{X}\right] = \mu$ and $\mathrm{Var}\left(\bar{X}\right) = \frac{\sigma^2}{N}$. □

C Lemmas

Lemma C.1. For $0 \leq k \leq n$,

$$\frac{1}{n+1} 2^{nH_2\left(\frac{k}{n}\right)} \leq \binom{n}{k} \leq 2^{nH_2\left(\frac{k}{n}\right)}$$

Proof. We consider the binomial distribution $B(n,p)$ with $p = \frac{k}{n}$. For the upper bound, we note

$$1 \geq B(n,p)(k) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k} 2^{n\left(\frac{k}{n}\log p + \frac{n-k}{n}\log(1-p)\right)} = \binom{n}{k} 2^{-nH_2\left(\frac{k}{n}\right)}$$

For the lower bound, since $np = k$ is an integer, the mode of $B(n,p)$ is np (see Wikipedia). Then

$$1 = \sum_{k=0}^{n} B(n,p)(k) \leq (n+1)\binom{n}{np} p^{np}(1-p)^{n-np} = (n+1)\binom{n}{k} p^k (1-p)^{n-k} = (n+1)\binom{n}{k} 2^{-nH_2\left(\frac{k}{n}\right)}$$

□

Lemma C.2. Pick any $0 < p < \frac{1}{2}$ and even $N \in \mathbb{N}$. Let $X \sim \mathrm{Ber}(\frac{1}{2})$ and $Z \in \{0,1\}^N$ where

$$Z_i = \begin{cases} X & \text{with probability } 1-p \\ \neg X & \text{with probability } p \end{cases} \qquad \forall i = 1 \ldots N, \text{ independently}$$

Then for any $Z = z$,

$$x^\star = \underset{x \in \{0,1\}}{\arg\max}\ \Pr(X = x | Z = z) = \mathbf{Vote}(z) \tag{17}$$

where $\mathbf{Vote}(z) = \mathbb{1}(> \frac{N}{2} \text{ bits in } z \text{ are } 1)$. Furthermore,

$$\Pr(\mathbf{Vote}(Z) \neq X) \approx (4p(1-p))^{N/2} \tag{18}$$

The approximation becomes exact as $p \to 0$ and $N \to \infty$.

[3] See this post for why it is called "weak".

Proof. For (17), by Bayes' rule and the uniformity of X,

$$x^\star = \arg\max_{x\in\{0,1\}} \Pr(Z = z|X = x) = \begin{cases} 1 & \text{if } \Pr(Z = z|X = 1) > \Pr(Z = z|X = 0) \\ 0 & \text{otherwise} \end{cases}$$

Since $Z_1 \dots Z_N$ are independent, $\Pr(Z = z|X = x) = p^{\text{count}_{\neg x}(z)}(1-p)^{\text{count}_x(z)}$. Thus

$$\begin{aligned} x^\star = 1 \quad &\Leftrightarrow \quad p^{\text{count}_0(z)}(1-p)^{\text{count}_1(z)} > p^{\text{count}_1(z)}(1-p)^{\text{count}_0(z)} \\ &\Leftrightarrow \quad \left(\frac{p}{1-p}\right)^{\text{count}_0(z)-\text{count}_1(z)} > 1 \\ &\Leftrightarrow \quad \text{count}_0(z) - \text{count}_1(z) < 0 \\ &\Leftrightarrow \quad \textbf{Vote}(z) = 1 \end{aligned}$$

using the fact that $0 < p < \frac{1}{2}$. For (18), $\textbf{Vote}(Z) \neq X$ iff at least $\frac{N}{2}(\pm 1)$ of the bits flip X. Thus

$$\begin{aligned} \Pr(\textbf{Vote}(Z) \neq X) &= \text{Bin}(N,p)\left(\frac{N}{2}\right) + \text{Bin}(N,p)\left(\frac{N}{2}+1\right) + \cdots + \text{Bin}(N,p)(N) \\ &\approx \text{Bin}(N,p)\left(\frac{N}{2}\right) && \text{(exact as } p \to 0 \text{ by (19))} \\ &= \binom{N}{N/2} p^{N/2}(1-p)^{N/2} \\ &\approx 2^N p^{N/2}(1-p)^{N/2} && \text{(exact as } N \to \infty \text{ by (13))} \\ &= (4p(1-p))^{N/2} \end{aligned}$$

For the first approximation, first note that the terms are monotonically decreasing since $\frac{N}{2} > Np$ (i.e., we are past the mean of the binomial distribution). The first term dominates the next term by

$$\frac{\text{Bin}(N,p)(N/2)}{\text{Bin}(N,p)(N/2+1)} = \left(\frac{\binom{N}{N/2}}{\binom{N}{N/2+1}}\right) \frac{p^{N/2}(1-p)^{N/2}}{p^{N/2+1}(1-p)^{N/2-1}} = \left(1 + \frac{2}{N}\right)\frac{1-p}{p} = \Omega\left(\frac{1}{p}\right) \tag{19}$$

So the approximation is justified for sufficiently small p. □

Lemma C.3. Pick any $\epsilon > 0$ and $0 < \delta < 1$. For each $N \in \mathbb{N}$, pick any subset $S_\delta(N) \subset \mathcal{X}^N$ satisfying $\Pr(X \in S_\delta(N)) \geq 1 - \delta$ with respect to $X \sim \textbf{pop}^N$. There is some $N_0 \in \mathbb{N}$ such that for all $N > N_0$

$$|S_\delta(N)| > 2^{N(H(\textbf{pop})-\epsilon)}$$

Proof. Suppose otherwise. Then there are infinitely many $N_1 < N_2 < \cdots$ such that $|S_\delta(N_i)| \leq 2^{N_i(H(\textbf{pop})-\epsilon)}$. For any constant $c > 0$, we may use the typical set $T_c(N_i)$ defined in (2) and its complement $T_c^{\complement}(N_i)$ to have

$$\begin{aligned} \Pr(X \in S_\delta(N_i)) &= \Pr\left(X \in S_\delta(N_i) \cap T_c^{\complement}(N_i)\right) + \Pr(X \in S_\delta(N_i) \cap T_c(N_i)) \\ &\leq \Pr(X \notin T_c(N_i)) + |S_\delta(N_i)| \max_{x' \in T_c(N_i)} \Pr(X = x') \qquad (20) \\ &< \frac{\sigma^2}{N_i c^2} + 2^{N_i(H(\textbf{pop})-\epsilon)} \cdot 2^{-N_i(H(\textbf{pop})-c)} \qquad (21) \\ &= \frac{\sigma^2}{N_i c^2} + 2^{N_i(c-\epsilon)} \qquad (22) \end{aligned}$$

(20) is a worst-case bound. The first term uses the fact that $X \in S_\delta(N_i) \cap T_c^{\complement}(N_i)$ implies $X \notin T_c(N_i)$. A more formal derivation of the second term is

$$\begin{aligned} \Pr(X \in S_\delta(N_i) \cap T_c(N_i)) = \sum_{x \in S_\delta(N_i)} \mathbb{1}(x \in T_c(N_i)) \Pr(X = x) &\leq \sum_{x \in S_\delta(N_i)} \max_{x' \in T_c(N_i)} \Pr(X = x') \\ &= |S_\delta(N_i)| \max_{x' \in T_c(N_i)} \Pr(X = x') \end{aligned}$$

where the inequality follows because for any $x \in \mathcal{X}$

$$\mathbb{1}\left(x \in T_c(N_i)\right)\Pr\left(X = x\right) = \begin{cases} \Pr\left(X = x\right) & \text{if } x \in T_c(N_i) \\ 0 & \text{otherwise} \end{cases} \leq \max_{x' \in T_c(N_i)} \Pr\left(X = x'\right)$$

(21) uses the coverage of the typical set (3), the smallness of $S_\delta(N_i)$, and the probability bound on a typical element (4). Now we select $c = \frac{\epsilon}{2} > 0$ to obtain

$$\Pr(X \in S_\delta(N_i)) < \frac{2\sigma^2}{N_i\epsilon^2} + 2^{-N_i(\epsilon/2)}$$

which grows strictly smaller for $N_1 < N_2 < \cdots$. Thus we can find a sufficiently large j such that

$$\Pr(X \in S_\delta(N_j)) < 1 - \delta$$

which contradicts the premise. □

Theorem C.4. Pick any $\epsilon > 0$ and $0 < \delta < 1$. For each $N \in \mathbb{N}$, pick a *smallest* subset $S^\star_\delta(N) \subset \mathcal{X}^N$ satisfying $\Pr(X \in S^\star_\delta(N)) \geq 1 - \delta$ with respect to $X \sim \mathbf{pop}^N$. There is some $N_0 \in \mathbb{N}$ such that for all $N > N_0$

$$\left|\frac{1}{N}\log|S^\star_\delta(N)| - H(\mathbf{pop})\right| < \epsilon$$

Proof. Since $S^\star_\delta(N)$ is a particular subset satisfying the condition in Lemma C.3, there is some $N'_0 \in \mathbb{N}$ such that $|S^\star_\delta(N)| > 2^{N(H(\mathbf{pop})-\epsilon)}$ for all $N > N'_0$. By Lemma 1.1, there is some $N''_0 \in \mathbb{N}$ such that $|S^\star_\delta(N)| < 2^{N(H(\mathbf{pop})+\epsilon)}$ for all $N > N''_0$. Thus for all $N > N_0 = \max(N'_0, N''_0)$,

$$\begin{aligned} 2^{N(H(\mathbf{pop})-\epsilon)} < |S^\star_\delta(N)| < 2^{N(H(\mathbf{pop})+\epsilon)} \quad &\Leftrightarrow \quad H(\mathbf{pop}) - \epsilon < \frac{1}{N}\log|S^\star_\delta(N)| < H(\mathbf{pop}) + \epsilon \\ &\Leftrightarrow \quad \left|\frac{1}{N}\log|S^\star_\delta(N)| - H(\mathbf{pop})\right| < \epsilon \end{aligned}$$

□