CO 432 Spring 2025:
Lecture Notes

Lecture notes taken, unless otherwise specified, by myself during the Spring 2025 offering of CO 432, taught by Vijay Bhattiprolu.

Lectures

Chapter 1

Introduction

Notation. I will be using my usual LaTeX typesetting conventions:

- $[n]$ means the set $\{1, 2, \dots, n\}$
- $\{0, 1\}^*$ means the set of bitstrings of arbitrary length (i.e., the Kleene star)
- $\sum_i$ is implicitly $\sum_{i=1}^{n}$
- $\mathsf{A}, \mathsf{B}, \dots, \mathsf{Z}$ are random variables (in sans-serif)
- $\mathsf{X} = (p_1, p_2, \dots, p_n)$ means X is a discrete random variable with n outcomes such that $\Pr[\mathsf{X} = 1] = p_1$, $\Pr[\mathsf{X} = 2] = p_2$, etc. (abbreviate further as $\mathsf{X} = (p_i)$)

1.1 Entropy

↓ *Lecture 1 adapted from Arthur* ↓ *Lecture 1 May 6*

Definition 1.1.1 (entropy)

For a random variable $\mathsf{X} = (p_i)$, the entropy $H(\mathsf{X})$ is

$$H(\mathsf{X}) = -\sum_i p_i \log p_i = \sum_i p_i \log \frac{1}{p_i}.$$

Convention. By convention, we usually use $\log_2$. Also, we define entropy such that $\log_2(0) = 0$ so that impossible values do not break the formula.

Example 1.1.2. If X takes on the values a, b, c, d with probabilities 1, 0, 0, 0, respectively, then $H(\mathsf{X}) = 1 \log 1 = 0$.

If X takes on those values instead with probabilities $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{8}$, respectively, then $H(\mathsf{X}) = \frac{7}{4}$.

Fact 1.1.3. $H(\mathsf{X}) = 0$ if and only if X is a constant.

Proof. Suppose X is constant. Then, $H(\mathsf{X}) = 1 \log 1 = 0$.

Suppose $H(\mathsf{X}) = 0$. Probabilities are in $[0,1]$, so $p_i \log \frac{1}{p_i} \geq 0$. Since $H(\mathsf{X}) = \sum_i p_i \log \frac{1}{p_i} = 0$ and each term is non-negative, each term must be zero. Thus, each p_i is either 0 or 1. We cannot have $\sum p_i > 1$, so exactly one $p_i = 1$ and the rest are zero. That is, X is constant. □

Theorem 1.1.4 (Jensen's inequality)
Let $f : \mathbb{R} \to \mathbb{R}$ be concave. That is, for any a and b in the domain of f and $\lambda \in [0,1)$, $f(\lambda a + (1-\lambda)b) \geq \lambda f(a) + (1-\lambda) f(b)$. For any discrete random variable X,

$$\mathbb{E}[f(\mathsf{X})] \leq f(\mathbb{E}[\mathsf{X}])$$

Proof. Consider a random variable X with two values a and b, each with probabilities λ and $1-\lambda$. Then, notice that

$$\mathbb{E}[f(\mathsf{X})] = \lambda f(a) + (1-\lambda) f(b) \leq f(\lambda a + (1-\lambda)b) = f(\mathbb{E}[\mathsf{X}])$$

by convexity of f.

TODO: This can be generalized by induction. □

Fact 1.1.5. Assume X is supported on $[n]$. Then, $0 \leq H(\mathsf{X}) \leq \log n$.

Proof. Start by claiming without proof that $\log n$ is concave, so we can apply Jensen's inequality.

Let $\mathsf{X}' = \frac{1}{p_i}$ with probability p_i. Then,

$$\begin{aligned} H(\mathsf{X}) &= \sum_i p_i \log \frac{1}{p_i} \\ &= \mathbb{E}\left[\log(\mathsf{X}')\right] \\ &\leq \log(\mathbb{E}[\mathsf{X}']) \\ &= \log\left(\sum p_i \frac{1}{p_i}\right) \\ &= \log n \end{aligned}$$ □

It is not a coincidence that $\log_2 n$ is the minimum number of bits to encode $[n]$.

1.2 Entropy as expected surprise

We want $S : [0,1] \to [0,\infty)$ to capture how "surprised" we are $S(p)$ that an event with probability p happens. We want to show that under some natural assumptions, this is the only function we could have defined as entropy. In particular:

1. $S(1) = 0$, a certainty should not be surprising
2. $S(q) > S(p)$ if $p > q$, less probable should be more surprising
3. $S(p)$ is continuous in p
4. $S(pq) = S(p) + S(q)$, surprise should add for independent events. That is, if I see something twice, I should be twice as surprised.

↑ *Lecture 1 adapted from Arthur* ↑ *Lecture 2 May 8*

Proposition 1.2.1

If $S(p)$ satisfies these 4 axioms, then $S(p) = c \cdot \log_2(1/p)$ for some $c > 0$.

Proof. Suppose a function $S : [0, 1] \to [0, \infty)$ exists satisfying the axioms. Let $c := S(\frac{1}{2}) > 0$.

By axiom 4 (addition), $S(\frac{1}{2^k}) = kS(\frac{1}{2})$. Likewise, $S(\frac{1}{2^{1/k}} \cdots \frac{1}{2^{1/k}}) = S(\frac{1}{2^{1/k}}) + \cdots + S(\frac{1}{2^{1/k}}) = kS(\frac{1}{2^{1/k}})$.

Then, $S(\frac{1}{2^{m/n}}) = \frac{m}{n} S(\frac{1}{2}) = \frac{m}{n} \cdot c$ for any rational m/n.

By axiom 3 (continuity), $S(\frac{1}{2^z}) = c \cdot z$ for all $z \in [0, \infty)$ because the rationals are dense in the reals. In particular, for any $p \in [0, 1]$, we can write $p = \frac{1}{2^z}$ for $z = \log_2(1/p)$ and we get

$$S(p) = S\left(\frac{1}{2^z}\right) = c \cdot z = c \cdot \log_2(1/p)$$

as desired. □

We can now view entropy as expected surprise. In particular,

$$\sum_i p_i \log_2 \frac{1}{p_i} = \mathop{\mathbb{E}}_{x \sim \mathsf{X}} [S(p_x)]$$

for a random variable $\mathsf{X} = i$ with probability p_i.

1.3 Entropy as optimal lossless data compression

Suppose we are trying to compress a string consisting of n symbols drawn from some distribution.

Problem 1.3.1

What is the expected number of bits you need to store the results of n independent samples of a random variable X?

We will show this is $nH(\mathsf{X})$.

Notice that we assume that the symbols we are drawn independently, which is violated by almost all data we actually care about.

Definition 1.3.2

Let $C : \Sigma \to (\Sigma')^*$ be a code. We say C is a uniquely decodable code (UDC) if there does not exist a collision $x, y \in \Sigma^*$, with identical encoding $C(x_1)C(x_2)\cdots C(x_k) = C(y_1)C(y_2)\cdots C(y_{k'})$.

Also, C is prefix-free (sometimes called instantaneous) if for any distinct $x, y \in \Sigma$, $C(x)$ is not a prefix of $C(y)$.

Proposition 1.3.3

Prefix-freeness is sufficient for unique decodability.

Example 1.3.4. Let $C : \{A, B, C, D\} \to \{0,1\}^*$ where $C(A) = 11$, $C(B) = 101$, $C(C) = 100$, and $C(D) = 00$. Then, C is prefix-free and uniquely decodable.

We can easily parse 1011100001100 unambiguously as 101.11.00.00.11.00 ($BADDAD$).

Recall from CS 240 that a prefix-free code is equivalent to a trie, and we can decode it by traversing the trie in linear time.

Theorem 1.3.5 (Kraft's inequality)

A prefix-free binary code $C : \{1, \dots, n\} \to \{0,1\}^*$ with codeword lengths $\ell_i = |C(i)|$ exists if and only if

$$\sum_{i=1}^{n} \frac{1}{2^{\ell_i}} \le 1.$$

Proof. Suppose $C : \{1, \dots, n\} \to \{0,1\}^*$ is prefix-free with codeword lengths ℓ_i. Let T be its associated binary tree and let W be a random walk on T where 0 and 1 have equal weight (stopping at either a leaf or undefined branch).

Define E_i as the event where W reaches i and $E_\varnothing$ where W falls off. Then,

$$\begin{aligned}
1 &= \Pr(E_\varnothing) + \sum_i \Pr(E_i) \\
&= \Pr(E_\varnothing) + \sum_i \frac{1}{2^{\ell_i}} && \text{(by independence)} \\
&\ge \sum_i \frac{1}{2^{\ell_i}} && \text{(probabilities are non-negative)}
\end{aligned}$$

Conversely, suppose the inequality holds for some ℓ_i. WLOG, suppose $\ell_1 < \ell_2 < \cdots < \ell_n$.

Start with a complete binary tree T of depth ℓ_n. For each $i = 1, \dots, n$, find any unassigned node in T of depth ℓ_i, delete its children, and assign it a symbol.

Now, it remains to show that this process will not fail. That is, for any loop step i, there is still some unassigned node at depth ℓ_i.

Let $P \leftarrow 2^{\ell_n}$ be the number of leaves of the complete binary tree of depth ℓ_n. After the i^{th} step, we decrease P by $2^{\ell_n - \ell_i}$. That is, after n steps,

$$\begin{aligned} P &= 2^{\ell_n} - \sum_{i=1}^{n} \frac{2^{\ell_n}}{2^{\ell_i}} \\ &= 2^{\ell_n} - 2^{\ell_n} \sum_{i=1}^{n} \frac{1}{2^{\ell_i}} \\ &\geq 0 \end{aligned}$$

by the inequality. □

Recall the problem we are trying to solve: *Lecture 3 May 13*

Problem 1.3.1

What is the expected number of bits you need to store the results of n independent samples of a random variable X?

Solution (Shannon & Faro). Consider the case where X is symbol i with probability p_i. We want to encode independent samples $x_i \sim \mathsf{X}$ as $C(x_i)$ for some code $C : [n] \to \{0,1\}^*$.

Suppose for simplification that $p_i = \frac{1}{2^{\ell_i}}$ for some integers ℓ_i. Since $\sum p_i = 1$, we must have $\sum \frac{1}{2^{\ell_i}} = 1$. Then, by Kraft's inequality, there exists a prefix-free binary code $C : [n] \to \{0,1\}^*$ with codeword lengths $|C(i)| = \ell_i$. Now,

$$\mathop{\mathbb{E}}_{x_i \sim \mathsf{X}} \left[\sum_i |C(x_i)| \right] = \sum_i p_i \ell_i = \sum_i p_i \log_2 \frac{1}{p_i} = H(\mathsf{X})$$

Proceed to the general case. Suppose $\log_2 \frac{1}{p_i}$ are non-integral. Instead, use $\ell_i' = \left\lceil \log_2 \frac{1}{p_i} \right\rceil$. We still satisfy Kraft since $\sum_i \frac{1}{2^{\ell_i'}} \leq \sum_i p_i = 1$. Then,

$$\mathop{\mathbb{E}}_{x_i \sim \mathsf{X}} \left[\sum_i |C(x_i)| \right] = \sum_i p_i \ell_i' = \sum_i p_i \left\lceil \log_2 \frac{1}{p_i} \right\rceil$$

which is bounded by

$$H(\mathsf{X}) = \sum_i p_i \log_2 \frac{1}{p_i} \leq \sum_i p_i \left\lceil \log_2 \frac{1}{p_i} \right\rceil < \sum_i p_i \left(1 + \log_2 \frac{1}{p_i} \right) = H(\mathsf{X}) + 1$$

We call the code C generated by this process the Shannon–Faro code. □

We can improve on this bound $[H(\mathsf{X}), H(\mathsf{X}) + 1)$ by amortizing over longer batches of the string.

Solution (batching). For Y defined on $[n]$ equal to i with probability q_i, define the random variable $\mathsf{Y}^{(k)}$ on $[n]^k$ equal to the string $i_1 \cdots i_k$ with probability $q_{i_1} \cdots q_{i_k}$. That is, $\mathsf{Y}^{(k)}$ models k independent samples of Y.

Apply the Shannon–Fano code to $\mathsf{Y}^{(k)}$ to get an encoding of $[n]^k$ as bitstrings of expected length ℓ

satisfying $H(\mathsf{Y}^{(k)}) \leq \ell \leq H(\mathsf{Y}^{(k)}) + 1$.

$$\begin{aligned}
H(\mathsf{Y}^{(k)}) &= \mathop{\mathbb{E}}_{i_1 \cdots i_k \sim \mathsf{Y}^{(k)}} \left[\log_2 \frac{1}{q_{i_1} \cdots q_{i_k}} \right] && \text{(by def'n)} \\
&= \mathop{\mathbb{E}}_{i_1 \cdots i_k \sim \mathsf{Y}^{(k)}} \left[\log_2 \frac{1}{q_{i_1}} + \cdots + \log_2 \frac{1}{q_{i_k}} \right] && \text{(log rules)} \\
&= \sum_{j=1}^{k} \mathop{\mathbb{E}}_{i_1 \cdots i_k \sim \mathsf{Y}^{(k)}} \left[\log_2 \frac{1}{q_{i_j}} \right] && \text{(linearity of expectation)} \\
&= \sum_{j=1}^{k} \mathop{\mathbb{E}}_{i \sim \mathsf{Y}} \left[\log_2 \frac{1}{q_i} \right] && \text{(q_{i_j} only depends on one character)} \\
&= kH(\mathsf{Y}) && \text{(by def'n, no j-dependence in sum)}
\end{aligned}$$

For every k symbols, we use ℓ bits, i.e., $\frac{\ell}{k}$ bits per symbol. From the Shannon–Faro bound, we have

$$\begin{aligned}
\frac{H(\mathsf{Y}^{(k)})}{k} \leq \frac{\ell}{k} < \frac{H(\mathsf{Y}^{(k)})}{k} + \frac{1}{k} \\
H(\mathsf{Y}) \leq \frac{\ell}{k} < H(\mathsf{Y}) + \frac{1}{k}
\end{aligned}$$

Then, we have a code for Y bounded by $[H(\mathsf{Y}), H(\mathsf{Y}) + \frac{1}{k})$.

Taking a limit of some sort, we can say that we need $H(\mathsf{Y}) + o(1)$ bits. □

Definition 1.3.6 (relative entropy)
Given two discrete distributions $p = (p_i)$ and $q = (q_i)$, the relative entropy

$$D(p \parallel q) := \sum p_i \log_2 \frac{1}{q_i} - \sum_i p_i \log_2 \frac{1}{p_i} = \sum p_i \log_2 \frac{p_i}{q_i}$$

This is also known as the KL divergence.

Lecture 4
May 15

The KL divergence works vaguely like a "distance" between distributions. (In particular, KL divergence is not a metric since it lacks symmetry and does not follow the triangle inequality, but it can act sorta like a generalized squared distance.)

Fact 1.3.7. $D(p \parallel q) \geq 0$ with equality exactly when $p = q$.

Proof. Observe that

$$-D(p \parallel q) = \sum_i p_i (-\log_2 \frac{p_i}{q_i}) = \sum_i p_i \log_2 \frac{q_i}{p_i}$$

and then define $\mathsf{X}' = \frac{q_i}{p_i}$ with probability p_i. By construction, we get

$$-D(p \parallel q) = \mathbb{E}[\log_2 \mathsf{X}'] \leq \log_2(\mathbb{E}[\mathsf{X}'])$$

by Jensen's inequality (as $f = \log_2$ is concave). Finally,

$$D(p \parallel q) \geq -\log_2(\mathbb{E}[\mathsf{X}']) = \log_2 \left(\sum_i p_i \frac{q_i}{p_i} \right) = \log_2 1 = 0$$ □

Proposition 1.3.8

Any prefix-free code has an expected length at least $H(\mathsf{X})$.

Proof. Let $\mathsf{X} = (p_i)$. Suppose C is a prefix-free code with codeword lengths ℓ_i.

Then, by Kraft's inequality, $\sum_i 2^{-\ell_i} \leq 1$. We want to show that $\sum_i p_i \ell_i \geq H(\mathsf{X})$, and we will prove this by showing that $\sum_i p_i \ell_i - H(\mathsf{X}) = D(p \parallel q)$ for some distribution q (then apply fact 1.3.7).

We will take q to be the random walk distribution corresponding to the binary tree associated to the candidate prefix-free code.

Let T be the binary tree associated to C. Consider the process of randomly going left/right at each node and stopping when either falling off the tree or hitting a leaf.

Let $q_i = 2^{-\ell_i}$ be the probability that this random walk reaches the leaf for the symbol i and let $q_{n+1} = 1 - \sum_i 2^{-\ell_i}$ be the probability that the random walk falls off the tree. Also, to keep ranges identical, let $\tilde{p}_i = p_i$ and $\tilde{p}_{n+1} = 0$. Now,

$$\begin{aligned} D(\tilde{p} \parallel q_C) &= \sum_{i=1}^{n+1} \tilde{p}_i \log_2 q_i^{-1} - \sum_{i=1}^{n+1} \tilde{p}_i \log_2 \frac{1}{p_i} \\ &= \sum_{i=1}^{n} p_i \log_2 2^{\ell_i} - \sum_{i=1}^{n} p_i \log_2 \frac{1}{p_i} && (\tilde{p}_{n+1} = 0) \\ &= \sum_{i=1}^{n} p_i \ell_i - H(\mathsf{X}) \end{aligned}$$

Therefore, by fact 1.3.7, $\sum_i p_i \ell_i \geq H(\mathsf{X})$. $\square$

This proof technique generalizes. Recall the distinction between UDCs and prefix-free codes:

Definition 1.3.2

Let $C : \Sigma \to (\Sigma')^*$ be a code. We say C is a uniquely decodable code (UDC) if there does not exist a collision $x, y \in \Sigma^*$, with identical encoding $C(x_1)C(x_2)\cdots C(x_k) = C(y_1)C(y_2)\cdots C(y_{k'})$.

Also, C is prefix-free (sometimes called instantaneous) if for any distinct $x, y \in \Sigma$, $C(x)$ is not a prefix of $C(y)$.

Example 1.3.9. The code $C(1,2,3,4) = (10, 00, 11, 110)$ is a uniquely decodable code.

The code $C'(1,2,3,4) = (0, 10, 110, 111)$ is a prefix-free code.

Remark 1.3.10. A natural additional requirement for unique decodability is that for any $k \in \mathbb{N}$, $x \in [n]^k$, $y \in [n]^k$, $C(x) \neq C(y)$.

Theorem 1.3.11

For any uniquely decodable code $C : [n] \to \{0,1\}^*$ of codeword lengths ℓ_i, there is also a prefix-free code $C' : [n] \to \{0,1\}^*$ of lengths ℓ_i.

We will show that for any UDC C, the lengths $\sum_i 2^{-\ell_i} \leq 1$. Then, Kraft's inequality applies and we have a prefix-free code C'.

Partition the code's codomain $C([n]) = C_1 \cup C_2 \cup C_3 \cup \cdots$ by the length of the codeword $C_j \subseteq \{0,1\}^j$. We must instead show $\sum_j \frac{|C_i|}{2^j} \leq 1$.

Consider the easy case $C([n]) = C_2 \cup C_3$. If there are no collisions of length 5, we have

$$2 \cdot |C_2| \cdot |C_3| \leq 2^5$$

because every string in $\{xy : x \in C_2, y \in C_3\} \cup \{yx : x \in C_2, y \in C_3\}$ is unique in $\{0,1\}^5$. That is, $|C_2| \cdot |C_3| \leq 2^4$.

Likewise, if there are no collisions of length $5k$, we get

$$\frac{(2k)!}{k! \cdot k!} \cdot |C_2|^k \cdot |C_3|^k \leq 2^{5k}$$

because the union $\bigcup\limits_{\substack{\alpha \in \{2,3\}^{2k}, \\ \alpha_i = 2 \text{ for} \\ k \text{ choices of } i}} C_{\alpha_i}$ consists of only unique strings.

In the limit, by Sterling's approximation,

$$\begin{aligned} \frac{2^{2k}}{\sqrt{k}} \cdot |C_2|^k \cdot |C_3|^k &\leq 2^{5k} \\ |C_2| \cdot |C_3| &\leq \frac{2^5}{2^2} (\sqrt{k})^{1/k} \approx 1 + \mathcal{O}(\log k / k) \end{aligned}$$

I have no idea where this was going.

Proof. Fix a $k \geq 1$. Let $\ell_{max} = \max \ell_i$. Write $C^{(k)}$ to be the set of encoded k-length strings.

Consider the distribution: sample a length m uniformly from the set $[k \cdot \ell_{max}]$. Also, sample a uniformly random string $s \in \{0,1\}^m$. For each $x \in C^{(k)}$, let E_x be the event where $s = x$.

Now, we can write

$$\sum_{x \in C^{(k)}} \Pr[E_x] \leq 1$$

because the events E_x are mutually exclusive. Then,

$$\begin{aligned} \sum_{x \in C^{(k)}} \frac{1}{k \cdot \ell_{max}} \cdot \frac{1}{2^{\ell(x)}} &\leq 1 \\ \sum_{x \in C^{(k)}} \frac{1}{2^{\ell(x)}} &\leq k \cdot \ell_{max} \end{aligned}$$

where $\ell(x)$ is the length of x. Since summing over each codeword $x \in C$ is the same as summing over each codeword ℓ_i,

$$\begin{aligned}
\left(\sum_i \frac{1}{2^{\ell_i}}\right)^k &= \left(\sum_{x\in C} \frac{1}{2^{\ell(x)}}\right)^k \\
&= \sum_{x_1,\dots,x_k\in C} \frac{1}{2^{\ell(x_1)}} \cdot \frac{1}{2^{\ell(x_2)}} \cdots \frac{1}{2^{\ell(x_k)}} \\
&= \sum_{x_1,\dots,x_k\in C} \frac{1}{2^{\ell(x_1)+\ell(x_2)+\cdots+\ell(x_k)}} \\
&= \sum_{x_1,\dots,x_k\in C} \frac{1}{2^{\ell(x_1 x_2\cdots x_k)}} \\
&= \sum_{x\in C^{(k)}} \frac{1}{2^{\ell(x)}}
\end{aligned}$$

where we can take the last step by uniquely decoding $x_1 x_2 \cdots x_k$ into x. Combining,

$$\begin{aligned}
\left(\sum_i \frac{1}{2^{\ell_i}}\right)^k &\leq k \cdot \ell_{max} \\
\sum_i \frac{1}{2^{\ell_i}} &\leq (k \cdot \ell_{max})^{\frac{1}{k}} \\
&\leq 1 + \mathcal{O}\left(\frac{\ell_{max} \cdot \log_2 k}{k}\right)
\end{aligned}$$

which tends to 1 as $k \to \infty$, as desired. $\square$

Chapter 2

Applications of KL divergence

Lecture 5
May 20

Notation. Write $H(p)$ to denote $H(\mathsf{X})$ for $\mathsf{X} \sim \text{Bernoulli}(p)$.

That is, $H(p) = p \log_2 \frac{1}{p} + (1-p)\log_2 \frac{1}{1-p}$.

Likewise, write $D(q \parallel p)$ to be $D(\mathsf{Y} \parallel \mathsf{X})$ where $\mathsf{Y} \sim \text{Bernoulli}(q)$.

Recall Sterling's approximation (which we have used before):

Theorem 2.0.1 (Sterling's approximation)
$m!$ behaves like $\sqrt{2\pi m}\left(\frac{m}{e}\right)^m\left(1 + \mathcal{O}\left(\frac{1}{m}\right)\right)$

2.1 The boolean k-slice

Consider the boolean k-slice (also known as the Hamming k-slice) of the hypercube $\{0,1\}^n$ defined by

$$B_k := \{x \in \{0,1\}^n : x \text{ has exactly } k \text{ ones}\}$$

Remark 2.1.1.

$$|B_k| \approx 2^{H(\frac{k}{n})\cdot n}$$

Proof. By Sterling's approximation, knowing that $|B_k| = \binom{n}{k}$:

$$\begin{aligned} |B_k| &= \binom{n}{k} \\ &= \frac{n!}{n!(n-k)!} \\ &\approx \frac{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n}{\sqrt{2\pi k}\left(\frac{k}{e}\right)^k \sqrt{2\pi(n-k)}\left(\frac{n-k}{e}\right)^{n-k}} \\ &= \sqrt{\frac{n}{2\pi k(n-k)}} \cdot \frac{n^k\left(\frac{n}{n-k}\right)^{n-k}}{k^k} \end{aligned}$$

Now, notice that $\left(\frac{n}{n-k}\right)^{n-k} = \left(1 + \frac{k}{n-k}\right)^{n-k} \approx e^k$ for $k \ll n-k$ because $1 + x \approx e^x$ for small x. Then, $\left(1 + \frac{k}{n-k}\right)^{n-k} \approx \left(e^{k/(n-k)}\right)^{n-k} = e^k$ and

$$\begin{aligned} |B_k| &\approx \left(\frac{ne}{k}\right)^k \\ &= 2^{k \log_2 \frac{ne}{k}} \\ &= 2^{k \log_2 \frac{n}{k} + k \log_2 e} \\ &= 2^{(\frac{k}{n} \log_2 \frac{n}{k})n + k \log_2 e} \\ &\approx 2^{(\frac{k}{n} \log_2 \frac{n}{k})n} \end{aligned} \tag{2.1}$$

for $1 \ll k \ll n$. Then, given that same assumption,

$$\begin{aligned} H\left(\frac{k}{n}\right) &= \frac{k}{n} \log_2 \frac{n}{k} + \cancelto{0}{\left(1 - \frac{k}{n}\right) \log_2 \frac{1}{1 - \frac{k}{n}}} \\ &\approx \frac{k}{n} \log_2 \frac{n}{k} \end{aligned}$$

because if $n \gg k$, $\frac{k}{n} \to 0$ and $1 \log_2 1 = 0$. Combining these approximations yields

$$|B_k| \approx 2^{H(\frac{k}{n})n}$$

□

Let X be a uniformly chosen point in B_k and $\mathsf{X}_1, \dots, \mathsf{X}_n \sim \text{Bernoulli}(\frac{k}{n})$.

This means that $H(\mathsf{X}) \approx H((\mathsf{X}_1, \dots, \mathsf{X}_n))$, which is remarkable because the latter could produce points in B_k or points with n ones or points with no ones.

This seems to imply that the majority of the mass of $(\mathsf{X}_1, \dots, \mathsf{X}_n)$ lies within the boolean k-slice. Formally, we make the following claim about the <u>concentration of measure</u>:[1]

[1]cf. Dvoretzky–Milman theorem

Proposition 2.1.2

Fix any $\varepsilon > 0$. The probability

$$\Pr\left[(\mathsf{X}_1, \ldots, \mathsf{X}_n) \notin \bigcup_{\ell=(1-\varepsilon k)}^{(i+\varepsilon)k} B_\ell\right] = \frac{1}{2^{n/\varepsilon^2}}$$

Informally, the probability of the randomly-drawn vector lying outside of the boolean k-slice is exponentially small.

We will prove a stronger claim:

Claim 2.1.3. Fix any $p \in (0,1)$ and consider any $q > p$. Then,

$$\Pr[w((\mathsf{X}_i)) > q \cdot n] \leq 2^{-D(q\|p)\cdot n}$$

where $w((\mathsf{X}_i))$ is the number of ones. Likewise, consider any $q < p$. Then,

$$\Pr[w((\mathsf{X}_i)) < q \cdot n] \leq 2^{-D(q\|p)\cdot n}$$

Consider a toy example first. Let X be the number of heads after n fair coin tosses.

Then, $\mathbb{E}[\mathsf{X}] = \frac{n}{2}$ and

$$\Pr[\mathsf{X} \geq 0.51n] = \frac{1}{2^n} \sum_{k \geq 0.51n}^{n} \binom{n}{k} \approx \frac{1}{2^n} \sum_{k \geq 0.51n}^{n} \left(\frac{ne}{k}\right)^k \to 0 \text{ very quickly}$$

by the same magic that we did in eq. (2.1) and because $\frac{1}{2^n}$ goes to 0 very quickly.

Now we can prove the claim.

Proof. Let $\theta_p(x)$ denote the probability of sampling a vector $x \in \{0,1\}^n$ where each bit is IID Bernoulli(p). Then,

$$\begin{aligned}
\frac{\theta_p(x)}{\theta_q(x)} &= \frac{p^k(1-p)^k}{q^k(1-q)^k} \\
&= \frac{(1-p)^n}{(1-q)^n} \left(\frac{\frac{p}{1-p}}{\frac{q}{1-q}}\right)^k \\
&\leq \frac{(1-p)^n}{(1-q)^n} \left(\frac{\frac{p}{1-p}}{\frac{q}{1-q}}\right)^{qn}
\end{aligned}$$

for any $k \geq qn$ because (1) if $q \geq p$, then $\frac{q}{1-q} \geq \frac{p}{1-p}$ and the ugly fraction is greater than 1 and (2) increasing the exponent increases the quantity if the base is greater than 1.

Let $B_{\geq k} := \bigcup_{\ell \geq k} B_\ell$. Then, for all $x \in B_{\geq qn}$, we must show that

$$\theta_p(x) \leq \frac{(1-p)^n}{(1-q)^n} \left(\frac{\frac{p}{1-p}}{\frac{q}{1-q}}\right)^{qn} \cdot \theta_q(x) = 2^{-nD(q\|p)\cdot\theta_q(x)}$$

Consider the right-hand expression:

$$\begin{aligned} 2^{n \cdot D(q \| p)} &= 2^{n \cdot (q \log_2 \frac{1}{p} + (1-q)\log_2 \frac{1}{1-p} - q \log_2 \frac{1}{q} - (1-q)\log_2 \frac{1}{1-q})} \\ &= \left(\frac{1}{p^q} \cdot \frac{1}{(1-p)^{1-q}} \cdot q^q \cdot (1-q)^{1-q} \right)^n \end{aligned}$$

and the left-hand expression:

$$\begin{aligned} \frac{(1-p)^n}{(1-q)^n} \left(\frac{\frac{p}{1-p}}{\frac{q}{1-q}} \right)^{qn} &= \left(\frac{(1-p)^{1-q} p^q}{(1-q)^{1-q} q^q} \right)^n \\ &= \left(p^q \cdot (1-p)^{1-q} \cdot \frac{1}{q^q} \cdot \frac{1}{(1-q)^{1-q}} \right)^n \end{aligned}$$

which is clearly the reciprocal of the right-hand expression.

Now, we know that $\theta_p(x) = 2^{-nD(q \| p)} \theta_q(x)$, so

$$\begin{aligned} &\Pr_{\mathsf{X}_1,\dots,\mathsf{X}_n \sim \mathrm{Bernoulli}(p)} [(\mathsf{X}_1, \dots, \mathsf{X}_n) \in B_{\geq qn}] \\ &= \sum_{x \in B_{\geq qn}} \theta_p(x) \\ &\leq 2^{-nD(q\|p)} \sum_{x \in B_{\geq qn}} \theta_q(x) \\ &\leq 2^{-nD(q\|p)} \end{aligned}$$

since the sum of the probabilities of x being any given entry in $B_{\geq qn}$ must be at most 1. □

2.2 Rejection sampling

The KL divergence can give us a metric of how accurately we can sample one distribution using another distribution.

> **Example 2.2.1.** Suppose $\mathsf{X} = \begin{cases} 0 & p = \frac{1}{2} \\ 1 & p = \frac{1}{2} \end{cases}$ and $\mathsf{Y} = \begin{cases} 0 & p = \frac{1}{4} \\ 1 & p = \frac{3}{4} \end{cases}$.
>
> How can we sample Y using X?

Solution (naive). Take IID X_1 and X_2. Return 0 if $x_1 = x_2 = 0$ and 1 otherwise. □

Solution (fancy). Take an infinite IID queue $\mathsf{X}_1, \mathsf{X}_2, \dots$

Starting at $i = 1$, if $\mathsf{X}_i = 0$, then output 0 with probability $\frac{1}{2}$, otherwise increment i until $\mathsf{X}_i = 1$. □

List of Named Results

Index of Defined Terms