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

Entropy

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)
- $\mathsf{X} \sim (p_1, p_2, \dots, p_n)$ 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 $\mathsf{X} \sim (p_i)$)

1.1 Definition

↓ *Lecture 1 adapted from Arthur* ↓

Lecture 1
May 6

Definition 1.1.1 (entropy)

For a random variable $\mathsf{X} \sim (p_i)$, 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 <u>independently</u>, 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

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

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,

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

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

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

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. $\square$

Chapter 2

Relative entropy

Definition 2.0.1 (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.

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.) *Lecture 4 May 15*

Fact 2.0.2. $D(p \parallel q) \geq 0$ 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 2.0.3

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

Proof. Let $\mathsf{X} \sim (p_i)$. 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 2.0.2).

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 2.0.2, $\sum_i p_i \ell_i \geq H(\mathsf{X})$. □

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 2.0.4. 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 2.0.5. 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 2.0.6

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 $\bigcup_{\substack{\alpha \in \{2,3\}^{2k}, \\ \alpha_i = 2 \text{ for} \\ k \text{ choices of } i}} C_{\alpha_i}$ 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,

$$\begin{aligned}
\left(\sum_i \frac{1}{2^{\ell_i}}\right)^k &= \left(\sum_{x\in C} \frac{1}{2^{\ell(x)}}\right)^k \\
&= \sum_{x_1,\dots,x_k\in C} \frac{1}{2^{\ell(x_1)}}\cdot\frac{1}{2^{\ell(x_2)}}\cdots\frac{1}{2^{\ell(x_k)}} \\
&= \sum_{x_1,\dots,x_k\in C} \frac{1}{2^{\ell(x_1)+\ell(x_2)+\cdots+\ell(x_k)}} \\
&= \sum_{x_1,\dots,x_k\in C} \frac{1}{2^{\ell(x_1x_2\cdots x_k)}} \\
&= \sum_{x\in C^{(k)}} \frac{1}{2^{\ell(x)}}
\end{aligned}$$

where we can take the last step by uniquely decoding $x_1x_2\cdots x_k$ into x. Combining,

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

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

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

2.1 The boolean k-slice

Consider the <u>boolean k-slice</u> (also known as the <u>Hamming k-slice</u>) 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 concentration of measure:[1]

[1]cf. Dvoretzky–Milman theorem

Proposition 2.1.2

Fix any $\varepsilon > 0$. The probability

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

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

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

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 $\mathsf{X}_1, \mathsf{X}_2, \ldots$

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

↓ *Lecture 6 adapted from Arthur* ↓

Lecture 6
May 22

> **Problem 2.2.2** (rejection sampling)
> Given access to a distribution $Q = (Q(x))_{x\in\mathcal{X}}$, how efficiently can you simulate $P = (P(x))_{x\in\mathcal{X}}$?

Example 2.2.3. Suppose $Q = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ and $P = (\frac{1}{2}, \frac{1}{2})$. We want to obtain the P distribution from Q.

Solution. Since Q and P are both uniform, we can just keep sampling from Q until we get something in P. That is, for $i = 1, \dots, \infty$:

1. Sample $\mathsf{X}_i \sim Q$.
2. If $\mathsf{X}_i \in \{1, 2\}$, accept and output $\mathsf{Y} \leftarrow \mathsf{X}_i$.
3. Otherwise, $i \leftarrow i + 1$.

This works because

$$\Pr[\mathsf{Y} = 1] = \Pr[\mathsf{X}_i = 1 \mid \mathsf{X}_i = 1 \lor \mathsf{X}_i = 2] = \frac{1/3}{2/3} = \frac{1}{2}$$

for the final round i, and similarly for $\mathsf{Y} = 2$. □

Example 2.2.4. Consider a slightly more complex distribution $P = (\frac{1}{3}, \frac{2}{3})$ and $Q = (\frac{1}{2}, \frac{1}{2})$.

Solution. We will create a more complex rejection sampling protocol with some cheating.

Again, iterate and draw independent X_i:

- If $\mathsf{X}_1 = 1$, accept with probability $\frac{2}{3}$. Otherwise, reject and continue to X_2 with probability $\frac{1}{3}$.
- If $\mathsf{X}_1 = 2$, accept.
- For $i \geq 2$, accept if $\mathsf{X}_i = 1$ and reject if $\mathsf{X}_i = 2$.

Then, the probability of accepting $\mathsf{X}_1 = 1$ is $\frac{1}{3}$, $\mathsf{X}_1 = 2$ is $\frac{1}{2}$, and rejecting X_1 is $\frac{1}{6}$.

Since later rounds only output 1, we output 1 with probability $\frac{1}{3} + \frac{1}{6} = \frac{1}{2}$ and 2 with probability $\frac{1}{2}$. □

Definition 2.2.5 (rejection sampler)

A rejection sampler is a procedure that reads sequentially independent random samples $\mathsf{X}_i \sim Q$ and in each round i either

- accepts the value of X_i and terminates with an index i^*, or
- rejects and continues.

The iteration we terminated on i^* is a random variable since it is a function of other random variables. It satisfies $\mathsf{X}_{i^*} \sim P$, which is weird since for all fixed i, $\mathsf{X}_i \sim Q$.

An interesting application is communication complexity. Suppose Alice has some hidden distribution P. Alice and Bob have access to a shared random IID sequence $\mathsf{X}_i \sim Q$.

Alice can send an encoding of i^* to Bob who outputs $\mathsf{X}_{i^*} \sim P$. This encoding i^* can be encoded using $\log i^*$ bits.

We will show that $\mathbb{E}[\log i^*] \leq D(P \parallel Q) + \mathcal{O}(1)$. You can also show that $D(P \parallel Q) \leq \mathbb{E}[\log i*]$.

For each round i and symbol x, we need to know whether x was sampled before round i, i.e., the probability assigned to x in previous rounds.

For round $i \geq 1$, define:

- $\alpha_i(x)$ to denote the probability that the procedure accepts X_i and that $\mathsf{X}_i = x$
- $p_i(x)$ to denote the probability that the procedure halts at round $i^* \leq i$ and $\mathsf{X}_{i^*} = x$

We want to construct our procedure such that

- for all x, $P(x) = \sum_{i=1}^{n} \alpha_i(x)$
- for all x and i, $p_i(x) = \sum_{k=1}^{i} \alpha_k(x)$
- the probability that we halt on or before round i is $p_i^* := \sum_{x \in \mathcal{X}} p_i(x)$

↑ *Lecture 6 adapted from Arthur* ↑

Lecture 7
May 27

Algorithm 1 $\textsc{RejectionSampling}(P, Q)$

Require: $\forall x \in \mathcal{X}, Q(x) > 0 \iff D(P \parallel Q) < \infty$

for $x \in \mathcal{X}$ **do** $p_0(x) \leftarrow 0$
$p_0^* \leftarrow 0$
for $i = 1, \dots, \infty$ **do**
  sample $\mathsf{X}_i \sim Q$
  if $P(\mathsf{X}_i) - P_{i-1}(\mathsf{X}_i) \leq (1 - p_{i-1}^*) \cdot Q(\mathsf{X}_i)$ **then**
    with probability $\beta_i(\mathsf{X}_i) = \frac{P(\mathsf{X}_i) - p_{i-1}(\mathsf{X}_i)}{(1-p_{i-1}^*)(Q(\mathsf{X}_i))}$ **do**
      ▷ *so that the net probability of sampling* X_i *will be* $\alpha_i(\mathsf{X}_i) = P(\mathsf{X}_i) - p_{i-1}(\mathsf{X}_i)$ ◁
      return X_i
  else
    with probability $\beta_i(\mathsf{X}_i) = 1$ **do**
      ▷ *so that the net probability of sampling* X_i *is* $\alpha_i(1 - p_{i-1}^*) \cdot Q(\mathsf{X}_i)$ ◁
      return X_i

In this case, for all x and for all i:

- the probability of accepting x in round i is $\alpha_i(x) = \min\{P(x) - p_{i-1}(x), (1 - p_{i-1}^*)Q(x)\}$
- the probability of accepting x on or before round i is $p_i(x) = p_{i-1}(x) + \alpha_i(x)$
- the probability of terminating on or before round i is $p_i^* = p_{i-1}^* + \sum_{x \in \mathcal{X}} \alpha_i(x) = \sum_{x \in \mathcal{X}} p_i(x)$

> **Example 2.2.6.** Let $P = (\frac{1}{2}, \frac{3}{8}, \frac{1}{8})$ and $Q = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Do the procedure.

Solution. In round 1, sample $\mathsf{X}_1 \sim Q$.

- If $\mathsf{X}_1 = 1$, accept with probability 1.
- If $\mathsf{X}_1 = 2$, accept with probability 1.
- If $\mathsf{X}_1 = 3$, accept with probability $\frac{3}{8}$.

Then, $p_1(1) = \frac{1}{3}$, $p_1(2) = \frac{1}{3}$, $p_1(3) = \frac{1}{8}$, and $p_1^* = \frac{19}{24}$.

In round 2, sample $\mathsf{X}_2 \sim Q$.

- If $\mathsf{X}_2 = 1$, accept with probability 1. There is a $\frac{5}{72}$ chance of getting here, but deficit probability is $\frac{1}{6}$, so no need to reduce.
- If $\mathsf{X}_2 = 2$, accept with probability $\frac{3}{5}$. There is a $\frac{5}{72}$ chance of getting here and deficit probability is $\frac{3}{8} - \frac{1}{3} = \frac{1}{24}$. For equality, use probability $\frac{3}{5} \cdot \frac{5}{72} = \frac{1}{24}$.
- If $\mathsf{X}_3 = 3$, accept with probability 0. We already fulfilled $P(3) = p_1(3)$.

Then, $p_2(1) = \frac{29}{72}$, $p_2(2) = \frac{3}{8}$, $p_3(2) = \frac{1}{8}$, and $p_2^* = \frac{19}{24} + \frac{5}{24} \cdot (\frac{1}{3} + \frac{3/5}{5}) = \frac{65}{72}$.

In round 3, sample $\mathsf{X}_3 \sim Q$.

- If $\mathsf{X}_3 = 1$, accept with probability 1.
- If $\mathsf{X}_3 = 2$ or 3, accept with probability 0.

Keep repeating until we accept a 1. $\square$

Proposition 2.2.7

$(p_i(x))_{x \in \mathcal{X}}$ converges to $P(x)$ as $i \to \infty$. In fact, the residual decays exponentially fast

$$P(x) - p_i(x) \leq P(x) \cdot (1 - Q(x))^i.$$

Proof. Begin with the claim that the probability of reaching round i is at least the residual at i for any x:

$$1 - p_{i-1}^* \geq P(x) - p_{i-1}(x) \quad \forall x$$

Intuitively, either you returned prior to round i (i.e., p_{i-1}^*) or you did not (i.e., the residual).

$$\begin{aligned} 1 - p_{i-1}^* &= \sum_{x \in \mathcal{X}} P(x) - \sum_{x \in \mathcal{X}} p_{i-1}(x) \\ &= \sum_{x \in \mathcal{X}} (P(x) - p_{i-1}(x)) \end{aligned} \tag{2.2}$$

Also, claim that

$$\alpha_i \geq (P(x) - p_{i-1}(x)) \cdot Q(x) \tag{2.3}$$

If $\alpha_i = P(x) - p_{i-1}(x)$, then clearly $\alpha_i \geq \alpha_i Q(x)$. Otherwise, if $\alpha_i = (1 - p_{i-1}^*)Q(x)$, then eq. (2.2) applies.

Proceed by induction.

Base case: exercise.

Inductive step: suppose that $P(x) - p_i(x) \leq P(x) \cdot (1 - Q(x))^i$. Then,

$$\begin{aligned} P(x) - p_{i+1}(x) &= P(x) - p_i(x) - \alpha_{i+1}(x) \\ &\leq (P(x) - p_{i-1}(x))(1 - Q(x)) && \text{(by eq. (2.3))} \\ &\leq (P(x) \cdot (1 - Q(x))^i)(1 - Q(x)) && \text{(by supposition)} \\ &\leq P(x) \cdot (1 - Q(x))^{i+1} \end{aligned}$$

$\square$

Now, we will prove that this is related to relative entropy.

Proposition 2.2.8

Let i^* be the iteration at which the procedure returns. Then, $\mathbb{E}[\log_2 i^*] \leq D(P \parallel Q) + 2\log_2 e$.

Proof. First, claim that for all $x \in \mathcal{X}$ and any $i \geq 2$ such that $\alpha_i(x) > 0$,

$$i \leq \frac{P(x)}{(1-p^*_{i-1}) \cdot Q(x)} + 1 \tag{2.4}$$

That is, if we reach a particular round i, the probability mass left must be sufficiently large.

We know that $P(x) \geq p_{i-1}(x)$ since we increase to $P(x)$. Then,

$$\begin{aligned}
P(x) &\geq p_{i-1}(x) \\
&= \alpha_1(x) + \dots + \alpha_{i-1}(x) \\
&\geq (1-p^*_1) \cdot Q(x) + \dots + (1-p_{i-1}) \cdot Q(x) \\
&\geq (1-p^*_{i-1}) \cdot Q(x) + \dots + (1-p_{i-1}) \cdot Q(x) \\
&= (i-1)(1-p^*_{i-1}) \cdot Q(x) \\
i &\leq \frac{P(x)}{(1-p^*_{i-1}) \cdot Q(x)} + 1
\end{aligned}$$

as long as $\alpha_{j-1} < \alpha_j$ for all j.

Lecture 8
May 29

Do a gigantic algebra bash:

$$\begin{aligned}
\mathbb{E}[\log_2 i^*] &= \sum_{i=1}^{\infty} (p^*_i - p^*_{i-1}) \cdot \log_2 i \\
&= \sum_{i=1}^{\infty} \sum_{x\in\mathcal{X}} \alpha_i(x) \cdot \log_2 i \\
&\leq \sum_{i=1}^{\infty} \sum_{x\in\mathcal{X}} \alpha_i(x) \cdot \log_2 \left[\frac{P(x)}{(1-p^*_{i-1})Q(x)} + 1\right] && \text{(by eq. (2.4))} \\
&\leq \sum_{i=1}^{\infty} \sum_{x\in\mathcal{X}} \alpha_i(x) \cdot \log_2 \left[\frac{1}{(1-p^*_{i-1})}\left(\frac{P(x)}{Q(x)} + 1\right)\right] \\
&= \underbrace{\sum_{i=1}^{\infty} \sum_{x\in\mathcal{X}} \alpha_i(x) \log_2 \frac{1}{(1-p^*_{i-1})}}_{A} + \underbrace{\sum_{i=1}^{\infty} \sum_{x\in\mathcal{X}} \alpha_i(x) \log_2 \left(\frac{P(x)}{Q(x)} + 1\right)}_{B}
\end{aligned}$$

Consider the first term A:

$$\begin{aligned}
A &= \sum_{i=1}^{\infty} \sum_{x\in\mathcal{X}} \alpha_i(x) \log_2 \frac{1}{(1-p^*_{i-1})} \\
&= \sum_{i=1}^{\infty} (p^*_i - p^*_{i-1}) \log_2 \frac{1}{(1-p^*_{i-1})}
\end{aligned}$$

Notice that this is a left-handed Riemann sum of $\log_2 \frac{1}{1-x}$:

$$\begin{aligned}
A &\leq \int_0^1 \log_2 \frac{1}{1-x} \,\mathrm{d}x \\
&= \log_2 e
\end{aligned}$$

Now, consider the second term B:

$$\begin{aligned}
B &= \sum_{i=1}^{\infty} \sum_{x \in \mathcal{X}} \alpha_i(x) \log_2 \left(\frac{P(x)}{Q(x)} + 1 \right) \\
&= \sum_{x \in \mathcal{X}} \sum_{i=1}^{\infty} \alpha_i(x) \log_2 \left(\frac{P(x)}{Q(x)} + 1 \right) && \text{(Fubini?)} \\
&= \sum_{x \in \mathcal{X}} P(x) \log_2 \left(\frac{P(x)}{Q(x)} + 1 \right) && (P(x) = \textstyle\sum_i \alpha_i(x)) \\
&= \sum_{x \in \mathcal{X}} P(x) \log_2 \left(\frac{P(x)}{Q(x)} \cdot \left(1 + \frac{Q(x)}{P(x)} \right) \right) \\
&= \sum_{x \in \mathcal{X}} P(x) \log_2 \left(\frac{P(x)}{Q(x)} \right) + \sum_{x \in \mathcal{X}} P(x) \log_2 \left(1 + \frac{Q(x)}{P(x)} \right) \\
&= D(P \parallel Q) + \sum_{x \in \mathcal{X}} P(x) \log_2 \left(1 + \frac{Q(x)}{P(x)} \right) \\
&\leq D(P \parallel Q) + \sum_{x \in \mathcal{X}} P(x) \log_2 \left(e^{Q(x)/P(x)} \right) && (1 + x \leq e^x \text{ for all } x \geq 0) \\
&= D(P \parallel Q) + \sum_{x \in \mathcal{X}} P(x) \frac{Q(x)}{P(x)} \log_2 e \\
&= D(P \parallel Q) + \log_2 e \sum_{x \in \mathcal{X}} Q(x) \\
&= D(P \parallel Q) + \log_2 e
\end{aligned}$$

Therefore,

$$\mathbb{E}[\log_2 i^*] \leq A + B \leq D(P \parallel Q) + 2 \log_2 e$$

completing the proof. $\square$

Intuition: for any $x \in \mathcal{X}$, if $\alpha_i(x) \leq Q(x) \lll P(x)$, then you need an expected amount of $\frac{P(x)}{Q(x)}$ steps to succeed, because you just won't roll x that often.

Also, if $\alpha_{i+1}(x) > 0$ (any round prior to termination), $(1 - p^*_{i-1}(x))Q(x) \leq \alpha_i(x)$.

Proposition 2.2.9

For any rejection sampler, let i^* be the index where it returns. Then,

$$\mathbb{E}[\ell(i^*)] \geq D(P \parallel Q)$$

Proof. For convenience, redefine $\alpha_i(x) := \Pr[i^* = i \wedge \mathsf{X}_i = x]$.

First, observe that for any $x \in \mathcal{X}$, a rejection sampler must have

$$\alpha_i(x) \leq Q(x)$$

because we only have a $Q(x)$ chance of rolling x to accept it in round i.

Now, fix $x \in \mathcal{X}$. Consider the random variable $i^*|_{\mathsf{X}_{i^*}=x}$. Then, by Kraft's inequality,

$$\begin{aligned}
\mathbb{E}[\ell(i^*) \mid \mathsf{X}_{i^*} = x] &\geq H(i^* \mid \mathsf{X}_{i^*} = x) \\
&= \sum_{i=1}^{\infty} \Pr[i^* = i \mid \mathsf{X}_{i^*} = x] \log_2 \frac{1}{\Pr[i^* = i \mid \mathsf{X}_{i^*} = x]} \\
&= \sum_{i=1}^{\infty} \frac{\alpha_i(x)}{P(x)} \log_2 \frac{P(x)}{\alpha_i(x)} \\
&\geq \sum_{i=1}^{\infty} \frac{\alpha_i(x)}{P(x)} \log_2 \frac{P(x)}{Q(x)} \\
&= \log_2 \frac{P(x)}{Q(x)} \cdot \sum_{i=1}^{\infty} \frac{\alpha_i(x)}{P(x)} \\
&= \log_2 \frac{P(x)}{Q(x)}
\end{aligned}$$

because $\sum_{i=1}^{\infty} \alpha_i(x) = P(x)$. Apply the law of total probability:

$$\begin{aligned}
\mathbb{E}[\ell(i^*)] &= \sum_{x \in \mathcal{X}} \Pr[\mathsf{X}_{i*} = x]\, \mathbb{E}[\ell(i^*) \mid \mathsf{X}_{i^*} = x] \\
&= \sum_{x \in \mathcal{X}} P(x)\, \mathbb{E}[\ell(i^*) \mid \mathsf{X}_{i^*} = x] \\
&\geq \sum_{x \in \mathcal{X}} P(x) \log_2 \frac{P(x)}{Q(x)} \\
&= D(P \parallel Q)
\end{aligned}$$

as desired. $\square$

Chapter 3

Mutual information

3.1 Definition and chain rules

Notation. Given two jointly distributed random variables (X, Y) over sample space $\mathcal{X} \times \mathcal{Y}$, write p_{xy} for $\Pr[\mathsf{X} = x, \mathsf{Y} = y]$.

Lecture 9
June 3

Definition 3.1.1

Given two jointly distributed random variables (X, Y) over sample space $\mathcal{X} \times \mathcal{Y}$, define the mutual information $I(\mathsf{X} : \mathsf{Y})$ by

$$\begin{aligned} I(\mathsf{X} : \mathsf{Y}) &= H(\mathsf{X}) + H(\mathsf{Y}) - H((\mathsf{X}, \mathsf{Y})) \\ &= H(\mathsf{X}) - H(\mathsf{X} \mid \mathsf{Y}) \\ &= H(\mathsf{Y}) - H(\mathsf{Y} \mid \mathsf{X}) \end{aligned}$$

where the conditional entropy $H(\mathsf{X} \mid \mathsf{Y})$ is

$$\sum_{y \in \mathcal{Y}} p_y \cdot H((\mathsf{X}|_{\mathsf{Y}=y}))$$

This is entirely analogous to saying that $|A \cap B| = |A| + |B| - |A \cup B| = |A| - |A \setminus B|$.

Theorem 3.1.2 (chain rule for entropy)

Given two jointly distributed random variables (X, Y) over a discrete sample space $\mathcal{X} \times \mathcal{Y}$,

$$H((\mathsf{X}, \mathsf{Y})) = H(\mathsf{X}) + H(\mathsf{Y} \mid \mathsf{X})$$

Proof. Do a bunch of algebra:

$$\begin{aligned}
H(\mathsf{X}) + H(\mathsf{Y} \mid \mathsf{X}) &= \sum_{x\in\mathcal{X}} p_x \log\frac{1}{p_x} + \sum_{x\in\mathcal{X}} p_x \sum_{y\in\mathcal{Y}} \Pr[\mathsf{Y}=y \mid \mathsf{X}=x] \log\frac{1}{\Pr[\mathsf{Y}=y \mid \mathsf{X}=x]} \\
&= \sum_{x\in\mathcal{X}} p_x \log\frac{1}{p_x} + \sum_{x\in\mathcal{X}} \cancel{p_x} \sum_{y\in\mathcal{Y}} \frac{p_{xy}}{\cancel{p_x}} \log\frac{p_x}{p_{xy}} \\
&= \sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p_{xy} \log\frac{1}{p_x} + \sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p_{xy} \log\frac{p_x}{p_{xy}} \\
&= \sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p_{xy} \left(\log\frac{1}{p_x} + \log\frac{p_x}{p_{xy}}\right) \\
&= \sum_{\substack{x\in\mathcal{X}\\ y\in\mathcal{Y}}} p_{xy} \log\frac{1}{p_{xy}} \\
&= H((\mathsf{X},\mathsf{Y}))
\end{aligned}$$

□

Corollary 3.1.3. For two independent variables, since $(\mathsf{Y} \mid \mathsf{X}) = \mathsf{Y}$, we have $H((\mathsf{X},\mathsf{Y})) = H(\mathsf{X}) + H(\mathsf{Y})$ as expected.

Corollary 3.1.4. $H((\mathsf{X}_1,\mathsf{X}_2,\mathsf{X}_3)) = H(\mathsf{X}_1) + H(\mathsf{X}_2 \mid \mathsf{X}_1) + H(\mathsf{X}_3 \mid (\mathsf{X}_1,\mathsf{X}_2))$

Proof. Consider $(\mathsf{X}_1,\mathsf{X}_2,\mathsf{X}_3) = ((\mathsf{X}_1,\mathsf{X}_2),\mathsf{X}_3)$. Then, by the chain rule for entropy,

$$H(((\mathsf{X}_1,\mathsf{X}_2),\mathsf{X}_3)) = H((\mathsf{X}_1,\mathsf{X}_2)) + H(\mathsf{X}_3 \mid (\mathsf{X}_1,\mathsf{X}_2))$$

and then by another application,

$$H(((\mathsf{X}_1,\mathsf{X}_2),\mathsf{X}_3)) = H(\mathsf{X}_1) + H(\mathsf{X}_2 \mid \mathsf{X}_1) + H(\mathsf{X}_3 \mid (\mathsf{X}_1,\mathsf{X}_2))$$

as desired. □

Theorem 3.1.5 (general chain rule for entropy)
For k random variables $\mathsf{X}_1,\ldots,\mathsf{X}_k$,

$$H((\mathsf{X}_1,\ldots,\mathsf{X}_k)) = \sum_{i=1}^{k} H(\mathsf{X}_i \mid (\mathsf{X}_1,\cdots,\mathsf{X}_{i-1}))$$

Proof. By induction on the chain rule for entropy. □

Notation. Although relative entropy is defined only on *distributions*, write $D(\mathsf{X} \parallel \mathsf{Y})$ to be $D(f_{\mathsf{X}} \parallel f_{\mathsf{Y}})$ where $\mathsf{X} \sim f_{\mathsf{X}}$ and $\mathsf{Y} \sim f_{\mathsf{Y}}$.

Theorem 3.1.6 (chain rule for relative entropy)
Let p and $q : \mathcal{X} \times \mathcal{Y} \to [0,1]$ be distributions. Let $p(x) := \sum_{y \in \mathcal{Y}} p(x,y)$ denote marginals of p and $p(y|x) := \frac{p(x,y)}{p(x)}$ denote conditionals of p. Then,

$$\begin{aligned} D(p(x,y) \parallel q(x,y)) &= D(p(x) \parallel q(x)) + D(p(y|x) \parallel q(y|x)) \\ &= D(p(x) \parallel q(x)) + \sum_{x \in \mathcal{X}} p(x) \cdot D((p(y|x))_{y \in \mathcal{Y}} \parallel (q(y|x))_{y \in \mathcal{Y}}) \end{aligned}$$

where $D(p(y|x) \parallel q(y|x))$ is the conditional relative entropy.

Equivalently, let $(\mathsf{X}_1, \mathsf{Y}_1)$ and $(\mathsf{X}_2, \mathsf{Y}_2)$ be two joint random variables. Then,

$$D((\mathsf{X}_1, \mathsf{Y}_1) \parallel (\mathsf{X}_2, \mathsf{Y}_2)) = D(\mathsf{X}_1 \parallel \mathsf{X}_2) + \sum_{x \in \mathcal{X}} \Pr[\mathsf{X}_1 = x] \cdot D(\mathsf{Y}_1|_{\mathsf{X}_1 = x} \parallel \mathsf{Y}_2|_{\mathsf{X}_2 = x})$$

Proof (for distributions). Do algebra:

$$\begin{aligned} &D(p(x) \parallel q(x)) + D(p(y \mid x) \parallel q(y \mid x)) \\ &= \sum_{x \in \mathcal{X}} p_x \log \frac{p_x}{q_x} + \sum_{x \in \mathcal{X}} p_x \sum_{y \in \mathcal{Y}} p(y \mid x) \log \frac{p(y \mid x)}{q(y \mid x)} \\ &= \sum_{x \in \mathcal{X}} p_x \log \frac{p_x}{q_x} + \sum_{x \in \mathcal{X}} p_x \sum_{y \in \mathcal{Y}} \frac{p_{xy}}{p_x} \log \frac{p_{xy} q_x}{q_{xy} p_x} \\ &= \sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} p_{xy} \log \frac{p_x}{q_x} + \sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} p_{xy} \log \frac{p_{xy} q_x}{q_{xy} p_x} \\ &= \sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} p_{xy} \left(\log \frac{p_x}{q_x} + \log \frac{p_{xy} q_x}{q_{xy} p_x} \right) \\ &= \sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} p_{xy} \log \frac{p_{xy}}{q_{xy}} \\ &= D(p(x,y) \parallel q(x,y)) \end{aligned}$$

as in the proof of chain rule for entropy. □

Fact 3.1.7.

$$I[\mathsf{X} : \mathsf{Y}] = \mathop{\mathbb{E}}_{x \leftarrow \mathsf{X}}[D(\mathsf{Y}|_{\mathsf{X}=x} \parallel \mathsf{Y})] = \sum_{x \in \mathcal{X}} p_x D(\mathsf{Y}|_{\mathsf{X}=x} \parallel \mathsf{Y})$$

Proof. First, claim that

$$I[\mathsf{X} : \mathsf{Y}] = D((\mathsf{X}, \mathsf{Y}) \parallel \tilde{\mathsf{X}} \otimes \tilde{\mathsf{Y}}) \tag{3.1}$$

where $\tilde{\mathsf{X}} \otimes \tilde{\mathsf{Y}}$ denotes a random variable consisting of $\tilde{\mathsf{X}}$ (resp. $\tilde{\mathsf{Y}}$) independently sampled according to the distribution of X (resp. Y) so that $\Pr\left[\tilde{\mathsf{X}} = x, \tilde{\mathsf{Y}} = y\right] = p_x p_y$. Expand the left-hand side:

$$\begin{aligned}
I[\mathsf{X} : \mathsf{Y}] &= \sum_{x \in \mathcal{X}} p_x \log \frac{1}{p_x} + \sum_{y \in \mathcal{Y}} p_y \log \frac{1}{p_y} - \sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} p_{xy} \log \frac{1}{p_{xy}} \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{xy} \log \frac{1}{p_x} + \sum_{y} \sum_{x} p_{xy} \log \frac{1}{p_y} - \sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} p_{xy} \log \frac{1}{p_{xy}} \\
&= \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{xy} \left(\log \frac{1}{p_x} + \log \frac{1}{p_y} - \log \frac{1}{p_{xy}} \right) \\
&= \sum_{\substack{x \in \mathcal{X} \\ y \in \mathcal{Y}}} p_{xy} \log \frac{p_{xy}}{p_x p_y} \\
&= D((\mathsf{X}, \mathsf{Y}) \parallel \tilde{\mathsf{X}} \otimes \tilde{\mathsf{Y}})
\end{aligned}$$

Now, apply the chain rule for relative entropy:

$$\begin{aligned}
D((\mathsf{X}, \mathsf{Y}) \parallel \tilde{\mathsf{X}} \otimes \tilde{\mathsf{Y}}) &= D(\mathsf{X} \parallel \tilde{\mathsf{X}}) + D((\mathsf{X}, \mathsf{Y}) \mid (\mathsf{X}, \tilde{\mathsf{X}}) \parallel (\tilde{\mathsf{X}} \oplus \tilde{\mathsf{Y}}) \mid (\mathsf{X}, \tilde{\mathsf{X}})) \\
&= 0 + \sum_{x} p_x D(\mathsf{Y}|_{\mathsf{X}=x} \parallel \mathsf{Y}) \\
&= \mathop{\mathbb{E}}_{x \leftarrow \mathsf{X}} D(\mathsf{Y}|_{\mathsf{X}=x} \parallel \mathsf{Y})
\end{aligned}$$

□

Lecture 11
June 10

Theorem 3.1.8 (chain rule for mutual information)
Let X_1, X_2, and Y be random variables. Then,

$$I((\mathsf{X}_1, \mathsf{X}_2) : \mathsf{Y}) = I(\mathsf{X}_1 : Y) + I(\mathsf{X}_2 : (Y \mid \mathsf{X}_1))$$

and in general

$$I((\mathsf{X}_1, \dots, \mathsf{X}_n) : \mathsf{Y}) = \sum_{i=1}^{n} I(\mathsf{X}_1 : (\mathsf{Y} \mid (\mathsf{X}_1, \dots, \mathsf{X}_{i-1})))$$

3.2 Markov chains and data processing

Definition 3.2.1
The random variables X, Y, and Z form a Markov chain if the conditional distribution of Z depends only on Y and is conditionally independent of X. Equivalently,

$$\Pr[\mathsf{X} = x, \mathsf{Y} = y, \mathsf{Z} = z] = \Pr[\mathsf{X} = x] \cdot \Pr[\mathsf{Y} = y \mid \mathsf{X} = x] \cdot \Pr[\mathsf{Z} = z \mid \mathsf{Y} = y]$$

Then, we write $\mathsf{X} \to \mathsf{Y} \to \mathsf{Z}$.

Example 3.2.2 (*Legend of the Drunken Master*)**.** In $\Omega = \mathbb{R}^2$, Jackie Chan is drunk and takes steps in random directions. He starts at $\mathsf{J}_0 = (0,0)$. Then, $\mathsf{J}_1 = \mathsf{J}_0 + d_1$ where d_1 is an independent random unit vector in $\mathbb{R}^2$, and $\mathsf{J}_2 = \mathsf{J}_1 + d_2$ and so on.

First, J_3 and J_1 are not independent. But if we fix $\mathsf{J}_2 = j_2 \in \mathbb{R}^2$, then $\mathsf{J}_1 \mid \mathsf{J}_2 = j_2$ and $\mathsf{J}_3 \mid \mathsf{J}_2 = j_2$ are independent. In fact, they are uniformly distributed random points on the circle of radius 1 centred at j_2.

Proposition 3.2.3 (Markov chain characterization)
Let X, Y, and Z be random variables. TFAE:

1. $\mathsf{X} \to \mathsf{Y} \to \mathsf{Z}$
2. X and Z are conditionally independent given Y. That is,

$$\Pr[\mathsf{X} = x, \mathsf{Z} = z \mid \mathsf{Y} = y] = \Pr[\mathsf{X} = x \mid \mathsf{Y} = y] \cdot \Pr[\mathsf{Z} = z \mid \mathsf{Y} = y]$$

3. Z is distributed according to $f(\mathsf{Y}, \mathsf{R})$ for some R independent of X and Y.

Exercise 3.2.4. Prove the definitions are equivalent.

Theorem 3.2.5 (data-processing inequality)
If $\mathsf{X} \to \mathsf{Y} \to \mathsf{Z}$, then $I(\mathsf{X} : \mathsf{Z}) \leq I(\mathsf{X} : \mathsf{Y})$.

Equality happens if and only if $\mathsf{X} \to \mathsf{Z} \to \mathsf{Y}$.

Proof. By the chain rule,

$$I(\mathsf{X} : (\mathsf{Y}, \mathsf{Z})) = I(\mathsf{X} : \mathsf{Y}) + \cancelto{0}{I(\mathsf{X} : \mathsf{Z} \mid \mathsf{Y})} = I(\mathsf{X} : \mathsf{Z}) + I(\mathsf{X} : \mathsf{Y} \mid \mathsf{Z})$$

so that

$$I(\mathsf{X} : \mathsf{Y}) = I(\mathsf{X} : \mathsf{Z}) + I(\mathsf{X} : \mathsf{Y} \mid \mathsf{Z})$$

One may show that the mutual information is always non-negative, so we have $I(\mathsf{X} : \mathsf{Y}) \geq I(\mathsf{X} : \mathsf{Z})$ as desired. We defer the proof of the equality case for section 3.4. ☐

3.3 Communication complexity

Lecture 10
June 5

Problem 3.3.1
Suppose there is a joint distribution (X, Y) that Alice and Bob wish to jointly compute. Alice and Bob have access to a shared random string $\mathsf{R} = (\mathsf{R}_i)$. Alice is given $x \in \mathcal{X}$ and wants to send Bob a prefix-free message of minimum length so that Bob can compute a sample from $\mathsf{Y} \mid \mathsf{X} = x$.

Definition 3.3.2

A protocol Π is a pair of functions (M, y) where $M : \mathcal{X} \times \Omega_{\mathsf{R}} \to \{0,1\}^*$ is the message Alice sends to Bob and $y : \{0,1\}^* \times \Omega_{\mathsf{R}} \to \mathcal{Y}$ is Bob's output.

The performance of Π is $\mathbb{E}_{\mathsf{X},\mathsf{R}} |M(\mathsf{X}, \mathsf{R})|$

Suppose X and Y are independent. Then, Bob needs no information so we can use the trivial protocol $M(\mathsf{X}, \mathsf{R}) = \varnothing$ with performance 0.

Otherwise, we can use a strategy of prefix-free encoding x so that $\mathbb{E}\,|M(\mathsf{X}, \mathsf{R})| \approx H(\mathsf{X})$.

Theorem 3.3.3

There exists a protocol $\Pi = (M, y)$ such that expected message length

$$\mathbb{E}\,|M(\mathsf{X}, \mathsf{R})| \leq I(\mathsf{X} : \mathsf{Y}) + \mathcal{O}(\log I(\mathsf{X} : \mathsf{Y}))$$

For all other protocols $\Pi' = (M', y')$,

$$\mathbb{E}\,|M'(\mathsf{X}, \mathsf{R})| \geq I(\mathsf{X} : \mathsf{Y})$$

Proof. Let X be a random point on the hypercube $\{\pm 1\}^n$. Let Y be a random point on $\{\pm 1\}^n$ that is ε-correlated with X. That is, $\mathsf{Y}_i = \mathsf{X}_i$ with probability ε and is uniformly random otherwise.

Observe that, individually, X and Y have the same distribution. In particular, in the ε case, then $\mathsf{Y}_i = \mathsf{X}_i$ is Uniform$\{\pm 1\}$. In the $1 - \varepsilon$ case, $\mathsf{Y}_i \sim$ Uniform$\{\pm 1\}$ by definition.

We can calculate $H(\mathsf{X}) = H(\mathsf{Y}) = n$.

Also, $H(\mathsf{Y} \mid \mathsf{X}) = \sum_x p_x H(\mathsf{Y} \mid \mathsf{X} = x) \approx (1 - \varepsilon)n$. One can show that $\mathsf{Y} \mid \mathsf{X} = x$ is approximately uniformly distributed over the vectors of length n that agree on εn coordinates with x. This sample space has size $2^{(1-\varepsilon)n}$.

Therefore, $I(\mathsf{X} : \mathsf{Y}) = H(\mathsf{Y}) - H(\mathsf{Y} \mid \mathsf{X}) \approx \varepsilon n$.

By prop. 2.2.8, there exists a rejection sampler such that $\mathbb{E}[\ell(i^*)] \leq D(P \parallel Q) + \mathcal{O}(\log D(P \parallel Q))$.

Recall from STAT 230 that we can transform R into any distribution with the change of variable bullshit. In particular, transform R_i to IID $\mathsf{Y}_i \sim \mathsf{Y}$ and the biased coins.

Alice will run RejectionSampler$(\mathsf{Y}|_{\mathsf{X}=x}, \mathsf{Y})$ to find a random index i^* such that Y_{i^*} has distribution $\mathsf{Y}|_{\mathsf{X}=x}$.

Alice sends a prefix-free encoding of i^*. Bob outputs Y_{i^*}. The performance is:

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathsf{X},\mathsf{R}} |M(\mathsf{X}, \mathsf{R})| &= \sum_{x \in \mathcal{X}} p_x \mathop{\mathbb{E}}_{i^*, \mathsf{Y}_1, \mathsf{Y}_2, \ldots} [\ell(i^*)] \\
&\leq \sum_{x \in \mathcal{X}} p_x (D(\mathsf{Y}|_{\mathsf{X}=x} \parallel \mathsf{Y}) + \mathcal{O}(\log D(\mathsf{Y}|_{\mathsf{X}=x} \parallel \mathsf{Y}))) \\
&= I(\mathsf{X} : \mathsf{Y}) + \sum_{x \in \mathcal{X}} p_x \mathcal{O}(\log D(\mathsf{Y}|_{\mathsf{X}=x} \parallel \mathsf{Y})) \\
&\leq I(\mathsf{X} : \mathsf{Y}) + \mathcal{O}(\log I(\mathsf{X} : \mathsf{Y}))
\end{aligned}$$

where the last step is by Jensen's inequality.

Now, let Π be any protocol. We will apply the data-processing inequality.

Lecture 11 June 10 (con'd)

Notice that $\mathsf{X} \to (M(\mathsf{X}, \mathsf{R}), \mathsf{R}) \to \mathsf{Y}$ if and only if Π is a valid protocol. If we sample $x \sim \mathsf{X}$ and Alice sends $M(x, \mathsf{R})$, then Bob outputs something distributed according to $\mathsf{Y} \mid \mathsf{X} = x$, i.e., just Y since x was arbitrary. Then,

$$\begin{aligned}
I(\mathsf{X} : \mathsf{Y}) &\leq I(\mathsf{X} : (M(\mathsf{X}, \mathsf{R}), \mathsf{R})) && \text{(data processing inequality)} \\
&= I(\mathsf{X} : \mathsf{R}) + \sum_{r \in \Omega_{\mathsf{R}}} p_r I(\mathsf{X}|_{\mathsf{R}=r} : M(\mathsf{X}, \mathsf{R})|_{\mathsf{R}=r}) && \text{(chain rule)} \\
&= 0 + I(\mathsf{X} : M(\mathsf{X}, \mathsf{R}) \mid \mathsf{R}) && \text{(independence)} \\
&\leq H(M(\mathsf{X}, \mathsf{R}) \mid \mathsf{R}) && (I(\mathsf{A} : \mathsf{B}) \leq \min\{H(\mathsf{A}), H(\mathsf{B})\}) \\
&\leq H(M(\mathsf{X}, \mathsf{R})) && (H(\mathsf{A} \mid \mathsf{B}) \leq H(\mathsf{A})) \\
&\leq \mathbb{E}\,|M(\mathsf{X}, \mathsf{R})| && \text{(Kraft inequality)}
\end{aligned}$$

completing the proof. □

3.4 Sufficient statistics

Lecture 12 June 12

We will develop the idea of sufficient statistics and data processing towards the asymptotic equipartition property. This is a warmup for the joint asymptotic equipartition property which we will use to prove one direction of Shannon's channel-coding theorem.

Problem 3.4.1

Suppose $\mathsf{X} = (\mathsf{X}_1, \dots, \mathsf{X}_n)$ are IID sampled according to Bernoulli(θ) for some fixed parameter $\theta \in [0, 1]$.

If we have a sample $x = (x_1, \dots, x_n)$, how can we recover θ?

The classical solution (recall from STAT 230) is the maximum likelihood estimator $\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i$ such that $\Pr\left[\left|\hat{\theta} - \theta\right| > \varepsilon\right] \leq 2^{-\Omega(\varepsilon^2 n)}$. In essence, we are reducing the number of bits to send θ from n to [whatever it is you need to send a float of desired accuracy lol].

Definition 3.4.2

A function $T(\mathsf{X})$ is a <u>sufficient statistic</u> relative to a family $\{f_\theta(x)\}$ if $\theta \to T(\mathsf{X}) \to \mathsf{X}$.

We are considering the case where f_θ is Bernoulli(θ). Clearly, $\theta \to \mathsf{X} \to T(\mathsf{X})$ is a Markov chain because X is distributed based on θ and T is a function of X which is not influenced θ.

Example 3.4.3. $T(\mathsf{X}) = \frac{1}{n} \sum_{i=1}^{n} \mathsf{X}_i$ is a sufficient statistic relative to the family {Bernoulli(θ)}.

Proof. We must show $\theta \to T(\mathsf{X}) \to \mathsf{X}$ is a Markov chain.

Fix $x = (x_1, \dots, x_n)$. Notice that

$$\Pr\left[\mathsf{X}_1 = 0, \dots, \mathsf{X}_n = 0 \mid \frac{1}{n}\sum \mathsf{X}_i = \frac{1}{2}\right] = 0$$

and

$$\Pr\left[\mathsf{X}_1 = 1, \dots, \mathsf{X}_n = 1 \mid \frac{1}{n}\sum \mathsf{X}_i = \frac{1}{2}\right] = 0$$

since we obviously cannot have half the X_i's be 1 if they are all 0s or all 1s.

But if we set exactly half of the X_i's to be 1, the distribution is uniform

$$\Pr\left[\mathsf{X}_1 = 1, \dots, \mathsf{X}_{\frac{n}{2}} = 1, \mathsf{X}_{\frac{n}{2}+1} = 0, \dots, \mathsf{X}_n = 0 \mid \frac{1}{n}\sum \mathsf{X}_i = \frac{1}{2}\right] = \Pr\left[\mathsf{X} = x \mid \frac{1}{n}\sum \mathsf{X}_i = \frac{1}{2}\right] = \frac{1}{\binom{n}{n/2}}$$

for all $x \in \{0,1\}^n$ such that $\frac{n}{2}$ entries are 1.

More generally, suppose x has exactly k ones where $k = n\bar{\theta}$. Then,

$$\Pr\left[\mathsf{X} = x \mid \frac{1}{n}\sum \mathsf{X}_i = \bar{\theta}\right] = \begin{cases} 1/\binom{n}{n\bar{\theta}} & \frac{1}{n}\sum x_i = \bar{\theta} \\ 0 & \text{otherwise} \end{cases}$$

so we have that $\mathsf{X} \mid \frac{1}{n}\sum \mathsf{X}_i = \bar{\theta}$ is independent of θ.

We can also see this by saying that $\mathsf{X} \sim \mathrm{Bernoulli}(\theta)^n$ can be equivalently sampled as:

1. first sampling $\mathsf{K} = k$ with probability $\Pr\left[\frac{1}{n}\sum \mathsf{X}_i = k\right]$,
2. then sampling a uniform random point that has exactly K ones.

which clearly shows that X can be sampled as $f(\frac{1}{n}\sum \mathsf{X}_i, \mathsf{R})$ for some new randomness R (the uniform randomness) independent of θ. □

> **Example 3.4.4** ("mostly unrelated *Drunken Master III*"). A public domain generic drunkard legally distinct from Jackie Chan begins at $(0,0)$ and takes steps in random directions d_i of length $\ell \sim \left|\mathcal{N}(0,\theta^2)\right|$.

Let X_n be the position at time n. We can show that

$$\|\mathsf{X}_n\|_2 = c(1 \pm o(1))\theta\sqrt{n}$$

with probability very close to 1. To be more precise,

$$\Pr[\text{length from origin} > (1 + o(1))(\text{expected length from origin})]$$

is exponentially small in n. That is, after n steps, the randomness cancels out, and we have a pretty good idea of where we end up.

The whole point of this exercise is to notice that if we have a sufficient statistic, the probability measure is extremely concentrated around some constant, and we can almost just treat the statistic as a constant itself.

Example 3.4.5. Consider IID Gaussians $\mathsf{X}_1, \dots, \mathsf{X}_n \sim \mathcal{N}(0,1)$. Then, what is the probability $\Pr\left[\mathsf{X}_1, \dots, \mathsf{X}_n > t\sqrt{n}\right]$ we overshoot the estimator by t times?

Solution. Apply simple properties of Gaussians from STAT 230:

$$\Pr\left[\mathsf{X}_1, \dots, \mathsf{X}_n > t\sqrt{n}\right] = \Pr\left[\sqrt{n}\mathcal{N}(0,1) > t\sqrt{n}\right] = \Pr[\mathcal{N}(0,1) > t] = \Phi(t) \approx e^{-t^2/2}$$

□

Lemma 3.4.6 (rotation invariance of the Gaussian)

Let X be a Gaussian and O be an orthonormal matrix. Then, $O\mathsf{X}$ is distributed identically to X.

Proof (super sketchy). Consider IID $\mathsf{X}_1, \dots, \mathsf{X}_n \sim \mathcal{N}(0,1)$. Then, since $p(x_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x_i^2}{2}\right)$, we have

$$p(x_1, \dots, x_n) = \frac{1}{\sqrt{2\pi}^n} \exp\left(-\frac{\|x\|_2^2}{2}\right)$$

Notice that this only depends on the length of x, so we are uniformly distributing on the n-ball of length $\|x\|_2$. □

Now consider what's going on with a summation. Notice that $\sum \mathsf{X}_i = \langle \mathsf{X}, \mathbb{1} \rangle$. There exists some rotation O such that $O\mathbb{1} = \sqrt{n}e_1$ (the first basis vector). Inner products preserve rotations, so $\sum \mathsf{X}_i = \langle O\mathsf{X}, O\mathbb{1} \rangle = \sqrt{n} \langle O\mathsf{X}, e_1 \rangle = \sqrt{n} O\mathsf{X}_1$. But by rotation invariance, this has the same distribution as $\sqrt{n}\mathsf{X}_1$, which is just a Gaussian.

Chapter 4

Coding theory

Chapter 5

Parallel repetition

List of Named Results

Index of Defined Terms