STATS 700, Fall 2024
LLMs and Transformers
HW 3

Ambuj Tewari

Due September 17

You are allowed to use LLMs provided you declare precisely how used them (but please be aware that they can hallucinate). You cannot search online or ask anyone. Discussing problems with other students in the course is allowed but all solutions should be your own work. If two students submit verbatim copies of any solution, it will be treated as a violation of academic integrity. Submit your solutions as a typeset PDF file (no handwritten scans please) on Canvas.

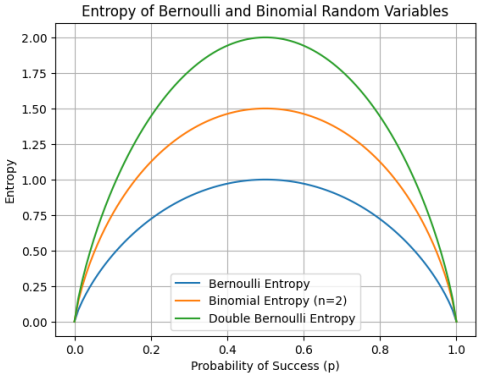
1 Entropy of Binomial vs Bernoulli: Special Case

Let $H_n(p)$ be the entropy of a Binomial(n, p) random variable. Note that $H_1(p)$ is therefore just the entropy of a Bernoulli(p) random variable which we will abbreviate to $H(p)$.

Note that a Bernoulli random variable's pmf has probability values $1-p$ and p, and that the pmf of a Binomial$(2, p)$ has probability values $(1-p)^2$, $2p(1-p)$, and p^2. We therefore have the explicit expressions:

$$H(p) = -(1-p)\log(1-p) - p\log p \ ,$$

$$H_2(p) = -(1-p)^2\log(1-p)^2 - 2p(1-p)\log(2p(1-p)) - p^2\log p^2 \ .$$



The figure above is a "proof by picture" that $H(p) \leq H_2(p) \leq 2H(p)$ for all $p \in [0, 1]$. Without using any information inequalities, provide a formal proof of these inequalities (using calculus, elementary algebra, etc.). To see the power of information inequalities, you will now derive much more general results with much less effort.

2 Entropy of Sums of Random Variables I

Suppose that $X_1, \ldots, X_n$ are independent (but not necessarily identically distributed) random variables. Is it always true that

$$H(X_1 + \ldots + X_n) \leq H(X_1) + \ldots + H(X_n) \ ?$$

Hint: Consider the simple case of two independent random variables X, Y first. Set $Z = X + Y$ and note that it is a function of X, Y.

3 Entropy of Sums of Random Variables II

Suppose that $X_1, \ldots, X_n$ are independent (but not necessarily identically distributed) random variables. Is it always true that

$$H(X_1 + \ldots + X_n) \geq \max\{H(X_1), \ldots, H(X_n)\} \ ?$$

Hint: Consider the simple case of two independent random variables X, Y first. Set $Z = X + Y$ and note that $I(Z; X) \geq 0$.

4 Entropy of Binomial vs Bernoulli: General Case

Using what you proved above, now show that for any positive integer n and $p \in [0, 1]$,

$$H(p) \leq H_n(p) \leq nH(p) \ ,$$

where we are re-using notation defined in Problem 1.

Do you think you can show this directly (without any appeals to information inequalities) using the Binomial pmf $p_{\text{binom}}(i) = \binom{n}{i} p^i (1-p)^{n-i}$ and the definition of entropy?

Remark: Actually, the upper bound you proved above is nowhere close to how fast $H_n(p)$ actually grows with n for fixed p. It grows logarithmically with n, not linearly! Approximating the Binomial(n, p) with a Normal distribution with mean $\mu = np$ and variance $\sigma^2 = np(1-p)$ and using the fact that the entropy of a Normal distribution is $\frac{1}{2}\log(2\pi e\sigma^2)$, we should expect the entropy of Binomial(n, p) to grow as $\frac{1}{2}\log(2\pi enp(1-p))$, By the way, Normal is not a discrete distribution, so we have to use what is called differential entropy but this is a heuristic argument anyway! The large gap makes intuitive sense: the joint entropy of n independent Bernoulli(p) outcomes $(X_1, \ldots, X_n)$ does scale (exactly!) as $nH(p)$ but there is a huge reduction of information in going from that to just the sum $S_n = X_1 + \ldots + X_n$. Entropies of sums are not easy to work with. For example, it was only proved in 2004[1] that entropy of the normalized sum $S_n/\sqrt{n}$ grows monotonically in n for any iid (and square integrable) $X_1, \ldots, X_n$!

5 Perplexity in a Simple IID World

Assume that a language has words $W_1, W_2, \ldots$ drawn iid from the fixed distribution p over a vocabulary V. You model the language as if words are drawn iid from a distribution q over V. Show that the perplexity

$$\sqrt[n]{\frac{1}{q(W_1)q(W_2)\cdots q(W_n)}}$$

converges almost surely to a deterministic limit as $n \to \infty$ and find the limit.

[1] Solution of Shannon's problem on the monotonicity of entropy, Shiri Artstein, Keith M. Ball, Franck Barthe and Assaf Naor. J. Amer. Math. Soc. **17** (2004), 975-982 DOI: `https://doi.org/10.1090/S0894-0347-04-00459-X`

6 Perplexity in a Simple Markov World

Assume that a language has words $W_1, W_2, \ldots$ drawn from the fixed Markov chain with kernel $p(x, x')$ (probability of the chain jumping from word x to word x') and initial distribution $p_0(x)$ over a vocabulary V. You model the language as if words are drawn from a Markov chain with kernel $q(x, x')$ and an initial distribution $q_0(x)$. Show that the perplexity

$$\sqrt[n]{\frac{1}{q_0(W_1)q(W_1, W_2)\cdots q(W_{n-1}, W_n)}}$$

converges almost surely to a deterministic limit as $n \to \infty$ and find the limit. Assume that the Markov chain induced by the kernel $p(\cdot, \cdot)$ mixes exponentially fast to a stationary distribution $\pi(x)$ over V.

7 Shannon Entropy as a Tsallis Entropy

For any real $q \neq 1$, define the Tsallis entropy as

$$\frac{1}{q-1}\left(1 - \sum_x p(x)^q\right).$$

Show that Tsallis entropy becomes Shannon entropy as $q \to 1$.

8 Shannon Entropy as a Rényi Entropy

For any $\alpha > 0, \alpha \neq 1$, define the Rényi entropy as

$$\frac{1}{1-\alpha}\log\left(\sum_x p(x)^\alpha\right).$$

Show that Rényi entropy becomes Shannon entropy as $\alpha \to 1$.

9 A Consistent Estimator of Entropy

Let $X_1, \ldots, X_n$ be iid copies of a discrete random variable X with finite support and entropy H. Define

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{X_i = x}$$

and

$$\hat{H} = -\sum_x \hat{p}(x)\log\hat{p}(x)$$

Show that $\hat{H} \to H$ almost surely.

10 Concentration of Entropy's Biased Estimate

The estimate mentioned in the problem above is not unbiased. Prove that $\mathbb{E}[\hat{H}] \leq H$. Also prove that an unbiased estimator of entropy does not exist. Finally, show that the estimator concentrates sharply around its expected value. In particular, show that there exist positive universal constants C, c such that for every distribution of X, and all $n \geq 1, \epsilon > 0$, we have

$$\mathbb{P}\left(|\hat{H} - \mathbb{E}[\hat{H}]| > \epsilon\right) \leq C\exp(-cn\epsilon^2/\log^2 n)\ .$$