Machine Learning Notes

This note is based on Machine Learning at DIKU

5 December 2021

Chapter 2. Concentration of Measure Inequalities

2.4 Basics of Information Theory: Entropy, Relative Entropy, and the Method of Types

In this section we briefly introduce a number of basic concepts from information theory that are very useful for deriving concentration inequalities. Specifically, we introduce the notions of entropy and relative entropy (Cover and Thomas 2006, Chapter 2)[7,8,9] and some basic tools from

[7] We first introduce the concept of *entropy*, which is a measure of the uncertainty of a random variable. Let X be a discrete random variable with alphabet $\mathcal{X}$ and probability mass function $p(x) = \mathbf{Pr}\{X = x\}, x \in \mathcal{X}$. We denote the probability mass function by $p(x)$ rather than $p_X(x)$, for convenience. Thus, $p(x)$ and $p(y)$ refer to two different random variables and are in fact different probability mass functions, $p_X(x)$ and $p_Y(y)$, respectively.

Remark (Definition). *The entropy* $\mathbf{H}(X)$ *of a discrete random variable* X *is defined by* $\mathbf{H}(X) = -\sum_{x\in\mathcal{X}} p(x) \log p(x)$.

We also write $H(p)$ for the above quantity. The $\log(\cdot)$ is to the base 2 and entropy is expressed in bits. For example, the entropy of a fair coin toss is 1 bit. We will use the convention that $0 \log 0 = 0$, which is easily justfied by continuity since $x \log x \to 0$ as $x \to 0$. Adding terms of zero probability does not change the entropy.

If the base of the logarithm is b, we denote the entropy as $\mathbf{H}_b(X)$. If the base of the logarithm is e, the entropy is measured in *nats*. Unless otherwise specified, we will take all logarithms to base 2, and hence all the entropies will be measured in bits. Note that entropy is a functional of the distribution of X. It does not depend on the actual values taken by the random variable X, but only on the probabilities.

We denote expectation by $\mathbf{E}[\cdot]$. Thus, if $X \sim p(x)$, the expected value of the random variable $g(X)$ is written as $\mathbf{E}_p[g(X)] = \sum_{x\in\mathcal{X}} g(x)p(x)$or more simply as $\mathbf{E}[g(X)]$ when the probability mass function is understood from the context. We shall take a peculiar interest in the eerily self-referential expectation of $g(X)$ under $p(x)$ when $g(X) = \log \frac{1}{p(X)}$. *NB.* When $g(X) = \log \frac{1}{p(X)}$, then $\mathbf{E}[g(X)] = \sum_{x\in\mathcal{X}} p(x) \log \frac{1}{p(x)} = -\mathbf{H}(p)$. When $f(x) = \log \frac{1}{x}$, then its derivates are $f'(x) = \frac{1}{x^{-1}}(-x^{-2}) = -x^{-1} = -\frac{1}{x}$. Note that $(e^x)' = e^x$ and $(\ln x)' = 1/x$.

Remark. *The entropy of* X *can also be interpreted as the expected value of the random variable* $\log(1/p(X))$*, where* X *is drawn according to probability mass function* $p(x)$*. Thus,* $\mathbf{H}(X) = \mathbf{E}_p\left[\log\left(1/p(X)\right)\right]$.

This definition of entropy is related to the definition of entropy in thermodynamics; some of the connections are explored later. It is possible to derive the definition of entropy axiomatically by defining certain properties that the entropy of a random variable must satisfy. This approach is illustrated in Problem 2.46. We do not use the axiomatic approach to justify the definition of entropy; instead, we show that it arises as the answer to a number of natural questions, such as "What is the average length of the shortest description of the random variable?"

[8] Wikipedia (nat (unit)). The *natural unit of information* (symbol: nat), sometimes also nit or nepit, is a unit of information, based on natural logarithms and powers of e, rather than the powers of 2 and base 2 logarithms, which define the shannon. This unit is also known by its unit symbol, the nat. One nat is the information content of an event when the probability of that event occurring is $1/e$, where e is the Euler's number. One nat is equal to $\frac{1}{\ln 2}$ shannons ≈ 1.44 Sh or, equivalently, $\frac{1}{\ln 10}$ hartleys ≈ 0.434 Hart.

[9] First, we derive some immediate consequences of the definition.

Remark (Lemma 2.1.1). $\mathbf{H}(X) \geqslant 0$. *Proof.* $0 \leqslant p(x) \leqslant 1$ *implies that* $\log(1/p(x)) \geqslant 0$ *due to* $1/p(x) \in [1, +\infty)$.

Remark (Lemma 2.1.2). $\mathbf{H}_b(X) = (\log_b a)\mathbf{H}_a(X)$. *Proof.* $\log_b p = \log_b a \log_a p$ *dule to* $\log_b a = \ln a / \ln b$.

the method of types (Cover and Thomas 2006, Chapter 11). We start with some definitions, and we have special interest in Bernoulli random variables.

Definition 2.7 (Entropy). *Let $p(x)$ be a distribution of a discrete random variable X taking values in a finite set $\mathcal{X}$. We define the* entropy *of p as*

$$\mathbf{H}(p) = -\sum_{x\in\mathcal{X}} p(x)\ln p(x)\,. \tag{2.68}$$

We use the convention that $0\ln 0 = 0$ (which is justified by continuity of $z\ln z$, since $z\ln z \to 0$ as $z \to 0$).

Definition 2.8 (Bernoulli random variable). *X is a Bernoulli random variable with bias p if X accepts values in $\{0,1\}$ with $\mathbf{Pr}(X=0) = 1-p$ and $\mathbf{Pr}(X=1) = p$.*

Note that expectation of a Bernoulli random variable is equal to its bias, that is,

$$\mathbf{E}[X] = 0\times\mathbf{Pr}(X=0) + 1\times\mathbf{Pr}(X=1) = \mathbf{Pr}(X=1) = p\,. \tag{2.69}$$

With a slight abuse of notation we specialise the definition of entropy to Bernoulli random variables. Note that when we talk about Bernoulli random variables p denotes the bias of the random variable and when we talk about more general random variables p denotes the complete distribution. Entropy is one of the central quantities in information theory and it has numerous applications. We start by using binary entropy to bound binomial coefficients.

Definition 2.9 (Binary entropy). *Let p be a bias of Bernoulli random variable X. We define the entropy of p as*

$$\mathbf{H}(p) = -p\ln p - (1-p)\ln(1-p)\,. \tag{2.70}$$

NB. As p is the bias of Bernoulli random variable X, we have that $\mathbf{E}[X] = p$, where $\mathbf{Pr}(X=0) = 1-p$, and $\mathbf{Pr}(X=1) = p$. Then $\mathbf{H}(p) = -\sum_{x=0}^{1}\mathbf{Pr}(X=x)\ln\mathbf{Pr}(X=x)$, that is, Eq. (2.70).

Lemma 2.10.

$$\frac{1}{n+1}e^{n\,\mathbf{H}\left(\frac{k}{n}\right)} \leqslant \binom{n}{k} \leqslant e^{n\,\mathbf{H}\left(\frac{k}{n}\right)}\,. \tag{2.71}$$

(Note that $\frac{k}{n}\in[0,1]$ and $\mathbf{H}(\frac{k}{n})$ in the lemma is the binary entropy.)

Proof. By the binomial formula we know that for any $p\in[0,1]$,

$$\sum_{i=0}^{n}\binom{n}{i}p^i(1-p)^{n-i} = 1\,. \tag{2.72}$$

We start with the upper bound. Take $p = \frac{k}{n}$. Since the sum is larger than any individual term, for the k-th term of the sum we get

$$\begin{aligned} 1 \geqslant \binom{n}{k}p^k(1-p)^{n-k} &= \binom{n}{k}\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k} = \binom{n}{k}\left(\frac{k}{n}\right)^k\left(\frac{n-k}{n}\right)^{n-k} \\ &= \binom{n}{k}e^{k\ln\frac{k}{n}+(n-k)\ln\frac{n-k}{n}} = \binom{n}{k}e^{n\left(\frac{k}{n}\ln\frac{k}{n}+\frac{n-k}{n}\ln\frac{n-k}{n}\right)} = \binom{n}{k}e^{-n\,\mathbf{H}\left(\frac{k}{n}\right)}\,. \end{aligned}$$

The second property of entropy enables us to change the base of the logarithm in the definition. Entropy can be changed from one base to another by multiplying by the appropriate factor.

By changing sides of the inequality we obtain the upper bound.

NB. Eq. (2.72) holds because the sum of all possibilities is one. Besides, for any $k \in \{0, 1, ..., n\}$,

$$\sum_{k'=0}^{n} \binom{n}{k'} p^{k'}(1-p)^{n-k'} \geqslant \binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k} \left(\frac{k}{n}\right)^k \left(1 - \frac{k}{n}\right)^{n-k} = \binom{n}{k} \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k} \tag{2.74a}$$

$$= \binom{n}{k} e^{\ln\left(\left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k}\right)} = \binom{n}{k} e^{k \ln \frac{k}{n} + (n-k) \ln \frac{n-k}{n}} \tag{2.74b}$$

$$= \binom{n}{k} e^{n\left(\frac{k}{n} \ln \frac{k}{n} + \frac{n-k}{n} \ln \frac{n-k}{n}\right)} = \binom{n}{k} e^{n\left(\frac{k}{n} \ln \frac{k}{n} + \left(1 - \frac{k}{n}\right) \ln \left(1 - \frac{k}{n}\right)\right)} \tag{2.74c}$$

$$= \binom{n}{k} e^{n \cdot \left(-\mathbf{H}\left(\frac{k}{n}\right)\right)} = \binom{n}{k} e^{-n\mathbf{H}\left(\frac{k}{n}\right)}, \tag{2.74d}$$

therefore,

$$\binom{n}{k} \leqslant e^{n\mathbf{H}\left(\frac{k}{n}\right)} = \exp\left(n\mathbf{H}\left(\frac{k}{n}\right)\right). \tag{2.75}$$

For the lower bound it is possible to show that if we fix $p = \frac{k}{n}$ then $\binom{n}{k} p^k (1-p)^{n-k} \geqslant \binom{n}{i} p^i (1-p)^{n-i}$ for any $i \in \{0, ..., n\}$, see (Cover and Thomas 2006, Example 11.1.3)[10,11,12] for details. We

[10] See the mentioned example (pp. 353) as follows. We give a slightly better approximation for the binary case. These bounds can be proved using Stirling's approximation for the factorial function (Lemma 17.5.1).

Remark (Example 11.1.3, Binary alphabet). *In this case, the type is defined by the number of* 1*'s in the sequence, and the size of the type class is therefore* $\binom{n}{k}$. *We show that* $\frac{1}{n+1} 2^{n\mathbf{H}\left(\frac{k}{n}\right)} \leqslant \binom{n}{k} \leqslant 2^{n\mathbf{H}\left(\frac{k}{n}\right)}$.

Remark (Lemma 17.5.1). *For* $0 < p < 1$, $q = 1 - p$, *such that* np *is an integer,* $\frac{1}{\sqrt{8npq}} \leqslant \binom{n}{np} 2^{-n\mathbf{H}(p)} \leqslant \frac{1}{\sqrt{\pi npq}}$.

Remark (Theorem 11.1.3, Size of a type class $T(P)$). *For any type* $P \in \mathcal{P}_n$, $\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{n\mathbf{H}(P)} \leqslant |T(P)| \leqslant 2^{n\mathbf{H}(P)}$.

[11] The AEP for discrete random variables (Chapter 3) focuses our attention on a small subset of typical sequences. The method of types is an even more powerful procedure in which we consider sequences that have the same empirical distribution. With this restriction, we can derive strong bounds on the number of sequences with a particular empirical distribution and the probability of each sequence in this set. It is then possible to derive strong error bounds for the channel coding theorem and prove a variety of rate distortion results. The method of types was fully developed by (Csiszár and Körner 2011), who obtained most of their results from this point of view. (see Chapter 11, pp. 347)

Let $X_1, X_2, ..., X_n$ be a sequence of n symbols from an alphabet $\mathcal{X} = \{a_1, a_2, ..., a_{|\mathcal{X}|}\}$. We use the notation x^n and $\mathbf{x}$ interchangeably to denote a sequence $x_1, x_2, ..., x_n$. The type of a sequence $\mathbf{x}$ is denoted as $P_{\mathbf{x}}$. It is a probability mass function on $\mathcal{X}$. (Note that in this chapter, we will use capital letters to denote types and distributions. We also loosely use the word *distribution* to mean a probability mass function.)

Remark (Definition). *The type* P_x *(or empirical probability distribution) of a sequence* $x_1, x_2, ..., x_n$ *is the relative proportion of occurrences of each symbol of* $\mathcal{X}$ *(i.e.,* $P_{\mathbf{x}}(a) = N(a|\mathbf{x})/n$ *for all* $a \in \mathcal{X}$, *where* $N(a|\mathbf{x})$ *is the number of times the symbol* a *occurs in the sequence* $\mathbf{x} \in \mathcal{X}^n$*).*

[12] In information theory, the analog of the law of large numbers is the asymptotic equipartition property (AEP). It is a direct consequence of the weak law of large numbers. The *law of large numbers* states that for independent, identically distributed (i.i.d.) random variables, $\frac{1}{n}\sum_{i=1}^{n} X_i$ is close to its expected value $\mathbf{E}[X]$ for large values of n. The AEP states that $\frac{1}{n} \log \frac{1}{p(X_1, X_2, ..., X_n)}$ is close to the entropy $\mathbf{H}(\cdot)$, where $X_1, X_2, ..., X_n$ are i.i.d. random variables and $p(X_1, X_2, ..., X_n)$ is the probability of observing the sequence $X_1, X_2, ..., X_n$. Thus, the probability $p(X_1, X_2, ..., X_n)$ assigned to an observed sequence will be close to $2^{-n\mathbf{H}}$. (see Chapter 3, pp. 57)

This enables us to divide the set of all sequences into two sets, the *typical set*, where the sample entropy is close to the true entropy, and the nontypical set, which contains the other sequences. Most of our attention will be on the typical sequences. Any property that is proved for the typical sequences will then be true with high probability and will determine the average behaviour of a large sample.

also note that there are $(n+1)$ elements in the sum in Eq. (2.72). Again, take $p = \frac{k}{n}$, then

$$1 \leqslant (n+1)\max_i \binom{n}{i}\left(\frac{k}{n}\right)^i\left(\frac{n-k}{n}\right)^{n-i} = (n+1)\binom{n}{k}\left(\frac{k}{n}\right)^k\left(\frac{n-k}{n}\right)^{n-k} = (n+1)\binom{n}{k}e^{-n\mathbf{H}\left(\frac{k}{n}\right)}, \tag{2.76}$$

where the last step follows the same step as in the derivation of the upper bound.

NB. According to (Cover and Thomas 2006, Example 11.1.3, pp. 353),[13] we have that:

$$\frac{\mathbf{Pr}(S=i+1)}{\mathbf{Pr}(S=i)} = \frac{C_n^{i+1}p^{i+1}(1-p)^{n-(i+1)}}{C_n^i p^i(1-p)^{n-i}} = \frac{\frac{n!}{(n-(i+1))!(i+1)!}p}{\frac{n!}{(n-i)!i!}(1-p)} = \frac{(n-i)!i!}{(n-i-1)!(i+1)!}\frac{p}{1-p} = \frac{n-i}{i+1}\frac{p}{1-p}, \tag{2.80a}$$

$$1 = \sum_{k'=0}^{n}\binom{n}{k'}p^{k'}(1-p)^{n-k'} \leqslant (n+1)\max_{k'}\binom{n}{k'}p^{k'}(1-p)^{n-k'} = (n+1)\binom{n}{k}p^k(1-p)^{n-k}, \tag{2.80b}$$

$$\text{s.t. } k \overset{\text{def}}{=} \operatorname*{argmax}_{k'\in\{0,1,\ldots,n\}}\binom{n}{k'}p^{k'}(1-p)^{n-k'}. \tag{2.80c}$$

Therefore,

$$1 \leqslant (n+1)\binom{n}{k}p^k(1-p)^{n-k} = (n+1)\binom{n}{k}\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k} = (n+1)\binom{n}{k}\left(\frac{k}{n}\right)^k\left(\frac{n-k}{n}\right)^{n-k} \tag{2.81a}$$

$$= (n+1)\binom{n}{k}e^{\ln\left(\left(\frac{k}{n}\right)^k\left(\frac{n-k}{n}\right)^{n-k}\right)} = (n+1)\binom{n}{k}e^{k\ln\frac{k}{n}+(n-k)\ln\frac{n-k}{n}} \tag{2.81b}$$

$$= (n+1)\binom{n}{k}e^{n\left(\frac{k}{n}\ln\frac{k}{n}+\frac{n-k}{n}\ln\frac{n-k}{n}\right)} = (n+1)\binom{n}{k}e^{n\left(\frac{k}{n}\ln\frac{k}{n}+\left(1-\frac{k}{n}\right)\ln\left(1-\frac{k}{n}\right)\right)} \tag{2.81c}$$

$$= (n+1)\binom{n}{k}e^{n\left(-\mathbf{H}\left(\frac{k}{n}\right)\right)} = (n+1)\binom{n}{k}e^{-n\mathbf{H}\left(\frac{k}{n}\right)}, \tag{2.81d}$$

and consequently,

$$\binom{n}{k} \geqslant \frac{1}{n+1}e^{n\mathbf{H}\left(\frac{k}{n}\right)}. \tag{2.82}$$

■

[13] For the lower bound, let S be a random variable with a binomial distribution with parameters n and p. The most likely value of S is $S = \langle np\rangle$. This can easily be verified from the fact that

$$\frac{\mathbf{Pr}(S=i+1)}{\mathbf{Pr}(S=i)} = \frac{n-i}{i+1}\cdot\frac{p}{1-p}, \tag{2.77}$$

and considering the cases when $i < np$ and when $i > np$. Then, since there are $(n+1)$ terms in the binomial sum,

$$1 = \sum_{k=0}^{n}\binom{n}{k}p^k(1-p)^{n-k} \leqslant (n+1)\max_k\binom{n}{k}p^k(1-p)^{n-k} \tag{2.78a}$$

$$= (n+1)\binom{n}{\langle np\rangle}p^{\langle np\rangle}(1-p)^{n-\langle np\rangle}. \tag{2.78b}$$

Now let $p = k/n$. Then we have $1 \leqslant (n+1)\binom{n}{k}\left(\frac{k}{n}\right)^k\left(1-\frac{k}{n}\right)^{n-k}$, which by the arguments in the upper bound is equivalent to $\frac{1}{n+1} \leqslant \binom{n}{k}2^{-n\mathbf{H}(k/n)}$, or $\binom{n}{k} \geqslant \frac{1}{n+1}2^{n\mathbf{H}(k/n)}$. Combining the two results, we see that $\binom{n}{k} \approx 2^{n\mathbf{H}\left(\frac{k}{n}\right)}$. A more precise bound can be found in theorem 17.5.1 when $k \neq 0$ or n.

Remark (Proof of Lemma 17.5.1). *We begin with a strong form of Stirling's approximation (Feller 1957), which states that*

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \leqslant n! \leqslant \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\frac{1}{12n}}. \tag{2.79}$$

Applying this to find an upper bound, and the lower bound is obtained similarly.

Lemma 2.10 shows that the number of configurations of chosing k out of n objects is directly related to the entropy of the imbalance $\frac{k}{n}$ between the number of objects that are selected (k) and the number of objects that are left out $(n-k)$. We now introduce one additional quantity, the *Kullback-Leibler (KL) divergence*, also known as *Kullback-Leibler distance* and as *relative entropy*.

Definition 2.11 (Relative entropy or Kullback-Leibler divergence). *Let $p(x)$ and $q(x)$ be two probability distributions of a random variable X (or two probability density functions, if X is a continuous random variable), the* Kullback-Leibler divergence *or* relative entropy *is defined as*

$$\mathbf{KL}(p\|q) = \mathbf{E}_p\left[\ln\frac{p(X)}{q(X)}\right] = \begin{cases} \sum_{x\in\mathcal{X}} p(x)\ln\frac{p(x)}{q(x)}\,, & \textit{if } \mathcal{X} \textit{ is discrete}\,; \\ \int_{x\in\mathcal{X}} p(x)\ln\frac{p(x)}{q(x)}dx\,, & \textit{if } \mathcal{X} \textit{ is continuous}\,. \end{cases} \tag{2.83}$$

We use the convention that $0\ln\frac{0}{0} = 0$ and $0\ln\frac{0}{q} = 0$ and $p\ln\frac{p}{0} = \infty$.

Definition 2.12 (Binary kl-divergence). *Let p and q be biases of two Bernoulli random variables. The* binary kl divergence *is defined as*

$$\mathbf{kl}(p\|q) = \mathbf{KL}\big([1-p,p]\big\|[1-q,q]\big) = p\ln\tfrac{p}{q} + (1-p)\ln\tfrac{1-p}{1-q}\,. \tag{2.84}$$

We specialise the definition to Bernoulli distributions. KL divergence is the central quantity in information theory. Although it is not a distance measure, because it does not satisfy the triangle inequality, it is the right way of measuring distances between probability distributions. This is illustrated by the following example.

Example 2.13. *Let $X_1, ..., X_n$ be an i.i.d. sample of n Bernoulli random variables with bias p and let $\frac{1}{n}\sum_{i=1}^n X_i$ be the empirical bias of the sample. (Note that $\frac{1}{n}\sum_{i=1}^n X_i \in \{0, \frac{1}{n}, \frac{2}{n}, ..., \frac{n}{n}\}$.) Then by Lemma 2.10,*

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^n X_i = \frac{k}{n}\right) = \binom{n}{k}p^k(1-p)^{n-k} \leqslant e^{n\mathbf{H}\left(\frac{k}{n}\right)}e^{n\left(\frac{k}{n}\ln p + \frac{n-k}{n}\ln(1-p)\right)} = e^{-n\mathbf{kl}\left(\frac{k}{n}\|p\right)}\,, \tag{2.85}$$

and

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^n X_i = \frac{k}{n}\right) \geqslant \frac{1}{n+1}e^{-n\mathbf{kl}\left(\frac{k}{n}\|p\right)}\,. \tag{2.86}$$

Thus, $\mathbf{kl}(\frac{k}{n}\|p)$ governs the probability of observing empirical bias $\frac{k}{n}$ when the true bias is p. It is easy to verify that $\mathbf{kl}(p\|p) = 0$ and it is also possible to show that $\mathbf{kl}(\hat{p}\|p)$ is convex in $\hat{p}$ and that $\mathbf{kl}(\hat{p}\|p) \geqslant 0$. Thus, the probability of empirical bias is maximised when it coincides with the true bias.

NB. We have known that by Lemma 2.10, $\binom{n}{k}e^{-n\mathbf{H}\left(\frac{k}{n}\right)} \in [\frac{1}{n+1}, 1]$, and $-\mathbf{H}\left(\frac{k}{n}\right) = \frac{k}{n}\ln\frac{k}{n} + \frac{n-k}{n}\ln\frac{n-k}{n}$. Then,

$$\binom{n}{k} \leqslant e^{n\mathbf{H}\left(\frac{k}{n}\right)} = e^{-n\left(\frac{k}{n}\ln\frac{k}{n} + \frac{n-k}{n}\ln\frac{n-k}{n}\right)} = e^{-k\ln\frac{k}{n} - (n-k)\ln\frac{n-k}{n}}\,, \tag{2.87a}$$

$$p^k(1-p)^{n-k} = e^{\ln\left(p^k(1-p)^{n-k}\right)} = e^{k\ln p + (n-k)\ln(1-p)}\,, \tag{2.87b}$$

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^n X_i = \frac{k}{n}\right) = \binom{n}{k}p^k(1-p)^{n-k} \leqslant e^{-k\left(\ln\frac{k}{n} - \ln p\right) - (n-k)\left(\ln\frac{n-k}{n} - \ln(1-p)\right)} \tag{2.87c}$$

$$= e^{-n\left(\frac{k}{n}\ln\frac{k/n}{p} + \frac{n-k}{n}\ln\frac{(n-k)/n}{1-p}\right)} = e^{-n\mathbf{kl}\left(\frac{k}{n}\|p\right)}\,. \tag{2.87d}$$

Similarly,

$$\binom{n}{k} \geqslant \frac{1}{n+1} e^{n\mathbf{H}\left(\frac{k}{n}\right)} = \frac{1}{n+1} e^{-n\left(\frac{k}{n}\ln\frac{k}{n}+\frac{n-k}{n}\ln\frac{n-k}{n}\right)} = \frac{1}{n+1} e^{-k\ln\frac{k}{n}-(n-k)\ln\frac{n-k}{n}}, \tag{2.88a}$$

$$\mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{k}{n}\right) = \binom{n}{k} p^k (1-p)^{n-k} \geqslant \frac{1}{n+1} e^{-k\left(\ln\frac{k}{n}-\ln p\right)-(n-k)\left(\ln\frac{n-k}{n}-\ln(1-p)\right)} \tag{2.88b}$$

$$= \frac{1}{n+1} e^{-n\left(\frac{k}{n}\ln\frac{k/n}{p}+\frac{n-k}{n}\ln\frac{(n-k)/n}{1-p}\right)} = \frac{1}{n+1} e^{-n\mathbf{kl}\left(\frac{k}{n}\middle\|p\right)}. \tag{2.88c}$$

Overall,

$$\frac{1}{n+1} e^{-n\mathbf{kl}\left(\frac{k}{n}\middle\|p\right)} \leqslant \mathbf{Pr}\left(\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{k}{n}\right) \leqslant e^{-n\mathbf{kl}\left(\frac{k}{n}\middle\|p\right)}. \tag{2.89}$$

Additionally, we could obtain that

$$\mathbf{kl}(p\|p) = p\ln\frac{p}{p} + (1-p)\ln\frac{1-p}{1-p} = p\ln 1 + (1-p)\ln 1 = 1\ln 1 = 0, \tag{2.90a}$$

$$\mathbf{kl}(\hat{p}\|p) = \hat{p}\ln\frac{\hat{p}}{p} + (1-\hat{p})\ln\frac{1-\hat{p}}{1-p}. \tag{2.90b}$$

Question: How to explain $\mathbf{kl}(\hat{p}\|p)$?

As we all know, a convex function means that for all $0 < t < 1$ and all $x_1, x_2 \in \mathcal{X}$ such that $x_1 \neq x_2$, it holds that $f\left(tx_1 + (1-t)x_2\right) \leqslant tf(x_1) + (1-t)f(x_2)$. Besides, $\mathbf{kl}(\hat{p}\|p) = \mathbf{kl}\left(\frac{k}{n}\|p\right)$ where $k/n, p \in [0,1]$, thus, we have $1 - k/n, 1-p \in [0,1]$. Because

$$\frac{1}{n+1} e^{-n\mathbf{kl}\left(\frac{k}{n}\|p\right)} \leqslant \mathbf{Pr}\left(\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{k}{n}\right) \leqslant 1, \tag{2.91}$$

then we get

$$-n\mathbf{kl}\left(\frac{k}{n}\middle\|p\right) \leqslant \ln(n+1), \tag{2.92a}$$

$$\mathbf{kl}\left(\frac{k}{n}\middle\|p\right) \leqslant \frac{1}{n}\ln(n+1) = \ln\sqrt[n]{n+1}. \tag{2.92b}$$

Because $e^{-n\mathbf{kl}\left(\frac{k}{n}\|p\right)} \geqslant 0$, then we have $\mathbf{Pr}(\hat{p} = k/n) \geqslant \frac{1}{n+1} e^{-n\mathbf{kl}(\hat{p}\|p)} \geqslant 0$.

Question: But I still have no idea how to demonstrate $\mathbf{kl}(\hat{p}\|p)$ is convex in $\hat{p}$.

If we would like to demonstrate that $\mathbf{kl}(\hat{p}\|p)$ is convex in $\hat{p}$, we need to demonstrate that: for all $0 < t < 1$ and all x_1, x_2 such that $x_1 \neq x_2$, it holds

$$\mathbf{kl}(tx_1 + (1-t)x_2\|p) \leqslant t\mathbf{kl}(x_1\|p) + (1-t)\mathbf{kl}(x_2\|p), \tag{2.93}$$

that is to say, for all $0 < t < 1$ and all $k_i \in [0,n], k_i \in \mathbb{Z}, \forall i \in \{1,2\}$ such that $k_1 \neq k_2$, it holds

$$\mathbf{kl}\left(t\tfrac{k_1}{n} + (1-t)\tfrac{k_2}{n}\|p\right) \leqslant t\mathbf{kl}\left(\tfrac{k_1}{n}\|p\right) + (1-t)\mathbf{kl}\left(\tfrac{k_2}{n}\|p\right). \tag{2.94}$$

References

Cover, Thomas M and Joy A Thomas (2006). *Elements of Information Theory*. Second Edition. Wiley Series in Telecommunications and Signal Processing.

Csiszár, Imre and János Körner (2011). *Information theory: coding theorems for discrete memoryless systems*. Cambridge University Press.

Feller, William (1957). *An Introduction to Probability Theory and Its Applications*. Second Edition. New York: John Wiley & Sons.