Problem Set 2 **(Graded)** —*Due Tuesday, Oct 1, before class starts*

For the Exercise Sessions on September 17 and 24

Last name	First name	SCIPER Nr	Points

Rules : You are allowed and encouraged to discuss these problems with your colleagues. However, you are expected to write the final solutions yourself. If you collaborated on a homework, write down the name of your collaborators and your sources. No points will be deducted for collaborations. Any evidence of *plagiarism*, however, will be treated very seriously.

Note : Only some of the problems in this problem set will be graded, but these will not be revealed beforehand. Therefore, you are expected to submit your solutions for all of the problems in this problem set.

Assume log is base 2 for this problem set unless the problem says otherwise.

Problem 1: Bounded random variables are subgaussian

This problem is a guided proof of a slightly weakened version of Lemma 2.4.

(a) Prove the following inequality:

$$\cosh(x) = (e^x + e^{-x})/2 \leq e^{x^2/2}. \tag{1}$$

(b) Using the previous inequality, give an upper bound on the moment generating function of a random variable S that only takes the values $+1$ and -1, with equal probability.

Hint: The upper bound should depend on the parameter of the moment generating function.

(c) Consider any random variable X and let X' be a random variable *independent of* X, but with exactly the same distribution. Show that

$$\mathbb{E}_X[e^{\lambda(X-\mathbb{E}[X])}] \leq \mathbb{E}_{X,X'}[e^{\lambda(X-X')}]. \tag{2}$$

(d) Show that the random variables $(X - X')$ and $S(X - X')$, where S is as in Part (b) and assumed independent of X and X', have the same distribution.

(e) From the previous part, we thus know that

$$\mathbb{E}_{X,X'}[e^{\lambda(X-X')}] = \mathbb{E}_{S,X,X'}[e^{\lambda S(X-X')}]. \tag{3}$$

Now assume that X is a bounded random variable, $X \in [a, b]$. Condition on $X = x$ and $X' = x'$, and take expectation over S. Observe that $(x - x')^2 \leq (b - a)^2$. Use this and your result from Part (b) to further upper bound $\mathbb{E}_{S,X,X'}[e^{\lambda S(X-X')}]$.

(f) Combine your results to give an upper bound on the moment generating function of a centered bounded random variable $X - \mathbb{E}[X]$, where $X \in [a, b]$.

Hint: The upper bound should depend on the parameter of the moment generating function as well as a and b.

(g) Compare your result to Lemma 2.4. Discuss the differences.

Solution 1. (a) To prove this inequality, we can proceed via the expansions of the exponential function. Specifically,

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \frac{x^5}{5!} + \frac{x^6}{6!} + \dots$$
$$e^{-x} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} - \frac{x^5}{5!} + \frac{x^6}{6!} - \dots$$

Adding up and dividing by 2,

$$\cosh(x) = 1 + \frac{x^2}{2!} + \frac{x^4}{4!} + \frac{x^6}{6!} \dots + \frac{x^{2n}}{(2n)!} + \dots$$

Expanding $e^{x^2/2}$ via the standard expansion for e^y,

$$e^{x^2/2} = 1 + \frac{x^2}{2} + \frac{x^4}{4 \cdot 2!} + \frac{x^6}{8 \cdot 3!} \dots + \frac{x^{2n}}{2^n n!} + \dots$$

A term-by-term comparison and noting that $2^n n! \leq (2n)!$ gives the claimed bound.

(b) Simply write out

$$\mathbb{E}\left[e^{\lambda S}\right] = \sum_s p_S(s) e^{\lambda s} = \frac{1}{2}e^{\lambda} + \frac{1}{2}e^{-\lambda} \leq e^{\lambda^2/2}.$$

(c) Write out using the independence of X and X'

$$\mathbb{E}_{X,X'}[e^{\lambda(X-X')}] = \mathbb{E}_X\left[\mathbb{E}_{X'}[e^{\lambda(X-X')}]\right].$$

Now, for the inner expectation, we apply Jensen's inequality, noting that the exponential function is convex:

$$\mathbb{E}_{X'}[e^{\lambda(X-X')}] \geq e^{\mathbb{E}_{X'}[\lambda(X-X')]} = e^{\lambda(X - \mathbb{E}_{X'}[X'])} = e^{\lambda(X - \mathbb{E}[X])},$$

where we have used the linearity of expectation and the fact that X and X' have the same distribution.

(d) For example, we can argue via the CDF. First, we observe that since X and X' are indistinguishable, we must have for any real number y

$$\mathbb{P}(X - X' \leq y) = \mathbb{P}(X' - X \leq y)$$

But then, by conditioning, we must have for any real number y

$$\begin{aligned}\mathbb{P}(S(X - X') \leq y) &= \mathbb{P}(S = 1)\mathbb{P}((X - X') \leq y) + \mathbb{P}(S = -1)\mathbb{P}(-(X - X') \leq y) \\ &= \mathbb{P}(X - X' \leq y),\end{aligned}$$

which proves the claim.

(e) Following the instruction, we write

$$\mathbb{E}_{S,X,X'}[e^{\lambda S(X-X')}] = \mathbb{E}_{X,X'}\left[\mathbb{E}_S[e^{\lambda S(X-X')}]|X, X'\right].$$

Now, for any fixed values $X = x$ and $X' = x'$, we have, using Part (b),

$$\mathbb{E}_S[e^{\lambda S(x-x')}] \leq e^{\lambda^2(x-x')^2/2} \leq e^{\lambda^2(b-a)^2/2}.$$

Hence,

$$\mathbb{E}_{X,X'}[e^{\lambda(X-X')}] = \mathbb{E}_{S,X,X'}[e^{\lambda S(X-X')}] \leq e^{\lambda^2(b-a)^2/2}.$$

(f) Combining everything, we have

$$\begin{aligned}\mathbb{E}_X[e^{\lambda(X-\mathbb{E}[X])}] &\leq \mathbb{E}_{X,X'}[e^{\lambda(X-X')}] \\ &= \mathbb{E}_{S,X,X'}[e^{\lambda S(X-X')}] \\ &\leq e^{\lambda^2(b-a)^2/2}.\end{aligned}$$

(g) In the lecture notes, we have shown that for a bounded random variable X, the moment generating function satisfies

$$\mathbb{E}_X[e^{\lambda(X-\mathbb{E}[X])}] \leq e^{\frac{\lambda^2(b-a)^2}{8}}.$$

So, the proof above establishes also that bounded random variables are subgaussian, but with a suboptimal parameter: the argument developed here says that bounded random variables are $(b-a)^2$-subgaussian, where with the more intricate argument from your homework, you can actually show that they are $(b-a)^2/4$-subgaussian. Needless to say, for many proofs, these two results are equally interesting, and there is only a small gain to be had from the factor of 4 improvement in the exponent.

Problem 2: Axiomatic definition of entropy

Let $(p_1, p_2, \ldots, p_m)$ be such that $p_i \geq 0$ for $i = 1, \ldots, m$ and $\sum_i p_i = 1$. Let

$$H_m(p_1, \ldots, p_m) = -\sum_{i=1}^{m} p_i \log p_i \tag{4}$$

be the entropy of $(p_1, p_2, \ldots, p_m)$.

(a) *(Grouping property)* Prove that

$$H_m(p_1, p_2, p_3, \ldots, p_m) = H_{m-1}(p_1 + p_2, p_3, \ldots, p_m) + (p_1 + p_2) H_2\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right).$$

Also prove it for grouping p_i and p_j for any arbitrary pair of indices (i, j). This property models the fact that the uncertainty in choosing among m objects should be equal to the uncertainty in first choosing a subgroup of the objects, and then choosing an object in the selected subgroup.

(b) Prove that if a sequence of functions F_m of probability vectors $(p_1, p_2, \ldots, p_m)$, is such that for every $m \geq 2$,

1. $F_m(p_1, p_2, \ldots, p_m)$ is continuous in the p_i's,
2. $F_m(p_1, p_2, \ldots, p_m)$ satisfies the grouping property (a),
3. $F_m(\frac{1}{m}, \ldots, \frac{1}{m}) = \log m$

then F_m must be equal to the entropy (4) (under the usual convention $0 \log 0 = 0$).

Hint: Suppose that the $p_i's$ are rational, i.e., $p_i = \frac{n_i}{n}$ for some positive integers $\{n_i\}_{i=1,\ldots,m}$. Show using (a) recursively that

$$F_n\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) = F_m\left(\frac{n_1}{n}, \ldots, \frac{n_m}{n}\right) + \sum_i \frac{n_i}{n} F_{n_i}\left(\frac{1}{n_i}, \ldots, \frac{1}{n_i}\right).$$

Solution 2. *(a)* Using (4), we can rewrite the right-hand side as

$$\begin{aligned}
&H(p_1+p_2,p_3,\dots,p_m)+(p_1+p_2)H\left(\frac{p_1}{p_1+p_2},\frac{p_2}{p_1+p_2}\right)\\
&\quad=-(p_1+p_2)\log(p_1+p_2)-\sum_{i=3}^{m}p_i\log p_i+(p_1+p_2)\left(-\frac{p_1}{p_1+p_2}\log\frac{p_1}{p_1+p_2}-\frac{p_2}{p_1+p_2}\log\frac{p_2}{p_1+p_2}\right)\\
&\quad=-(p_1+p_2)\log(p_1+p_2)-\sum_{i=3}^{m}p_i\log p_i-p_1\log p_1-p_2\log p_2+(p_1+p_2)\log(p_1+p_2)\\
&\quad=-\sum_{i=1}^{m}p_i\log p_i=H(p_1,p_2,p_3,\dots,p_m).
\end{aligned}$$

(b) It can be proved by induction that the grouping property holds for grouping an arbitrary number of elements. Hence, using it recursively on $F\left(\frac{1}{m},\dots,\frac{1}{m}\right)$, we get

$$F\left(\frac{1}{m},\dots,\frac{1}{m}\right)=F\left(\frac{m_1}{m},\dots,\frac{m_k}{m}\right)+\sum_i\frac{m_i}{m}F\left(\frac{1}{m_i},\dots,\frac{1}{m_i}\right).$$

Using property 3 on $F\left(\frac{1}{m},\dots,\frac{1}{m}\right)$ and on each $F\left(\frac{1}{m_i},\dots,\frac{1}{m_i}\right)$, we get

$$\log m=F\left(\frac{m_1}{m},\dots,\frac{m_k}{m}\right)+\sum_i\frac{m_i}{m}\log m_i.$$

Rearranging the last equation gives

$$F\left(\frac{m_1}{m},\dots,\frac{m_k}{m}\right)=-\sum_i\frac{m_i}{m}\log\frac{m_i}{m}.$$

This proves the result for every rational probability vector. By using the continuity of F (property 1), we can extend the result to any probability vector.

Problem 3: Conditional KL divergence

We saw in class that a *probability kernel* $P_{Y|X}:\mathcal{X}\to\mathcal{Y}$ is a matrix $P_{Y|X}=P_{Y|X}(y|x):x\in\mathcal{X},y\in\mathcal{Y}$ such that $P_{Y|X}(y|x)\geq 0$, and for each $x\in\mathcal{X},\sum_y P_{Y|X}(y|x)=1$. Let $P_X\in\Pi(\mathcal{X})$ be a probability distribution on $\mathcal{X}$. We define the *conditional KL divergence* between two probability kernels $P_{Y|X}:\mathcal{X}\to\mathcal{Y}$ and $Q_{Y|X}:\mathcal{X}\to\mathcal{Y}$ given P_X to be

$$D(P_{Y|X}\|Q_{Y|X}|P_X)\triangleq\sum_{x\in\mathcal{X}}P_X(x)D(P_{Y|X}(\cdot|x)\|Q_{Y|X}(\cdot|x))$$

where for every x, $D(P_{Y|X}(\cdot|x)\|Q_{Y|X}(\cdot|x))$ is the standard KL divergence between the two distributions $P_{Y|X}(\cdot|x)$ and $Q_{Y|X}(\cdot|x)$ over $\mathcal{Y}$.

(a) *(Chain rule of the KL divergence)* Show that

$$D(P_{X,Y}\|Q_{X,Y})=D(P_X\|Q_X)+D(P_{Y|X}\|Q_{Y|X}|P_X)$$

where $P_{X,Y}$ and $Q_{X,Y}$ are two joint distributions on $\mathcal{X}\times\mathcal{Y}$ such that $P_{X,Y}(x,y)=P_X(x)P_{Y|X}(y|x)$ and $Q_{X,Y}(x,y)=Q_X(x)Q_{Y|X}(y|x)$.

(b) Using (a), show that

$$D(P_{Y|X}\|Q_{Y|X}|P_X)=D(P_{X,Y}\|Q_{X,Y})$$

where $P_{X,Y}(x,y)=P_X(x)P_{Y|X}(y|x)$ and $Q_{X,Y}(x,y)=P_X(x)Q_{Y|X}(y|x)$.

(c) *(Conditioning increases divergence)* Using (b) and the Data Processing Inequality seen in class, show that

$$D(P_Y\|Q_Y) \le D(P_{Y|X}\|Q_{Y|X}|P_X)$$

where $P_Y(y) = \sum_{x\in\mathcal{X}} P_X(x)P_{Y|X}(y|x)$ and $Q_Y(y) = \sum_{x\in\mathcal{X}} P_X(x)Q_{Y|X}(y|x)$.

Solution 3. *(a)*

$$\begin{aligned}
D(P_{XY}\|Q_{XY}) &= \sum_{x,y} P_{XY}(x,y)\log\frac{P_{XY}(x,y)}{Q_{XY}(x,y)} \\
&= \sum_{x,y} P_X(x)P_{Y|X}(y|x)\log\frac{P_X(x)P_{Y|X}(y|x)}{Q_X(x)Q_{Y|X}(y|x)} \\
&= \sum_{x,y} P_X(x)P_{Y|X}(y|x)\log\frac{P_X(x)}{Q_X(x)} + \sum_{x,y} P_X(x)P_{Y|X}(y|x)\log\frac{P_{Y|X}(y|x)}{Q_{Y|X}(y|x)} \\
&= D(P_X\|Q_X) + \sum_x P_X(x)D(P_{Y|X}(\cdot|x)\|Q_{Y|X}(\cdot|x)) = D(P_X\|Q_X) + D(P_{Y|X}\|Q_{Y|X}|P_X).
\end{aligned}$$

(b)

$$D(P_{XY}\|Q_{XY}) = D(P_X\|P_X) + D(P_{Y|X}\|Q_{Y|X}|P_X) = D(P_{Y|X}\|Q_{Y|X}|P_X).$$

(c) Define the kernel

$$W(\tilde{y}|x,y) = \begin{cases} 1, & \text{if } \tilde{y} = y, \\ 0, & \text{otherwise.} \end{cases}$$

Then we have $P_{\tilde{Y}}(\tilde{y}) = \sum_{x,y} P_{XY}(x,y)W(\tilde{y}|x,y) = P_Y(\tilde{y})$ and $Q_{\tilde{Y}}(\tilde{y}) = \sum_{x,y} Q_{XY}(x,y)W(\tilde{y}|x,y) = Q_Y(\tilde{y})$. Hence, we have

$$D(P_{Y|X}\|Q_{Y|X}|P_X) = D(P_{XY}\|Q_{XY}) \ge D(P_{\tilde{Y}}\|Q_{\tilde{Y}}) = D(P_Y\|Q_Y).$$

where the equality follows from part (b) and the inequality follows from DPI.

Problem 4: Geometrical interpretation of mutual information

In the previous problem, we introduced the conditional KL divergence between two probability kernels $P_{Y|X} : \mathcal{X} \to \mathcal{Y}$ and $Q_{Y|X} : \mathcal{X} \to \mathcal{Y}$ given a distribution P_X over $\mathcal{X}$ as

$$D(P_{Y|X}\|Q_{Y|X}|P_X) \triangleq \sum_{x\in\mathcal{X}} P_X(x)D(P_{Y|X}(\cdot|x)\|Q_{Y|X}(\cdot|x)),$$

where for every $x \in \mathcal{X}$, $D(P_{Y|X}(\cdot|x)\|Q_{Y|X}(\cdot|x))$ is the standard KL divergence between the two distributions $P_{Y|X}(\cdot|x)$ and $Q_{Y|X}(\cdot|x)$ over $\mathcal{Y}$.

(a) Let X and Y be two random variables with joint distribution $P_{XY} = P_X P_{Y|X}$. Show that

$$I(X;Y) = \sum_{x\in\mathcal{X}} P_X(x)D(P_{Y|X}(\cdot|x)\|P_Y)$$

where P_Y is the marginal distribution of Y. This formula shows that the mutual information can be interpreted as a weighted average of the distances between the conditional distributions $P_{Y|X}(\cdot|x)$ and the marginal distribution P_Y.

(b) Show that for any distribution Q_Y on $\mathcal{Y}$,

$$I(X;Y) = D(P_{Y|X}\|Q_Y|P_X) - D(P_Y\|Q_Y).$$

You can think of this formula as a KL equivalent of the classical $I(X;Y) = H(Y) - H(Y|X)$.

(c) Show that

$$I(X;Y) = \min_{Q_Y} D(P_{Y|X}\|Q_Y|P_X).$$

According to this formula, the minimizing Q_Y can be interpreted as the "center of gravity" of the conditional distributions $P_{Y|X}(\cdot|x)$, and the mutual information as its radius.

Solution 4. All the results can be proved working directly with the definitions of KL divergence and mutual information. The following is a simple solution that makes use of the results proved in Homework 2, Problem 3.

(a)

$$I(X;Y) = D(P_X P_{Y|X}\|P_X P_Y) = D(P_{Y|X}\|P_Y|P_X) = \sum_{x\in\mathcal{X}} P_X(x) D(P_{Y|X}(\cdot|x)\|P_Y),$$

where the second equality is due to Homework 2, Problem 3(b).

(b)

$$\begin{aligned} D(P_Y\|Q_Y) + I(X;Y) &= D(P_Y\|Q_Y) + D(P_{X|Y}\|P_X|P_Y) \\ &= D(P_{XY}\|P_X Q_Y) \\ &= D(P_{Y|X}\|Q_Y|P_X) \end{aligned}$$

where the first equality is due to part (a) by exchanging the roles of X and Y, the second equality is due to the chain rule of the KL divergence (Homework 2, Problem 3(a)), and the third equality is again due to Homework 2, Problem 3(b).

(c) By part (b) we know that $I(X;Y) \le D(P_{Y|X}\|Q_Y|P_X)$ for every Q_Y, since $D(P_Y\|Q_Y) \ge 0$. Hence, $I(X;Y) \le \min_{Q_Y} D(P_{Y|X}\|Q_Y|P_X)$. The equality is achieved by picking $Q_Y = P_Y$, for which $D(P_{Y|X}\|Q_Y|P_X) = D(P_{Y|X}\|P_Y|P_X) = I(X;Y)$.

Problem 5: Entropy and combinatorics

Let $n \ge 1$ and fix some $0 \le k \le n$. Let $p = \frac{k}{n}$ and let $T_p^n \subset \{0,1\}^n$ be the set of all binary sequences with exactly np ones (assume that np is an integer).

(a) Show that

$$\log|T_p^n| = nh(p) + O(\log_e n)$$

where $h(p) = -p\log_e p - (1-p)\log_e(1-p)$ is the binary entropy function.

Hint: Stirling's approximation states that for every $n \ge 1$,

$$e^{\frac{1}{12n+1}}\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \le n! \le e^{\frac{1}{12n}}\sqrt{2\pi n}\left(\frac{n}{e}\right)^n$$

(b) Let $Q^n = \text{Bernoulli}(q)^n$ be the i.i.d. Bernoulli distribution on $\{0,1\}^n$. Show that

$$\log Q^n[T_p^n] = -nd(p\|q) + O(\log_e n)$$

where $d(p\|q) = p\log_e\frac{p}{q} + (1-p)\log_e\frac{1-p}{1-q}$ is the binary KL divergence.

Solution 5. *(a)* When $p = 0$ or 1, we have $|T_p^n| = 1$, or equivalently $\log|T_p^n| = 0$, so the result holds trivially, since $h(p) = 0$ for $p = 0, 1$. For $p \ne 0, 1$, we have that $|T_p^n| = \binom{n}{np} = \frac{n!}{(np)!(n(1-p))!}$. Using Stirling's approximation on the three factorials we get

$$\begin{aligned} &\frac{1}{\sqrt{2\pi np(1-p)}} p^{-np}(1-p)^{-n(1-p)} e^{\frac{1}{12n+1} - \frac{1}{12np} - \frac{1}{12n(1-p)}} \le |T_p^n| \\ &\qquad\qquad \le \frac{1}{\sqrt{2\pi np(1-p)}} p^{-np}(1-p)^{-n(1-p)} e^{\frac{1}{12n} - \frac{1}{12np+1} - \frac{1}{12n(1-p)+1}}. \end{aligned}$$

By taking the log on each side, we get

$$nh(p) - \frac{1}{2}\log(2\pi np(1-p)) + \frac{1}{12n+1} - \frac{1}{12np} - \frac{1}{12n(1-p)} \le \log|T_p^n| \le nh(p) - \frac{1}{2}\log(2\pi np(1-p)) + \frac{1}{12n} - \frac{1}{12np+1} - \frac{1}{12n(1-p)+1}.$$

Since $\frac{1}{n} \le p \le \frac{n-1}{n}$ and the same holds for $1-p$, we can obtain the following (loose) bounds:

$$-\frac{1}{2}\log n + \frac{1}{2}\log(2\pi) \le \frac{1}{2}\log(2\pi np(1-p)) \le \frac{1}{2}\log n + \frac{1}{2}\log(2\pi)$$

$$\frac{1}{12n+1} - \frac{1}{12np} - \frac{1}{12n(1-p)} \ge -2$$

$$\frac{1}{12n} - \frac{1}{12np+1} - \frac{1}{12n(1-p)+1} \le 1$$

so that we get

$$nh(p) - \frac{1}{2}\log n - \frac{1}{2}\log(2\pi) - 2 \le \log|T_p^n| \le nh(p) + \frac{1}{2}\log n - \frac{1}{2}\log(2\pi) + 1$$

i.e., $\log|T_p^n| = nh(p) + O(\log n)$.

(b) We have

$$Q^n[T_p^n] = \binom{n}{np} q^{np}(1-q)^{n(1-p)} = |T_p^n| q^{np}(1-q)^{n(1-p)}$$

and therefore

$$\begin{aligned}\log Q^n[T_p^n] &= \log|T_p^n| + np\log q + n(1-p)\log(1-q) \\ &= nh(p) + np\log q + n(1-p)\log(1-q) + O(\log n) \\ &= -nd(p\|q) + O(\log n)\end{aligned}$$

where in the last step we used (a).

Problem 6: Sum of binomials

Looking at the part (a) previous problem, it can be seen that the entropy function is related to the asymptotic value of the binomial coefficient by:

$$\log\binom{n}{np} = nh(p) + O(\log_e n),$$

for $n \ge 1$ and $0 \le p \le 1$, where $h(p) \triangleq -p\log_e p - (1-p)\log_e(1-p)$ is the binary entropy function. We want to derive a similar bound for the sum of binomial coefficients.

(a) Fix $0 \le p \le 1/2$ and let $\mathcal{C}$ be the set of all subsets of $\{1, 2, \ldots, n\}$ of size at most np. Let X be a random variable uniformly distributed over $\mathcal{C}$. Show that

$$H(X) \le nh(p).$$

Hint: Let $(X_1, X_2, \ldots, X_n)$ be a random vector such that for every i, $X_i = 1$ if $i \in X$, and $X_i = 0$ otherwise. Argue that $H(X) = H(X_1, X_2, \ldots, X_n)$.

(b) Using part (a), conclude that

$$\sum_{i=0}^{\lfloor np \rfloor} \binom{n}{i} \le 2^{nh(p)}.$$

(c) Using part (b), show that if $Z \sim \text{Binomial}(n, p = \frac{1}{2})$, then

$$\Pr\left(\left|Z - \frac{n}{2}\right| \geq c\sigma\right) \leq 2^{1-c^2/2}$$

for every $c \geq 0$, where $\sigma = \frac{\sqrt{n}}{2}$ is the standard deviation of Z. Compare this bound with the Hoeffding inequality for a σ^2-subgaussian random variable we derived in class.

Hint: you can use (without proving it) the bound $h(p) \leq 1 - 2\left(\frac{1}{2} - p\right)^2$.

Solution 6. (a) There is a one-to-one correspondence between X and $(X_1, X_2, \ldots, X_n)$: from the value of X we can uniquely determine the value of $(X_1, X_2, \ldots, X_n)$, and viceversa. Hence, $H(X) = H(X_1, X_2, \ldots, X_n)$. Then,

$$H(X) = H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i) = nH(X_1)$$

where the last equality is due to symmetry.

Now, X takes values from the set C, therefore the expected cardinality of X is less than or equal to np. Since the cardinality of X is equal to the sum of the indicator functions $\sum_{i=1}^{n} 1(X_i = 1)$, we have $np \geq E(|X|) = E(\sum_{i=1}^{n} 1(X_i = 1)) = \sum_{i=1}^{n} E(1(X_i = 1)) = \sum_{i=1}^{n} P(X_i = 1)$ using the linarity of the expectation and properties of the indicator function.

Then, due to the symmetry, $np \; geq \sum_{i=1}^{n} P(X_i = 1) = nP(X_1 = 1)$. And, $P(X_1 = 1) \leq p$ follows.

Now, $\Pr(X_1 = 1) \leq p \leq \frac{1}{2}$, and therefore $H(X_1) \leq h(p)$. Hence, $H(X) \leq nh(p)$.

(b)

$$H(X) = \log|\mathcal{C}| = \log \sum_{i=0}^{\lfloor np \rfloor} \binom{n}{i} \leq nh(p).$$

Hence,

$$\sum_{i=0}^{\lfloor np \rfloor} \binom{n}{i} \leq 2^{nh(p)}.$$

(c)

$$\begin{aligned}
\Pr\left(\left|Z - \frac{n}{2}\right| \geq c\frac{\sqrt{n}}{2}\right) &= 2\left(\frac{1}{2}\right)^n \sum_{i=0}^{\left\lfloor n\left(\frac{1}{2} - \frac{c}{2\sqrt{n}}\right)\right\rfloor} \binom{n}{i} \\
&\leq 2^{nh\left(\frac{1}{2} - \frac{c}{2\sqrt{n}}\right) - n + 1} \\
&\leq 2^{n\left(1 - \frac{c^2}{2n}\right) - n + 1} \\
&= 2^{1 - c^2/2}.
\end{aligned}$$

Now, let's consider Hoeffding bound for comparison. Assume we are applying Hoeffding bound to a σ^2-subgaussian random variable Z. We get:

$$\Pr\left(\left|Z - \frac{n}{2}\right| \geq c\sigma\right) \leq 2e^{-\frac{c^2}{2}}$$

Then, the bound that we showed is looser than the the Hoeffding bound.