

# Information Theory: Problem Set

General instructions:

- Please hand in your homework via Gradescope (entry code: KDPE8G) before 11:59 PM.
- Numbered exercises are taken from the book “Information Theory: From Coding to Learning” by Y. Polyanskiy and Y. Wu, available online at <https://people.lids.mit.edu/yp/homepage/data/itbook-export.pdf>.
- Unless otherwise specified, all logarithms (including those in entropy, mutual information, and KL divergence) are in base  $e$ .

## Homework 1 (Due on Oct 1, 2025)

Required problems:

R1. I.13

R2. III.19

- R3. (a) Show that  $I(X;Y) \geq I(X;Y|U)$  for a Markov chain  $U - X - Y$ . Conclude that  $I(X;Y)$  is concave in  $P_X$  for fixed  $P_{Y|X}$ .
- (b) Show that  $I(X;Y) \leq I(X;Y|U)$  if  $X$  and  $U$  are independent. Conclude that  $I(X;Y)$  is convex in  $P_{Y|X}$  for fixed  $P_X$ .

R4. Prove Tao’s inequality: for random variables  $X, Y, Z$  with  $X \in [-1, 1]$  almost surely,

$$\mathbb{E}|\mathbb{E}[X|Y] - \mathbb{E}[X|Y, Z]| \leq \sqrt{2I(X;Z|Y)}.$$

Optional problems (solve three of them):

O1. I.49 (Note: the claimed limit  $1/\sqrt{1-\tau}$  is incorrect and should be replaced by

$$\frac{e^{-\tau/2-\tau^2/4}}{\sqrt{1-\tau}} - 1. \quad )$$

O2. I.51

O3. I.53

O4. I.59 (Note: there is an obvious typo in Part (d):  $\varepsilon^{\lambda-1}$  should be  $\varepsilon^{1-\lambda}$ )

O5. I.63

O6. III.28

O7. *Shearer for sums.* Let  $X, Y, Z$  be independent random integers. Prove that

$$2H(X + Y + Z) \leq H(X + Y) + H(X + Z) + H(Y + Z).$$

O8. *Pinning lemma.* Let  $(X_1, \dots, X_n)$  be  $\{\pm 1\}^n$ -valued random vector. For  $2 \leq k \leq n$ , let  $S$  be a uniformly random subset of  $[n]$  of size  $k$ , and  $i, j \in S$  be two uniformly random draws from  $S$  without replacement. Define the quantity

$$f_k = \mathbb{E}[I(X_i; X_j | X_{S \setminus \{i,j\}})].$$

- (a) Prove that  $\sum_{k=2}^n f_k \leq \log 2$ .
- (b) Deduce that for  $m \geq 0$ , there exists a subset  $T \subseteq [n]$  with  $|T| \leq m$  such that

$$\mathbb{E}[\text{Cov}(X_i, X_j | X_T)^2] \leq \frac{2 \log 2}{m+1}.$$

Here the expectation is taken over the randomness in both the uniformly random pair  $(i, j) \in \binom{[n]}{2}$  and  $X_T$ .

O9. *Coin weighing.* There is an unknown subset  $X \subseteq [n]$ . You must choose in advance  $k$  subsets  $S_1, \dots, S_k \subseteq [n]$ , and receive the cardinalities  $|X \cap S_i|$  for all  $i \in [k]$ . We wish to determine the smallest number  $k$  needed to recover the unknown subset  $X$ .

- (a) Prove that if  $k \leq 1.99n / \log_2 n$  and  $n$  is sufficiently large, any strategy cannot guarantee the recovery of every subset  $X \subseteq [n]$ .
- (b) Propose a successful strategy if  $k \geq 3.17n / \log_2 n$  and  $n$  is sufficiently large (here the constant is chosen such that  $3.17 > 2 \log_2 3$ ).
- (c) (*challenging and not graded*) Prove (b) with 3.17 replaced by 2.01.

O10. *Information bound on variance.* Let  $X_1, \dots, X_n$  be i.i.d., and  $(\phi_t)_{t \in \mathcal{T}}$  be a collection of functions  $\phi_t : \mathcal{X} \rightarrow [0, 1]$ . For  $t \in \mathcal{T}$ , let  $\sigma^2(\phi_t) = \text{Var}(\phi_t(X_1))$  be the true variance, and

$$s_n^2(\phi_t) = \frac{1}{n} \sum_{i=1}^n \phi_t(X_i)^2 - \left( \frac{1}{n} \sum_{i=1}^n \phi_t(X_i) \right)^2$$

be the sample variance. Show that for any  $C > 0$  and random index  $T$ , it holds that

$$\mathbb{E}\left[\frac{s_n^2(\phi_T)}{\max\{C, \sigma^2(\phi_T)\}}\right] \leq \frac{I(T; X^n)}{nC} + 2.$$

(Hint: use  $\mathbb{E}[e^X] \leq \mathbb{E}[1 + 2X] \leq e^{2\mathbb{E}[X]}$  for  $X \in [0, 1]$ .)

## Homework 2 (Due on Nov 1, 2025)

Required problems:

R1. VI.8

R2. VI.14, Part (a) - (c)

R3. Suppose  $X_1, \dots, X_n \sim \text{Bern}(p)$  with unknown  $p \in [0, 1]$ . Using the two-point method, argue that if there is an estimator  $T$  such that

$$\sup_{p \in [0,1]} \mathbb{P}_p(|T(X) - p| > \varepsilon) \leq \delta$$

with  $\varepsilon, \delta \in (0, 1/4)$ , then

$$n \geq c \cdot \frac{\log(1/\delta)}{\varepsilon^2}$$

for a universal constant  $c > 0$ . (*Hint:  $1 - \text{TV} \geq \frac{1}{2} \exp(-\text{KL})$ .*)

R4. Let  $X_1, \dots, X_n$  be i.i.d. drawn from a discrete distribution  $P = (p_1, \dots, p_k)$ , the learner aims to estimate the entropy  $H(P) = \sum_{i=1}^k -p_i \log p_i$ . With a slight abuse of notation, we also use the same letter  $P$  to denote the free parameter  $(p_1, \dots, p_{k-1})$ , which belongs to the parameter set  $\mathcal{P}_k = \{(p_1, \dots, p_{k-1}) \in \mathbb{R}_+^{k-1} : \sum_{i=1}^{k-1} p_i \leq 1\}$  with a non-empty interior in  $\mathbb{R}^{k-1}$ .

- (a) For a fixed  $P$  in the interior of  $\mathcal{P}_k$ , find the expression of the Fisher information  $I(P)$  and the inverse Fisher information  $I(P)^{-1}$  in the above model with  $n = 1$ . (*Hint: for  $I(P)^{-1}$ , use Woodbury matrix identity.*)
- (b) Use the local asymptotic minimax theorem to show that for any  $P_0$  in the interior of  $\mathcal{P}_k$  and any sequence of estimators  $\hat{H}_n$  based on  $n$  samples, it holds that

$$\lim_{C \rightarrow \infty} \liminf_{n \rightarrow \infty} n \cdot \sup_{P \in \mathcal{P}_k : \|P - P_0\|_2 \leq C/\sqrt{n}} \mathbb{E}_P[(\hat{H}_n - H(P))^2] \geq \text{Var}_{X \sim P_0}(\log P_0(X)).$$

- (c) Find a suitable  $P_0$  in (b) to conclude that

$$\liminf_{n \rightarrow \infty} n \cdot \inf_{\hat{H}_n} \sup_{P \in \mathcal{P}_k} \mathbb{E}_P[(\hat{H}_n - H(P))^2] \geq c \cdot \log^2 k,$$

where  $c > 0$  is a universal constant.

Optional problems (solve three of them):

O1. I.65 (In Part (c),  $n$  should be  $d$ )

O2. I.66

O3. VI.14, Part (d)

O4. VI.16

O5. VI.17

O6. *Monotonic CLT.* Let  $X_1, X_2, \dots$  be i.i.d. random variables with  $\mathbb{E}[X_1] = 0, \text{Var}(X_1) = 1$ , and  $h(X_1) > -\infty$ . Let  $S_n = \sum_{i=1}^n X_i$  and  $T_n = \frac{1}{\sqrt{n}}S_n$ .

- (a) Let  $S_{\sim i} = S_n - X_i$ , and  $\rho_i$  be the score function of  $S_{\sim i}$ , where we recall that the score function for a random variable  $X$  with density  $f$  is  $\rho(x) = (\log f)' = \frac{f'(x)}{f(x)}$ . Show that the score function  $\rho$  of  $S_n$  is  $\rho(S_n) = \mathbb{E}[\rho_i(S_{\sim i})|S_n]$ .
- (b) Prove the following lemma: for independent  $Z_1, \dots, Z_n$  and functions  $f_1, \dots, f_n$  such that  $f_i$  depends only on  $Z_{\sim i}$  and  $\mathbb{E}[f_i(Z_{\sim i})] = 0$ , it holds that

$$\mathbb{E} \left[ \left( \sum_{i=1}^n f_i(Z_{\sim i}) \right)^2 \right] \leq (n-1) \sum_{i=1}^n \mathbb{E}[f_i(Z_{\sim i})^2].$$

(Hint: ANOVA decomposition.)

O7. *Bernoulli EPI: Mrs. Gerber's Lemma.* Let  $h_2(p) = -p \log p - (1-p) \log(1-p)$  be the binary entropy function, and  $h_2^{-1} : [0, \log 2] \rightarrow [0, \frac{1}{2}]$  be its inverse.

- (a) Show that for any fixed  $p \in [0, 1]$ , the function  $v \mapsto h_2(h_2^{-1}(v)*p)$  is convex, where  $p * q = p(1-q) + (1-p)q$  denotes the convolution.
- (b) Use (a) to show that for any  $(X, U)$  with  $X \in \{0, 1\}$  and  $Y = X \oplus \text{Bern}(p)$ ,

$$H(Y|U) \geq h_2(h_2^{-1}(H(X|U)) * p).$$

(c) Use (b) to show that for any  $X^n \in \{\pm 1\}^n$  and  $Y^n = X^n \oplus \text{Bern}(p)^{\otimes n}$ ,

$$\frac{H(Y^n)}{n} \geq h_2 \left( h_2^{-1} \left( \frac{H(X^n)}{n} \right) * p \right).$$

O8. *Tree-based lower bound.* This problem proves another lower bound for the test error of testing multiple hypotheses. Let  $T = ([m], E)$  be an undirected graph with vertex set  $[m]$  and edge set  $E$ , and be a tree in the sense that  $T$  is both connected and acyclic.

- (a) Show that for any real numbers  $x_1, \dots, x_m$ , it holds that

$$\sum_{i=1}^m x_i - \max_{i \in [m]} x_i \geq \sum_{(i,j) \in E} \min\{x_i, x_j\}.$$

(b) Use the result in Part (a), show that for probability distributions  $P_1, \dots, P_m$ ,

$$\min_{\Psi} \frac{1}{m} \sum_{i=1}^m P_i(\Psi \neq i) \geq \frac{1}{m} \sum_{(i,j) \in E} (1 - \text{TV}(P_i, P_j)),$$

where the minimum is over all possible tests  $\Psi : \mathcal{X} \rightarrow [m]$ .

- (c) Evaluate the terms on both sides of (b) under  $P_i = \mathcal{N}(i\Delta, 1)$  and a line tree with edge set  $E = \{(1, 2), (2, 3), \dots, (m-1, m)\}$ , and show that they are equal.

O9. *VC class with small oracle risk.* We have a function class  $\mathcal{F}$  with VC dimension  $d$ , and  $n$  training data  $(x_1, y_1), \dots, (x_n, y_n)$  drawn from an unknown joint distribution  $P_{XY}$ , with  $\mathcal{Y} = \{0, 1\}$ . Define the following class  $\mathcal{P}(\mathcal{F}, \varepsilon)$  of joint distributions where the best classifier has an error at most  $\varepsilon$ :

$$\mathcal{P}(\mathcal{F}, \varepsilon) = \left\{ P_{XY} : \inf_{f^* \in \mathcal{F}} P_{XY}(Y \neq f^*(X)) \leq \varepsilon \right\}.$$

So  $\varepsilon = 0$  corresponds to the well-specified case, and  $\varepsilon = 1$  corresponds to the misspecified case. Define the minimax excess risk  $R^*(\mathcal{F}, \varepsilon)$  over  $\mathcal{P}(\mathcal{F}, \varepsilon)$  as

$$R^*(\mathcal{F}, \varepsilon) = \inf_{\hat{f}} \sup_{P_{XY} \in \mathcal{P}(\mathcal{F}, \varepsilon)} \mathbb{E} \left[ P_{XY}(Y \neq \hat{f}(X)) - \inf_{f^* \in \mathcal{F}} P_{XY}(Y \neq f^*(X)) \right].$$

Show that for all  $\varepsilon \in [0, 1]$ ,

$$R^*(\mathcal{F}, \varepsilon) = \Omega \left( \min \left\{ \sqrt{\frac{d}{n} \cdot \varepsilon} + \frac{d}{n}, 1 \right\} \right).$$

O10. *Bias-variance analysis for orthogonal polynomials.* The concept of orthogonal polynomials is useful not only in proving lower bounds, but also in constructing and analyzing unbiased estimators. Let  $(P_\theta)_{\theta \in [\theta_0 - \varepsilon, \theta_0 + \varepsilon]}$  be a one-dimensional family of probability distributions with the following local expansion:

$$\frac{P_{\theta_0+u}(x)}{P_{\theta_0}(x)} = \sum_{m=0}^{\infty} p_m(x; \theta_0) \frac{u^m}{m!}, \quad \forall |u| \leq \varepsilon, x \in \mathcal{X}.$$

In addition, assume that the quantity  $\sum_{x \in \mathcal{X}} P_{\theta_0+u}(x) P_{\theta_0+v}(x) / P_{\theta_0}(x)$  depends only on  $\theta_0$  and  $uv$ , for all  $u, v \in [-\varepsilon, \varepsilon]$ . In class we showed that  $\{p_m(x; \theta_0)\}_{m \geq 0}$  are orthogonal in  $L^2(P_{\theta_0})$ , i.e.,

$$\mathbb{E}_{X \sim P_{\theta_0}} [p_m(X; \theta_0) p_n(X; \theta_0)] = A_m(\theta_0) \cdot \mathbb{1}(m = n)$$

for some constants  $\{A_m(\theta_0)\}_{m \geq 0}$ .

- (a) Show that for  $X \sim P_{\theta_0+u}$  with  $u \in [-\varepsilon, \varepsilon]$ ,

$$\mathbb{E}_{X \sim P_{\theta_0+u}} [p_m(X; \theta_0)] = c_m u^m$$

for some constant  $c_m$  independent of  $u$ . Find the expression of  $c_m$  using  $A_m(\theta_0)$ .

- (b) Suppose the same local expansion and orthogonality condition also hold under  $\theta_0 + u$ , and

$$p_m(x; \theta_0) = \sum_{\ell=0}^m b(m, \ell, \theta_0, u) \cdot p_\ell(x; \theta_0 + u), \quad \forall |u| \leq \varepsilon, x \in \mathcal{X},$$

Show that

$$\mathbb{E}_{X \sim P_{\theta_0+u}} [p_m(X; \theta_0)^2] = \sum_{\ell=0}^m b(m, \ell, \theta_0, u)^2 \cdot A_\ell(\theta_0 + u).$$

- (c) Show that in the Poisson model  $\mathcal{X} = \mathbb{N}$ ,  $P_\theta = \text{Poi}(\theta)$ ,

$$b(m, \ell, \theta_0, u) = \binom{m}{\ell} \frac{(\theta_0 + u)^\ell u^{m-\ell}}{\theta_0^m}.$$

### Homework 3 (Due on Dec 1, 2025)

Required problems:

R1. II.20

R2. VI.25

R3. Let  $A \in \mathbb{R}^{m \times n}$  be a random matrix with i.i.d.  $\mathcal{N}(0, 1)$  entries. The operator norm of  $A$  is defined as  $\|A\|_{\text{op}} = \max_{v \in S^{n-1}} \|Av\|_2$ , where  $S^{n-1}$  denotes the unit sphere in  $\mathbb{R}^n$ .

- (a) Let  $\mathcal{U} = \{u_1, \dots, u_M\}$  and  $\mathcal{V} = \{v_1, \dots, v_N\}$  be an  $\varepsilon$ -net (under the  $\ell_2$  norm) for  $S^{m-1}$  and  $S^{n-1}$  respectively. Show that for  $\varepsilon < 1/2$ ,

$$\|A\|_{\text{op}} \leq \frac{1}{1 - 2\varepsilon} \max_{u \in \mathcal{U}, v \in \mathcal{V}} u^\top Av.$$

- (b) Deduce from (a) that  $\mathbb{E}\|A\|_{\text{op}} \lesssim \sqrt{m} + \sqrt{n}$ .

- (c) Show a matching lower bound  $\mathbb{E}\|A\|_{\text{op}} \gtrsim \sqrt{m} + \sqrt{n}$  using Sudakov minoration.

R4. Recall that  $M(A, d, \varepsilon)$  denotes the maximum number  $m$  of points  $x_1, \dots, x_m$  such that  $d(x_i, x_j) \geq \varepsilon$  for every  $i \neq j \in [d]$ .

- (a) Let  $A$  be the set of all non-decreasing functions  $f : [0, 1] \rightarrow [0, 1]$ . Show that for  $\varepsilon \in [0, 1]$ , there exist universal constants  $c_1, c_2 > 0$  such that

$$\log M(A, L_2([0, 1]), c_1 \varepsilon) \geq \frac{c_2}{\varepsilon}.$$

- (b) Now let  $A$  be the set of all convex functions  $f : [0, 1] \rightarrow [0, 1]$ . Show that for  $\varepsilon \in [0, 1]$ , there exist universal constants  $c_1, c_2 > 0$  such that

$$\log M(A, L_2([0, 1]), c_1 \varepsilon) \geq \frac{c_2}{\sqrt{\varepsilon}}.$$

*Hint: you may try to break into several small intervals, find two possible function constructions in each interval, and concatenate them. Use the Gilbert–Varshamov bound in class.*

Optional problems (solve three of them):

O1. VI.19

O2. VI.24

O3. VI.26 (*A typo in the problem: the inequality should be*

$$\frac{1}{n - |T| - 1} \sum_{i \neq j \in T^c} D(\nu_{X_i, X_j} \| \mu_{X_i, X_j}^{(\sigma_T)}) \geq \left(2 - \frac{c}{n - |T| - 1}\right) \sum_{i \in T^c} D(\nu_{X_i} \| \mu_{X_i}^{(\sigma_T)}).$$

*Additional hint: for any function  $h : 2^{[n]} \rightarrow \mathbb{R}$  and  $S \sim \text{Bern}(\tau)^{\otimes n}$ , show that*

$$\begin{aligned} \frac{d}{d\tau} \mathbb{E}[h(S)] &= \sum_{i=1}^n \mathbb{E}[h(S \cup \{i\}) - h(S)], \\ \frac{d^2}{d\tau^2} \mathbb{E}[h(S)] &= \sum_{i \neq j} \mathbb{E}[h(S \cup \{i, j\}) - h(S \cup \{i\}) - h(S \cup \{j\}) + h(S)]. \end{aligned}$$

O4. *Redundancy bound with general Beta mixing.* Let  $P_\theta = \text{Bern}(\theta)$ , and  $\theta \sim \text{Beta}(\alpha, \beta)$  follow a Beta distribution. For  $x^n \in \{0, 1\}^n$ , let  $Q(x^n) = \mathbb{E}_\theta[P_\theta^{\otimes n}(x^n)]$ , and  $n_0, n_1$  be the number of 0's and 1's in  $x^n$  (with  $n_0 + n_1 = n$ ).

(a) Let  $B(\alpha, \beta)$  be the Beta function. Show that

$$\max_{\theta \in [0, 1]} \frac{P_\theta^{\otimes n}(x^n)}{Q(x^n)} = \frac{\frac{n_0^{n_0} n_1^{n_1}}{n^n}}{\frac{B(\alpha+n_0, \beta+n_1)}{B(\alpha, \beta)}}.$$

(b) By Stirling's approximation  $1 \leq \Gamma(z)/(\sqrt{2\pi} z^{z-1/2} e^{-z}) \leq e^{\frac{1}{12z}}$  for  $z > 0$ , show that

$$\max_{\theta \in [0, 1]} D_{\text{KL}}(P_\theta^{\otimes n} \| Q) \leq \max_{\theta \in [0, 1]} \max_{x^n} \log \frac{P_\theta^{\otimes n}(x^n)}{Q(x^n)} \leq \max \left\{ \frac{1}{2}, \alpha, \beta \right\} \log n + O_{\alpha, \beta}(1).$$

In other words, there is a range of parameters for the Beta prior that can achieve the optimal regret up to first order.

O5. *Last-iterate convergence for Hellinger.* For probability distributions  $P_1, \dots, P_m$  on the same space  $\mathcal{X}$ , let  $Q_{X^n} = \sum_{i=1}^m \pi_i P_i^{\otimes n}$  be the Yang–Barron type mixture with prior  $\pi$ . Show that for  $X_1, \dots, X_{n-1} \sim P_1$  and a universal constant  $C$ ,

$$\mathbb{E}[H^2(P_1, Q_{X_n|X^{n-1}})] \leq \frac{C}{n} \log \frac{m}{\pi_1}.$$

(*Hint: Express  $Q_{X_n|X^{n-1}} = \sum_i w_i P_i$  and upper bound  $\mathbb{E}[w_i]$  based on  $H^2(P_1, P_i)$ .*)

O6. *Jeffreys prior.*

- (a) I.57, Part (a)
- (b) Let  $\pi$  be a prior on  $\Theta \subseteq \mathbb{R}^d$  with density  $\pi(\theta)$ , and  $\theta_0 \in \text{int}(\Theta)$ . Let  $(P_\theta)_{\theta \in \Theta}$  be a family of distributions, with Fisher information matrix  $I(\theta)$  at  $\theta = \theta_0$ . Show that for  $Q = \mathbb{E}_{\theta \sim \pi}[P_\theta^{\otimes n}]$ , by Laplace's method and the local behavior of KL divergence by Fisher information, one has the approximate upper bound

$$D_{\text{KL}}(P_{\theta_0}^{\otimes n} \| Q) \leq \frac{d}{2} \log \frac{n}{2\pi} - \log \frac{1}{\pi(\theta_0)} + \frac{1}{2} \log \det I(\theta_0) + O(1).$$

Therefore, choosing  $\pi(\theta) \propto (\log \det I(\theta))^{-1/2}$  (known as *the Jeffreys prior*) achieves a redundancy upper bound  $\frac{d}{2} \log n + O(1)$  assuming regularity conditions.

- O7. *Redundancy of uniform family.* Let  $\mathcal{P} = \{\text{Unif}(0, \theta) : \theta \in [\frac{1}{2}, 1]\}$ . Show that  $\text{Red}(\mathcal{P}^{\otimes n}) \sim \log n$  by proving redundancy upper and lower bounds. Explain the difference from the usual relation  $\text{Red}(\mathcal{P}^{\otimes n}) \sim \frac{1}{2} \log n$ .
- O8. *Branching number.* For a countable rooted tree  $T$  where each vertex has a finite degree, a *flow* is a function  $f : V(T) \rightarrow \mathbb{R}_+$  such that  $f_u = \sum_{v \in \text{children}(u)} f_v$  for all vertices  $u$ . The *branching number*  $\text{br}(T)$  is defined as the supremum of  $\lambda \in \mathbb{R}$  such that there exists a flow  $f$  with  $f_u > 0$  for some  $u$ , and  $f_u \leq \lambda^{-d(u)}$  for all vertices  $u$ , where  $d(u)$  denotes the depth of  $u$ .

- (a) Show that for broadcasting on tree  $T$ , if each edge represents a channel  $P_{Y|X}$  with  $\eta_{\text{KL}}(P_{Y|X}) \leq \eta$ , then the model has non-reconstruction when  $\text{br}(T)\eta < 1$ .
- (b) For  $p \in [0, 1]$ , let  $T_p$  be the connected component containing the root in a random graph where each edge of  $T$  is removed independently with probability  $1 - p$ . Let  $p_c = p_c(T) \in [0, 1]$  be the critical (percolation) probability:

$$p_c = \sup\{p \in [0, 1] : \mathbb{P}(T_p \text{ has infinitely many vertices}) = 0\}.$$

Show that  $\text{br}(T) \leq p_c^{-1}$ . (*Hint: for  $p > \text{br}(T)^{-1}$ , construct a flow  $\{f_u\}$  and define  $M_n = \sum_{u \in V(T_p) : \text{depth}(u)=n} f_u p^{-n}$ . Show that  $\{M_n\}$  is a martingale, and that  $\sup_n \mathbb{E}[M_n^2] < \infty$ . Therefore  $M_n \rightarrow M_\infty$  in  $L^1$ .*)

- (c) Show that  $\text{br}(T) \geq p_c^{-1}$ . (*Hint: show that if  $p < \text{br}(T)^{-1}$ , find a sequence of cuts  $\{C_n\}$  of  $T$  such that*

$$\lim_{n \rightarrow \infty} \mathbb{E}[\text{number of edges in } T_p \text{ crossed by the cut } C_n] = 0;$$

*the max-flow min-cut theorem might be useful.)*

- (d) Conclude that if the offspring distribution of a branching process has mean  $m > 1$ , then given the event that this process does not become extinct, the corresponding Galton–Watson tree  $T$  has branching number  $m$  almost surely. (*Hint: recall that the extinction probability in a branching process with  $m > 1$  is the unique solution in  $[0, 1]$  to the equation  $x = \sum_{i=0}^{\infty} p_i x^i$ . How about the probability that  $T$  has branching number at least  $\lambda$ , for different choices of  $\lambda$ ?*)

O9.  $F_I$  curve. For a joint distribution  $P_{XY}$  and  $t > 0$ , define

$$F_I(P_{XY}, t) = \sup\{I(U; Y) : I(U; X) \leq t, U - X - Y \text{ forms a Markov chain}\}.$$

- (a) Show that  $t \mapsto F_I(P_{XY}, t)$  is concave, and  $\frac{d}{dt}F_I(P_{XY}, t)|_{t=0} = \eta_{\text{KL}}(P_X, P_{Y|X})$ .
- (b) Using EPI and Bernoulli EPI (Mrs. Gerber's lemma in HW2 O7), find the expression of  $F_I(P_{XY}, t)$  in the following two scenarios:
  - i.  $(X, Y)$  is zero-mean and jointly Gaussian with correlation  $\rho \in [-1, 1]$ ;
  - ii.  $(X, Y)$  is zero-mean and jointly Bernoulli with correlation  $\rho \in [-1, 1]$ .
- (c) Conclude that in both scenarios, the maximal correlation between  $X$  and  $Y$  is  $|\rho|$ .

O10. SDPI for Fisher information. Let  $(P_\theta)_{\theta \in \Theta \subseteq \mathbb{R}^n}$  be a family of distributions, with score function  $s_\theta(\cdot)$  and Fisher information matrix  $I(\theta)$ .

- (a) Show that for  $\theta - X - Y$ , the score function for  $Y$  is  $s_\theta^Y(y) = \mathbb{E}[s_\theta(X)|Y = y]$ .
- (b) If  $P_\theta$  is a discrete pmf  $(\frac{1}{2n} + \theta_1, \frac{1}{2n} - \theta_1, \dots, \frac{1}{2n} + \theta_n, \frac{1}{2n} - \theta_n)$ , show that

$$\sup\{\text{trace}(I^Y(0)) : \theta - X - Y, |\mathcal{Y}| \leq \ell\} \asymp \min\{n(\ell - 1), n^2\},$$

where  $I^Y(\theta)$  denotes the Fisher information matrix for  $Y$ , and  $\ell \in \mathbb{N}$ .

- (c) If  $P_\theta = \mathcal{N}(\theta, I_n)$ , show that

$$\sup\{\text{trace}(I^Y(0)) : \theta - X - Y, |\mathcal{Y}| \leq \ell\} \asymp \min\{\log \ell, n\},$$