散度，集中不等式

参考答案

1. (3 分) 对于随机变量 X, Y，我们定义了信息熵、条件 (信息) 熵、互信息

$$H[X] = \sum_x \Pr[X = x] \log \frac{1}{\Pr[X = x]} = \mathbb{E}\Big[\log \frac{1}{P_X(X)}\Big],$$
$$H[X|Y] = \sum_{x,y} \Pr[X = x, Y = y] \log \frac{1}{\Pr[X = x|Y = y]} = \mathbb{E}\Big[\log \frac{1}{P_{X|Y}(X|Y)}\Big],$$
$$I[X;Y] = H[X] - H[X|Y].$$

考虑事件 E 或另一随机变量 Z，还定义了条件互信息

$$I[X;Y|E] = H[X|E] + H[Y|E] - H[X,Y|E]$$
$$I[X;Y|Z] = \sum_z \Pr[Z = z] I[X;Y|Z = z].$$

如果我们定义三个随机变量之间的互信息为

$$I[X;Y;Z] = H[X] + H[Y] + H[Z] - H[X,Y] - H[X,Z] - H[Y,Z] + H[X,Y,Z].$$

(1) 证明 $I[X;Y;Z] = I[X;Y] - I[X;Y|Z]$.

(2) 举例说明 $I[X;Y;Z]$ 可以大于零，也可以小于零. 也就是说，泄露额外信息 Z 后，X, Y 之间的互信息可能增加，也可能减少.

解

(1) 展开整理即得.

(2) 令 X, Y 是独立同分布的均匀伯努里 $\mathrm{Bern}(1/2)$. 令 $Z = X \oplus Y$. 容易验证，$I[X;Y] = 0$，$I[X;Y|Z] = \log 2$，因此 $I[X;Y;Z] = -\log 2 < 0$.

另一方面，$I[X;X;X] = I[X;X] - I[X;X|X] = \log 2 > 0$.

2. (8 分) 完成以下关于 Chernoff bound 的证明.

(1) (0 分) 两个 Bernoulli 分布间的 KL 散度可以简记为 $d(p\|q) := D(\mathrm{Bern}(p)\|\,\mathrm{Bern}(q))$. 证明 $d(p\|q)/\log e \geq 2(p-q)^2$.

Remark: 除以 $\log e$ 等价于使用 e 作底数.

(2) (0 分) 设随机变量 $(X_1, \ldots, X_n) \sim (\mathrm{Bern}(p))^n$，即它们独立地服从 $\mathrm{Bern}(p)$. 对任意 $t > 0$，

$$\Pr\Big[\frac{X_1 + \cdots + X_n}{n} \geq q\Big] = \Pr[e^{t(X_1+\cdots+X_n)} \geq e^{tqn}] \underset{\text{Markov's bound}}{\leq} \frac{\mathbb{E}[e^{t(X_1+\cdots+X_n)}]}{e^{tqn}} = \Big(\frac{\mathbb{E}[e^{tX_1}]}{e^{tq}}\Big)^n.$$

当 $0 \le p \le q \le 1$ 时，请选取合适的 t 使得上式最紧. 得到的结果应为

$$\Pr\Big[\frac{X_1 + \cdots + X_n}{n} \ge q\Big] \le \exp(-n \cdot d(q\|p)).$$

Remark: 对称地，当 $0 \le q \le p \le 1$ 时，可以证明

$$\Pr\Big[\frac{X_1 + \cdots + X_n}{n} \le q\Big] \le \exp(-n \cdot d(q\|p)).$$

(3) 设 $(X_1, \ldots, X_n) \sim P_1 P_2 \ldots P_n$，即它们相互独立. 每个 P_i 都是 $[0,1]$ 上的期望等于 p 的分布. 证明当 $0 \le p \le q \le 1$ 时，

$$\Pr\Big[\frac{X_1 + \cdots + X_n}{n} \ge q\Big] \le \exp(-n \cdot d(q\|p)).$$

提示：比较 $\mathbb{E}_{X\sim P_i}[e^{tX}]$ 和 $\mathbb{E}_{X\sim \mathrm{Bern}(p)}[e^{tX}]$ 的大小.

(4) 有 $m > n$ 个球，其中 pm 个是白球. 从中无放回的随机选取 n 个球. 用随机变量 $(X_1, \ldots, X_n)$ 表示这 n 次选取的结果. $X_i = 1$ 表示第 i 个球是白球，$X_i = 0$ 表示第 i 个球不是白球. 显然 $\mathbb{E}[X_i] = p$. 证明当 $0 \le p \le q \le 1$ 时，

$$\Pr\Big[\frac{X_1 + \cdots + X_n}{n} \ge q\Big] \le \exp(-n \cdot d(q\|p)).$$

解

(1) 定义 $f_q(p) = d(p\|q)/\log e$.

$$\begin{aligned} f_q(p) &= p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q} \\ f_q'(p) &= \ln \frac{p}{q} - \ln \frac{1-p}{1-q} \\ f_q''(p) &= \frac{1}{p} + \frac{1}{1-p} \end{aligned}$$

可以看出，$f_q(q) = 0$，$f_q'(q) = 0$ 而且 $f_q''(p) \ge 4$. 这足以说明 $f_q(p) \ge 2(p-q)^2$.

(2) 定义 $f(t) = \ln \frac{\mathbb{E}[e^{tX_1}]}{e^{tq}} = \ln(pe^t + 1 - p) - tq$. 对 f 求导

$$f'(t) = \frac{pe^t}{pe^t + 1 - p} - q.$$

f' 单调递增，且存在唯一 t^* 使得 $f'(t^*) = 0$. 这说明 f 的最小值点为 $t^* = \ln(\frac{q}{1-q}\frac{1-p}{p}) > 0$.

$$\min_{t>0} f(t) = f(t^*) = \ln\Big(p\frac{q}{1-q}\frac{1-p}{p} + 1 - p\Big) - q\ln\Big(\frac{q}{1-q}\frac{1-p}{p}\Big) = -d(p\|q).$$

(3) 只需说明 $\mathbb{E}_{X\sim P_i}[e^{tX}] \le \mathbb{E}_{X\sim \mathrm{Bern}(p)}[e^{tX}]$. 证明的其余部分和 Bernoulli 分布的情况相同.

不妨定义一个从 $[0,1]$ 到 $\{0,1\}$ 的 kernel $P_{Y|X}$ 使得 $P_{Y|X=x} = \mathrm{Bern}(x)$. 换言之，

$$P_{Y|X}(y|x) = \begin{cases} x, & \text{if } y = 1 \\ 1 - x, & \text{if } y = 0 \end{cases}$$

对任意 $[0,1]$ 上期望等于 p 的分布 P_X，考虑 $(X,Y)\sim P_XP_{Y|X}$. 因为 $\mathbb{E}[Y]=\mathbb{E}[\mathbb{E}[Y|X]]=\mathbb{E}[X]=p$，所以 $Y\sim\mathrm{Bern}(p)$. 对任何 $x\in\mathrm{Supp}(P_i)$，因为指数函数的凸性，

$$\mathbb{E}[e^{tY}|X=x]\geq e^{t\,\mathbb{E}[Y|X=x]}=e^{tx}.$$

进而

$$\mathbb{E}[e^{tY}]=\mathbb{E}[\mathbb{E}[e^{tY}|X]]\geq\mathbb{E}[e^{tX}].$$

(4) 只需证明对任何 $t>0$，

$$\mathbb{E}[e^{t(X_1+\cdots+X_n)}]\leq\mathop{\mathbb{E}}_{X\sim\mathrm{Bern}(p)}[e^{tX}]^n.$$

证明的其余部分和 Bernoulli 分布的情况相同.

我们递归地证明这个命题. 当 $n=1$ 时，命题显然成立. 下面假设命题对 $n-1$ 成立，我们证明命题对 n 也成立. 考虑 X_n 的条件分布，

$$\Pr[X_n=1|X_1+\cdots+X_{n-1}=s]=\frac{pm-s}{m-n+1}.$$

不难看出 s 越大，X_n 的条件期望就越小. 类似地，$e^{t(X_1+\cdots+X_{n-1})}$ 越大，$e^{t(X_n)}$ 的条件期望就越小. 因此

$$\begin{aligned}\mathbb{E}[e^{t(X_1+\cdots+X_n)}]&=\mathbb{E}[e^{t(X_1+\cdots+X_{n-1})}\,\mathbb{E}[e^{t(X_n)}|e^{t(X_1+\cdots+X_{n-1})}]]\\&\leq\mathbb{E}[e^{t(X_1+\cdots+X_{n-1})}]\,\mathbb{E}[e^{t(X_n)}]\leq\mathop{\mathbb{E}}_{X\sim\mathrm{Bern}(p)}[e^{tX}]^n.\end{aligned}$$

第一个不等号可以抽象化为一个类似排序不等式的引理：对任意非负实数上的分布 P 和任意单调递减的函数 f，$\mathbb{E}_{X\sim P}[Xf(X)]\leq\mathbb{E}_{X\sim P}[X]\,\mathbb{E}_{X\sim P}[f(X)]$. 其中 X 对应 $e^{t(X_1+\cdots+X_n)}$，$f(v):=\mathbb{E}[e^{t(X_n)}|e^{t(X_1+\cdots+X_{n-1})}=v]$. 引理的证明如下：

$$\begin{aligned}\mathop{\mathbb{E}}_{X\sim P}[X]\mathop{\mathbb{E}}_{X\sim P}[f(X)]&=\sum_x P(x)x\sum_y P(y)f(y)\\&=\sum_x P^2(x)xf(x)+\sum_{x<y}P(x)P(y)xf(y)+\sum_{x>y}P(x)P(y)xf(y)\\&=\sum_x P^2(x)xf(x)+\sum_{x<y}\Big(P(x)P(y)xf(y)+P(x)P(y)yf(x)\Big)\\&\leq\sum_x P^2(x)xf(x)+\sum_{x<y}\Big(P(x)P(y)xf(x)+P(x)P(y)yf(y)\Big)\\&=\sum_x P^2(x)xf(x)+\sum_{x\neq y}P(x)P(y)xf(x)\\&=\sum_x\sum_y P(x)P(y)xf(x)\\&=\sum_x P(x)xf(x)\\&=\mathop{\mathbb{E}}_{X\sim P}[Xf(X)]\end{aligned}$$

3. (10 分) 根据 Sanov's Theorem 我们可以看出，Chernoff bound 对于

$$\Pr_{(X_1,\dots,X_n)\sim(\mathrm{Bern}(p))^n}\Big[\frac{X_1+\cdots+X_n}{n}\geq q\Big]$$

的估计已经很精确，指数上的系数是紧的. 这个估计对非 Bernoulli 分布是否也同样精确?

考虑有限个正实数上的分布 P. 记 $\mathrm{Supp}(P)=\{v_1,\dots,v_T\}\subseteq\mathbb{R}^+$. 记 $p_i:=P(v_i)>0$. 这个分布的期望是 $\bar{v}=\sum p_iv_i$. 考虑任意 $b\in(\bar{v},\max_i v_i)$，定义

$$Q^*=\arg\min_{\substack{\text{分布 } Q\\ \mathbb{E}_{X\sim Q}[X]\geq b}} D(Q\|P).$$

根据 Sanov's Theorem，

$$\Pr_{(X_1,\dots,X_n)\sim P^n}\Big[\frac{X_1+\cdots+X_n}{n}\geq b\Big]\leq(n+1)^T\cdot\exp(-n\cdot D(Q^*\|P)).$$

而根据 Chernoff bound，

$$\Pr_{(X_1,\dots,X_n)\sim P^n}\Big[\frac{X_1+\cdots+X_n}{n}\geq b\Big]\leq\min_{t>0}\Big(\frac{\mathbb{E}_{X\sim P}[e^{tX}]}{e^{tb}}\Big)^n.$$

请问是否存在 P,b 使得 Chernoff bound 的估计要弱于 Sanov's Theorem?

提示：拉格朗日乘数.

解 不存在.

首先考虑 Sanov's Theorem 一边. 用 $q_1,\dots,q_T$ 表示 Q 分布下的概率. Q^* 是如下优化问题的解

$$\begin{aligned}
&\text{最小化}\quad f(q_1,\dots,q_T):=\sum_i q_i\ln\Big(\frac{q_i}{p_i}\Big)=D(Q\|P)/\log e\\
&\text{约束：}\quad (1)\ \sum_i q_iv_i\geq b\\
&\qquad\quad (2)\ \sum_i q_i=1\\
&\qquad\quad (3)\ \forall i\ \ q_i\geq 0
\end{aligned}$$

约束条件是有界闭集而 f 连续，所以最小值一定存在. 对最小值点 Q^* 来说，条件 (2) 显然是紧的. 条件 (1) 也是紧的，因为从任何 P 到 Q 的连线上，$f(\varepsilon Q+(1-\varepsilon)P)$ 都单调地增长 ($\varepsilon\in[0,1]$). 而条件 (3) 是松的，不妨考虑任何一个满足 (1)(2) 的分布 Q 且满足 $q_i=0$，我们来说明 Q 不是极小值点. 计算偏导

$$\frac{\partial f}{\partial q_i}=\ln\Big(\frac{q_i}{p_i}\Big)+1,$$

在 $q_i=0$ 的位置，$\frac{\partial f}{\partial q_i}(Q)=-\infty$，这提示我们微调 q_i 的值会使函数值更小. 严格来说，可以分两种情况讨论

- 如果存在 j,k 使得 $q_j,q_k>0$：定义 Q_ε 为

$$Q_\varepsilon(v_x)=\begin{cases}\varepsilon, & \text{if } x=i\\ q_j+C_j\varepsilon, & \text{if } x=j\\ q_k+C_k\varepsilon, & \text{if } x=k\\ Q(v_x)=q_x, & \text{otherwise}\end{cases}\qquad\text{其中 } C_j,C_k \text{ 是 } \begin{matrix}1+C_j+C_k=0\\ v_i+v_jC_j+v_kC_k=0\end{matrix} \text{ 的解}$$

这样对足够小的 $\varepsilon \geq 0$，Q_ε 满足 (1)(2)(3). 同时 $Q_0 = Q$, $\frac{d}{d\varepsilon} f(Q_\varepsilon)|_{\varepsilon=0} = -\infty$. 因此 Q 不是极小值点.

- 如果存在 k 使得 $q_k = 1$：根据现有的条件，这说明 $v_k = b \in (\min_x v_x, \max_x v_x)$. 因此一定存在 j 使得 $v_i < v_k < v_j$ 或 $v_i > v_k > v_j$. 定义 Q_ε 为

$$Q_\varepsilon(v_x) = \begin{cases} C_i\varepsilon, & \text{if } x = i \\ C_j\varepsilon, & \text{if } x = j \\ 1-\varepsilon, & \text{if } x = k \\ Q(v_x) = 0, & \text{otherwise} \end{cases} \qquad \text{其中 } C_i, C_j \text{ 是} \begin{matrix} C_i + C_j - 1 = 0 \\ v_iC_i + v_jC_j - v_k = 0 \end{matrix} \text{的解}$$

  同样的论证可以说明 Q 不是极小值点.

使用拉格朗日乘数法，存在 A, B 使得

$$\frac{\partial}{\partial q_i}\Big(f(Q) + A\sum_j q_j v_j + B\sum_j q_j\Big) = \ln\Big(\frac{q_i}{p_i}\Big) + 1 + Av_i + B$$

在极小值点等于 0. 不妨令 $t = -A$，那么在极小值点处

$$q_i = p_i e^{-1-Av_i-B} \propto p_i e^{tv_i}.$$

因为 Q 是概率，所以 B 的取值一定会令

$$q_i = \frac{p_i e^{tv_i}}{\sum_j p_j e^{tv_j}}.$$

定义 $C_t = \sum_j p_j e^{tv_j}$，定义 Q_t 为 $Q_t(v_i) = p_i e^{tv_i}/C_t$. 已经证明存在 t 使得 $Q^* = Q_t$. 注意到 Q_t 的期望随着 t 严格单调增加，因此存在唯一的 $t^* > 0$ 满足

$$b = \mathop{\mathbb{E}}_{X\sim Q_{t^*}}[X] = \frac{\sum_i v_i p_i e^{t^* v_i}}{\sum_i p_i e^{t^* v_i}}$$

同一个 t 使得 $Q^* = Q_{t^*}$ 成立.

再考虑 Chernoff bound 一边.

$$\Pr_{(X_1,\ldots,X_n)\sim P^n}\Big[\frac{X_1 + \cdots + X_n}{n} \geq b\Big] \leq \min_{t>0}\Big(\frac{\mathbb{E}_{X\sim P}[e^{tX}]}{e^{tb}}\Big)^n \leq \Big(\frac{\mathbb{E}_{X\sim P}[e^{t^*X}]}{e^{t^*b}}\Big)^n.$$

只需再证明

$$\ln\Big(\frac{\mathbb{E}_{X\sim P}[e^{t^*X}]}{e^{t^*b}}\Big) = -D(Q^*\|P)/\log e.$$

简单验证即可

$$\text{左边} = \ln(\mathop{\mathbb{E}}_{X\sim P}[e^{t^*X}]) - t^*b = \ln\Big(\sum_i p_i e^{t^*v_i}\Big) - t^*b = \ln C_{t^*} - t^*b$$

$$\text{右边} = -\sum_i Q^*(v_i)\ln\frac{Q^*(v_i)}{p_i} = -\sum_i Q^*(v_i)\ln\frac{e^{t^*v_i}}{C_{t^*}} = \ln C_{t^*} - t^*\sum_i Q^*(v_i)v_i = \ln C_{t^*} - t^*b$$

4. (3 分) 证明散度的 data-processing 不等式. 对任意 P_X, Q_X 和 kernel $P_{Y|X}$, 令 $P_Y = P_X \circ P_{Y|X}$, $Q_Y = Q_X \circ P_{Y|X}$ (也就是说, P_Y, Q_Y 分别是 $P_{XY} = P_X P_{Y|X}, Q_{XY} = Q_X P_{Y|X}$ 的边缘分布). 证明

$$D(P_X \| Q_X) \geq D(P_Y \| Q_Y).$$

Remark: 互信息的 data-processing 不等式可以由散度的 data-processing 不等式推出. 如果 X, Y, Z 的依赖关系可以用有向图 $X \to Y \to Z$ 表示 (即 $P_{XYZ} = P_X P_{Y|X} P_{Z|Y}$), 注意到

$$I(X;Y) = D(P_{XY} \| P_X P_Y), \qquad I(X;Z) = D(P_{XZ} \| P_X P_Z).$$

只需定义一个合适的 kernel 便证明了互信息的 data-processing 不等式 $I(X;Y) \geq I(X;Z)$.

解 对 $D(P_{XY} \| Q_{XY})$ 使用散度的 chain rule 两次.

$$D(P_{XY} \| Q_{XY}) = D(P_X \| Q_X) + \underbrace{D(P_{Y|X} \| Q_{Y|X} | P_X)}_{=0},$$
$$D(P_{XY} \| Q_{XY}) = D(P_Y \| Q_Y) + \underbrace{D(P_{X|Y} \| Q_{X|Y} | P_Y)}_{\geq 0}.$$

5. (3 分) 使用 data-processing 不等式, 证明

$$\sqrt{\frac{1}{2 \log e} D(P \| Q)} \geq \Delta(P, Q)$$

这里 $\Delta(P, Q)$ 表示 P, Q 之间的统计距离 (statistical distance, 也可以更精确地称为 total variation distance)

$$\Delta(P, Q) = \frac{1}{2} \sum_x |P(x) - Q(x)| = \max_{\text{事件 } E} (P(E) - Q(E)).$$

解 用 E^* 表示统计距离的定义中最能区分分布 P_X, Q_X 的事件

$$E^* = \arg \max_{\text{事件 } E} (P_X(E) - Q_X(E)).$$

定义条件概率 $P_{Z|X}$

$$P_{Z|X}(1|x) = 1 \text{ if } x \in E \qquad P_{Z|X}(0|x) = 1 \text{ if } x \notin E$$

因此 $\mathrm{Bern}(P_X(E)), \mathrm{Bern}(Q_X(E))$ 分别是 $P_X P_{Y|X}$ 和 $Q_X P_{Y|X}$ 的边缘分布. 根据 data-processing 不等式,

$$D(P_X \| Q_X) \geq D(\mathrm{Bern}(P_X(E)) \| \mathrm{Bern}(Q_X(E))) \geq 2 \log e \cdot (P_X(E) - Q_X(E))^2 = 2 \log e \cdot \Delta(P_X, Q_X)^2.$$

6. (4 分) 证明对于服从任意联合分布的随机变量 X, Y, Z,

$$2H[X, Y, Z] \leq H[X, Y] + H[X, Z] + H[Y, Z].$$

据此证明 Shearer 引理：令 Ω 是 $\mathbb{R}^3$ 上 n 个点组成的集合，Ω 向三个坐标平面投影分别有 n_1, n_2, n_3 个像，那么 $n^2 \le n_1 n_2 n_3$. 并说明何时可以取到等号.

解 先证明这个熵不等式.

$$\begin{aligned}
&H[X,Y] + H[X,Z] + H[Y,Z] - 2H[X,Y,Z] \\
&= H[X,Y] - H[Y|X,Z] - H[X|Y,Z] \\
&\ge H[X,Y] - H[Y|X] - H[X|Y,Z] \\
&= H[X] - H[X|Y,Z] \ge 0.
\end{aligned}$$

不等式取等号的必要条件是 $H[X] = H[X|Y,Z]$，也就是 X 与 (Y,Z) 独立. 由对称性，不等式取等号的必要条件 X, Y, Z 相互独立. 不难验证这也是充分条件.

令随机变量 (X,Y,Z) 是从 Ω 中随机选取的一个点的坐标，那么 $H[X,Y,Z] = \log n$. 根据题目条件，$(X,Y),(X,Z),(Y,Z)$ 的支撑集大小分别为 n_1, n_2, n_3，所以 $H[X,Y], H[X,Z], H[Y,Z]$ 分别不超过 $\log n_1, \log n_2, \log n_3$. 根据刚刚证明的熵不等式，

$$2\log n = H[X,Y,Z] \le H[X,Y] + H[X,Z] + H[Y,Z] \le \log n_1 + \log n_2 + \log n_3.$$

也就是题目要求的 $n^2 \le n_1 n_2 n_3$.

不等式取等号，首先需要熵不等式取等号. 根据之前的讨论，必要条件是 X, Y, Z 相互独立，这时支撑集 Ω 一定是一个笛卡尔积 $\Omega = \mathrm{Supp}(X) \times \mathrm{Supp}(Y) \times \mathrm{Supp}(Z)$. 不难验证，$\Omega$ 是笛卡尔积也是 $n^2 = n_1 n_2 n_3$ 的充分条件.

7. (4 分) 考虑 Markov kernel $P_{Y|X} : \{0,1\} \times \{0,1\} \to \mathbb{R}$，$P_{Y|X}(0|0) = 1$，$P_{Y|X}(0|1) = P_{Y|X}(1|1) = \frac{1}{2}$.
   (1) 找到 P_X^* 使得 $I(X;Y)$ 最大，其中 $(X,Y) \sim P_X^* P_{Y|X}$. 这个最大值被称作 $P_{Y|X}$ 的容量.
   (2) 令 P_X^* 是前一问找到的分布. 定义 P_Y^* 为 $P_X^* P_{Y|X}$ 的边缘分布. 请计算 $I(X;Y), D(P_{Y|X=0} \| P_Y^*)$ 和 $D(P_{Y|X=1} \| P_Y^*)$ 的值.
   (3) (0 分) 现在考虑任意 Markov kernel $P_{Y|X} : \mathcal{Y} \times \mathcal{X} \to \mathbb{R}$. 为了保证最值存在，我们要求 $\mathcal{X}, \mathcal{Y}$ 都是有限集合. 证明

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

   课上我们定义了

$$D(P_{Y|X} \| Q_{Y|X} | P_X) = \sum_x P_X(x) D(P_{Y|X=x} \| Q_{Y|X=x}) = D(P_X P_{Y|X} \| P_X Q_{Y|X}).$$

   类似地，可以自然地定义 $D(P_{Y|X} \| Q_Y | P_X)$，只需将 Q_Y 视作一个退化的 kernel

$$D(P_{Y|X} \| Q_Y | P_X) = \sum_x P_X(x) D(P_{Y|X=x} \| Q_Y) = D(P_X P_{Y|X} \| P_X Q_Y).$$

解

(1) 不难证明 $P_X^* = \mathrm{Bern}(\frac{2}{5})$ 使 $I(X;Y)$ 最大.

(2) $I(X;Y) = D(P_{Y|X=0}\|P_Y^*) = D(P_{Y|X=1}\|P_Y^*) = \log\frac{5}{4}$.

(3) 首先证明第一个等号. 我们知道

$$D(P_{Y|X}\|Q_Y|P_X) = D(P_{Y|X}\|P_Y|P_X) + D(P_Y\|Q_Y)$$

因而

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

用 P_X^* 表示使 $I(X;Y)$ 最大的的 P_X，定义 P_Y^* 为 $P_X^*P_{Y|X}$ 的边缘分布，有

$$\max_{P_x} I(X;Y) = D(P_{Y|X}\|P_Y^*|P_X^*)$$

为了证明第二个等号，只需证明

$$\forall P_X,\ D(P_{Y|X}\|P_Y^*|P_X) \le D(P_{Y|X}\|P_Y^*|P_X^*), \tag{*}$$

因为 (*) 可以推出

$$\begin{aligned}\min_{Q_Y}\max_{P_X} D(P_{Y|X}\|Q_Y|P_X) &\le \max_{P_X} D(P_{Y|X}\|P_Y^*|P_X) \le D(P_{Y|X}\|P_Y^*|P_X^*)\\ &= \max_{P_x} I(X;Y) \le \max_{P_X}\min_{Q_Y} D(P_{Y|X}\|Q_Y|P_X) \le \min_{Q_Y}\max_{P_X} D(P_{Y|X}\|Q_Y|P_X).\end{aligned}$$

为了证明 (*)，对于任意 $\lambda\in[0,1]$，定义随机变量 Z, X_λ, Y_λ 满足 $Z\to X_\lambda\to Y_\lambda$. 其中

- $Z\sim\mathrm{Bern}(\lambda)$.
- X_λ 条件于 Z 的分布是 $P_{X_\lambda|Z=0} = P_X^*$ 而 $P_{X_\lambda|Z=1} = P_X$. 因而 X_λ 的边缘分布是 $P_{X_\lambda} = \lambda P_X + (1-\lambda)P_X^*$.
- Y_λ 条件于 X_λ 的分布是 $P_{Y|X}$.

$$\begin{aligned}I(P_X^*;P_{Y|X}) &\ge I(X_\lambda;Y_\lambda) = I(Z,X_\lambda;Y_\lambda)\\ &= I(Z;Y_\lambda) && + I(X_\lambda;Y_\lambda|Z)\\ &= D(P_{Y_\lambda|Z}\|P_{Y_\lambda}|P_Z) && + I(X_\lambda;Y_\lambda|Z)\\ &= \lambda D(P_Y\|P_{Y_\lambda}) + (1-\lambda)D(P_Y^*\|P_{Y_\lambda}) && + \lambda I(X;Y) + (1-\lambda)I(P_X^*;P_{Y|X})\\ &\ge \lambda D(P_Y\|P_{Y_\lambda}) && + \lambda I(X;Y) + (1-\lambda)I(P_X^*;P_{Y|X})\end{aligned}$$

也就是

$$\lambda I(P_X^*;P_{Y|X}) \ge \lambda D(P_Y\|P_{Y_\lambda}) + \lambda I(P_X;P_{Y|X}).$$

两边除以 λ，再取 $\lambda \to 0$ 的极限，得到

$$I(P_X^*; P_{Y|X}) \geq D(P_Y \| P_Y^*) + I(P_X; P_{Y|X}).$$

现在可以推出 (*)

$$\begin{aligned} D(P_{Y|X} \| P_Y^* | P_X) = D(P_{Y|X} \| P_Y | P_X) + D(P_Y \| P_Y^*) &= I(P_X; P_{Y|X}) + D(P_Y \| P_Y^*) \\ &\leq I(P_X^*; P_{Y|X}) = D(P_{Y|X} \| P_Y^* | P_X^*). \end{aligned}$$