散度，集中不等式

请在 12 月 5 日课前提交纸质作业.

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.

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

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

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

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

(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，

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

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

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

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 $\mathrm{Supp}(P) = \{v_1, \ldots, v_T\} \subseteq \mathbb{R}^+$. 记 $p_i := P(v_i) > 0$. 这个分布的期望是 $\bar{v} = \sum p_i v_i$. 考虑任意 $b \in (\bar{v}, \max_i v_i)$，定义

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

根据 Sanov's Theorem，

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

而根据 Chernoff bound，

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

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.

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

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

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 $I(X;Y) \ge I(X;Z)$.

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)

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

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 $n^2 \leq n_1n_2n_3$. 并说明何时可以取到等号.

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

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

   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

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