集中不等式, 离散傅里叶变换

请在 12 月 4 日课前提交纸质作业.

1. (10 分) 完成以下关于 Chernoff bound 的证明.

   (1) (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

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

   (2) 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 $0 \leq p \leq q \leq 1$ 时,

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

   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.

   (3) 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 $0 \leq p \leq q \leq 1$ 时,

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

2. (10 分) 根据 Sanov's Theorem 我们可以看出, Chernoff bound 对于

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

   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 $\mathrm{Supp}(P) = \{v_1, \ldots, v_T\} \subseteq \mathbb{R}^+$. 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

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

   根据 Sanov's Theorem,

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

而根据 Chernoff bound,

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

请问是否存在 P 和 $b\in(\bar{v},\max_i v_i)$ 使得 Chernoff bound 的估计要弱于 Sanov's Theorem?

提示：拉格朗日乘数.

3. (6 分) 对于一个布尔函数 $f:\{0,1\}^n\to\{0,1\}$, 函数第 i 位的影响被定义为

$$\mathrm{Influence}_i(f):=\Pr_{x\leftarrow\{0,1\}^n}\big[f(x)\neq f(x\oplus e_i)\big],$$

其中 $e_i=(0,\ldots,0,1,0,\ldots,0)$ 只在第 i 位等于 1.

(1) 布尔函数 f 被称为单调（monotone）, 如果 $x\geq y\implies f(x)\geq f(y)$. （这里 $x\geq y$ 表示 $\forall i,x_i\geq y_i$.）请寻找一个单调的布尔函数 f, 使得 $\sum_i\mathrm{Influence}_i(f)$ 最大, 并证明.

(2) 布尔函数 f 被称为平衡（balanced）, 如果 $\Pr_{x\leftarrow\{0,1\}^n}[f(x)=1]=\frac{1}{2}$. 请寻找一个平衡的布尔函数 f, 使得 $\max_i\mathrm{Influence}_i(f)$ 尽量小, 可以忽略常数系数.

Remark: 证明 $\max_i\mathrm{Influence}_i(f)$ 的下界需要非常有技巧地使用傅里叶变换. 本题不需要证明结果最优.

4. (12 分) 对于函数 $f:\{0,1\}^n\to\mathbb{R}$, 用 f 的傅里叶系数表示如下量

(1) $\widehat{g_s}(x)$, 其中 $g_s(x):=f(x\oplus s)$.

(2) $\widehat{g_y}(x)$, 其中 $g_y(x):=(-1)^{\langle x,y\rangle}f(x)$.

(3) $\widehat{f_i}(x)$, 其中 $f_i(x):=f(x)-f(x\oplus e_i)$.

(4) $\widehat{g_k}(x)$, 其中 $g_k(x_1,\ldots,x_k):=f(x_1,\ldots,x_k,0,\ldots,0)$.

(5) $\widehat{g_a}(x)$, 其中 $g_a(x_1,\ldots,x_k):=\mathbb{E}_{y\leftarrow\{0,1\}^{n-k}}[f(x,y)\chi_a(y)]$.

(6) $\mathrm{Var}[f(X)]$, 其中 X 服从均匀分布.

这里约定 $\hat{f}(a)=\mathbb{E}_x[f(x)\overline{\chi_a(x)}]$, $f(x)=\sum_a\hat{f}(a)\chi_a(x)$.

5. (5 分) 对于函数 $f:\{0,1\}^n\to\mathbb{R}$, 它的傅里叶系数 $\hat{f}:\{0,1\}^n\to\mathbb{R}$ 满足

$$\hat{f}(y)=\frac{1}{2^n}\sum_x f(x)\chi_y(x),\qquad f(x)=\sum_y\hat{f}(y)\chi_y(x).$$

考虑一个“噪音算子”$T_\rho:(\{0,1\}^n\to\mathbb{R})\to(\{0,1\}^n\to\mathbb{R})$, 其中 $\rho\in[0,1]$.

$$T_\rho(f)(x)=\mathop{\mathbb{E}}_{y\sim(\mathrm{Bern}(\rho))^n}\Big[f(x\oplus y)\Big].$$

求 $\widehat{T_\rho(f)}(y)$. 化简后的表达式不应该出现 $\sum$ 或 $\mathbb{E}$.

6. (10 分) 令 $X_1, \ldots, X_{2n}$ i.i.d. 服从 Bern(1/2) 分布. 我们要选取一组系数 $c_1, \ldots, c_{2n} \in \mathbb{Z}$, 使得 $\sum_i c_i X_i$ 接近均匀分布. 当然, 并不存在 $\mathbb{Z}$ 上的均匀分布, 我们实际的要求是统计距离

$$\Delta\Big(\sum_i c_i X_i, \sum_i c_i X_i + 1\Big) \le 2^{-\lambda}. \tag{*}$$

一种显然的做法, 是令 $n = \lambda/2$, 令 $c_i = 2^{i-1}$; 这样 $\sum_i c_i X_i$ 服从 $\{0, 1, \ldots, 2^\lambda - 1\}$ 上的均匀分布, 而 $\sum_i c_i X_i + 1$ 服从 $\{1, 2, \ldots, 2^\lambda\}$ 上的均匀分布, 满足我们对统计距离的要求.

进一步, 假设 $X_1, \ldots, X_n$ 中有一半值被泄漏. 要求即使已经看到泄露值, (*) 仍然成立.

为了方便分析, 我们令 $c_1, \ldots, c_n$ 是 i.i.d. 从某个分布 P_C 中选取的. 这样不管哪部分值泄露, 分析都相同. 不失一般性, 可以假设前一半值没有泄露. 定义函数

$$\mathrm{Err}(c_1, \ldots, c_n) = \Delta\Big(\sum_i c_i X_i, \sum_i c_i X_i + 1\Big)$$

我们要求当 $c_1, \ldots, c_n$ 是从 $(P_C)^n$ 选取时, 以 $1 - 2^{-\lambda}$ 的概率（这个随机性只依赖于 $c_1, \ldots, c_n$）

$$\mathrm{Err}(c_1, \ldots, c_n) \le 2^{-\lambda}.$$

请根据 λ, 选取合适的 n 以及分布 P_C, 使得要求被满足. 请让 n 的取值尽量小, 可以忽略常数系数. 建议选取 P_C 为 $\{1, 2, 3, \ldots, B\}$ 上的均匀分布, 其中 $B = 2^{O(\lambda)}$ 根据 λ 选取.