[CSS.413.1] Pseudorandomness 15 Sep, 2025

# Problem Set 1

- Due Date: **28 Sep 2025**
- The points for each problem is indicated on the side. The total for this set is **70** points.
- Turn in your problem sets electronically (PDF; either LaTeXed or scanned etc.) via email.
- Collaboration is encouraged, but all writeups must be done individually and must include names of all collaborators.
- You are not allowed to use the assistance of LLMs for this problem set.
- Referring to sources other than the text book and class notes is strongly discouraged. But if you do use an external source (eg., other text books, lecture notes, or any material available online), ACKNOWLEDGE all your sources (including collaborators) in your writeup. This will not affect your grades. However, not acknowledging will be treated as a serious case of academic dishonesty.
- Be clear in your writing.

1. **[Derandomising Turán's theorem]** (3 + 7)

   Let $G = (V, E)$ be an undirected graph. For a vertex $v \in V$, let $d(v)$ denote the degree of the vertex $v$ in $G$. Let $d_{\text{avg}} = 2|E|/|V|$ denote the average degree.

   (a) Show that any such graph $G$ has an independent set (a subset of vertices such that no two of them are connected) of size at least

   $$\sum_{v \in V} \frac{1}{d(v)+1} \geq \frac{|V|}{d_{\text{avg}+1}}$$

   [Hint: Consider the set of vertices in a random order and pick an independent set greedily. What size do you get on expectation? AM-HM should be helpful for the inequality.]

   (b) Come up with a deterministic polynomial time algorithm to compute an independent set of size of the above size.

2. **[Derandomising discrepancy bounds]** (3 + 7)

   Suppose you are given a collection $n$ sets $S_1, \ldots, S_n \subseteq [m]$. The goal is to colour the universe $[m]$ into red and blue elements so that the 'imbalance / discrepancy' in each set $S_i$ is as small as possible. Formally, we wish to compute $x_1, \ldots, x_m = \pm 1$ such that $\max_i \left|\sum_{j \in S_i} x_j\right|$ is as small as possible.

   (a) Use the probabilistic method to show that there *exists* such a colouring $x_1, \ldots, x_m \in \pm 1$ such that $\left|\sum_{j \in S_i} x_j\right| = O(\sqrt{m \log n})$.

Base version: (2025-09-15 22:46:29 +0530) , 5227344

(b) Come up with a *deterministic* polynomial time algorithm to compute a colouring satisfying maximum discrepancy of $O(\sqrt{m \log n})$.

[Hint: Let $p_i = \Pr[S_i$ is $t$-imbalanced] and let $R = \sum_{i=1}^{n} p_i$.]

3. **[Some candidate constructions of pairwise independent hash families]** (10)

Which of the following family of functions of the form $\{h : \{0,1\}^n \to \{0,1\}^n\}$ constitute a pairwise independent hash family? Support your answer with a proof of pairwise independence (if yes), or provide a counter-example (if no).

(a) $\mathcal{H} = \left\{h_A(x) = Ax \; : \; A \in \mathbb{F}_2^{n \times n}\right\}$. That is, each hash function is specified by a matrix $A$ and the hash function is just matrix-vector multiplication (over $\mathbb{F}_2$).

A random function from the family is chosen by picking the matrix $A$ uniformly at random.

(b) $\mathcal{H} = \left\{h_{A,b}(x) = Ax + b \; : \; A \in \mathbb{F}_2^{n \times n} \; , \; b \in \mathbb{F}_2^n\right\}$. That is, each hash function is given by multiplication by a matrix $A$ followed by adding $b$ (again, over $\mathbb{F}_2$).

A random function from the family is chosen by picking the matrix $A$ and vector $b$ uniformly at random.

4. **[Lower bound for $k$-wise independent families]** (10)

For this problem, we will only consider families of the form $\mathcal{H} = \{h : [n] \to \{0,1\}\}$. Each such $h : [n] \to \{0,1\}$ can be thought of as just a string in $\{0,1\}^n$ and hence $\mathcal{H}$ is just some (multi-)set of strings in $\{0,1\}^n$.

Rephrasing the definition of $k$-wise independent in this setting, we have that for any distinct $i_1, \ldots, i_k \in [n]$ and (not necessarily distinct) $a_1, \ldots, a_k \in \{0,1\}$,

$$\Pr_{x \in \mathcal{H}}[x_{i_1} = a_1 \; , \; \ldots \; , \; x_{i_k} = a_k] = \frac{1}{2^k}.$$

For any $T \subseteq [n]$, define $\chi_T : \{0,1\}^n \to \mathbb{R}$ as $\chi_T(x) = (-1)^{\sum_{i \in T} x_i}$.

(a) Suppose $\mathcal{H}$ was a $k$-wise independent (multi-)set. Consider the following collection of vectors in $\mathbb{R}^{|\mathcal{H}|}$:

$$\{(\chi_T(x) \; : \; x \in \mathcal{H})\}_{T \subseteq [n] \; , \; |T| \leq (k/2)}$$

That is, there is a vector for each $T \subseteq [n]$ of size at most $k/2$, and each such vector consists of the evaluation of $\chi_T$ on the points in $\mathcal{H}$.

Show that the above set of vectors are linearly independent over $\mathbb{R}$.

(b) Conclude that $|\mathcal{H}| \geq \sum_{i=0}^{k/2} \binom{n}{i}$.

5. **[Better tail bounds with higher independence]** (7 + 3)

Suppose $X_1, \ldots, X_t$ are random variables taking values in $[0,1]$ and let $X = X_1 + \cdots + X_t$. Let $\mu_i = \mathbb{E}[X_i]$, and $\mu = \sum \mu_i = \mathbb{E}[X]$. Suppose that these random variables are 4-wise independent, i.e. for any set of 4-distinct indices $i_1, i_2, i_3, i_4$ and any *events* $A_1, A_2, A_3, A_4 \subseteq [0,1]$, we have

$$\Pr\left[X_{i_1} \in A_1 \; , \; \ldots \; , \; X_{i_4} \in A_4\right] = \prod_{j=1}^{4} \Pr[X_{i_j} \in A_j].$$

(a) Prove that $\mathbb{E}[(X-\mu)^4] \leq O(t+t^2)$

[Hint: Rewrite $(X-\mu)^4 = (Y_1 + \cdots + Y_t)^4$ where $Y_i = X_i - \mu_i$. What happens to terms that have $Y_i$ with a single power (i.e. not terms of the form $Y_1^2 Y_2^2$, but terms such as $Y_1 Y_2^3$)?]

(b) Conclude that $\Pr[|X-\mu| \geq t\varepsilon] \leq O\left(\frac{1}{t^2\varepsilon^4}\right)$ in the 4-wise independent case.

(c) **[extra credit]** Can you generalise this to $k$-wise independence (for even $k$)? That is, show that if $X_1, \ldots, X_t$ are $k$-wise independent and $X = \sum X_i$, then

$$\Pr[|X-\mu| > t\varepsilon] \leq O\left(\frac{k^k}{t^{k/2}\varepsilon^k}\right)$$

[Hint: Once again, expand out $\mathbb{E}[(Y_1 + \ldots + Y_t)^k]$ as earlier and argue that the only terms that matter are those where each $Y_i$ in that terms appears at least with an exponent of 2. Use this to show $\mathbb{E}[(Y_1 + \cdots + Y_t)^k] \leq O(k^k \cdot t^{k/2})$. ]

6. **[Existence of bipartite biregular left-vertex-expanders]** (10)

In class, we saw the proof of existence of bipartite left-vertex-expanders that were left-regular for appropriate setting of parameters. To do this, we let each vertex on the left pick $D$ random neighbours on the right, and argued that the size of the neighbourhood of any set $S$ of size $K \leq \alpha n$ was at least $(D-2) \cdot K$ with high probability (for a suitable choice of $\alpha$).

Extend the proof to show the existence of bipartite biregular graphs that are left-vertex-expanding by picking such a graph $G$ by taking a union of $D$ perfect matchings.

7. **[Not an averaging sampler]** (5 + 5)

The following template known as "median-of-averages" is often used to improve a general sampler. Let $\mathcal{A}(\delta, \varepsilon)$ be an arbitrary $(\delta, \varepsilon)$-sampler for $m$-coordinate functions and suppose this sampler makes $q(\delta, \varepsilon)$ queries to the function and uses $r(\delta, \varepsilon)$ random bits. From $\mathcal{A}$, consider the following alternate sampler:

> Let $t$ be a positive integer (to be chosen by you). Run $t$ independent runs of $\mathcal{A}(0.1, \varepsilon)$ to get estimates $\mu_1, \ldots, \mu_t$. Return the median of $\mu_1, \ldots, \mu_t$.

(a) What should you choose $t$ to be so that the above gives a $(\delta, \varepsilon)$-sampler?

(b) If $\mathcal{A}$ was instantiated to be the pairwise independent sampler that we saw in class, how many queries does the above sampler make and how many random bits does it use?