Problem 1 - (20pt)

Let $X \sim \mathcal{N}(\mu, \sigma^2)$ be a Gaussian random variable (r.v):

$$\text{Prob}(X) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\Big(- \frac{(x-\mu)^2}{2\sigma^2} \Big).$$

(a) [10pt] Prove the following statement: For any $s \in \mathbb{R}$,

$$\mathbb{E}\big[\exp(sX)\big] = \exp\Big(\mu s + \frac{\sigma^2 s^2}{2}\Big).$$

Solution.

$$\begin{aligned}
\mathbb{E}\big[\exp(sX)\big] &= \int_{-\infty}^{\infty} \exp(sx) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big(- \frac{(x-\mu)^2}{2\sigma^2} \Big) dx : \text{definition of expectation.} && (1) \\
&= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big(- \frac{(x-\mu)^2 - 2\sigma^2 sx}{2\sigma^2} \Big) dx, && (2) \\
&= \exp\Big(\mu s + \frac{\sigma^2 s^2}{2}\Big) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} exp\Big(- \frac{(x-(\mu+\sigma^2 s))^2}{2\sigma^2} \Big) dx, \ \textbf{[5pt]} && (3) \\
&= \exp\Big(\mu s + \frac{\sigma^2 s^2}{2}\Big). \ \textbf{[5pt]} && (4)
\end{aligned}$$

Note that the RHS of (3) equals to 1, since it is the integral of probability distribution $\mathcal{N}(\mu + \sigma^2 s, \sigma^2)$.

(b) [10pt] Prove the following statement: For any $t > 0$,

$$\text{Prob}(|X - \mu| \geq t) \leq 2\exp\Big(-\frac{t^2}{2\sigma^2}\Big).$$

Solution.

Since $X - \mu \sim \mathcal{N}(0, \sigma^2)$ is a symmetrical distribution on 0,

$$\text{Prob}(|X - \mu| \geq t) = \text{Prob}(X - \mu \geq t) + \text{Prob}(X - \mu \leq -t) = 2\text{Prob}(X - \mu \geq t). \tag{5}$$

To bound the RHS of (5),

$$\begin{aligned}
\text{Prob}(X - \mu \geq t) &= \text{Prob}\Big(\exp(sX) \geq \exp\big(s(\mu + t)\big)\Big) \text{ for } s \geq 0, &(6)\\
&\leq \frac{\mathbb{E}\big(\exp(sX)\big)}{\exp\big(s(\mu + t)\big)} \ : \text{ Markov's inequ}, &(7)\\
&= \exp\Big(\frac{s^2\sigma^2}{2} - st\Big) \ : \text{ by using (4)} \quad \textbf{[5pt]} &(8)
\end{aligned}$$

For a tight bound, we minimize the RHS of (8) w.r.t s by differentiating the term in the exponential and set to zero:

$$\frac{\partial}{\partial s}\Big(\frac{s^2\sigma^2}{2} - st\Big) = s\sigma^2 - t = 0, \tag{9}$$

which results in $s = \frac{t}{\sigma^2}$. By applying $s = \frac{t}{\sigma^2}$ into (8),

$$\text{Prob}(X - \mu \geq t) \leq \exp\Big(-\frac{t^2}{2\sigma^2}\Big). \tag{10}$$

By applying (10) into (5), we result in

$$\text{Prob}(|X - \mu| \geq t) \leq 2\exp\Big(-\frac{t^2}{2\sigma^2}\Big) \ \textbf{[5pt]}. \tag{11}$$

Problem 2 - (20pt) Consider a 1-dimensional lattice with n vertices. The transition probability of each neighbor is $\frac{1}{2}$. Each boundary point has a self-loop with transition probability of $\frac{1}{2}$. Find the ε-mixing time for this graph in the following way:

(a) [7pt] Find the stationary distribution.

(b) [8pt] Compute the normalized conductance Φ.

(c) [5pt] Compute the ε-mixing time.

solution.

(a) By lemma 4.3, it suffices to find $\boldsymbol{\pi}$ such that $\pi_x p_{xy} = \pi_y p_{yx}$ for all x and y and $\sum_x \pi_x = 1$. We only need to consider neighboring x and y because the transition probability of non-neighboring vertices are 0. When x and y are neighboring, $p_x = p_y = \frac{1}{2}$. Therefore, the stationary probabilities of all vertices are equal. As there are n vertices, the stationary distribution is $\frac{1}{n}$ for all vertices.

(b) The set with minimum normalized conductance is the set S with probability $\pi(S) \leq \frac{1}{2}$ having the smallest ratio of probability mass exisiting it, $\sum_{(x,y)\in(S,\bar{S})} \pi_x p_{xy}$, to probability mass inside it, $\pi(S)$. This set consists of the first $\frac{n}{2}$ vertices. Therefore, $\Phi = \frac{1}{2n}/\frac{1}{2} = \frac{1}{n}$.

(c) By theorem 4.5, ε-mixing time is $O(n^2 \log n/\varepsilon^3)$.

Problem 3 - (20pt)

For $k = 1, 2, ..., M$, let r_k be some positive integer and $\mathcal{F}_k : \{-1, 1\}^{r_k} \to \{-1, 1\}$ be some function class with VC-dimension d_k. Define $\mathcal{G}_k : \{-1, 1\}^{r_k} \to \{-1, 1\}^{r_{k+1}}$ with $r_{M+1} = 1$ as:

$$\mathcal{G}_k = \{h(x) = (f_1(x), ..., f_{r_{k+1}}(x)) | f_1, ..., f_{r_{k+1}} \in \mathcal{F}_k\}.$$

Our hypothesis class $\mathcal{H}$ is M-layer feedforward network defined by:

$$\mathcal{H} = \{h_M \circ ... \circ h_1 : \{-1, 1\}^{r_1} \to \{-1, 1\} | h_1 \in \mathcal{G}_1, ..., h_M \in \mathcal{G}_M\}.$$

(a) [6pt] Prove that the growth function of $\mathcal{H}$ (denoted by $\mathcal{H}[n]$) is bounded as:

$$\mathcal{H}[n] \leq \Pi_{k=1}^{M} (\mathcal{F}_k[n])^{r_{k+1}}$$

where $\mathcal{F}_k[n]$ is the growth function of $\mathcal{F}_k$.

(b) [6pt] Prove $\mathcal{H}[n] \leq (en)^d$, where $d = \sum_{k=1}^{M} r_{k+1} d_k$.

Hint: Use Sauer's Lemma: If d is the VC-dimension of $\mathcal{H}$, then $\mathcal{H}[n] \leq (\frac{en}{d})^d$.

(c) [8pt] Show that VC-dimension of $\mathcal{H}$ is $O(d \ln d)$.

Hint: Use the result of (b), when $n \geq 16$, we have $\log_2 n \leq \sqrt{n}$.

solution.

(a) Firstly, we prove useful lemmas.

Lemma 1. *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two function classes with the same domain X. Let $\mathcal{H} = \mathcal{H}_1 \times \mathcal{H}_2$ be their cartesian product. Then,*

$$\mathcal{H}[n] \leq \mathcal{H}_1[n]\mathcal{H}_2[n].$$

Proof. Fix $x^n \in X^n$. Then,

$$|\mathcal{H}(x^n)| = |\mathcal{H}_1(x^n)||\mathcal{H}_2(x^n)| \leq \mathcal{H}_1[n]\mathcal{H}_2[n].$$

As $x^n \in X^n$ was arbitrary, we have $\mathcal{H}[n] \leq \mathcal{H}_1[n]\mathcal{H}_2[n]$. Note that $\mathcal{H}(x^n) := \{(h(x_1), ..., h(x_n) | x^n = (x_1, ..., x_n) \in X^n, h \in \mathcal{H}\}$. □

Lemma 2. *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two function classes where the range of $\mathcal{H}_1$ and the domain of $\mathcal{H}_2$ coincide. Let $\mathcal{H} = \mathcal{H}_1 \circ \mathcal{H}_2$ be their composition. Then,*

$$\mathcal{H}[n] \leq \mathcal{H}_1[n]\mathcal{H}_2[n].$$

Proof. Let the domain of $\mathcal{H}_1$ and $\mathcal{H}_2$ be X and Y, respectively. Fix $x^n \in X^n$. Then,

$$\begin{aligned}\mathcal{H}(x^n) &= \{(h_2(h_1(x_1)), ..., h_2(h_1(x_n))) | h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2\} \\ &= \cup_{y^n \in \mathcal{H}_1(x^n)} \{(h_2(y_1), ..., h_2(y_n)) | h_2 \in \mathcal{H}_2\}.\end{aligned}$$

Therefore,

$$
\begin{aligned}
|\mathcal{H}(x^n)| &\leq \sum_{y^n \in \mathcal{H}_1(x^n)} |\{(h_2(y_1), ..., h_2(y_n)) | h_2 \in \mathcal{H}_2\}| \\
&\leq \sum_{y^n \in \mathcal{H}_1(x^n)} \mathcal{H}_2[n] \\
&= |\mathcal{H}_1(x^n)| \mathcal{H}_2[n] \\
&\leq \mathcal{H}_1[n] \mathcal{H}_2[n].
\end{aligned}
$$

As $x^n \in X^n$ was arbitrary, we have $\mathcal{H}[n] \leq \mathcal{H}_1[n]\mathcal{H}_2[n]$. □

Because $\mathcal{H} = \mathcal{G}_M \circ \mathcal{G}_{M-1} \circ ... \circ \mathcal{G}_1 = \mathcal{F}_M^{r_{M+1}} \circ \mathcal{F}_{M-1}^{r_M} \circ ... \circ \mathcal{F}_1^{r_2}$, by lemma 1 and 2, we have $\mathcal{H}[n] \leq \Pi_{k=1}^{M} (\mathcal{F}_k[n])^{r_{k+1}}$.

(b) By Sauer's lemma, $\mathcal{H}[n] \leq (en/d_1)^{r_2 d_1} \cdot ... \cdot (en/d_{M-1})^{r_M d_{M-1}} \cdot (en/d_M)^{d_M} \leq (en)^{r_2 d_1 + ... + r_M d_{M-1} + r_{M+1} d_M} = (en)^d$.

(c) If the set of size n is shattered by $\mathcal{H}$, then $\mathcal{H}[n] = 2^n$. Note that the largest n with $\mathcal{H}[n] = 2^n$ is the VC-dimension of $\mathcal{H}$.

For $\mathcal{H}[n] = 2^n$, it suffices to have $n \leq d\log_2(en) \leq d\sqrt{en}$ (i.e., $n \leq d^2 e$), for sufficiently large n. By plugging in the original inequality, $n \leq d\log_2(e^2 d^2)$. Hence, the VC-dimension of $\mathcal{H}$ is $O(d\ln d)$.

Problem 4 - (20pt)

Prove the following statement.

Let $b_1, b_2, \ldots, b_d$ be the distinct values that appear in the input. Select h from the 2-universal family of hash functions. Then the set $S = \{h(b_1), h(b_2), \ldots, h(b_d)\}$ is a set of d random and pairwise independent values from the set $\{0, 1, 2, \ldots, M-1\}$. Show that with probability at least $\frac{2}{3} - \frac{d}{M}$, we have $\frac{d}{6} \le \frac{M}{\min} \le 6d$, where min is the smallest element of S.

Solution.

(10pt) First, show that $\text{Prob}\left(\frac{M}{\min} > 6d\right) < \frac{1}{6} + \frac{d}{M}$.

$$\begin{aligned}\text{Prob}\left(\frac{M}{\min} > 6d\right) &= \text{Prob}\left(\min < \frac{M}{6d}\right) = \text{Prob}\,(\exists k, h(b_k) < \frac{M}{6d}) \\ &\le \sum_{i=1}^{d} \text{Prob}\left(h(b_i) < \frac{M}{6d}\right) \le d\left(\frac{\lceil \frac{M}{6d} \rceil}{M}\right) \le d\left(\frac{1}{6d} + \frac{1}{M}\right) \le \frac{1}{6} + \frac{d}{M}.\end{aligned}$$

(10pt) Next, show that $\text{Prob}(\frac{M}{\min} < \frac{d}{6})$. This part will use pairwise independence. First, $\text{Prob}\left(\frac{M}{\min} < \frac{d}{6}\right) = \text{Prob}(\min > \frac{6M}{d}) = \text{Prob}(\forall k, h(b_k) > \frac{6M}{d})$. For $i = 1, 2, \ldots, d$, define the indicator variable

$$y_i = \begin{cases} 0 & \text{if} \quad h(b_i) > \frac{6M}{d} \\ 1 & \text{otherwise} \end{cases}$$

and let

$$y = \sum_{i=1}^{d} y_i.$$

For 2-way independent random variables, the variance of their sum is the sum of their variances. So $\text{Var}(y) = d\text{Var}(y_1)$. Further, since y_1 is 0 or 1, $\text{Var}(y_1) = E[(y_1 - E(y_1))^2] = E(y_1^2) - E^2(y_1) \le E(y_1)$. Thus $\text{Var}(y) \le E(y)$. By the Chebyshev inequality,

$$\begin{aligned}\text{Prob}\left(\frac{M}{\min} < \frac{d}{6}\right) &= \text{Prob}(\min > \frac{6M}{d}) = \text{Prob}\left(\forall k\ h(b_k) > \frac{6M}{d}\right) \\ &= \text{Prob}(y = 0) \\ &\le \text{Prob}(|y - E(y)| \ge E(y)) \\ &\le \frac{\text{Var}(y)}{E^2(y)} \le \frac{1}{E(y)} \le \frac{1}{6}\end{aligned}$$

since $\frac{M}{\min} > 6d$ with probability at most $\frac{1}{6} + \frac{d}{M}$ and $\frac{M}{\min} < \frac{d}{6}$ with probability at most $\frac{1}{6}$, $\frac{d}{6} \le \frac{M}{\min} \le 6d$ with probability at least $\frac{2}{3} - \frac{d}{M}$.

Problem 5 - (20pt)

Prove the following properties (a, b, c, d) of a graph Laplacian matrix L, which plays an important role in spectral clustering algorithms.

Consder a undirected weighted graph $G = (V, E)$ (with no self-loop), where $V = \{1, 2, \ldots, n\}$ is the set of vertices, $E \subset V \times V$ is the set of edges, and $[w_{ij} \geq 0 : (i, j) \in E]$ are weights on edges. The graph Laplacian matrix L is defined as $D - W$, where $D = [D_{ij}]$ is the diagonal matrix such that D_{ii} is the degree of vertex i, defined as $D_{ii} = \sum_{j=1}^{n} w_{ij}$ in the weighted edge case, and W is the weighted adjacency matrix such that $W_{ii} = 0$ and $W_{ij} = \begin{cases} w_{ij} & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases}$, i.e., the weight of the edge between $i \in V$ and $j \in V$.

(a) [5pt] For every vector $f \in R^n$, we have $f^T L f = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij}(f_i - f_j)^2$.

Solution.

$$\begin{aligned}
f^T L f &= f^T D f - f^T W f \\
&= \sum_{i=1}^{n} f_i^2 d_{ii} - \sum_{i=1}^{n} \sum_{j=1}^{n} f_i f_j w_{ij} \\
&= \frac{1}{2} \left(\sum_{i=1}^{n} f_i^2 d_{ii} - 2 \sum_{i=1}^{n} \sum_{j=1}^{n} f_i f_j w_{ij} + \sum_{j=1}^{n} f_j^2 d_{jj} \right)
\end{aligned}$$

Now use that $d_{ii} = \sum_{j=1}^{n} w_{ij}$ to get:

$$\begin{aligned}
&= \frac{1}{2} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} f_i^2 - 2 \sum_{i=1}^{n} \sum_{j=1}^{n} f_i f_j w_{ij} + \sum_{j=1}^{n} \sum_{i=1}^{n} w_{ij} f_j^2 \right) \\
&= \frac{1}{2} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} (w_{ij} f_i^2 - 2 f_i f_j w_{ij} + w_{ij} f_j^2) \right) \\
&= \frac{1}{2} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (f_i - f_j)^2 \right).
\end{aligned}$$

(b) [5pt] L is symmetric and positive semi-definite.

Solution.

Since D and W are symmetric, so it $L = D - W$.

Since $f^T L f = \frac{1}{2} \left(\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (f_i - fj)^2 \right)$ for all $f \in R^n$, this quantity can never be negative due to the requirement that weights w_{ij} be nonnegative. The quantity can equal to 0 if $f_i = f_j$. Thus, $f_T L f \geq 0$ for all f, which is the definition of a positive semi-definite matrix L.

(c) [5pt] The smallest eigenvalue of L is 0, and the corresponding eigenvector is the constant $\mathbf{1}$.

Solution.

If we multiply L by the constant vector $\mathbf{1}$, we get:

$$\begin{aligned} L\mathbf{1} &= D\mathbf{1} - W\mathbf{1} \\ &= \begin{pmatrix} d_{11} \\ d_{22} \\ \vdots \\ d_{nn} \end{pmatrix} - \begin{pmatrix} \sum_{j=1}^{n} w_{1j} \\ \sum_{j=1}^{n} w_{2j} \\ \vdots \\ \sum_{j=1}^{n} w_{nj} \end{pmatrix} \end{aligned}$$

Here we use again that $d_{ii} = \sum_{j=1}^{n} w_{ij}$ to get:

$$= \begin{pmatrix} d_{11} \\ d_{22} \\ \vdots \\ d_{nn} \end{pmatrix} - \begin{pmatrix} d_{11} \\ d_{22} \\ \vdots \\ d_{nn} \end{pmatrix} = \mathbf{0}.$$

(d) [5pt] L has n non-negative, real-valued eigenvalues $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$.

Solution.

By properties 2 and 3, L is positive semi-definite and has an eigenvalue of 0, so 0 must be the smallest eigenvalue and all other eigenvalues must be non-negative. Since L is real and symmetric, it must have real eigenvalues and n independent eigenvectors (since by the spectral theorem there exists a Q such that $L = Q^T \Lambda Q$, where Q is an orthogonal basis for R^n and Λ is the diagonal matrix with Λ_{ii} is the i-th eigenvalue of L.) Therefore, L has n real-valued, non-negative eigenvalues with $0 = \lambda_1 \leq \lambda_2 \leq \cdots \lambda_n$.