CPSC 483/583: Deep Learning on Graph-Structured Data Yale University

Assignment 4

Out: October 30, 2024 Due: November 13, 2024

General Instructions

These questions require thought, but do not require long answers. Please be as concise as possible. You are allowed to take a maximum of 1 late period (see the course website or slides about the definition of a late period).

Submission instructions: You should submit your answers in a *single* PDF file. LaTeX is highly preferred due to the need of formatting equations.

Submitting answers: Prepare answers to your homework in a *single* PDF file. Make sure that the answer to each sub-question is on a *separate page*. The number of the question should be at the top of each page.

Honor Code: When submitting the assignment, you agree to adhere to the Yale Honor Code. Please read carefully to understand what it entails!

1 Hyperbolic Graph Embedding

1. Hyperboloid model (aka Minkowski model or Lorentz model) is a model of n-dimensional hyperbolic geometry in which points are represented in the upper sheet of a two-sheeted Hyperboloid in $(n+1)$-dimensional Minkowski space as shown in Figure 1. Consider the 3D Minkowski space, 2D Hyperboloid satisfies the Cartesian coordinates equation $x^2+y^2-z^2=-1$, where $z>0$.

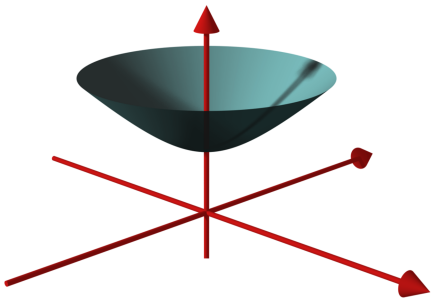

Figure 1: Hyperboloid Model
s

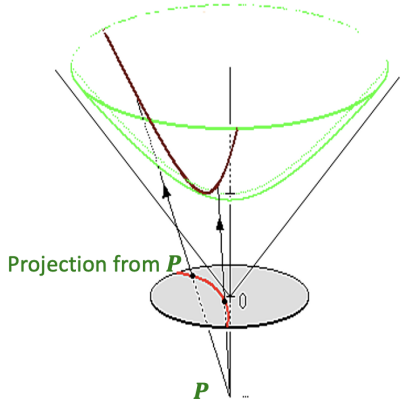

Figure 2: Hyperboloid Projection

The Poincaré model can be obtained with a projection of the Hyperboloid model from the focal point $P=(0,0,-1)$ onto the $z=0$ plane. The projection is shown in Figure 2. For point (x^h,y^h,z^h) in the 2D Hyperboloid model, let (x^p,y^p) denote the projected point in the corresponding 2D Poincaré disk on the $z=0$ plane. Use x^h,y^h,z^h to represent x^p and y^p.

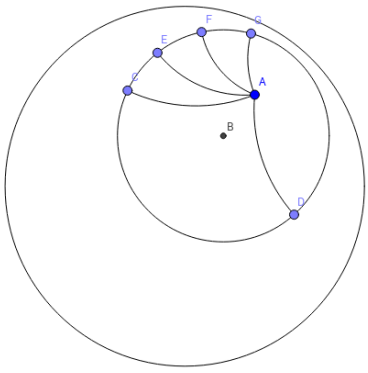

Figure 3: Hyperbolic Circle

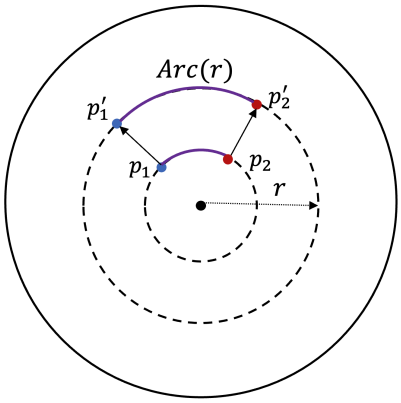

Figure 4: Poincaré Disk

2. Consider a hyperbolic circle in the 2D Poincaré model in which all points have a fixed distance from the center of the circle. The fixed distance is called the Hyperbolic radius r. Due to the negative curvature of the Poincaré model, the distance of points increases as they are closer to the boundary. Therefore, the center of the hyperbolic circle is different from the euclidean center. As shown in Figure 3, the center of the hyperbolic circle is A and all points on the circle ($e.g., C, D, E, F, G$) have the same distance from the center A. In hyperbolic geometry, the circumference of a hyperbolic circle is $C = 2\pi\sinh(r)$ with a hyperbolic radius r.

   (a) Compare the circumference of a hyperbolic circle with the one in Euclidean space of the same radius, which one has a larger circumference? Prove your answer.

   (b) If we fix the circle to be centered at the origin of the Poincaré model, prove that the arc length between any pairs of two points on the circle increases exponentially as the radius r increases (**Note**: As Figure 4 shows, point p_1 moves to p_1' and p_2 moves to p_2'. The two points move away from the origin along the radial direction of the Poincaré disk as the circle radius increases. Prove that $Arc(r)$ increases exponentially with respect to r for any pairs of points p_1 and p_2.)

3. Hyperbolic space is a good embedding space for a tree-like graph $G = (V, E)$ with hierarchical structures. $d_{\mathcal{H}}(u, v)$ denotes the hyperbolic distance between any two nodes $u, v \in V$. The corrupted set of node u is defined as $S(u) = \{v|(u, v) \notin E\}$. Our goal is to ensure that connected nodes are close in the embedding space, and also push the nodes in $S(u)$ far away from u for any $u \in V$ in the embedding space. Similar to **Random Walk Optimization** introduced in Lecture 18, write out the expression of Loss function based on $d_{\mathcal{H}}(u, v)$ and the corrupted sample distances $d_{\mathcal{H}}(u, v')$ where $v' \in S(u)$.

   (**Hint**: Apply softmax function over distance $d_{\mathcal{H}}(u, v)$ to parameterize the likelihood of edge $(u, v) \in E$. The goal of embedding is to maximize the likelihood of each edge in E. You don't need to approximate the loss with negative sampling method in your answer.)

2 Discrete Denoising Diffusion Models for Graph Generation

Markov Chain. Define a Markov Chain of discrete random variables $\{X_t\}$, where $t = 0, 1, 2, \cdots$ indicate timestep. Each X_t denotes the present state at timestep t in the Markov Chain. Let $x_t \in \{1, 2, \cdots, K\}$ be an instantiation of the random variable X_t from its state space (i.e. $\{1, 2, \cdots, K\}$). Let $\boldsymbol{x}_t$ be the corresponding one-hot encoding of x_t. That is to say, $\boldsymbol{x}_t$ is a row vector where the i-th element is 1 while others are 0s when $x_t = i$.

The transition probability from X_{t-1} to X_t is given by the transition matrix $Q_t \in \mathbb{R}^{K \times K}$, where the entry at (i, j) is defined as

$$[Q_t]_{ij} := p(X_t = j | X_{t-1} = i) \tag{1}$$

, which denotes the transition probability from $X_{t-1} = i$ to $X_t = j$.

1. Suppose we have obrserved $X_{t-1} = x_{t-1}$, write out the expression $p(X_t|X_{t-1} = x_{t-1})$ as a row vector in terms of $\boldsymbol{x}_{t-1}$ and Q_t, where the j-th element means the transition probability from x_{t-1} to state j at timestep t.

2. Define $\bar{Q}_t = Q_1 Q_2 \cdots Q_t$. Suppose we have obrserved $X_0 = x_0$, write out the expression $p(X_t|X_0 = x_0)$ as a row vector in terms of $\boldsymbol{x}_0$ and $\bar{Q}_t$.

3. Suppose we have observed $X_t = x_t$ and $X_{t-1} = x_{t-1}$, write out the expression of $p(X_t = x_t|X_{t-1} = x_{t-1})$ as a scalar in terms of $\boldsymbol{x}_t$, $\boldsymbol{x}_{t-1}$ and Q_t.

4. Suppose we have observed $X_t = x_t$, write out the expression of $p(X_t = x_t|X_{t-1})$ as a column vector in terms of $\boldsymbol{x}_t$ and Q_t, where the i-th element means the transition probability from state i at timestep $t-1$ to x_t.

5. Write out the expression of $p(X_{t-1}|X_t = x_t, X_0 = x_0)$ in terms of all the known variables.
   Hint 1: Use Bayes' rule.
   Hint 2: You can use $\odot$ for element-wise multiplication.

6. Suppose $Q_t = (1-\beta)I + \beta \mathbb{1}^T \mathbb{1}/K$, where $\beta \in (0,1)$, I is the identity matrix, and $\mathbb{1}$ is an all-one row vector. What distribution does $\boldsymbol{x}_t$ follow when $t \to \infty$.
   Hint: Use mathematical induction. i.e. First show the equation of Q_t holds for $t = 1$, and then show the equation holds for $t = n+1$ if it holds for $t = n$. You will find the final distribution of x_t is independent of the initial state x_0.

7. Let G_0 be the initial state of an undirected graph with N nodes. We send G_0 to the Markov chain as described before to obtain $G_1, G_2, \cdots G_t$, where G_t is the state of the graph at timestep t. $G_t = (\{X_{i,t}\}, \{E_{ij,t}\}), 1 \le i < j \le N$, where $X_{i,t}$ denote the state of the i-th node at timestep t, and $E_{ij,t}$ denotes the edge between the i-th node and the j-th node at timestep t (note that i, j in this problem is the index of nodes rather than the index of states). There are K states for nodes and 2 states for edges (existence or non-existence).

   To simplify the problem, we make the following *i.i.d assumption*: we assume the Markov process is conducted on each node and each edge independently and identically. All nodes share the same set of transition matrices $\{Q_t^X : t = 1, 2, \cdots\}$ and all edges share the same set of transition matrices $\{Q_t^E : t = 1, 2, \cdots\}$.

   Suppose we want to reverse the process of the Markov chain, i.e. $p(G_{t-1}|G_t = g_t)$, where $g_t \in \{1, \cdots, K\}^N \times \{1, 2\}^{\frac{N(N-1)}{2}}$ is a realization of G_t. Due to the *i.i.d assumption*, we have

$$p(G_{t-1}|G_t = g_t) = \prod_i p(X_{i,t-1}|X_{i,t} = x_{i,t}) \prod_{i<j} p(E_{ij,t-1}|E_{ij,t} = e_{ij,t}), \tag{2}$$

Show that for the i-th node:

$$p(X_{i,t-1}|X_{i,t}=x_{i,t})=\sum_{k=1}^{K}\frac{\boldsymbol{x}_{i,t}Q_t^{X^T}\odot\mathbf{1}_k\bar{Q}_{t-1}^X}{\mathbf{1}_k\bar{Q}_t^X\boldsymbol{x}_{i,t}^T}p(X_{i,0}=k|X_{i,t}=x_{i,t}) \tag{3}$$

where $\mathbf{1}_k$ denotes the one-hot vector where the k-th element is 1. And expand $p(E_{ij,t-1}|E_{ij,t}=e_{ij,t})$ to write out the expression for $p(G_{t-1}|G_t=g_t)$.

After deriving the closed form of $p(G_{t-1}|G_t=g_t)$, you will find we can generate a graph by sampling the nodes and edges from discrete uniform distributions. We consider this as a graph at $t=T$ for a sufficiently large T, then iteratively denoise it until $t=0$, with the learnable components $p(X_{i,0}=k|X_{i,t}=x_{i,t})$ and $p(E_{ij,0}=k|E_{ij,t}=e_{ij,t})$.

Hint: Use the law of total probability with conditioning (e.g., $P(A|B)=\sum_i P(A|B,C_i)P(C_i)$), replace A,B,C with the correct probabilities.