



计算方法

刘景铖

计算机软件新技术国家重点实验室
南京大学



Random walk on graphs

给定图 $G = (V, E)$

图上的随机游走:

- 从一个给定的顶点出发
- 接下来，每一步都从当前顶点，移动到一个均匀随机选取的邻居
- 不断重复

这个随机过程的“长期表现”是怎么样的呢？

1. 重复 t 步之后，当前顶点是某个顶点 u 的概率是多少？
2. 是否存在一个极限的随机分布，随机游走会收敛到它？（稳态）
3. 多久才会收敛？(混合时间，mixing time)
4. 从点 s 出发，多久才会到达点 t ? (hitting time)
5. 多久才会遍历每个顶点至少一次？(遍历时间，cover time)



Random walk on graphs

这个随机过程的“长期表现”是怎么样的呢？

1. 重复 t 步之后，当前顶点是某个顶点 u 的概率是多少？
2. 是否存在一个极限的随机分布，随机游走会收敛到它？（稳态）
3. 多久才会收敛？(混合时间，mixing time)
4. 从点 s 出发，多久才会到达点 t ? (hitting time)
5. 多久才会遍历每个顶点至少一次？(遍历时间，cover time)

前3个问题可以通过纯概率的方法(coupling)来研究；我们将讨论如何利用与幂迭代的联系，通过马尔可夫链(Markov chains)转移矩阵的特征值来研究。

后2个问题我们将留到之后讨论“图与电阻电路网络”的时候



幂迭代与无向图上的随机游走

例子：考虑无向图3-cycle

从顶点1出发

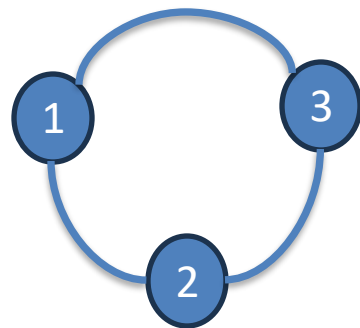
以概率 $\frac{1}{2}$ 走到顶点2

以概率 $\frac{1}{2}$ 走到顶点3

矩阵和向量的形式表示：

起始概率分布 $p_0 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$

随机游走一步之后，概率分布变成 $p_1 = \begin{pmatrix} 0 \\ 1/2 \\ 1/2 \end{pmatrix}$





幂迭代与无向图上的随机游走

令 A 为无向图的邻接矩阵, D 为度数的对角矩阵. (注: 因为是无向图, 所以 A 是对称的)

记 $p_t(v)$ 为时间 t 在点 v 的概率, 则有

$$p_{t+1}(v) = \sum_{u:uv \in E} p_t(u) \cdot \frac{1}{\deg(u)}$$

可以写出矩阵的形式: $\overrightarrow{p_{t+1}} = (AD^{-1})\overrightarrow{p_t}$, 进而有 $\overrightarrow{p_t} = (AD^{-1})^t \overrightarrow{p_0}$; $\overrightarrow{p_t}$ 为列向量

注意: Markov chain 里面习惯使用行向量; 我们进行谱分析, 更适合使用列向量的记号

转移矩阵 $W := AD^{-1}$

要研究 W^t , 只需要研究 $\mathcal{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$ 的幂迭代, 因为:

$$\mathcal{A} = D^{-\frac{1}{2}} W D^{\frac{1}{2}} \Rightarrow W^t = D^{\frac{1}{2}} \mathcal{A}^t D^{-\frac{1}{2}}$$

对于实数对称矩阵 $\mathcal{A} = V\Sigma V^T$ 的幂迭代:

$$\mathcal{A}^k = V\Sigma^k V^T = \sum_{1 \leq i \leq n} \lambda_i^k v_i v_i^T$$

幂迭代的收敛性? 对称矩阵 $\mathcal{A}$ 的特征值是怎样分布的? 下节课将讨论。



图的谱

在传统的算法课上，**无向图**的邻接矩阵往往只是作为一个数据结构

在接下来的几节课里面，会把邻接矩阵作为矩阵研究：特征值，行列式，秩等等，并把这些代数性质与大家习惯的组合性质相联系

这节课，首先给大家介绍随机游走与马尔可夫链的基本概念



Markov chains

考虑一般的有向图上的随机游走，记 X_t 为时间 t 随机游走所处的状态
转移矩阵 $P_{i,j} = \Pr[X_{t+1} = j | X_t = i]$

记 $p_t(i)$ 为时间 t 在状态 i 的概率. 那么, 对于所有 $0 \leq j \leq n-1$,

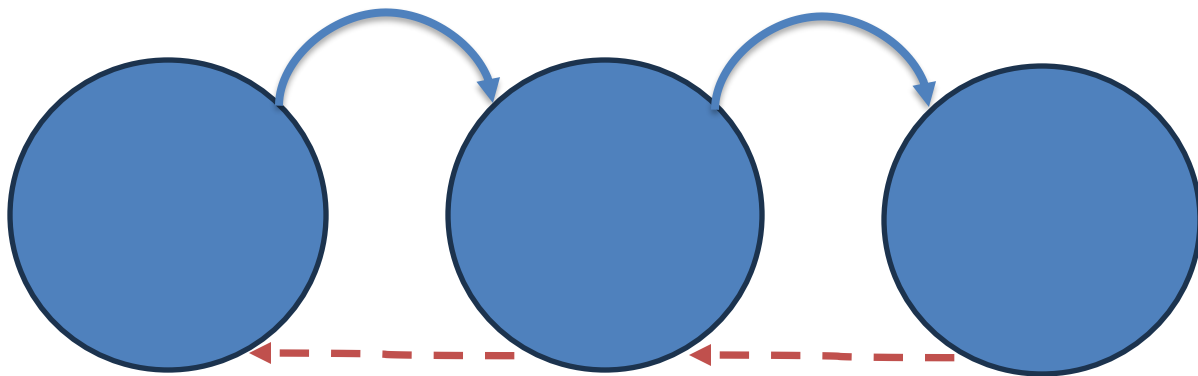
- $p_{t+1}(j) = \sum_{i=0}^{n-1} p_t(i) \cdot P_{i,j}$
- 更一般地, $\overrightarrow{p_{t+1}} = \overrightarrow{p_t} \cdot P$, 并且 $\overrightarrow{p_t} = \overrightarrow{p_0} \cdot P^t$. 这里 $\overrightarrow{p_t}$ 为行向量

注意: 转移只与时间 t 在各个状态的概率相关, 与过去无关



Irreducible Markov chains

对于有限的马尔可夫链，如果它对应的有向图是强连通的，则称它为不可约的(irreducible)。





Aperiodic Markov chains

给定一个马尔可夫链，其中状态 i 的周期定义为

$$period(i) := \gcd\{t \mid P_{i,i}^t > 0\}$$

- 如果一个状态 i 的周期 $period(i) = 1$ ，则称它为非周期性的(aperiodic)
- **如果**马尔可夫链中每一个状态都是非周期性的话，则称该马尔可夫链是非周期性的；否则称它是周期性的。

例子：二分图，3-cycle，self-loops



Irreducible & aperiodic Markov chains

引理：如果一个马尔可夫链是不可约，且非周期的话，则一定存在足够大的常数 $T < \infty$ ，使得

$$(P^t)_{i,j} > 0, \forall i, j, \forall t \geq T$$

证明思路：首先证明 $(P^t)_{i,i} > 0$ ，之后使用不可约的性质。



稳态分布 (stationary distribution)

如果分布 $\vec{\pi}$ 满足 $\vec{\pi} = \vec{\pi} \cdot P$, 则称 $\vec{\pi}$ 为稳态分布

稳态分布是一个 “平衡态” / 不动点

特别地: $\vec{\pi} = \vec{\pi} \cdot P^t, \forall t$

注意: 如果极限分布存在, 则一定是一个稳态分布



概率分布的距离度量与收敛

要研究收敛，需要一个距离的度量；这里我们使用的是TV距离(total variation distance)。给定概率分布 $\vec{p}$ 与 $\vec{q}$ ，它们的TV距离定义为

$$d_{TV}(\vec{p}, \vec{q}) := \frac{1}{2} \|p - q\|_1 = \frac{1}{2} \sum_{i=1}^n |p_i - q_i|$$

注意 $0 \leq d_{TV}(\vec{p}, \vec{q}) \leq 1$

如果 $\lim_{t \rightarrow \infty} d_{TV}(\vec{p}_t, \vec{q}) = 0$ ，则说 $\{\vec{p}_t\}$ 收敛到 $\vec{q}$

(选讲) 也可以选用2范数

$$\left\| \frac{p}{\pi} - 1 \right\|_{2,\pi}^2 := \sum_{i=1}^n \pi_i \left(\frac{p_i}{\pi_i} - 1 \right)^2$$

类似于“方差”

由Cauchy-Schwarz不等式,

$$\sum_{i=1}^n |p_i - \pi_i| = \sum_{i=1}^n \pi_i \left| \frac{p_i}{\pi_i} - 1 \right| \leq \sqrt{\sum_{i=1}^n \pi_i \left| \frac{p_i}{\pi_i} - 1 \right|^2} = \left\| \frac{p}{\pi} - 1 \right\|_{2,\pi}$$

这意味着，如果2-norm 是小的, 则1-norm 也是小的 (但反过来不一定).



回归时间 (Return time)

从状态 i 出发，第一次返回到状态 i 的时间，称之为状态 i 的回归时间，定义为

$$H_i := \min\{t \geq 1 \mid X_t = i, X_0 = i\}$$

期望回归时间定义为 $h_i := E[H_i]$.



Fundamental Theorem of Markov Chains

马尔可夫链的基本定理

对于有限，不可约，非周期的马尔可夫链，有以下事实：

1. 存在一个稳态分布 $\vec{\pi}$.
2. 随着 $t \rightarrow \infty$, $\vec{p}_t$ 都会收敛到 $\vec{\pi}$, 无论从什么样的 $\vec{p}_0$ 开始.
3. 稳态分布是唯一的
4. $\pi(i) = \frac{1}{h_i}$



Intuition

马尔可夫链的基本定理的一些直观理解

对于有限，不可约，非周期的马尔可夫链，有以下事实：

1. 存在一个稳态分布 $\vec{\pi}$.
2. 随着 $t \rightarrow \infty$, $\vec{p}_t$ 都会收敛到 $\vec{\pi}$, 无论从什么样的 $\vec{p}_0$ 开始.
3. 稳态分布是唯一的
4. $\pi(i) = \frac{1}{h_i}$

考虑两个随机游走，当它们相遇之后，就变得“无法区别”了

回忆之前的引理：存在足够大的常数 T ，使得 $(P^t)_{i,j} > 0 \forall i, j, \forall t \geq T$

这意味着无论它们现在在何方，经过一段时间之后，都有非零的概率相遇

最终会相遇的概率是1，此时将收敛到同一个稳态分布

通过概率的coupling argument，可以把上述想法变成严格的证明

$\pi(i) = \frac{1}{h_i}$ ：期望回归时间为 h_i



Pagerank

考虑一个描述网页超链接的有向图

点集：网页

边集：如果网页 i 到网页 j 存在超链接，则 $i \rightarrow j$

搜索引擎想要给网页一个关于其内容质量/重要程度的打分

Pagerank的想法：

- 如果一个页面被很多人转载/引用，它是比较重要的；
 - 衡量标准：存在超链接指向它；
- 如果一个页面被其它重要的页面所指向，它本身可能也是比较重要的；
 - 衡量标准：？？？



Pagerank

一个迭代算法:

- 一开始每个页面都有pagerank分数为 $1/n$
- **接下来的每一步**, 每个页面都会把自己的分数平均分开, 并发送给它所指向的邻居
- 每个页面会更新自己的分数为, 它所收到的分数的和
- 不断重复直到收敛



Pagerank

Pagerank分数收敛到的平衡态，正是随机游走的稳态分布。

$$Pagerank_{t+1}(j) = \sum_{i:ij \in E} Pagerank_t(i) / \deg_{out}(i)$$

$$\text{令 } P_{i,j} = \begin{cases} \frac{1}{\deg_{out}(i)}, & \text{if } ij \in E \\ 0, & \text{otherwise} \end{cases}, \text{ 则有 } \overrightarrow{Pagerank_{t+1}} = \overrightarrow{Pagerank_t} \cdot P$$

Pagerank满足与随机游走相同的迭代关系

- 当图是有限，不可约（强连通），非周期的时候，Pagerank分数是唯一的。（马尔可夫链的基本定理）
- Pagerank分数只与图的结构有关，与初始值无关
- Pagerank分数将会是随机游走期望回归时间的倒数
- 可以通过有向图上的随机游走模拟



Pagerank

实践中的图不一定是不可约（强连通），非周期的。

可以作以下修改：固定 ϵ , 把 $1 - \epsilon$ 的pagerank分数分给邻居， ϵ 的pagerank分数均匀分给所有点

对应的随机游走则是：

以概率 $1 - \epsilon$ ，走到一个均匀随机的邻居；

以概率 ϵ ，走到一个均匀随机的顶点。

可以证明，这个修改将保证图是强连通且非周期的
而且不会改变网页的相对分数



Perfect Matchings in d -regular Bipartite Graphs (选讲)

Random walks can also be used to design faster algorithms.

Perfect Matchings in $O(n \log n)$ Time in Regular Bipartite Graphs, by Goel, Kapralov, Khanna

- Traditional: find augmenting paths by e.g. BFS/DFS
- Idea: replace BFS/DFS by a random walk (on a different Eulerian directed graph, i.e. indegree=outdegree for every vertex)
- Expected time to find an augmenting path $\Rightarrow$ expected return time of the random walk $\Rightarrow$ stationary distribution value
- Open research problem: Can you extend this to non-regular bipartite graphs?



随机采样

随机游走的应用之一，在于进行随机采样

- 扑克牌中，如何进行洗牌？
 - 52张牌
 - 把每个排列顺序看成图的一个节点
 - 每次洗牌看成一次随机游走
 - 是否收敛？
 - 多少步收敛？
 - 7次“Rifle” shuffle, New York Times
- 随机的生成树 (Random spanning tree)
 - Basis exchange walk
- 随机的完美匹配，近似计数等等



无向图上随机游走的稳态分布

如果分布 $\vec{\pi}$ 满足 $\vec{\pi} = W\vec{\pi}$, 则称 $\vec{\pi}$ 为稳态分布

稳态分布是一个“平衡态” / 不动点

特别地: $\vec{\pi} = W^t \vec{\pi}, \forall t$

注意: 如果极限分布存在, 则一定是一个稳态分布

给定无向图 G , 令 $\vec{d} \in \mathbb{R}^n$ 为顶点度数的向量, $m = |E|$

定理. 概率分布 $\vec{\pi} = \frac{\vec{d}}{2m}$ 是 G 上随机游走的一个稳态分布.

换言之, 向量 $\vec{\pi} = \frac{\vec{d}}{2m}$ 是 AD^{-1} 的一个右特征向量, 且特征值为1



无向图上的马尔可夫链基本定理

是否不管从任意 p_0 出发，都有 $p_t \rightarrow \vec{\pi} = \frac{\vec{d}}{2m}$?

不一定: 如果Markov chain是reducible（不连通），或者周期性的（周期不为1）

对于无向的连通图, 非周期性等价于非二分图

(To see why, recall what happens to an undirected 3-cycle)

因此无向图上的马尔可夫链基本定理如下:

定理. 对任意有限的，连通的，非二分图，都有 $p_t \rightarrow \vec{\pi} = \frac{\vec{d}}{2m}$ 随着 $t \rightarrow \infty$ ，与 p_0 的选取无关



Lazy Random Walks

对于二分图，可以考虑惰性随机游走：

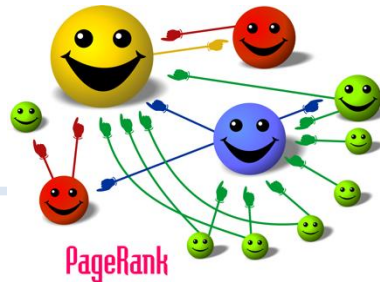
- 每一步会以1/2的概率停留在当前状态；
- 以1/2的概率，随机地从邻居里面均匀地选取一个，作为下一步的状态。

用矩阵表示的话则有 $p_t = \left(\frac{1}{2}I + \frac{1}{2}AD^{-1}\right)^t p_0$.

定理. 对于任意有限的，连通的，不管从什么样的 p_0 开始， $p_t = \left(\frac{1}{2}I + \frac{1}{2}AD^{-1}\right)^t p_0 \rightarrow \frac{\vec{d}}{2m}$



谱图理论 (Spectral graph theory)



谱分析: 特征值 + 特征向量
+ 相关的线性代数

图论与组合结构:

- 连通性 (Cheeger不等式)
- 图染色
- 聚类(Clustering)
- Mixing of random walks
- Expander graphs (efficient network, superconcentrators)

计算机科学:

- Pagerank
- 稀疏化Sparsification
- 迭代法解线性方程
- 电阻网络
- Expander codes (LDPC, Tanner codes)
- 不可近似性(Dinur's proof of the PCP theorem)
- 去随机化 (Derandomization)