不考

STAT0008: Stochastic Processes

Lecture 8 - Mixing Times

Lecturer: Weichen Zhao　　Spring 2025

Key concepts:

- 全变差距离；
- *Dobrushin*准则；
- 混合时间。

本节我们讨论Markov链收敛到不变分布的收敛速度，这在随机算法的分析中尤为重要。本节我们只考虑有限状态空间$E = \{1, 2, \ldots, n\}$。

8.1 全变差距离

记有限状态空间E上概率分布的全体为$\mathcal{M}$, 那么转移矩阵$\mathbf{P}$诱导出$\mathcal{M}$上的一个变换

$$P : \mathcal{M} \to \mathcal{M}, \quad \nu \mapsto \nu P.$$

不变分布就是该变换的不动点。

在分析收敛性之前，我们需要一个刻画分布之间距离的数学量。

Definition 8.1 (全变差距离) 分布$\mu, \nu \in \mathcal{M}$的全变差距离*(total variation distance)* 定义为

$$\|\mu - \nu\|_{TV} = d_{TV}(\mu, \nu) = \max_{A \subset E} |\mu(A) - \nu(A)|.$$

这个定义具有明确的概率含义：μ和ν之间的距离是由两个分布在单个事件的概率之间的最大差异。

注. 可以验证d_{TV}是$\mathcal{M}$上的度量，即$(\mathcal{M}, d_{TV})$是一个度量空间。

下面我们不加证明的列出一些命题，详细证明请参考：Levin D A, Peres Y. Markov chains and mixing times[M]. American Mathematical Soc., 2017. Section 4.1

Proposition 8.2 设$\mu, \nu \in \mathcal{M}$，则

$$d_{TV}(\mu,\nu) = \frac{1}{2}\sum_{i\in E}|\mu_i - \nu_i| = \sum_{i\in E, \mu_i \geq \nu_i}(\mu_i - \nu_i).$$

Proposition 8.3 设$\mu, \nu \in \mathcal{M}$，则

$$d_{TV}(\mu,\nu) = \frac{1}{2}\sup\left\{\sum_{i\in E} f(i)\mu_i - \sum_{i\in E} f(i)\nu_i : f\text{满足}\max_{i\in E}|f(i)| \leq 1\right\}.$$

Proposition 8.4 设$\mu, \nu \in \mathcal{M}$，则

$$d_{TV}(\mu,\nu) = \inf\left\{P\{X \neq Y\} : (X,Y)\text{是}\mu\text{和}\nu\text{的耦合}\right\}.$$

Theorem 8.5 (收敛定理) 设P不可约非周期，具有不变分布π，那么存在一个常数$\alpha \in (0,1)$以及$C > 0$，使得

$$\max_{i\in E}\left\|P_{i,\cdot}^k - \pi\right\|_{\mathrm{TV}} \leq C\alpha^k.$$

指数压缩

Proof: 证明请参考：Levin D A, Peres Y. Markov chains and mixing times[M]. American Mathematical Soc., 2017. Theorem 4.9 ■

注. 由强遍历定理，我们知道

$$\lim_{k\to\infty} P_{ij}^k = \pi_j$$

但这只是一个渐近结果，定理8.5给出了一个非渐近的结果，可以看出$P_{i,\cdot}^k$会以α为参数指数压缩到π。之后我们能够给出来这里常数的具体形式。

8.2 Dobrushin 准则

设不可约非周期的Markov链的初分布为μ，不变分布为π，由强遍历定理，

$$\lim_{k\to\infty} d_{TV}(\mu P^k, \pi) = 0$$

本节讨论$d_{TV}(\mu P^k, \pi)$趋于0的速度。

对于非周期，这个容易达到

Proposition 8.6 (Dobrushin 准则) 若转移矩阵P满足$P_{ij} > 0$，$\forall i, j$，则存在常数$0 \leqslant \alpha < 1$，对任意$\mu, \nu \in \mathcal{M}$，

两个初分布

$$d_{TV}(\mu P, \nu P) \leqslant \alpha d_{TV}(\mu, \nu)$$

特别地，若π是P的不变分布，则

$$d_{TV}(\nu P^k, \pi) \leqslant \alpha^k, \quad \forall k \geq 0$$

Proof: 证明请参考：《应用随机过程》，陈大岳、章复熹，北京大学出版社，2023，命题1.10.1 ■

注意到Dobrushin 准则里要求$P_{ij} > 0$，$\forall i, j$，其实这个条件是可以放松的，有如下命题：

Proposition 8.7 设P不可约非周期，π为P的不变分布，则存在常数 $C > 0, \beta > 0$，使得对任意$\mu \in \mathcal{M}$，

$$d_{TV}(\mu P^k, \pi) \leqslant C\mathrm{e}^{-\beta k}, \forall\, k \geqslant 0.$$

Proof: 证明请参考：《应用随机过程》，陈大岳、章复熹，北京大学出版社，2023，命题1.10.2 ■

命题8.7表明：从任何初分布出发，Markov链的状态分布都将收敛到不变分布，而且收敛速度是指数阶的。

8.3 混合时间

Markov链的状态分布收敛到不变分布的行为称为混合(Mixing)，本节我们引入一个数学量——混合时间，来刻画Markov链离平稳状态的距离。本节假设Markov链是可逆的。为了不引起记号混乱，本节将Markov链的状态记到圆括号( )中，例如$P(x,y) = P_{xy}$，$\pi(x) = \pi_x$，$x, y \in E$. 将时间参数记为t.

Definition 8.8 (混合时间) *设不可约非周期Markov链*$\{X_k\}$*的转移矩阵为*P*，不变分布为*π*，定义它的混合时间(Mixing Time)为*

$$t_{\min}(\varepsilon) \triangleq \min\{t : d(t) \le \varepsilon\}$$

其中$d(t) \triangleq \sup_{\mu\in\mathcal{M}} d_{TV}(\mu P^t, \pi)$.

矩阵的特征值的集合称为谱(Spectral)，接下来我们将利用谱方法(Spectral Methods)给出不可约非周期可逆Markov链混合的结果。下面列出一些关于可逆转移矩阵P的谱的命题。

Lemma 8.9 *设*P*是关于*π*可逆的转移矩阵，则*

P 1 = 1·1

(1) P*的所有特征值*λ*满足*$|\lambda| \le 1$;

(2) *若*P*不可约，我们对*P*的特征值排序：*$1 \ge \lambda_1 \ge \cdots \ge \lambda_n \ge -1$*，那么*$\lambda_1 = 1$*，且*$\lambda_2 < 1$. *我们还能得到特征值*$\lambda_1$*对应的特征空间是由全一特征向量*$(1, 1, \ldots, 1)$*张成的一维空间;*

(3) *若*P*非周期，*-1*不是*P*的特征值;* 正负-来回变换

(4) *记*λ_j*对应的特征函数(向量)为*f_j*，*$j = 1, 2, \ldots, n$*，那么对所有*$j \ne 1$*，有*

$$\pi(f_j) \triangleq \sum_{x\in E} f_j(x)\pi(x) = 0$$

(5) *对任意*x, y

$$\sum_{j=1}^{n} f_j(x) f_j(y) = \pi(x)^{-1}\delta_x(y),$$

其中$\delta_x(y) := 1_{\{x=y\}}$.

Proof: 证明涉及泛函分析Hilbert空间的相关基础，有兴趣的同学请参考 Roch S. Modern discrete probability: An essential toolkit[M]. Cambridge University Press, 2024. Section 5.2.1 ■

Theorem 8.10 (P^t**的谱分解**) 设P是不可约的关于π可逆的转移矩阵，$\{\lambda_j\}_{j=1}^n$和$\{f_j\}_{j=1}^n$分别为相应的特征值和特征向量，设$1 \geq \lambda_1 \geq \cdots \geq \lambda_n \geq -1$，那么

$P^t(x,y) = P^t_{xy}$

$$\frac{P^t(x,y)}{\pi(y)} = \sum_{j=1}^{n} f_j(x) f_j(y) \lambda_j^t = 1 + \sum_{j=2}^{n} f_j(x) f_j(y) \lambda_j^t.$$

由于1对于特征向量为 $\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix}$ 即 $f_1(x)=f_1(y)=1$

Proof: 证明参考Levin D A, Peres Y. Markov chains and mixing times[M]. American Mathematical Soc., 2017. Lemma 12.2 ■

λ_2

从P^t的谱分解可以看出，$P^t(x,y)$收敛到$\pi(y)$的速度主要由不是1的最大的特征值控制，所以自然的，我们关注下面叫做谱隙(Spectral gap)的数学量。

Definition 8.11 (谱隙) 设转移矩阵的特征值为$\lambda_1 \geq \cdots \geq \lambda_n$，绝对谱隙*(absolute spectral gap)*定义为

$$\gamma_* \triangleq 1 - \lambda_*$$

其中$\lambda_* = \max\{|\lambda_2|, |\lambda_n|\}$，谱隙*(Spectral gap)*定义为

$$\gamma \triangleq 1 - \lambda_2$$

下面通过谱隙我们可以给出关于混合时间的估计

谱隙越大收敛越快

Theorem 8.12 设P不可约非周期，关于π可逆，记$\pi_{min} = \min_{x \in E} \pi(x)$, 那么对于任意$\varepsilon > 0$，

$$(\gamma_*^{-1} - 1) \log\left(\frac{1}{2\varepsilon}\right) \leq t_{\text{mix}}(\varepsilon) \leq \log\left(\frac{1}{\varepsilon \pi_{min}}\right) \gamma_*^{-1}.$$

Proof: 证明参考 Roch S. Modern discrete probability: An essential toolkit[M]. Cambridge University Press, 2024. Theorem 5.2.14 ■

从定理8.12可以看出P的绝对谱隙γ_*越大，则Markov链混合时间的上界越小，即Markov链收敛到平稳分布的速度越快。

下面我们给一个谱方法分析MCMC算法的量化结果。

在高维空间中，例如Hypercube $E \triangleq \{-1,1\}^n$ → 2^n个，计算

$$\pi(f) = \sum_{x\in E} \pi(x)f(x), \quad f: E \to \mathbb{R}$$

$= E(f(x))$

是非常复杂的，例如统计物理中配分函数的计算。一种主要的计算方法是Monte Carlo方法：从分布π中采样$X_1, X_2, \ldots, X_T$，用经验和近似期望和

$$\frac{1}{T}\sum_{t=1}^{T} f(X_t) \approx \pi(f)$$

当π比较复杂时，不容易通过简单方法获得它的样本，这时可以构造一个Markov链$\{X_k\}$，使得它的不变分布为π，例如Metropolis链，那么由平均遍历定理6.11，随着$T\to\infty$

$$\frac{1}{T}\sum_{t=1}^{T} f(X_t) \to \pi(f)$$

这种方法即是Markov Chain Monte Carlo方法，但是遍历定理的结果是渐近的，在算法的分析中，我们希望能够显式量化MCMC方法估计$\pi(f)$的误差与迭代时间T的关系。

需要满足细致平衡条件

Theorem 8.13 设P不可约非周期，关于π可逆，那么对于任何$\varepsilon > 0$，

$$P\left[\left|\frac{1}{T}\sum_{t=1}^{T} f(X_t) - \pi(f)\right| < \varepsilon\right] \geq 1 - \frac{9\pi_{\min}^{-1}\|f\|_\infty^2 \gamma_*^{-1}\frac{1}{T}}{(\varepsilon - \pi_{\min}^{-1}\|f\|_\infty \gamma_*^{-1}\frac{1}{T})^2},$$

其中$\|f\|_\infty \triangleq \sup_{x\in E}|f(x)|$为无穷范数。

Proof: 证明参考 Roch S. Modern discrete probability: An essential toolkit[M]. Cambridge University Press, 2024. Theorem 5.2.27 ∎