Applications of Nonstandard Analysis to Markov Processes

Haosui Duanmu
Department of Economics
UC Berkeley

September 18, 2018

List of Contents

- Basic Concepts for Markov Processes.
- Mixing Times and Hitting Times for Markov Processes.
- Introduction to Nonstandard Analysis.
- Hyperfinite Probability Spaces.
- Asymptotical Equivalence between Mixing Times and Hitting Times.
- Application to Metropolis-Hasting Algorithm.

Introduction to Markov Process

A discrete-time Markov process $\{X_t\}_{t\in\mathbb{N}}$ is characterized by the following ingredients:

1. A σ-compact metric state space X with Borel σ-algebra $\mathcal{B}[X]$.
2. Transition kernel given by $\{g(x,1,\cdot)\}_{x\in X}$.

For every $x \in X$, $g(x,1,\cdot)$ is a probability measure on $(X,\mathcal{B}[X])$. We define higher-order transition probabilities inductively by

1. For every $x \in X$, $A \in \mathcal{B}[X]$ and every $n \geq 1$, $g(x,n+1,A) = \int_X g(y,n,A)g(x,1,\mathrm{d}y)$.

Definition

A stationary distribution $\pi(\cdot)$ is a probability measure on $(X,\mathcal{B}[X])$ such that $\pi(A) = \int_X g(x,1,A)\pi(\mathrm{d}x)$ for all $A \in \mathcal{B}[X]$.

Concepts in Markov Processes

The transition kernel g can be viewed as a function from $X \times \mathbb{N} \times \mathcal{B}[X]$ to $[0,1]$.

Definition

A transition kernel $\{g(x,1,\cdot)\}_{x\in X}$ with stationary distribution π is reversible if $\int_A g(x,1,B)\pi(\mathrm{d}x) = \int_B g(x,1,A)\pi(\mathrm{d}x)$ for every $A, B \in \mathcal{B}[X]$.

Definition

Given a transition kernel $\{g(x,1,\cdot)\}_{x\in X}$, its lazy transition kernel $\{g_L(x,1,\cdot)\}_{x\in X}$ is

$$g_L(x,1,\cdot) = \frac{1}{2}g(x,1,\cdot) + \frac{1}{2}\delta_x(\cdot),$$

where $\delta_x(\cdot)$ is the point mass at x.

Mixing Times for Markov Processes

Let $\{g(x,1,\cdot)\}_{x\in X}$ be a transition kernel with stationary distribution π.

Definition

For $\epsilon > 0$, the mixing time with respect to ϵ is

$$t_m(\epsilon) = \min\{t \in \mathbb{N} : \sup_{x\in X} \parallel g(x,t,\cdot) - \pi(\cdot) \parallel \leq \epsilon\}.$$

The lazy mixing time with respect to ϵ is

$$t_L(\epsilon) = \min\{t \in \mathbb{N} : \sup_{x\in X} \parallel g_L(x,t,\cdot) - \pi(\cdot) \parallel \leq \epsilon\}.$$

The average mixing time with respect to ϵ is

$$t_a(\epsilon) = \min\{t \in \mathbb{N} : \sup_{x\in X} \parallel \frac{g(x,t,\cdot) + g(x,t+1,\cdot)}{2} - \pi(\cdot) \parallel \leq \epsilon\}.$$

Hitting Times for Markov Processes

The maximal hitting time measures the time it takes for a Markov process to get into "big" sets.

Definition

Let $0 < \alpha \leq 1$. The maximum hitting time with respect to α is

$$t_H(\alpha) = \sup\{\mathbb{E}_x(\tau_A) : x \in X, A \in \mathcal{B}[X] \text{ such that } \pi(A) \geq \alpha\}$$

where $\tau_A = \min\{t \in \mathbb{N} : X_t \in A\}$.

Alternatively, one can consider the large hitting time with respect to α is $\tau_g(\alpha)$

$$\min\{t \in \mathbb{N} : \inf\{\mathbb{P}_x(\tau_A \leq t) : x \in X, A \in \mathcal{B}[X] \ \wedge \pi(A) \geq \alpha\} > 0.9\} \tag{1}$$

For every $0 < \alpha \leq 1$, we have $0.1\tau_g(\alpha) \leq t_H(\alpha) \leq 2\tau_g(\alpha)$.

Mixing Times and Hitting Times

We are often interested in mixing times but they are hard to compute directly. Hitting times are easier to compute.

Theorem (Y.Peres and P.Sousi; R.Oliveira)

Let $0 < \alpha < \frac{1}{2}$. Then there exist universal constants c_α, c'_α so that for every **finite** *reversible Markov process*

$$c'_\alpha t_H(\alpha) \leq t_L \leq c_\alpha t_H(\alpha).$$

Theorem (Y.Peres and P.Sousi)

For every $0 < \epsilon \leq \frac{1}{4}$. There exist universal constants c_ϵ and c'_ϵ so that for every **finite** *reversible Markov process*

$$c'_\epsilon t_L(\epsilon) \leq t_a(\epsilon) \leq c_\epsilon t_L(\epsilon).$$

Strong Feller Condition

Definition

A transition kernel $\{g(x,1,\cdot)\}_{x\in X}$ satisfies the *Strong Feller* condition if for every $x \in X$ and every $\epsilon > 0$ there exists $\delta > 0$ such that

$$(\forall y \in X)(|y-x| < \delta \implies (\| g(x,1,\cdot) - g(y,1,\cdot) \| < \epsilon))$$

Let $\mathcal{C}$ be the collection of discrete-time reversible transition kernels with compact metric state space satisfying the Strong Feller condition.

Main Results on Compact Spaces

We have asymptotical equivalence between lazy mixing times and hitting times.

Theorem (R.Anderson, H.Duanmu and A.Smith)

Let $0 < \alpha < \frac{1}{2}$. *Then there exist universal constants* $0 < a_\alpha, a'_\alpha < \infty$ *such that, for every* $\{g(x,1,\cdot)\} \in \mathcal{C}$, *we have*

$$a'_\alpha t_H(\alpha) \leq t_L \leq a_\alpha t_H(\alpha).$$

We have asymptotical equivalence between lazy mixing times and average mixing times.

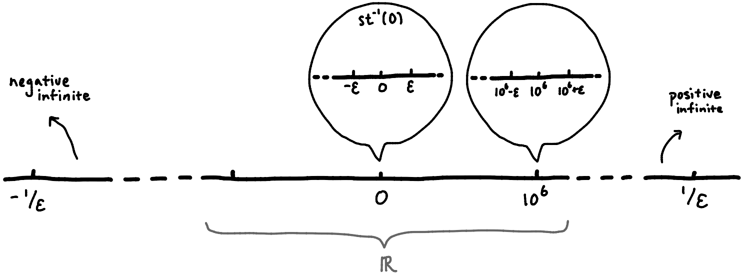
Theorem (R.Anderson, H.Duanmu and A.Smith)

For $0 < \epsilon \leq \frac{1}{4}$, *Then there exist universal constants* $0 < d_\epsilon, d'_\epsilon < \infty$ *such that for every* $\{g(x,1,\cdot)\} \in \mathcal{C}$

$$d'_\epsilon t_L(\epsilon) \leq t_a(\epsilon) \leq d_\epsilon t_L(\epsilon).$$

$$
\begin{tikzcd}
\text{nonstandard structures} \arrow[r] & \text{nonstandard theorem} \arrow[d, "\text{push down}"] \\
\text{standard structures} \arrow[u, "\text{push up}"] & \text{standard theorem}
\end{tikzcd}
$$

Nonstandard real line



1. **Transfer Principle**:
   $\mathbb{R} \models \phi(x_1, ..., x_n) \Leftrightarrow {}^*\mathbb{R} \models {}^*\phi({}^*x_1, ..., {}^*x_n)$.
2. κ-**Saturation Principle**: Let $\mathcal{F}$ be a family of internal sets with cardinality less than κ. If any finite intersection of elements in $\mathcal{F}$ is nonempty, then the total intersection of $\mathcal{F}$ is non-empty.

A subset of ${}^*\mathbb{R}$ is internal if it can be described using logic formulas. We can define extension $*$ for any space, not just for $\mathbb{R}$.

Infinite and infinitesimal numbers

Lemma

There is a $k \in {}^*\mathbb{R}_{>0}$ such that $k \geq n$ for all $n \in \mathbb{N}$.

Proof.

For $n \in \mathbb{N}$, let $A_n = \{k \in {}^*\mathbb{R}_{>0} : k > n\}$. Note that every A_n is internal and $A_n \supset A_{n+1}$. By saturation, $\bigcap_{n \in \mathbb{N}} A_n$ is non-empty. $\square$

Definition

- $x \in {}^*\mathbb{R}$ is **infinite** if $|x| > n$ for all $n \in \mathbb{N}$.
- $x \in {}^*\mathbb{R}$ is **infinitesimal** if $\frac{1}{x}$ is infinite.
- $x, y \in {}^*\mathbb{R}$ are **infinitely close**, written $x \approx y$, if $|x - y|$ is infinitesimal.

These notions can be generalized to arbitrary metric spaces.

Standard Part Map

Definition

An element $x \in {}^*X$ is near-standard if there is an element $y \in X$ such that ${}^*d(x, y) \approx 0$. The point y is called the standard part of x and is denoted by $\mathrm{st}(x)$.

Example

Every finite element in ${}^*\mathbb{R}$ has standard part. Infinite elements in ${}^*\mathbb{R}$ do not have standard part. Infinitesimals do not have standard part in ${}^*(0, 1)$.

A set $A \subset X$ is usually closely connected to $\mathrm{st}^{-1}(A) = \{x \in {}^*X | (\exists y \in A)(x \approx y)\}$.

Hyperfinite Probability Space

Definition

A set A is **hyperfinite** if and only if there exists an internal bijection between A and $\{0, 1, ..., N-1\}$ for some $N \in {}^*\mathbb{N}$. This N is unique and is called the internal cardinality of A.

A hyperfinite probability space is a triple $(\Omega, I(\Omega), P)$ such that

1. Ω is a hyperfinite set.
2. $I(\Omega)$ is the collection of all hyperfinite subsets of Ω.
3. $P : I(\Omega) \to {}^*[0, 1]$ such that $P(\emptyset) = 0$, $P(\Omega) = 1$ and P is hyperfinitely additive.

Theorem (Peter A.Loeb)

Given a hyperfinite probability space $(\Omega, I(\Omega), P)$, *we can extend it to a standard countably additive probability space* $(\Omega, \overline{I(\Omega)}, \overline{P})$.

$(\Omega, \overline{I(\Omega)}, \overline{P})$ is called the **Loeb space** of $(\Omega, I(\Omega), P)$.

Hyperfinite Representation of Borel Probability Space

Theorem (Robert Anderson)

Let $(X, \mathcal{B}[X], \mu)$ be a Borel probability space. Then there is a hyperfinite probability space $(S_X, I(S_X), P)$ with $S_X \subset {}^*X$ such that

$$\overline{P}(\mathrm{st}^{-1}(E) \cap S_X) = \mu(E)$$

for all $E \in \mathcal{B}[X]$. $(S_X, I(S_X), P)$ is called the hyperfinite representation of $(X, \mathcal{B}[X], \mu)$.

Example

Let $T = \{\frac{1}{N}, \frac{2}{N}, \dots, 1\}$ for some $N \in {}^*\mathbb{N} \setminus \mathbb{N}$. Let $P(\{\omega\}) = \frac{1}{N}$ for every $\omega \in T$. Then the hyperfinite probability space $(T, I(T), P)$ is a hyperfinite representation of the Lebesgue measure on $[0, 1]$.

Hyperfinite Markov chain

The hyperfinite time line $T = \{0, 1, 2, \ldots, K\}$ for some infinite number K. The hyperfinite Markov chain with hyerfinite time line T is characterized by 3 ingredients:

1. A hyperfinite set $S \subset {}^*X$.
2. An initial internal probability measure on S.
3. An internal "one-step" transition probability $\{G_i(j)\}_{i,j \in S}$ such that $\sum_{j \in S} G_i(j) = 1$.

Definition

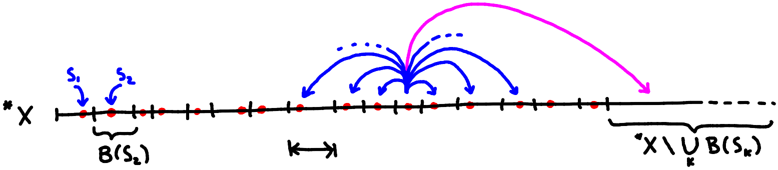
A distribution Π on state space $(S, I(S))$ is said to be a ***stationary distribution** if, for any $A \in I(S)$, we have $\Pi(A) = \sum_{s \in S} \Pi(s) G_s(A)$ where $G_s(A) = \sum_{j \in A} G_s(j)$.

Three Main Steps

$$\begin{tikzcd} \text{nonstandard structures} \arrow[r] & \text{nonstandard theorem} \arrow[d, "\text{push down}"] \\ \text{standard structures} \arrow[u, "\text{push up}"] & \text{standard theorem} \end{tikzcd}$$

1. Construct a hyperfinite Markov chain $\{X'_t\}$ which is closely related to $\{X_t\}$.
2. Mixing times, hitting times and average mixing times of $\{X'_t\}$ are asymptotically equivalent.
3. Mixing times, hitting times and average mixing times of $\{X_t\}$ are asymptotically equivalent.

Construction of $\{X'_t\}_{t \in T}$ from $\{X_t\}_{t \geq 0}$



1. Cut ${}^*X$ (extension of the state space) into hyperfinitely pieces of infinitesimal radius.
2. Pick a point from each piece in order to form the hyperfinite state space S_X for $\{X'_t\}_{t \in T}$.
3. The one-step transition probability for $\{X'_t\}_{t \in T}$ is $G(s_i, s_j) = \frac{\int_{B(i)} {}^*g(x,1,{}^*B(j))^*\pi(\mathrm{d}x)}{{}^*\pi(B(i))}$ for $s_i, s_j \in S_X$.

Relationship with the standard chain

The internal transition kernel $\{G(s,\cdot)\}_{s\in S_X}$ is closely related to the standard transition kernel $\{g(x,1,\cdot)\}_{x\in X}$. Let $\pi'(\{s\}) = {}^*\pi(B(s))$ for every $s \in S_X$.

1. For every $A \in \mathcal{B}[X]$, every $s \in S$ and every $t \in \mathbb{N}$, we have $\overline{G}_s^{(t)}(\mathrm{st}^{-1}(A) \cap S_X) = g(\mathrm{st}(s), t, A)$.
2. π' is a *stationary distribution of $\{G(s,\cdot)\}_{s\in S_X}$.
3. The transition kernel $\{G(s,\cdot)\}_{s\in S_X}$ is *reversible with respect to π'.
4. The standard lazy mixing time $t_L(\epsilon)$ is no greater than the hyperfinite lazy mixing time $T_L(\epsilon)$.
5. The standard large hitting time $\tau_g(\alpha)$ is no less than the hyperfinite large hitting time $T_g(\alpha)$.
6. The standard average mixing time $t_a(\epsilon)$ is no greater than the hyperfinite average mixing time $T_a(\epsilon)$.

Proof of Main Result for Compact Case

Theorem

Let $0 < \alpha < \frac{1}{2}$. Then there exist universal constants $0 < a_\alpha, a'_\alpha < \infty$ such that, for every $\{g(x, 1, \cdot)\} \in \mathcal{C}$, we have

$$a'_\alpha t_H(\alpha) \leq t_L \leq a_\alpha t_H(\alpha).$$

Proof.

By elementary standard argument, there exists $0 < a'_\alpha < \infty$ such that $a'_\alpha t_H(\alpha) \leq t_L$.
By the transfer principle, there exists $0 < e_\alpha < \infty$ such that $T_L \leq e_\alpha T_g(\alpha)$.
Note that $t_L \leq T_L \leq e_\alpha T_g(\alpha) \leq e_\alpha \tau_g(\alpha)$.
As $0.1\tau_g(\alpha) \leq t_H(\alpha) \leq 2\tau_g(\alpha)$, we have $t_L \leq a_\alpha t_H(\alpha)$, where $a_\alpha = 20e_\alpha$. □

Trace Chain

Let $\{g(x,1,\cdot)\}_{x\in X}$ be a transition kernel of a Markov chain with stationary distribution π. Let $K \in \mathcal{B}[X]$ have measure $\pi(K) > 0$. Fix $x \in K$ and let $\{X_t\}_{t\in\mathbb{N}}$ be a Markov process with transition kernel g and starting point $X_0 = x$. Define a sequence of random variables $\{\eta_t\}_{t\in\mathbb{N}}$ by writing

$$\eta_0 = \min\{t \geq 0 : X_t \in K\}$$

and recursively setting

$$\eta_{t+1} = \min\{t > \eta_t : X_t \in K\}.$$

We define the *trace* of $\{X_t\}_{t\in\mathbb{N}}$ on K to be the Markov chain with transition kernel

$$g^{(K)}(x,t,A) = \mathbb{P}_x(X_{\eta_t} \in A).$$

Properties of the Trace Chain

Theorem

Let g be a transition kernel with stationary distribution π satisfying the Strong Feller condition. Let $K \in \mathcal{B}[X]$ be a set with $\pi(K) > 0$. Then the trace $g^{(K)}$ also satisfies the Strong Feller condition.

Let $t_m^{(K)}, \tau_g^{(K)}$ denote the mixing times and large hitting times of the trace chain, respectively. Let $\mathcal{K}[X]$ be the collection of compact subsets of X.

Theorem

Let g be a transition kernel of a Markov process on a σ-compact state space X with stationary distribution π. Let $0 < \alpha < \frac{1}{2}$. Then we have

$$t_m \leq 2 \sup_{K \in \mathcal{K}[X]} t_m^{(K)} \text{ and } \tau_g(\alpha) \geq \sup_{K \in \mathcal{K}[X]} \tau_g^{(K)}(\alpha).$$

Extension to σ-Compact Spaces

Let $\mathcal{M}$ denote the collection of discrete time reversible transition kernels with a stationary distribution on a σ-compact metric state space satisfying the Strong Feller condition.

Theorem (R.Anderson, H.Duanmu and A.Smith)

Let $0 < \alpha < \frac{1}{2}$. Then there exist universal constants $0 < a_\alpha, a'_\alpha < \infty$ such that, for every $\{g(x,1,\cdot)\}_{x \in X} \in \mathcal{M}$, we have

$$a'_\alpha t_H(\alpha) \leq t_L \leq a_\alpha t_H(\alpha).$$

Proof.

There exists an universal constant $0 < a'_\alpha < \infty$ such that $a'_\alpha t_H \leq t_L$. There exists an universal constant $0 < c_\alpha < \infty$

$$t_L \leq 2 \sup_{K \in \mathcal{K}[X]} t_m^{(K)} \leq 2c_\alpha \sup_{K \in \mathcal{K}[X]} \tau_g^{(K)}(\alpha) \leq 2c_\alpha \tau_g(\alpha) \leq 20 c_\alpha t_H(\alpha).$$

$\square$

Skeleton Chain

Definition

Let $\{g(x,1,\cdot)\}_{x\in X}$ be the transition kernel of a Markov process. For every $k \in \mathbb{N}$, the *k-skeleton chain*, denoted by $g^{(k)}$ the transition kernel

$$g^{(k)}(x,t,A) = g(x,kt,A)$$

for every $x \in X$, $t \in \mathbb{N}$ and $A \in \mathcal{B}[X]$.

Let g_L denote the lazy transition kernel of g, and let $g_L^{(k)}$ denote the k-skeleton chain of the lazy transition kernel g_L.

Almost Strong Feller Property

Very few Markov chains in statistical computation satisfies the Strong Feller Condition. We introduce the following property.

Definition ((C, R)-Almost Strong Feller)

For $C, R \in \mathbb{N}$, a transition kernel g is (C, R)-almost strong Feller if there exists transition kernels G_1, G_2 so that the following is satisfied

1. G_1 is reversible and satisfies the Strong Feller Property.
2. For some $0 \leq p \leq \frac{1}{CRt_L}$, there exists $k \leq Rt_L$ such that

$$g_L^{(k)} = (1-p)G_1 + pG_2.$$

We use $\mathcal{E}(C, R)$ to denote the collection of all (C, R)-almost strong Feller transition kernels on a σ-compact metric state space X.

Almost Strong Feller Result

We establish asymptotical equivalence between mixing times and hitting times for almost strong Feller chains.

Theorem (R.Anderson, H.Duanmu and A.Smith)

There exists an universal constant C_0 such that, for every $0 < \alpha < \frac{1}{2}$, there exist universal constants d_α, d'_α so that for all $C > C_0$, all $R \geq 1$ and all $g \in \mathcal{E}(C, R)$, we have

$$d_\alpha t_L \leq k\ell_H^{(k)}(\alpha) \leq d'_\alpha R t_L$$

where $\ell_H^{(k)}(\alpha)$ denote the maximum hitting time of the transition kernel $g_L^{(k)}$.

The mixing time of the lazy chain is asymptotically equivalent to the maximum lazy hitting time of the k-skeleton chain.

Metropolis-Hasting Chain

Fix a distribution π with continuous density ρ supported on $\mathbb{R}^d$. Also fix a reversible kernel $\{q(x,1,\cdot)\}_{x\in\mathbb{R}^d}$ on $\mathbb{R}^d$ with stationary measure ν. For every $x \in \mathbb{R}^d$, assume that $q(x,1,\cdot)$ has continuous density q_x. We define the *acceptance function* by the formula

$$\beta(x,y) = \min\{1, \frac{\rho(y)q_y(x)}{\rho(x)q_x(y)}\}.$$

Finally, define g to be the kernel given by the formula

$$g(x,1,A) = \int_{y\in A} q_x(y)\beta(x,y)dy + \delta(x,A)\int_{\mathbb{R}^d} q_x(y)(1-\beta(x,y))dy.$$

Note that g would be reversible with respect to the stationary distribution π.

Mixing Times and Hitting Times for MH Chain

Lemma

Let $\mathcal{E}$ be the collection of Metropolis-Hasting Chains with finite mixing time, and for which $q_x(y)$ is uniformly continuous jointly in x, y. Then, for all $0 < \alpha < \frac{1}{2}$ and $0 < C < \infty$, there exists a universal constant $R_{\alpha,C}$ so that all $\{g(x,1,\cdot)\}_{x\in X} \in \mathcal{E}$ are (C,R)-almost strong Feller with constant $R \leq R_{\alpha,C}$.

Theorem

Let $\mathcal{E}$ be the collection of Metropolis-Hasting Chains with finite mixing time, and for which $q_x(y)$ is jointly continuous in x, y. Then, for all $0 < \alpha < \frac{1}{2}$, there exist constants $0 < c_\alpha, c'_\alpha < \infty$ so that for all $\{g(x,1,\cdot)\}_{x\in X} \in \mathcal{E}$,

$$c_\alpha t_L \leq k\ell_H^{(k)}(\alpha) \leq c'_\alpha t_L.$$

Conclusion

In this talk, we establish the asymptotical equivalence between lazy mixing times and hitting times for large classes of Markov processes.

1. $t_L \sim t_H(\alpha)$ for finite reversible Markov processes.
2. By nonstandard analysis, $t_L \sim t_H(\alpha)$ for reversible Markov processes on compact state space satisfying the Strong feller condition.
3. By trace chain, $t_L \sim t_H(\alpha)$ for reversible Markov processes on σ-compact state space satisfying the Strong feller condition.
4. $t_L \sim t_H(\alpha)$ for reversible Markov processes on σ-compact state space satisfying the almost Strong feller condition.
5. Application for MH-algorithm.