STA4502: Topics in Stochastic Processes Winter 2018

Lecture 5: Introduction — March 28, 2018

Lecturer: Jeffrey Rosenthal *Scribe: Jeffrey Negrea*

1 Last Example of Random Walks on Groups

1.1 Bit-Flipping

Recall the bit-flipping example:

$$\begin{aligned}\mathcal{X} &= (\mathbb{Z}/2)^d \\ Q &= \text{Unif}(\{\text{id}\} \cup \{e_j : j \in [d]\})\end{aligned}$$

where e_j frobnicates the jth bit, leaving the other $j-1$ bits unchanged, and $\text{id} = 0$ is the identity element.

We already derived the characters for this group:

$$\chi_m(x) = \exp\left(2\pi i \sum_{i=d}^{n} \frac{m_i x_i}{2}\right) = (-1)^{\langle m,\ x\rangle}$$

The eigenvalues of P may then be computed

$$\begin{aligned}\lambda_m &= \mathop{\mathbb{E}}_{X\sim Q}[\chi_m(X)] \\ &= \sum_{x\in\mathcal{X}} Q(x)\chi_m(x) \\ &= Q(0)\chi_m(0) + \sum_{i=1}^{d} Q(e_i)\chi_m(e_i) \\ &= \frac{1}{d+1}\left(1 + \sum_{i=1}^{d}(-1)^{\langle m,\ e_i\rangle}\right) \\ &= \frac{1}{d+1}\left(1 - N(m) + (d - N(m))\right) \\ &= 1 - \frac{2N(m)}{d+1}\end{aligned}$$

where $N(m) = \langle m,\ \mathbf{1}\rangle$ is the number of 1s in m, since $\langle m,\ e_i\rangle$ is 1 if $m_i = 0$ and is -1 otherwise.

Thus $\lambda_\star = 1 - \frac{2}{d+1}$, which is realised when $N(m) = 1$, for example when $m = e_1$.

Using the crude $\lambda - \star$-based method, we have the following bound for the total variation distance of the marginal distribution of the chain to stationarity:

$$\|\mu_k - \pi\|_{\text{TV}} \leq \frac{\sqrt{|\mathcal{X}|}}{2}\lambda_\star^k = \frac{\sqrt{|\mathcal{X}|}}{2}\left(1 - \frac{2}{d+1}\right)^k$$

For $d = 1000$ we get that $k_\star = 175243$ is sufficient for $\|\mu_{k_\star} - \pi\|_{\mathrm{TV}} \leq 0.01$

Using the more refined summation-based method, we have the following bound for the total variation distance of the marginal distribution of the chain to stationarity:

$$\begin{aligned}
\|\mu_k - \pi\|_{\mathrm{TV}} &\leq \frac{1}{2}\sqrt{\sum_{m \in \mathcal{X}\setminus\{0\}} |\lambda_m|^{2k}} \\
&\leq \frac{1}{2}\sqrt{\sum_{m \in \mathcal{X}\setminus\{0\}} \left|1 - \frac{2N(m)}{d+1}\right|^{2k}} \\
&\leq \frac{1}{2}\sqrt{\sum_{n=1}^{d} \binom{d}{n} \left|1 - \frac{2n}{d+1}\right|^{2k}}
\end{aligned}$$

For $d = 1000$ we get that $k_\star = 3684$ is sufficient for $\|\mu_{k_\star} - \pi\|_{\mathrm{TV}} \leq 0.01$. This was calculated with the following `R` script:

```
bins = choose(1000,1:1000)
pows = abs(1-2*(1:1000)/1001)
tv.bound = function(k){1/2 * sqrt(sum(bins * pows ^ (2 * k))) -0.01}
k.star = ceiling(uniroot(tv.bound,c(0,176000))$root)
```

We can also get a lower bound for the total variation distance from stationarity, which gives a necessary number of steps through the chain:

$$\begin{aligned}
\|\mu_k - \pi\|_{\mathrm{TV}} &\geq \frac{1}{2}\left|\mathop{\mathbb{E}}_{X\sim Q}[\chi_{e_1}]\right|^k \\
&= \frac{1}{2}\left(1 - \frac{2}{d+1}\right)^k
\end{aligned}$$

For $d = 1000$ we get that $k_\star \geq 1957$ is necessary for $\|\mu_{k_\star} - \pi\|_{\mathrm{TV}} \leq 0.01$.

Putting these together we get that, for $d = 1000$, the true $k_\star$ is between 1957 and 3684.

2 Drift and Minorisation Conditions

Recall the uniform minorisation condition:

If $P(x, \cdot) \geq \epsilon\rho(\cdot)$ for all $x \in \mathcal{X}$ for some $\epsilon > 0$ and some probability measure ρ on $\mathcal{X}$, then the markov chain is *uniformly geometrically ergodic*. That is to say, for any initial probability measure μ_0 and for any $k \in \mathbb{N}$:

$$\|\mu_k - \pi\|_{\mathrm{TV}} \leq (1 - \epsilon)^k \ .$$

The universal quantification over all initial measures seems to be nice mathematically, but restricts our analysis to Markov chains which converge to stationarity uniformly. In order to be able to analyse markov chains without uniform convergence properties we need to develop new tools. Te following example will be used to illustrate non-uniform ergodicity in this section:

Example 1 (Canonical non-uniformly ergodic example: AR(1)-process). . A particular gaussian autoregresive process of order 1 is given by:

Let $\mathcal{X} = \mathbb{R}$ and let $P(x, \cdot) \equiv \mathcal{N}(\cdot\,; \frac{x}{2}, \frac{3}{4})$.

This kernel "pulls" the chain back to 0 on each step. The farther the chain is from 0 the longer it will take to return to a neighbourhood of 0.

Does this process have a stationary distribution? Yes! The stationary distribution is $\mathcal{N}(0, 1)$. We verify this below. Suppose that $X_n \sim \mathcal{N}(0, 1)$, then;

$$\begin{aligned} X_n \perp\!\!\!\perp Z = X_{n+1} - \frac{X_n}{2} &\sim \mathcal{N}(0, \frac{3}{4}) \\ \implies X_{n+1} &\sim \mathcal{N}(0, 1)\ . \end{aligned}$$

Since this example is so simple, we could directly bound its total variation distance from stationarity. Since the methods used wouldn't generalise, this would not be instructive. In this section we will develop generally applicable techniques, and then apply them to this example.

In this example, we cannot get uniform minorisation for all $x \in \mathcal{X}$ since, taking any two x sufficiently far apart w,e could show that no uniform minorising probability measure exists for any $\epsilon > 0$.

2.1 Drift and minorisation derivation

Instead of minorising uniformly over the whole state space, we may instead attempt to minorise only uniformly over some subset of the state space. More precisely we will attempt to find $C \subset \mathcal{X}$ such that $P(x, \cdot) \geq \epsilon\rho(\cdot)$ for all $x \in C$. We will call such a C a "small set".

We cannot use the same construction as in the uniform case — we need to first allow both copies of the chain to reach the small set, then hope that the chains couple.

2.1.1 Coupling Construction

The coupling is constructed as follows. Let $X_0 \sim \mu_0$ and let $Y_0 \sim \pi$. At stage n, given the coupled X_n and Y_n and $Z_n \sim \text{Bernoulli}(\epsilon)$ we determine the coupled X_{n+1} and Y_{n+1} by:

if	$X_n = Y_n$	then	$X_{n+1} = Y_{n+1} \sim P(X_n, \cdot)$	
else if	$(X_n, Y_n) \in C^2 \ \wedge \ Z_n = 1$	then	$X_{n+1} = Y_{n+1} \sim \rho(\cdot)$	
else if	$(X_n, Y_n) \in C^2 \ \wedge \ Z_n = 0$	then	$X_{n+1} \sim \frac{P(X_n, \cdot) - \epsilon\rho(\cdot)}{1-\epsilon} = R(X_n, \cdot)$	
			$\perp\!\!\!\perp\ Y_{n+1} \sim \frac{P(Y_n, \cdot) - \epsilon\rho(\cdot)}{1-\epsilon} = R(Y_n, \cdot)$	
else	$(X_n, Y_n) \notin C^2$	then	$X_{n+1} \sim P(X_n, \cdot)$	
			$\perp\!\!\!\perp\ Y_{n+1} \sim P(Y_n, \cdot)$	

2.1.2 Coupling Inequality

To bound the distance from stationarity we use the coupling inequality:

$$\|\mu_k - \pi\|_{\mathrm{TV}} \leq \mathbb{P}(X_k \neq Y_k) \ .$$

Choose $j \in [k]$. Let $N_k = |\left\{m \in [k] : (X_m, Y_m) \in C^2\right\}|$. We can then decompose the RHS above as:

$$\begin{aligned}\mathbb{P}(X_k \neq Y_k) &= \mathbb{P}(X_k \neq Y_k, N_{k-1} \geq j) + \mathbb{P}(X_k \neq Y_k, N_{k-1} < j) \\ &\leq (1-\epsilon)^j + \mathbb{P}(X_k \neq Y_k, N_{k-1} \leq j-1)\end{aligned}$$

We need some new techniques to bound $\mathbb{P}(X_k \neq Y_k, N_{k-1} \leq j-1)$, the probability that we have an insufficient number of chances to couple and do not couple.

2.1.3 Drift conditions

The new trick will be to introduce a "drift condition". There are univariate and bivariate versions of drift conditions. In this course we will only examine bivariate versions.

We introduce the forward expectation operator $\overline{P}$ defined by

$$\overline{P}h(x,y) = \mathbb{E}[h(X_1, Y_1)\big|(X_0, Y_0)] = (x,y)$$

Suppose that we have a function $h : \mathcal{X}^2 \to [1, \infty)$ and $\alpha > 1$ such that

$$\overline{P}h(x,y) \leq \frac{h(x,y)}{\alpha} \qquad \forall (x,y) \notin C^2 \tag{1}$$

Then h is called a drift function and (**??**) is called a drift . The key idea is that, on average, h gets smaller per step of the Markov chain. Often times we will use an additive, symmetric drift function of the form $h(x,y) = 1 + V(x) + V(y)$ for $V : \mathcal{X} \to [0, \infty)$. For the AR(1) example, we will use $V(x) = x^2$ and so $h(x,y) = 1 + x^2 + y^2$.

2.1.4 Final coupling bound

Suppose we have a drift condition as well as a local minorisation condition. Then, for $B \geq 1$,

$$\begin{aligned}&\mathbb{P}(X_k \neq Y_k, N_{k-1} \leq j-1) \\ \text{equality if } B \neq 1 \qquad &\leq \mathbb{P}(X_k \neq Y_k, B^{-N_{k-1}} \geq B^{-(j-1)}) \\ &= \mathbb{P}\left((\mathbf{1}_{X_k \neq Y_k} B^{-N_{k-1}}) \geq B^{-(j-1)}\right) \\ \text{Markov's Ineq.} \qquad &\leq B^{j-1}\mathbb{E}\left[\mathbf{1}_{X_k \neq Y_k} B^{-k-1}\right] \\ &\leq B^{j-1}\alpha^{-k}\mathbb{E}\left[\mathbf{1}_{X_k \neq Y_k}\alpha^k B^{-N_{k-1}} h(X_k, Y_k)\right]\end{aligned}$$

Let $M_k = \mathbf{1}_{X_k \neq Y_k}\alpha^k B^{-N_{k-1}} h(X_k, Y_k)$and let $B = 1 \vee \left[\alpha(1-\epsilon)\sup_{(x,y)\in C^2} \mathbb{E}_{\substack{X_1 \sim R(x,\cdot) \\ Y_1 \sim R(y,\cdot)}}[h(X_1, Y_1)]\right]$.

We claim that M_k is a supermartingale with respect to the filtration generated by $\{(X_k, Y_k) : k \in \mathbb{N}\}$. In light of this claim, we have:

$$\begin{aligned}\mathbb{P}(X_k \neq Y_k, N_{k-1} \leq j-1) &\leq B^{j-1}\alpha^{-k}\mathbb{E}[M_0] \\ &= B^{j-1}\alpha^{-k}\mathbb{E}[h(X_0, Y_0)]\end{aligned}$$

We verify the submartingale claim below. The proof relies on the fact the the N_{k-1} term in the definition of M_k is predictable.

Case 1: The chain coupled by time $k+1$, so that $X_{k+1} = Y_{k+1}$. Then $M_{k+1} = 0 \leq M_k$.

Case 2: The chains were not coupled and did not have a chance to couple at time $k+1$, i.e.: $(X_k, Y_k) \notin C^2$ and $X_{k+1} \neq Y_{k+1}$. Then $N_{k-1} = N_k$ (no new chance to couple) so that

$$\begin{aligned}&\mathbb{E}[M_{k+1}\big|(X_k, Y_k)] \\ &\quad= \mathbb{E}[\mathbf{1}_{X_k = Y_k} M_{k+1}\big|(X_k, Y_k)] \\ &\qquad+ \mathbb{E}[\mathbf{1}_{(X_k,Y_k)\notin C^2}\mathbf{1}_{X_k \neq Y_k}\ M_{k+1}\big|(X_k, Y_k)] \\ &\qquad+ \mathbb{E}[\mathbf{1}_{(X_k,Y_k)\in C^2}\mathbf{1}_{X_k \neq Y_k}\ M_{k+1}\big|(X_k, Y_k)]\end{aligned}$$

The first term is 0 since if the chain is coupled at time k then it is coupled at time $k+1$ so $M_{k+1} = 0$.

The second term can be bounded by:

$$\begin{aligned}\mathbb{E}[\mathbf{1}_{(X_k,Y_k)\notin C^2}\mathbf{1}_{X_k\neq Y_k} M_{k+1}\big|(X_k, Y_k)] &\leq \mathbb{E}[\mathbf{1}_{(X_k,Y_k)\notin C^2 \wedge X_k \neq Y_k}\ \alpha^{k+1}B^{-N_k}h(X_{k+1}, Y_{k+1})\big|(X_k, Y_k)] \\ &= \alpha\mathbb{E}[\mathbf{1}_{(X_k,Y_k)\notin C^2 \wedge X_k \neq Y_k}\ \alpha^{k}B^{-N_{k-1}}h(X_{k+1}, Y_{k+1})\big|(X_k, Y_k)] \\ &= \mathbf{1}_{(X_k,Y_k)\notin C^2} M_k\ \frac{\mathbb{E}[h(X_{k+1}, Y_{k+1})\big|(X_k, Y_k)]}{h(X_k, Y_k)/\alpha} \\ &\leq \mathbf{1}_{(X_k,Y_k)\notin C^2} M_k\end{aligned}$$

The third term can be bounded by:

$$\begin{aligned}\mathbb{E}[\mathbf{1}_{(X_k,Y_k)\in C^2}\mathbf{1}_{X_k\neq Y_k} M_{k+1}\big|(X_k, Y_k)] &\leq \mathbb{E}[\mathbf{1}_{(X_k,Y_k)\in C^2 \wedge X_k \neq Y_k \wedge Z_k = 0}\ \alpha^{k+1}B^{-N_k}h(X_{k+1}, Y_{k+1})\big|(X_k, Y_k)] \\ &\leq \frac{\alpha}{B}\mathbb{E}[\mathbf{1}_{(X_k,Y_k)\in C^2 \wedge X_k \neq Y_k \wedge Z_k = 0}\ \alpha^{k}B^{-N_{k-1}}h(X_{k+1}, Y_{k+1})\big|(X_k, Y_k)] \\ &= \mathbf{1}_{(X_k,Y_k)\in C^2} M_k \frac{\mathbb{E}[\mathbf{1}_{Z_k=0}h(X_{k+1}, Y_{k+1})\big|(X_k, Y_k)]}{Bh(X_k, Y_k)/\alpha} \\ &= \mathbf{1}_{(X_k,Y_k)\in C^2} M_k \frac{(1-\epsilon)\mathbb{E}[h(X_{k+1}, Y_{k+1})\big|(X_k, Y_k, Z_k = 0)]}{Bh(X_k, Y_k)/\alpha} \\ &\leq \mathbf{1}_{(X_k,Y_k)\in C^2} M_k\end{aligned}$$

The last step follows from the choice of B. Combining these, we get the supermartingale property for M_k.

2.1.5 Drift and minorisation theorem

Theorem 2. *If a Markov chain has a stationnary distribution, π, and a local minorisation condition of the form:*

$$\exists(C \in \Sigma_{\mathcal{X}},\ \epsilon > 0,\ \rho \in \mathcal{M}(\Sigma_{\mathcal{X}}))\ :\ (\pi(C) > 0 \text{ and } (x \in C \implies P(x,\cdot) \geq \epsilon\rho(\cdot)))\ ,$$

and a drift condition of the form:

$$(\exists h : \mathcal{X}^2 \to [1,\infty),\ \alpha > 0)\ :\ ((x,y) \notin C \times C \implies \overline{P}h(x,y) \leq \alpha^{-1}h(x,y))$$

then for any $j \in [k]$ we have

$$\|\mu_k - \pi\|_{TV} \leq (1-\epsilon)^j + \alpha^{-k}B^{j-1}\mathbb{E}h(x_0,y_0)\ ,$$

where $B = 1 \vee \left[\alpha(1-\epsilon)\sup_{(x,y)\in C^2} \mathop{\mathbb{E}}_{\substack{X_1\sim R(x,\cdot)\\ Y_1\sim R(y,\cdot)}}[h(X_1,Y_1)]\right]$.

Example 3 (Canonical non-uniformly ergodic example: AR(1)-process — continued). We will choose $C = [-\sqrt{3},\sqrt{3}]$ and $h(x,y) = 1 + x^2 + y^2$. Then we get:

$$\begin{aligned}
\epsilon &= \int_{\mathbb{R}} \inf_{x\in C} \mathcal{N}(\mathrm{d}y\ ;\ \frac{x}{2}, \frac{3}{4}) \\
&= \mathbb{P}(|\mathcal{N}(0,1)| \geq 1) \\
&= 0.3173105\ ,
\end{aligned}$$

and, for $(x,y) \notin C^2$ we have $h(x,y) \geq 4$, so then:

$$\begin{aligned}
\overline{P}h(x,y) &= 1 + \left(\frac{x^2}{4} + \frac{3}{4}\right) + \left(\frac{y^2}{4} + \frac{3}{4}\right) \\
&= \frac{9 + h(x,y)}{4} = \frac{\frac{9}{4}4 + h(x,y)}{4} \\
&\leq \frac{h(x,y)}{4}\left(1 + \frac{9}{4}\right) \\
&= \frac{13}{16}h(x,y)
\end{aligned}$$

which means we can take

$$
\begin{aligned}
\alpha &= \frac{16}{13} \\
B &= \frac{16}{13}(1-0.3173105) \sup_{(x,y)\in C^2} \mathop{\mathbb{E}}_{\substack{X_1\sim R(x,\cdot)\\ Y_1\sim R(y,\cdot)}} [h(X_1,Y_1)] \\
&\le \frac{16}{13}(1-0.3173105) \sup_{(x,y)\in C^2} \left(1+\overline{R}[x^2]+\overline{R}[y^2]\right) \\
&= \frac{16}{13}(1-0.3173105) \sup_{(x,y)\in C^2} \left(1+\frac{\overline{P}[x^2]-\epsilon \mathop{\mathbb{E}}_{W\sim\rho}[W]}{1-\epsilon}+\frac{\overline{P}[y^2]-\epsilon \mathop{\mathbb{E}}_{W\sim\rho}[W]}{1-\epsilon}\right) \\
&= \frac{16}{13}(1-0.3173105) \sup_{(x,y)\in C^2} \left(1+\frac{\overline{P}[x^2]}{1-\epsilon}+\frac{\overline{P}[y^2]}{1-\epsilon}\right) \\
&= \frac{16}{13}(1-0.3173105) \left(1+\frac{3/4+3/4}{1-\epsilon}+\frac{3/4+3/4}{1-\epsilon}\right) \\
&= \frac{16}{13}(1-0.3173105) \left(1+\frac{3}{1-\epsilon}\right) \\
&= \frac{16}{13}(4-0.3173105) \\
&= 4.532541 \le 4.6
\end{aligned}
$$

Take $j = \lfloor k/10 \rfloor$

Then we get the following bound for $\mu_0 = \delta_0$:

$$
\|\mu_k - \pi\|_{\mathrm{TV}} \le (0.683)^{\lfloor k/10 \rfloor} + \left(\frac{13}{16}\right)^k 4.6^{\lfloor k/10 \rfloor - 1}\ 2 \tag{2}
$$

Since $\mathbb{E}h(X_0, y_0) = 1 + 0 + \mathbb{E}Y_0^2 = 2$

Finally, for $k_\star = 130$ we have $\|\mu_{k_\star} - \pi\|_{\mathrm{TV}} \le 0.01$, which was computed with the following `R` script:

```
tv.bound = function(k){(1-2*pnorm(-1))^floor(k/10) +
                       (13/16)^k * ((16/13) * (4- 0.3173105))^(floor(k/10)-1)*2 -0.01 }
k.star = ceiling(uniroot(tv.bound,c(0,1000))$root)
```