Eric Fischer
UID: 303 759 361
emfischer712@ucla.edu
Stat 202C - Homework 1

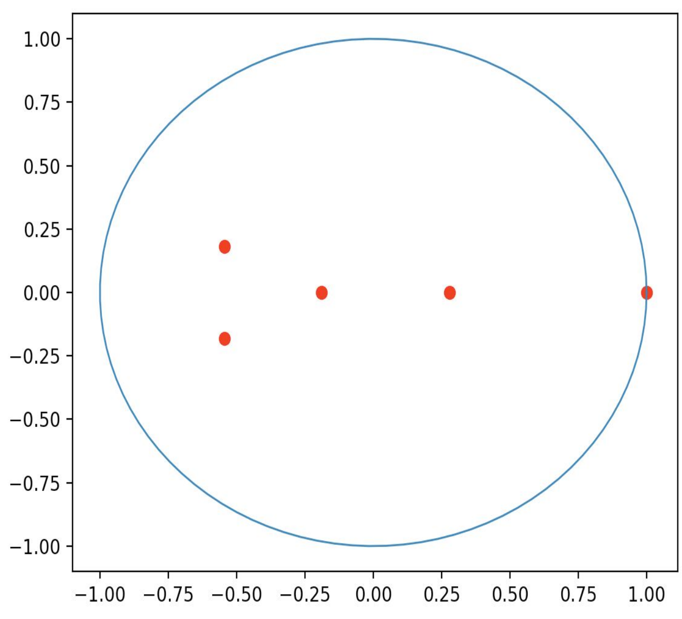
Markov Chain Theory and Markov Chain Monte Carlo Methods

Markov Kernel and the TV-norm, KL-divergence, and Diaconis-Hanlon Bound

Problem 1a: Here we consider the Markov kernel below for the five families living in an island.

$$K = \begin{pmatrix} 0.3, & 0.6, & 0.1, & 0.0, & 0.0 \\ 0.2 & 0.0, & 0.7, & 0.0, & 0.1 \\ 0.0, & 0.5, & 0.0, & 0.5, & 0.0 \\ 0.0, & 0.0, & 0.4, & 0.1, & 0.5 \\ 0.4, & 0.1, & 0.0, & 0.4, & 0.1 \end{pmatrix}$$

A plot of the 5 eigenvalues (complex numbers) in a 2D plane is below. They are represented as dots in a unit circle.



Note the invariant probability of the distribution is **[0.1488, 0.2353, 0.2635, 0.2098, 0.1427]**. The value of $\boldsymbol{\lambda}_{\text{slem}} = |\boldsymbol{\lambda}_2|$, the second largest eigenvalue modulus, is **0.7833**. The convergence rate is decided by $\boldsymbol{\lambda}_{\text{slem}}$.

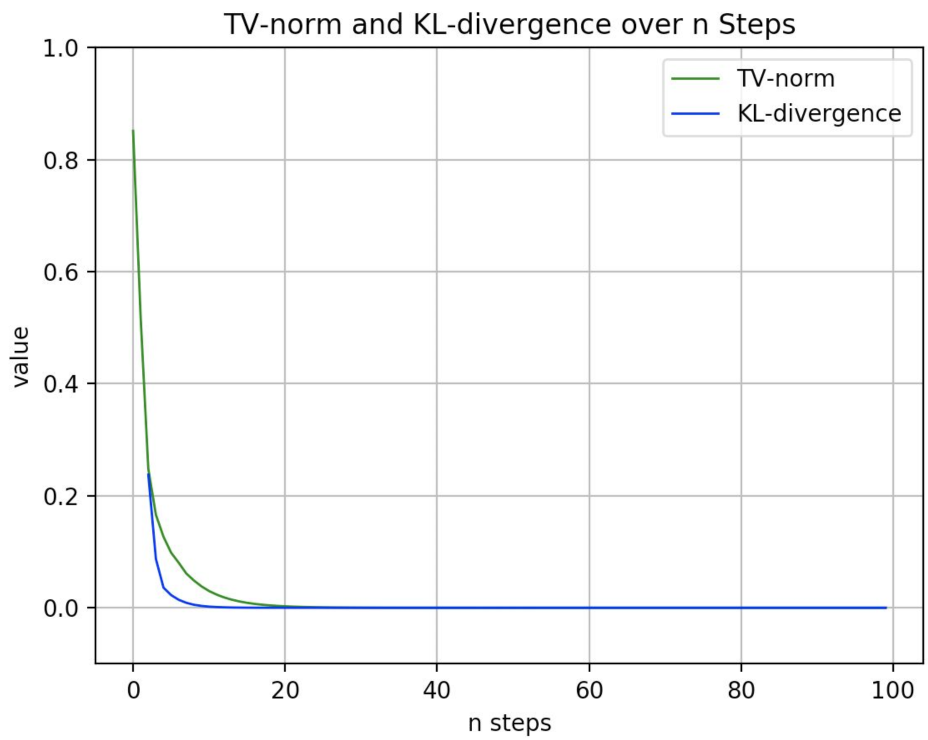
<u>Problem 1b</u>: Now suppose we start with initial probabilities v = (1,0,0,0,0), i.e. we know that the initial state of the Markov chain is at $x_0 = 1$. So at step n, the Markov chain state follows a distribution $\mu_n = v * K^n$. We compute the distance between μ_n and π by the **TV-norm**,

$$d_{\mathrm{TV}}(n) = ||\pi - \mu_n||_{\mathrm{TV}} = \frac{1}{2}\sum_{i=1}^{5} |\pi(i) - \mu_n(i)|;$$

or the **KL-divergence**,

$$d_{\mathrm{KL}}(n) = \sum_{i=1}^{5} \pi(i) \log \frac{\pi(i)}{\mu_n(i)}.$$

A plot of $d_{\mathrm{TV}}(n)$ and $d_{\mathrm{KL}}(n)$ for the first 100 steps is below. We can see that KL-divergence convergences faster to 0.



<u>Problem 1c</u>: Now we perform a calculation of the contraction coefficient for K. Note that the contraction coefficient is the maximum TV-norm between any two rows in the transition kernel,

$$C(K) = \max_{x,y} ||K(x, \cdot) - K(y, \cdot)||_{\mathrm{TV}}$$

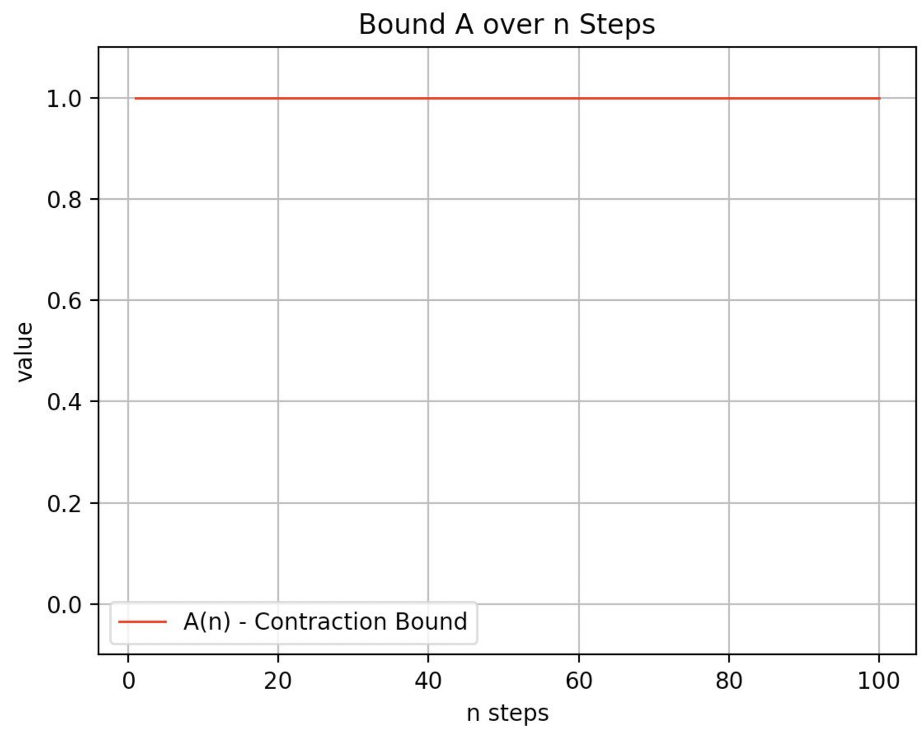
One can prove that

$$||\nu_1 \cdot K - \nu_2 \cdot K||_{\mathrm{TV}} \leq C(K)||\nu_1 - \nu_2||_{\mathrm{TV}}$$

As $||\nu_1 - \nu_2||_{\mathrm{TV}} \leq 1$, if $C(K) < 1$ then the convergence rate could be upper bounded by

$$||\nu_1 \cdot K^n - \nu_2 \cdot K^n||_{\mathrm{TV}} \leq C^n(K)||\nu_1 - \nu_2||_{\mathrm{TV}} \leq C^n(K) = A(n), \quad \forall \nu_1, \nu_2$$

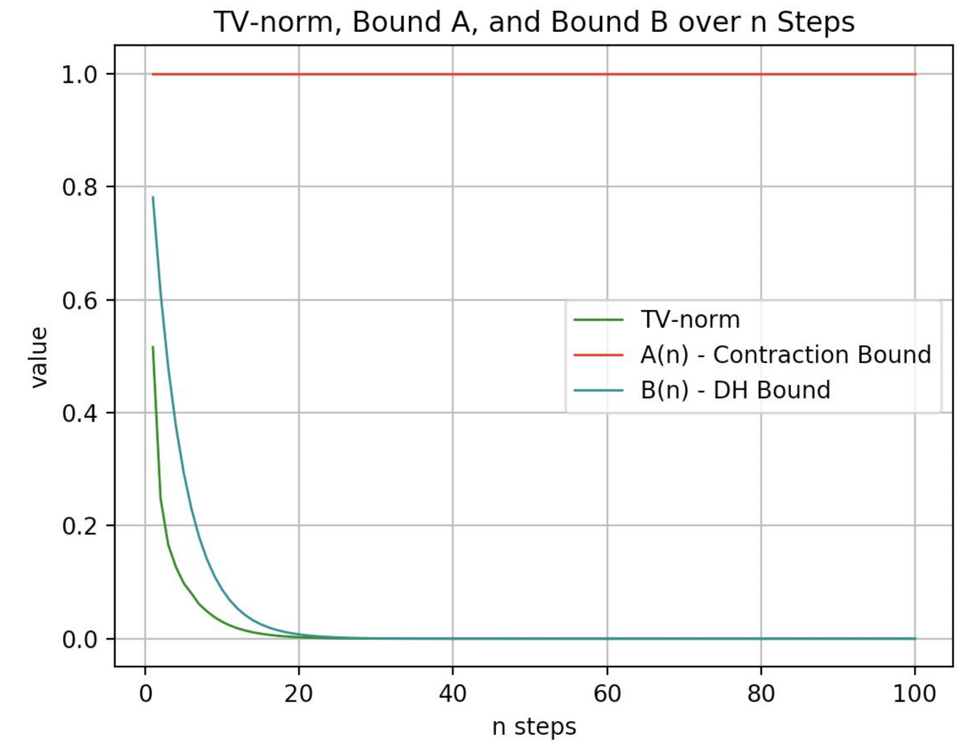
A plot of the bound $C^n(K)$, which we refer to as $A(n)$, over $n = 1, \ldots, 100$ is below. As we can see, it is not very useful.



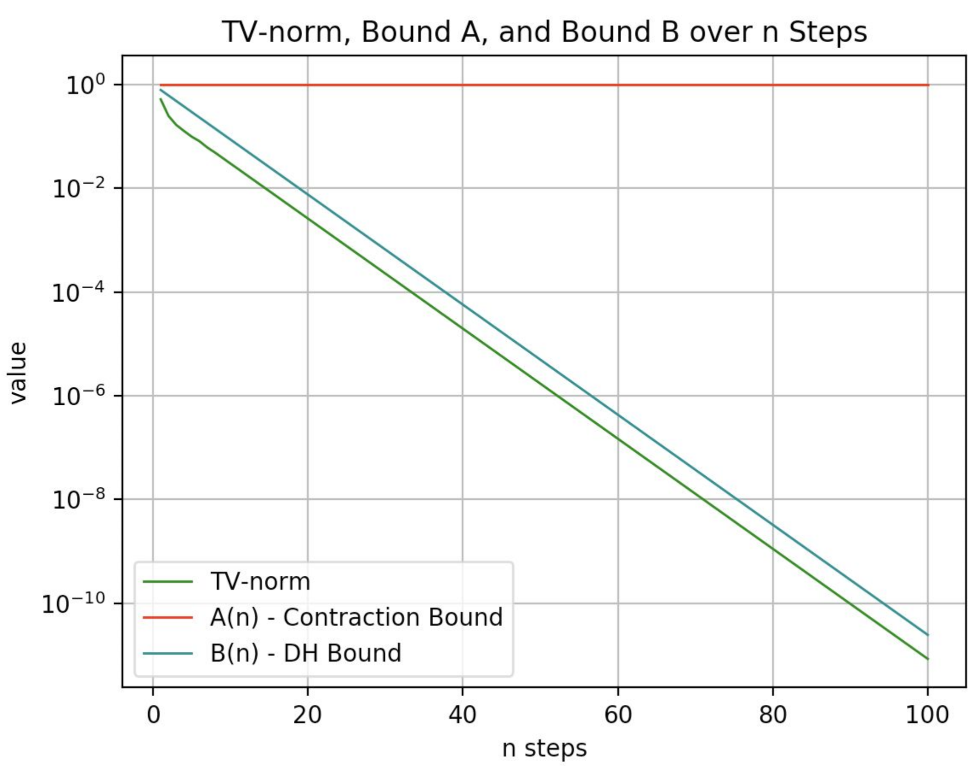
Problem 1d: There is also another bound — the Diaconis-Hanlon bound which we refer to as *B(n)* below,

$$||\pi - \nu K^n||_{\mathrm{TV}} \leq \sqrt{\frac{1 - \pi(x_0)}{4\pi(x_0)}} \lambda_{\mathrm{slem}}^n = B(n)$$

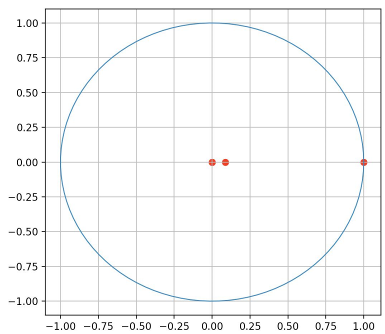
where $x_0 = 1$ is the initial state and $\pi(x_0)$ is a target probability at $x = 1$. A plot of the real convergence rate $d_{\mathrm{TV}}(n)$ in comparison with $A(n)$ and $B(n)$ is below. We can see that *B(n)*, the Diaconis-Hanlon bound, is a useful bound on the convergence, unlike *A(n)*, the bound based on the contraction coefficient.

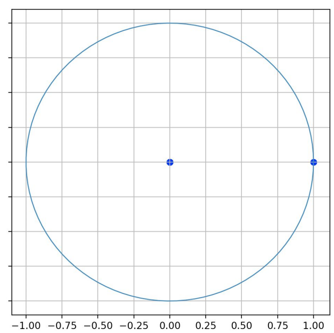


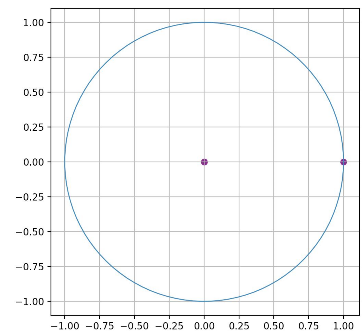
Also included is a 2nd figure to compare the log-plots of the TV-norm, $A(n)$, and $B(n)$, as they are exponential rates.



<u>Problem 1e</u>: Now we define a new Markov chain with transition kernel $P = K^n$ and then draw the 5 eigenvalues of P on a 2D complex plane as we did in **Problem 1a**. We show how these eigenvalues move on the plane at 3 stages $n = 10, 100, 1000$ below by showing a drawing of each stage. (Drawing traces of the 5 dots to show their trajectories makes it difficult to see the movement of each dot separately.)







If we examine the matrix P for $n = 1000$ to see whether it becomes the "ideal" transition kernel, in which each row is the stationary distribution π, we can see that it indeed does. The matrix $P = K^{1000}$ is:

[[0.1488, 0.2353, 0.2635, 0.2098, 0.1427],
[0.1488, 0.2353, 0.2635, 0.2098, 0.1427],
[0.1488, 0.2353, 0.2635, 0.2098, 0.1427],
[0.1488, 0.2353, 0.2635, 0.2098, 0.1427],
[0.1488, 0.2353, 0.2635, 0.2098, 0.1427]]

Irreducibility, Aperiodicity, and Invariant Probabilities for 2 Markov Kernels

<u>Problem 2</u>: Now we consider two more transition matrices K_1 and K_2 below.

$\boldsymbol{K_1}$ = [[0.1, 0.4, 0.3, 0.0, 0.2],
[0.5, 0.3, 0.2, 0.0, 0.0],
[0.0, 0.4, 0.5, 0.1, 0.0],
[0.0, 0.0, 0.0, 0.5, 0.5],
[0.0, 0.0, 0.0, 0.7, 0.3]]

$\boldsymbol{K_2}$ = [[0.0, 0.0, 0.0, 0.4, 0.6],
[0.0, 0.0, 0.0, 0.5, 0.5],
[0.0, 0.0, 0.0, 0.9, 0.1],
[0.0, 0.2, 0.8, 0.0, 0.0],
[0.3, 0.0, 0.7, 0.0, 0.0]]

<u>Problem 2a</u>: If we ask whether K_1 and K_2 are irreducible and aperiodic, we should recall the definitions. A Markov chain is ***irreducible*** if its transition kernel K has only 1 communication class, i.e. it is possible to get to any state from any other state. An irreducible Markov chain with transition kernel K is ***aperiodic*** if its largest period is $d = 1$. (An irreducible Markov chain has ***period*** d if there is a (unique) partition of graph G into d cyclic classes — in other words, it will always take a multiple of d timesteps to return any one of d cyclic classes. There is no connection between states within a cyclic class in a periodic ($d > 1$) Markov chain.)

If we draw the transition diagram for K_1, we can see that in some cases there is a zero probability of reaching a state from another state. For example, once we are in state 4 or 5, we can transition between states 4 and 5 but cannot leave those states. So K_1 is reducible. We can also see from the diagram that any state can return to itself in 1 timestep, indicating K_1 is aperiodic.

Drawing the transition diagram for K_2, we can see that there is a nonzero probability to reach any state from any other state, so K_2 is irreducible. Interestingly, K_2 has a period of 2 because upon close examination, it takes at least 2 timesteps for any state to be able to return to itself. This means no state in the K_2 matrix possesses a nonzero probability value for returning to itself — the diagonal of kernel K_2 is all zeros.

<u>Problem 2b</u>: The 5 eigenvalues of K_1 are [1, 0.915, -0.2, -0.198, 0.182].The 5 eigenvalues of K_2 are [-1, 1, -0.265, 0.265, -3.270e-18]. The eigenvectors for K_1 and K_2 are below.

$\boldsymbol{K_1}$: [[-4.274e-15, -2.560e-01, -6.185e-14, 1.300e-02, -6.323e-01],
[-6.523e-15, -4.238e-01, 3.675e-14, -7.740e-03, -1.042e-01],
[-5.912e-15, -3.919e-01, 1.586e-14, -3.372e-03, 6.623e-01],
[-8.137e-01, 6.340e-01, 7.071e-01, -7.080e-01, 3.081e-01],
[-5.812e-01, 4.356e-01, -7.071e-01, 7.061e-01, -2.345e-01]]

K_2: [[5.041e-02, -5.041e-02, 5.669e-01, 5.669e-01, 6.172e-01],
[1.400e-01, -1.400e-01, -3.780e-01, -3.780e-01, -7.715e-01],
[6.778e-01, -6.778e-01, -1.890e-01, -1.890e-01, 1.543e-01],
[-7.002e-01, -7.002e-01, 5.000e-01, -5.000e-01, 1.737e-16],
[-1.680e-01, -1.680e-01, -5.000e-01, 5.000e-01, -2.078e-17]]

Problem 2c: For each matrix K_1 and K_2 there is one invariant probability. The invariant probability for K_1 is [3.064e-15, 4.676e-15, 4.238e-15, 5.833e-01, 4.167e-01]. And the invariant probabilities for K_2 is [0.029, 0.081, 0.390, -0.403, -0.097].

Markov Chain Returning Time

<u>Problem 3</u>: A Markov chain returning time $\tau_{ret}(i)$ is the minimum steps that a Markov chain returns to state i after leaving i. Suppose we consider a random walk in the countable set of non-negative numbers $\Omega = \{0,1,2, \ldots\}$. At a step, the Markov chain state $x_t = n$, it has probability α to go up (i.e. $x_{t+1} = n + 1$) and probability $1 - \alpha$ to return to $x_{t+1} = 0$. We calculate the probability of returning to state 0 in finite steps and the expected return time $E[\tau_{ret}(0)]$ below.

$$Prob(\tau_{ret}(0) < \infty) = \sum_{\tau(0)=1}^{\infty} Prob(\tau(0))$$

Every step the Markov chain has probability $1-\alpha$ to return to state 0 ($x_{t+1} = 0$).
α to not return to state 0.

So the $p(\tau_{ret}(0) < \infty) = \alpha + \alpha(1-\alpha) + \alpha^2(1-\alpha) + \ldots$ and so forth, ie the calculation for the probability of returning to state 0 in a finite number of steps can be expressed as the infinite summation of the probability of success at each timestep. So $\rightarrow$

$$p(\tau_{ret}(0) < \infty) = (1-\alpha)(1+\alpha+\alpha^2+\ldots)$$
$$= 1-\alpha * \frac{1}{1-\alpha}$$
$$= 1 \quad (\text{if } \alpha < 1)$$

So the probability to return to state 0 in a finite number of steps is 1 if $\alpha < 1$. This means that the expected return time $E[\tau_{ret}(0)]$, or the expected number of timesteps it will take to return to step 0, is just the idea of the standard mean number of failures we see before the first success in geometric distributions. So,

$$E[\tau_{ret}(0)] = \begin{cases} \frac{1}{1-\alpha} & 0 \le \alpha < 1 \\ \infty & \alpha \ge 1 \end{cases}$$

Problem 4:

4) Let P be a transition matrix over a finite space Ω with stationary dist. π. Suppose P satisfies detailed balance; $\pi(i)K_{ij} = \pi(j)K_{ji}$, $\forall i,j \in \Omega$. Let μ be any dist. over Ω and let $\nu = \mu P$ be the dist. obtained by applying a single Markov transition to μ.

Prove $KL(\pi \| \nu) \leq KL(\pi \| \mu)$.

$$Q(y,x) = \frac{p(x,y)\mu(x)}{\nu(y)} \rightarrow \text{reverse transition probability}$$

We want to show $KL(\pi\|\mu) - KL(\pi\|\nu) = E_{Y\sim\pi}[KL_x(p(y,x)\| q(y,x))] \geq 0$, ie that the KL divergence decreases monotonically.

KL-divergence definition:
(relative entropy)
"measure of surprise"
dist.-wise assymetric measure

$$KL(P\|Q) = \sum_{x\in X} P(x)\log\left(\frac{P(x)}{Q(x)}\right)$$

← expectation of the logarithmic difference between prob. P and Q; where expectation is taken using prob. P

$$KL(\pi\|\nu) = \sum_y \pi(y)\log\frac{\pi(y)}{\nu(y)} = \underbrace{\boxed{\sum_y \pi(y)\,\frac{\pi(y)}{u(x)p(x,y)}}}_{KL(\pi\|\nu)} \leftarrow \nu = \mu P$$

$$KL(\pi\|\nu) \leq -\sum_y \pi(y)\log\frac{u(x)}{\pi(y)}$$

$$\leq -\sum_y \pi(y)\log\left[q(y,x)\frac{u(x)}{\pi(x)}\right] \qquad \left(q(y,x) = \frac{p(x,y)u(x)}{v(y)}\right)$$

$$\leq -\sum_y \pi(y)\,q(y,x)\log\frac{u(x)}{\pi(x)} = \sum_x \pi(x)\log\frac{\pi(x)}{u(x)} = KL(\pi\|u)$$

Thus $KL(\pi\|v) \leq KL(\pi\|u)$ which means $KL(\pi\|u) - KL(\pi\|v) = E_{Y\sim\pi}[KL_x(p(y,x)\| q(y,x))] \geq 0$