

On the spectral radius and stiffness of Markov jump process rate matrices

Peter Glynn & Alex Infanger

To cite this article: Peter Glynn & Alex Infanger (2020): On the spectral radius and stiffness of Markov jump process rate matrices, Stochastic Models, DOI: [10.1080/15326349.2020.1815546](https://doi.org/10.1080/15326349.2020.1815546)

To link to this article: <https://doi.org/10.1080/15326349.2020.1815546>



Published online: 29 Sep 2020.



Submit your article to this journal [↗](#)



Article views: 43




View related articles [↗](#)



View Crossmark data [↗](#)



On the spectral radius and stiffness of Markov jump process rate matrices

Peter Glynn  and Alex Infanger 

Stanford University, Stanford, CA, USA

ABSTRACT

It is well known that the numerical stability of many finite difference time-stepping algorithms for solving the Kolmogorov differential equations for Markov jump processes depends on the magnitude of the spectral radius of the rate matrix. In this paper, we develop bounds on the spectral radius that rigorously establish that the spectral radius typically scales in proportion to the maximal jump rate. Our analysis also provides rigorous bounds on the stiffness of the rate matrix, when the process is reversible.

ARTICLE HISTORY

Received 25 December 2019
Accepted 24 August 2020

KEYWORDS

Markov jump process;
rate matrices; bounds;
spectral radius; stiffness
ratio; Kolmogorov
differential equations

1. Introduction

Let $X = (X(t) : t \geq 0)$ be an irreducible Markov jump process with finite state space S and rate matrix $Q = (Q(x, y) : x, y \in S)$. It is well known that 0 is an eigenvalue of Q , and that the remaining eigenvalues cannot have positive real parts; see the proof of Proposition 2.9, Anderson^[2], for example. A great deal of effort has been expended in the literature on obtaining bounds on the spectral gap, defined informally as the distance between 0 and the next largest eigenvalue. A more careful discussion will be provided in Section 2.

In this paper, our main concern will be the study of the spectral radius ρ of Q , namely the largest modulus amongst the eigenvalues of Q . This quantity plays a key role in studying numerical methods for solving the Kolmogorov forwards and backwards equations for X . In particular, suppose we wish to solve the forwards equations

$$\mu'(t) = \mu(t)Q \quad (1.1)$$

subject to $\mu(0) = \mu_0$, for the unknown (row vector) probability solution $(\mu(t) : t \geq 0)$, given an initial (row vector) distribution μ_0 . If we implement the forwards Euler time-stepping algorithm to compute $(\mu(t) : t \geq 0)$, then $\mu'(t)$ is replaced by a forward finite difference, namely

$$\frac{\mu_h((k+1)h) - \mu_h(kh)}{h} = \mu_h(kh)Q,$$

for $k \geq 0$, where h is the time-increment. Consequently,

$$\mu_h((k+1)h) = \mu_h(kh)(I + hQ)$$

for $k \geq 0$, so that

$$\mu_h(nh) = \mu_0(I + hQ)^n \tag{1.2}$$

for $n \geq 0$. We note that the discretized solution $(\mu_h(nh) : n \geq 0)$ blows up if $I + hQ$ has eigenvalues with a complex modulus greater than 1. In particular, if there exists an eigenvalue λ of Q for which $|\lambda| > 2/h$, (1.2) will blow up as $n \rightarrow \infty$. So, the magnitude of the spectral radius of Q plays a key role in the numerical stability of the forwards Euler method for solving (1.1). Furthermore, the stability of the widely used explicit Runge-Kutta methods depends upon whether $|f(\lambda_i)| < 1$ for each eigenvalue λ_i of hQ , where $f(\cdot)$ is a polynomial that can be associated to the specific method. Having bounds on the magnitude of the eigenvalues of hQ is therefore of significant interest in assessing the stability of such schemes. The location of the eigenvalues also plays a role in explicit linear multistep methods; see Theorem 1.2 on p. 241 in Hairer and Wanner^[9] and also Jeltsch and Nevanlinna^[11].

We note, therefore, that for many methods, a large value of ρ necessitates taking small time steps, thereby increasing the computational time needed to solve the Kolmogorov equations over a given time horizon. This is closely connected to the notion of stiffness for the corresponding differential equation (1.1). A ‘‘folk result’’ in the literature on numerical solvers for the Kolmogorov equations is that stiffness tends to arise when jump processes have widely varying jump rates; see, for example, Section 3, Clarotti^[5], p. 21, Reibman and Trivedi^[16], and Section 3.2, Malhotra et al.^[14].

As far as we are aware, this paper is the first to develop bounds on ρ , with the goal of relating ρ , and the closely related real spectral spread (to be defined in Section 2), to jump rates and other problem-specific data of the underlying stochastic model. Section 2 is concerned with developing upper and lower bounds on ρ that hold for general jump processes, whereas Section 3 develops bounds for reversible jump processes, specifically for birth-death models. We also establish there that the rate matrices for many-server queues are provably stiff when the number of servers is large.

2. The spectrum of a general rate matrix

A rate matrix $Q = (Q(x, y) : x, y \in S)$ has non-negative off-diagonal entries, with row sums that are zero. The quantity $\eta(x) = \Delta - Q(x, x)$ has the

interpretation as the *jump rate* out of state $x \in S$. We set $\eta^* = \max\{\eta(x) : x \in S\}$, so that η^* is the maximum jump rate for Q .

For any function $v : S \rightarrow \mathbb{C}$, we can encode $v = (v(x) : x \in S)$ as a column vector. Similarly, given a probability mass function $\nu = (\nu(x) : x \in S)$, we may encode ν as a row vector. Given a function v , let $\|v\|_p$ be the p -norm defined by $\|v\|_p = (\sum_{x \in S} |v(x)|^p)^{1/p}$ for $1 \leq p < \infty$ and $\|v\|_p = \max\{|v(x)| : x \in S\}$ for $p = \infty$. For a square matrix A , we can then define the induced matrix norm $\|A\|_p$ via

$$\|A\|_p = \sup \left\{ \frac{\|Av\|_p}{\|v\|_p} : \|v\|_p \neq 0 \right\} \quad (2.1)$$

for $p \geq 1$. Note that if λ is an eigenvalue of A with associated eigenvector v , it follows from the definition (2.1) that

$$|\lambda| \|v\|_p = \|Av\|_p \leq \|A\|_p \|v\|_p, \quad (2.2)$$

so that $|\lambda| \leq \|A\|_p$ for $p \geq 1$. Finally, for $p = \infty$, it is well-known that $\|A\|_\infty$ can be computed explicitly, namely

$$\|A\|_\infty = \max_x \sum_{y \in S} |A(x, y)|; \quad (2.3)$$

see p. 72, Golub and Van Loan^[8], for example.

Let $\mathcal{S} = \{\lambda \in \mathbb{C} : \lambda \text{ is an eigenvalue of } Q\}$ be the *spectrum* of Q . As a consequence of (2.3),

$$\begin{aligned} \|\eta^*I + Q\|_\infty &= \sup_{x \in S} \left\{ |\eta^* - \eta(x)| + \sum_{y \neq x} |Q(x, y)| \right\} \\ &= \sup_{x \in S} \left\{ \eta^* - \eta(x) + \sum_{y \neq x} Q(x, y) \right\} \\ &= \sup_{x \in S} \{\eta^* - \eta(x) + \eta(x)\} = \eta^*, \end{aligned}$$

so that the eigenvalues of $\eta^*I + Q$ are contained in $\{w \in \mathbb{C} : |w| \leq \|\eta^*I + Q\|_\infty\} = \{w \in \mathbb{C} : |w| \leq \eta^*\}$. Hence $\mathcal{S} \subseteq \{w \in \mathbb{C} : |w + \eta^*| \leq \eta^*\} \triangleq \mathcal{D}$.

Recall that $0 \in \mathcal{S}$ and $e = (1, 1, \dots, 1)^T$ is a column eigenvector associated with eigenvalue 0, while $\pi = (\pi(x) : x \in S)$ is its associated row eigenvector, where π is an equilibrium distribution of X ; see p. 343, Brémaud^[4]. Because $\mathcal{D} \cap \{w \in \mathbb{C} : \operatorname{Re}(w) \geq 0\} = \{0\}$, it follows that $\operatorname{Re}(\lambda) < 0$ for $0 \neq \lambda \in \mathcal{S}$. Furthermore, $|w| \leq 2\eta^*$ for $w \in \mathcal{D}$. We summarize our discussion with the following result.

Theorem 2.1. *If Q is a rate matrix, then:*

- a) $\mathcal{S} \subseteq \{w \in \mathbb{C} : |w + \eta^*| \leq \eta^*\}$
- b) $|\lambda| \leq 2\eta^*$ for $\lambda \in \mathcal{S}$
- c) $0 \in \mathcal{S}$
- d) $Re(\lambda) < 0$ for $0 \neq \lambda \in \mathcal{S}$.

Remark 1. Let $\rho = \max\{|\lambda| : \lambda \in \mathcal{S}\}$ be the *spectral radius* of Q and let $r = \max\{|Re(\lambda)| : \lambda \in \mathcal{S}\}$ be the (*real*) *spectral spread* of Q . Note that $r \leq \rho$. **Theorem 2.1** establishes an upper bound on r , namely $2\eta^*$. This upper bound can be attained. In particular if

$$Q = \begin{pmatrix} -\mu & \mu \\ \mu & -\mu \end{pmatrix},$$

then $\mathcal{S} = \{-2\mu, 0\}$ so that $\rho = r = 2\mu$.

Remark 2. The inclusion $\mathcal{S} \subseteq \mathcal{D}$ is precisely the same inclusion as would be obtained by applying the Gershgorin circle theorem to the rows of Q ; see Section 7.2.1 of Golub and Van Loan^[8] for a discussion of Gershgorin's circle theorem.

Remark 3. Note that Q and Q^T have the same spectrum. So, (2.2) implies that $\| \eta^* I + Q^T \|_\infty$ is another easily computable upper bound on r .

If ν_1, ν_2 are two probability vectors on S , the total variation distance between ν_1 and ν_2 , denoted, $\|\nu_1 - \nu_2\|_{tv}$ is given by $\|\nu_1 - \nu_2\|_1/2$. We write $P_\mu(\cdot) = \sum_{x \in S} \mu(x) P(\cdot | X(0) = x)$ and let $P(t, x, y) = P(X(t) = y | X(0) = x)$ for $t \geq 0, x, y \in S$. If $P(t)$ is the matrix $(P(t, x, y) : x, y \in S)$, it is well-known that

$$P(t) = \exp(Qt) \tag{2.4}$$

for $t \geq 0$; see p. 339, Brémaud^[4].

The following result describes the role of the spectral gap for Markov jump processes. We provide a quick proof for the general case, when X need not be reversible.

Proposition 2.1. *Suppose that Q is an irreducible rate matrix, and let $\gamma = \max\{Re(\lambda) : 0 \neq \lambda \in \mathcal{S}\}$. Then, for any probability vector μ on S ,*

$$\overline{\lim}_{t \rightarrow \infty} \frac{1}{t} \log (\|P_\mu(X(t) \in \cdot) - \pi(\cdot)\|_{tv}) \leq \gamma. \tag{2.5}$$

Furthermore, there exists a probability vector ν on S such that

$$\underline{\lim}_{t \rightarrow \infty} \frac{1}{t} \log (\|P_\nu(X(t) \in \cdot) - \pi(\cdot)\|_{tv}) \geq \gamma. \quad (2.6)$$

In view of the above result, we may refer to γ as the rate of convergence of the jump process to its equilibrium. The quantity $|\gamma|$ is called the *spectral gap* of the process. The great majority of the literature on the eigenvalues of a rate matrix Q focuses on bounding $|\gamma|$. Many bounds on the spectral gap have appeared in the literature on reversible processes, with some results also extending to nonreversible processes; see, for example, Diaconis and Stroock^[6], Montenegro and Tetali^[15].

Proof of Proposition 2.1. Because $I + Q/\eta^*$ is an irreducible stochastic matrix, its eigenvalue 1 has algebraic (and geometric) multiplicity 1; see p. 64, Gantmacher^[7]. Consequently, 0 has algebraic (and geometric) multiplicity 1 for the matrix Q . For the other eigenvalues, they may have algebraic and geometric multiplicity greater than 1. Using the Jordan form of Q in (2.4) we find (see also p. 133 in Hirsch et al.^[10])

$$P(t, x, y) = e(x)\pi(y) + \sum_{0 \neq \lambda \in \mathcal{S}} O(t^{\tilde{d}(\lambda)} e^{\lambda t}) \quad (2.7)$$

as $t \rightarrow \infty$, where $\tilde{d}(\lambda) + 1$ is the maximal degree of nilpotency associated with λ 's Jordan block(s) and $O(h(t))$ is a function for which $O(h(t))/h(t)$ remains bounded as $t \rightarrow \infty$. From (2.7), we see that

$$\|P_\mu(X(t) \in \cdot) - \pi(\cdot)\|_{tv} = O(t^{\tilde{d}} e^{\gamma t})$$

as $t \rightarrow \infty$, where $\tilde{d} = \max\{\tilde{d}(\lambda) : \operatorname{Re}(\lambda) = \gamma\}$. Hence (2.5) is immediate.

For (2.6), let ζ be a (possibly complex-valued) row eigenvector of Q associated with any eigenvalue λ of Q having real part γ . Write $\zeta = \alpha + i\omega$. Note that ζe must vanish, since any row eigenvector of $0 \neq \lambda \in \mathcal{S}$ must be orthogonal to the column eigenvector e associated with eigenvalue 0; see, for example, Theorem 26 of Brauer^[3]. Hence, $\alpha e = 0 = \omega e$. At least one of α and ω must be non-zero. Suppose it is α . (A similar argument works if $\omega \neq 0$.) Since Q is irreducible, π is strictly positive and so $\nu = \pi + \delta\alpha$ must be stochastic for δ sufficiently small and positive.

Note that

$$P_\nu(X(t) = y) = (\nu e^{Qt})(y) \quad (2.8)$$

for $y \in S$, and

$$\nu e^{Qt} - \pi = \delta\alpha e^{Qt}.$$

Since ζ is an eigenvector of Q associated with eigenvalue λ ,

$$\zeta e^{Qt} = e^{\lambda t} \zeta \quad (2.9)$$

for $t \geq 0$. Taking the real parts of both sides of (2.9), we find that

$$\alpha e^{Qt} = e^{\gamma t} (\alpha \cos(\operatorname{Im}(\lambda)t) - \omega \sin(\operatorname{Im}(\lambda)t))$$

for $t \geq 0$. If $\operatorname{Im}(\lambda) = 0$ (so that λ is real), (2.8)-(2.9) imply that

$$\lim_{t \rightarrow \infty} \frac{1}{t} \log (\|P_\nu(X(t) \in \cdot) - \pi(\cdot)\|_{tv}) = \gamma.$$

If $\operatorname{Im}(\lambda) \neq 0$, we send $t \rightarrow \infty$ through multiples of $2\pi/|\operatorname{Im}(\lambda)|$, and conclude from (2.8)-(2.9) that

$$\underline{\lim}_{t \rightarrow \infty} \frac{1}{t} (\log \|P_\nu(X(t) \in \cdot) - \pi(\cdot)\|_{tv}) \geq \gamma,$$

thereby proving (2.6). □

Because our upper bound $2\eta^*$ on r is tight (without further assumptions), we now focus on lower bounds for r . We recall that for any square $d \times d$ matrix A , the *trace* of A , denoted as $\operatorname{tr}(A)$, is given by $\operatorname{tr}(A) = \sum_{x \in S} A(x, x)$. An important fact regarding the trace is that $\operatorname{tr}(A) = \sum_{i=1}^d \lambda_i$ where we repeat the eigenvalue $\lambda_i \in \mathcal{S}$ according to its algebraic multiplicity; see p. 348, Golub and Van Loan^[8]. Set

$$\bar{\eta} = \frac{1}{d} \sum_{x \in S} \eta(x).$$

Theorem 2.2. *For any rate matrix Q , the spectral spread r and the spectral radius ρ are lower bounded by $\bar{\eta}$.*

Proof. Observe that

$$\operatorname{tr}(Q) = - \sum_{x \in S} \eta(x) = \sum_{i=1}^d \operatorname{Re}(\lambda_i),$$

so that there exists at least one eigenvalue $\lambda_i \in \mathcal{S}$ for which

$$\operatorname{Re}(\lambda_i) \leq -\bar{\eta}.$$

It follows that $|\operatorname{Re}(\lambda_i)| \geq \bar{\eta}$, proving that $r \geq \bar{\eta}, \rho \geq \bar{\eta}$. □

Theorem 2.3. *If Q is a rate matrix with $Q \neq 0$, then*

$$r \geq \frac{\operatorname{tr}(Q^2)}{\operatorname{tr}(-Q)}.$$

Proof. For each eigenvalue λ_j , we write $\lambda_j = a(\lambda_j) + ib(\lambda_j)$, so that $a(\lambda_j)$ and $b(\lambda_j)$ are the real and imaginary parts of the eigenvalues of λ_j . Then,

$$\operatorname{tr}(Q^2) = \sum_{j=1}^d \operatorname{Re}(\lambda_j^2) = \sum_{j=1}^d (a^2(\lambda_j) - b^2(\lambda_j)) \leq \sum_{j=1}^d a^2(\lambda_j).$$

But

$$\sum_{j=1}^d a^2(\lambda_j) \leq \max_{1 \leq k \leq d} |a(\lambda_k)| \sum_{j=1}^d |a(\lambda_j)| \leq r \operatorname{tr}(-Q).$$

□

Remark 4. We note that when Q has real eigenvalues, $\operatorname{tr}(Q^2) = \sum_{j=1}^d \lambda_j^2$ and $\operatorname{tr}(Q) = \sum_{j=1}^d \lambda_j$. The Cauchy-Schwarz inequality implies that

$$d \sum_{j=1}^d \lambda_j^2 \geq \left(\sum_{j=1}^d \lambda_j \right)^2$$

with strict inequality unless the λ_j 's are all identical. But 0 is an eigenvalue of Q , and at least one other eigenvalue is non-zero, since $\operatorname{tr}(Q) < 0$. Hence, $\operatorname{tr}(Q^2)/\operatorname{tr}(-Q) > \bar{\eta}$, so that [Theorem 2.3](#) is then a strict improvement of [Theorem 2.2](#).

We now provide an improved upper bound on r for rate matrices with real eigenvalues. Note that

$$\begin{aligned} \frac{1}{d}(r - \bar{\eta})^2 &\leq \frac{1}{d} \sum_{i=1}^d (\lambda_i - \bar{\eta})^2 \\ &= \frac{1}{d} \sum_{i=1}^d \lambda_i^2 - \bar{\eta}^2 \\ &= \frac{\operatorname{tr}(Q^2)}{d} - \left(\frac{\operatorname{tr}(Q)}{d} \right)^2 \\ &\triangleq \sigma^2. \end{aligned}$$

We therefore find that $\sigma^2/(r - \bar{\eta})^2 \geq \frac{1}{d}$, yielding the upper bound $r \leq \bar{\eta} + \sigma\sqrt{d}$. With more effort, one can improve the bound slightly to the following.

Proposition 2.2. For any $d \times d$ rate matrix Q with real eigenvalues,

$$r \leq \bar{\eta} + \sqrt{(d-1)\sigma^2}$$

where $\sigma^2 = d^{-1}\operatorname{tr}(Q^2) - (d^{-1}\operatorname{tr}(Q))^2$.

We refer to Wolkowicz and Styan^[17] for the proof. The presence of the factor $d-1$ in the square root means that [Proposition 2.2](#)'s bound typically becomes very loose when d is large.

$$\bar{\eta} \rightarrow \sum_{i=1}^s \mu_i,$$

$$\eta^* \rightarrow \sum_{i=1}^s \mu_i,$$

and

$$\frac{1}{|S|} \text{Tr}(Q^2) \rightarrow \left(\sum_{i=1}^s \mu_i \right)^2$$

as $n \rightarrow \infty$. Our bounds then yield the inequalities

$$\sum_{i=1}^s \mu_i \leq \underline{\lim}_{n \rightarrow \infty} r \leq \overline{\lim}_{n \rightarrow \infty} \rho \leq 2 \sum_{i=1}^s \mu_i.$$

In this example, [Theorem 2.3](#)'s lower bound offers no asymptotic improvement to that of [Theorem 2.2](#). Again, r and ρ are within a constant factor of $\bar{\eta}$ and η^* .

All the examples of this section have the property that $\bar{\eta}$ and η^* are within a constant factor of one another. This seems typical of most stochastic modeling applications. For rate matrices in which $\bar{\eta}$ and η^* are within a constant factor of one another, our theory establishes that r and ρ are then within a constant factor of both $\bar{\eta}$ and η^* .

3. Bounds for reversible rate matrices

An irreducible rate matrix Q with $|S| = d < \infty$ is said to be *reversible* if its unique equilibrium distribution $\pi = (\pi(x) : x \in S)$ satisfies the *detailed balance* relationship

$$\pi(x)Q(x, y) = \pi(y)Q(y, x)$$

for all $x, y \in S$. In this case, we can define $R = (R(x, y) : x, y \in S)$ via

$$R(x, y) = \sqrt{\frac{\pi(x)}{\pi(y)}} Q(x, y).$$

A key fact is that R is a symmetric matrix for which

$$Q = D^{-1/2} R D^{1/2},$$

where $D^{1/2}$ is the diagonal matrix in which the x 'th diagonal entry is $\pi(x)^{1/2}$. Because Q and R are similar matrices, they share the same spectrum \mathcal{S} . Furthermore, the spectrum of R (and hence Q) can consist only of real eigenvalues; see [Theorem 8.1.1](#) and its proof in [Golub and Van Loan](#)^[8], for example. This implies that

$$-\rho = -r = \min\{\lambda : \lambda \in \mathcal{S}\}.$$

We now analyze the spectral radius ρ by using the Rayleigh characterization of R 's eigenvalues. In particular, $\min\{\lambda : \lambda \in \mathcal{S}\}$ is given by

$$\min\{\lambda : \lambda \in \mathcal{S}\} = \inf_{v \neq 0} \frac{v^T R v}{v^T v};$$

see Theorem 8.1.2 of Golub and Van Loan^[8], for example. Hence, for any choice of column vector $u_0 \neq 0$, we obtain the lower bound

$$\rho \geq \frac{u^T R u}{u^T u} \quad (3.1)$$

on ρ . Furthermore, we have equality in (3.1) if and only if u is an eigenvector with eigenvalue ρ (a consequence of Theorem 8.1.1 in Golub and Van Loan^[8]).

The easiest choice for u is $u_x(w) = \delta_{xw}$ for $w \in S$. In this case, (3.1) takes the form

$$\rho \geq \eta(x),$$

yielding the lower bound

$$\rho \geq \max_{x \in S} \eta(x) = \eta^*.$$

If Q is irreducible, then $\eta(x) > 0$ for each $x \in S$, and there exists $w_x \neq x$ for each x so that $Q(w_x, x) > 0$. Hence, for each $x \in S$, $(Q u_x)(w_x) = Q(w_x, x) \neq 0 = -\eta(x) u_x(w_x)$, so that u_x can not be an eigenvector associated with $-\rho$. Consequently, $\rho > \eta^*$ and we have proved the following result.

Theorem 3.1. If Q is an irreducible and reversible rate matrix, then

$$\rho = r > \eta^*.$$

Hence, for reversible rate matrices, ρ and r must lie in $(\eta^*, 2\eta^*]$.

A better bound can be obtained with a better choice of u in (3.1). Here, we consider vectors u supported on two states of S , so that u is of the form

$$u_{x,y}(w) = c\delta_{xw} + d\delta_{yw}$$

for $c, d \in \mathbb{R}$ and $x \neq y$. In this case,

$$\begin{aligned} \rho = r &\geq - \inf_{c^2 + d^2 > 0} \frac{u_{x,y}^T R u_{x,y}}{u_{x,y}^T u_{x,y}} \\ &= - \inf_{\tilde{u} \neq 0} \frac{\tilde{u}^T \tilde{R}_{x,y} \tilde{u}}{\tilde{u}^T \tilde{u}}, \end{aligned} \quad (3.2)$$

where \tilde{u} is a 2×1 column vector and $\tilde{R}_{x,y}$ is the 2×2 matrix

$$\tilde{R}_{x,y} = \begin{pmatrix} -\eta(x) & R(x,y) \\ R(y,x) & -\eta(y) \end{pmatrix}.$$

Since $\tilde{R}_{x,y}$ is symmetric, the minimum and maximum of the ratio in (3.2) is attained at the two eigenvalues of $\tilde{R}_{x,y}$ (via Rayleigh's characterization). The eigenvalues γ_1 and γ_2 of the matrix $\tilde{R}_{x,y}$ can be explicitly computed. We find that

$$\begin{aligned} \gamma_1 &= -\left(\frac{\eta(x) + \eta(y)}{2}\right) - \frac{1}{2} \sqrt{(\eta(x) - \eta(y))^2 + 4(R(x,y))^2}, \\ \gamma_2 &= -\left(\frac{\eta(x) + \eta(y)}{2}\right) + \frac{1}{2} \sqrt{(\eta(x) - \eta(y))^2 + 4(R(x,y))^2}. \end{aligned}$$

We have therefore proved the following result.

Theorem 3.2. If Q is an irreducible and reversible rate matrix, then

$$\rho = r \geq \max_{\substack{x,y \in S \\ x \neq y}} \left(\frac{\eta(x) + \eta(y)}{2} + \frac{1}{2} \sqrt{(\eta(x) - \eta(y))^2 + 4 \frac{\pi(x)}{\pi(y)} (Q(x,y))^2} \right).$$

In particular, if Q corresponds to an irreducible birth-death process on $\{0, 1, \dots, m\}$ with birth rates $\alpha_i (0 \leq i < m)$ and death rates $\beta_i (0 < i \leq m)$, we find that

$$\rho = r \geq \max_{1 \leq i \leq m-1} \left(\frac{\alpha_i + \alpha_{i-1}}{2} + \frac{\beta_i + \beta_{i+1}}{2} + \frac{1}{2} \sqrt{(\alpha_i + \beta_i - \alpha_{i+1} - \beta_{i+1})^2 + 4\alpha_i\beta_{i+1}} \right). \quad (3.3)$$

We now apply these bounds to the examples from Section 2 corresponding to birth-death processes.

Example 1 (continued). Here we find that (3.3) translates into

$$\liminf_{m \rightarrow \infty} r \geq \left(\sqrt{\lambda} + \sqrt{\mu} \right)^2 - \sqrt{\lambda\mu}.$$

Example 2 (continued). In the setting of our asymptotic regime, for this example,

$$\liminf_{m \rightarrow \infty} \frac{r}{m} \geq \mu \left(\phi + 1 + \sqrt{\phi} \right). \quad (3.4)$$

It is a simple exercise to establish that the right-hand side of (3.4) is larger than the lower bound in Section 2.

Example 3 (*continued*). As in [Example 2](#), the bound

$$\liminf_{m \rightarrow \infty} \frac{r}{m} \geq \mu \left(\phi + 1 + \sqrt{\phi} \right)$$

holds in the corresponding asymptotic regime for this example.

This paper has thus far been primarily concerned with deriving upper and lower bounds on the spectral radius ρ and the real spectral spread r of Q . As discussed in the introduction, the quantity ρ provides valuable information on the time step needed to guarantee that explicit time-stepping methods will remain numerically stable.

We conclude this section with a discussion of a measure that is intended to indicate how challenging the discretization of [\(1.1\)](#) is likely to be, as a function of the problem instance Q . We note that if we re-scale time in [\(1.1\)](#) (by a factor τ , say) then Q is replaced by τQ , thereby increasing the spectral radius by a factor of τ . We seek a measure that is independent of such a time-rescaling. Such a measure is given by $\rho/|\gamma|$ or $r/|\gamma|$, where $|\gamma|$ is the spectral gap defined in [Section 2](#). We refer to the ratio $r/|\gamma|$ as the *stiffness ratio* of Q ; see Aiken^[1] for further discussion of stiffness. Some papers focus instead on the spectral radius in measuring stiffness, (e.g., using the ratio $\rho/|\gamma|$ rather than $r/|\gamma|$ or related measures – see for example, Clarotti^[5]). In this section, these notions coincide because all the eigenvalues of reversible processes are real. Furthermore, our bounds of [Section 2](#) establish that r and ρ are often within a constant factor of one another. Consequently the two stiffness ratios are often within a constant factor of one another. We choose, in our remaining discussion of stiffness, to frame stiffness in terms of $r/|\gamma|$.

An upper bound on this stiffness ratio is given by $2\eta^*/\ell$, where ℓ is a lower bound on the spectral gap $|\gamma|$. Such lower bounds have been studied extensively; see, for example, Diaconis and Stroock^[6] and Montenegro and Tetali^[15]. A lower bound on the stiffness ratio is obtained from \tilde{r}/u , where \tilde{r} is a lower bound on the real spectral spread r and u is an upper bound on the spectral gap. As we have already discussed a number of lower bounds \tilde{r} on r , we will focus our attention on the upper bound on $|\gamma|$.

We again employ the Rayleigh quotient, this time making use of the well-known minimax characterization of γ ,

$$\gamma = \max_{\substack{v^T u = 0 \\ v \neq 0}} \frac{v^T R v}{v^T v} \tag{3.5}$$

where u is the column eigenvector $u = (\sqrt{\pi(x)} : x \in S)$ for the matrix R associated with eigenvalue 0. (This is a special case of [Theorem 8.1.2](#), the

Courant-Fisher minimax theorem, in Golub and Van Loan^[8].) Characterization (3.5) of γ is sometimes called the variational characterization of γ . With model-specific choices of the “test vector” v , Landau and Odlyzko^[12] have produced a lower bound on γ for the random walk on the barbell graph, while Diaconis and Stroock^[6] found a lower bound for the random walk on the full binary tree of finite depth. We consider v of the form $v = v_{x,y}$, where $v_{x,y}(w) = c\delta_{xw} + d\delta_{yw}$ for $c, d \in \mathbb{R}$ and $x \neq y$. In view of the fact that $v_{x,y}^T \mu = 0$, we find that

$$c = \sqrt{\frac{\pi(y)}{\pi(x) + \pi(y)}}, \quad d = -\sqrt{\frac{\pi(x)}{\pi(x) + \pi(y)}}$$

if we additionally require that $v_{x,y}^T v_{x,y} = 1$.

The ratio (3.5) then specializes to

$$\frac{v_{x,y}^T R v_{x,y}}{v_{x,y}^T v_{x,y}} = -\left(\frac{\pi(y)\eta(x) + \pi(x)\eta(y) + 2\pi(x)Q(x,y)}{\pi(x) + \pi(y)}\right).$$

We therefore have the following upper bound on $|\gamma|$.

Theorem 3.3. Suppose that Q is an irreducible and reversible rate matrix. Then,

$$|\gamma| \leq \min_{\substack{x,y \in S \\ x \neq y}} \left(\frac{\pi(y)\eta(x) + \pi(x)\eta(y) + 2\pi(x)Q(x,y)}{\pi(x) + \pi(y)}\right).$$

We now specialize this upper bound on the spectral gap to the birth-death setting. By choosing $y = x + 1$, we find that

$$|\gamma| \leq \alpha_x + \beta_{x+1} + \frac{\alpha_x \beta_x + \alpha_{x+1} \beta_{x+1}}{\alpha_x + \beta_{x+1}}.$$

If we apply this to Examples 1 through 3 with $x = 0$, we find that

$$|\gamma| \leq \lambda + \mu + \frac{2\lambda\mu}{\lambda + \mu}. \quad (3.6)$$

Remark 5. Alternative upper bounds on $|\gamma|$ exist in the literature. For example, a Cheeger-like inequality for Markov chains (see p.52, Diaconis and Stroock^[6], for example) establishes that for birth-death processes,

$$|\gamma| \leq \frac{2\pi(x)\beta_x}{\sum_{y \geq x} \pi(y)} \quad (3.7)$$

provided that we select x so that $\sum_{y \geq x} \pi(y) \leq 1/2$. To see how this bound compares to (3.6), we consider Example 1. In that setting, (3.7) translates into

$$|\gamma| \leq \frac{2\mu\pi(x)}{\sum_{y \geq x} \pi(y)}.$$

When $\lambda < \mu$, $\pi(x) / \sum_{y \geq x} \pi(y) \rightarrow 1 - \lambda/\mu$ as $m \rightarrow \infty$, so that the Cheeger upper bound on $|\gamma|$ becomes $2(\mu - \lambda)$ in this setting. Our upper bound (3.6) is superior when λ is small, while (3.7) is better in the “heavy traffic” setting.

Given the lower bounds on ρ developed earlier, we conclude that the stiffness ratios for Examples 2 and 3 grow at rate m in the asymptotic regimes described there. So, it follows that many-server queues are inherently numerically challenging, at least in so far as solving the Kolmogorov equations (1.1) is concerned.

Acknowledgements

The authors would like to thank the referees for their careful reading of the paper, and excellent suggestions. The authors particularly appreciate the shorter proofs for Theorem 2.3 and the bound immediately preceding Proposition 2.2 that were supplied by one of the referees.

ORCID

Peter Glynn  <http://orcid.org/0000-0003-1370-6638>

Alex Infanger  <http://orcid.org/0000-0001-5873-5382>

References

- [1] Aiken, R. C. 1985. *Stiff Computation*, volume 169. Oxford: Oxford University Press.
- [2] Anderson, W. J. 2012. *Continuous-Time Markov Chains: An Applications-Oriented Approach*; New York: Springer Science & Business Media.
- [3] Brauer, A. Limits for the Characteristic Roots of a Matrix. IV: Applications to Stochastic Matrices. *Duke Math. J.* 1952, 19, 75–91. DOI: 10.1215/S0012-7094-52-01910-8.
- [4] Brémaud, P. 1999. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues, Volume 31*; New York: Springer Science & Business Media.
- [5] Clarotti, C. A. 1986. The Markov Approach to Calculating System Reliability: Computational Problems. In Serra, A., Barlow, R. E., Eds.; *Proc. of the International School of Physics, Course XCIV*, Amsterdam: North-Holland Physics Publishing; 55–66.
- [6] Diaconis, P.; Stroock, D. Geometric Bounds for Eigenvalues of Markov Chains. *Ann. Appl. Probab.* 1991, 1, 36–61. DOI: 10.1214/aoap/1177005980.
- [7] Gantmacher, F. R. 1959. *The Theory of Matrices*, Volume 2; New York: Chelsea.
- [8] Golub, G. H.; Van Loan, C. F. 2013. *Matrix Computations*, 4th Edition; Baltimore, MD: The Johns Hopkins University Press.

- [9] Hairer, E.; Wanner, G. 2002. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Springer Series in Computational Mathematics. Verlag, Berlin, second revised edition.
- [10] Hirsch, M. W.; Devaney, R. L.; Smale, S. 1974. *Differential Equations, Dynamical Systems, and Linear Algebra, Volume 60*; New York: Academic Press.
- [11] Jeltsch, R.; Nevanlinna, O. Largest Disk of Stability of Explicit Runge-Kutta Methods. *BIT* 1978, 18, 500–502. DOI: [10.1007/BF01932030](https://doi.org/10.1007/BF01932030).
- [12] Landau, H. J.; Odlyzko, A. M. Bounds for Eigenvalues of Certain Stochastic Matrices. *Linear Algebra Appl.* 1981, 38, 5–15. DOI: [10.1016/0024-3795\(81\)90003-3](https://doi.org/10.1016/0024-3795(81)90003-3).
- [13] Ledermann, W.; Reuter, G. E. H. Spectral Theory for the Differential Equations of Simple Birth and Death Processes. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical and Physical Sciences* 1954, 246, 321–369.
- [14] Malhotra, M.; Muppala, J. K.; Trivedi, K. S. Stiffness-Tolerant Methods for Transient Analysis of Stiff Markov Chains. *Microelectron. Reliab.* 1994, 34, 1825–1841. DOI: [10.1016/0026-2714\(94\)90137-6](https://doi.org/10.1016/0026-2714(94)90137-6).
- [15] Montenegro, R.; Tetali, P. Mathematical Aspects of Mixing Times in Markov Chains. *Foundations and Trends in Theoretical Computer Science* 2006, 1, 237–354. DOI: [10.1561/0400000003](https://doi.org/10.1561/0400000003).
- [16] Reibman, A.; Trivedi, K. Numerical Transient Analysis of Markov Models. *Comput. Oper. Res.* 1988, 15, 19–36. DOI: [10.1016/0305-0548\(88\)90026-3](https://doi.org/10.1016/0305-0548(88)90026-3).
- [17] Wolkowicz, H.; Styan, G. P. H. Bounds for Eigenvalues Using Traces. *Linear Algebra Appl.* 1980, 29, 471–506. DOI: [10.1016/0024-3795\(80\)90258-X](https://doi.org/10.1016/0024-3795(80)90258-X).