Exponential Convergence Rates for Stochastically Ordered Markov Processes Under Perturbation *

Julia Gaudio, Saurabh Amin, and Patrick Jaillet

Abstract

In this technical note we find computable exponential convergence rates for a large class of stochastically ordered Markov processes. We extend the result of Lund, Meyn, and Tweedie (1996), who found exponential convergence rates for stochastically ordered Markov processes starting from a fixed initial state, by allowing for a random initial condition that is also stochastically ordered. Our bounds are formulated in terms of moment-generating functions of hitting times. To illustrate our result, we find an explicit exponential convergence rate for an M/M/1 queue beginning in equilibrium and then experiencing a change in its arrival or departure rates, a setting which has not been studied to our knowledge.
Keywords: Markov processes, Queueing

1 Introduction

This paper is concerned with parametrized stochastically ordered Markov processes. Consider, for example, a stable M/M/1 queue with service rate μ and arrival rate $\lambda < \mu$. For a fixed μ, let $\{X_t(\pi, \lambda)\}_{t\geq 0}$ be the queue-length process with service rate λ and initial distribution π. Then $X_t(\pi, \lambda)$ is stochastically increasing in λ, for all $t \geq 0$. That is,

$$\mathbb{P}\left(X_t(\pi, \lambda) \geq x\right) \leq \mathbb{P}\left(X_t(\pi, \lambda') \geq x\right)$$

for all $x \in \mathbb{Z}_+$ if $\lambda \leq \lambda'$. Similarly, $X_t(\pi, \mu)$ is stochastically decreasing in μ for fixed λ and π. The focus of our paper is to analyze the convergence of a parametrized stochastically ordered Markov process to its stationary distribution, when its initial state is distributed according to a stationary distribution for another parameter choice. This will be stated more precisely below.

The Markov process is described by its transition kernel and its initial distribution. We assume that the initial distribution is the stationary distribution associated with setting the parameter equal to r_0, and we let r be the parameter setting of the transition kernel. The parameter change happens once, at $t = 0$. In other words, if $r = r_0$, the process is always in equilibrium, and when $r \neq r_0$, the system starts in the equilibrium associated with r_0 and transitions over time to the one associated with r. The equilibrium distributions are denoted by $\pi(r_0)$ and $\pi(r)$. When $r \neq r_0$ we say the system is "perturbed." These Markov processes will be denoted by $X_t\left(r_0, r\right)$. We sometimes refer to the collection $\{X_t\left(r_0, r\right)\}_{r_0,r}$

*Accepted for publication to *Systems and Control Letters* on August 26, 2019.

as a "system." Note that there could be multiple parameters. For example, to study an M/M/1 queue starting in the stationary distribution associated with (λ_0, μ_0), operating under parameters (λ, μ), we would have $r_0 = (\lambda_0, \mu_0)$ and $r = (\lambda, \mu)$. In the M/M/1 setting, we say that $r = r_0$ if $\lambda_0 = \lambda$ and $\mu_0 = \mu$. Otherwise, $r_0 \neq r$. As in [9], we consider Markov processes that take value in $[0, \infty)$. In this paper, we consider the total variation distance between a parametrized continuous time Markov process and its stationary distribution. Recall the definition of total variation distance:

Definition 1. *The total variation distance between two measure P and Q on state space Ω is given by*

$$\|P - Q\|_{TV} = \sup_{A \subset \Omega} |P(A) - Q(A)| .$$

For a given random variable X, let $\mathcal{L}(X)$ denote the distribution law of X. We seek a convergence bound of the form

$$\|\mathcal{L}\left(X_t(r_0, r)\right) - \pi(r)\|_{\mathrm{TV}} \leq Ce^{-\alpha t}.$$

The value α is referred to as the "convergence rate."

Prior work in the area of the convergence of continuous-time Markov processes focuses on convergence assuming a particular deterministic initial state. However, this type of analysis is limiting, because the initial state of a process is often unknown. In situations where the initial state is unobservable, it may be more reasonable to assume a particular initial distribution rather than a particular initial state. Our extension of the result by [9] allows one to analyze a system in equilibrium that undergoes a perturbation of its parameters, pushing it towards another equilibrium. For example, one might wish to analyze the effect of a disruption on a queue of customers waiting for service. The bounds in this paper would allow one to study how quickly the queue length process reaches the new equilibrium after being perturbed.

We start by reviewing the existing literature on the convergence of stochastically ordered Markov processes, focusing on a paper by Lund, Meyn and Tweedie ([9]). We extend the result of [9], allowing the initial state of the system to be distributed according to a stationary distribution from the family of distributions parametrized by the system parameters. To illustrate the value of our result, we apply it to the analysis of perturbed M/M/1 queues. We also analyze a control system of parallel M/M/1 queues, in which the controller seeks to equalize the queue lengths in response to perturbations of the service rates. More importantly, our result applies to a broader class of Markov processes, namely any parametrized Markov process whose initial distribution is a stationary distribution.

2 Related work

Lund, Meyn, and Tweedie ([9]) establish convergence rates for nonnegative Markov processes that are stochastically ordered in their initial state, starting from a fixed initial state. Examples of such Markov processes include: M/G/1 queues, birth-and-death processes, storage processes, insurance risk processes, and reflected diffusions. We reproduce here the main theorem, Theorem 2.1 from [9], which will be extended in this paper.

Theorem 1. *([9]) Suppose that $\{X_t\}$ is a Markov process on $\Omega = [0, \infty)$ that is stochastically increasing in its initial state, with parameter setting r. Let $\tau_0(x)$ be the hitting time to zero* *of X_t given that $X_0 = x$, and let $\tau_0(\pi(r))$ be the hitting time to zero of X_t given that X_0 is distributed according to the stationary distribution $\pi(r)$.*

Let $\mathcal{L}(X_t(x,r))$ be the distribution law of X_t given that $X_0 = x$. If $\mathbb{E}\left[e^{\alpha\tau_0(x)}\right] < \infty$ for some $\alpha > 0$ and some $x > 0$, then

$$\left\|\mathcal{L}(X_t(x,r)) - \pi(r)\right\|_{TV} \leq \left(\mathbb{E}\left[e^{\alpha\tau_0(x)}\right] + \mathbb{E}\left[e^{\alpha\tau_0(\pi(r))}\right]\right) e^{-\alpha t} \tag{1}$$

for every $x \geq 0$ and $t \geq 0$.

The significance of this theorem is that it provides computable rates of convergence for a large class of Markov processes by relating the total variation distance from equilibrium to moment generating functions of hitting times to zero. We extend this result to the situation where X_0 is distributed according to the stationary distribution corresponding to a different parameter choice. The proof is analogous to the one given in [9] and is based on a coupling approach.

The second major result in [9] is to connect a drift condition to the convergence rate in (1), which is Theorem 2.2 (i) in [9], reproduced below:

Theorem 2. *([9]) Suppose that $\{X_t\}$ is a Markov process that is stochastically increasing in its initial state. Let $\mathcal{A}$ be the extended generator of the process. If there exists a drift function $V : \Omega \to [1, \infty)$ and constants $c > 0$ and $b < \infty$ such that for all $x \in \Omega$*

$$\mathcal{A}V(x) \leq -cV(x) + b\mathbb{1}_{\{0\}}(x) \tag{2}$$

then $\mathbb{E}\left[e^{c\tau_0(x)}\right] < \infty$ for all $x > 0$, which implies that (1) holds for $\alpha \leq c$.

We also connect Theorem 2 to our extension of Theorem 1. Theorem 1 is applied to several univariate systems in [9]: finite capacity stores, dam processes, diffusion models, periodic queues, and M/M/1 queues. Additionally, one multivariate system is considered in [9]: two M/M/1 queues in series.

The paper by Lund et al (1996) has inspired numerous related papers, some of which we reference here. Several directly apply the main results; for example Novak and Watson (2009) used Theorem 1 to derive the convergence rate of an M/D/1 queue. In a more applied work, Kiessler (2008) used the result of [9] to prove the convergence of an estimator for traffic intensity.

Other works build on the derivation of bounds for other processes, or more general bounds. For example, Liu et al (2008) applied the main theorem in order to bound the best uniform convergence rate for strongly ergodic Markov chains. Other processes studied are Langevin diffusions ([12]) and jump diffusions ([15]). Hou et al (2005) also used a coupling method, focusing on establishing subgeometric convergence rates. In related work to [4], Liu et al (2010) established subgeometric convergence rates via first hitting times and drift functions. Douc et al (2004) were able to generalize convergence bounds to time-inhomogeneous chains using coupling and drift conditions. Baxendale (2005) derives convergence bounds for geometrically ergodic Markov processes with an alternate approach to [9], though also using a drift condition.

Few papers allow for a random initial condition. Roberts and Tweedie (2000) found convergence bounds for stochastically ordered Markov processes with a random initial condition, allowing for no minimal reachable element. However, their bound is stated in terms of a drift condition, which may be challenging to verify because it requires finding a drift function. Rosenthal (2002) also derives a convergence bound for an initial distribution for more general chains, using drift and minorization conditions, via a coupling approach.

3 Main result

We begin with some definitions that we utilize in this paper. We let $\{X_t(r_0, r)\}$ denote the process governed by r with initial distribution corresponding to r_0. Similarly, we let $\{X_t(r)|X_0(r) = x\}$ denote the process governed by r with initial state x.

Definition 2. *A set A is said to be* increasing *if*

$$\forall x \in A, y \geq x \implies y \in A.$$

Definition 3. *For a family of nonnegative Markov processes $\{X_t(r_0, r)\}$ with transition kernel parametrized by r with starting stationary distribution parametrized by r_0, we say that X_t is* stochastically increasing in r_0 *if for all $t \geq 0$ and all increasing sets $A \subset \Omega$,*

$$\mathbb{P}\left(X_t(r_0, r) \in A\right) \leq \mathbb{P}\left(X_t(r_0', r) \in A\right)$$

whenever $r_0 \leq r_0'$. Note that for a univariate process, A is of the form $\{y \in \Omega : y \geq x\}$ for some x.

Definition 4. *Define $\tau_0(r_0, r)$ to be the hitting time to the zero state of $\{X_t(r_0, r)\}$. Similarly, define $\tau_0(x, r)$ to be the hitting time to the zero state of $\{X_t(r)|X_0(r) = x\}$ For a Markov process $\{X_t(r_0, r)\}$, let*

$$G(r_0, r, \alpha) = \mathbb{E}\left[e^{\alpha \tau_0(r_0, r)}\right]$$

and similarly for a Markov process $\{X_t(r)|X_0(r) = x\}$, define $G(x, r, \alpha) = \mathbb{E}\left[e^{\alpha \tau_0(x, r)}\right]$.

We now extend Theorem 1 to allow for a random initial condition.

Theorem 3. *Consider a family of nonnegative Markov processes $\{X_t(r_0, r)\}$ that is stochastically increasing in r, where $r = r_0$ corresponds to the system being in equilibrium. Let $r_m = \max\{r_0, r\}$. If $G(r_m, r, \alpha) < \infty$ for some $\alpha > 0$, then*

$$\|\mathcal{L}\left(X_t(r_0, r)\right) - \pi(r)\|_{TV} \leq G(r_m, r, \alpha) e^{-\alpha t} \tag{3}$$

Proof. Note that $X_t(r, r) \stackrel{\mathrm{d}}{=} \pi(r)$. Using the coupling inequality, we have

$$\|\mathcal{L}\left(X_t(r_0, r)\right) - \pi(r)\|_{\mathrm{TV}} \leq \mathbb{P}\left(X_t(r_0, r) \neq X_t(r, r)\right)$$

where $(X_t(r_0, r), X_t(r, r))$ is any coupling.
Either $\{X_t(r_m, r)\} \stackrel{\mathrm{d}}{=} \{X_t(r_0, r)\}$ or $\{X_t(r_m, r)\} \stackrel{\mathrm{d}}{=} \{X_t(r, r)\}$. We can create copies $X_t(r_0, r)'$, $X_t(r, r)'$ so that $X_t(r_m, r) = X_t(r_0, r)' \geq X_t(r, r)'$ if $r_m = r_0$, and $X_t(r_m, r) = X_t(r, r)' \geq X_t(r_0, r)'$ if $r_m = r$. This is possible by an extension of Strassen's Theorem to stochastic processes, developed in [5] and as cited by [9]. We take $(X_t(r_0, r)', X_t(r, r)')$ as the coupling. Then, the process $X_t(r_m, r)$ acts as a bounding process. Observe that

$$\{X_t(r_m, r) = 0\} \implies \{X_t(r_0, r) = X_t(r, r) = 0\}$$

and the coupling occurs at or before time t. So then we have

$$\mathbb{P}\left(X_t(r_0, r) \neq X_t(r, r)\right) \leq \mathbb{P}\left(\tau_0\left(r_m, r\right) > t\right).$$

Exponentiating and using Markov's inequality, we obtain the desired result:

$$\begin{aligned} \|\mathcal{L}\left(X_t(r_0, r)\right) - \pi(r)\|_{\mathrm{TV}} &\le \mathbb{P}\left(e^{\alpha\tau_0(r_m, r)} > e^{\alpha t}\right) \text{ for } \alpha > 0 \\ &\le \mathbb{E}\left[e^{\alpha\tau_0(r_m, r)}\right] e^{-\alpha t} \\ &= G(r_m, r, \alpha)e^{-\alpha t} \end{aligned}$$

□

However, the challenge in applying Theorem 3 is finding $\alpha > 0$ for which $G(r_m, r, \alpha)$ is finite. Note that $G(r_m, r, \alpha)$ is a moment generating function (MGF), so $\{\alpha : G(r_m, r, \alpha) < \infty\}$ is an interval containing zero, typically referred to as the *domain*. For some Markov processes, the domain is precisely known. One such example is the M/M/1 queue with fixed service rate, where the arrival rate is perturbed from $r_0 = \lambda_0$ to $r = \lambda$. For processes where the domain is difficult to find but $r_m = r$, we can apply Theorem 2.

Corollary 1. *If $r_m = r$ and the drift condition (2) holds for a Markov process $X_t(x, r)$ with some $V(\cdot), b, c$, then (3) holds with $\alpha = c$.*

Proof. If the drift condition holds then $G(x, r, \alpha) < \infty$, by Theorem 2. Applying Lemma 3.1 from [9], we also have that $G(r, r, c) < \infty$. □

We now apply Theorem 3 to the analysis of a single M/M/1 queue.

4 M/M/1 queues

4.1 Queue length process

We study the queue length process and consider perturbing the arrival and service rates from $r_0 = (\lambda_0, \mu_0)$ to $r = (\lambda, \mu)$. Throughout, we assume the stability conditions $\lambda_0 < \mu_0$ and $\lambda < \mu$. First we consider the case of changing the arrival rate while keeping the service rate fixed. We then show how to find bounds for any change of the two parameters, as long as the service rate is greater than the arrival rate.

Suppose that $\mu = \mu_0$. The two processes are then $X_t\left((\lambda_0, \mu_0), (\lambda, \mu_0)\right)$ and $X_t\left((\lambda, \mu_0), (\lambda, \mu_0)\right)$. For clarity of presentation, we omit the service rate in the notation, and refer to the two processes as $X_t(\lambda_0, \lambda)$ and $X_t(\lambda, \lambda)$, respectively. Let $\lambda_m = \max\{\lambda_0, \lambda\}$. From Theorem 3, we have

$$\|\mathcal{L}\left(X_t(\lambda_0, \lambda)\right) - \pi(\lambda)\|_{\mathrm{TV}} \le G(\lambda_m, \lambda, \alpha)e^{-\alpha t}. \tag{4}$$

Let us analytically compute $G(\lambda_m, \lambda, \alpha)$. Let $\tau_y(x, \lambda)$ be the hitting time to y of the M/M/1 queue with parameters set to (λ, μ), started from a queue length of x, and write

$$G(x, \lambda, \alpha) = \mathbb{E}\left[e^{\alpha\tau_0(x)}\right].$$

Then by conditioning on the initial state, we obtain

$$G(\lambda_m, \lambda, \alpha) = \mathbb{E}\left[e^{\alpha\tau_0(\lambda_m, \lambda)}\right]$$

$$= \sum_{x=0}^{\infty} \left(1 - \frac{\lambda_m}{\mu}\right) \left(\frac{\lambda_m}{\mu}\right)^x G(x, \lambda, \alpha)$$

Now by decomposing the hitting time and noting the independence and stationarity of the incremental hitting times,

$$\begin{aligned} G(x, \lambda, \alpha) &= \mathbb{E}\left[e^{\alpha \tau_0(x,\lambda)}\right] \\ &= \mathbb{E}\left[\prod_{i=1}^{x} e^{\alpha \tau_{x-i}(x-i+1,\lambda)}\right] \\ &= \prod_{i=1}^{x} \mathbb{E}\left[e^{\alpha \tau_{x-i}(x-i+1,\lambda)}\right] \\ &= \prod_{i=1}^{x} \mathbb{E}\left[e^{\alpha \tau_0(1,\lambda)}\right] \\ &= (G(1, \lambda, \alpha))^x \end{aligned}$$

Therefore

$$\begin{aligned} G(\lambda_m, \lambda, \alpha) &= \sum_{x=0}^{\infty} \left(1 - \frac{\lambda_m}{\mu}\right) \left(\frac{\lambda_m}{\mu}\right)^x (G(1, \lambda, \alpha))^x \\ &= \frac{1 - \frac{\lambda_m}{\mu}}{1 - \frac{\lambda_m}{\mu} G(1, \lambda, \alpha)} \end{aligned} \tag{5}$$

as long as $\frac{\lambda_m}{\mu} G(1, \lambda, \alpha) < 1$. Next we compute $G(1, \lambda, \alpha)$.

Theorem 4. *Assume $\lambda < \mu$. For $\alpha \leq \left(\sqrt{\mu} - \sqrt{\lambda}\right)^2$,*

$$G(1, \lambda, \alpha) = \mathbb{E}\left[e^{\alpha \tau_0(1,\lambda)}\right] = \frac{1}{2\lambda} \left(\lambda + \mu - \alpha - \sqrt{(\lambda + \mu - \alpha)^2 - 4\lambda\mu}\right) \tag{6}$$

Proof. In order to calculate the MGF, we condition on whether a departure or an arrival happens first. Let E_A be the event that an arrival happens first and let E_D be the event that a departure happens first. Let $\tau(A, \lambda)$ be the time required for the arrival, conditioned on the an arrival happening first; we define $\tau(D, \lambda)$ similarly. Using properties of exponential random variables, we have

$$\begin{aligned} \mathbb{E}\left[e^{\alpha \tau_0(1,\lambda)}\right] &= \mathbb{E}\left[e^{\alpha \tau_0(1,\lambda)} | E_A\right] \mathbb{P}(E_A) + \mathbb{E}\left[e^{\alpha \tau_0(1,\lambda)} | E_D\right] \mathbb{P}(E_D) \\ &= \mathbb{E}\left[e^{\alpha(\tau_0(2,\lambda) + \tau(A,\lambda))}\right] \frac{\lambda}{\lambda + \mu} + \mathbb{E}\left[e^{\alpha \tau(D,\lambda)}\right] \frac{\mu}{\lambda + \mu} \\ &= \mathbb{E}\left[e^{\alpha \tau_0(1,\lambda)}\right]^2 \mathbb{E}\left[e^{\alpha \tau(A,\lambda)}\right] \frac{\lambda}{\lambda + \mu} + \mathbb{E}\left[e^{\alpha \tau(D,\lambda)}\right] \frac{\mu}{\lambda + \mu} \end{aligned}$$

Now since $\tau(A, \lambda) \overset{\mathrm{d}}{=} \tau(D, \lambda) \sim exp(\lambda + \mu)$,

$$\mathbb{E}\left[e^{\alpha \tau(A,\lambda)}\right] = \mathbb{E}\left[e^{\alpha \tau(D,\lambda)}\right] = \frac{\lambda + \mu}{\lambda + \mu - \alpha}$$

so long as $\alpha < \lambda + \mu$. In fact, this is the case: $\alpha \le \left(\sqrt{\mu} - \sqrt{\lambda}\right)^2 = \mu + \lambda - 2\sqrt{\lambda\mu} < \lambda + \mu$. Now in order to find $\mathbb{E}\left[e^{\alpha\tau_0(1,\lambda)}\right]$ we solve the resulting quadratic to obtain

$$\mathbb{E}\left[e^{\alpha\tau_0(1,\lambda)}\right] = \frac{1}{2\lambda}\left(\lambda + \mu - \alpha \pm \sqrt{(\lambda+\mu-\alpha)^2 - 4\lambda\mu}\right) \tag{7}$$

In order for the discriminant to be nonnegative, we need $\alpha \le \left(\sqrt{\mu} - \sqrt{\lambda}\right)^2$ or $\alpha \ge \left(\sqrt{\mu} + \sqrt{\lambda}\right)^2$. However, the second condition is overruled by the condition $\alpha < \lambda + \mu$. To identify the correct root, we use the differentiation property of moment generating functions:

$$\mathbb{E}\left[\tau_0(1,\lambda)\right] = \frac{d}{d\alpha}\mathbb{E}\left[e^{\alpha\tau_0(1,\lambda)}\right]\bigg|_{\alpha=0}$$

Again conditioning on whether an arrival or departure happens first, we have

$$\begin{aligned}
&\mathbb{E}\left[\tau_0(1,\lambda)\right] \\
&= \left(\mathbb{E}\left[\tau_0(2,\lambda)\right] + \frac{1}{\lambda+\mu}\right)\frac{\lambda}{\lambda+\mu} + \left(\frac{1}{\lambda+\mu}\right)\frac{\mu}{\lambda+\mu} \\
&= \mathbb{E}\left[\tau_0(1,\lambda)\right] = \left(2\mathbb{E}\left[\tau_0(1,\lambda)\right] + \frac{1}{\lambda+\mu}\right)\frac{\lambda}{\lambda+\mu} + \left(\frac{1}{\lambda+\mu}\right)\frac{\mu}{\lambda+\mu} \\
&\implies \mathbb{E}\left[\tau_0(1,\lambda)\right] = \frac{1}{\mu-\lambda}
\end{aligned}$$

The + root of Equation (7) gives $\frac{d}{d\alpha}\mathbb{E}\left[e^{\alpha\tau_0(1,\lambda)}\right]\Big|_{\alpha=0} = \frac{\mu}{\lambda(\lambda-\mu)} < 0$ and the − root gives $\frac{d}{d\alpha}\mathbb{E}\left[e^{\alpha\tau_0(1,\lambda)}\right]\Big|_{\alpha=0} = \frac{1}{\mu-\lambda} = \mathbb{E}\left[\tau_0(1,\lambda)\right]$. This concludes the proof. □

Remark 1. *After proving Theorem 4, we came to know of an alternate proof in [11], pp. 92-95.*

Now we apply Theorem 3 to the convergence of M/M/1 queue with arrival rate perturbed from $r_0 = \lambda_0$ to $r = \lambda$, using Theorem 4. There are two cases:
Case 1: $\lambda_m = \lambda \ge \lambda_0$
Set $\alpha = \left(\sqrt{\mu} - \sqrt{\lambda}\right)^2$ in Equation (6) to obtain

$$G(1,\lambda,\alpha) = \sqrt{\frac{\mu}{\lambda}}.$$

To substitute into Equation (5), we need to verify that $\frac{\lambda_m}{\mu}G(1,\lambda,\alpha) < 1$.

$$\frac{\lambda_m}{\mu}G(1,\lambda,\alpha) = \frac{\lambda}{\mu}\sqrt{\frac{\mu}{\lambda}} = \sqrt{\frac{\lambda}{\mu}} < 1.$$

Thus, we obtain

$$G(\lambda_m,\lambda,\alpha) = \frac{1-\frac{\lambda}{\mu}}{1-\sqrt{\frac{\lambda}{\mu}}} = 1 + \sqrt{\frac{\lambda}{\mu}}$$

and

$$\|\mathcal{L}\left(X_t(\lambda_0,\lambda)\right)-\pi(\lambda)\|_{\mathrm{TV}} \leq \left(1+\sqrt{\frac{\lambda}{\mu}}\right) e^{-\left(\sqrt{\mu}-\sqrt{\lambda}\right)^2 t}.$$

Case 2: $\lambda_m = \lambda_0 > \lambda$

We need to pick α for which 1) $\frac{\lambda_0}{\mu} G(1,\lambda,\alpha) < 1$ and 2) $\alpha \leq \left(\sqrt{\mu}-\sqrt{\lambda}\right)^2$. Condition 1) is equivalent to

$$\frac{\lambda_0}{\mu}\left(\frac{1}{2\lambda}\left(\lambda+\mu-\alpha-\sqrt{\left(\lambda+\mu-\alpha\right)^2-4\lambda\mu}\right)\right) < 1$$

$$\sqrt{\left(\lambda+\mu-\alpha\right)^2-4\lambda\mu} > -\frac{2\lambda\mu}{\lambda_0}+\lambda+\mu-\alpha$$

To determine when Condition 1) holds, we set these quantities equal to each other.

$$\begin{aligned}\sqrt{\left(\lambda+\mu-\alpha\right)^2-4\lambda\mu} &= -\frac{2\lambda\mu}{\lambda_0}+\lambda+\mu-\alpha\\ \left(\lambda+\mu-\alpha\right)^2-4\lambda\mu &= \left(-\frac{2\lambda\mu}{\lambda_0}+\lambda+\mu-\alpha\right)^2\\ \alpha &= \lambda+\mu-\lambda_0-\frac{\lambda\mu}{\lambda_0}\end{aligned}$$

Squaring may have introduced additional solutions. With this value of α, the left side is equal to

$$\begin{aligned}\sqrt{\left(\lambda+\mu-\alpha\right)^2-4\lambda\mu} &= \sqrt{\left(\lambda_0+\frac{\lambda\mu}{\lambda_0}\right)^2-4\lambda\mu}\\ &= \sqrt{\left(\lambda_0-\frac{\lambda\mu}{\lambda_0}\right)^2}\\ &= \left|\lambda_0-\frac{\lambda\mu}{\lambda_0}\right|\end{aligned}$$

The right side is equal to $\lambda_0-\frac{\lambda\mu}{\lambda_0}$. If $\lambda_0 > \sqrt{\lambda\mu}$ there is a solution, otherwise there is no solution. Setting $\alpha = 0$, the left side is equal to $\mu-\lambda$, while the right side is less than $\mu-\lambda$ (setting $\lambda_0 = \mu-\epsilon$). Therefore when $\lambda_0 > \sqrt{\lambda\mu}$, we pick $\alpha < \lambda+\mu-\lambda_0-\frac{\lambda\mu}{\lambda_0}$. We verify that $\lambda+\mu-\lambda_0-\frac{\lambda\mu}{\lambda_0} \leq \left(\sqrt{\mu}-\sqrt{\lambda}\right)^2$. Otherwise, when $\lambda_0 \leq \sqrt{\lambda\mu}$, we are free to pick $\alpha = \left(\sqrt{\mu}-\sqrt{\lambda}\right)^2$.

Therefore Theorem 3 is satisfied by substituting either $\alpha = \lambda+\mu-\lambda_0-\frac{\lambda\mu}{\lambda_0}-\epsilon$ or $\alpha = \left(\sqrt{\mu}-\sqrt{\lambda}\right)^2$, depending on the value of λ_0. Intuitively, large values of λ_0 correspond to more "contraction" when the system goes to equilibrium, and therefore the convergence rate α should be smaller.

Remark 2. *The function*

$$f(\lambda_0) = \begin{cases}\left(\sqrt{\mu}-\sqrt{\lambda}\right)^2 & \text{if } \lambda_0 \leq \sqrt{\lambda\mu}\\ \lambda+\mu-\lambda_0-\frac{\lambda\mu}{\lambda_0} & \text{if } \lambda_0 > \sqrt{\lambda\mu}\end{cases}$$

is continuous in λ_0. In other words, the convergence rate changes continuously in λ_0.

Remark 3. *The rate $\alpha^\star = \left(\sqrt{\mu} - \sqrt{\lambda}\right)^2$ is well-known as the best convergence rate for the M/M/1 queue length process starting in a fixed initial condition (see e.g. [2] in addition to [9]). However, it is not immediately clear that the same result would hold in our setting where the initial state of the queue has a distribution:*

$$\left\|\mathcal{L}\left(X_t(\lambda_0, \lambda)\right) - \pi(\lambda)\right\|_{TV} \not\geq \mathbb{E}_{X\sim\pi(\lambda_0)}\left[\left\|\mathcal{L}\left(X_t(\lambda)|X_0 = X\right) - \pi(\lambda)\right\|_{TV}\right].$$

In other words, we cannot simply go from quenched to annealed convergence.

In the Appendix, we show another technique that gives a convergence rate of

$$\overline{\alpha} = \frac{\log \frac{\mu}{\lambda_0}}{\log \sqrt{\frac{\mu}{\lambda}}}\left(\sqrt{\mu} - \sqrt{\lambda}\right)^2$$

when $\lambda_0 > \sqrt{\lambda\mu}$. Therefore, the best known convergence rate in the $\lambda_0 > \sqrt{\lambda\mu}$ case is

$$\max\left\{\lambda + \mu - \lambda_0 - \frac{\lambda\mu}{\lambda_0}, \frac{\log \frac{\mu}{\lambda_0}}{\log \sqrt{\frac{\mu}{\lambda}}}\left(\sqrt{\mu} - \sqrt{\lambda}\right)^2\right\}.$$

We now consider a more general perturbation. Suppose the parameters of the M/M/1 queue change from (λ_0, μ_0) to (λ, μ). We can relate these parameters by $ab\lambda_0 = \lambda$ and $b\mu_0 = \mu$. First, observe that $\pi\left((b\mu_0, ab\lambda_0)\right) \equiv \pi\left((\mu_0, a\lambda_0)\right)$. Second, observe that the process $X_t((\mu_0, \lambda_0), (b\mu_0, ab\lambda_0))$ is a sped-up version of the process $X_t((\mu_0, \lambda_0), (\mu_0, a\lambda_0))$, by a factor of b. Therefore,

$$\mathcal{L}\left(X_t((\mu_0, \lambda_0), (b\mu_0, ab\lambda_0))\right) \equiv \mathcal{L}\left(X_{bt}((\mu_0, \lambda_0), (\mu_0, a\lambda_0))\right).$$

These two observations allow us to write

$$\begin{aligned}&\left\|\mathcal{L}\left(X_t((\mu_0, \lambda_0), (b\mu_0, ab\lambda_0)\right) - \pi\left((b\mu_0, ab\lambda_0)\right)\right\|_{\text{TV}}\\ &= \left\|\mathcal{L}\left(X_{bt}((\mu_0, \lambda_0), (\mu_0, a\lambda_0)\right) - \pi\left((\mu_0, a\lambda_0)\right)\right\|_{\text{TV}}.\end{aligned}$$

We then conclude

$$\left\|\mathcal{L}\left(X_t((\mu_0, \lambda_0), (\mu, \lambda)\right) - \pi\left((\mu, \lambda)\right)\right\|_{\text{TV}} \leq G\left((\mu_0, \lambda_0), (\mu_0, a\lambda_0), \alpha\right)e^{-\alpha bt}.$$

Thus, we are left with $G\left((\mu_0, \lambda_0), (\mu_0, a\lambda_0), \alpha\right)$ which is of the same form as Equation (4), allowing us to apply Theorem 3 and Theorem 4 in order to calculate a bound.

Example 1. *We now apply our work to a simple control system. Consider two parallel M/M/1 queues with arrival rates λ_1 and λ_2, and service rates μ_1 and μ_2, respectively. This queueing system could be a model for two parallel road segments, for example. Let $\lambda = \lambda_1 + \lambda_2$ be the total arrival rate. The controller chooses λ_1 and λ_2 so that the expected queue lengths are equal, by setting*

$$\lambda_1 = \frac{\mu_1\lambda}{\mu_1 + \mu_2}, \lambda_2 = \frac{\mu_2\lambda}{\mu_1 + \mu_2}.$$

We assume that $\lambda < \mu_1 + \mu_2$, so that the queueing system is stable.

Suppose that at $t = 0$, the queueing system is in equilibrium, meaning that each queue is in equilibrium. Suddenly, the service rates are perturbed to μ_1' and μ_2', which are assumed to satisfy $\lambda < \mu_1' + \mu_2'$, and are known to the controller. The controller responds by setting the new arrival rates (λ_1', λ_2') to be

$$\lambda_1' = \frac{\mu_1'\lambda}{\mu_1' + \mu_2'}, \lambda_2' = \frac{\mu_2'\lambda}{\mu_1' + \mu_2'}.$$

Our methods can be used to analyze the rate of convergence of the queue length of each queue to equilibrium. Consider the first queue. Let $b = \frac{\mu_1'}{\mu_1}$ and $a = \frac{\lambda_1'}{b\lambda_1}$. Then by the above analysis,

$$\left\|\mathcal{L}\left(X_t((\mu_1, \lambda_1), (\mu_1', \lambda_1')\right) - \pi\left((\mu_1', \lambda_1')\right)\right\|_{TV} \leq G\left((\mu_1, \lambda_1), (\mu_1, a\lambda_1), \alpha\right) e^{-\alpha b t}$$

Therefore the convergence rate is $\alpha^\star b$, where

$$\alpha^\star = \begin{cases} \left(\sqrt{\mu_1} - \sqrt{a\lambda_1}\right)^2 & \text{if } \lambda_1 \leq \sqrt{a\lambda_1\mu_1} \\ \alpha' & \text{if } \lambda_1 > \sqrt{a\lambda_1\mu_1} \end{cases},$$

and

$$\alpha' = \max\left\{ a\lambda_1 + \mu_1 - \lambda_1 - \frac{a\lambda_1\mu_1}{\lambda_1}, \frac{\log\frac{\mu_1}{\lambda_1}}{\log\sqrt{\frac{\mu_1}{a\lambda_1}}}\left(\sqrt{\mu_1} - \sqrt{a\lambda_1}\right)^2 \right\}.$$

The first condition can be rewritten as $\lambda_1 \leq \mu_1\sqrt{\frac{\lambda_1'}{\mu_1'}} = \mu_1\sqrt{\frac{1}{\mu_1' + \mu_2'}}$. The analysis for the second queue is analogous.

4.2 Workload process

Next we consider the workload process, $\{W_t\}$, for an M/M/1 queue. The value $W_t \in \mathbb{R}_{\geq 0}$ is the time remaining until the queue is empty, starting from time t. As for the queue length process, we consider changing the arrival rate from λ_0 to λ while keeping the service rate fixed at μ_0. The process $\{W_t\}$ is stochastically increasing in λ. Applying Theorem 3, we need to calculate $G_W(\lambda_m, \lambda, \alpha)$ for the process $\{W_t\}$. But $\{W_t = 0\} = \{X_t = 0\}$ since the workload is zero if and only if the queue length is zero. Therefore $G_W(\lambda_m, \lambda, \alpha) = G_X(\lambda_m, \lambda, \alpha)$, and the same convergence results follow.

In [9], it is shown that $\alpha^\star = \left(\sqrt{\mu} - \sqrt{\lambda}\right)^2$ is the best possible convergence rate for the M/M/1 workload process beginning with initial condition $W_0 = 0$. Precisely, [9] show that if $\alpha > \alpha^\star$ and $W_0 = 0$,

$$\limsup_{t\to\infty} e^{\alpha t} \left\|\mathcal{L}(W_t) - \pi\right\|_{\text{TV}} = \infty.$$

We investigate whether a similar property holds when W_0 is distributed according to the parameters (μ_0, λ_0). When $\lambda_0 \leq \sqrt{\lambda\mu}$, it turns out that $\alpha^\star = \left(\sqrt{\mu} - \sqrt{\lambda}\right)^2$ is in fact the best rate. We use the bounding process idea again with $W_t(\lambda_0, \lambda)$ and $W_t(\lambda, \lambda)$, which is analogous to the proof of Theorem 2.3 in [9]. Let $T = \inf_t\{t : W_t(\lambda_0, \lambda) = W_t(\lambda, \lambda)\}$.

$$\begin{aligned} &\left\|\mathcal{L}(W_t(\lambda_0, \lambda)) - \pi(\lambda)\right\|_{\text{TV}} \\ &= \sup_A \left|\mathbb{P}\left(W_t(\lambda_0, \lambda) \in A\right) - \pi(A; \lambda)\right| \end{aligned}$$

$$
\begin{aligned}
&\geq \left|\mathbb{P}\left(W_t(\lambda_0, \lambda) = 0\right) - \pi(0; \lambda)\right| \\
&= \mathbb{P}\left(W_t(\min\{\lambda_0, \lambda\}, \lambda) = 0, T > t\right) \\
&\geq \mathbb{P}\left(W_t(\min\{\lambda_0, \lambda\}, \lambda) = 0, T > t | W_0(\min\{\lambda_0, \lambda\}, \lambda) = 0\right) \\
&\quad \times \mathbb{P}\left(W(\min\{\lambda_0, \lambda\}, \lambda) = 0\right) \\
&= \mathbb{P}\left(W_t(\lambda) = 0, T > t | W_0(\lambda) = 0\right) \left(1 - \frac{\lambda_0}{\mu}\right)
\end{aligned}
$$

It is shown in the proof of Theorem 2.3 in [9] that for $\alpha > \left(\sqrt{\mu} - \sqrt{\lambda}\right)^2$,

$$
\limsup_{t \to \infty} e^{\alpha t} \mathbb{P}\left(W_t(\lambda) = 0, T > t | W_0(\lambda) = 0\right) = \infty.
$$

Multiplying the left side by the constant $\left(1 - \frac{\lambda_0}{\mu}\right)$,

$$
\limsup_{t \to \infty} e^{\alpha t} \mathbb{P}\left(W_t(\lambda) = 0, T > t | W_0(\lambda) = 0\right) \left(1 - \frac{\lambda_0}{\mu}\right) = \infty,
$$

and we conclude that

$$
\limsup_{t \to \infty} e^{\alpha t} \left\| \mathcal{L}(W_t(\lambda_0, \lambda)) - \pi(\lambda) \right\|_{\mathrm{TV}} = \infty
$$

when $\alpha > \left(\sqrt{\mu} - \sqrt{\lambda}\right)^2$.

When $\lambda_0 \geq \sqrt{\lambda\mu}$, we have a gap between the best known rate

$$
\alpha = \max\left\{ \frac{\log \frac{\mu}{\lambda_0}}{\log \sqrt{\frac{\mu}{\lambda}}} \left(\sqrt{\mu} - \sqrt{\lambda}\right)^2, \lambda + \mu - \lambda_0 - \frac{\lambda\mu}{\lambda_0} \right\}
$$

and the upper bound on the rate, $\alpha^\star = \left(\sqrt{\mu} - \sqrt{\lambda}\right)^2$.

5 Conclusion

In this paper we presented a method for finding exponential convergence rates for stochastically ordered Markov processes with a random initial condition. This method of analysis is useful for perturbation analysis of Markov processes, such as various queueing systems. Furthermore, we provided an explicit exponential bound for convergence in total variation distance of an M/M/1 queue that begins in an equilibrium distribution, and applied it in the analysis of a control system. The method developed in this paper can certainly be applied to other systems, such as M/G/1 queues, as long as one can identify the domain of the moment generating function of the hitting time to the zero state.

Appendix

Using a truncation technique, we can improve the convergence of the M/M/1 queue-length process (and therefore the workload process as well) in the case $\lambda_0 > \sqrt{\lambda\mu}$.

Theorem 5. *There exists a computable C such that*

$$\|\mathcal{L}(X_t(\lambda_0,\lambda)) - \pi(\lambda)\|_{TV} \le Ce^{-\overline{\alpha}t}$$

where

$$\overline{\alpha} = \frac{\log \frac{\mu}{\lambda_0}}{\log \sqrt{\frac{\mu}{\lambda}}} \left(\sqrt{\mu} - \sqrt{\lambda}\right)^2.$$

Proof.

$$\begin{aligned}
\|\mathcal{L}(X_t(\lambda_0,\lambda)) - \pi(\lambda)\|_{TV} &= \sup_A \left|\mathbb{P}\left(X_t(\lambda_0,\lambda) \in A\right) - \pi(A;\lambda)\right| \\
&= \sup_A \left(\mathbb{P}\left(X_t(\lambda_0,\lambda) \in A\right) - \pi(A;\lambda)\right) \\
&= \sup_A \left(\sum_{x=0}^{\infty} \mathbb{P}\left(X_t(\lambda) \in A | X_0 = x\right) \pi(x;\lambda_0) - \pi(A;\lambda)\right) \\
&= \sup_A \sum_{x=0}^{\infty} \pi(x;\lambda_0) \left[\mathbb{P}\left(X_t(\lambda) \in A | X_0 = x\right) - \pi(A;\lambda)\right]
\end{aligned}$$

The first equality is due to the fact that for any A,

$$\mathbb{P}\left(X_t(\lambda_0,\lambda) \in A\right) - \pi(A;\lambda) = -\left(\mathbb{P}\left(X_t(\lambda_0,\lambda) \in A^c\right) - \pi(A^c;\lambda)\right).$$

Since one of these differences of probabilities must be nonnegative, the absolute value can be dropped.

We now truncate $\pi(\lambda_0)$. Let $N(\epsilon) = \min\left\{N : \sum_{x=N+1}^{\infty} \pi(x;\lambda_0) \le \epsilon\right\}$. Continuing,

$$\le \sup_A \sum_{x=0}^{N(\epsilon)} \left[\pi(x;\lambda_0)\left(\mathbb{P}\left(X_t(\lambda) \in A | X_0 = x\right) - \pi(A;\lambda)\right)\right] + \epsilon \tag{8}$$

$$\le \sum_{x=0}^{N(\epsilon)} \left[\pi(x;\lambda_0) \sup_A \left(\mathbb{P}\left(X_t(\lambda) \in A | X_0 = x\right) - \pi(A;\lambda)\right)\right] + \epsilon \tag{9}$$

Let $\alpha^\star = \left(\sqrt{\mu} - \sqrt{\lambda}\right)^2$. Applying Theorem 2.1 from [9], we can write

$$\begin{aligned}
&\le \sum_{x=0}^{N(\epsilon)} \left[\pi(x;\lambda_0)\left(G(x,\lambda,\alpha) + G(\lambda,\lambda,\alpha^\star)\right) e^{-\alpha^\star t}\right] + \epsilon \\
&\le (1-\epsilon)\left(1 + \sqrt{\frac{\lambda}{\mu}}\right) e^{-\alpha^\star t} + \epsilon + \sum_{x=0}^{N(\epsilon)} \left[\pi(x;\lambda_0) G(x,\lambda,\alpha) e^{-\alpha^\star t}\right] \\
&= (1-\epsilon)\left(1 + \sqrt{\frac{\lambda}{\mu}}\right) e^{-\alpha^\star t} + \epsilon + \sum_{x=0}^{N(\epsilon)} \left[\pi(x;\lambda_0)\left(G(1,\lambda,\alpha)\right)^x e^{-\alpha^\star t}\right] \\
&= (1-\epsilon)\left(1 + \sqrt{\frac{\lambda}{\mu}}\right) e^{-\alpha^\star t} + \epsilon + \sum_{x=0}^{N(\epsilon)} \left[\left(1 - \frac{\lambda_0}{\mu}\right)\left(\frac{\lambda_0}{\mu}\right)^x \left(\sqrt{\frac{\mu}{\lambda}}\right)^x e^{-\alpha^\star t}\right]
\end{aligned}$$

$$= (1-\epsilon)\left(1 + \sqrt{\frac{\lambda}{\mu}}\right) e^{-\alpha^\star t} + \epsilon + \left(1 - \frac{\lambda_0}{\mu}\right)\left(\frac{\left(\frac{\lambda_0}{\sqrt{\lambda\mu}}\right)^{N(\epsilon)+1} - 1}{\frac{\lambda_0}{\sqrt{\lambda\mu}} - 1}\right) e^{-\alpha^\star t} \tag{10}$$

Set $\epsilon = e^{-\overline{\alpha}t}$ in order to fold in the ϵ term into a convergence bound. Then $N(\epsilon)$ must satisfy

$$\left(1-\frac{\lambda_0}{\mu}\right)\sum_{x=0}^{N(\epsilon)}\left(\frac{\lambda_0}{\mu}\right)^x \geq 1-e^{-\overline{\alpha}t}$$

$$N(\epsilon) \geq \frac{1}{\log\frac{\mu}{\lambda_0}}\overline{\alpha}t - 1.$$

Substituting the value $N(\epsilon) = \frac{1}{\log\frac{\mu}{\lambda_0}}\overline{\alpha}t \geq \left\lceil \frac{1}{\log\frac{\mu}{\lambda_0}}\overline{\alpha}t - 1\right\rceil$ back into the bound (10), the last term in the bound becomes

$$\left(1-\frac{\lambda_0}{\mu}\right)\left(\frac{\left(\frac{\lambda_0}{\sqrt{\lambda\mu}}\right)^{\frac{1}{\log\frac{\mu}{\lambda_0}}\overline{\alpha}t+1}-1}{\frac{\lambda_0}{\sqrt{\lambda\mu}}-1}\right)e^{-\alpha^\star t}$$

$$=\left(1-\frac{\lambda_0}{\mu}\right)\left(\frac{\frac{\lambda_0}{\sqrt{\lambda\mu}}e^{\log\left(\frac{\lambda_0}{\sqrt{\lambda\mu}}\right)\frac{1}{\log\frac{\mu}{\lambda_0}}\overline{\alpha}t}-1}{\frac{\lambda_0}{\sqrt{\lambda\mu}}-1}\right)e^{-\alpha^\star t}$$

If $\log\left(\frac{\lambda_0}{\sqrt{\lambda\mu}}\right)\frac{1}{\log\frac{\mu}{\lambda_0}}\overline{\alpha} < \alpha^\star$, then we get convergence at rate

$$\min\left\{\overline{\alpha}, \alpha^\star - \frac{\log\left(\frac{\lambda_0}{\sqrt{\lambda\mu}}\right)}{\log\frac{\mu}{\lambda_0}}\overline{\alpha}\right\}.$$

Let $\overline{\alpha} = c\alpha^\star$ with $c < \frac{\log\frac{\mu}{\lambda_0}}{\log\left(\frac{\lambda_0}{\sqrt{\lambda\mu}}\right)}$. Then we seek to maximize

$$\min\left\{c\alpha^\star, \alpha^\star - \frac{\log\left(\frac{\lambda_0}{\sqrt{\lambda\mu}}\right)}{\log\frac{\mu}{\lambda_0}}c\alpha^\star\right\}$$

over c. When $\lambda_0 > \sqrt{\lambda\mu}$ the factor $\frac{\log\left(\frac{\lambda_0}{\sqrt{\lambda\mu}}\right)}{\log\frac{\mu}{\lambda_0}}$ is positive, and the optimal c is found by setting the two quantities equal to each other, leading to $c = \frac{\log\frac{\mu}{\lambda_0}}{\log\sqrt{\frac{\mu}{\lambda}}}$. We verify that this value is less than $\frac{\log\frac{\mu}{\lambda_0}}{\log\left(\frac{\lambda_0}{\sqrt{\lambda\mu}}\right)}$. Therefore the best rate obtained by this method is

$$\overline{\alpha} = \frac{\log\frac{\mu}{\lambda_0}}{\log\sqrt{\frac{\mu}{\lambda}}}\left(\sqrt{\mu}-\sqrt{\lambda}\right)^2.$$

□

Remark 4. *The function*

$$g(\lambda_0) = \begin{cases}\left(\sqrt{\mu}-\sqrt{\lambda}\right)^2 & \text{if } \lambda_0 \leq \sqrt{\lambda\mu}\\ \frac{\log\frac{\mu}{\lambda_0}}{\log\sqrt{\frac{\mu}{\lambda}}}\left(\sqrt{\mu}-\sqrt{\lambda}\right)^2 & \text{if } \lambda_0 > \sqrt{\lambda\mu}\end{cases}$$

is continuous. In other words, the convergence rate changes continuously in λ_0.

For certain values of $(\lambda_0, \lambda, \mu)$ this rate is better than the rate previously computed, $\alpha = \lambda + \mu - \lambda_0 - \frac{\lambda\mu}{\lambda_0}$. However, $\overline{\alpha} < \alpha^\star$ when $\lambda_0 > \sqrt{\lambda\mu}$, so there is still a gap, and we do not know the best convergence rate in this case. We suspect that the rate $\overline{\alpha}$ is not the best possible, since the step from expression (8) to expression (9), which exchanges the order of a supremum with a sum, can be quite loose.

Acknowledgment

This work was partially supported by the Singapore National Research Foundation through the Singapore-MIT Alliance for Research and Technology (SMART) Centre for Future Urban Mobility (FM). Julia Gaudio is supported by a Microsoft Research PhD Fellowship.

References

[1] P. H. Baxendale. Renewal theory and computable convergence rates for geometrically ergodic Markov chains. *The Annals of Applied Probability*, 15(1B):700–738, 2005.

[2] E. A. Van Doorn. Conditions for exponential ergodicity and bounds for the decay parameter of a birth-death process. *Advances in Applied Probability*, 17(3), 1985.

[3] R. Douc, E. Moulines, and J. S. Rosenthal. Quantitative bounds for geometric convergence rates of Markov chains. *Annals of Applied Probability*, 14(4):1643–1665, 2004.

[4] Z. Hou, Y. Liu, and H. Zhang. Subgeometric rates of convergence for a class of continuous-time Markov process. *Journal of Applied Probability*, 42(3):698–712, 2005.

[5] T. Kamae, U. Krengel, and G. L. O'Brien. Stochastic inequalities on partially ordered spaces. *The Annals of Probability*, 5(899-912), 1977.

[6] P. C. Kiessler and R. Lund. Traffic intensity estimation. *Naval Research Logistics*, 56(4):385–387, 2008.

[7] Y. Liu, H. Zhang, and Y. Zhao. Computable strongly ergodic rates of convergence for continuous-time Markov chains. *ANZIAM Journal*, 49(4):463–478, 2008.

[8] Y. Liu, H. Zhang, and Y. Zhao. Subgeometric ergodicity for continuous-time Markov chains. *Journal of Mathematical Analysis and Applications*, 368:178–189, 2010.

[9] R. B. Lund, S. P. Meyn, and R. L. Tweedie. Computable exponential convergence rates for stochastically ordered Markov processes. *The Annals of Applied Probability*, 61(1):218–237, 1996.

[10] A. Novak and R. Watson. Determining an adequate probe separation for estimating the arrival rate in an M/D/1 queue using single-packet probing. *Queueing Systems*, 61:255–272, 2009.

[11] N.U. Prabhu. *Stochastic Storage Processes: Queues, Insurance Risk, Dams, and Data Coomunication*. Springer-Verlag, 2 edition, 1998.

[12] G. O. Roberts and J. S. Rosenthal. Hitting time and convergence rate bound for symmetric Langevin diffusions. *Methodology and Computing in Applied Probability*, 2017.

[13] G. O. Roberts and R. L. Tweedie. Rates of convergence of stochastically monotone and continuous time Markov models. *Journal of Applied Probability*, 37:359–373, 2000.

[14] J. S. Rosenthal. Quantitative convergence rates of Markov chains: A simple account. *Electronic Communications in Probability*, 7:123–128, 2002.

[15] A. Sarantsev. Explicit rates of exponential convergence for reflected jump-diffusions on the half-line. *Latin American Journal of Probability and Mathematical Statistics*, 13:1069–1093, 2015.