A Linear Programming Algorithm for Computing the Stationary Distribution of Heavy Traffic Approximations

Peter W. Glynn* Stanford University
Assaf Zeevi† Columbia University
Denis Saure‡ Columbia University

January 2008

Abstract

This paper proposes a linear programming algorithm for computing the stationary distribution of multidimensional diffusions that arise as approximate models of queueing networks. Our algorithm is based on the necessity and sufficiency of a Basic Adjoint Relationship (BAR), which characterizes the stationary distribution. To use the BAR characterization , we approximate the state space with a finite grid of points, and use a finite set of "test" functions. With these approximations, the BAR reduces to a set of linear equations that can be solved using standard linear programming techniques. We show that the sequence of stationary distributions which arise as solutions to the algorithm converges, in a suitable sense, to the stationary distribution of the original diffusion processes. We present applications to diffusions arising under the standard heavy-traffic regime (Semi-martingale Reflected Brownian Motions) , and to diffusions arising under the Halfin-Whitt heavy-traffic regime. Extensive computational experiences are reported. The algorithm is shown to produce good estimates of the stationary moments, as well as the entire distribution.

Short Title: a LP Algorithm for S-S Distribution of Heavy Traffic Approximations

Keywords: queueing, diffusion approximations, reflected Brownian motion, Halfin-Whitt regime, numerical methods.

2008 AMS Subject Classification (Primary): XXXX - XXXX - XXXX

*Management Science and Engineering, e-mail: glynn@stanford.edu
†Columbia Graduate School of Business, e-mail: assaf.zeevi@columbia.edu
‡Columbia Graduate School of Business, e-mail: dsaure05@gsb.columbia.edu

1 Introduction

Multiclass queueing networks are widely used as mathematical models of many real-world complex systems, e.g., communications networks, manufacturing system, and service operations. That said, most relevant queueing network models are rather intractable for purposes of exact performance analysis. Consequently, considerable research efforts have been placed on deriving performance bounds for these networks, the typical goal being to bound moments of the steady-state queue lengths or workload [see, e.g., Kumar and Kumar [29], Sigman and Yao [30], Bertsimas, Gamarnik and Tsitsiklis [28], Gamarnik and Zeevi [2], as well as references therein].

An alternative path is to approximate the queueing network model with a more simplified structure. The idea then is to use the stationary distribution of the approximating model, whose dynamics are "close" to the ones of the original network as a proxy for the original model.Perhaps the most prevalent approach in this context has focused on the study of diffusion approximations that arise under the so-called heavy-traffic operating regimes. The main objective of this paper is to provide a method to approximate the steady-state distribution of these diffusion approximations.

Consider a diffusion process $X = \{X(t), t \geq 0\}$ with state-space $S \subset \mathbb{R}^k$ that arises as a heavy-traffic limit for some queueing network, and assume that X admits a unique stationary distribution π. In most cases, π it is impossible to compute π in closed form, and a plausible alternative is to compute it numerically. To this end, let $\mathcal{A}$ denote the generator of the process X, and let $\mathcal{D}_{\mathcal{A}}$ denote its domain. Under suitable technical conditions on $\mathcal{A}$ and S, we have that

$$\int_S (\mathcal{A}f)\, \pi(dx) = 0 \quad \text{for all} f \in \mathcal{D}_{\mathcal{A}} \tag{1}$$

is necessary and sufficient to characterize π. The equation above is known as the *basic adjoint relationship* (BAR). In the context of queueing and heavy-traffic theory, it was first used by Harrison and Williams [9] to characterize the stationary distribution of a reflected Brownian motion in the nonnegative orthant. Since (1) characterizes π as the solution to an infinite dimensional system of linear equations, it is impractical to try to solve this directly. An alternative involves solving approximating S and $\mathcal{D}_{\mathcal{A}}$ by suitable chosen sequences $\{S_n\}$ and $\{\mathcal{D}_m\}$ of finite subsets of S and finite subspaces of $\mathcal{D}_{\mathcal{A}}$ respectively, such that $S_n \uparrow S$ and $\mathcal{D}_m \uparrow \mathcal{D} = \mathcal{D}_{\mathcal{A}}$ in a precise sense. For fixed values of n and m, (1) can be stated as a finite-dimensional linear program (LP), that can be solve efficiently. The hope is then, that π arises as a limit of a suitable sequence of optimal solutions to the aforementioned LP's. The main contribution of this papers is to articulate mathematical conditions under which this holds true, for the classes of diffusions processes arising in common heavy-traffic approximations, and to investigate the performance of the algorithm numerically, contrasting it with existing methods.

Related literature. A related algorithm was proposed by Dai and Harrison [6] for solving sta-

tionary distribution of Semimartingale Reflected Brownian Motions (SRBM). In there, the authors view the BAR as an orthogonality condition between a infinite dimensional functional space and the stationary distribution. By consider an increasing sequence of finite dimensional approximations of the infinite dimensional functional space, they obtain a sequence of approximating densities by means of orthogonal projections into a sequence of finite dimensional spaces. They prove that this sequence of densities converges to the stationary density of the SRBM, but their proof relies on a certain conjecture that concerns the behavior of solutions to the BAR. Later, Shen et al. [24] considered a variant of the Dai-Harrison algorithm for the case of a SRBM on a hypercube using a finite element method (or piecewise polynomials) to form the finite dimensional approximations of the functional space of "test" functions. Shen and Chen [25] extend this algorithm for SRBM on the positive orthan. Our work differs from these previous algorithms along three dimensions: First, it allows for one to impose further constraints on the structure of the underlying stationary distribution, which results in reasonable estimates not only for stationary moments but also for the distribution itself. Since the tails of the queue-length/work-load distribution play important roles as performance indicators, obtaining good approximations to the distribution itself is of obvious value. Second, the convergence of our algorithm does not rely on the structural conjecture that appears in the Dai-Harrison and follow-up papers. Finally, our method can be easily extended to derive approximations to the stationary distribution under parameter uncertainty using robust optimization (see section 5).

Harrison and Nguyen [CITE] were among the first to propose the use of the stationary distribution of the diffusion model as a mean to approximate the one corresponding to the original network, mostly on the context of moment calculation. It should be noted, however that even if the approximating diffusion model is arrived at rigourously (i.e., as a formal heavy-traffic limit), its stationary distribution may not provide a rigorous approximation of that of the underlying model. To date, the validity of this interchange-of-limits has not been established in full generality (see Gamarnik and Zeevi [2] and Gurvich and Zeevi [3] to some analysis supporting this interchange argument).

The linear programming approach pursued in the present paper originates with the work of Manne [26] in the context of discrete time and finite space Markov processes. Hernandez and Lasserre [1] extend this to analyze the convergence of linear-programming approximations for discrete time controlled Markov processes in metric spaces. In [1], the authors approximate a discrete time analogous of the BAR using a discrete probability distribution and a finite subspace of test functions. The main objective there is not to characterize the steady-state distribution but rather to minimize a steady-state cost function. Mendiondo and Stockbridge [27] extends this work to the continuous time setting in the context of long-term average and discounted control problems, when the state space and control space are assumed to be compact. The key features that distinguish

our approach relative to these papers are: (i) Our focus is on computing the steady-state distribution rather than a long-run average cost;(ii) Since our focus is on diffusions arising as heavy-traffic limits, the state space is not necessarily compact, which creates a potential problem with the sequence of approximating measures; (iii) we explicitly show our algorithm's application in the case of diffusions with reflecting boundaries, and (iv) feasibility of each linear program in our sequence of approximating problems is guaranteed, while the work cited above assumes feasibility in the respective measure-control space.

The remainder of the paper. Section 2 formulates our algorithm in general form, and its convergence is established under general conditions. Section 3 presents the application of the algorithm to the SRBM, which arises as an approximating diffusion model under the classical heavy-traffic scaling, and Section 4 illustrates its application to a Ornstein-Uhlenbeck type diffusion that arises in the so-called many server heavy-traffic regime. Extensive computational experiences are reported for both applications. Finally, Section 5 extends the algorithm for the case of uncertain parameters on the underlying queueing system, by formulating the "robust" counterpart of the LP approximation. Computational experiences are reported for the SRBM case.

2 Proposed Algorithm

Consider a positive recurrent continuous-time Markov chain (CTMC) with infinite state space S and infinitesimal generator matrix Q. A vector π is a stationary distribution for this CTMC if and only if $\pi^t Q = 0$ and $e^t \pi = 1$, where e (abusing notation) is a column vector with all entries being 1. One way of finding π is to solve the following infinite-dimensional *linear program* (LP),

$$\left\{\min u \;|\pi^t Q \leq e^t u\,;\; -\pi^t Q \leq e^t u\,;\; e^t \pi = 1\,;\; \pi \geq 0\right\} \tag{2}$$

Solving (2) exactly is virtually impossible. One possible approach is trying to "approximate" π by solving a finite-dimensional LP that approximate in some sense the original LP.

When S is a locally compact separable metric space, the Banach space $\mathcal{C}_0(S)$ contains a *countable* dense subspace (is *separable*)

$$H \equiv \{h_1, h_2, \ldots\} \subset \mathcal{C}_0(S)$$

By the denseness of H in $\mathcal{C}_0(S)$, for any two probability measures μ, ν we have

$$\begin{aligned} \mu = \nu \quad &\Longleftrightarrow \quad \langle \mu, h \rangle = \langle \nu, h \rangle \quad \forall\, h \in \mathcal{C}_0(S) \\ &\Longleftrightarrow \quad \langle \mu, h \rangle = \langle \nu, h \rangle \quad \forall\, h \in H \end{aligned}$$

Hence, solving (2) is equivalent to solving

$$\left\{\min u \mid \left\langle \pi^t Q, h \right\rangle \leq e^t u \, \forall \, h \in H \, ; \, -\left\langle \pi^t Q, h \right\rangle \leq e^t u \, \forall \, h \in H \, ; \, e^t \pi = 1 \, ; \, \pi \geq 0 \right\} \tag{3}$$

Notice that (3) has a *countable* set of constraints. The condition $\left\langle \pi^t Q, h \right\rangle = 0 \forall \, h \in \mathcal{C}_0(S); \quad e^t \pi = 1$ corresponds to the *basic adjoint relationship* (BAR), introduced by Harrison and Williams [9] to characterize the stationary distribution of a reflected Brownian motion in the nonnegative orthant.

Under the same assumptions, there exists a countable subset $\hat{S}$ dense in S.

$$\hat{S} \equiv \{x_1, x_2, \ldots\}$$

The key feature here is that, for any probability measure π on S, there exist a sequence of measures with finite support, that converges weakly to π. We can try to approximate π by solving a sequence of finite dimensional LPs.

$$\mathbb{P}_{nm} \equiv \left\{ \min u \mid \quad \left| \sum_{i=1}^{n} \lambda_i [Q(h(x_i))] \right| \leq u \, \forall \, h \, \in H_m \, ; \, e^t \lambda = 1 \, ; \, \lambda \geq 0 \right\} \tag{4}$$

where $H_m = \{h_1, \ldots, h_m\}$. The intuition here is that, as n grows large we are able to give better a approximation of an arbitrary probability distribution in S. On the other hand, as k grows large, we are giving a better approximation to the BAR condition. Hernandez and Lasserre [1] proved the validity of the approach when approximating the moment of an inf-compact nonnegative function under the stationary distribution. The approach remains valid when estimating π, provided that a tightness condition is imposed on each LP.

In what follows, we will adapt this idea to approximate stationary distributions for diffusions for which the equivalent BAR condition is known to be necessary and sufficient, and will prove the validity of the approach under certain assumptions.

Suppose the k-dimensional process $\left\{X_t \in S \subset \mathbb{R}^k : t \geq 0\right\}$ arises as a heavy-traffic limit for a stochastic process describing some queueing system. In generality X_t could be decomposed as the sum of a "free" time-homogeneous diffusion process plus some finite-variational process, reflecting the behavior of X on the boundary of S, ∂S. We will assume that X_t is a diffusion process, ignoring the boundary behavior, although our results are easily extended to the general case, as we will illustrate in section 3, when dealing with the semimartingale reflected Brownian motion case.

Assume $\left\{X_t \in S \subset \mathbb{R}^k : t \geq 0\right\}$ solves the following stochastic differential equation

$$dX_t = b(X_t)dt + \sigma(x_t)dB_t \quad t \geq 0, \quad X_0 = x \tag{5}$$

where $b(x) \in \mathbb{R}^k$, $\sigma(x) \in \mathbb{R}^{k \times l}$ are continuous functions , and B_t is a l-dimensional Brownian motion. The infinitesimal generator $\mathcal{A}$ of X_t is defined by

$$\mathcal{A}f(x) = \lim_{t \downarrow 0} \frac{E^x[f(X_t)] - f(x)}{t} \quad x \in S \tag{6}$$

The set of functions $f : S \to \mathbb{R}$ such that the limit exits at x is denoted by $\mathcal{D}_{\mathcal{A}}(x)$, while $\mathcal{D}_{\mathcal{A}}$ denotes the set of functions for which the limit exists for all $x \in S$. We will assume there exits a further set of functions $H \subset C^2(S)$ such that $f \in \mathcal{D}_{\mathcal{A}}$ and

$$\mathcal{A}f(x) = \sum_i b_i(x)\frac{\partial f}{\partial x_i} + \frac{1}{2}\sum_{i,j}(\sigma\sigma^t)_{i,j}(x)\frac{\partial^2 f}{\partial x_i \partial x_j} \tag{7}$$

For example, if σ is bounded, then we can take $H \equiv C_0^2$. Suppose that X_t is a positive recurrent process and π is its unique stationary distribution, then

$$\int_S \mathcal{A}f(x)\pi(dx) = 0 \quad \forall\, f \in H \tag{8}$$

Unfortunately, for the diffusion case the BAR condition (8) is not always both necessary and sufficient to characterize the stationary distribution π, and further assumptions are required. For the purpose of this paper, we will just assume that the BAR condition is both necessary and sufficient, and we will elaborate from it.

Suppose that $\left\{X_t \in S \subset \mathbb{R}^k : t \geq 0\right\}$ solves (5), that exists a countable subset $\hat{S}$ dense in S, and that $H \subset C_0^2(S)$ contains a *countable* dense subspace $\hat{H} \equiv \{h_1, h_2, \ldots\} \subset H$. Using the aggregation-relaxation-inner approximation procedure described by Hernandez and Lasserre [1] we can set the following LP

$$\mathbb{P}_{nm} \equiv \left\{ \min u | \sum_{i=1}^{n} \lambda_i \mathcal{A}h(x_i) \leq u \,\forall\, h \in \hat{H}_m \,;\, -\sum_{i=1}^{n} \lambda_i \mathcal{A}h(x_i) \leq u \,\forall\, h \in \hat{H}_m \,;\, \lambda \in \Lambda_n \right\} \tag{9}$$

where $\hat{H}_m = \{h_1, \ldots, h_m\}$, and $\Lambda_n \equiv \left\{\lambda \in \mathbb{R}^n | \sum_{i=1}^n \lambda_i g(x_i) \leq M \,;\, e^t\lambda = 1 \,;\, \lambda \geq 0\right\}$ with $g : S \to \mathbb{R}$ nonnegative and continuous and $M > 0$ such that $\exists\, K > 0$, $r > 0$ for which $g(x) \geq K \,\forall\, |x| > r$, $x \in S$, and $E_\pi[g(x)] < M$. The idea here is to ensure tightness for a sequence of solutions to $\mathbb{P}_{nm}$. Let Θ_{nm} denote the set of optimal solutions to $\mathbb{P}_{nm}$.

Theorem 1 *Let $\left\{X_t \in S \subset \mathbb{R}^k : t \geq 0\right\}$ be a positive recurrent process that solves (5) and let π be its unique stationary distribution. Suppose that (8) is both necessary and sufficient, and that both $\hat{S}$ and $\hat{H}$ exists. Then, there exist sequences of integers $n(i)$ and $m(i)$ such that $\pi_{n(i)m(i)} \to \pi$ as $i \uparrow \infty$, with $\pi_{mn} \in \Theta_{nm}$.*

Proof We know there exists a sequence $\{\mu_n\}$ of distribution functions on S such that

(a) For every $n = 1, \ldots,$ μ_n has finite support $\{x_1, \ldots, x_n\}$, that is, μ_n is of the form $\mu_n = \sum_{i=1}^n \beta_i^n \delta_{x_i}$, with $\beta_i^n \geq 0 \,\forall\, i = 1 \ldots, n$, and $\sum_{i=1}^n \beta_i^n = 1$.

(b) The sequence $\{\mu_n\}$ converges weakly to π.

From the definition of weak convergence (plus a truncation argument) we know that $E_{\mu_n}[g(x)] \to E_\pi[g(x)]$, therefore there exists a n^* for which $\beta_n \in \Lambda_n \ \forall\, n \geq n^*$. Fix $m > 0$. Consider $\pi_{mn} \equiv \{\lambda_{mni}\, i = 1 \dots n\}$. For $h \in \hat{H}_m$ we have that

$$\lim_{n\uparrow\infty} |\sum_{i=1}^{n} \lambda_{mni} \mathcal{A}h(x_i)| \leq \lim_{n\uparrow\infty} |\sum_{i=1}^{n} \beta_i^n \mathcal{A}h(x_i)| = \lim_{n\uparrow\infty} |\int_S \mathcal{A}h(x)\mu_n(dx)| \to |\int_S \mathcal{A}h(x)\pi(dx)| = 0 \quad (10)$$

The first inequality comes from the optimality of π_{mn} and the feasibility of μ_n for $n \geq n^*$. The limit holds since $\mathcal{A}h(x)$ is bounded on the support of h by its continuity. Due to the tightness condition we have that there exists a subsequence $\{n^m(i)\}$ of integers such that $\pi_{nm} \Rightarrow \pi_m$, with π_m a proper distribution function. From the reasons above we can also conclude that

$$\lim_{n\uparrow\infty} |\sum_{i=1}^{n} \lambda_{mni} \mathcal{A}h(x_i)| = \int_S \mathcal{A}h(x)\pi_m(dx) = 0 \quad (11)$$

Now consider the sequence $\{\pi_m\}$. Again, this sequence is tight, and therefore $\pi_m \Rightarrow \hat{\pi}$ along a further subsequence. Finally we have that, for an arbitrary function $h \in \hat{H}$ $\exists\, m^*$ such that $h \in \hat{H}_m$ $\forall\, m \geq m^*$ and we have that $|\int_S \mathcal{A}h(x)\pi_m(dx)| = 0$. Therefore

$$|\int_S \mathcal{A}h(x)\pi(\hat{dx})| = \lim_{m\uparrow\infty} |\int_S \mathcal{A}h(x)\pi_m(dx)| = 0 \quad (12)$$

Since h was arbitrary, we conclude that $\hat{\pi} = \pi$ by the necessity and sufficiency of (8), and the uniqueness of the stationary distribution. ■

3 Applications: Semimartingale Reflected Brownian Motions

This section illustrates the application of the algorithm to a class of diffusion processes that plays a central role in queuing theory: The Semimartingale Reflected Brownian Motions (abbreviated as SRBM). These processes have been shown to arise as diffusion limits of open multiclass queueing networks operating under "conventional" Heavy-Traffic conditions: Consider a sequence of queueing networks indexed by n. The key idea of conventional Heavy-Traffic theory is to scale amounts to express time in multiples of n and space in multiples of $n^{1/2}$ so that functional central limit theorem applies to a properly normalized queue length process while the sequence of traffic intensities converges to 1 as n increases. A formal for the SRBM definition is the following:

Definition 1 *Let $S \equiv \mathbb{R}_+^d$ (the positive d-dimensional orthant). Let μ be a constant vector in $\mathbb{R}^d$, σ a $d \times d$ non-degenerate covariance matrix (symmetric and strictly positive definite), and R a $d \times d$ matrix. For each $x \in S$, a SRBM associated with the data (S, μ, σ, R, x) is a $\mathcal{F}_t$-adapted, d-dimensional process $Z = (Z(t) : t \geq 0)$ defined on some filtered probability space $(\Omega, \mathcal{F}.\mathcal{F}_t, \mathbb{P})$ such that:*

(i) $Z = X + RY$, $\mathbb{P}_x$*-a.s.,*

(ii) $\mathbb{P}_x$*-a.s.,* Z *has continuous paths and* $Z(t) \in S$ *for all* $t \geq 0$*,*

(iii) X *is a* d*-dimensional Brownian motion with drift vector* μ*, covariance matrix* σ *and* $X(0) = x$*. In addition* $X(t) - \mu t$ *is a* $\mathcal{F}_t$*-martingale,*

(iv) Y *is an* $\mathcal{F}_t$*-adapted* d*-dimensional process such that under* $\mathbb{P}$ *it satisfies for each* $j = 1, \ldots, d$ *:*

a). $Y(0) = 0$

b). $(Y_i(t) : t \geq 0)$ *is continuous and non-decreasing,*

c). $Y_i(t)$ *can increase only when* Z *hits the face* $F_i = \{x \in S : x_i = 0\}$*.*

Loosely speaking, SRBM behaves like Brownian motion in the interior of S, and is confined to the orthant by instantaneous "reflection" at the boundary faces, where the direction of reflection is dictated by the matrix R. The most general condition currently known to ensure existence and uniqueness (in law) of SRBM in the orthant is the matrix R to be *completely S*.

Definition 2 *A* $d \times d$ *matrix* R *is said to be* S *if there exits a* d*-dimensional vector* $u \geq 0$ *such that* $Ru > 0$*, and to be a* completely S *matrix if each of its principal submatrices is an* S *matrix.*

This *completely S* condition is in fact necessary (Reiman and Williams [16]) and sufficient (Taylor and Williams [12]). Regarding its stationary distribution it has been shown (Dupuis and Williams [17]) that a sufficient condition for its existence and uniqueness is that all solutions of an associated deterministic Skorohod problem are attracted to the origin in finite time. Define $\gamma = R^{-1}\mu$. A more tractable condition, $\gamma < 0$ is known to be necessary when $R^{-1} \geq 0$ (Harrison and Williams [9], Dai [6]) but not sufficient (Dai and Harrison [18]).

If we want to apply our algorithm to this class of processes we need to check that (i) The BAR condition is sufficient and necessary, and that (ii) we can provide a tightness bound for the interior distribution and tightness and finiteness bounds for boundary measures.

3.1 Bar Condition

For $f \in C_b^2(S)$, define

$$Lf \equiv \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}\sigma_{ij}\frac{\partial^2 f}{\partial x_i \partial x_j} + \sum_{i=1}^{d}\mu_i\frac{\partial f}{\partial x_i} \tag{13}$$

$$D_i f \equiv R_{\cdot i}\nabla f \text{ for } x \in F_i \, (i = 1, \ldots, d) \tag{14}$$

For these processes, BAR takes the following form:

$$\int_{\pi} Lf d\pi + \sum_{i=1}^{d} \int_{F_i} D_i f d\nu_i = 0 \quad \forall\, f \in C_b^2(S) \tag{15}$$

where π is the stationary distribution for Z associated with boundary measures $\nu_i\,(i = 1, \ldots, d)$.

Necessity of BAR was first derived by Harrison and Williams [9] when the matrix R is Minkowski ($I - R \geq 0$ and $I - R$ is transient), and later by Dai [6] for the *completely S* case. Sufficiency was proven by Dai and Kurtz [8] through the following theorem:

Theorem 2 *Assume R is a* completely S *matrix. Suppose that π_0 is a probability measure on S with support in the interior of S, and $\pi_1, \ldots, \pi_d$ are positive finite measures with supports on $F_1, \ldots, F_d$ respectively. If they jointly satisfies (15), then π_0 is the stationary distribution for a (σ, μ, R)-SRBM Z.*

3.2 Tightness condition

The use of Lyapunov functions to bound expectations of Markov processes is a widely used technique (for a clear exposition see Glynn and Zeevi [19]). Given a function f for which one is interested in computing stationary moments, the key idea is to find a Lyapunov-like function g that satisfies a certain "mean" drift inequality with respect to f, which in turn leads to bounds on the stationary expectation of the later. This idea has been used to prove the existence of exponential moments of the interior stationary distribution for the SRBM: Glynn and Zeevi [19] provided an explicit construction for the case of R being symmetric and positive definite. Budhiraja and Lee [15] have proven the existence of such function for the *completely S* case. We prove that same result holds for the boundary measures. In what follows we will use the following proposition, proven by Dai [6]

Proposition 1 *If π is the stationary distribution, associated with boundary measures $\nu_i\,(i = 1, \ldots, d)$, then*

$$E_{\pi}\left[\int_0^t f(Z(s))dY_i(s)\right] = t\int_{F_i} f\nu_i, \quad (i = 1, \ldots, d) \tag{16}$$

where σ_i denote the (d−1)-dimensional Lebesgue measure on the face F_i.

Theorem 3 *Assume R is a* completely S *matrix, and suppose π is the stationary distribution for Z associated with boundary measures $\nu_i\,(i = 1, \ldots, d)$. There exits a vector $\upsilon > 0$ such that*

$$\int_{F_i} e^{\upsilon^t x}\,\nu_i(dx) < \infty \quad , i = 1, \ldots, d$$

These results indicates that moment bounds exists, but we still need practical bounds to be used in the setting of the algorithm, for tightness and finiteness of the boundary measures.

3.3 Practical Bounds

Finiteness of boundary measures

We will use a Lyapunov-function argument. Consider a d-dimensional vector $v > 0$ such that $R^t v > 0$ and set $f(x) = x^t v$. Proceeding as in proof of Theorem 3 (see appendix) we have:

$$E_x[f(Z(t))] - E_x[f(x)] + \alpha = E_x[\sum_{i=1}^{d} \beta_i \int_0^t dY_i(s)]$$

where $\alpha \equiv v^t \mu > 0$ and $\beta_i \equiv v^t R_{\cdot i} > 0 \quad i = 1, \ldots, d$. Taking expectations with respect to π, and using Proposition 1 we have

$$\int_{F_i} \nu_i(ds) \leq \frac{\alpha}{\beta_i} \quad i = 1 \ldots, d$$

Tightness of stationary distribution

Here we will assume that R is symmetric to facilitate the explicit construction of a simple Lyapunov function. This can be relaxed by means of a more clever choice of such function. Consider $f(x) = x^t R^{-1} x$. Proceeding as in the proof of Theorem 3, we have

$$-f(x) + E_x[\int_0^t Z(s)^t \gamma ds] \leq C + E_x[\sum_{i=1}^{d} \int_0^t Z_i(s) dY_i(s)]$$

where $C = \sum_{ij} \sigma_{ij} R_{ij}$. Notice that $E_x[\sum_{i=1}^{d} \int_0^t Z_i(s) dY_i(s)] = 0$ by definition of the SRBM. Taking expectations with respect to π we have

$$\int_S \langle e, s \rangle \pi(ds) \leq \frac{C}{\min_j \{\gamma_j\}}$$

For the Minkowski case, a bound can be derived directly from lemma 8.4 in Harrison-Williams [9].

Tightness of boundary measures

Take a d-dimensional vector $v > 0$ such that $R^t v > 0$ and set the $d \times d$ symmetric matrix V such that $V_{ij} = v_j v_i > 0$. Notice that $V > 0$ and $VR > 0$. Let $f(x) = x^t V x$. Proceeding as in the proof of Theorem 3, we have

$$E_x[f(Z(t))] + E_x[\int_0^t Z(s)^t V R \gamma ds] \geq E_x[\sum_{i=1}^{d} \int_0^t Z(s)^t V R_{\cdot i} dY_i(s)]$$

Taking expectations with respect to π, and using Proposition 1 we have

$$\int_{F_i} \langle e, s\rangle \nu_i(ds) \leq \frac{\alpha}{\beta_i} \int_S \langle e, s\rangle \pi(ds),\ i = 1, \ldots, d$$

where $\alpha \equiv \max_j \{(VR\gamma)_j\} > 0$ and $\beta_i \equiv \min_j \{(VR_{\cdot i})_j\} > 0$, $i = 1, \ldots, d$. We have implicitly used the finiteness of second moments for the stationary distribution.

3.4 Algorithm setting

To apply the algorithm we still need to specify (i) $\{S_n\}$, the sequence of grid for which we will approximate (15), and (ii) H_m, the sequence of finite dimensional subspaces approximating $C_b^2(S)$.

Grid choice

Harrison and Williams [10] proved that the stationary distribution for a "standard" SRBM has a separable density function (in the usual Cartesian coordinates) if and only if the covariance matrix σ satisfies the "skew symmetric" condition

$$2\sigma_{ij} = R_{ij} + R_{ji} \quad \text{for} \quad i \neq j \tag{17}$$

In this case, the marginal distribution for coordinate i is exponential with rate $2\gamma_i$. More over, the boundary measures are the restriction of the join distribution to the corresponding faces of S. In practice the matrix R does not fulfill this condition, but the closer it is to fulfill it, the closer the marginal distributions should be to be exponentials.

Recently, Budhiraja and Lee [15] established the finiteness of the moment generating function of the steady state distribution in a neighborhood of zero, proving the exponential decay of it.

With this in mind we choose the grid to have an "exponential spacing on the marginal":

$$S_n = \{x \in S | x_i \in \{[\log(n) - \log(j)]/\lambda_i \quad j = 1, \ldots, n\} \quad i = 1, \ldots, d\} \tag{18}$$

For the boundary, we just project this grind on the corresponding face

$$F_i^n = \{x \in S_n | x_i = 0\} \quad i = 1, \ldots, n \tag{19}$$

This is somehow related to the choice of a reference density on Harrison and Dai's algorithm [7].

Remark. In practice, the values for λ were selected as follows: First we perform a first run with the algorithm using low values for λ. Then these values were adjusted according to the first moments on the marginal distributions.

m	3	4	5	6	7	8	9	10
$E_\pi[x]$	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5
$E_\pi[y]$	0.793	0.821	0.764	0.768	0.762	0.758	0.760	0.750

Table 1: Moment estimates Two-dimensional SRBM ($n = 100$)

Subspace choice

For each $m \geq 1$, we choose $H_m = \{f \text{ polynomial of degree } \leq m\}$. The convergence of $H_m \to C_b^2(S)$ is a well known result (Proper citation).

3.5 Numerical Results

In this section we will compare numerical results from our algorithm with some known analytical results of particular instances of SRBM. Also, we will compare our algorithm with previous algorithms using instances for which either simulation or good approximation results are known.

A Two-Dimensional SRBM

Consider a two-dimensional SRBM associated to the following data:

$$R = \begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix} \qquad \sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \mu = \begin{pmatrix} \mu_1 \\ 0 \end{pmatrix}$$

For this SRBM the stationary condition $\gamma < 0$ reduces to $\mu_1 < 0$. Harrison [14] computed a closed form solution for the stationary distribution density in polar coordinates:

$$p(x, y) = (2|\mu|)^{3/2}/(\pi^{1/2})r^{-1/2} \exp(\mu_1 r(1 + \cos(\theta))) \cos(\theta/2) \tag{20}$$

where $(x, y) = (r\cos(\theta), r\sin(\theta))$

Without loss of generality consider $\mu_1 = -1$. It can be shown (Greenberg [13]) that

$$E_\pi[x] = 0.5 \qquad E_\pi[y] = 0.75 \tag{21}$$

Taking $n = 100$ we used our algorithm to get estimates for these moments. The results are shown on Table 1. We see that our algorithm provides good estimates even for low values of m. Each one of the instances shown on Table 1 ran in less than 3 minutes (on a regular Desktop PC) on a MATLAB implementation of the algorithm. This type of SRBM is the only one (beside the skew symmetric type) for which the actual distribution is know, so we can use this knowledge to test our algorithm. Figure 1 shows our marginal distribution estimates for $n = 100$ and $m = 6$,

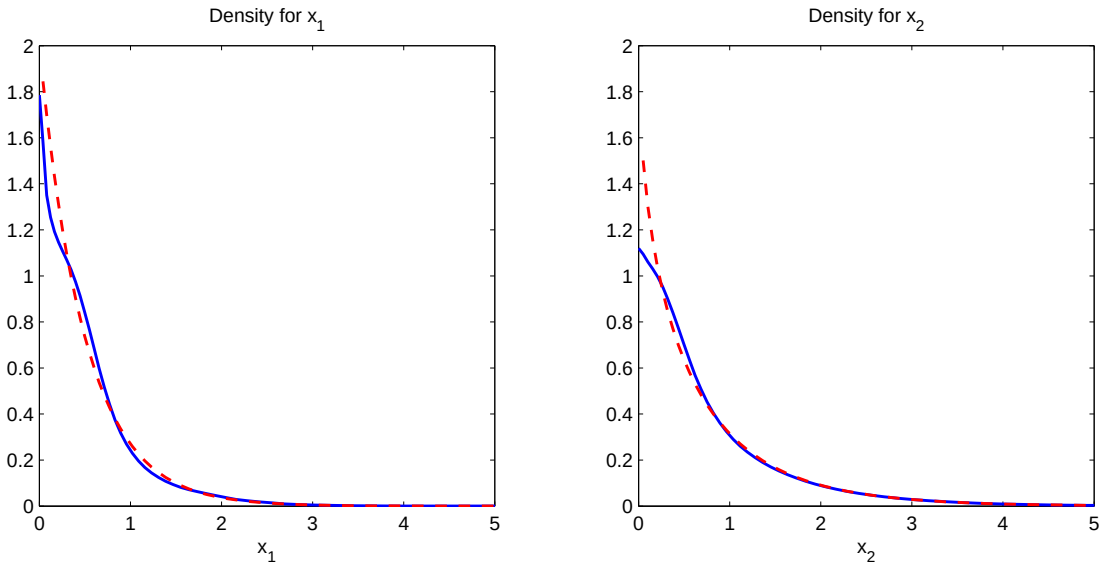

Figure 1: Marginal distribution estimates Two-dimensional SRBM ($n = 100$, $m = 6$)

compared with the actual marginal distributions (doted line). It is proven (Harrison [14]) that the one dimensional marginal distribution for x is exponential, however the same result does not hold for y. We see that our algorithm provides good estimates for the marginal distributions.

Symmetric SRBM's

A standard SRBM is said to be symmetric if its data has the following properties: $\sigma_{ij} = \rho$ for $1 \leq i < j \leq d$, $\mu_i = -1$ for $i = 1, \ldots, d$ and $R_{ij} = R_{ji} = -r \leq 0$ for $1 \leq i < j \leq d$. Positiveness of σ implies $-1/(d-1) < \rho < 1$ and the *completely S* condition reduces to $r(d-1) < 1$. This type of SRBM arises as a Heavy-Traffic limit of a symmetric generalized Jackson network. Manipulating the BAR condition, Dai [6] showed that

$$m_1 = \ldots = m_d = \frac{1 - (d-2)r + (d-1)r\rho}{2(r+1)} \tag{22}$$

where $m_i = E_\pi[x_i]$ $i = 1, \ldots, d$. We use our algorithm to compute estimates for these moments for the case $d = 2$. Imitating Dai's work, we let ρ range through $\{-0.9, -0.5, 0.0, 0.5, 0.9\}$ and r range through $\{0.2, 0.4, 0.6, 0.8, 0.9, 0.95\}$. Table 2 shows the relative errors between our estimates and the exact values for $m = 6$ and $n = 100$. We see that in each case the relative error is lower that 1%.

r/ρ	-0.9	-0.5	0.0	0.5	0.9
0.2	1.98e-4	1.65e-4	1.26e-3	2.32e-3	6.66e-4
0.4	3.90e-4	4.88e-5	2.52e-5	2.92e-3	2.00e-4
0.6	1.76e-5	5.97e-4	2.05e-4	1.06e-6	5.70e-4
0.8	1.53e-5	4.27e-5	2.22e-5	1.17e-3	2.16e-3
0.9	1.65e-5	4.26e-5	5.43e-4	1.00e-3	3.24e-3
0.95	4.57e-6	3.42e-5	2.46e-4	5.98e-4	4.82e-3

Table 2: Relative errors for Symmetric SRBM

m	3	4	5	6	7	8	9	10
$E_\pi[x_1]$	0.4942	0.4943	0.4968	0.5014	0.5005	0.5011	0.4996	0.5002
$E_\pi[x_2]$	2.0095	2.0094	2.0053	1.9976	1.9990	1.9981	2.0005	1.9996

Table 3: Moment estimates Skew-symmetric SRBM

Skew-symmetric SRBM

Harrison and Williams [10] proved that a standard SRBM has a product form stationary distribution if and only if $\gamma < 0$ and condition 17 holds. In this, case we know that the marginal distribution for x_i is exponential with mean $1/(2\gamma_i)$. Consider a two-dimensional SRBM associated to the following data:

$$R = \begin{pmatrix} 1 & -0.6 \\ -0.25 & 1 \end{pmatrix} \qquad \sigma = \begin{pmatrix} 1 & -0.425 \\ -0.425 & 1 \end{pmatrix} \qquad \mu = \begin{pmatrix} -0.85 \\ 0 \end{pmatrix}$$

One can check that condition 17 holds, and that $\gamma^t = (-1, -0.25)$. This implies that $E_\pi[x_1] = 0.5$ and $E_\pi[x_2] = 2$. Taking $n = 100$ we used our algorithm to get estimates for these moments. The results are shown on Table 3. We see that our algorithm provides good estimates even for low values of m. Each one of the instances shown on Table 3 ran in less than 3 minutes (on a regular desktop PC) on a MATLAB implementation of the algorithm. Since we know that each marginal distribution is exponential, we can compare our estimated marginal distribution against the actual marginal distribution for each coordinate. Figure 2 shows our marginal distribution estimate for $n = 100$ and $m = 6$, compared with the actual marginal distributions (doted line). We see that our algorithm provides good estimates for the marginal distributions.

Suresh and Whitt's Experiments

This section follows closely the analysis presented in Chapter 4 of Dai's dissertation [6]. Consider a network of d queues in tandem. Let ρ_i denote the mean service time at station i, $C_{s_i}^2$ denote the squared coefficient of variation of the service time distribution at station i, and C_a^2 denote the

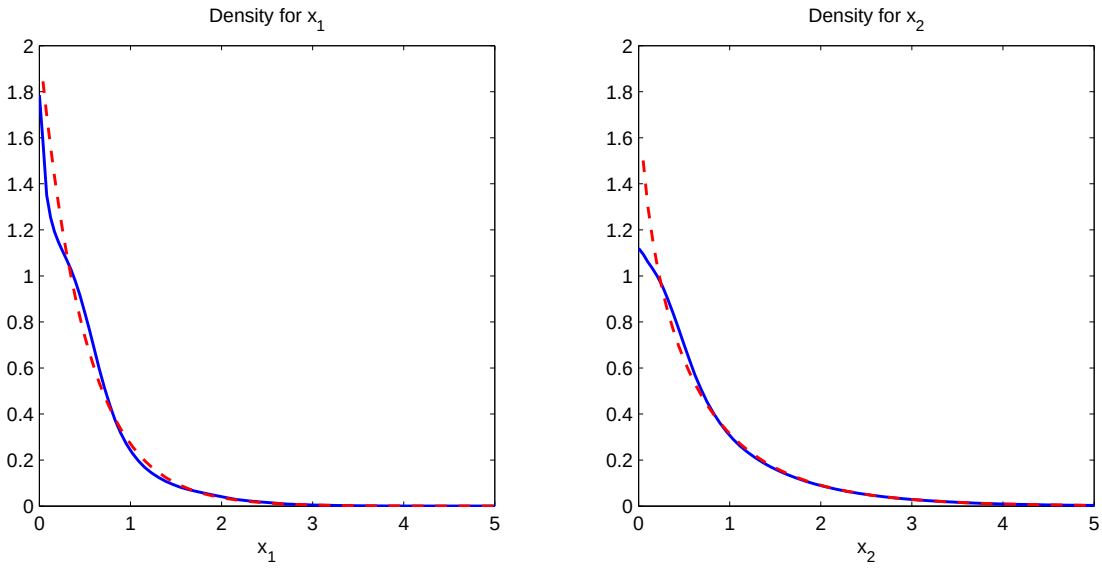

Figure 2: Marginal distribution estimates Skew-symmetric SRBM

squared coefficient of variation for the interarrival time distribution. Using the method proposed by Harrison and Nguyen [11], Dai [6] shows that the d-dimensional current workload process can be approximated by a d-dimensional SBRM associated with the data (σ, R, μ). When $d = 2$ we have

$$R = \begin{pmatrix} 1 & 0 \\ -\rho_2/\rho_1 & 1 \end{pmatrix} \qquad \sigma = \begin{pmatrix} \rho_1^2(C_a^2 + C_{s_1}^2) & -rho_1\rho_2 C_{s_1}^2 \\ \rho_2^2(C_{s_1}^2 + C_{s_2}^2) & -rho_1\rho_2 C_{s_1}^2 \end{pmatrix} \qquad \mu = \begin{pmatrix} \rho_1 - 1 \\ \rho_2/\rho_1 - 1 \end{pmatrix}$$

The arrival rate is assumed to be 1, so that ρ_i represents the traffic intensity at station i.

Suresh and Whitt [20] studied a system of two queues in tandem with the purpose of finding the best order for the queues to minimize the average total waiting time. They consider various variability parameter triples $(C_a^2, C_{s_1}^2, C_{s_2}^2)$ for all combinations of the traffic intensities ρ_1 and ρ_2 in a representative range. For $C_{s_1}^2 \neq C_{s_2}^2$ they considered five variability triples, namely, (0.5,0.5,2.0), (1.0,0.5,8.0), (1.0,2.0,4.0), (4.5,0.5,1.0) and ,(4.0,1.0,4.0). We will refer them as Case 1 to Case 5, respectively. For $C_{s_1}^2 = C_{s_2}^2$ they consider two variability triples, namely, (1.0,0.5,0.5) and (1.0,4.0,4.0). We will refer them as Case 6 and Case 7 respectively. When $C_{s_1}^2 \neq C_{s_2}^2$, for each queue they consider four values of ρ_i : 0.3, 0.6, 0.8, 0.9. When $C_{s_1}^2 \neq C_{s_2}^2$, for each queue they consider five values of ρ_i : 0.1, 0.2, 0.3, 0.6, 0.9.

Extensive simulation experiments were conducted in order to obtain estimates for the expected steady-state waiting times. When $C^2 = 0.5$, the E_2 distribution was used. When $C^2 = 1.0$, the exponential distribution was used. When $C^2 > 1.0$, the H_2 distribution with balanced means was used.

			Order I			Order II		
Case	ρ_1	ρ_2	QNET-D	QNA-D	LP-D	QNET-D	QNA-D	LP-D
1	0.9	0.9	0.01	0.01	0.00	0.02	0.13	0.01
	0.8	0.8	0.08	0.08	0.08	0.06	0.08	0.07
2	0.9	0.9	0.03	0.06	0.03	0.01	0.21	0.01
	0.8	0.8	0.04	0.01	0.02	0.08	0.07	0.07
3	0.9	0.9	0.04	0.02	0.04	0.04	0.19	0.05
	0.8	0.8	0.00	0.04	0.01	0.02	0.02	0.01
4	0.9	0.9	0.13	0.30	0.14	0.19	0.29	0.19
	0.8	0.8	0.14	0.08	0.11	0.09	0.00	0.07
5	0.9	0.9	0.01	0.12	0.01	0.01	0.01	0.00
	0.8	0.8	0.09	0.14	0.08	0.05	0.05	0.05
6	0.9	0.9	0.01	0.08	0.02	0.01	0.10	0.01
	0.8	0.8	0.01	0.04	0.01	0.01	0.04	0.01
7	0.9	0.9	0.04	0.18	0.05	0.06	0.07	0.04
	0.8	0.8	0.02	0.05	0.03	0.01	0.06	0.01
Average			0.05	0.09	0.04	0.05	0.09	0.04

Table 4: Overall Comparisons with QNA and QNET Approximations in Heavy Traffic

They also compare their simulation results with QNA approximations. Dai [6] use this set of experiments to test the QNET method. There, simulation and approximation results for the expected waiting time at the second station are compared. We used these results to compare our algorithm against the QNA and QNET methods. Table X-Y (appendix) give simulation estimates, QNET estimates, QNA estimates, and our "LP" estimates. Table 4 summarizes all balanced heavy traffic cases and gives an overall comparison of QNA estimates, QNET estimates and LP estimates under heavy traffic. There, quantities shown represent the minimum between absolute relative difference and absolute difference, with respect to simulation estimates.

We used $n = 100$ and $m = 6$ for all experiments. Each run took less than 3 minutes on regular Desktop PC. As expected, QNET and LP methods give "better" estimates than the ones from QNA method, under balanced heavy traffic conditions. Also, LP method performs as good as the QNET method, most of the time better.

4 Applications: Two Class Queue

Consider a single queue system with m classes, with class 1 denoting a guaranteed-rate or high-priority class, and classes $2, \ldots, m$ being different best-effort classes that are labeled according to their priority level i.e., class i has higher priority than all classes $j > i$. The guaranteed-rate class, assumed to arrive according to a Poisson process with rate λ_1, engage *one unit of capacity* each for i.i.d. exponentially distributed amounts of time with rate μ_1, provided that the total number of guaranteed users connected is less than the system capacity C; otherwise they are denied service.

The *best-effort users of class* i $(i = 2, \ldots, m)$, assumed to arrive according to a Poisson process with rate λ_i (independent of everything else), are always admitted into the system, requiring a total processing time exponentially distributed with rate μ_i. When there is enough capacity not used up by users from classes $j < i$, class i users are allocated a nominal processing rate corresponding to one unit of capacity, and when capacity is not sufficient, they share the available capacity in an *egalitarian manner*:

$$\text{Class } i \text{ service rate at time } t = \frac{(C - \sum_{j<i} Q_j(t))^+}{Q_i(t)} \wedge 1$$

where $Q_i(t)$ denotes the number of class i users in the system.

We see that Q_1 evolves according to the state of an M/M/C/C system, while Q_i $(i = 2, \ldots, m)$ evolve as a M/M/C(t), where the available capacity, $C(t)$, is stochastically modulated by the number of higher priority users present in the system. Despite its simple structure, exact analysis of this multiclass system is not straightforward and relies either on simulation or on numerical methods that offer little insight as to its structural behavior.

Maglaras and Zeevi [21] studied this system and derived a diffusion approximation under the Halfin and Whitt [22] asymptotic heavy traffic regime. This regime is defined by letting capacity grow large and concurrently letting the system utilization approach 1 at an appropriate rate. Specifically, Maglaras and Zeevi considered a sequence of systems with capacity $C^n = n$ and arrival rates $\lambda_i^n = n\kappa_i\mu_i - \gamma_i\sqrt{n}\mu_i$ for some constants $\kappa_i > 0$ and $\gamma_i \in \mathbb{R}$ for $i = 1, \ldots, m$ such that $\sum_i \kappa_i = 1$. There κ denotes the vector of *relative workload contributions*

$$\kappa_i = \lim_{n\to\infty} \frac{\lambda_i^n/\mu_i}{\sum_{j=1^m} \lambda_j^n/\mu_j} \quad i = 1, \ldots, m \tag{23}$$

Defining the normalized state processes

$$X_i^n(t) \equiv \frac{Q_i^n(t) - \kappa_i n}{\sqrt{n}} \quad i = 1, \ldots, m$$

they proved the following theorem:

Theorem 4 *Suppose that for some* $\xi \in \mathbb{R}^m$, $Q_i^n(0) = \lfloor n\kappa_i + \sqrt{n}\xi_i \rfloor$ *for* $i = 1, \ldots, m$. *Then,* $X^n \to X$ *in* $D^m[0,\infty)$ *as* $n \to \infty$, *where* X *is a diffusion process. Specifically,* X *is the unique strong solution of the following stochastic differential equation:*

$$dX(t) = b(X(t)dt + \Sigma dW(t), \qquad X(0) = \xi, \tag{24}$$

where $W = (W(t) : t \geq 0)$ *is a standard Brownian motion in* $\mathbb{R}^m$, *the infinitesimal drift function* $b_i(\cdot)$ *for the* i*th component is*

$$\begin{aligned} b_i(x) &= -\mu_i\gamma_i - \mu_i x_i \qquad i = 1, \ldots, m-1 \\ b_m(x) &= \begin{cases} -\mu_m\gamma_m - \mu_m x_m & \sum_{i=1}^m x_i \leq 0 \\ -mu_m\gamma_m + \sum_{i=1}^{m-1} x_i & \sum_{i=1}^m x_i > 0 \end{cases} \end{aligned}$$

and $\Sigma \equiv diag(\sigma_1, \sigma_2, \ldots, \sigma_m)$, *with* $\sigma_i^2 = 2\mu_i\kappa_i$.

The first $m-1$ components of X are simple O-U processes and clearly admits a steady-state, however the last component has more complicated structure, and thus is not clear *a priori* under what conditions a steady-state will exist for X. Maglaras and Zeevi proved X is positive recurrent and that it admits a unique stationary distribution if and only if $\gamma = \sum_{i=1}^m \gamma_i > 0$, which is also the steady-state distribution.

Our objective in this section is to apply our algorithm to approximate the stationary distribution of these type of processes. For this we need first to check the sufficiency and necessity of the BAR condition, and second, to provide tightness bounds for the steady-state distribution.

4.1 Bar Condition

Necessity of the BAR condition comes from the following proposition.

Proposition 2 *Suppose* π *is the stationary distribution for* X. *Then for each* $f \in C_b^2(\mathbb{R}^m)$,

$$\int_{\mathbb{R}^m} \left(\sum_{i=1}^m b_i(x) \frac{\partial f(x)}{\partial x_i} + \frac{\sigma_i^2}{2} \frac{\partial^2 f(x)}{\partial x_i^2} \right) \pi(dx) = 0 \tag{25}$$

Proof First we need to prove that $P\left[\int_0^t |b_i(x)|dx < \infty \quad \forall\, t \geq 0\right] = 1$. If this is the case, then we can apply Ito's lemma, and take expectation E_x to obtain

$$E_x[f(X(t))] - f(x) = E_x\left[\int_0^t \left(\sum_{i=1}^m b_i(X(s)) \frac{\partial f(x)}{\partial x_i} + \frac{\sigma_i^2}{2} \frac{\partial^2 f(x)}{\partial x_i^2} \right) ds)\right] + E_x\left[\int_0^t \langle \nabla f(X(s))\Sigma, dW(s)\rangle\right]$$

Since $f \in C_b^2(\mathbb{R}^m)$ the last term is a martingale, so its expectation is 0. Integrating both sides with respect to the stationary distribution π, we obtain (25) when applying Fubini's lemma. ■

Sufficiency of the BAR condition comes from the following proposition (see appendix for a proof).

Proposition 3 *Suppose that π is a probability measure on $\mathbb{R}^m$. If π satisfies (25), then π is the stationary distribution for X.*

4.2 Tightness condition

Since X lives on $\mathbb{R}^m$, this application does not involve any boundary measure associated to the stationary distribution. This leaves us with just one "tightness" constraint, the one for the "interior" distribution. In exchange the simple sum of first moments of each coordinate under the stationary distribution does not imply tightness.

For $i = 1, \ldots, m-1$ consider $f_i(x) = \mu_1^{-1}(x_i - \gamma_i)$. Applying Ito, taking expectation w.r.t to π, dividing by t, and taking $t \to \infty$ we have that $E_\pi[x_i^2] = \gamma_i^2 + \kappa_i$ $(i = 1, \ldots, m-1)$. Taking $f_i(x) = x_i$ and repeating the same reasoning we have that $E_\pi[x_i] = -\gamma_i$ $(i = 1, \ldots, m-1)$. Therefore we can conclude that $X_i(\infty) \sim N(-\gamma_i, \kappa_i)$. In order to state the tightness constraint, we still need to analyze the last component of X. Consider:

$$f(x) = \sum_{i<m} (1/2) C_{1,i} x_i^2 + \sum_{i<m} (1/2) C_{2,i} x_i + C_3 \sqrt{1 + x_m^2} + C_4$$

where C_4 is set so that f is nonnegative, $c > 0$ is a constant that ensures that $\mathcal{A}f(x) < -\epsilon$ for $x \in K^c$, where K is a compact set, and

$$C_3 = \frac{c + \sum_{i<m} \gamma_i^2}{\gamma \mu_m} \qquad C_{2,i} = (C_3 \mu_m - \gamma_i)/\mu_i \qquad C_{1,i} = 1/\mu_i$$

Maglaras and Zeevi [21] shown that this function is a valid Lyapunov function for X. This function can be used, for example, to establish the finiteness of the moment generating function of the steady state distribution. In particular it can be shown (see Glynn and Zeevi [19]) that

$$E[\sum_{i=1}^{m} |x_i|] \leq \max_{x \in K} \{|\mathcal{A}f(x)|\}$$

which the tightness constraint that we will use.

4.3 Algorithm setting

To apply the algorithm we still need to specify (i) $\{S_n\}$, the sequence of grid for which we will approximate (15), and (ii) H_m, the sequence of finite dimensional subspaces approximating $C_b^2(S)$.

Grid choice

We know that $X_i(\infty) \sim N(-\gamma_i, \kappa_i)(i = 1, \ldots, m-1)$, but we don't have any prior information about the distribution $X_m(\infty)$, but that it has an exponential decay. With this in mind we choose the grid to have an "Normal" spacing on the marginal distribution for X_i $i = 1, \ldots, m-1$, and an "Uniform" spacing on the marginal distribution for X_m:

$$S_n = \{x \in S | \{x_i = (\Phi^{-1}(1/2 + j/(2n)) + \gamma_i)/\sqrt{\kappa_i} \quad j = 0, \pm 1 \ldots, \pm n - 1 \tag{26}$$

$$i = 1, \ldots, m-1\} \times \{x_m = \alpha j/n \quad j = -n, \ldots, n\}\} \tag{27}$$

Subspace choice

For each $m \geq 1$, we choose $H_m = \{f$ polynomial of degree $\leq m\}$.

4.4 Numerical Results

The following example was studied by Maglaras and Zeevi [21], and restrict attention to a two-class system with a single guaranteed and best-effort class. Consider a system with capacity $C = 100$, $\mu_1 = 1$, $\mu_2 = 2$, $\lambda_1 = 47.5$ and $\lambda_2 = 95$ (here $\rho = 0.95$). The idea is to approximate this system using the two dimensional diffusion described in this section. For that purpose we can solve for κ_i and γ_i $(i = 1, \ldots, m)$ using (23). In this example:

$$\kappa_i = \frac{\lambda_i/\mu_i}{\lambda_1/\mu_1 + \lambda_2/\mu_2} \Rightarrow \kappa_1 = 47.5/(47.5 + 95/2) = 0.5 \quad \text{and} \quad \kappa_1 = 1 - \kappa_2 = 0.5$$

and

$$\gamma_i = \frac{\kappa_i \mu_i C - \lambda_i}{\mu_i \sqrt{C}} \Rightarrow \gamma_1 = \frac{0.5 \cdot 1 \cdot 100 - 47.5}{\sqrt{100}} = 0.25 \quad \text{and} \quad \gamma_2 = \frac{0.5 \cdot 2 \cdot 100 - 95}{2 \cdot \sqrt{100}} = 0.25$$

Queue length steady state distribution can be approximated as follows

$$Q_i^C(\infty) = \kappa_i C + \sqrt{C} X_i(\infty) \quad i = 1, \ldots, m$$

Figure 3 compares the approximated marginal distributions resulting from the algorithm (in what follows we will use $m = 100$, $n = 4$) with the actual marginal distribution computed via Monte Carlo simulation (doted line).

This does not look like good approximation to the marginal distributions. However, when comparing moments for the second coordinate, we have that the approximation does a good job. See table 5.

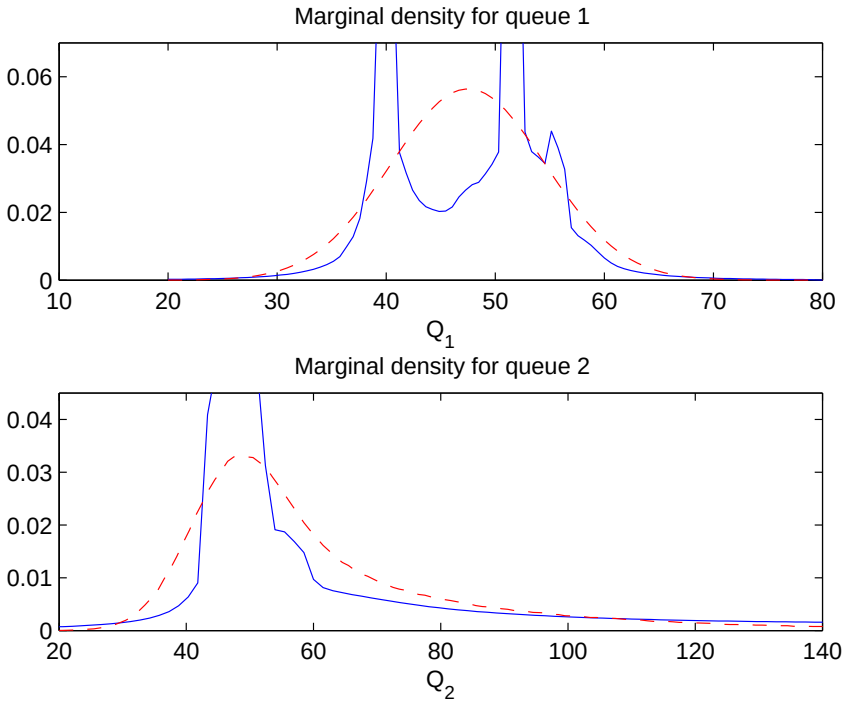

Figure 3: Marginal distribution estimates Two Class Queue ($n = 100$, $m = 6$)

	$E[x_2^1]$	$E[x_2^2]$	$E[x_2^3]$	$E[x_2^4]$
Simulation	1.19	8.31	70.67	854
Algorithm	1.22	9.14	68.2	592
Relative error	0.03	0.10	0.03	0.30

Table 5: Moment approximation for second coordinate Two Class Queue

	$E[x_2^1]$	$E[x_2^2]$	$E[x_2^3]$	$E[x_2^4]$
Simulation	1.19	8.31	70.67	854
Algorithm	1.22	9.14	68.2	592
Smooth Alg.	1.23	8.92	68.7	618
Relative error	0.03	0.07	0.03	0.28

Table 6: Smooth moment approximation for second coordinate Two Class Queue

We would expect for the actual steady state distribution to have "smooth" marginal distributions, but this condition is not explicitly imposed on the LP formulation. It turns out that the approximated BAR condition can not rule out such erratic behavior on the marginal distributions, at least not for such low values of m and n. Also, we would expect for a discretization of the stationary distribution to be both feasible and to achieve an objective vale close to the optimal one, and to have an "smooth" behavior. With this in mind we decided to impose "smoothness" on the algorithm's output. Consider a neighborhood parameter $r > 0$, and a "smoothness" parameter δ. We will impose that the probability assigned to two points on S^n whose distance (euclidian norm) from each other is less or equal to r should not be greater in absolute value to δ

$$|\pi_m^n(x) - \pi_m^n(y)| \leq \delta \quad \forall\, (x, y) \in S^n \quad s.t.\, ||x - y|| < r \tag{28}$$

Remark The choice of r and δ will be case dependent. The greater the value of r is, the greater is the number of constraints added. This will affect the complexity of the LP. On the other hand, low values for δ will reduce the space of feasible solutions, and at some point it could even discard a true steady-state distribution discretization.

In practice we introduce these constraints on each marginal distribution separately. We choose r in order to include just one neighbor in each marginal, and δ low enough to rule out peaks as in figure 3. For $\delta_1 = 0.0026$ (smoothness parameter for the first coordinate marginal) and $\delta_2 = 0.0011$ (smoothness parameter for the first coordinate marginal) we obtain the moment approximation shown on Table 6. Figure 4 shows the marginal obtained including the smoothing constraints.

5 Robust Steady State Estimates (?)

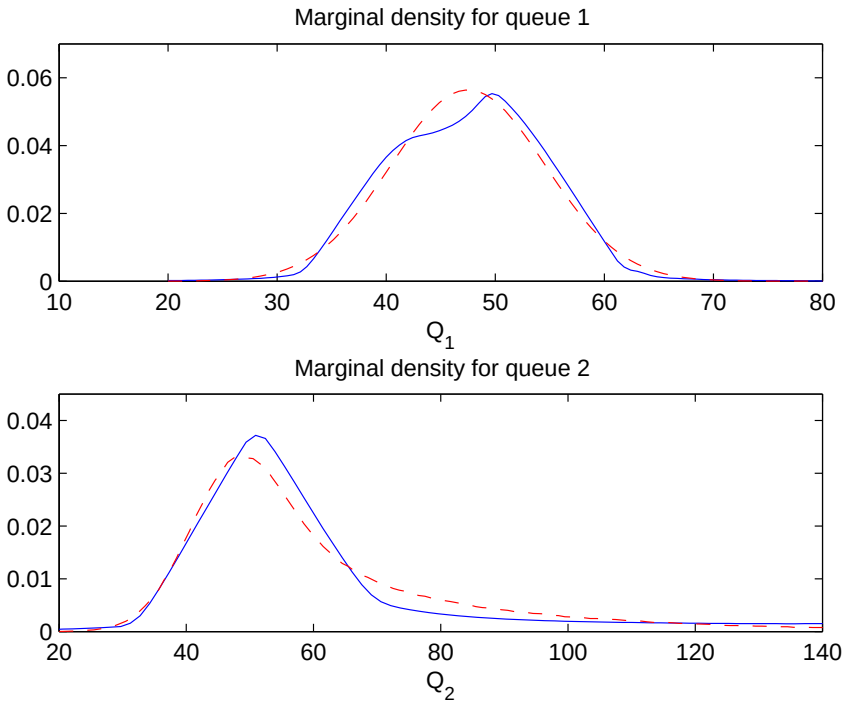

Figure 4: Smooth marginal distribution estimates Two Class Queue ($n = 100$, $m = 6$)

A Proofs

A.1 Proof of Theorem 3

Define $\gamma = -R^{-1}\mu > 0$. Now, take a d-dimensional vector $v > 0$ such that $R^t v > 0$ (the *completely* S is close under the transpose and inverse operations). Fix $a > 0$ and let $f(x) = \exp(ax^t v)$. Now let $f_n \in C_b^2(S)$ be a positive function that agrees with f for $x \in B_S(n) \equiv \{x \in S, |x| \leq n\}$. Applying Itô's formula, we have that

$$f_n(Z(t)) - f_n(z(0)) - \int_0^t Lf_n(Z(s))ds - \sum_{i=1}^{d} \int_0^t D_i f_n(Z(s))dY_i(s)$$

is a P_x martingale. Let $T_n \equiv \inf\{t \geq 0, |Z(t)| \geq n\}$. Since $f_n > 0$ we have

$$E_x[f(Z(t \wedge T_n))] \geq E_x[\int_0^{t\wedge T_N} Lf(Z(s))ds] + E_x[\sum_{i=1}^{d} \int_0^{t\wedge T_N} D_i f(Z(s))dY_i(s)]$$

For this particular function we have $Lf(x) = f(x)\left[-av^t R \cdot \gamma + \frac{a^2}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}\Gamma_{ij}v_i v_j\right]$ and $D_i f(x) = af(x)[v^t R_{\cdot i}]$. Taking $a < \frac{2v^t R\cdot\gamma}{\sum_{i=1}^{d}\sum_{j=1}^{d}\Gamma_{ij}v_i v_j}$ we have

$$E_x[f(Z(t\wedge T_n))]+\alpha E_x[\int_0^{t\wedge T_N} f(Z(s))ds]\geq E_x[\sum_{i=1}^{d}\beta_i\int_0^{t\wedge T_N} f(Z(s))dY_i(s)]$$

where $\alpha\equiv av^tR\cdot\gamma-\frac{a^2}{2}\sum_{i=1}^{d}\sum_{j=1}^{d}\Gamma_{ij}v_iv_j>0$ and $\beta_i\equiv v^tR_{.i}>0\quad i=1,\ldots,d$. Since $f>0$ and taking $\{f_n\}_{n\in\mathbb{N}}$ increasing, the Monotone convergence theorem implies the following:

$$E_x[f(Z(t))]+\alpha E_x[\int_0^{t} f(Z(s))ds]\geq E_x[\sum_{i=1}^{d}\beta_i\int_0^{t} f(Z(s))dY_i(s)]$$

Taking expectations with respect to π (and considering $a>0$ small enough for $E_\pi[e^{av^tx}]$ to exist) we have that

$$E_\pi[f]+\alpha tE_\pi[f]\geq t\sum_{i=1}^{d}\beta_i\int_{F_i} f\nu_i(ds)$$

where we have used proposition 1, the definition of the stationary distribution and Tonelli's theorem. Dividing by t and taking $t\to\infty$ we conclude the following

$$\int_{F_i} e^{av^tx}\nu_i(ds)\leq\frac{\alpha}{\beta_i}E_\pi[e^{av^tx}]<\infty,\ i=1\ldots,d$$

A.2 Proof of Proposition 3

First we will need some side results. For $f\in C^2(\mathbb{R}^m)$ define the operator $\mathcal{A}$ as follows

$$\mathcal{A}f\equiv\sum_{i=1}^{m}b_i(x)\frac{\partial f(x)}{\partial x_i}+\frac{\sigma_i^2}{2}\frac{\partial^2 f(x)}{\partial x_i^2}$$

Lemma 1 *The operator $(\mathcal{A},C_b^2(\mathbb{R}^m)$ satisfies the positive maximum principle, i.e., whenever $f\in C_b^2(\mathbb{R}^m)$, $x_0\in\mathbb{R}^m$, and $\sup_{x\in\mathbb{R}^m}f(x)=f(x_0)\geq 0$, we have $\mathcal{A}f(x_0)\leq 0$.*

Proof Consider $f\in C_b^2(\mathbb{R}^m)$. Applying Ito

$$f(X(t))-\int_0^t\mathcal{A}f(X(s))ds=f(x_0)+\int_0^t\langle\nabla f(X(s))\Sigma,dW(s)\rangle \tag{29}$$

Now take expectations w.r.t. P_{x_0}, we have

$$E_{x_0}[f(X(t))]-f(x_0)=E_{x_0}\left[\int_0^t\mathcal{A}f(X(s))ds\right] \tag{30}$$

Dividing by t and taking $t\to 0$ by the continuity of $\mathcal{A}f$ and the continuity of X, we get $\mathcal{A}f(x_0)\leq 0$, since the left hand side above is non-positive. ■

Proposition 4 *For any probability measure π on $\mathbb{R}^m$, the measure $P_\pi \equiv \int_{\mathbb{R}^m} P_x \pi(dx)$ is a solution for the martingale problem for $(\mathcal{A}, \pi)$.*

Proof It is enough to show that for $f \in C_b^2(\mathbb{R}^m)$ and each $x \in \mathbb{R}^m$,

$$f(X(t)) - \int_0^t \mathcal{A}f(X(s))ds \tag{31}$$

is a P_x-martingale. However, this is direct a consequence of (29). ■

Proof of Proposition 3. Clearly $C_b^2(\mathbb{R}^m)$ is an algebra, and also it is dense in $C(\mathbb{R}^m)$ (**CITE**). By Lemma 1, the operator $(\mathcal{A}, C_b^2(\mathbb{R}^m)$ satisfies the positive maximum principle, and therefore Echeverria's theorem (see [23], Theorem 9.14 of Chapter 4) applies to assert that π is a stationary distribution for a solution of the martingale problem for $(\mathcal{A}, \pi)$. By proposition 4 we know that P_π is a solution of the martingale problem for $(\mathcal{A}, \pi)$, and this solution is unique (since (24) admits a unique strong solution, see [21]. This implies that π is a stationary distribution for X.The results follows from the uniqueness of the stationary distribution.

B Numerical Results

7 Pages of numerical results for the SRBM?????

Acknowledgements. We are indebted to...

References

[1] O. Hernandez-Lerma and J. B. Lasserre. Markov Chains and Invariant Probabilities. *Progress in Mathematics*, Vol 211, Birkhäuser.

[2] D. Gamarnik and A. Zeevi. Validity of heavy traffic steady-state approximations in generalized Jackson networks, in *Annals of Applied Probability*, Vol 16, No. 1, 2006.

[3] I. Gurvich and A. Zeevi. Validity of Heavy-Traffic Steady-State Approximations in Open Queueing Networks: Sufficient conditions involving state-space collapse, forthcoming.

[4] J. M. Harrison and M. Reiman. Reflected Brownian Motion on an Orthan, in *The Annals of Probability*, Vol. 9, No. 2, 1981.

[5] M. I. Reiman. Open Queueing Networks in Heavy Traffic, in *Mathematics of Operations Research*, Vol. 9, No. 3, 1984.

[6] J. G. Dai. Steady-state analysis of reflected Brownian motions: Characterization, numerical methods and queueing applications. Ph.D. dissertation, Dept. Math. Stanford University, 1990.

[7] J. G. Dai and J. M. Harrison. Reflected Brownian motion in an orthant: Numerical methods for steady-state analysis, in *The Annals of Applied probability*, Vol 2, No. 1, 1992.

[8] J.G. Dai and T. G. Kurtz. The sufficiency of the basic adjoint relationship. Unpublished manuscript.

[9] J. M. Harrison and R. J. Williams. Brownian models of open queueing networks with homogeneous customer populations, in *Stochastics*, 1987.

[10] J. M. Harrison and R. J. Williams. Multidimensional reflected Brownian motions having exponential stationary distributions, in *Annals of Applied Probability*, 1987.

[11] J. M. Harrison and V. Nguyen. The QNET method for two-moment analysis of open queueing networks, in *Queueing Systems: Theory and Applications*, 1990.

[12] L.M. Taylor and R. J. Williams. Existence and uniqueness of semimartingale reflecting Brownian motion in an orthant, in *Probability Theory and Related Fields*, 1993.

[13] B. S. Greenberg. Queueing systems with returning customers and the order of tandem queues. Ph.D. dissertation, University of California, Berkeley, 1986.

[14] J. M. Harrison. The diffusion approximation for tandem queues in heavy traffic, in *Advances in Applied Probability*, 1978.

[15] A. Budhiraja and C. Lee. Long time asymptotics for constrained diffusions in polyhedral domains, in *Stochastic Processes and their Applications*,doi:10.1016/j.spa.2006.11.007.

[16] M. I. Reiman and R. J. Williams. A boundary property of semimartingale reflecting Brownian motions, in *Probability Theory and Related Fields*, 1988.

[17] P. Dupuis and R. J. Williams. Lyapunov functions for semimartingale Brownian motions, in *Annals of Probability*, 1994.

[18] J. G. Dai and J. M. Harrison. Reflecting Brownian motion in the orthan: An illuminating example of inestability, forthcoming.

[19] P. W. Glynn and A. Zeevi. Bounding Stationary Expectations of Markov Procesess, forthcomming.

[20] S. Suresh and W. Whitt. Arranging queues in series: a simulation experiment, in *Management Sciences*, Vol. 9, No. 3, 1990.

[21] C. Maglaras and A. Zeevi. Diffusion Approximations for a Multiclass Markovian Service System, in *Mathematics of Operations Research*, Vol. 29, No. 4, 2004.

[22] S. Halfin and W. Whitt. Heavy-traffic limits for queues with many exponential servers, in *Operations Research*, Vol 29, 1981.

[23] S. N. Either and T. G. Kurtz. Markov Processes, Characterization and Convergence. *Wiley Series in Probability and Mathematical Statistics.*

[24] X. Shen, H. Chen, J. Dai and W. Dai. The Finite Element method for Computing the Stationary Distribution of an SRBM in a Hypercube with Applications to Finite Buffer Queueing Networds, in *Queueing Systems: Theory and Applications*, Vol 42, 2002

[25] X. Shen, H. Chen. Computing the stationary distribution of an SRBM in an orthant with applications to queueing networks, in *Queueing Systems*, Vol 45, 2003

[26] A. Manne. Linear programming and sequencial desicions, in *Management Science*, Vol. 6, 1960.

[27] M. Mendiondo, R. Stockbridge. Approximation of Infinite-Dimensional Linear Programming Problems which Arise in Stochastic Control, in *Journal of Control Optimization*, Vol. 36, 1998.

[28] D. Bertsimas, J. Tsitsiklis and D. Gamarnik. Performance of Multiclass Markovian Queueing Networks via Piecewise Linear LYapunov Functions, in *Annals of Applied Probability,* Vol. 11, 2001.

[29] S. Kumar and P. Kumar. Performance bounds for queueing networks and scheduling policies , in *IEEE Trans. on Automatic Control,* Vol. 39, 1994.

[30] K. Sigman and D. Yao. Finite Moments for Inventory Processes, in *Annals of Applied Probability*, Vol. 3, 1993.

[31] M.Reiman, A. Mandelbaum and W. Massey. Strong Approximations for Markovian Service Networks, in *Queueing Systems*, Vol. 30, 1998.