Semidefinite Optimization for Transient Analysis of Queues

Takayuki Osogami
IBM Research - Tokyo
1623-14 Shimotsuruma
Yamato-shi, Kanagawa 242-8502, Japan
osogami@jp.ibm.com

Rudy Raymond
IBM Research - Tokyo
1623-14 Shimotsuruma
Yamato-shi, Kanagawa 242-8502, Japan
raymond@jp.ibm.com

ABSTRACT
We derive an upper bound on the tail distribution of the transient waiting time for the GI/GI/1 queue from a formulation of semidefinite programming (SDP). Our upper bounds are expressed in closed forms using the first two moments of the service time and the interarrival time. The upper bounds on the tail distributions are integrated to obtain the upper bounds on the corresponding expectations. We also extend the formulation of the SDP, using the higher moments of the service time and the interarrival time, and calculate upper bounds and lower bounds numerically.

Categories and Subject Descriptors
D.4.8 [Performance]: Stochastic Analysis

General Terms
Theory

Keywords
Semidefinite programming, Duality, G/G/1 queue, Transient, Bounds, Moments, Occupation measure.

1. INTRODUCTION
Bounds on the steady-state waiting time, $W$, have been studied extensively in the literature. For example, Bertsimas and Natarajan [1] propose a computational approach for calculating the bounds on the moments and the tail distribution of $W$ in the GI/G1/1/FCFS queue. Their bounds are calculated based on the moments of the service time and the interarrival time. The bounds on the tail distribution, $\Pr(W > y)$, in terms of the moments are in contrast to the traditional form of exponential bounds.

We derive bounds on the transient waiting times in the GI/GI/1 queue. In particular, we derive an upper bound on $\Pr(W_n > y)$ in a closed form using the first two moments of the service time and the interarrival time (see Theorem 1). The upper bound on $\Pr(W_n > y)$ can be integrated to obtain an upper bound on $\mathbb{E}[W_n]$ (see Corollary 1). These upper bounds in closed forms and our approach of deriving the bounds constitute our primary contributions.

Our upper bound on a tail distribution is derived by first formulating a (primal) semidefinite programming (SDP) problem, where the objective function is the tail distribution, and the objective value of the optimal solution to the maximization problem corresponds to the upper bound. We then construct a feasible solution to the dual problem associated with the SDP and obtain a closed-form expression for the objective value of the feasible solution. Since the objective value of the feasible solution of the dual problem is no smaller than the objective value of the optimal solution to the primal SDP, we thus find an upper bound on the tail distribution in a closed form. We can extend the formulation of the primal SDP, using the higher moments of the service time and the interarrival time, which however make it difficult to derive a closed-form solution to the dual problem.

Alternatively, we can solve the primal SDP numerically, which gives us numerical bounds on the tail distribution. Note that the SDP can be solved in polynomial time with interior point methods, and reliable solvers are available. Our secondary contributions are numerical evaluations of the upper and lower bounds on $\Pr(W_n > y)$. We find that our bounds can be made tighter than the bounds derived from the standard theory of large deviations or martingales with Wald’s identity when sufficiently many moments are incorporated in the SDP (see Figure 2).

The approach of formulating an SDP to obtain a bound on performance has been applied for a few queueing models in the literature, including [1]. Our approach differs from the existing approaches in that we use the positive semidefinite conditions for the moments of an occupation measure. The use of the occupation measure allows us to obtain bounds on transient waiting time. Also, we obtain bounds in closed forms from a dual SDP, while the bounds are calculated numerically from a primal SDP in the prior work.

2. MAIN RESULTS

Theorem 1. Consider the GI/GI/1 queue. Let $C_A$ (respectively, $C_S$) be the coefficient of variation for the interarrival time (respectively, service time), $\lambda$ be the arrival rate, and $\rho$ be the load of the system. Let $W_i$ be the waiting time of the $i$-th job when the waiting time of the 0-th job is $W_0 = 0$. Let $x_1 = -(1 - \rho)/\lambda$ and $\sigma_X^2 = (C_S^2 \rho^2 + C_A^2)/\lambda^2$. If $\rho \geq 1$ and $y > n x_1$, then

$$\Pr(W_n \geq y) \leq \frac{n \sigma_X^2}{n \sigma_X^2 + (n x_1 - y)^2}. \tag{1}$$

If $\rho < 1$, then

$$\Pr(W_n \geq y) \leq \frac{1}{2} + \sqrt{\left(\frac{1}{2} + \eta\right)^2 + \frac{4 y x_1}{\sigma_X^2} - \eta}. \tag{2}$$
where \( \eta \equiv 2 x_1 (n x_1 - y) / \sigma_X^2 \). Improved bounds exist if additional conditions are satisfied: If \( \sigma_X^2 \leq 4 n x_1^2 - 2 y x_1 \), then

\[
\Pr(W_n \geq y) \leq 1 - \frac{y^2}{2 n \sigma_X^2} \left( \sqrt{1 + \frac{4 n \sigma_X^2}{y^2}} - 1 \right). \tag{3}
\]

If \( \sigma_X^2 \geq -2 x_1 (y - n x_1)^2 / (y + n x_1) \) and \( y + n x_1 > 0 \), then (1) holds.

If \( y \leq n x_1 \), then our approach finds only a trivial bound: \( \Pr(W_n \geq y) \leq 1 \).

Corollary 1. Consider the GI/GI/1 queue as defined in Theorem 1 with \( \rho < 1 \). Let \( \kappa \equiv (C_3^2 \rho^2 + C_4^2)/(1 - \rho)^2 \). If \( \gamma / 4 \leq n \), then

\[
E[W_n] \leq \frac{4 \sqrt{n} (C_3^2 \rho^2 + C_4^2)}{3 \lambda}. \tag{4}
\]

If \( \gamma / 16 \leq n \leq \gamma / 4 \), then

\[
E[W_n] \leq \frac{\xi}{4} + \frac{1 - \rho}{\lambda} \ln \left( \frac{\xi^2 \gamma}{4 n} \right) n + \frac{2 (1 - \rho) n^2}{3 \lambda \gamma}. \tag{5}
\]

where \( \xi \equiv (C_3^2 + C_4^2) / (2 \lambda (1 - \rho)) \) is Kingman’s upper bound on \( E[W] \). Let \( U \) be the right-hand side of (5). If \( n \leq \gamma / 16 \), then

\[
E[W_n] \leq U - \frac{(1 - \rho) \sqrt{n}}{\lambda} \left( \frac{1}{2} \phi_n - \arctan \sqrt{\phi_n} \right) - \frac{2 n (1 - \rho)}{\lambda} \ln \left( \sqrt{\phi_n + \sqrt{\phi_n + 1}} \right), \tag{6}
\]

where \( \phi_n = \gamma / (16 n) - 1 \).

In Figure 1, the solid line shows our upper bound on \( E[W_n] \) as a function of \( n \). We set \( C_3^2 = C_4^2 = 16 \) and \( \lambda = 1 \). The load, \( \rho \), is varied as specified in each column. Because we assume \( W_0 = 0 \), we have \( E[W_n] \leq E[W] \), so that an upper bound on \( E[W] \) is also an upper bound on \( E[W_n] \). The dashed line shows Kingman’s upper bound, \( \xi \), on \( E[W] \), and the dashed-dotted line shows Daley’s upper bound, \( \xi' \).

When \( \rho = 0.9 \), \( \xi \) and \( \xi' \) are close to each other and indistinguishable in the figure. Observe that our upper bound on \( E[W_n] \) is smaller than the existing upper bound on \( E[W] \) for a sufficiently small \( n \).

Figure 2 studies our upper bounds on \( \Pr(W_n \geq 4) \) as functions of \( n \). Let \( A \) be the interarrival time and \( S \) be the service time. In the figure, we assume that \( X \equiv S - A \) is a normal random variable with mean \( x_1 = -1/4 \) in Column (a) and \( x_1 = 1/4 \) in Column (b). The variance of \( X \) is fixed at \( \sigma_X^2 = 1 \). The dashed lines show \( \Pr(W_n \geq 4) \) estimated with simulations, where a simulation is repeated 100,000 times for each data point. Dotted lines show the upper bounds from large deviations, which is equivalent to those from martingales with Walski’s identity for the settings under consideration. Solid lines show the upper bounds calculated with the SDP, where \( \kappa \) moments of \( X \) are used to calculate the upper bounds for \( \kappa = 2, 4, 8, 12 \). The bound with a large \( \kappa \) is close to the simulated value.

Observe that our bound is often closer to the simulated value than the bound from large deviations. This is surprising, since the bound from large deviations and the simulated value depend on the distribution (i.e., all moments) of \( X \), while our bound depends only on its several moments. In Column (a), our bound with \( \kappa = 12 \) is closer to the simulated value than the bound from large deviations for \( n \leq 32 \). For a large \( n \), the bound from large deviations is superior to our bound. In Column (b), our bound is closer to the simulated value than that from large deviations for a wider range of \( n \). The bound from large deviations is trivial (i.e., 1.0) for \( 16 \leq n \), while the bound from SDP with \( \kappa = 8 \) is 0.70 for \( n = 16, 0.90 \) for \( n = 32, \) and 0.98 for \( n = 64 \). Overall, we find that our upper bound can be made smaller than that from large deviations by setting \( \kappa \) sufficiently large under the condition that \( n \) is sufficiently small.

Acknowledgments

This work was supported by “Promotion program for Reducing global Environmental load through ICT innovation (PREDICT)” of the Ministry of Internal Affairs and Communications, Japan. Details of the results in this abstract are provided with extensions in a technical report [2].

3. REFERENCES
