Conservation Law and Achievable Region for Tail Probability in 2-class M/G/1 Queue

by

Prof. N. Hemachandra

Joint work with Manu K. Gupta

May 25, 2015

Outline

1. System Description
2. Conservation Law and Achievable Region
   - Approximate Conservation Law
   - Approximate Achievable Region
   - Bounds Computation
3. Numerical Experiments
   - Static Priority Region
   - Dynamic Priority Region

Introduction

- Multi-class queueing systems
  - Customers may differ in arrival and service process
  - To model complex systems.
  - Performance measures of interest: mean waiting time, tail probability, variance of waiting time etc.
  - Applications in wireless and computer communications, transportation and job shop manufacturing systems.
  - Optimal control for efficient system design (See [3] and [7]).

Introduction

- Achievable region and completeness for mean waiting time
- Nice geometric structure (Polytope) for mean waiting time driven by Kleinrock's conservation law [1].
- A parametrized policy is mean waiting time *complete* if it sweeps the entire achievable region.
- Some mean waiting time *complete* policies do exist [2].
- Useful tool in solving optimal control problem (see [3], [8]).

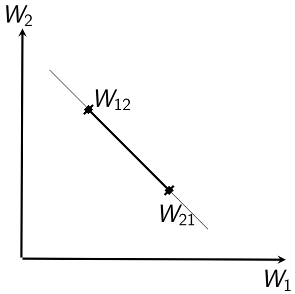

Figure: Achievable region for mean waiting time

Problem Statement

Purpose of the talk

To explore the conservation law and achievable region for waiting time tail probability in two class queues.

Notations

$\mathcal{F}$: Set of all work conserving, non pre-emptive and non anticipative scheduling policies.
π : Scheduling policy in $\mathcal{F}$.
$\bar{W}_i^\pi$: Mean waiting time of class i under scheduling policy π.
N : Number of classes.
λ_i : Independent Poisson arrival rate of class i.
$1/\mu_i$: Mean of the general service distribution of class i.
Achievable region for mean waiting time $\mathcal{W}$:

$$\mathcal{W} = \{(\bar{W}_1^\pi, \bar{W}_2^\pi, \cdots, \bar{W}_N^\pi) : \pi \in \mathcal{F}\}$$

Kleinrock's conservation law [6] is given by

$$\sum_{i=1}^{N} \rho_i \bar{W}_i^\pi = \frac{\rho W_0}{1-\rho} \quad \text{(constant)} \tag{1}$$

Problem Description

- Achievable region for mean waiting time forms a Polytope in N classes (see [1]).
- In case of two classes, it's a line segment.
- A parametrized policy is called mean *complete* if it achieves all achievable vectors of mean waiting time.
- To find conservation law and achievable region for waiting time tail probability in two class queue.
- Mathematically, to study the following space

$$\mathcal{T}_x = \{(P(W_1^\pi > x), P(W_2^\pi > x)) : \pi \in \mathcal{F}\}$$

We study the approximate conservation law and approximate achievable region related to the above set.

Approximate Conservation Law

Approximation of tail probability for class i is given by Yuming Jiang, Chen-Khong Tham, and Chi-Chung Ko [5]:

$$P(W_i^\pi > x) \approx \rho e^{-\rho x/\bar{W}_i^\pi}, i = 1, 2 \tag{2}$$

Approximate waiting time tail probability conservation law is given by:

$$\begin{aligned} & \rho_2 \log P(W_1^\pi > x) + \rho_1 \log P(W_2^\pi > x) \\ + & \frac{\rho^2 x W_0}{(1-\rho) \int_0^\infty P(W_1^\pi > y)dy \int_0^\infty P(W_2^\pi > y)dy} = \rho \log \rho \end{aligned}$$

- Proof follows by approximate tail probability and Kleinrock's conservation law.
- RHS is independent of scheduling policy.

Approximate Achievable Region

- No explicit expression for tail probability of waiting times.
- Subset in unit square, $[0, 1] \times [0, 1]$.
- Achievable region by log transformation.

$$\rho_2 \log P(W_1^\pi > x) + \rho_1 \log P(W_2^\pi > x) =$$

$$\rho \log \rho - \frac{\rho^2 x W_0}{\bar{W}_1^\pi \bar{W}_2^\pi (1 - \rho)} \tag{3}$$

A uniform bound, independent of scheduling policy, can be obtained by solving certain optimization problems.

Upper and Lower Bounds Calculation

P1: $\max\limits_{\mathcal{F}} \quad \bar{W}_1^{\pi} \bar{W}_2^{\pi}$

Subject to

$$\rho_1 \bar{W}_1^{\pi} + \rho_2 \bar{W}_2^{\pi} = \frac{\rho W_0}{1-\rho}$$

P2: $\min\limits_{\mathcal{F}} \quad \bar{W}_1^{\pi} \bar{W}_2^{\pi}$

Subject to

$$\rho_1 \bar{W}_1^{\pi} + \rho_2 \bar{W}_2^{\pi} = \frac{\rho W_0}{1-\rho}$$

- u^* and l^* be the optimal objectives of above optimization problems.
- Upper bound $ub(x)$ and lower bound $lb(x)$ can be obtained as a function of u^* and l^*.

For a given x, approximate achievable region turns out to be included in trapezium by changing scheduling policies.

Illustration

$lb(x)$ and $ub(x)$ are the bounds for

$$\rho_2 \log P(W_1^\pi > x) + \rho_1 \log P(W_2^\pi > x)$$

where

$$lb(x) = \rho \log \rho - \frac{\rho^2 x W_0}{(1-\rho) l^*}$$

$$ub(x) = \rho \log \rho - \frac{\rho^2 x W_0}{(1-\rho) u^*}$$

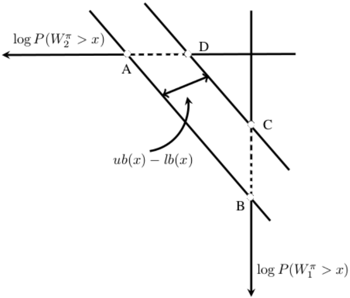

Figure: Approximate achievable performance vectors for tail probability of waiting time

Solve optimization problems to obtain $lb(x)$ and $ub(x)$

Mean Completeness of Relative Priority

Relative priority was first introduced by Moshe Haviv and Van Der Wal [4].

- Each class has independent Poisson arrival rate.
- p_i be the parameter associated with class i.
- Next job is from class i, with probability

$$\frac{n_i p_i}{\sum_{j=1}^{N} n_j p_j}, \quad 1 \leq i \leq N$$

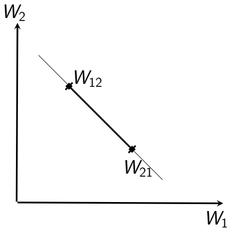

Figure: Achievable region for mean waiting time

Relative priority is mean complete in two classes (See [2]).

Upper Bound Computation

P1: $\max_{\mathcal{F}} \quad \bar{W}_1^{\pi} \bar{W}_2^{\pi}$

Subject to

$$\rho_1 \bar{W}_1^{\pi} + \rho_2 \bar{W}_2^{\pi} = \frac{\rho W_0}{1-\rho}$$

T1: $\max_{0 \leq p \leq 1} \quad \bar{W}_1^{p} \bar{W}_2^{p}$

Subject to

$$\rho_1 \bar{W}_1^{p} + \rho_2 \bar{W}_2^{p} = \frac{\rho W_0}{1-\rho}$$

- $\bar{W}_i^p$ is the mean waiting time of class i with p as relative priority parameter.
- p and $1-p$ are the parameters associated with class 1 and class 2 respectively.

On using the expression of mean waiting time [4],

$$\max_{0 \leq p \leq 1} \frac{(1-\rho p)(1-\rho(1-p))W_0^2}{((1-\rho_1-(1-p)\rho_2)(1-\rho_2-p\rho_1)-p(1-p)\rho_1\rho_2)^2}$$

- Conservation law is trivially satisfied.
- Unconstrained optimization problems as function of p.
- Problem is solved by finding derivatives.
- Stability region is decomposed based on nature of optimizer (pure dynamic or static).

Dynamic Priority Optimality Region

Theorem 3

Pure dynamic policy will be the optimal solution to problem P1 with $p^* = -C_1/C_2$ if $\lambda_1,\ \lambda_2$ and μ are in following stability region D:

$$D \equiv \{\lambda_1,\ \lambda_2,\ \mu : \beta_1 < Y < \beta_2\}$$

where $Y = \rho_1(1-\rho_1) - \rho_2(1-\rho_2)$, $\beta_1 = -\rho^2(1-\rho_2)/2$ and $\beta_2 = \dfrac{\rho^2(1-\rho_2)(1-\rho)}{\rho^2 + 2(1-\rho)}$. And objective function is concave in nature.

- D is obtained by imposing $p^* \in (0,1)$.
- Second derivative of objective function decides nature of objective.

Decomposition of Stability Region

Given $\rho_2 > \rho_1$,

$$\begin{aligned} S_1 &\equiv \{\lambda_1,\ \lambda_2,\ \mu : Y \in (-\infty, \beta_1]\} \\ D_1 &\equiv \{\lambda_1,\ \lambda_2,\ \mu :\ \beta_1 < Y < \beta_2\} \end{aligned}$$

Given $\rho_2 < \rho_1$,

$$\begin{aligned} S_2 &\equiv \{\lambda_1,\ \lambda_2,\ \mu : Y \in [\beta_2, \infty)\} \\ D_2 &\equiv \{\lambda_1,\ \lambda_2,\ \mu :\ \beta_1 < Y < \beta_2\} \end{aligned}$$

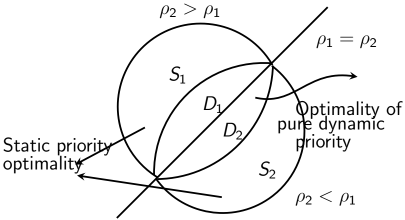

Figure: Decomposition of stability region

$Y,\ \beta_1,\ \beta_2$ are appropriate function of ρ_1 and ρ_2.

Nature of Objective Function

Theorem 4

Objective of optimization problem P1 is monotonically decreasing and increasing in stability region S_1 and S_2 respectively.

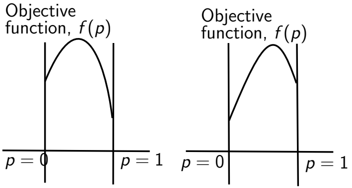

Figure: Nature in Region D_1 and D_2

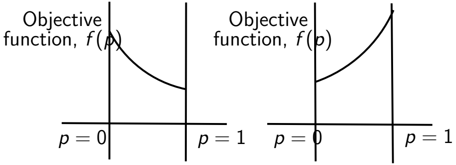

Figure: Nature in Region S_1 and S_2

Tightness of Trapezium

Theorem 5

For a given $x > 0$, approximate achievable region for tail probability, $(P(W_1 > x), P(W_2 > x))$, is a semi open trapezium in 3rd orthand of $\mathbb{R}^2$ bounded by $lb(x) \leq \rho_2 \log P(W_1 > x) + \rho_1 \log P(W_2 > x) \leq ub(x)$ where $lb(x)$ and $ub(x)$ are calculated by ARB algorithm.

- Input date determines the stability region.
- l^* and u^* are computed and accordingly.
- $lb(x)$ and $ub(x)$ are calculated using l^* and u^*.

ARB Algorithm I

Inputs: $\lambda_1,\ \lambda_2,\ \mu,\ x$

1: Determine the stability region among $S_1,\ S_2,\ D_1,\ D_2$ for given input parameters
2: **if** $\lambda_1,\ \lambda_2,\ \mu \in S_1$ **then**
3: $\quad l^* = \bar{W}_1\bar{W}_2|_{p=1}$ and $u^* = \bar{W}_1\bar{W}_2|_{p=0}$
4: **else if** $\lambda_1,\ \lambda_2,\ \mu \in S_2$ **then**
5: $\quad l^* = \bar{W}_1\bar{W}_2|_{p=1}$ and $u^* = \bar{W}_1\bar{W}_2|_{p=0}$
6: **else if** $\lambda_1,\ \lambda_2,\ \mu \in D_1$ **then**
7: $\quad l^* = \bar{W}_1\bar{W}_2|_{p=1}$ and $u^* = \bar{W}_1\bar{W}_2|_{p=-C_1/C_2}$
8: **else if** $\lambda_1,\ \lambda_2,\ \mu \in D_2$ **then**
9: $\quad l^* = \bar{W}_1\bar{W}_2|_{p=0}$ and $u^* = \bar{W}_1\bar{W}_2|_{p=-C_1/C_2}$
10: **else if** $\rho_1 = \rho_2$ **then**
11: $\quad l^* = \bar{W}_1\bar{W}_2|_{p=0}$ or $\bar{W}_1\bar{W}_2|_{p=1}$ and $u^* = \bar{W}_1\bar{W}_2|_{p=1/2}$

ARB Algorithm II

12: Compute $\bar{W}_1\bar{W}_2$ as below to calculate l^* and u^*

$$\bar{W}_1^\pi \bar{W}_2^\pi|_{p=0} = \frac{W_0^2}{(1-\rho)(1-\rho_2)^2}, \ \bar{W}_1^\pi \bar{W}_2^\pi|_{p=1} =$$

$$\frac{W_0^2}{(1-\rho)(1-\rho_1)^2} \text{ and } \bar{W}_1^\pi \bar{W}_2^\pi|_{p=\frac{1}{2}} = \frac{W_0^2}{(1-2\rho_1)^2}$$

To compute u^* for region D_1 or D_2, calculate $p^* = -C_1/C_2$

Output: $lb(x) = \rho \log \rho - \dfrac{\rho^2 x W_0}{l^*(1-\rho)}$ and

$$ub(x) = \rho \log \rho - \frac{\rho^2 x W_0}{u^*(1-\rho)}$$

Error in Approximation

- Use simulator to check error in approximation.
- Build a simulator for two class queue with relative priority across classes.
- Simulator is build in SimPy, a python based simulator
- Validated using theoretical mean waiting times.
- Computed tail probability via simulation and approximation to check error.

Tail Probability via Simulation

Settings	Priority	Simulation		Approximation		Absolute Difference	
		P($W_1 > 0.5$)	P($W_2 > 0.5$)	P($W_1 > 0.5$)	P($W_2 > 0.5$)	Class 1	Class 2
	$p = 0.1$	0.00463	0.00217	0.00431	0.00204	0.00063	0.00053
$\lambda_1 = 1.5$	$p = 0.4$	0.00406	0.00355	0.00382	0.00319	0.00049	0.00077
$\lambda_2 = 0.5$	$p = 0.8$	0.00353	0.00642	0.00317	0.00532	0.00065	0.00144
	$p = 0.1$	0.14259	0.03869	0.16041	0.04049	0.01782	0.00200
$\lambda_1 = 2$	$p = 0.4$	0.09530	0.06851	0.10022	0.07165	0.00492	0.00314
$\lambda_2 = 4$	$p = 0.8$	0.03021	0.09542	0.03226	0.10666	0.00204	0.01124
	$p = 0.1$	0.55027	0.14350	0.62282	0.15435	0.07254	0.01084
$\lambda_1 = 6$	$p = 0.4$	0.51010	0.42166	0.57207	0.47912	0.06196	0.05746
$\lambda_2 = 3$	$p = 0.8$	0.36855	0.58024	0.39582	0.67987	0.02726	0.09963

Table: Error calculation in tail probability via simulation for $x = 0.5$

Approximations are quiet accurate.

Tightness of Bounds

Difference between upper and lower bound in tail probability conservation law.

$$t(x) := ub(x) - lb(x) = \frac{\rho^2 x W_0 (u^* - l^*)}{(1-\rho) l^* u^*}$$

- Linear in x.
- Closed form expressions for region S_1 and S_2
  - Static policies optimality.

For Stability region S_1,

$$t(x) = (\rho_2 - \rho_1)(2 - \rho_1 - \rho_2)\frac{\rho^2 x}{W_0}$$

Further Results

Stability region S_1

- Log scale axis.
- Green and red points are approximate and simulated tail probabilities respectively.
- Parallel red lines are drawn using above analysis.
- Blue line are the projection of extreme green points on parallel red lines.
- Square is the achievable region

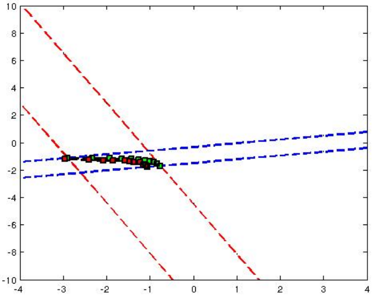

Figure: $\lambda_1 = 1.5,\ \lambda_2 = 5.5,\ \mu = 10$ and tail value, $x = 0.3$

Stability Region S_2

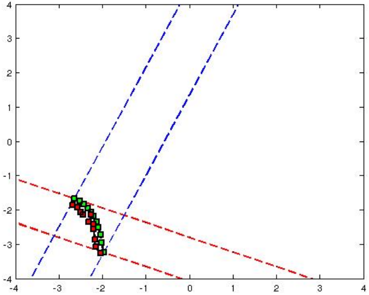

Figure: Tail value, $x = 0.3$

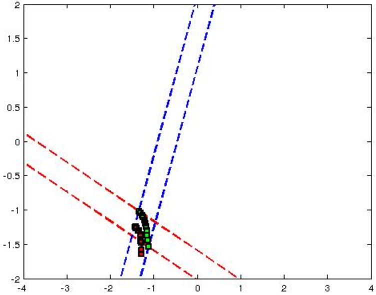

Figure: Tail value, $x = 0.1$

- Linearity of tightness with x.
- $\lambda_1 = 3.5, \ \lambda_2 = 1.5, \ \mu = 10.$

Stability Region D_1 and D_2

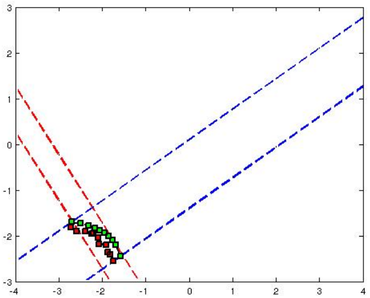

Figure: In region D_1, tail value 0.5, $\lambda_1 = 1,\ \lambda_2 = 1.5,\ \mu = 5$

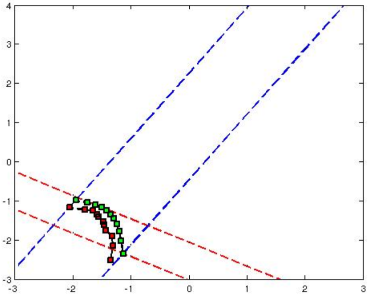

Figure: In region D_2, $\lambda_1 = 2,\ \lambda_2 = 1.2,\ \mu = 5,$ $x = 0.5$

Results

- Tail probability is a non linear curve unlike line segment for mean waiting time.
- Nature of non linearity depends on stability region.
- Relative priority is a complete class for tail probability approximation.

$$P(W_i^\pi > x) \approx \rho e^{-\rho x/\bar{W}_i^\pi}, i = 1, 2 \tag{4}$$

- Any mean waiting time complete class will be tail probability complete for approximation.
- Few extreme points may be outside approximate achievable region.

Some pre-emptive, anticipative queue discipline

- PLIFO and LRPT have variance beyond 2-moment complete range for any ρ.
- PS is beyond 2-moment complete range for $\rho \in (0, \frac{3-\sqrt{5}}{2})$.

Remark

Variance of waiting time can be beyond 2-moment complete range if scheduling policy violates any of the conditions on queue discipline.

Some Applications

Variance minimization problem with lower bound on variance

P1: $\min_{\mathcal{F}} \quad Var(W)$

Subject to

$$Var(W) \geq \gamma$$

T1: $\min_{0 \leq \delta \leq 1} \quad Var(W)$

Subject to

$$Var(W) \geq \gamma$$

- $\mathcal{F}$ is the set of all non pre-emptive, non anticipative and work conserving scheduling policies for $M/M/1$ queue.
- P1 and T1 are equivalent as parametrized queue discipline is 2-moment complete.
- Problem T1 is easy to solve.

Solution depends on γ.

Summary

- Conservation law for tail probabilities.
- Approximate achievable region.
- Achievable Region Bound (ARB) algorithm to compute bounds.
- Error in approximation.
- Expanding these results for multi-class queues.
- Explore optimal control problems using approximate achievable region.

References I

EG Coffman Jr and I Mitrani.
A characterization of waiting time performance realizable by single-server queues.
Operations Research, 28(3-part-ii):810–821, 1980.

Manu K. Gupta, N. Hemachandra, and J. Venkateswaran.
On mean waiting time completeness and equivalence of EDD and HOL-PJ dynamic priority in 2-class M/G/1 queue.
In *8th international conference on performance methodology and tools (Valuetools)*, 2014.

Refael Hassin, Justo Puerto, and Francisco R Fernández.
The use of relative priorities in optimizing the performance of a queueing system.
European Journal of Operational Research, 193(2):476–483, 2009.

Moshe Haviv and Jan van der Wal.
Waiting times in queues with relative priorities.
Operations Research Letters, 35:591 – 594, 2007.

References II

Yuming Jiang, Chen-Khong Tham, and Chi-Chung Ko.
An approximation for waiting time tail probabilities in multiclass systems.
IEEE Communications letters, 5(4):175–177, 2001.

Leonard Kleinrock.
A delay dependent queue discipline.
Naval Research Logistics Quarterly, 11:329–341, 1964.

Chih-ping Li and Michael J Neely.
Delay and rate-optimal control in a multi-class priority queue with adjustable service rates.
In *INFOCOM, Proceedings IEEE*, pages 2976–2980, 2012.

S. K. Sinha, N. Rangaraj, and N. Hemachandra.
Pricing surplus server capacity for mean waiting time sensitive customers.
European Journal of Operational Research, 205:159–171, August 2010.