CONTROL FORMULATIONS IN THE NONDEGENERATE SLOWDOWN DIFFUSION REGIME

Rami Atar
Technion, Israel
atar@ee.technion.ac.il

with

Itai Gurvich
Northwestern, USA

Nir Solomon
Technion, Israel
I. Introduction
Heavy traffic diffusion regimes

Consider a queue with multiple servers

Parametrize by letting

\[ \lambda_n \approx n, \quad N_n \approx n^\alpha, \quad \mu_n \approx n^{1-\alpha}, \]

where \( 0 \leq \alpha \leq 1 \), so that \( \lambda_n \approx N_n \mu_n \).

Obtain:

\[ \alpha = 0 \quad \text{Conventional} \]
\[ \alpha = 1 \quad \text{Halfin-Whitt} \]

Define slowdown = sojourn time / service time

Slowdown is degenerate at both endpoints
When is the slowdown nondegenerate?

Consider $\alpha = 1/2$.

$$\lambda_n \approx n, \quad N_n \approx n^{1/2}, \quad \mu_n \approx n^{1/2}$$

Clearly the service time $\approx n^{-1/2}$

Obtain

$$\text{DELAY} \sim \text{SERVICE TIME}$$
Earlier work (the case of M/M/N):

* Whitt (Oper. Res., 2003): Convergence of queuelength and delay processes to a RBM ($\alpha = 1/2$)

* Mandelbaum and Shaikhet (Mandelbaum’s EURANDOM lecture notes, 2003): independently, a similar result, ($\alpha = 1/2$); observe that the delay and the time in service are of the same order

* Gurvich (M.Sc. Thesis, 2004): Convergence of queuelength/delay processes to a RBM for $\alpha \in [\frac{1}{2}, 1)$.

The above works regard this as a part of the **Efficiency Driven regime** (the diffusion being RBM, the probability of delay being close to 1)
Our point of view

* The joint law of delay and time in service is interesting

* $\alpha = 1/2$ is the only case where the limit is a nondegenerate pair of processes

* The limiting joint law (and in particular the limiting sojourn time law) is distinct from that under the other two diffusion regimes

We will refer to it as the Non-Degenerate Slowdown (NDS) regime
II. Some diffusion limit results
Model

renewal arrivals each requiring a single non-interruptible service

routing mechanism

$N_n$ heterogenous exponential servers

$\mu_1 \quad \mu_2 \quad \cdots \quad \mu_{N_n}$
Assumptions. The NDS Regime ($\alpha = 1/2$)

- **Arrivals**: $\lambda_n = \lambda n + \hat{\lambda} n^{1/2} + o(n^{1/2})$
- **Number of servers**: $N_n = n^{1/2} + o(n^{1/2})$
- **Individual service rates**: $\mu_1 n, \mu_2 n, \ldots, \mu_N n$
- With $\mu_n = \sum_{k=1}^{N_n} \mu_{kn}$, $n^{-1} \mu_n \to \mu \in (0, \infty)$  
  $\hat{\mu}_n = n^{-1/2} (\mu_n - n \mu) \to \hat{\mu} \in (-\infty, \infty)$
- **Critical load condition**: $\lambda = \mu$
Assumptions (cont.)

• The empirical measure of \( \{ \hat{\mu}_{kn} := \mu_{kn} n^{-1/2} \} \) converges weakly, namely

\[
\frac{1}{N_n} \sum_{k=1}^{N_n} \delta \hat{\mu}_{kn} \to m,
\]

for some probability measure \( m \) on \( \mathbb{R}_+ \).
Assumptions on the routing policy

- Work conserving
- Nonanticipating

Includes, for example,

- Always route to the slowest available server
- Always route to the fastest available server
Processes of interest

\( \Delta_n(t) = \) delay experienced by the first customer to arrive at or after time \( t \)

\( \Sigma_n(t) = \) time in service of the same customer

Diffusion scaling:

\[ \hat{\Delta}_n = n^{1/2} \Delta_n \quad \hat{\Sigma}_n = n^{1/2} \Sigma_n \]
AT = Arrival Time
RT = Routing Time
DEP = Departure Time
AB = Abandonment Time
\( \Delta \) = Delay
\( \Sigma \) = Service Time
THEOREM: The joint law of \((\hat{\Delta}_n, \hat{\Sigma}_n)\) converges to

\((\text{RBM}, f\text{-White noise})\)

in finite dimensional distributions. That is, given \(j\) and \(0 < t_1 < \cdots < t_j < \infty\), we have

\[
(\hat{\Delta}_n(t_1), \hat{\Sigma}_n(t_1), \ldots, \hat{\Delta}_n(t_j), \hat{\Sigma}_n(t_j)) \Rightarrow (\bar{\xi}(t_1), \eta_1, \ldots, \bar{\xi}(t_j), \eta_j),
\]

were, \(\bar{\xi} = \xi / \mu\), \(\xi\) is the RBM

\[
\xi(t) = \xi_0 + (\hat{\lambda} - \hat{\mu})t + \sigma w(t) + l(t),
\]

and \(\eta_i\) are independent of \(\xi\), i.i.d., with p.d.f.

\[
f(x) = \frac{1}{\mu} \int y^2 e^{-y x} m(dy), \quad x \in [0, \infty).
\]
Interpretation of $f$

* Draw a random variable $Y$ from the distribution

$$\frac{ym(dy)}{\int zm(dz)}$$

* Let $\eta$ be exponentially distributed with mean $Y$. 
Extension to case with abandonment

Customers abandon the queue while waiting to be served, at fixed rate \( \gamma \) (according to an exponential clock).

The result holds, with

\[
\xi(t) = \xi_0 + (\hat{\lambda} - \hat{\mu})t - \gamma \int_0^t \xi(s)ds + \sigma w(t) + l(t)
\]
Expressions for slowdown (formal)

Without abandonment ($\gamma = 0$) need to assume $\hat{\lambda} - \hat{\mu} < 0$, and then

$$\text{slowdown} = 1 + \frac{\sigma^2}{2(\hat{\mu} - \hat{\lambda})}$$

With abandonment ($\gamma > 0$)

$$\text{slowdown} = 1 + \frac{\int_0^\infty x e^{-\frac{(x-b)^2}{2c^2}} \, dx}{\int_0^\infty e^{-\frac{(x-b)^2}{2c^2}} \, dx}$$

where $(b, c^2) = \left(\frac{\lambda - \mu}{\gamma}, \frac{\sigma^2}{2\gamma}\right)$. 

III. Control formulations
Control to minimize sojourn time

- As a diffusion-limited control problem, this set us is meaningful only in the NDS regime
The heavy traffic condition

Following Harrison and Lopez (1999), consider the linear program

Minimize $\rho \in [0, 1]$ s.t. $\sum_j \mu_{ij} \xi_{ij} = \lambda_i, \ \forall i, \ \xi_{ij} \geq 0, \ \forall (i, j), \ \sum_i \xi_{ij} \leq \rho, \ \forall j$

The HT condition: There exists a unique optimal solution $(\xi^*, \rho^*)$, $\rho^* = 1$. Moreover, $\sum_i \xi_{ij}^* = 1$
The complete resource pooling condition

\[ i \sim j \] — an activity
\[ \xi^*_{ij} > 0 \] — a basic activity

The CRP condition:
* Uniqueness of solutions to a dual program (Harrison and Lopez 1999)
* The graph \( G_b \), of basic activities, is connected (Harrison and Lopez 1999)
* The graph \( G_b \) is a tree (Williams 2000)

Significance:
* High level of cooperation between service stations, so stations work like a single super-server
* Workload is one-dimensional
The diffusion scaling

Denote

\[ Q^n_i(t) = \text{number of class-}i \text{ customers in the queue at time } t \]

\[ X^n_i(t) = \text{number of class-}i \text{ customers in the system at time } t \]

\[ \hat{Q}^n_i(t) = n^{-1/2} Q^n(t), \quad i = 1, 2, \ldots, I \]

\[ \hat{X}^n_i(t) = n^{-1/2} \left( X^n_i(t) - \sum_j \xi^*_i j N^n_j \right), \quad i = 1, 2, \ldots, I \]
The diffusion control problem (Harrison-Lopez 1999)

The DCP consists of r.v.s $X_{0,i}$, BMs $W_i$, and processes $X_i, I_j, Y_{ij}$:

$$X_i(t) = X_{0,i} + W_i(t) + \sum_j \mu_{ij} Y_{ij}(t) \geq 0, \quad t \geq 0, i = 1, 2, \ldots, I,$$

$$I_j := \sum_i Y_{ij} \text{ is non-decreasing and } I_j(0) \geq 0, \quad j = 1, 2, \ldots, J,$$

$$Y_{ij} \text{ is non-increasing and } Y_{ij} \leq 0, \quad (i, j) \in \mathcal{E}_{nb}.$$

REM: $Y_{ij}$ are further required in Harrison-Lopez to be adapted; one can drop this requirement (Bell-Williams 2000)
An equivalent DCP

Harrison-Lopez 1999, Mandelbaum-Stolyar 2004

\[ X(t) = X_0 + W(t) + Z(t) \in \mathbb{R}^I_+, \quad t \geq 0, \]

\[ \theta' Z \text{ is nondecreasing, and } \theta' Z(0) \geq 0 \]

Here, \( \theta \in \mathbb{R}^I_+ \) is a fixed vector (the workload vector).

**THEOREM (with Itai Gurvich):** The two diffusion control problems are equivalent.
IV. DCP for sojourn time -
an explicit solution
DCP for sojourn time

CASE OF A SINGLE POOL

* Nonlinear cost is of interest

We will consider \( \text{COST} = \sum_i c_i E \left[ \left( \frac{X_i(t)}{\mu_i} + \Sigma_i \right)^2 \right] \)

\( \Sigma_i \)-r.v.s representing service time

* Easy to reduce to \( E[C(X(t))] \)
Solution of DCP

Denote

\[
\rho_i = \frac{\lambda_i}{\mu_i}, \quad \beta_i = \frac{\rho_i^2}{c_i}, \quad i = 1, 2, \ldots, I
\]

THEOREM (with Nir Solomon): The DCP is solved by bringing \( X(t) \) to \( X^*(t) \) s.t.

\[
\frac{X_i^* + \rho_i}{\mu_i} = \frac{\beta_i}{\sum_k \beta_k} \sum_k \frac{X_k^* + \rho_k}{\mu_k}, \quad \text{for all } i
\]
V. Asymptotics
Asymptotics, the conventional regime

BACK TO THE GENERAL CASE (general number of pools, $J \geq 1$; general cost $C$)

In conventional heavy traffic:

* Ata-Kumar (2005) - a discretization approach

* Bell-Williams (2001, 2005) - a threshold policy

* Mandelbaum-Stolyar (2004) - a generalized $c\mu$ rule
Asymptotics, the NDS regime

Let $C : \mathbb{R}_+^I \rightarrow \mathbb{R}_+$ be a continuous function, increasing wrt usual partial order

$$C^*(a) = \min\{C(q) : q \in \mathbb{R}_+^I, \theta'q = a\}$$

Let $q(a)$ be a minimizer. Assumption: $q$ is Lipschitz continuous.

PROPOSED POLICY:

<table>
<thead>
<tr>
<th>$X_1$</th>
<th>$X_2$</th>
<th>$X_3$</th>
<th>$X_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>*</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

LEGEND: $X$ --- $q(X)$ *

Priority to overloaded classes

In addition, (i) No use of nonbasic activities, (ii) Work conservation.
THEOREM (with Itai Gurvich): Assume $C$ is convex. Fix a finite $T$. Then under any policy,

$$\liminf_{n \to \infty} \int_0^T C(\hat{Q}^n(t))dt \geq \int_0^T C^*(Q^*(t))dt,$$

where $Q^*$ is the RBM $\Gamma(\theta'X_0 + \theta'W)$.

Moreover, under the proposed policy,

$$\limsup_{n \to \infty} \int_0^T C(\hat{Q}^n(t))dt = \int_0^T C^*(Q^*(t))dt$$
About the lower bound

The LB does not hold in non-integral form.

* Minimality of the Skorohod map is well-known:

Let \( \zeta \in D \). Let \( \eta \in D \) be non-decreasing, \( \eta(0) \geq 0 \). Assume \( \zeta(t) + \eta(t) \geq 0 \), for all \( t \geq 0 \). Then

\[
\zeta(t) + \eta(t) \geq \Gamma[\zeta](t) \equiv \zeta(t) + \sup_{s \leq t} \zeta(s)^-, \quad t \geq 0.
\]
About the lower bound

The integral LB uses the following perturbation lemma about the Skorohod map:

**LEMMA (with Itai Gurvich):** Let $T > 0$ and $\varepsilon > 0$, $\varepsilon < T$, be given. Let $\zeta \in D$ and assume $\zeta(0) \geq 0$. Let

$$\alpha = \zeta + \eta + \beta,$$

where $\eta \in D$ is non-decreasing, $\eta(0) \geq 0$, $\beta \in D$ satisfies

$$-\varepsilon^2 \leq \int_0^t \beta(s) \, ds \leq \varepsilon^2 \quad t \in [0, T],$$

and $\alpha(t) \geq 0$, $t \in [0, T]$. Then

$$\alpha(t) \geq \Gamma[\zeta](t) + \beta(t) - \text{Osc}(\zeta|_{[0,T]}, \varepsilon) - 3\varepsilon, \quad t \in [0, T].$$
Thank you!