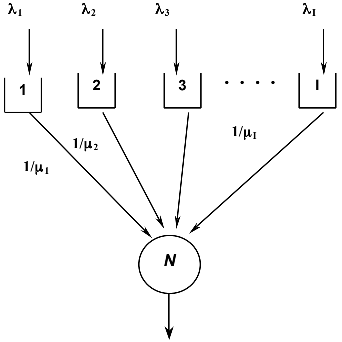
# On optimality gaps in heavy-traffic

Itai Gurvich
Northwestern University

Joint work with Baris Ata

# A multiclass queue

- I customer classes
- $N$ servers
- $Poisson(\lambda_i)$ arrivals
- $Exp(\mu_i)$ service time
- linear holding costs $c_i$




$$V(x) := \inf_{\pi\in\Pi} \mathbb{E}_x \int_0^\infty e^{-\gamma s} \sum_{i=1}^{I} c_i Q_i(s) ds$$

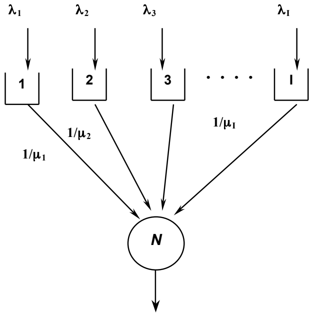
$\Pi$ = Non-preemptive non-anticipative policies

# A multiclass queue in heavy-traffic

The heavy-traffic regime:

$$N^n = \sum_{i=1}^{I} \frac{n\lambda_i}{\mu_i} + \beta \sqrt{\sum_{i=1}^{I} \frac{n\lambda_i}{\mu_i}}$$

(Halfin-Whitt regime)




$$V^n(x) := \inf_{\pi^n \in \Pi^n} \mathbb{E}_x \int_0^\infty e^{-\gamma s} \sum_{i=1}^{I} c_i Q_i^n(s) ds$$

$$(1-\rho^n) \sim \frac{1}{\sqrt{n}} \quad \text{hence} \quad \sum_i Q_i^n \sim \sqrt{n} \quad \text{hence} \quad V^n(x) \sim \sqrt{n}$$

# A multi-class queue in heavy-traffic cont.

Find a sequence $\{\pi^n\}$ so that

$$\frac{1}{\sqrt{n}}\mathbb{E}_x \int_0^\infty e^{-\gamma s} c \cdot Q^{n,\pi^n}(s) ds \leq \frac{V^n(x)}{\sqrt{n}} + o(1)$$

Asymptotic optimality established in Atar et. al (04'):

via optimal control of diffusion limit (as $n \to \infty$)

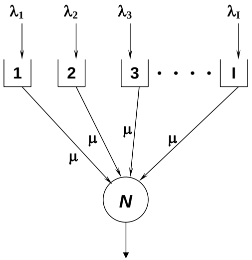
- Optimality gap $= o(\sqrt{n})$
- How to improve gap? (sufficient conditions?)
- Tradeoff simplicity of prescription vs. optimality gap

# Approximation errors in heavy-traffic

- Capacity optimization for multi-server queues: Janssen, van Leeuwaarden and Zwart (08'), Zhang et. al. (09')
- Mandelbaum et. al. (98'), Chen (96')– $O(\log n)$ *performance* bounds via strong approximations

Can strong approximations be preserved under the optimal control process?

# Motivation: A simple case (common service rates)




- Preemptive priorities is optimal

- Non-preemptive static priority is optimal in heavy-traffic

- Optimality gap $= O(1) =$ Non-preemptive - Preemptive

$$\mathbb{E}_x \int_0^\infty e^{-\gamma s} c \cdot Q^{n,\pi^n}(s) ds \leq V^n(x) + K$$

- Proof (almost) ad-hoc for this simple case.

# Main Result

## Theorem

*Fix $n$. We can find a non-preemptive control $\pi^{*,n}$ so that*

$$\mathbb{E}_x \int_0^\infty e^{-\gamma s} c \cdot Q^{n,\pi^{*,n}}(s) ds \leq V^n(x) \left(1 + C \frac{log^m(n)}{\sqrt{n}}\right),$$

*$C$ and $m$ are independent of $n$ and explicitly identifiable.*

- With linear costs: $V^n(x) \sim \sqrt{n}$ so that gap $\sim \log^m(n)$
- Bound in terms of system parameters

# Some preliminaries

Atar et. al: A sequence of controls $\{\pi^n\}$ so that

$$\text{Optimality gap in } n^{th} \text{ system} = o(\sqrt{n}).$$

Proof of asymptotic optimality based on:

(1) Preemptive - Diffusion limit $= o(\sqrt{n})$

(2) Non-Preemptive - Preemptive $= o(\sqrt{n})$

# Some preliminaries

- $X_i^n(t) :=$ # of class-$i$ customers in the system
- $Z_i^n(t) = X_i^n(t) - Q_i^n(t)$ # of class-$i$ customers in service

$$\begin{aligned} X_i^n(t) &= \mathcal{N}_i^a(n\lambda_i t) - \mathcal{N}_i^s\left(\mu_i \int_0^t Z_i^n(s)ds\right) \\ &= \mathcal{N}_i^a(n\lambda_i t) - \mathcal{N}_i^s\left(\mu_i \int_0^t X_i^n(s) - Q_i^n(s)ds\right) \end{aligned}$$

Control= Controlling $Q_i^n(t)$

# Identifying the source of the gap

Departures: $S_i^n(t) = \mathcal{N}_i^s\left(\mu_i \int_0^t Z_i^n(s)ds\right)$

- Strong App: $S_i^n(t) \approx \mu_i \int_0^t Z_i^n(s)ds + W_i\left(\mu_i \int_0^t Z_i^n(s)ds\right)$
- Under $\sqrt{n}$ scaling only fluid appears in Brownian motion
- $Z_i^n(s) - \frac{n\lambda_i}{\mu_i} = O(\sqrt{n})$ so that $\frac{Z_i^n}{n} \to \frac{\lambda_i}{\mu_i}$ as $n \to \infty$
- Diffusion limit: $S_i^n(t) \approx \mu_i \int_0^t Z_i^n(s)ds + W_i\left(\mu_i \frac{n\lambda_i}{\mu_i} t\right)$

$$S_i^n(t) - \mu_i \int_0^t Z_i^n(s)ds - W_i(n\lambda_i) = O(n^{1/4})$$

# A sequence of Diffusion control problems

$$\inf_{\pi \in \hat{\Pi}^n} \mathbb{E}_x \int_0^\infty e^{-\gamma s} c \cdot \tilde{Q}^{n,\pi}(s) ds$$

$$\begin{aligned}
\text{s.t.} \quad (1) \quad \tilde{X}_i^{n,\pi}(t) &= \tilde{X}_i^n(0) + n\lambda_i t - \mu_i \int_0^t (\tilde{X}_i^{n,\pi}(s) - \tilde{Q}_i^{n,\pi}(s)) ds \\
&+ W_i \left( n\lambda_i t + \mu_i \int_0^t \tilde{X}_i^{n,\pi}(s) - \tilde{Q}_i^{n,\pi}(s) ds \right) \\
(2) \quad e \cdot \tilde{Q}^{n,\pi}(t) &= [e \cdot \tilde{X}^{n,\pi}(t) - N^n]^+, \ \tilde{Q}_i^{n,\pi}(t) \geq 0.
\end{aligned}$$

- Key: preserve state and control dependence in Brownian term
- Solve up to a hitting time

# A sequence of HJB equations

For each $n$ we have a different HJB equation.

$$\begin{aligned} 0 &= \inf_{u \geq 0,\ e \cdot u = 1} \left\{ (e \cdot x)^+ \sum_i (c_i + V_i^n(x) - \frac{1}{2} V_{ii}^n(x)) u_i \right\} \\ &+ \sum_i (l_i^n - \mu_i x_i) V_i^n(x) + \frac{1}{2} \sum_i (n\lambda_i + \mu_i(n\rho_i + x_i)) V_{ii}^n - \gamma V^n(x) \end{aligned}$$

These are fully non-linear second order PDEs and non-smooth

# Existence, uniqueness and verification

### Theorem

*Fix $n$. The HJB equation (1) considered on $\Omega^n = B(0, M\sqrt{n}\log n)$ with the boundary condition $\phi = 0$ on $\partial\Omega^n$ is uniquely solvable in $C^2(\Omega^n)\bigcup C^0(\bar{\Omega}^n)$.*

### Theorem

*There exists a unique classical solution $\phi \in C^2_{pol}(\bar{\Omega}^n)$ to the HJB equation. Moreover, the value up to hitting of $\partial\Omega^n$ is equal to $\phi$. Finally, there exists a Markov policy which is optimal.*

# The Markovian control

A proportion function $u_i^n(\cdot)$, such that $\pi^*$ satisfies

$$\frac{Q_i^{n,\pi*}}{\sum_k Q_k^{n,\pi^*}} = u_i^n(\tilde{X}^{n,\pi*}(t)),$$

For the linear-cost case: $u_i^n(x) = 1$ for $i = i^*(x)$

$$i^*(x) := \min \operatorname*{argmin}_i \left\{ (e \cdot x)^+ (c_i + \phi_i^n(x) - \frac{1}{2}\phi_{ii}^n(x)) \right\}$$

where $\phi^n(\cdot)$ is the solution to the $n^{th}$ HJB equation.

Implementable directly to original system via preemption.

# Non-preemptive tracking of Preemptive

Non-preemptive tracking: given Lipschitz(???) functions $u_i^n$, serve $i^*$

$$i^* \in \underset{i}{\operatorname{argmax}} \left\{ \frac{Q_i^n(t)}{\sum_k Q_k^n(t)} - u_i^n(X^n(t))h \right\}$$

# Non-preemptive tracking of Preemptive

## Theorem (a "state-space collapse" result)

*Fix $T > 0$ and use the tracking policy. Then, there exists $C > 0$ s.t.*

$$\mathbb{E}\left[\sup_{0\le t\le T\log n}\left|Q_i^n(t) - u_i^n(X^n(t))\sum_k Q_k^n(t)\right|\right] \le C\log n.$$

Corollary: Let $\pi$ be the tracking policy. Then,

$$\left|\mathbb{E}_x\int_0^\infty e^{-\gamma s}c\cdot Q^{n,\pi}(s)ds - \phi^n(x)\right| \le C\log n.$$

This is not enough

# Towards combining the pieces

"Standard" argument for asymptotic optimality:

(1) $\hat{V}(x)$= value function for *limit* control problem

(2) Show that $\hat{V}(\cdot)$ "almost" solves DP equation for all $n$ large enough.

(3) Uses only continuity of $\hat{V}$ and its derivatives.

We need to bound the gap for fixed $n$

# Gradient Estimates

## Proposition

*Let* $\phi^n$ *be the solution of the* $n^{th}$ *HJB equation. Then, there exists* $M$ *such that with* $\tilde{\Omega}^n = B\left(0, \frac{M}{2}\sqrt{n}\log n\right)$ *s.t.*

(i) $\sup_{x \in \tilde{\Omega}^n} |D\phi^n(x)| \leq C \log n$

(ii) $\sup_{x \in \tilde{\Omega}^n} |D^2\phi^n(x)| \leq \dfrac{C \log n}{\sqrt{n}}$

(iii) $\sup_{x,y \in \tilde{\Omega}^n} \dfrac{|D^2\phi^n(x) - D^2\phi^n(y)|}{|x-y|^\alpha} \leq \dfrac{C \log n}{n}$

*where* $C > 0$ *and* $0 < \alpha \leq 1$ *are independent of* $n$

# Completing the proof

(1) Write Taylor expansion for $\phi^n(X^n(t))$

(2) Plug estimates back into the Taylor expansion, to show that $\phi^n(\cdot)$ is appropriately close to $V^n(\cdot)$.

(3) Use preemptve vs. non-preemptive bounds.

# Summary

- Logarithmic optimality gaps
- A specific case–the V model with linear costs
- Strictly convex cost: we can generate a solution with cost

$$V^n\left(1+C\frac{\log^m n}{\sqrt{n}}\right)$$

- Analysis highlights
  (1) Sources of gaps
  (2) How and when can be tightened.

Questions?