

Queueing systems with many servers: Control theory and heavy traffic asymptotics

Rami Atar

Department of Electrical Engineering

Technion, Haifa

Outline

Queueing model and corresponding diffusion model

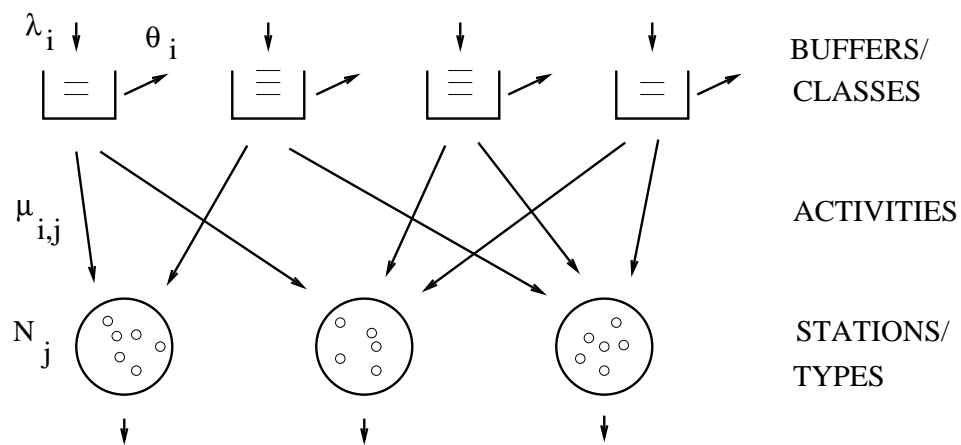
Control theoretic approach

Treelike case: PDE analysis and asymptotic optimality

Beyond the treelike case: null-controllability results

Joint work with A. Mandelbaum, M. Reiman and G. Shaikhet

The model



Classes indexed by $\mathcal{I} = \{1, \dots, I\}$.

Types indexed by: $\mathcal{J} = \{1, \dots, J\}$

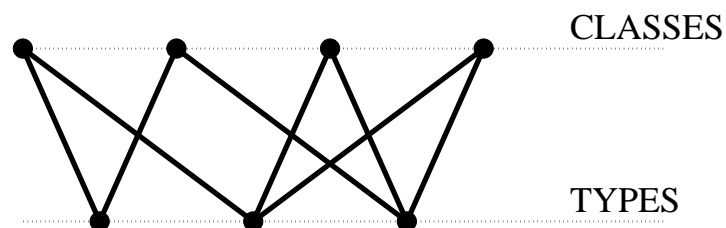
Arrivals: Renewal processes $A_i, i \in \mathcal{I}$, mean interarrival time λ_i^{-1}

Abandonments: At rate θ_i from buffer i .

Service of a class- i customer by type- j server exponential, rate μ_{ij}

Servers per station: N_j servers in station $j \in \mathcal{J}$

\mathcal{G} - graph with **vertex set:** classes and types, **edge set:** activities (pairs (i, j) with $\mu_{ij} > 0$)



Associated with the queueing model are: **a fluid model in heavy traffic** and **a diffusion model**, introduced next.

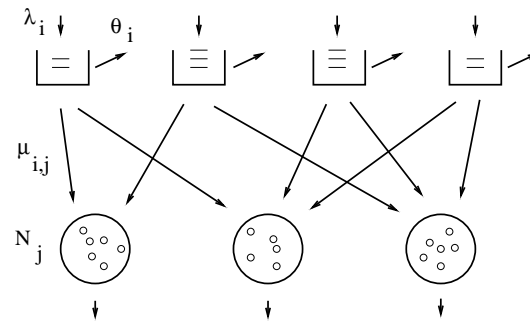
Many server parametrization (first order)

$$\lambda_i^n \sim n\lambda_i$$

$$\theta_i^n \sim \theta_i$$

$$\mu_{ij}^n \sim \mu_{ij}$$

$$N_j^n \sim n\nu_j$$



The total potential service capacity at activity (i, j) : $N_j^n \mu_{ij}^n \sim n\mu_{ij}\nu_j \equiv n\bar{\mu}_{ij}$.
 The fluid model will involve only λ_i and $\bar{\mu}_{ij}$.

Static allocation problem [Harrison (2000), Harrison & López (1999)]:

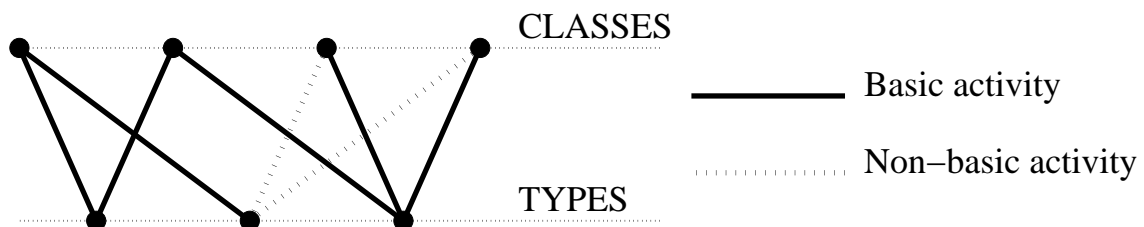
Minimize ρ subject to

$$\begin{aligned} \sum_{j \in \mathcal{J}} \bar{\mu}_{ij} \xi_{ij} &= \lambda_i, & i \in \mathcal{I}, \\ \sum_{i \in \mathcal{I}} \xi_{ij} &\leq \rho, & j \in \mathcal{J}, \\ \xi_{ij} &\geq 0, & i \in \mathcal{I}, j \in \mathcal{J}. \end{aligned}$$

Heavy traffic condition [Harrison (2000)]: There exists a unique optimal solution (ξ^*, ρ^*) to the linear program. Moreover, $\rho^* = 1$, and $\sum_{i \in \mathcal{I}} \xi_{ij}^* = 1$ for all $j \in \mathcal{J}$.

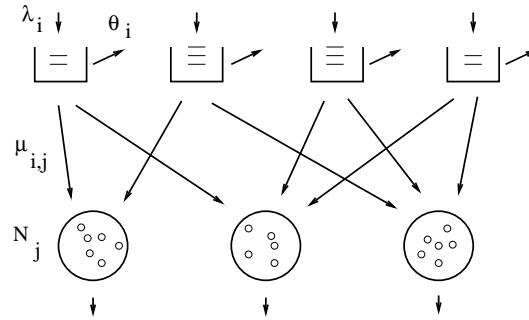
Basic activities: Those activities (i, j) for which $\xi_{ij}^* > 0$.

Complete resource pooling condition [Harrison & López (1999)]: The basic activities form a tree.



Notation: A = set of activities (i, j) ,
 BA = set of basic activities,
 NBA = set of nonbasic activities.

Quantities and relations



Ψ_{ij}^n = number of class- i customers in station j ; Note: $\Psi_{ij}^n = 0$ for $(i, j) \notin A$
 X_i^n = total number of class- i customers in the system
 Y_i^n = number of class- i customers in the queue
 Z_j^n = number of idle servers in station j

$$Y_i^n + \sum_{j \in \mathcal{J}} \Psi_{ij}^n = X_i^n, \quad i \in \mathcal{I}$$

$$Z_j^n + \sum_{i \in \mathcal{I}} \Psi_{ij}^n = N_j^n, \quad j \in \mathcal{J}$$

For the fluid model, denote $\psi_{ij}^* = \nu_j \xi_{ij}^*$, $x_i^* = \sum_j \psi_{ij}^*$, $y_i^* = 0$, $z_j^* = 0$.
Obtain analogously:

$$y_i^* + \sum_{j \in \mathcal{J}} \psi_{ij}^* = x_i^*, \quad i \in \mathcal{I}$$

$$z_j^* + \sum_{i \in \mathcal{I}} \psi_{ij}^* = \nu_j, \quad j \in \mathcal{J}$$

With S_{ij}^n Poisson of rate μ_{ij}^n and R_i^n Poisson of rate θ_i^n ,

$$X_i^n(t) = X_i^{0,n} + A_i^n(t) - \sum_j S_{ij}^n \left(\int_0^t \Psi_{ij}^n(s) ds \right) - R_i^n \left(\int_0^t Y_i^n(s) ds \right)$$

customers = initial + # arrivals - # service completions - # abandonments

Work conservation assumption: $Y_i^n \wedge Z_j^n = 0$ whenever $(i, j) \in A$.

Joint work conservation (JWC) assumption: $\mathbf{1} \cdot Y^n \wedge \mathbf{1} \cdot Z^n = 0$ (this is “nearly always” possible to achieve)

We assume JWC throughout.

Parametrization (second order)

Assume

$$n^{1/2}(n^{-1}\lambda_i^n - \lambda_i) \rightarrow \hat{\lambda}_i, \quad n^{1/2}(\mu_{ij}^n - \mu_{ij}) \rightarrow \hat{\mu}_{ij}, \quad n^{1/2}(n^{-1}N_j^n - \nu_j) \rightarrow 0$$

Perform centering about fluid and rescaling:

$$\begin{aligned} \hat{X}_i^n(t) &= n^{1/2}(n^{-1}X_i^n(t) - x_i^*), \\ \hat{Y}_i^n(t) &= n^{1/2}(n^{-1}Y_i^n(t) - 0), \\ \hat{Z}_j^n(t) &= n^{1/2}(n^{-1}Z_j^n(t) - 0), \\ \hat{\Psi}_{ij}^n(t) &= n^{1/2}(n^{-1}\Psi_{ij}^n - \psi_{ij}^*) \end{aligned}$$

to obtain

$$\hat{X}_i^n(t) = \hat{X}_i^{0,n} + \hat{W}_i^n(t) - \sum_j \mu_{ij}^n \int_0^t \hat{\Psi}_{ij}^n(s) ds - \theta_i^n \int_0^t Y_i(s) ds, \quad i \in \mathcal{I}$$

$$\sum_j \hat{\Psi}_{ij}^n = \hat{X}_i^n - \hat{Y}_i^n, \quad i \in \mathcal{I},$$

$$\sum_i \hat{\Psi}_{ij}^n = -\hat{Z}_j^n, \quad j \in \mathcal{J},$$

$$\mathbf{1} \cdot \hat{Y}^n \wedge \mathbf{1} \cdot \hat{Z}^n = 0$$

\hat{W}_i^n depends on A_i^n , S_{ij}^n and R_i^n , and converges weakly to a Brownian motion.

As a result, obtain the **Diffusion Model**:

$$X_i(t) = x_i + W_i(t) - \sum_j \mu_{ij} \int_0^t \Psi_{ij}(s) ds - \theta_i \int_0^t Y_i(s) ds, \quad i \in \mathcal{I}$$

$$\sum_j \Psi_{ij} = X_i - Y_i, \quad i \in \mathcal{I},$$

$$\sum_i \Psi_{ij} = -Z_j, \quad j \in \mathcal{J},$$

$$\mathbf{1} \cdot Y \wedge \mathbf{1} \cdot Z = 0.$$

Controlled diffusion: X ; **Control:** Ψ ; **Constraints:** last 3 equations.

Additional constraints: $\Psi_{ij} \geq 0$ for $(i, j) \in NBA$ (however, $\Psi_{ij} \in \mathbb{R}$ for $(i, j) \in BA$).

Control theory

A control theoretic approach to queues in heavy traffic analyzes an optimal control problem for a diffusion model so as to construct from it asymptotically optimal controls for the queueing model.

Examples:

Diffusion model 1: $dX = b(X, U)dt + dW$,

Diffusion model 2: $dX = b(X, U)dt + d\eta + dW$,

where η is a process that increases in a certain cone of directions.

Infinite horizon, discounted cost: $E \int_0^\infty e^{-\gamma t} L(X_t, U_t) dt$

Finite horizon cost: $E \int_0^T L(X_t, U_t) dt + g(X_T)$

Ergodic cost: $\limsup_{T \rightarrow \infty} T^{-1} E \int_0^T L(X_t, U_t) dt$

- Characterize optimal cost as the unique solution to a Hamilton-Jacobi-Bellman PDE
- Find optimal controls for the diffusion model
- Find asymptotically optimal controls for the queueing model

Obstacles

1) The domain, \mathbb{R}^I , is unbounded; The drift $b(x, u)$ is unbounded (linear growth); If the function L represents, say, linear combinations of queue lengths, it is unbounded.

Under these conditions, **uniqueness** for the PDE (under polynomial growth conditions) **does not hold** in general.

2) Consider the ergodic cost problem. For PDE analysis one needs either

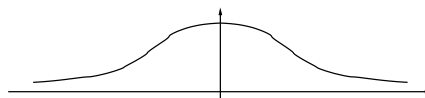
A) Criteria for stability of the diffusion model. However, **stability issues are poorly understood** (for example, it is not clear how to adopt a workload formulation)

Or: B) Borkar's near-monotone condition: When the dynamics deviates considerably from a fixed bounded set, large cost is incurred. **This does not occur here.** For queue length related cost,

$$L(x, u) = (\mathbf{1} \cdot x)^+ \ell(u) \not\leq \varepsilon \|x\|$$

Suppose one can find a control that stabilizes the system, is there an optimal control that is stationary?

Counterexample: Linear dynamics $dX = UXdt + dW$, $U \in [-1, 1]$. Cost function $L = L(X)$:



A similar situation occurs in the current model.

3) Consider infinite horizon, discounted cost. PDE analysis requires either

A) Large time estimates: $E_x^U f(X_t)$ is polynomial in t , for polynomial f ; However, this is **not known** in general.

Or: B) Condition similar to 2)B.

Remark: Abandonments make the queueing model stable, but **they do not make the diffusion model stable.**

4) Interesting applications have nonpreemptive disciplines; the diffusion model is designed for preemptive disciplines. A bridge is possible, but nontrivial.

The treelike case

We assume: \mathcal{G} itself is a tree. As a result there are no nonbasic activities.

In this case the diffusion model has the form: $dX = b(X, U)dt + dW$

PROOF: The system

$$\begin{cases} \sum_j \psi_{ij} = \alpha_i, & i \in \mathcal{I} \\ \sum_i \psi_{ij} = \beta_j, & j \in \mathcal{J} \end{cases}$$

has a unique solution (ψ_{ij}) for every (α, β) with $\mathbf{1} \cdot \alpha = \mathbf{1} \cdot \beta$. Denote this linear map by G : $\psi = G(\alpha, \beta)$. Then $\Psi = G(X - Y, -Z)$.

Since $\mathbf{1} \cdot X = \mathbf{1} \cdot Y - \mathbf{1} \cdot Z$, and $\mathbf{1} \cdot Y \wedge \mathbf{1} \cdot Z = 0$, we have $\mathbf{1} \cdot Y = (\mathbf{1} \cdot X)^+$, $\mathbf{1} \cdot Z = (\mathbf{1} \cdot X)^-$. Hence Y and Z can be represented as

$$Y_i(t) = (\mathbf{1} \cdot X(t))^+ u_i(t), \quad Z_j(t) = (\mathbf{1} \cdot X(t))^- v_j(t).$$

Thus **considering** $U := (u, v)$ **as a control process** taking values in

$$\mathbb{U} := \{(u, v) \in \mathbb{R}^{I+J} : u_i, v_j \geq 0, \mathbf{1} \cdot u = \mathbf{1} \cdot v = 1\}.$$

we have $\Psi = G(X - (\mathbf{1} \cdot X)^+ u, -(\mathbf{1} \cdot X)^- v) =: \hat{G}(X, U)$.

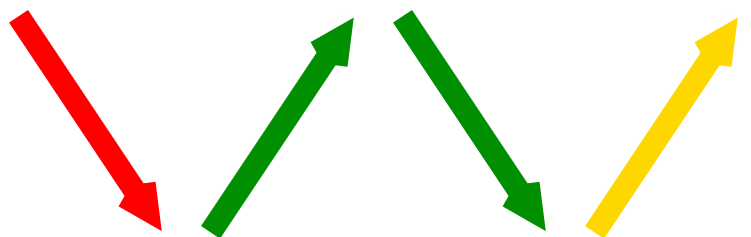
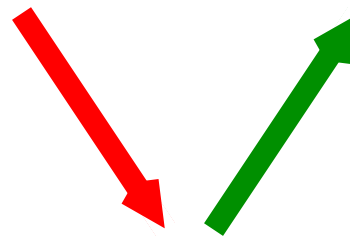
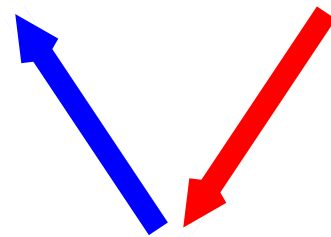
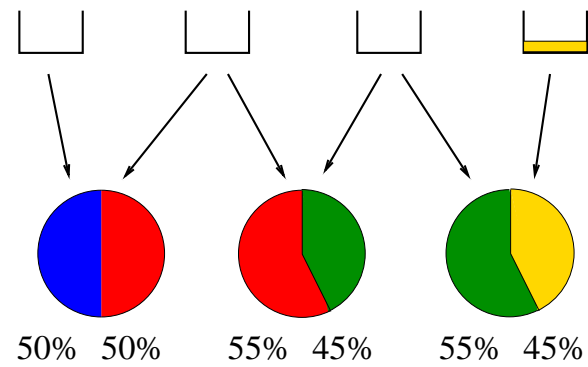
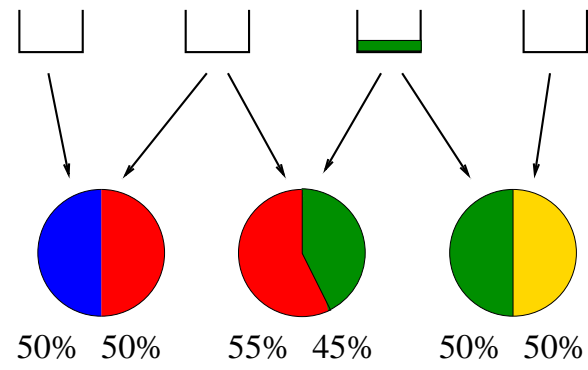
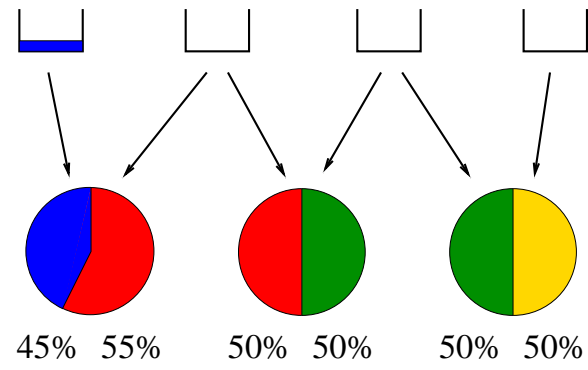
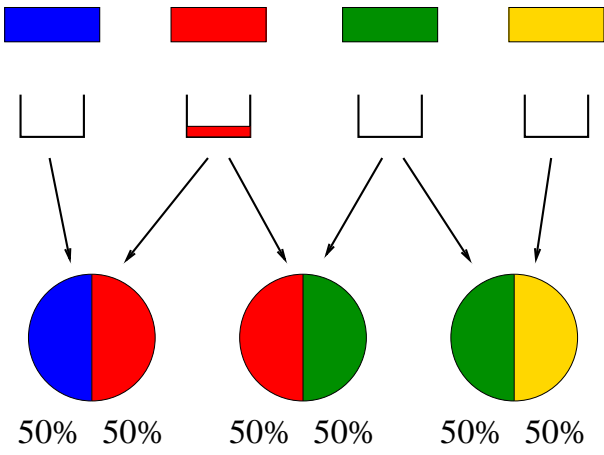
Denoting $b = -\mu \circ \hat{G}$, the equation

$$X_i(t) = x_i + W_i(t) - \sum_j \mu_{ij} \int_0^t \Psi_{ij}(s) ds$$

becomes

$$X(t) = x + W(t) + \int_0^t b(X(s), U(s)) ds.$$

The role of the control Ψ in the treelike case



Results on the HJB equation

The equations for the infinite and resp., finite horizon problems are

$$(\Delta - \gamma)f + H(x, \nabla f) = 0; \quad H(x, p) = \inf_{U \in \mathbb{U}} [b(x, U) \cdot p + L(x, U)]$$

$$\left(\frac{\partial}{\partial t} + \Delta_x\right)f + \tilde{H}(x, \nabla_x f) = 0; \quad \tilde{H}(x, p) = \inf_{U \in \mathbb{U}} [b(x, U) \cdot p + L(x, U)];$$

with growth condition

$$\exists C \quad |f(x)|, |f(t, x)| \leq C(1 + \|x\|^C), \quad x \in \mathbb{R}^I$$

and terminal condition

$$f(T, \cdot) = g$$

Assumptions on L and g : Some regularity, and polynomial growth in x .

Infinite horizon discounted problem:

Theorem [Harrison & Zeevi]: In the single station case, unique solvability of the HJB equation by the value function.

More general models:

Theorem: Unique solvability by the value function, under each of the following conditions:

- 1) μ_{ij} depends only on i ; or μ_{ij} depends only on j ; no abandonments.
- 2) The tree \mathcal{G} is of diameter 3 at most, and $\forall (i, j) \in A \quad \theta_i \leq \mu_{ij}$
- 3) $L(x, U) \sim \|x\|^m$, and $\exists (i, j) \in A \quad \theta_i \leq \mu_{ij}$
- 4) $L(x, U)$ is bounded.

Finite horizon problem:

Theorem: Unique solvability by the value function holds in general.

Uniform estimates on the Lyapunov exponent

To prove uniqueness in C_{pol}^2 use Ito's formula for for $f \in C_{\text{pol}}^2$ to get

$$f(x) = \int_0^t e^{-\gamma s} L(X_s, U_s) ds + e^{-\gamma t} f(X_t) - \int_0^t e^{-\gamma s} \nabla f(X_s) \cdot dW_s,$$

with equality iff U_s is the optimal control. To show: $e^{-\gamma t} f(X_t) \rightarrow 0$.

In a deterministic setting, consider:

$$x_i(t) = w_i(t) - \sum_j \mu_{ij} \int_0^t \psi_{ij}(s) ds, \quad i \in \mathcal{I},$$

$$\sum_j \psi_{ij} = x_i - y_i, \quad i \in \mathcal{I},$$

$$\sum_i \psi_{ij} = -z_j, \quad j \in \mathcal{J},$$

$$\mathbf{1} \cdot y \wedge \mathbf{1} \cdot z = 0.$$

Q: When is it true that the above relations imply the following

$$\|x\|_t^* \leq c(1+t)^m(1+\|w\|_t^*)^m.$$

Remark: Using the relation $x = w + \int b(x, U)$ we have exponential bounds.

FACT: The treelike assumption is necessary (counterexamples exist).

Notation: $\mathfrak{I}f = \int_0^\cdot f$

LEMMA:

$$\|x\| \leq c(\|w\|^* + \mathfrak{I}(\mathbf{1} \cdot y) + \mathfrak{I}(\mathbf{1} \cdot z)).$$

An integral formula for y and z

For $\alpha \in \mathbb{R}$ denote

$$\mathfrak{T}_\alpha f = f + \alpha \mathfrak{I}f.$$

Note: \mathfrak{T}_α is invertible; $\mathfrak{T}_\alpha, \mathfrak{T}_\beta$ commute.

If $A = (\alpha_1, \dots, \alpha_k)$ is a finite real-valued sequence, denote

$$\mathfrak{T}_A = \mathfrak{T}_{\alpha_1} \circ \dots \circ \mathfrak{T}_{\alpha_k}.$$

Then \mathfrak{T}_A does not depend on the order of the elements of A , but it depends on the multiplicity of each element. Let $\mathfrak{T}_\emptyset = \text{identity map}$. Then

$$\sum_j \mathfrak{T}_{\mu_{ij}} \psi_{ij} = w_i - y_i, \quad i \in \mathcal{I},$$

$$\sum_i \psi_{ij} = -z_j, \quad j \in \mathcal{J}.$$

Theorem: y and z solve the following integral equation

$$\sum_{i \in \mathcal{I}} \mathfrak{T}_{A_i}(w_i - y_i) + \sum_{j \in \mathcal{J}} \mathfrak{T}_{B_j} z_j = 0,$$

where A_i and B_j are finite (possibly empty) sequences with values in $\{\mu_{ij}\}$.

- Writing \mathfrak{I}_n for the n -power of the operator \mathfrak{I} , there are **positive** $a_{i,n}, b_{j,n}$ s.t.

$$e \cdot w - e \cdot y + e \cdot z + \sum_{i \in \mathcal{I}} \sum_{n=1}^{m_i} a_{i,n} \mathfrak{I}_n(w_i - y_i) + \sum_{j \in \mathcal{J}} \sum_{n=1}^{m_j} b_{j,n} \mathfrak{I}_n z_j = 0.$$

- Gives relation between w, y and z **alone**.

The dynamics can be put in terms of the controls (u, v) and a **one dimensional** quantity $e \cdot x$ alone.

- Gives bounds on y if bounds on z are available, and vice versa.

- Treat case (I) of the infinite horizon result, where the equation simplifies, implying uniform polynomial estimates.

The non-idling property

Say that the system incurs no idleness on $[0, T]$ if $\mathbf{1} \cdot z(t) = 0$ for $t \in [0, T]$.

The non-idling property: If the system starts with $w_i(0) > 0$ and $w_i, i \in \mathcal{I}$, are strictly increasing on $[0, T]$, then the system incurs no idleness on $[0, T]$.

Heuristic: There is data (corresponding to “large amount of work”) that will keep all servers busy for a while under **any** jointly work conserving control.

This is relevant for us because:

FACT 1: For trees of diameter not exceeding 3, the nonidling property implies the estimate

$$\|x\|_t^* \leq c(1+t)^m \|w\|_t^*$$

Sketch: - Show that modifying w_i to the **increasing functions** w_i^* does not affect the state x by too much;

- By the nonidling property, $z = 0$;

- Now use the integral formula to dominate y .

FACT 2: For trees of diameter not exceeding 3, the nonidling property holds.

(In general: OPEN).

As a result, get estimates on the state for trees of diameter not exceeding 3 (case (II) of the infinite horizon result).

Scheduling control problem

Preemptive scheduling: Service to a customer can be stopped and resumed at a later time, possibly in a different station. Scheduling decisions are made by continuously selecting Ψ .

Nonpreemptive scheduling: Customers complete service with the server they are first assigned. Scheduling decisions are made by selecting a server for each customer, and when to begin service.

Nonanticipating controls in presence of renewal processes:

$$\text{Let } \tau_i(t) = \inf\{u \geq t : A_i(u) - A_i(u-) > 0\}$$

Past information

$$\mathcal{F}_t = \sigma\{A_i(s), S_{ij}(T_{ij}(s)), \Psi_{ij}(s), X_i(s), Y_i(s), Z_j(s) : s \leq t\}$$

Future information

$$\mathcal{H}_t = \sigma\{A_i(\tau_i(t)+u) - A_i(\tau_i(t)), S_{ij}(T_{ij}(t)+u) - S_{ij}(T_{ij}(t)) : u \geq 0\}$$

A control is nonanticipating if

- i) For each t , \mathcal{F}_t and \mathcal{H}_t are independent
- ii) for each t , the process $S_{ij}(T_{ij}(t)+\cdot) - S_{ij}(T_{ij}(t))$ is equal in law to $S_{ij}(\cdot)$.

Scheduling control problem:

Minimize cost over admissible, nonanticipating control processes Ψ

Results on asymptotic optimality

Cost for the queueing scheduling problem:

$$E \int_0^\infty e^{-\gamma t} \tilde{L}(\hat{X}^n(s), \hat{\Psi}^n(s)) ds,$$
$$E \int_0^T \tilde{L}(\hat{X}^n(s), \hat{\Psi}^n(s)) ds + g(\hat{X}^n(T)).$$

Theorem: In each of the above cases where the PDE analysis is possible, the optimal cost of the queueing scheduling problem converges to the diffusion control problem's value. Moreover, one can construct **preemptive** and **nonpreemptive** scheduling controls under which the cost converges to the same limit.

The optimal control U for the diffusion problem is given as

$$\Psi_s = h(X_s),$$

where h is a function determined by the PDE solution.

For the preemptive problem, it is possible to let

$$\hat{\Psi}_s^n = h(\hat{X}_s^n)$$

and argue by weak convergence.

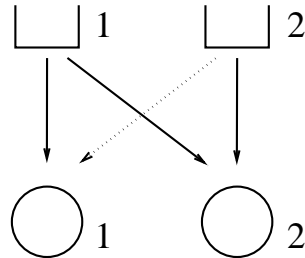
For the nonpreemptive problem there is no direct control over the $\hat{\Psi}$ process. Apply a **tracking mechanism:**

- Compute $h(\hat{X}_s^n)$
- Declare activities with $\hat{\Psi}_s^n > h(\hat{X}_s^n)$ as “over-populated”
- Block over-populated activities. The population drops rapidly; population in the other activities increases rapidly
- As a result

$$\hat{\Psi}_s^n \sim h(\hat{X}_s^n)$$

Beyond the treelike condition: The role of NBA

Consider now a model containing NBA.

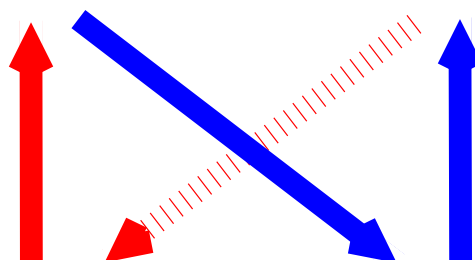
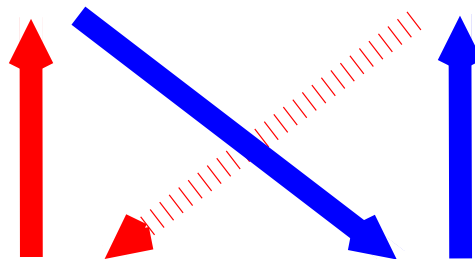
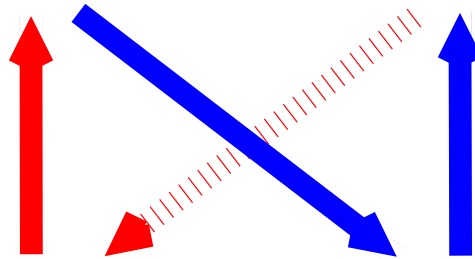
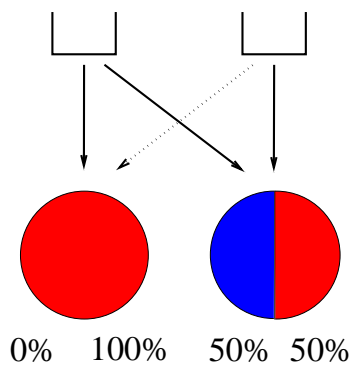
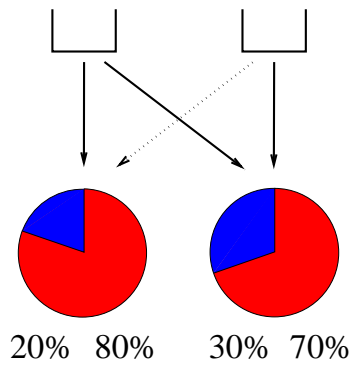
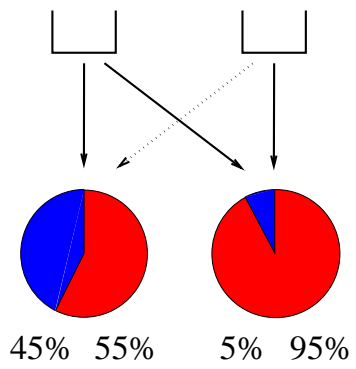
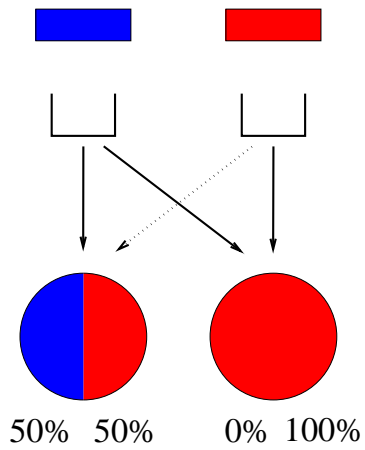


The diffusion model now has the form

$$dX = b(X, U)dt + md\eta + dW,$$

where $m = (\mu_{11} - \mu_{12}, \mu_{22} - \mu_{21})'$, and η is an increasing process. The control is the pair (U, η) .

The role of the control η in a model with cycles

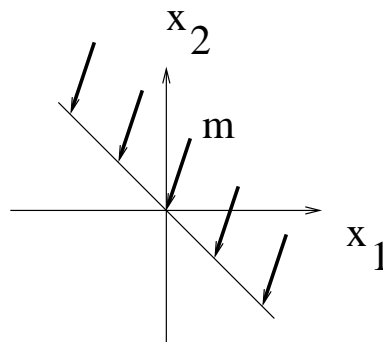


Null-controllability

Theorem: Assume $\mathbf{1} \cdot m < 0$, i.e. $\mu_{11} + \mu_{22} < \mu_{12} + \mu_{21}$.

Then the system is **null-controllable** in the sense that there exists a control under which $Y(t) = 0, t > 0$, w.p.1.

Proof: $\sum Y_i = (\sum X_i)^+$



For the general model, if there are K cycles, there is a direction m_k per cycle:

$$dX = b(X, U)dt + \sum_{k=1}^K m_k d\eta_k + dW$$

Theorem: Assume $\mathbf{1} \cdot m_k < 0$ for some $k = 1, \dots, K$. Then the system is null-controllable.

Asymptotic null-controllability:

Theorem: In the null-controllable case one can construct preemptive and nonpreemptive controls for the queueing model under which \hat{Y}^n converges weakly to zero.

Future work will address PDE analysis (viscosity solutions) and heavy traffic asymptotics for the general case.