Large deviations and calculus of variations for some pure jump interacting particle systems

Ruoyu Wu

Iowa State University

(Based on joint works with Rami Atar, Amarjit Budhiraja, Paul Dupuis, Eric Friedlander, and Zhenhua Wang)

Robust Optimization and Simulation of Complex Stochastic Systems, ICERM

(In honor of Paul Dupuis' 65th birthday)

September 13–15, 2024

Introduction

First meeting with Paul: Summer School 2016

Second meeting with Paul: Fall 2016

- Large deviation principles (LDP) of component sizes of configuration models

Introduction

Discrete-time Markov chain with:

- Infinite dimensional dynamics.
- Vanishing jump rates (near the boundary).
- Discontinuous statistics (at the boundary).

LDP (local) rate functions usually have poor regularity behavior (unbounded / non-Lipschitz)
— mollification might work.

Formulate as a continuous-time problem:
— apply weak convergence and stochastic control approach (Dupuis-Ellis '97, Budhiraja-Dupuis '19)

- Tightness for upper bound.
- Uniqueness of ODEs for lower bound.

Related calculus of variations problems

- Optimal paths are not fully explicit / tractable.

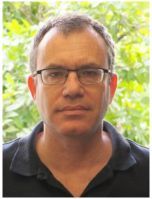
Introduction

Three subtle features of the dynamics:

- Infinite dimensional dynamics.
- Vanishing jump rates.
- Discontinuous statistics.

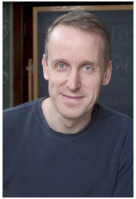
Some / all of these arise in a few LDP problems:

- Configuration models
  (Bhamidi, Budhiraja, Dupuis, W. '22)
- M/M/1 queue with Markovian abandonment
  (Atar, Budhiraja, Dupuis, W. '21)
- Join the shortest queue
  (Budhiraja, Friedlander, W. '21)
- Join the shortest queue(d) / power-of-d / supermarket model
  (Wang, W. '24+)

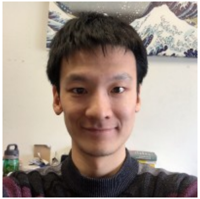

Rami Atar


Shankar Bhamidi


Amarjit Budhiraja


Paul Dupuis


Eric Friedlander


Zhenhua Wang

Overview

1. M/M/1 queue with Markovian abandonment
2. Join the shortest queue
3. Join the shortest queue(d)

Model

M/M/1 queue with abandonment:

- Single server queue
- Jobs/Customers arrival rate $n\lambda$
- First-come-first-serve with service rate $n\mu$
- Each arriving job comes with a "patience" random variable, i.i.d. exponential with mean θ^{-1}
- The job abandons the queue at the time its patience expires — at rate θ
- Inter-arrival times, service times, patience times are mutually independent

Goal 1: LDP of (scaled) queue length process and total abandonment process as $n \to \infty$ and $t \to \infty$.

Goal 2: Asymptotic probability of large abandonment numbers. (assuming overloaded system $\lambda \geq \mu$)

(LDP for M/M/n queue with abandonment can also be obtained, with minor adjustments)

(LDP estimate for G/G/n queue can be obtained)

LLN

$Q^n(t)$ = queue length at time t. $V^n(t)$ = total abandonment by time t. Consider scaled processes

$$X^n(t) = \frac{Q^n(t)}{n}, \quad Y^n(t) = \frac{V^n(t)}{n}, \quad t \in [0, T].$$

Assume $(X^n(0), Y^n(0)) = (x_n, 0) \to (x_0, 0)$ as $n \to \infty$.

Law of Large Numbers (LLN):

As $n \to \infty$, $(X^n, Y^n) \to (x, y)$ in $\mathbb{D}([0, T] : \mathbb{R}^2_+)$ in probability:

$$x(t) = x_0 + (\lambda - \mu)t - \theta \int_0^t x(s)ds, \quad y(t) = \theta \int_0^t x(s)ds, \quad t \in [0, T].$$

- The equilibrium point is $\bar{x} = (\lambda - \mu)/\theta$.
- As $T \to \infty$, $y(T) \sim \theta \bar{x} T = (\lambda - \mu)T$.
- Total abandonment rate is $\lambda - \mu$, independent of θ.

State dynamics

$Q^n(t)$ = queue length at time t. $V^n(t)$ = total abandonment by time t,

$$X^n(t) = \frac{Q^n(t)}{n}, \quad Y^n(t) = \frac{V^n(t)}{n}, \quad t \in [0, T].$$

Given that $X^n(t-) = x$, $Y^n(t-) = y$, possible transitions:

- arrival: $(x, y) \to (x + \frac{1}{n}, y)$ at rate $n\lambda$
- departure: $(x, y) \to (x - \frac{1}{n}, y)$ at rate $n\mu\mathbf{1}_{\{x>0\}}$ — discontinuous statistics
- abandonment: $(x, y) \to (x - \frac{1}{n}, y + \frac{1}{n})$ at rate θnx — vanishing rates

State dynamics

Let N_1, N_2, N_3 be three mutually independent Poisson Random Measures on $[0,T]\times\mathbb{R}_+$, $[0,T]\times\mathbb{R}_+$ and $[0,T]\times\mathbb{R}_+^2$ respectively with intensities $\lambda\, dsdy$, $\mu\, dsdy$ and $\theta\, dsdydz$, respectively.

$$
\begin{aligned}
X^n(t) = x_n &+ \frac{1}{n}\int_{[0,t]\times\mathbb{R}_+} \mathbf{1}_{[0,n]}(y) N_1(ds\, dy) \\
&- \frac{1}{n}\int_{[0,t]\times\mathbb{R}_+} \mathbf{1}_{[0,n]}(y)\mathbf{1}_{\{X^n(s-)\neq 0\}} N_2(ds\, dy) \\
&- \frac{1}{n}\int_{[0,t]\times\mathbb{R}_+\times\mathbb{R}_+} \mathbf{1}_{[0,n]}(y)\mathbf{1}_{[0,X^n(s-)]}(z) N_3(ds\, dy\, dz). \\
Y^n(t) = \frac{1}{n}&\int_{[0,t]\times\mathbb{R}_+\times\mathbb{R}_+} \mathbf{1}_{[0,n]}(y)\mathbf{1}_{[0,X^n(s-)]}(z) N_3(ds\, dy\, dz).
\end{aligned}
$$

State dynamics

One can rewrite the evolution of (X^n, Y^n) using the one-dimensional Skorokhod map Γ as

$$\begin{aligned} X^n(t) = \Gamma \Bigg(x_n &+ \frac{1}{n} \int_{[0,\cdot]\times\mathbb{R}_+} \mathbf{1}_{[0,n]}(y) N_1(ds\,dy) \\ &- \frac{1}{n} \int_{[0,\cdot]\times\mathbb{R}_+} \mathbf{1}_{[0,n]}(y) N_2(ds\,dy) \\ &- \frac{1}{n} \int_{[0,\cdot]\times\mathbb{R}_+\times\mathbb{R}_+} \mathbf{1}_{[0,n]}(y) \mathbf{1}_{[0,X^n(s-)]}(z) N_3(ds\,dz\,dy) \Bigg)(t) \\ Y^n(t) = \frac{1}{n} &\int_{[0,t]\times\mathbb{R}_+\times\mathbb{R}_+} \mathbf{1}_{[0,n]}(y) \mathbf{1}_{[0,X^n(s-)]}(z) N_3(ds\,dz\,dy). \end{aligned}$$

where $\Gamma : \mathbb{D}([0,T] : \mathbb{R}) \to \mathbb{D}([0,T] : \mathbb{R}_+)$ is

$$\Gamma(\psi)(t) \doteq \psi(t) - \inf_{0\le s\le t} [\psi(s) \wedge 0], \quad t \in [0,T], \quad \psi \in \mathbb{D}([0,T] : \mathbb{R}).$$

LDP and rate function

Theorem (Atar-Budhiraja-Dupuis-W. '21)

$\{(X^n, Y^n)\}$ satisfies a LDP on $\mathbb{D}([0,T]:\mathbb{R}_+^2)$ with rate function I_T.

Form of the rate function I_T:

For $(\xi,\zeta) \in \mathbb{C}([0,T]:\mathbb{R}_+^2)$, define

$$I_T(\xi,\zeta) = \inf_{\varphi \in \mathcal{U}(\xi,\zeta)} \left\{ \lambda \int_0^T \ell(\varphi_1(s))ds + \mu \int_0^T \ell(\varphi_2(s))ds + \theta \int_0^T \xi(s)\ell(\varphi_3(s))ds \right\}.$$

Here $\ell(x) := x\log x - x + 1 \geq 0$ and $\mathcal{U}(\xi,\zeta)$ is the collection of all non-negative functions $\varphi = (\varphi_1, \varphi_2, \varphi_3)$ such that

$$\xi(t) = \Gamma\left(x_0 + \lambda \int_0^{\cdot} \varphi_1(s)ds - \mu \int_0^{\cdot} \varphi_2(s)ds - \theta \int_0^{\cdot} \varphi_3(s)\xi(s)ds\right)(t), \quad \zeta(t) = \theta \int_0^t \varphi_3(s)\xi(s)ds.$$

Set $I_T(\xi,\zeta) = \infty$ if $\mathcal{U}(\xi,\zeta) = \emptyset$ or $(\xi,\zeta) \notin \mathbb{C}([0,T]:\mathbb{R}_+^2)$.

LDP and rate function

Key step in the proof of LDP lower bound:

Show that given a near optimal path (ξ^*, ζ^*) and associated near optimal control $\varphi^* = (\varphi_1^*, \varphi_2^*, \varphi_3^*)$ with finite cost

$$\lambda \int_0^T \ell(\varphi_1^*(s))\, ds + \mu \int_0^T \ell(\varphi_2^*(s))\, ds + \theta \int_0^T \xi^*(s)\ell(\varphi_3^*(s))\, ds < \infty,$$

the ODE

$$\xi(t) = \Gamma\left(x_0 + \lambda \int_0^\cdot \varphi_1^*(s)ds - \mu \int_0^\cdot \varphi_2^*(s)ds - \theta \int_0^\cdot \varphi_3^*(s)\xi(s)ds\right)(t), \quad \zeta(t) = \theta \int_0^t \varphi_3^*(s)\xi(s)ds$$

has a unique solution, which must be (ξ^*, ζ^*).

- Discontinuous statistics is taken care of by the Skorokhod map Γ.
- Vanishing rates cannot be treated via Gronwall + Lipschitz property, as one may not have L^2 (or even L^1) bound on φ_3^*. — It is treated via monotonicity arguments (around the boundary $\xi(s) = 0$).

Rare event probability

$Q^n(t)$ = queue length at time t. $V^n(t)$ = total abandonment by time t,

$$X^n(t) = \frac{Q^n(t)}{n}, \quad Y^n(t) = \frac{V^n(t)}{n}, \quad t \in [0, T].$$

Recall LLN: $Y^n(T) \sim (\lambda - \mu)T$ for large n and T.

Question: Given $\gamma > \lambda - \mu$, what is $\mathbb{P}(Y^n(T) \geq \gamma T)$ for large n, T?

$$\limsup_{T\to\infty} \limsup_{n\to\infty} \frac{1}{T}\frac{1}{n} \log \mathbb{P}(Y^n(T) \geq \gamma T) = ?$$

$$\liminf_{T\to\infty} \liminf_{n\to\infty} \frac{1}{T}\frac{1}{n} \log \mathbb{P}(Y^n(T) \geq \gamma T) = ?$$

Short answer: $? = -C(\gamma)$ for some simple explicit function $C(\cdot)$ that depends on λ, μ, but not θ.

Rare event probability

Since $\{(X^n, Y^n)\}$ satisfies a LDP on $\mathcal{D}([0,T]:\mathbb{R}^2_+)$ with rate function I_T, contraction principle gives

$$\limsup_{T\to\infty}\limsup_{n\to\infty}\frac{1}{T}\frac{1}{n}\log\mathbb{P}(Y^n(T)\geq\gamma T)\leq\limsup_{T\to\infty}-\frac{1}{T}\inf\{I_T(\xi,\zeta):\zeta(T)\geq\gamma T\},$$

$$\liminf_{T\to\infty}\liminf_{n\to\infty}\frac{1}{T}\frac{1}{n}\log\mathbb{P}(Y^n(T)\geq\gamma T)\geq\liminf_{T\to\infty}-\frac{1}{T}\inf\{I_T(\xi,\zeta):\zeta(T)>\gamma T\}.$$

Difficult to get simple and tractable forms for RHS infimum.

But nice asymptotics can be obtained as $T\to\infty$.

Calculus of variations

First consider long-time analysis of

$$\inf\{I_T(\xi,\zeta) : \zeta(T) = \gamma T\}, \quad \gamma \geq 0.$$

Write

$$I_T(\xi,\zeta) = \int_0^T L(\xi(s), \xi'(s), \zeta(s), \zeta'(s))\, ds$$

in terms of some non-negative convex local rate function L on $\mathbb{R}_+ \times \mathbb{R} \times \mathbb{R}_+^2$.

Solve the Euler-Lagrange equations ($L_i := \partial_i L$)

$$L_1 = \frac{d}{dt} L_2, \quad L_3 = \frac{d}{dt} L_4$$

with boundary conditions

$$\xi(0) = x_0, \quad \zeta(0) = 0, \quad \zeta(T) = \gamma T$$

and transversality condition

$$L_2|_{t=T} = 0, \quad \text{(because no terminal constraint on } \xi(T))$$

to get a candidate minimizer $(\bar{\xi}, \bar{\zeta})$.

Calculus of variations

Remains to prove

- $(\bar{\xi}, \bar{\zeta})$ is the minimizer of
$$\inf\{I_T(\xi,\zeta) : \zeta(T) = \gamma T\}, \quad \gamma \geq 0,$$
- and $\lim_{T\to\infty} \frac{1}{T} I_T(\bar{\xi}, \bar{\zeta}) = C(\gamma)$.

Difficulties:

- $(\bar{\xi}, \bar{\zeta})$ is not well defined unless $x_0 > 0$ and T is sufficiently large.
  — Otherwise $\bar{\xi}(t) < 0$ for some $t \in [0, T]$.
- $(\xi, \xi', \zeta, \zeta')$ takes values in unbounded set, and the local rate function L is not bounded.
  — Involves $\zeta' \log \frac{\zeta'}{\xi}$ etc.
- $(\bar{\xi}, \bar{\zeta})$ is not given explicitly.

Candidate minimizer

There exists a unique $A \in (-\infty, e^{-\theta T})$ such that

$$\frac{1}{1 - Ae^{\theta T}} = \frac{1}{2\lambda\left[\theta T - 1 + e^{-\theta T}\right]}\Bigg\{\theta\left[\gamma T - x_0 + x_0\frac{e^{-\theta T} - A}{1 - A}\right] \\ + \Bigg(\theta^2\left[\gamma T - x_0 + x_0\frac{e^{-\theta T} - A}{1 - A}\right]^2 \\ -4\lambda\left[\theta T - 1 + e^{-\theta T}\right]\mu\left[\log\frac{e^{-\theta T} - A}{1 - A} - \frac{e^{-\theta T} - A}{1 - A} + 1\right]\Bigg)^{1/2}\Bigg\}.$$

Let $B = 1/(1 - Ae^{\theta T})$. Then

$$\bar{\zeta}(t) = \frac{\lambda B}{\theta}\left[\theta t - 1 + e^{-\theta t}\right] + \frac{\mu}{\theta B}\left[\log\frac{e^{-\theta t} - A}{1 - A} - \frac{e^{-\theta t} - A}{1 - A} + 1\right] + \frac{1 - e^{-\theta t}}{1 - A}x_0.$$

~~Painful~~ Careful analysis is needed.

Long-time asymptotics

Theorem (Atar-Budhiraja-Dupuis-W. '21)

Let $C(\gamma) := \lambda\left(1 - z_\gamma^{-1}\right) + \mu\left(1 - z_\gamma\right) - \gamma \log z_\gamma$, $z_\gamma := \dfrac{\sqrt{\gamma^2 + 4\lambda\mu} - \gamma}{2\mu}$. For all $x_0 \geq 0$,

$$
\begin{aligned}
\lim_{T\to\infty} \frac{1}{T} \inf\{I_T(\xi,\zeta) : \zeta(T) = \gamma T\} &= C(\gamma), \quad \gamma \geq 0,\\
\lim_{T\to\infty} \frac{1}{T} \inf\{I_T(\xi,\zeta) : \zeta(T) \geq \gamma T\} &= C(\gamma), \quad \gamma \geq \lambda - \mu,\\
\lim_{T\to\infty} \frac{1}{T} \inf\{I_T(\xi,\zeta) : \zeta(T) \leq \gamma T\} &= C(\gamma), \quad 0 \leq \gamma \leq \lambda - \mu.
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\limsup_{T\to\infty} \limsup_{n\to\infty} \frac{1}{T}\frac{1}{n} \log \mathbb{P}(Y^n(T) \geq \gamma T) &\leq \limsup_{T\to\infty} -\frac{1}{T} \inf\{I_T(\xi,\zeta) : \zeta(T) \geq \gamma T\} = -C(\gamma),\\
\liminf_{T\to\infty} \liminf_{n\to\infty} \frac{1}{T}\frac{1}{n} \log \mathbb{P}(Y^n(T) \geq \gamma T) &\geq \liminf_{T\to\infty} -\frac{1}{T} \inf\{I_T(\xi,\zeta) : \zeta(T) > \gamma T\} = -C(\gamma).
\end{aligned}
$$

1. M/M/1 queue with Markovian abandonment
2. Join the shortest queue
3. Join the shortest queue(d)

Large-scale load-balancing queueing systems

Most basic setup:

- 1 dispatcher, n servers
- Jobs arrive at the dispatcher at rate $n\lambda_n$, $\lim_{n\to\infty} \lambda_n = \lambda > 0$
- Each job is routed by the dispatcher to some queue
- Each server maintains a First-In-First-Out queue
- Jobs processed at rate 1 at each server
- Service times and inter-arrival times are independent exponential random variables

Check out lines at supermarkets, cloud computing ...

Large-scale load-balancing queueing systems

Aims:

- Good delay performance, such as low average waiting time
- Economical in implementation, such as low communication cost among dispatcher/servers

Popular load-balancing algorithms:

- Route the incoming job into the shortest queue — Join the Shortest Queue (JSQ)
- Upon job's arrival, choose d queues uniformly at random and route the incoming job into the shortest queue among these d queues — Power-of-d (JSQ(d))
- Route the incoming job into the idle queue, if any, as if implementing JSQ. Otherwise, route the incoming job in a different way, such as JSQ(d) — Join the Idle Queue (JIQ)

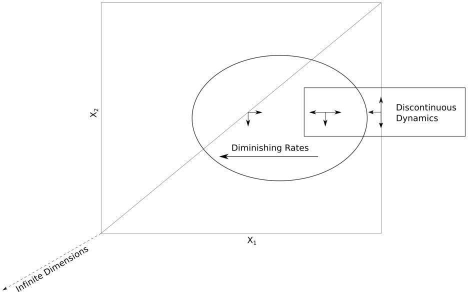
JSQ state process

$\boldsymbol{X}^n(t) := (X_i^n(t))_{i\geq 0}$ denotes the occupancy measure process.

$X_i^n(t)$ = proportion of queues of length at least i
$= (\#$ servers with queue length at least i at time $t)/n$.

$X_0^n(t) \equiv 1 \geq X_1^n(t) \geq X_2^n(t) \geq \cdots \geq 0.$

Assume $X_i^n(0) = x_i$, arrival rate $n\lambda_n$ with $\lambda_n \to \lambda \in (0,\infty)$.



$$X_1^n(t) = x_1 - \frac{1}{n}\int_{[0,t]\times[0,1]} \mathbf{1}_{[0,X_1^n(s-)-X_2^n(s-)]}(y)D_1^n(ds\,dy) + \frac{1}{n}\int_{[0,t]\times[0,1]} D_0^{n\lambda_n}(ds\,dy) - \eta_1^n(t),$$

$$X_i^n(t) = x_i - \frac{1}{n}\int_{[0,t]\times[0,1]} \mathbf{1}_{[0,X_i^n(s-)-X_{i+1}^n(s-)]}(y)D_i^n(ds\,dy) + \eta_{i-1}^n(t) - \eta_i^n(t), \quad i\geq 2,$$

$$\eta_i^n(t) := \frac{1}{n}\int_{[0,t]\times[0,1]} \mathbf{1}_{\{X_i^n(s-)=1\}} D_0^{n\lambda_n}(ds\,dy), \quad i \geq 1.$$

$D_i^\theta(ds\,dy)$: i.i.d. Poisson random measure on $[0,T]\times[0,1]$ with intensity measure $\theta\,dsdy$.

State process

View $\boldsymbol{X}^n$ as the solution of infinite-dimensional Skorokhod problem for $\boldsymbol{Y}^n$ with respect to the region $[0,1]^\infty$ and the reflection matrix R_∞:

$$\boldsymbol{X}^n(t) = \boldsymbol{Y}^n(t) + R_\infty \boldsymbol{\eta}^n(t).$$

Here R_∞ is given by $R_\infty(i,i) = -1,\ R_\infty(i,i-1) = 1.$

$$Y_1^n(t) = x_1 - \frac{1}{n}\int_{[0,t]\times[0,1]} \mathbf{1}_{[0,X_1^n(s-)-X_2^n(s-)]}(y) D_1^n(ds\,dy) + \frac{1}{n}\int_{[0,t]\times[0,1]} D_0^{n\lambda_n}(ds\,dy),$$

$$Y_i^n(t) = x_i - \frac{1}{n}\int_{[0,t]\times[0,1]} \mathbf{1}_{[0,X_i^n(s-)-X_{i+1}^n(s-)]}(y) D_i^n(ds\,dy), \quad i \geq 2,$$

$$\eta_i^n(t) := \frac{1}{n}\int_{[0,t]\times[0,1]} \mathbf{1}_{\{X_i^n(s-)=1\}} D_0^{n\lambda_n}(ds\,dy), \quad i \geq 1,$$

with

$$X_1^n(t) = Y_1^n(t) - \eta_1^n(t), \qquad X_i^n(t) = Y_i^n(t) + \eta_{i-1}^n(t) - \eta_i^n(t), \quad i \geq 2.$$

Skorokhod problem

For each $\omega \in \Omega$, there are finite arrivals.

So it suffices to consider some $M = M(\omega)$ and the finite-dimensional Skorokhod problem on $[0,1]^M$ with matrix $R_M = -I_{M\times M} + P_M$:

$$\begin{aligned} \boldsymbol{X}^n(t) &= \boldsymbol{Y}^n(t) + R_M \boldsymbol{\eta}^n(t) \quad \text{(first } M \text{ coordinates)}, \\ X_i^n(t) &= Y_i^n(t), \quad i > M. \end{aligned}$$

The spectral radius of P_M is less than 1 $\Rightarrow$

- The Skorokhod problem is well-defined;
- unique solution $\boldsymbol{X}^n(t) = \Gamma_M(\boldsymbol{Y}^n)(t)$;
- $\Gamma_M \colon \mathcal{D}^\infty \to \mathcal{D}^\infty$ is Lipschitz on the trajectory, $\mathcal{D} = \mathbb{D}([0,T] : \mathbb{R})$.

Rate function and LDP

For $(\zeta, \psi) \in \mathcal{C}^\infty \times \mathcal{C}^\infty$, where $\mathcal{C} := \mathbb{C}([0,T] : \mathbb{R})$, with ζ solving the Skorokhod problem for ψ with respect to the region $[0,1]^\infty$ and matrix R_∞:

$$\zeta_i(t) = \psi_i(t) + \eta_{i-1}(t) - \eta_i(t),$$

$$\eta_0(t) \equiv 0, \eta_i(0) = 0, \quad \eta_i(t) \text{ is non-decreasing}, \quad \int_0^t \mathbf{1}_{\{\zeta_i(s)<1\}}\, \eta_i(ds) = 0, \quad i \geq 1,$$

let

$$I_T(\zeta,\psi) := \inf_{\varphi} \left\{ \int_{[0,T]\times[0,1]} \lambda \ell(\varphi_0(s,y))\, ds\, dy + \sum_{i=1}^{\infty} \int_{[0,T]\times[0,1]} \ell(\varphi_i(s,y))\, ds\, dy \right\},$$

where $\ell(x) := x \log x - x + 1$, and the infimum is taken over all φ such that

$$\psi_1(t) = x_1 - \int_{[0,t]\times[0,1]} \mathbf{1}_{[0,\zeta_1(s)-\zeta_2(s)]}(y)\varphi_1(s,y)\, ds\, dy + \int_{[0,t]\times[0,1]} \varphi_0(s,y)\, ds\, dy,$$

$$\psi_i(t) = x_i - \int_{[0,t]\times[0,1]} \mathbf{1}_{[0,\zeta_i(s)-\zeta_{i+1}(s)]}(y)\varphi_i(s,y)\, ds\, dy, \quad i \geq 2.$$

Let $I_T(\zeta,\psi) := \infty$ otherwise.

Rate function and LDP

Theorem (Budhiraja-Friedlander-W. '21)

The sequence $(\boldsymbol{X}^n, \boldsymbol{Y}^n)$ satisfies a LDP on $\mathcal{D}^\infty \times \mathcal{D}^\infty$ with rate function I_T.

Key step in the proof of LDP lower bound: Find a near optimal path $(\zeta, \psi) \in \mathcal{D} \times \mathcal{D}$ and a near optimal control φ, such that,

given φ, the pair $(\zeta, \psi) \in \mathcal{D} \times \mathcal{D}$ is the unique solution to the above ODEs. $(16 + 1$ pages)

Both discontinuous statistics and vanishing rates are subtle here.

"Suitably smoothing out small excursions" + "introducing ε-gaps in φ"

Calculus of variations

Suppose all queues are of length 1 at time 0 (i.e. $X_1^n(0) = x_1 = 1$ and $X_j^n(0) = x_j = 0$ for $j \geq 2$).

Consider the critical regime arrival rate $\lambda_n \to \lambda = 1 =$ service rate.

Consider the rare event: Fixing $j \geq 3$, let

$$A_j^{n,T} := \{\text{There is a queue with length} \geq j \text{ at some time } t \in [0,T]\} = \{(\boldsymbol{X}^n, \boldsymbol{Y}^n) \in F_j^{n,T}\}.$$

Relate $A_j^{n,T}$ to open and closed sets:

$$\{(\boldsymbol{X}^n, \boldsymbol{Y}^n) \in G_j\} = A_j^{n,T} \subset \{(\boldsymbol{X}^n, \boldsymbol{Y}^n) \in F_j\},$$

$$G_j := \{(\zeta, \psi) : \sup_{t \in [0,T]} \zeta_j(t) > 0\}, \quad F_j := \{(\zeta, \psi) : \sup_{t \in [0,T]} \zeta_{j-1}(t) = 1\}.$$

Then

$$-I_T(G_j) \leq \liminf_{n \to \infty} \frac{1}{n} \log \mathbb{P}(A_j^{n,T}) \leq \limsup_{n \to \infty} \frac{1}{n} \log \mathbb{P}(A_j^{n,T}) \leq -I_T(F_j).$$

Intuition: start from all queues of length 1 at time 0; end with all queues of length $j-1$; convexity of local rate function $\ell(\cdot)$ suggests linearly having more customers in the system, at rate $(j-2)/T$.

Calculus of variations

Theorem (Budhiraja-Friedlander-W. '21)

For every $j \geq 3$,

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(A_j^{n,T}) &= -I_T(G_j) = -I_T(F_j) \\ &= -T\ell\left(\frac{\frac{j-2}{T} + \sqrt{4 + (\frac{j-2}{T})^2}}{2}\right) - T\ell\left(\frac{-\frac{j-2}{T} + \sqrt{4 + (\frac{j-2}{T})^2}}{2}\right),\end{aligned}$$

$$\lim_{T\to\infty} \lim_{n\to\infty} \frac{T}{n} \log \mathbb{P}(A_j^{n,T}) = -\frac{(j-2)^2}{4}.$$

As a special case, if $j - 2 = T$ (e.g., $j = 3,\ T = 1$), then the probability depends on the golden ratio

$$\mathbb{P}(A_j^{n,T}) \approx \exp\left[-nT\left(\ell\left(\frac{1+\sqrt{5}}{2}\right) + \ell\left(\frac{-1+\sqrt{5}}{2}\right)\right)\right], \text{ for large } n.$$

1. M/M/1 queue with Markovian abandonment
2. Join the shortest queue
3. Join the shortest queue(d)

Model

Recall the difference of Joint-the-shortest-queue(d) (JSQ(d)) from JSQ:

Upon job's arrival at the dispatcher, choose d queues uniformly at random and route the incoming job into the shortest queue among these d queues.

$X_i^n(t) =$ proportion of queues of length at least i at time t.

$X_0^n(t) \equiv 1 \geq X_1^n(t) \geq X_2^n(t) \cdots \geq 0$.

$$
\begin{aligned}
X_i^n(t) &= X_i^n(0) + \frac{1}{n} \int_{[0,t]\times[0,1]} \mathbf{1}_{[0, R_i^n(X^n(s-))]}(y)\, N_i^{n\lambda}(ds\, dy) \\
&\quad - \frac{1}{n} \int_{[0,t]\times[0,1]} \mathbf{1}_{[0, X_i^n(s-) - X_{i+1}^n(s-)]}(y)\, \bar{N}_i^n(ds\, dy), \\
R_i^n(z) &= \left[\binom{nz_{i-1}}{d} - \binom{nz_i}{d} \right] \Big/ \binom{n}{d}, \quad i \geq 1.
\end{aligned}
$$

$N_i^{n\lambda}$ and $\bar{N}_i^n$ are independent Poisson random measures on $[0, T] \times [0, 1]$ with intensity $n\lambda\, ds\, dy$ and $n\, ds\, dy$, respectively.

Conjectured rate function

For $\boldsymbol{\psi} \in \mathcal{C}^\infty$, let

$$I(\boldsymbol{\psi}) := \inf_{\boldsymbol{\varphi}} \sum_{i=1}^{\infty} \int_{[0,T]\times[0,1]} \left(\lambda \ell(\varphi_i(s,y)) + \ell(\bar{\varphi}_i(s,y))\right) ds\, dy,$$

where $\ell(x) := x \log x - x + 1$, and the infimum is taken over all $\boldsymbol{\varphi} = (\varphi_i, \bar{\varphi}_i)_{i=1}^{\infty}$ such that

$$\begin{aligned} \psi_i(t) &= x_i + \lambda \int_{[0,t]\times[0,1]} \mathbf{1}_{[0, R_i(\boldsymbol{\psi}(s))]}(y)\varphi_i(s,y)\, ds\, dy \\ &\quad - \int_{[0,t]\times[0,1]} \mathbf{1}_{[0,\psi_i(s)-\psi_{i+1}(s)]}(y)\bar{\varphi}_i(s,y)\, ds\, dy, \\ R_i(z) &= z_{i-1}^d - z_i^d, \quad i \geq 1. \end{aligned}$$

Let $I(\psi) := \infty$ otherwise.

Conjecture: $\boldsymbol{X}^n$ satisfies a LDP on $\mathcal{D}^\infty$ with rate function I.

Main challenge: Lower bound. Nonlinear vanishing rates.

Moderate deviation principle

The LLN limit of $\boldsymbol{X}^n$ is given by the deterministic limit $\boldsymbol{q}$ as the unique solution to the set of ODEs

$$\frac{dq_i(t)}{dt} = \lambda[(q_{i-1}(t))^d - (q_i(t))^d] - (q_i(t) - q_{i+1}(t)), \quad i = 1, 2, \ldots$$

One can analyze the moderate deviation principle (MDP) of $\boldsymbol{X}^n$ from $\boldsymbol{q}$, by analyzing the LDP of

$$\boldsymbol{Y}^n := a(n)\sqrt{n}(\boldsymbol{X}^n - \boldsymbol{q}).$$

Theorem (Wang-W. '24+)

The sequence $\boldsymbol{Y}^n$ satisfies a MDP on $\mathbb{D}([0, T] : \ell^2)$ with speed $a^2(n)$ and rate function $\mathcal{I}$.

Calculus of variations problems are quite challenging: non-explicit LLN $\boldsymbol{q}$.

Summary

- Sample path LDP are established for some pure jump stochastic processes arising from queueing systems: M/M/1+M (M/M/n+M) and Join-the-Shortest-Queue.
- Three challenging features: infinite dimensional dynamics, vanishing jump rates, discontinuous statistics.
- Rare events of interest can be analyzed by solving the related calculus of variations problems written in terms of the LDP rate functions.
- Ongoing and future works on LDP and MDP of Join-the-Shortest-Queue-d (power-of-d) queueing system.

Thank you!