Asymptotic analysis of the power of choice phenomenon for queuing models

Miheer Dewaskar

Statistics and Operations Research, UNC Chapel Hill.

Probability Seminar, Jan 30th 2020.

Outline

1. Balls and bins
   - Power of choice ($d = 1$ vs. $d = 2$)
   - Dependence on $d \geq 1$
   - How to choose d?
2. Supermarket model
   - Introduction
   - Analysis of join the shortest queue
   - Fluid limit for $\texttt{JSQ}(d_N)$ as $d_N \to \infty$
   - Diffusion limit theorem
3. Summary

Outline

1. Balls and bins
   - Power of choice ($d = 1$ vs. $d = 2$)
   - Dependence on $d \geq 1$
   - How to choose d?

2. Supermarket model
   - Introduction
   - Analysis of join the shortest queue
   - Fluid limit for $\mathtt{JSQ}(d_N)$ as $d_N \to \infty$
   - Diffusion limit theorem

3. Summary

The balls and bins problem

Simplest model to describe the power-of-choice.

Aim

Sequentially place n balls into n bin to minimize conflicts when a centralized dispatcher is absent and $n \in \mathbb{N}$ is large.

Strategy `Smallest(d)`:

Each incoming ball

- samples d bins uniformly at random with replacement,
- selects the least loaded among these d bin.

Compare: `Smallest(1)`, `Smallest(2)` and `Smallest(`∞`)`.

Outline

1. Balls and bins
   - Power of choice ($d = 1$ vs. $d = 2$)
   - Dependence on $d \geq 1$
   - How to choose d?

2. Supermarket model
   - Introduction
   - Analysis of join the shortest queue
   - Fluid limit for $\mathtt{JSQ}(d_N)$ as $d_N \to \infty$
   - Diffusion limit theorem

3. Summary

The power of choice

Choice ($d = 2$) is much better than no choice ($d = 1$).

Maximum load is monotonically decreasing in d (coupling argument).

(Mitzenmacher, 2001) As $n \to \infty$, w.h.p:

	`Smallest(1)`	`Smallest(2)`	$\texttt{Smallest}(\infty)$
Max. load	$O(\log n)$	$O(\log \log n)$	1

The power of choice

Drastic improvement of $d = 2$ over $d = 1$.

Applications (Mitzenmacher, Richa, Sitaraman, 2001)

Hashing, distributed computing, circuit routing and more.

The power of choice

Choice ($d = 2$) is much better than no choice ($d = 1$).

Maximum load is monotonically decreasing in d (coupling argument).

(Mitzenmacher, 2001) As $n \to \infty$, w.h.p:

	`Smallest(1)`	`Smallest(2)`	`Smallest`(∞)
Max. load	$O(\log n)$	$O(\log \log n)$	1

The power of choice

Drastic improvement of $d = 2$ over $d = 1$.

Applications (Mitzenmacher, Richa, Sitaraman, 2001)

Hashing, distributed computing, circuit routing and more.

Outline

1. Balls and bins
   - Power of choice ($d = 1$ vs. $d = 2$)
   - Dependence on $d \geq 1$
   - How to choose d?
2. Supermarket model
   - Introduction
   - Analysis of join the shortest queue
   - Fluid limit for $\mathtt{JSQ}(d_N)$ as $d_N \to \infty$
   - Diffusion limit theorem
3. Summary

Dependence of maximum load on d

Max. load is $\frac{\log\log n}{\log d} + O(1)$

Theorem : Assume $1 < d_n < \text{Poly}(\log n)$ and $n \to \infty$

The maximum load for the n Balls-and-Bins problem using strategy $\texttt{Smallest}(\mathtt{d_n})$ is between

$$\left[\frac{\log\log n}{\log d_n} - 4, \frac{\log\log n}{\log d_n} + 4\right] \qquad \text{w.h.p}$$

Proof formulation (using empirical distribution of bin sizes)

Scale time $t = \{0, \frac{1}{n}, \ldots \frac{n}{n}\} \subseteq [0,1]$ and let

$$G_n(i,t) = \frac{\#\text{ of bins with } \geq i \text{ balls at time } t}{n} \quad \text{and} \quad g_n(i,t) = \mathbb{E}G_n(i,t).$$

Then

- Fixed t : $\{G_n(i,t)\}_{i\geq 1}$ is the distribution of bin sizes at time t.
- Max. bin load is $M^* = \min\{\, i \mid G_n(i+1,1) = 0 \,\}$.

Dependence of maximum load on d

Max. load is $\frac{\log\log n}{\log d} + O(1)$

Theorem : Assume $1 < d_n < \mathrm{Poly}(\log n)$ and $n \to \infty$

The maximum load for the n Balls-and-Bins problem using strategy $\mathtt{Smallest(d_n)}$ is between

$$\left[\frac{\log\log n}{\log d_n} - 4, \frac{\log\log n}{\log d_n} + 4\right] \qquad \text{w.h.p}$$

Proof formulation (using empirical distribution of bin sizes)

Scale time $t = \{0, \frac{1}{n}, \dots \frac{n}{n}\} \subseteq [0,1]$ and let

$$G_n(i,t) = \frac{\#\text{ of bins with } \geq i \text{ balls at time } t}{n} \quad \text{and} \quad g_n(i,t) = \mathbb{E}G_n(i,t).$$

Then

- Fixed t : $\{G_n(i,t)\}_{i\geq 1}$ is the distribution of bin sizes at time t.
- Max. bin load is $M^* = \min\{\, i \mid G_n(i+1,1) = 0\,\}$.

Proof (concentration)

Concentration (Luczak and McDiarmid)

$$\boldsymbol{P}\left(\sup_t \sup_i |G_n(i,t) - g_n(i,t)| > \frac{\log n}{\sqrt{n}}\right) \le 2\exp\left(-\frac{1}{2}\log^2 n\right)$$

No dependence on d.

Concentration for maximum (Luczak and McDiarmid)

W.h.p. the maximum bin load M^* is concentrated on the two values $\{i_n^*, i_n^*+1\}$ where

$$i_n^* = \min\left\{ i \mid g_n(i,n) \le \frac{\ln n}{\sqrt{n}} \right\}.$$

Final step (to show)

$$\frac{\log\log n}{\log d_n} - 3 \le i_n^* \le \frac{\log\log n}{\log d_n} + 3 \qquad \text{eventually as } n \to \infty.$$

Proof continued (properties of the process)

Recall

We scaled time $t = \{0, \frac{1}{n}, \dots \frac{n}{n}\} \subseteq [0,1]$ and defined

$$G_n(i,t) = \frac{\#\text{ of bins with } \geq i \text{ balls at time } t}{n} \quad \text{and} \quad g_n(i,t) = \mathbb{E}G_n(i,t).$$

Let $\boldsymbol{G_n}(t) = (G_n(i,t))_{i \geq 1}$ be the total configuration at time t.

$\boldsymbol{G_n}$ is discrete time markov-chain

For any t and $i \geq 1$

$$\mathbb{E}[G_n(i, t+1/n) - G_n(i,t) \mid \boldsymbol{G_n}(t)] = \frac{1}{n}(G_n(i-1,t)^{d_n} - G_n(i,t)^{d_n})$$

$g_n(i,t)$ satisfies recursion

$$g_n(i,t) = \int_0^t g_n(i-1,s)^{d_n} - g_n(i,s)^{d_n} ds + O\left(\frac{d_n^2}{n}\right)$$

Completing the proof (analyze the recursion)

Approximating g_n using an ODE

Suppose $\{g(i,t)\}$ satisfy:

$$g(i,t) = \int_0^t g(i-1,s)^{d_n} - g(i,s)^{d_n} ds \qquad \forall t \in [0,1] \text{ and } i \geq 1$$

Then $\sup_{s\in[0,1]} |g_n(i,s) - g(i,s)| \leq \frac{15e^i d_n^{i+2}}{n}$ for any $i \geq 1$.

Estimates on the growth of the ODE

$$\exp(-d_n^{i+1}) \leq g(i,1) \leq \exp(-d_n^{i-1}).$$

Double exponential decay in i.

Recall $d_n < \text{Poly}(\log n)$ and $i_n^* = \min\left\{ i \mid g_n(i,n) \leq \frac{\ln n}{\sqrt{n}} \right\}$. Then

$$\frac{\log\log n}{\log d_n} - 3 \leq i_n^* \leq \frac{\log\log n}{\log d_n} + 3 \quad \text{eventually as } n \to \infty .$$

Outline

How to choose d_n?

Recall: The maximum for the n Balls-and-Bins problem using strategy $\texttt{Smallest}(\mathrm{d_n})$ is between

$$\left[\frac{\log\log n}{\log d_n} - 4, \frac{\log\log n}{\log d_n} + 4\right] \qquad \text{w.h.p,}$$

provided that $1 < d_n < \text{Poly}(\log n)$.

- Need $d_n \to \infty$ to keep the maximum load bounded.
- Choose $d_n = (\log n)^\delta$ to keep the maximum load under $4 + \frac{1}{\delta}$, w.h.p.

We can get near optimal performance using $\texttt{Smallest}(\log \mathrm{n})$.

Outline

Outline

1. Balls and bins
   - Power of choice ($d = 1$ vs. $d = 2$)
   - Dependence on $d \geq 1$
   - How to choose d?
2. Supermarket model
   - Introduction
   - Analysis of join the shortest queue
   - Fluid limit for $\texttt{JSQ}(d_N)$ as $d_N \to \infty$
   - Diffusion limit theorem
3. Summary

The Supermarket Model

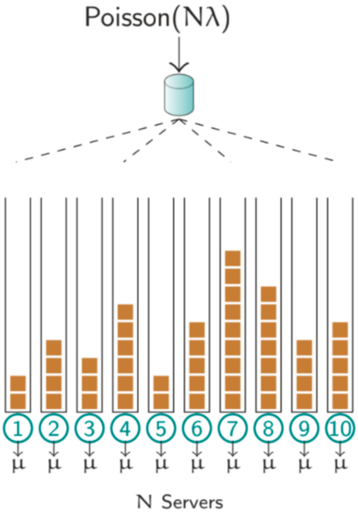


Image credit : Debankur Mukherjee

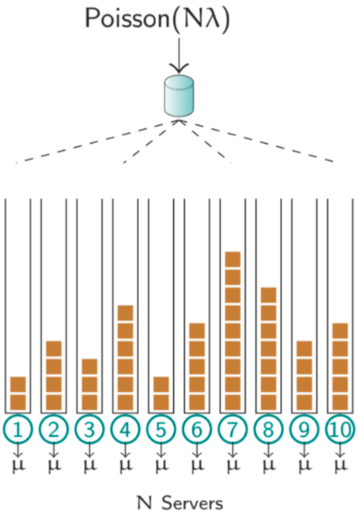
How to route these customers?

- At random.
  Overhead = 0. JSQ(1)
- Join the shortest queue (JSQ).
  Overhead = N. JSQ(N)
- JSQ(d) for $d \geq 2$.
  Overhead = d.

JSQ(d) : Choose a random size-d subset of servers and join the shortest queue among that subset.

Application to load balancing

JSQ is optimal



Data centers : customers are connections and computers are the N servers.

- Customers can't be queued at the dispatcher.
- Keep queues balanced to make best use of resources.
- Need efficiency ($\frac{\lambda_N}{\mu} \uparrow 1$).

JSQ:

- Optimial among all non-anticipating policies (Winston, 1977).

Image credit : Debankur Mukherjee

Outline

Asymptotic performance of JSQ

Halfin-Whitt regime : $\lambda_N = 1 - \frac{\beta}{\sqrt{N}}$

Let

- $G_{N,i}(t) = \dfrac{\#\text{ of servers with } \geq i \text{ customers at time } t}{N}$.
- $Z_{N,1} = \sqrt{N}(G_{N,1} - 1)$ and $Z_{N,i} = \sqrt{N} G_{N,i}$ for $i = 2, 3 \dots$.

Diffusion limit for JSQ (Eschenfeldt and Gamarnik, 2015)

If $(Z_{N,1}(0), Z_{N,2}(0)) \xrightarrow{P} (z_1, z_2)$ with $z_1 \leq 0$, $Z_{N,3}(0) = 0$ as $N \to \infty$, then $(Z_{N,1}, Z_{N,2}, Z_{N,3}) \Rightarrow (Z_1, Z_2, 0)$ in $\mathbb{D}^3$ where

$$Z_1(t) = z_1 + \sqrt{2}B(t) - \beta t - \int_0^t Z_1(s) - Z_2(s) ds - U(t)$$

$$Z_2(t) = z_2 + U(t) - \int_0^t Z_2(s) ds$$

B is a brownian motion, and U is the unique non-decreasing process so that $U(0) = 0$, $\int_0^t \mathbb{I}_{\{Z_1(s) < 0\}} dU(s) = 0$ and $Z_1 \leq 0$.

Can JSQ(d) be as good as JSQ?

letting $d \to \infty$

Need $d_N \to \infty$ for a typical customer's waiting time to vanish like in JSQ (Gamarnik, Tsitsiklis, Zubeldia, 2016).

(Mukherjee, Borst, Leeuwaarden, Whiting 2018)

- As long as $d_N \to \infty$, the first order (fluid-scale) limiting behaviors of JSQ(d_N) and JSQ agree. (Universality of fluid limit.)
- The second order (diffusion-scale) behavior of JSQ(d_N) in the Halfin-Whitt regime is the same as JSQ if $\frac{d_N}{\sqrt{N}\log N} \to \infty$. (Diffusion level optimality.)

We provide explicit limit theorems for first and second order behavior of JSQ(d_N), as $d_N \to \infty$ and $\lambda_N \to 1$.

Outline

Crash course on Skorokhod map

If $x \in \mathbb{D}_+ \doteq \{ f \in \mathbb{D} \mid f(0) \geq 0 \}$,
then $\exists!\, y \in \mathbb{D}_+$ so that

- $z(t) = x(t) + y(t)$
- $z(t) \geq 0$
- y satisfies
  - $y(0) = 0$
  - y is non-decreasing
  - $\int_{[0,\infty)} z(s)dy(s) = 0$

Explicit Skorokhod map

Define $\Phi : \mathbb{D}_+ \to \mathbb{D}_+^2$ by
$\Phi(x) = (z, y)$ where

$$y(t) = \sup_{0 \leq s \leq t} x^-(s)$$

$$z(t) = x(t) + y(t)$$

Φ is Lipscitz with respect to the supremum norm

$$\|\Phi(x) - \Phi(y)\|_{*,t} \leq 2\,\|x - y\|_{*,t}$$

where $\|f\|_{*,t} = \sup_{s \in [0,t]} |f(s)|$.

Fluid behavior of $JSQ(d_N)$

Recall: $\boldsymbol{G_N}(t) = (G_{N,1}(t), G_{N,2}(t), G_{N,3}(t), \ldots)$

Fluid limit as $d_N \to \infty$ and $\lambda_N \to \lambda$

If $\boldsymbol{G_N}(0) \xrightarrow{P} (r_1, r_2, \ldots)$ in l_1, then $\boldsymbol{G_N} \xrightarrow{P} \boldsymbol{g}$ in $D([0,\infty) : l_1)$

where $\boldsymbol{g} = (g_1, g_2, \ldots)$ is the unique solution to

$$(g_i, v_i) = \Phi_1\left(r_i - \int_0^{\cdot} g_i(s) - g_{i+1}(s)ds + v_{i-1}(\cdot)\right) \qquad i = 1, 2, \ldots$$

and $v_0(t) = \lambda t$. $\Phi_1 : \mathbb{D}_{\leq 1} \to \mathbb{D}^2$ is the Skorokod map at 1. Universality.

- Proof uses tightness + uniqueness argument.
- (Mukherjee, Borst, Leeuwaarden, Whiting 2018) identify limiting equations but can't show uniqueness.
- Formulation using Skorokhod map shows uniqueness.

Fluid behavior of $JSQ(d_N)$

Recall: $\boldsymbol{G_N}(t) = (G_{N,1}(t), G_{N,2}(t), G_{N,3}(t), \dots)$

Fluid limit as $d_N \to \infty$ and $\lambda_N \to \lambda$

If $\boldsymbol{G_N}(0) \xrightarrow{P} (r_1, r_2, \dots)$ in l_1, then $\boldsymbol{G_N} \xrightarrow{P} \boldsymbol{g}$ in $D([0,\infty) : l_1)$

where $\boldsymbol{g} = (g_1, g_2, \dots)$ is the unique solution to

$$(g_i, v_i) = \Phi_1\left(r_i - \int_0^{\cdot} g_i(s) - g_{i+1}(s)ds + v_{i-1}(\cdot)\right) \qquad i = 1, 2, \dots$$

and $v_0(t) = \lambda t$. $\Phi_1 : \mathbb{D}_{\leq 1} \to \mathbb{D}^2$ is the Skorokod map at 1. Universality.

- Proof uses tightness + uniqueness argument.
- (Mukherjee, Borst, Leeuwaarden, Whiting 2018) identify limiting equations but can't show uniqueness.
- Formulation using Skorokhod map shows uniqueness.

Proof overview (fluid limit)

Representation as Poission time-change

For $i = 1, 2, \dots$

$$G_{N,i}(t) = G_{N,i}(0) - \frac{1}{N} D_i\left(N \int_0^t G_{N,i}(s) - G_{N,i+1}(s) ds\right) + \frac{1}{N} A_i\left(\lambda_N N \int_0^t G_{N,i-1}(s)^{d_N} - G_{N,i}(s)^{d_N} ds\right)$$

where $\{A_i\}_{i\geq 1}, \{D_i\}_{i\geq 1}$ are independent rate-1 poission processes.

Subtract compensators:

$$G_{N,i}(t) = G_{N,i}(0) - \int_0^t G_{N,i}(s) - G_{N,i+1}(s) ds + \lambda_N \int_0^t G_{N,i-1}(s)^{d_N} - G_{N,i}(s)^{d_N} ds + M_{N,i}(t)$$

$\boldsymbol{M_N}(t) = (M_{N,i}(t))_{i\geq 1}$ is a collection of martingales with $\mathbb{E} \left\| \boldsymbol{M_N} \right\|_{*,T} \to 0$

Outline

Ingredient : diffusion centering

Fix N. Omiting the martingale term, the previous ODE is:

$$\begin{aligned} G_{N,i}(t) = G_{N,i}(0) + & \int_0^t (\lambda_N G_{N,i-1}(s)^{d_N} - G_{N,i}(s))ds \\ & - \int_0^t (\lambda_N G_{N,i}(s)^{d_N} - G_{N,i+1}(s))ds \end{aligned}$$

Unique fixed point : $\boldsymbol{\mu_N} = (\lambda_N, \lambda_N^{1+d_N}, \lambda_N^{1+d_N+d_N^2}, \ldots) \in l_1$.

Diffusion scaled process

$$\boldsymbol{Z_N} = \sqrt{N}(\boldsymbol{G_N} - \boldsymbol{\mu_N})$$

This is different from the usual fluid limit centering, which may not be stable.

Diffusion behavior for JSQ(d_N) : reflected case

Recall: $\boldsymbol{Z_N} = \left(\sqrt{N}(G_{N,1} - \lambda_N), \sqrt{N}(G_{N,2} - \lambda_N^{1+d_N}), \ldots\right)$.

Diffusion limit as $\lambda_N = 1 - \left(\frac{\log d_N}{d_N} + \frac{\alpha}{\sqrt{N}}\right)$ and $\sqrt{N} \ll d_N \ll N^{2/3}$

If $\boldsymbol{Z_N}(0) \xrightarrow{P} (z_1, z_2, 0, 0 \ldots)$ in l_2 with $z_1 \le \alpha$, then $\boldsymbol{Z_N} \Rightarrow (Z_1, Z_2, 0, 0 \ldots)$ in $D([0,\infty) : l_2)$ where (Z_1, Z_2) satisfy

$$Z_1, U_1 = \Phi_\alpha\left(z_1 + \sqrt{2}B(\cdot) - \int_0^{\cdot} (Z_1(s) - Z_2(s))ds\right)$$

$$Z_2(t) = z_2 + U_1(t) - \int_0^t Z_2(s)ds,$$

B is a standard Brownian motion and $\Phi_\alpha : \mathbb{D}_{\le\alpha} \to \mathbb{D}^2$ is reflection at α.

- When $d_N \gg \sqrt{N}\log N$, limit agrees with JSQ (Eschenfeld and Gamarnik, 2015) and (Mukherjee, Borst, Leeuwarden, Whiting 2018).

Proof idea (diffusion limit)

Center and scale the generating equation

$$Z_{N,1}(t) = Z_{N,1}(0) - \int_0^t Z_{N,1}(s) - Z_{N,2}(s)ds$$
$$+ \sqrt{N} M_{N,1}(t) - \int_0^t t_{N,1}(Z_{N,1}(s))ds$$
$$Z_{N,2}(t) = Z_{N,2}(0) + \int_0^t t_{N,1}(Z_{N,1}(s))ds - \int_0^t Z_{N,2}(s)ds + o_p(1).$$

Under hyothesis $\sqrt{N} M_{N,1} \Rightarrow \sqrt{2}B$.

Reflection term

Fix any $M > 0$. Then uniformly on $z \in [-M, M]$

$$t_{N,1}(z) = (1 + o(1)) \exp\left(\frac{d_N}{\sqrt{N}}(z - \alpha)\right) \frac{\sqrt{N}}{d_N}$$

Proof outline (diffusion limit)

Choose $M > 0$: $T_{N,M} = \inf \{ t \mid \|\boldsymbol{Z_N}(t)\|_2 \geq M \} \wedge T$

$Z_{N,1}$ will not exceed α on $[0, T_{N,M}]$

$$\sup_{t \in [0, T_{N,M}]} (Z_{N,1} - \alpha)^+ \xrightarrow{P} 0$$

Rewrite using skorokhod map

$$Z_{N,1}, U_N = \Phi_\alpha\left(Z_{N,1}(0) - \int_0^{\cdot} Z_{N,1}(s) - Z_{N,2}(s) + \sqrt{2} B_N(\cdot) \right) + o_p(1)$$

$$Z_{N,2}(t) = Z_{N,2}(0) + U_N(t) - \int_0^t Z_{N,2}(s) ds + o_p(1)$$

where $B_N \Rightarrow B$, and the $o_p(1)$ terms converge uniformly on $[0, T_{N,M}]$.

Tightness

Choose M large enough so that $T_{N,M} \geq T$ eventually.

Diffusion behavior for $\texttt{JSQ}(d_N)$: non-reflection case

Diffusion limit as $\frac{d_N}{\sqrt{N}} \to 0$ and $d_N \mu_{N,k+1} \to \alpha$

If $\boldsymbol{Z_N}(0) \xrightarrow{P} (z_1, \ldots, z_{k+1}, 0, 0, \ldots)$ in l_2, then
$\boldsymbol{Z_N} \Rightarrow (0, \ldots, 0, Z_k, Z_{k+1}, 0, 0, \ldots)$, where

$$Z_k(t) = z_k - (\alpha + \mathbb{I}_{\{k=1\}}) \int_0^t Z_k(s)ds + \int_0^t Z_{k+1}(s)ds + \sqrt{2}B(t)$$

$$Z_{k+1}(t) = z_{k+1} + \alpha \int_0^t Z_k(s)ds - \int_0^t Z_{k+1}(s)ds$$

Here B is a standard Brownian motion.

Outline

Future direction : distributed load balancing

Blogs interested in distributed balancing using Power-of-d scheme:

- Nginx
- Haproxy
- Mark's

Network model from (Budhiraja, Mukherjee, Wu, 2019).

Acknowledgement

Mentors and collaborators: Shankar Bhamidi and Amarjit Budhiraja.

Supporting Grants

- NIH R01 HG009125-01
- National Science Foundation, DMS-1613072