**Submitted to *Stochastic Systems***
**manuscript MS-0001-1922.65**



# State Dependent Control of Closed Queueing Networks

**(Authors' names blinded for peer review)**

We study the design of state-dependent control for a closed queueing network model inspired by shared transportation systems such as those for ridesharing and bike-sharing. We focus on the assignment policy, where the platform can choose which supply unit to assign when a customer request comes in, and assume that this is the exclusive control lever available. The supply unit once again becomes available at the destination after dropping the customer. We consider the proportion of dropped demand in steady state as the performance measure.

We propose a family of simple and explicit state-dependent policies called Scaled MaxWeight (SMW) policies and prove that under the complete resource pooling (CRP) condition (analogous to the condition in Hall's marriage theorem), each SMW policy leads to exponential decay of demand-dropping probability as the number of supply units scales to infinity. We further show that there is an SMW policy that achieves the optimal exponent among all assignment policies, and analytically specify this policy in terms of the customer arrival rates for all source-destination pairs. The optimal SMW policy protects structurally under-supplied locations.

*Key words*: maximum weight policy, closed queueing network, control, Lyapunov function, shared transportation systems, large deviations

## 1. Introduction

Recently there is an increasing interest in the control of shared transportation systems such as Uber and Lyft. These platforms are dynamic two-sided markets where customers (demands) arrive at different physical locations stochastically over time, and vehicles (supplies) circulate in the system as a result of driving demands to their destinations. The platform's goal is to maximize throughput (proportion of demands fulfilled), revenue or other objectives by employing various types of controls.

The main inefficiency in such systems comes from the geographic mismatch of supply units and customers: when a customer arrives, he has to be matched immediately with a nearby supply unit, otherwise the customer will abandon the request due to impatience. There are two sources of spatial supply-demand asymmetry: *structural imbalance* and *stochasticity*. In ridesharing, the former is dominant during rush hours in the city when most of the demand pickups concentrate in a particular region of the city while dropoffs concentrate on others, otherwise the latter source often dominates, see Hall et al. (2015). *We propose an assignment (scheduling) policy that deals with both sources simultaneously (if this is possible): the choice of policy parameters accounts for the structural imbalance, while its state dependence nature manages stochasticity optimally.*

Many control mechanisms have been proposed and analyzed in the literature. *Pricing*, for example, enjoys great popularity in both academia and industry, see, e.g., Waserhole and Jost (2016), Banerjee et al. (2016). By adjusting prices of rides, the system can indirectly re-balance supply and demand. *Empty-vehicle routing* (Braverman et al. 2016) focus on sending available supply units to under-supplied locations in order to meet more demand.

In this paper, we study another important form of control, *assignment*. When a customer request comes in, the platform can decide from where to assign a supply unit, which will in turn influence the platform's future ability to fulfill demands. Previous work has studied assignment decisions based on fluid limits – by optimizing the system on fluid scale, the platform can calculate the probability of assigning from any compatible location when a demand arises, and realize it by randomization (Ozkan and Ward 2016) (hereafter referred to as fluid-based policy). However, this approach requires exact knowledge of customer arrival rates (which is infeasible in practice), fails to react to the stochastic variation in the system and creates additional variance due to randomization. Although this control guarantees asymptotic optimality in the Law of Large Numbers sense, it converges only slowly to the fluid limit (Banerjee et al. 2016). One might expect that by implementing fluid-based policy in closed loop (i.e. allow resolving once in a while, hereafter referred to as resolving fluid-based policy) it will perform better, but in our setting (where the objective is minimizing dropped demand) this is not the case [1]. To counter these issues, *we study the state-dependent assignment control of closed queueing networks*[2].

We model the system as a closed queueing network with $n$ servers representing physical locations, and $K$ "jobs" that stand for supply units. This is a common model for ridesharing systems, see for example Banerjee et al. (2016), Braverman et al. (2016), Waserhole and Jost (2016). For each location $i$ there are some compatible supply locations that are close enough, from where the platform can assign supply units to serve customer at $i$. Customers arrive at the system stochastically, each has a destination in mind. Each time a customer arrives, the platform makes an assignment decision from a compatible supply location based on the current spatial distribution of available supplies. After a supply unit picks up a customer, it drops her at the destination and becomes available again. (Supplies do not enter or leave the system.) The platform's goal is to maintain adequate supply in all neighborhoods and hence meet as much customer demands as possible, therefore we adopt the (global) proportion of dropped demands as a measure of efficiency. For the formal description of our model, see section 2.

To study state-dependent spatial rebalancing of supply while keeping the size of the state space manageable, we make a key simplification – we assume that pickup and service of customers are both instantaneous. This allows us to get away from the complexity of tracking the positions of in transit supply units, while retaining the essence of our focal challenge, that of ensuring that all neighborhoods have supplies at (almost) all times. To obtain tight characterizations, we further consider the asymptotic regime where the number of supply units in the system $K$ goes to infinity. It's worth noting that the large supply regime (demand arrival rates stay fixed while $K \to \infty$) and the large market regime (demand arrival rate scales with $K$ as $K \to \infty$) are equivalent in our model. The reason is that we assume instantaneous relocation of assigned supply units, hence the large market regime is simply a speed-up of the infinite supply regime.

A main assumption in our model is the complete resource pooling (CRP) condition. CRP is a standard assumption in the heavy traffic analysis of queueing systems (see e.g. Harrison and López 1999, Dai and Lin 2008, Shi et al. 2015). It can be interpreted as requiring enough overlapping in the processing ability of servers so that they form a "pooled server". For the model considered in this paper, the CRP condition is closely related to the condition in Hall's marriage theorem in bipartite matching theory. We show that the CRP condition is necessary for any assignment policy to have demand dropping probability that converges to zero. In ridesharing the demand arrival rates are endogenously determined by system manager's pricing/admission control decisions

[1] In fact, any static fluid-based policy that keeps the number of demands at each location unchanged in the fluid limit is also an optimal resolving fluid-based policy.

[2] Note that {fluid-based policy}$\subset${resolving fluid-based policy}$\subset${state-dependent policy}.

(though we don't focus on that here). If CRP condition doesn't hold for base prices, the platform may use spatially varying prices as in Bimpikis et al. (2016), Banerjee et al. (2016) to make it true, and then use our assignment control to manage the unpredictable fluctuations in the system.

One key difficulty in the analysis is the necessity to deal with a multi-dimensional system in the limit. In many existing works that seek to minimize the workload process or holding costs of a queueing system, asymptotic optimality of a certain policy relies on "collapse" of the system state to a lower dimensional space in the heavy traffic limit under the CRP condition. This is not the case for the objective we are considering, i.e., minimizing demand dropping probability, or maximizing throughput. Since the exponent of demand dropping probability depends on the most likely event that leads to demand dropping event, we need to "protect" all the subsets of locations simultaneously. An interesting observation drawn from our analysis is a "critical subset" property: given the current state, the most likely way a demand may be dropped is via draining a certain critical subset of locations. The critical subset changes with the state of the system.

**Formulated as classic closed queueing network scheduling problem.** Using terminology of classic queueing theory, the $K$ supply units are "jobs", each demand location is a "server", each supply location is a "buffer", inter-arrival times of customers with origin $i$ are "service times" at server $i$. The distribution of customers' destinations given origin captures "routing probabilities". "Servers" are flexible (i.e. can serve multiple queues), and assignment is equivalent to "scheduling". Our model deviates a little from classic model in that a scheduled job relocates to destination queue at the start of its service time, instead of the end.

### 1.1. Main Contributions

As a function of system primitives, we derive a large deviation rate-optimal assignment policy that minimizes demand dropping (maximizes throughput). Our optimal policy is strikingly simple and its parameters depend in a natural way on demand arrival rates. Our contribution is threefold:

1. **Achievability**: We propose a family of state-dependent assignment policies called Scaled MaxWeight (SMW) policies, and prove that all of them guarantee exponential decay of demand-dropping probability under CRP condition. The proof is based on a family of novel Lyapunov functions (a different one for each SMW policy) which are used to analyze a multi-dimensional variational problem. An SMW policy is parameterized by an $n$-dimensional vector consisting of a scaling factor for each location; each demand is served by assigning a supply from the compatible location with the largest scaled number of cars. We obtain an explicit specification for the optimal scaling factors based on location compatibilities and demand arrival rates. Further, we obtain the surprising finding that the optimal SMW policy is, in fact, exponent-optimal among all state-dependent policies (Theorem 1). SMW policies are simple, explicit and appear promising for practical applications.

2. **Converse bounds**: We provide lower bound of demand dropping probability for any assignment policy using random-walk related inequalities. We first show that no fluid-based assignment policy can achieve exponential decay rate (Proposition 4), which demonstrates the value of state-dependent control — even a naive state-dependent assignment policy with no knowledge of demand arrival rates beats the best fluid-based assignment policy asymptotically. Then we justify the CRP assumption by showing that it is a necessary condition for exponential convergence (Proposition 1; in fact if any of the inequalities is reversed, a positive fraction of demands must be dropped for that instance even as $K \to \infty$). Finally, we obtain an upper bound on the demand dropping probability exponent for any state-dependent policy that matches the achievable exponent of the optimal SMW policy, thus proving that the best SMW policy is, in fact, exponent-optimal.

3. **Qualitative insights**: We characterize the system behavior under SMW policies as $K \to \infty$, which is technically challenging since the problem remains $n$ dimensional even in the limit. We establish the *critical subset* property of the problem: given a system state (in the limit), there exists a (state-dependent) subset $J$ of demand locations that are most likely to be depleted of supply in compatible locations, hence leading to demand dropping. The optimal SMW policy avoids using supply from locations compatible with $J$ for demand arising outside of $J$.

### 1.2. Literature Review

**MaxWeight scheduling.** MaxWeight policy is a simple scheduling policy in constrained queueing networks that exhibits many good properties. It attaches a weight to each schedule activity $u$, which is a function of current queue-lengths $\{f_u(\mathbf{X})\}$; in many cases $f_u(\mathbf{X})$ is simply the queue length at the scheduled server (source). At each time period it activates the admissible activities with the largest total weight.

MaxWeight scheduling has been studied intensively since the seminal work Tassiulas and Ephremides (1992), which showed MaxWeight (with weight defined as a constant multiple of source-destination queue-length difference) achieves the entire stability region for an open one-pass system. Dai and Lin (2005) showed that MaxWeight (with the same choice of weight as Tassiulas and Ephremides 1992) achieves throughput optimality in open stochastic processing networks in fluid limit under mild conditions.

Stolyar (2004) showed that for one-hop open network, any MaxWeight policy with weight $f_u(\mathbf{X})$ chosen as a positive multiple of the $\beta$-th ($\beta > 0$) power of source queue length, minimizes workload (weighted sum of queue lengths) under Resource Pooling (RP) condition in diffusion limit. A similar result was obtained in Dai and Lin (2008), where they showed that for a more general family of open networks, MaxWeight policy with weight defined as in Dai and Lin (2005) minimizes workload in diffusion limit under CRP condition.

Eryilmaz and Srikant (2012) showed that for one-hop network, MaxWeight policy with weight $f_u(\mathbf{X}) = c_{\text{source}(u)}\mathbf{X}_{\text{source}(u)}$ minimizes the expectation of $\sum_i c_i\mathbf{X}_i$ in heavy traffic limit with one-dimensional state-space collapse. Maguluri and Srikant (2016) showed that for cross-bar switch MaxWeight policy with $f_u(\mathbf{X}) = \mathbf{X}_{\text{source}(u)}$ achieves the optimal scaling of total queue-length in heavy traffic with multi-dimensional state space collapse. Shi et al. (2015) considered a model similar to Eryilmaz and Srikant (2012) with focus on design of network topology.

In contrast of the many of the above works where asymptotic optimality is guaranteed for any choice of constant $c_{\text{source}(u)} > 0$ in weight $f_u(\mathbf{X})$, we show that in our model the coefficients need to be chosen carefully in order to achieve optimal exponent, though any choice of positive constants will result in exponential decay under CRP condition. For example, vanilla MaxWeight ($f_u(\mathbf{X}) = \mathbf{X}_{\text{source}(u)}$) can result in non-trivial sub-optimality of system performance. This is because for open queueing networks under CRP condition, though the queue lengths can collapse to different subspaces under different MaxWeight policies, the workload process always converges to the same weak limit which is uniquely defined by the model primitives. In our case, however, different state space collapse usually leads to different system performance (exponent); and evaluation of each MaxWeight policy requires analysis of policy-specific most-likely sample paths.

There are other variants of MaxWeight policy that are less related to our work, e.g. Meyn (2009).

**Large deviations in queueing systems.** There is a large literature on characterizing the decay rate of probability of building up large queue lengths in open queueing networks. To the best of our knowledge, we are the first to consider large deviation of controlled closed queueing networks.

Stolyar and Ramanan (2001) showed the exponent optimality of Largest Weighted Delay First scheduling in minimizing the probability of waiting times exceeding large values for multi-class single server queueing system, and Stolyar (2003) extended the result to multi-class multi-server queueing networks (the choice of weights corresponds to the objective at hand, in contrast with our work, where the optimal choice of weights will be determined by the model primitives). Bodas et al. (2014) considered the large deviation optimal scheduling of parallel servers, but the asymptotic regime is different in that they are scaling up the size of network while keeping buffer size fixed. Compared with these works, the difficulty of analyzing our model comes from its complex dynamics: supply units circulate in the system endlessly (in contrast to one-hop system in Bodas et al. 2014), and each supply unit can be matched to demands at different nodes (in contrast to multi-class model in Stolyar and Ramanan 2001, Stolyar 2003). As a result, the techniques used in these works cannot be directly applied here. Our result is also qualitatively different: in Stolyar (2003) the analysis of queueing networks with arbitrary topology reduces to studying one node in isolation, but in our work the optimal exponent include terms for each subset of nodes.

The closest work to ours is that of Venkataramanan and Lin (2013), who established the relationship between Lyapunov function and buffer overflow probability for open queueing networks. Our Lyapunov function approach is inspired by their work, and we devise a family of Lyapunov functions such that the decay rate of demand dropping probability is the same as that of Lyapunov function exceeding certain threshold. The key difficulty of extending the Lyapunov approach to closed queueing networks is the lack of natural reference state where the Lyapunov function equals to 0 (in open queueing network it's simply $\mathbf{0}$). It turns out when optimizing the MaxWeight parameters we are also solving for the best reference state.

There are also works that study the large deviation behavior of queueing networks without control aspect, see e.g. Majewski and Ramanan (2008), Blanchet (2013).

**Shared Transportation Systems.** Optimization of shared transportation systems such as those for ridesharing and bike-sharing has drawn attention in recent years. Ozkan and Ward (2016) studied revenue-maximizing fluid-based assignment control by solving a minimum cost flow problem in the fluid limit. Braverman et al. (2016) modeled the system by a closed queueing network and derived the optimal static routing policy that sends empty vehicles to under-supplied locations. Banerjee et al. (2016) adopted Gordon-Newell closed queueing network model and considered static pricing policy that maximize throughput, welfare or revenue. In contrast to our work, which studies state-dependent control, these works consider static control that completely relies on system parameters. In terms of convergence rate to the fluid-based solution, Ozkan and Ward (2016) did not show the order of convergence rate of their policy, Braverman et al. (2016) proved an $O(1/\sqrt{K})$ rate as number of supply units in the closed system $K$ goes to infinity, and Banerjee et al. (2016) showed an $O(1/K)$ convergence rate as $K \to \infty$. When CRP condition holds, we show that no fluid-based policy can do better than $O(1/K^2)$, while our state-dependent policy achieves an $O(e^{-\gamma K})$ convergence rate with optimal $\gamma > 0$.

We remark briefly that our CRP condition has some similarity with the notion of "balancedness" in Bimpikis et al. (2016), although balancedness requires an exact balance between inflows and outflows for each location[3], whereas the CRP condition does not, due to the flexibility from having multiple compatible supply locations for a demand location.

There are many other works that also studied ridesharing platforms but focus on pricing aspects, see, e.g., Adelman (2007), Bimpikis et al. (2016), Waserhole and Jost (2016), Cachon et al. (2017), Hall et al. (2015).

**Online Stochastic Bipartite Matching.** There is a stream of research on online stochastic bipartite matching (OSBM) that are related to our work, see e.g. Caldentey et al. (2009), Adan

[3] To compare with our closed network model, we consider here the special case of the model in Bimpikis et al. (2016) where drivers never exit or enter, i.e., their parameter $\beta = 1$.

and Weiss (2012), Bušić and Meyn (2015), Mairesse and Moyal (2016). In OSBM, different types of supplies and demands arrive over time, and the system manager matches supplies with demands of compatible types using some matching policy then discharge the matched pairs from the system. Many matching policies considered in the literature have the flavor of MaxWeight, if not exactly the same: in Caldentey et al. (2009) and Adan and Weiss (2012) demands are matched to the oldest compatible supplies available, Mairesse and Moyal (2016) considered "match the longest" policy. Our model is different from OSBM in the following ways: (1)our model is closed hence supply units never leave enter or leave the system, while in OSBM the network is open; (2)many OSBM works focus on the stability and other properties under a given policy instead of looking for the optimal control, e.g. Caldentey et al. (2009), Adan and Weiss (2012), Mairesse and Moyal (2016); (3)in terms of dynamics, in Caldentey et al. (2009), Bušić and Meyn (2015) exactly one supply and one demand enter the system each time, which is different from our model.

### 1.3. Organization

The remainder of our paper is organized as follows. In section 2 we introduce the basic notation used throughout the paper and formally describe our model together with performance measure. Some background on sample path large deviation principle will also be provided. In section 3 we introduce the family of Scaled MaxWeight policies. In section 4 we present our main theoretical result, i.e., exponent optimality of SMW policy. In section 5 we prove the achievability bounds of of SMW policies. In section 6 we prove the converse bounds and show the exponent optimality of SMW policy. In section 7 we conclude the paper with further discussion of results and future directions. We also present there the converse bounds for fluid-based policies and cases where CRP condition is violated.

## 2. Model and Preliminaries

### 2.1. Notation

Wherever possible, we reserve capital letters for random quantities and small letters for their realizations; we also use boldface letters to indicate column vectors. We use $\mathbf{e}_i$ to denote the $i$-th unit vector, and $\mathbf{1}$ the all-1 vector. If vector $\mathbf{b}$ is strictly larger than $\mathbf{a}$ component-wise, we write $\mathbf{b} > \mathbf{a}$. For index set $A$, define $\mathbf{1}_A \triangleq \sum_{i\in A} \mathbf{e}_i$. For a set $\Omega$ in Euclidean space $\mathbb{R}^n$, denote its relative interior by relint($\Omega$). We use $B(\mathbf{x},\epsilon)$ to denote a ball centered at $\mathbf{x}\in\mathbb{R}^n$ with radius $\epsilon > 0$. For event $C$, we define the indicator random variable $\mathbb{1}\{C\}$ to equal 1 when $C$ is true, else 0.

### 2.2. Basic Setting

**Underlying Model and Simplifications**: We model the shared transportation system as a finite-state Markov process, comprising of a fixed number of identical *supply units* circulating among $n$

*nodes* (i.e., a given partition of a city into neighborhoods). Customers (i.e., prospective passengers) arrive at each node $i$ with desired destination $j$ according to independent Poisson processes with rate $\hat{\phi}_{ij}$. To serve an arriving customer, the platform immediately assigns a supply unit from a "neighboring" station of $i$ (i.e., one among a set of nearby stations, defined formally below), and subsequently, after serving the customer, the supply unit becomes available at the destination node $j$. If however there are no supplies available in the neighboring nodes of $i$, then we experience a *demand drop*, wherein the customer leaves the system without being served. Customers do not wait. The aim of the platform is to assign supplies so as to minimize the fraction of demands dropped. Intuitively, to achieve this objective, the platform should ensure that it maintains adequate supply in (or near) all neighborhoods, i.e., it needs to manage the spatial distribution of supply.

To study the design of assignment rules in the above model, we make two simplifications. First, we assume that pickup and service are instantaneous. This allows us to reformulate the above model as a *discrete-time* Markov chain (the so-called *jump chain* of the continuous-time process), where in each time-slot $t \in \mathbb{N}$, with probability proportional to $\hat{\phi}_{ij}$, exactly one customer arrives to the system at node $i$ with desired destination $j$. The customer is then served by an assigned supply unit from a neighboring station of $i$, which then becomes available at node $j$ at the beginning of time-slot $t+1$. This simplification removes the high-dimensionality required for tracking the positions of all in-transit supply units, while still retaining the complex supply externalities between stations, which is the hallmark of ridesharing systems. Unfortunately, however, even after this simplification, the setting still does not admit any amenable way to characterize the performance of complex assignment policies. To circumvent this we make a second simplification, we study the performance of assignment policies as the number of supply units $K$ grow to infinity, while fixing all other parameters.

**Formal System Definition**: We define $\phi \in \mathbb{R}^{n\times n}$ to be the arrival rate matrix with a row for each origin and a column for each destination, normalized[4] such that $\mathbf{1}^T\phi\mathbf{1} = 1$. We denote the $i$-th column (i.e., the arrival rates from different origins of customers to destination $i$) as $\phi_{(i)}$, and the transpose of the $i'$-th row of the arrival matrix (i.e., the arrival rates of customers to node $i'$ with different destination nodes) as $\phi_{i'}$. Thus, the probability a customer arrives at node $i'$ is $\mathbf{1}^T\phi_{i'}$, and, assuming all customers are matched, the rate of supply units arriving at node $i$ is $\mathbf{1}^T\phi_{(i)}$.

As mentioned above, we consider a sequence of systems parameterized by the number of supplies $K$. For the $K$-th system, its state at any time $t \in \mathbb{N}$ is given by $\mathbf{X}^K[t]$, a vector that tracks the

[4] This can always be achieved by appropriately re-scaling the arrival rates $\{\hat{\phi}_{ij}\}$, which produces an equivalent setting since pickups and dropoffs are instantaneous.

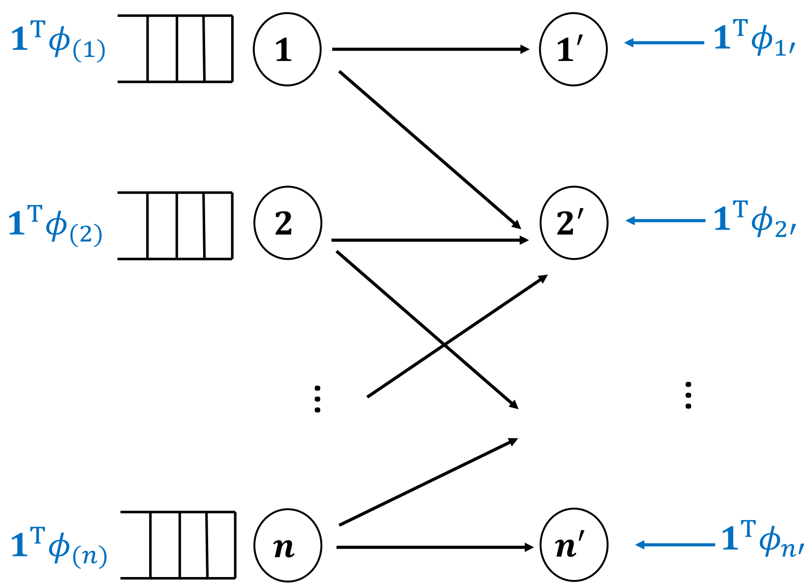


**Figure 1 The bipartite compatibility graph for the assignment problem: On the left are supply nodes $i$, and on the right are demand nodes $i'$. Customers arrive to $i'$ with distributions chosen according to $\phi_{i'}$; the total probability of a customer arrival at $i'$ is thus $1^T\phi_{i'}$. Similarly, assuming no demand is dropped, the total probability a supply unit arrives at $i$ is $1^T\phi_{(i)}$. The edges entering a node $i'$ encode compatible (i.e., nearby) nodes that can supply node $i'$.**

number of supplies at each location in time-slot $t$. The state space of the $K$-th system is thus given by:

$$\Omega_K \triangleq \{\mathbf{x} \in \mathbb{R}^n | \mathbf{1}^T\mathbf{x} = K\} \cap \mathbb{N}^n.$$

Note that the normalized state $\mathbf{X}^K/K$ lies in the $n$-simplex $\Omega = \{\mathbf{x} \in \mathbb{R}^n | \mathbf{x} \geq 0, \mathbf{1}^T\mathbf{x} = 1\}$. Henceforth, we drop explicit dependence on $t$ when clear from context.

## 2.3. Assignment Policies and System Dynamics

The main idea behind the assignment problem in ridesharing is that most arriving customers have a maximum tolerance (say 7 minutes) for the pickup time or ETA (i.e., expected time of arrival) of a matched supply unit, but are essentially indifferent if the ETA is less than that. Thus when a customer arrives at a node $i$, then any supply unit located at a node which is within 7 minutes of $i$ is a feasible match, while other vehicles which are further away are infeasible. This suggests a natural model for assignment via a bipartite compatibility graph, as depicted in Figure 1.

**Compatibility Graph**: For pedagogical reasons, we move to a setup where we distinguish formally between demand locations $V_D$ where customers arrive and supply locations $V_S$ where supply units wait (and where customers are dropped off). We add a prime symbol to the indices of nodes in $V_D$ to distinguish between the two.

We encode the compatibility graph as a bipartite graph $G(V_S \cup V_D, E)$, wherein each station $i \in V$ is replicated as a supply node $i \in V_S$ and a demand node $i' \in V_D$.

An edge $(i, j') \in E$ represents a compatible pair of supply and demand nodes, i.e., supplies stationed at $i$ can serve demand arriving at $j'$. We denote the neighborhood of a supply node $i \in V_S$ (resp. demand node $j' \in V_D$) in $G$ as $\partial(i)$ (resp. $\partial(j')$); thus, for a supply node $i$, its compatible demands are given by $\partial(i) = \{j' \in V_D | (i, j') \in E\}$, and similarly for each demand node. Moreover, for any set of supply nodes $A \subseteq V_S$, we also use $\partial(A)$ to denote its demand neighborhood (and vice versa).

We make some mild assumptions on arrival rates $\phi$:

**Assumption 1** *The following holds:*

1. *Connectedness: Matrix* $[\phi_{ij}]$ *is irreducible.*[5]
2. *Non-triviality: There exists an origin-destination pair* $i' \in V_D$ *and* $j \in V_S$ *such that* $j \notin \partial(i')$ *and* $\phi_{i'j} > 0$.

**Remark 1** *The irreducibility assumption is useful for charactering the system's steady state behavior under our policies. The second assumption is made to ensure that the assignment control problem at hand is non-trivial. (If it is violated, and if the system starts with at least one supply unit in each location, then we can "reserve" a supply unit for serving each demand origin location* $i' \in V_D$*, and each such car will never leave the corresponding neighborhood* $\partial(i')$*, ensuring that no demand is ever dropped.)*

**Assignment policies**: Given the above setting, the problem we want to study is how to design assignment policies which minimize the probability of dropping demand. For fixed $K$, this problem can be formulated as an average cost Markovian decision process on finite state space, and is thus known to admit an optimal stationary policy (i.e., ones where assignment decisions at time $t$ only depend on the system state $\mathbf{X}^K[t]$; see Proposition 5.1.3 in Bertsekas (1995)). Let $\mathcal{U}^K$ be the set of stationary policies for the $K$-th system.

For each $t \in \mathbb{N}$, $i' \in V_D$, an assignment policy $U \in \mathcal{U}$ includes a series of mappings $U^K \in \mathcal{U}^K$ for each system $K = 1, 2, \cdots$, which maps the current queue-length to $U^K[\mathbf{X}^K[t]](i') \in \partial(i') \cup \{\emptyset\}$. Here $U^K[\mathbf{X}^K[t]](i') = j$ denotes that given the current state $\mathbf{X}^K[t]$, we assign a supply unit from $j \in \partial(i')$ to fulfill demand at $i'$, and $U^K[\mathbf{X}^K[t]](i') = \emptyset$ means that the platform does not assign supplies to $i'$ and hence any arriving demand at $i'$ is dropped. For simplicity of notation, we refer to the policies by $U$ instead of $U^K$.

[5] Replacing non-zero entries in the matrix by one, and viewing the matrix as the adjacency matrix of a directed graph, the matrix is irreducible if and only if this directed graph is strongly connected.

**System Evolution**: We can now formally define the evolution of the Markov chain we want to study. At the beginning of time-slot $t$, the state of the system is $\mathbf{X}^K[t-1]$; note that this incorporates the state-change due to serving the demand in time-slot $t-1$. Now suppose the platform uses an assignment policy $U$, and in time-slot $t$, a customer arrives at origin node $o[t]$ with destination $d[t]$ (chosen from arrival matrix $\phi$). If $U^K[\mathbf{X}^K[t]](o[t]) \neq \emptyset$, then a supply unit from $U^K[\mathbf{X}^K[t]](o[t])$ will pick up the demand and relocate to $d[t]$ instantly. Let $S[t] \triangleq U^K[\mathbf{X}^K[t]](o[t])$ be the chosen supply node (potentially $\emptyset$). Formally, we have

$$\mathbf{X}^K[t] = \begin{cases} \mathbf{X}^K[t-1] - \mathbf{e}_{S[t]} + \mathbf{e}_{d[t]}, & \text{if } S[t] \in V_S. \\ \mathbf{X}^K[t-1], & \text{if } S[t] = \emptyset. \end{cases}$$

### 2.4. Performance Measure

The platform's goal is to find an assignment policy that drops as few demands as possible in steady-state. In this section we formally define the performance measure for all assignment policies. We first introduce some necessary notations.

For each subset $A \subseteq V_S$ (resp. $A' \subseteq V_D$), define:

$$A^c \triangleq \{v \in V_S : v \notin A\} \quad (\text{resp. } (A')^c \triangleq \{v \in V_D : v \notin A'\}). \tag{1}$$

Define the scaled state as $\bar{\mathbf{X}}^K \triangleq \frac{1}{K}\mathbf{X}^K \in \Omega$. Now for a given state $\mathbf{x} \in \Omega$, we define the set of empty stations as

$$I_{\text{em}}(\mathbf{x}) \triangleq \{v \in V_S : \mathbf{x}_v = 0\}.$$

We now have the following simple observation:

**Observation 1** *If the scaled state at time $t$ is $\bar{\mathbf{X}}^K[t] = \mathbf{x}$, then*[6]

$$\mathbb{P}[\textit{demand dropped at } t+1] = \sum_{i':\partial(i') \subseteq I_{\text{em}}(\mathbf{x}), j \in V_S} \phi_{i'j} = \mathbf{1}^T_{\{i':\partial(i') \subseteq I_{\text{em}}(\mathbf{x})\}} \phi \mathbf{1}.$$

This follows from observing that demand arriving in period $t$ to stations in $\{i : \partial(i) \subseteq I_{\text{em}}(\bar{\mathbf{X}}^K[t])\}$ must be dropped.

Given the above setting, a natural performance measure is the *long-run average demand-drop probability*. Formally, for $U \in \mathcal{U}$ we define

$$P_o^{K,U} \triangleq \min_{\mathbf{X}^{K,U}[0] \in \Omega_K} \mathbb{E}\left( \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T} \left[ \mathbf{1}^{\mathrm{T}}_{\{i:\partial(i) \subseteq I_{\text{em}}(\mathbf{X}^{K,U}[t])\}} \phi \mathbf{1} \right] \right), \tag{2}$$

[6] The equality holds for non-idling policies, i.e. policies that don't drop a customer at $i$ whenever there are supplies in $\partial(i)$. For other policies, the equality becomes large than or equal to. This doesn't affect our result because (1)for achievability result, the policies we propose are non-idling; (2)for converse benchmark, the inequality will only make our result stronger.

$$P_p^{K,U} \triangleq \max_{\mathbf{X}^{K,U}[0]\in\Omega_K} \mathbb{E}\left(\lim_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T}\left[\mathbf{1}^{\mathrm{T}}_{\{i:\partial(i)\subseteq I_{\mathrm{em}}(\mathbf{X}^{K,U}[t])\}}\phi\mathbf{1}\right]\right). \tag{3}$$

Here (2) is an optimistic (subscript 'o' for optimistic) performance measure (under-estimates demand-drop probability), (3) is a pessimistic (subscript 'p' for pessimistic) performance measure (over-estimates demand-drop probability). When establishing the optimality of our policy, we will compare its pessimistic measure against other policies' optimistic measure, so we are not 'cheating'. Since $U^K \in \mathcal{U}^K$ is a stationary policy, the limits in (2) and (3) exist. We will drop the expectation sign.

One issue however is that the exact expression of (2) and (3) may be complicated for a fixed $K$. To this end, the main performance measures of interest in this work are the decay rates of $P_o^{K,U}$ and $P_p^{K,U}$ as $K\to\infty$:

$$\gamma_o(U) = -\liminf_{K\to\infty}\frac{1}{K}\log P_o^{K,U}, \tag{4}$$

$$\gamma_p(U) = -\limsup_{K\to\infty}\frac{1}{K}\log P_p^{K,U}. \tag{5}$$

For brevity, we henceforth refer to these as the *demand-drop exponents*. Again the definition is suited to strong converse results, whereas for our positive results the lim sup will, in fact, be the limit, and hence the same as the lim inf.

For any given policy $U$, the demand-drop exponent can be simplified further in terms of the long-run average behavior of $\bar{\mathbf{X}}^{K,U}[t]$.

Now let $D_A \triangleq \{\mathbf{x}\in\Omega : \mathbf{x}_i = 0 \text{ for } i\in A, \mathbf{x}_j > 0 \text{ for } j\notin A\}$, and let $\mathcal{A} = \{A \subsetneq V_S : \exists i' \in V_D \text{ such that } A \supseteq \partial(i')\}$ and let $D \triangleq \cup_{A\in\mathcal{A}} D_A$. In words, $D$ is the set of (normalized) states satisfying the following requirement — that there is at least one location where demand arrives with positive rate which is currently starved of supply at compatible locations. The demand-drop exponent can now be simplified by the following lemma:

**Lemma 1**

$$\gamma_o(U) = -\liminf_{K\to\infty}\frac{1}{K}\log\left(\min_{\mathbf{X}^{K,U}[0]\in\Omega_K}\left(\liminf_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T}\mathbb{1}\left\{\bar{\mathbf{X}}^{K,U}[t]\in D\right\} \text{ a.s.}\right)\right) \tag{6}$$

$$\gamma_p(U) = -\limsup_{K\to\infty}\frac{1}{K}\log\left(\max_{\mathbf{X}^{K,U}[0]\in\Omega_K}\left(\limsup_{T\to\infty}\frac{1}{T}\sum_{t=0}^{T}\mathbb{1}\left\{\bar{\mathbf{X}}^{K,U}[t]\in D\right\} \text{ a.s.}\right)\right) \tag{7}$$

Lemma 1 says that the probability of demand dropping has the same exponential decay rate as the probability that there exists a node without any supply unit at a compatible location. The key idea of the proof is to bound the ratio of the two probabilities by a constant that doesn't scale with $K$.

Finally, the benchmark we use over all policies is:

$$\max_{U \in \mathcal{U}} \left(\gamma_o(U)\right), \tag{8}$$

since no policy can achieve a larger demand-drop exponent.

## 2.5. Sample Path Large Deviation Principle

Our result relies on classical large deviation theory, which we briefly introduce in this subsection.

For each fixed $K \in \mathbb{N}_+$, define $\bar{\mathbf{A}}^K[\cdot] \in (L^\infty[0,T])^{n^2}$ where $\bar{\mathbf{A}}^K[0] = \mathbf{0}$. For $t = 1/K, 2/K, \dots, \lceil KT \rceil / K$,

$$\bar{\mathbf{A}}^K_{ij}[t] \triangleq \frac{1}{K} \sum_{\tau=1}^{Kt} \mathbb{1}\{o[\tau] = i, d[\tau] = j\}.$$

$\bar{\mathbf{A}}^K[t]$ is defined by linear interpolation for other $t$'s. Here $\bar{\mathbf{A}}^K[t]$ is the fluid-scale accumulated demand arrival process of the $K$-th system.

Let $\mu_K$ be the law of $\bar{\mathbf{A}}^K[\cdot]$ in $(L^\infty[0,T])^{n^2}$. Let $\Lambda(\lambda)$ be the cumulant generating function of $\bar{\mathbf{A}} \triangleq \bar{\mathbf{A}}^1[1]$:

$$\Lambda(\lambda) \triangleq \log \mathbb{E} e^{\langle \lambda, \bar{\mathbf{A}} \rangle} = \log \left( \sum_{i=1}^n \sum_{j=1}^n \phi_{ij} e^{\lambda_{ij}} \right) \qquad \lambda \in \mathbb{R}^{n \times n}.$$

Let $\Lambda^*(\mathbf{f})$ be the Legendre-Fenchel transform of $\Lambda(\cdot)$, then

$$\Lambda^*(\mathbf{f}) \triangleq \sup_{\lambda \in \mathbb{R}^n} \langle \lambda, \mathbf{f} \rangle - \Lambda(\lambda) = \begin{cases} D_{KL}(\mathbf{f} || \phi) & \text{if } \mathbf{f} \geq 0, \mathbf{1}^T \mathbf{f} \mathbf{1} = 1 \\ \infty & \text{otherwise.} \end{cases}$$

Here $D_{KL}(\mathbf{f}||\phi)$ is Kullback-Leibler divergence defined as:

$$D_{KL}(\mathbf{f}||\phi) = \sum_{i=1}^n \sum_{j=1}^n \mathbf{f}_{ij} \log \frac{\mathbf{f}_{ij}}{\phi_{ij}}.$$

For any set $\Gamma$, define $\bar{\Gamma}$ as its closure, $\Gamma^o$ as its interior.

Below is the sample path large deviation principle (a.k.a. Mogulskii's theorem) (see Dembo and Zeitouni 1998):

**Fact 1** *For measures $\{\mu_K\}$ defined above, and any arbitrary measurable set $\Gamma \subseteq (L^\infty[0,T])^{n^2}$, we have*

$$-\inf_{\bar{\mathbf{A}} \in \Gamma^o} I_T(\bar{\mathbf{A}}) \leq \liminf_{K \to \infty} \frac{1}{K} \log \mu_K(\Gamma) \leq \limsup_{K \to \infty} \frac{1}{K} \log \mu_K(\Gamma) \leq -\inf_{\bar{\mathbf{A}} \in \bar{\Gamma}} I_T(\bar{\mathbf{A}}),$$

*where the rate function is:*

$$I_T(\bar{\mathbf{A}}) = \begin{cases} \int_0^T \Lambda^*\left(\frac{d}{dt} \bar{\mathbf{A}}(t)\right) dt, & \text{if } \bar{\mathbf{A}}(\cdot) \in \mathrm{AC}[0,T], \bar{\mathbf{A}}(0) = \mathbf{0} \\ \infty, & \text{otherwise.} \end{cases}$$

*Here* $\mathrm{AC}[0,T]$ *is the space of absolutely continuous functions on* $[0,T]$.

**Remark 2** *Since absolutely continuous functions are differentiable almost everywhere, the rate function is well-defined.*

## 3. Scaled MaxWeight Policies

We now introduce the family of scaled MaxWeight (SMW) policies.

The traditional MaxWeight policy (hereafter referred to as vanilla MaxWeight) is a dynamic scheduling rule that allocates the service capacity to the queue(s) with largest "weight" (where weight can be any relevant parameter such as queue-length, sum of queue-lengths, head-of-the-line waiting time, etc.). In our setting, vanilla MaxWeight would correspond to assigning from the compatible location with most supplies (with appropriate tie-breaking rules).

The popularity of MaxWeight scheduling stems from the fact that it is known to be optimal for different metrics in various problem settings (e.g., Stolyar 2003, 2004, Shi et al. 2015, Maguluri and Srikant 2016). However, in our setting, vanilla MaxWeight is suboptimal (and further, we will show that it does not achieve the optimal exponent). The suboptimality of vanilla MaxWeight can be seen from the following simple example.

**Example 1** *Consider a network with two nodes* $\{1,2\}$*, compatibility graph* $G=(V_S\cup V_D,E)=(\{1,2\}\cup\{1',2'\},\{11',12',22'\})$ *and demand arrival rates* $\phi_{1'1}=3/8, \phi_{1'2}=1/8, \phi_{2'1}=\phi_{2'2}=1/4$*, as shown in Figure 2. Suppose at time* $t$ *we have* $\mathbf{X}_1[t]>\mathbf{X}_2[t]$ *and a demand arrives at node* $2'$*.*

*Under vanilla MaxWeight policy, we would assign from node* 1 *since there are more supply units there. However, we claim that vanilla MaxWeight is dominated (in terms of minimizing demand dropping probability) by another policy where one always assign from node* 2 *to serve demand at* $2'$ *as long as* $\mathbf{X}_2>0$*. We call this policy the priority policy. To see this intuitively, note that under both policies demand at node* $2'$ *will never be dropped, hence demand dropping happens if and only if supply at node* 1 *is depleted. The priority policy tries to keep all the servers at node* 1 *while vanilla MaxWeight tries to equal the number of servers on both nodes, hence the priority policy drops less demand. In fact, as we will show formally later, the exponent of demand dropping under the priority policy is twice as large as the exponent under vanilla MaxWeight.*

To deal with this issue, we slightly generalize vanilla MaxWeight by attaching a positive scale-down parameter $w_i$ to each queue $i\in V$, and assign from the compatible queue with largest *weighted* queue length $\mathbf{X}_i/w_i$. Without loss of generality, we normalize $\mathbf{w}$ s.t.

$$\mathbf{1}^T\mathbf{w}=1, \quad \text{or equivalently,} \quad \mathbf{w}\in\mathrm{relint}(\Omega).$$

We call this family of policies *Scaled MaxWeight (*SMW*) policies*, and denote SMW with parameter $\mathbf{w}$ as SMW($\mathbf{w}$). Going back to Example 1, we can approximate the priority policy by attaching a much larger scale-down parameter at node 1 than at node 2.

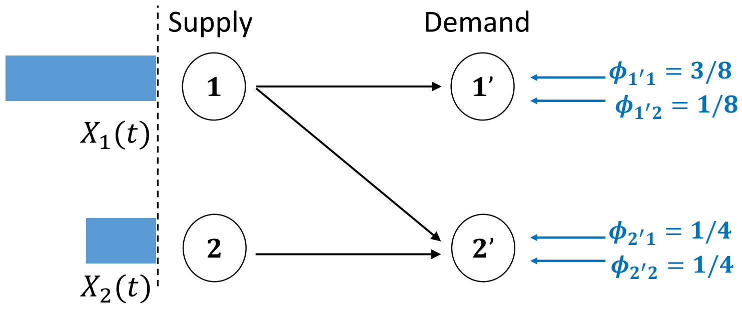


**Figure 2** An example of the sub-optimality of vanilla MaxWeight policy.

The formal definition of SMW is as follows.

**Definition 1 (Scaled MaxWeight)** *Given system state* $\mathbf{X}[t-1]$ *at the start of* $t$*-th period and for demand arriving at* $o[t]$*,* SMW$(\mathbf{w})$ *assigns from*

$$argmax_{i\in\partial(o[t])}\frac{\mathbf{X}_i[t-1]}{w_i}$$

*if* $\max_{i\in\partial(o[t])}\frac{\mathbf{X}_i[t-1]}{w_i}>0$*; otherwise the demand is dropped. If there are ties when determining the argmax, assign from the location with highest index.*

The following fact (which we formalize in Section 5, on the way to proving our main result) gives some intuition about SMW policies.

**Remark 3 (Resting point of state under SMW(w))** *If Assumption 2 holds, the* SMW$(\mathbf{w})$ *policy causes the normalized system state* $\mathbf{X}^K/K$ *to drift towards* $\mathbf{w}$*, where all the scaled queue lengths are equal (to* $1$*).*

## 4. Optimality of Scaled MaxWeight

In this section we present our main result.

### 4.1. Complete Resource Pooling Condition

The following is the main assumption of this paper.

**Assumption 2** *We assume that for any* $J\subsetneq V_D$ *where* $J\neq\emptyset$*,*

$$\sum_{i\in\partial(J)}\mathbf{1}^T\phi_{(i)}>\sum_{j'\in J}\mathbf{1}^T\phi_{j'}. \tag{9}$$

The intuition behind this assumption is clear: it assumes the system is "balanceable" in that for each subset $J \subsetneq V_D$ of demand locations we have enough supply at neighboring locations to meet the demand. Assumption 2 is equivalent to a strict version of the condition in Hall's marriage theorem. It is also closely related to Complete Resource Pooling (CRP) condition in queueing literature (Harrison and López 1999, Stolyar 2004, Ata and Kumar 2005, Gurvich and Whitt 2009). This assumption marks the limit of assignment policies — no assignment policy can achieve exponentially decaying demand dropping probability when Assumption 2 is violated. (if the sign of the inequality is *reversed* in (9) for any subset $J$, then it is easy to see that, in fact, an $\Omega(1)$ fraction of demand must be dropped under any policy).

**Proposition 1** *For any $G$ and $\phi$'s such that Assumption 2 is violated, it holds that for any policy $U$, the demand dropping probability does not decay exponentially, i.e., $\gamma_o(U) = \gamma_p(U) = 0$ where $\gamma_o(U)$ and $\gamma_p(U)$ was defined in* (4) *and* (5). *In fact, if the inequality* (9) *in Assumption 2 is strictly* reversed *for some $J \subsetneq V_D$, then we have $\limsup_{K\to\infty} P_p^{K,U} \geq \liminf_{K\to\infty} P_o^{K,U} \geq \epsilon > 0$ for $\epsilon = \sum_{j'\in J} \mathbf{1}^T \phi_{j'} - \sum_{i\in\partial(J)} \mathbf{1}^T \phi_{(i)}$.*

In other words, if Assumption 2 is violated, this means the system has significant spatial imbalance of demand and stronger forms of control like pricing or repositioning should be employed to restore spatial balance.

### 4.2. Rate Optimality of Scaled MaxWeight

In the fluid limit, it is easy to find an assignment rule such that no demand is dropped (see, e.g., Banerjee et al. 2016). Our goal here is to approach this utopian world as fast as possible as the number of supply units $K$ grows. Define

$$\mathcal{J} \triangleq \left\{ J \subsetneq V_D : \sum_{i'\in J} \sum_{j\notin\partial(J)} \phi_{i'j} > 0 \right\}, \tag{10}$$

we have $\mathcal{J} \neq \emptyset$ by Assumption 1. The following result characterizes the rate at which the fraction of dropped demand falls as $K \to \infty$.

**Theorem 1** *Under Assumptions 1 and 2, the following statements are true:*

1. ***Achievability**: For any $\mathbf{w} \in \mathrm{relint}(\Omega)$, SMW($\mathbf{w}$) achieves exponential decay of the demand dropping probability with exponent*[7,8]

$$\gamma(\mathbf{w}) = \min_{J\in\mathcal{J}} \left(\mathbf{1}^T_{\partial(J)}\mathbf{w}\right) \log\left(\frac{\sum_{i'\notin J}\sum_{j\in\partial(J)}\phi_{i'j}}{\sum_{i'\in J}\sum_{j\notin\partial(J)}\phi_{i'j}}\right) > 0\,. \tag{11}$$

[7] We emphasize that for SMW policies, the lim inf and lim sup in (4) and (5) are limits, and we show that they are equal. Thus, we are not "cheating" on either count.

[8] Note that the argument of the logarithm has a larger numerator than denominator for every $J \subsetneq V_D$ since Assumption 2 holds, implying that $\gamma(\mathbf{w})$ is the minimum of several positive numbers, and hence is positive. (Also see Remark 4.)

2. ***Converse:*** *Under any policy $U$, it must be that*

$$\gamma_p(U) \leq \gamma_o(U) \leq \gamma^*,$$

*where*

$$\gamma^* = \sup_{\mathbf{w} \in \mathrm{relint}(\Omega)} \gamma(\mathbf{w}). \tag{12}$$

*In words, there is an* SMW *policy that achieves an exponent arbitrarily close to the optimal one.*

We will prove claim 1 of the theorem in section 5.3, claim 2 in section 6.1. The first part of the theorem states that for any SMW policy with $\mathbf{w} \in \mathrm{relint}(\Omega)$, the policy achieves an explicitly specified positive exponent $\gamma(\mathbf{w})$ such that the demand dropping probability decreases at the rate $\Theta(e^{-\gamma(\mathbf{w})K})$ as $K \to \infty$. The second part of the theorem provides a universal upper bound $\gamma^*$ on the exponent that any policy can achieve, i.e., for any assignment policy $U$, demand dropping probability must be $\Omega(e^{-\gamma^* K})$. Crucially, $\gamma^*$ is in fact identical to the supremum over $\mathbf{w}$ of $\gamma(\mathbf{w})$. In other words, *there is an (almost) exponent optimal* SMW *policy*, and moreover, this policy can be obtained as the solution to an analytically specified optimization problem.

We note that this result is somewhat different from the numerous results showing optimality of maximum weight matching in various open queuing network settings. Here, our objective is a natural objective that is symmetric in all the queues. Our result says that there is an optimal maximum weight policy, but it is *not* symmetric; rather, it is asymmetric via specific scaling factors in a way that optimally accounts for the primitives of the setting at hand by protecting structurally undersupplied locations.

The following remark provides some intuition regarding the expression for $\gamma(\mathbf{w})$.

**Remark 4 (Intuition for $\gamma(\mathbf{w})$)** *Consider the expression for $\gamma(\mathbf{w})$ in* (11). *It is a minimum of a "robustness" terms for each subset $J \in \mathcal{J}$ of demand locations. For subset $J$, the robustness of* SMW($\mathbf{w}$)*'s ability to serve demand arising in $J$ is the product of two terms:*

- *Robustness arising from $\mathbf{w}$: At the resting point $\mathbf{w}$ (see Remark 3) of* SMW($\mathbf{w}$)*, the supply at neighboring locations is $(\mathbf{1}^T_{\partial(J)}\mathbf{w})$, and the larger that is, the more unlikely it is that the subset will be deprived of supply.*
- *Robustness arising from excess supply: The logarithmic term captures how vulnerable that subset is to being drained of supply. The numerator of the argument of the logarithm can be interpreted as maximum*[9] *average rate at which supply can come in to $\partial(J)$ from $V_S \backslash \partial(J)$, whereas the denominator can be interpreted as the minimum*[9] *average rate at which supply must go from $\partial(J)$ to*

[9] Which arises if the assignment policy uses supply at $\partial(J)$ only to serve demand in $J$, i.e., the policy is maximally "protecting" $J$.

$V_S \backslash \partial(J)$, *unless demand is dropped (the larger the ratio, the more oversupplied and hence robust J is). To see why the term appears in this form, recall that the equilibrium distribution of the length of a stable M/M/1 queue is geometric, and hence has an exponentially decaying tail with the exponent corresponding to* log(service rate/arrival rate). *The "magic" here is that* SMW($\mathbf{w}$) *achieves an exponent such that it suffers no loss from the need to protecting multiple J's simultaneously. (The converse result is somewhat more intuitive, pursuing this reasoning further.)*

Similarly, we also try and give some intuition regarding the optimal choice of the resting state $\mathbf{w}$.

**Remark 5 (Intuition for optimal w)** *To develop some intuition regarding the optimal choice of* $\mathbf{w}$*, consider (informally) the special case of a "heavy traffic" type setting where there is just one subset* $J$ *which has a vanishing logarithmic term (because it is only very slightly oversupplied, the numerator being only slightly more than the denominator), whereas each other subset of* $V_D$ *has a logarithmic term that converges to some positive number. Then the optimal choice of* $\mathbf{w}$ *will satisfy* $\mathbf{1}_{\partial(J)}^T \mathbf{w} \to 1$*, i.e., all but a vanishing fraction of the weight will go to supply locations that can serve* $J$*. The intuition is that the random walk for the supply at* $\partial(J)$ *has only slightly positive drift even if the assignment protects it, and hence it is optimal to keep the total supply at these locations at a high resting point, to minimize the likelihood that the supply at these locations will be depleted. We think an optimal policy for such a special case is itself interesting; what is more remarkable is that the optimal policy characterized in Theorem 1 solves the general* $n$*-dimensional problem considering* all *subsets simultaneously.*

Another feature of our result is the *novel Lyapunov analysis for a closed queueing network.* A key technical challenge we face in our closed queueing network setting is that it is a priori unclear what the "best" state is for the system to be in. This is in contrast to open queueing network settings in which the best state is typically the one in which all queues are empty, and the Lyapunov functions considered typically achieve their minimum at this state. We get around this issue via an innovative approach where we define a tailored Lyapunov function that achieves its minimum at the resting point of the SMW policy we are analyzing, and use this Lyapunov function to characterize its exponent $\gamma(\mathbf{w})$. Moreover, given the optimal choice of $\mathbf{w}$, our tailored Lyapunov function corresponding to this choice of helps us establish our converse result. Our analysis is described in the next section.

## 5. Analysis of Scaled MaxWeight Policies

In this section, we first formally characterize the system behavior under SMW in fluid scale through fluid sample paths (section 5.1) and fluid limits (section 5.2). In section 5.3 we prove achievability bound of SMW policies based on a novel family of Lyapunov functions and large deviation principle (Fact 1) of fluid sample paths. Finally, we provide the performance guarantee of vanilla MaxWeight in section 5.4. Proofs are included in the appendices.

### 5.1. Fluid Sample Paths

Under any stationary assignment policy $U \in \mathcal{U}$ defined in section 2, the system dynamics is as follows:

$$\begin{aligned}\mathbf{X}_i^K[t+1] - \mathbf{X}_i^K[t] = &\sum_{j' \in V_D} \mathbb{1}\{o[t+1]=j', d[t+1]=i\}\mathbb{1}\{U[\mathbf{X}^K[t]](j') \neq \emptyset\} \\ &- \sum_{k' \in \partial(i), j \in V_S} \mathbb{1}\{o[t+1]=k', d[t+1]=j\}\mathbb{1}\{U[\mathbf{X}^K[t]](k')=i\}.\end{aligned} \tag{13}$$

For SMW($\mathbf{w}$), we write the assignment mapping as $U^{\mathbf{w}}[\bar{\mathbf{X}}](i')$ for scaled queue length $\bar{\mathbf{X}} \in \Omega$ and $i' \in V_D$ since it only depends on the scaled state. Let $\mathbf{A}_{j'i}[t]$ be the total number of type $(j', i)$ demands arriving in system during the first $t$ periods ($t = 1, 2, \cdots$), and let

$$\bar{\mathbf{A}}_{j'i}^K[t] \triangleq \frac{1}{K}\mathbf{A}_{j'i}[Kt], \quad \bar{\mathbf{X}}_{j'i}^{K,U}[t] \triangleq \frac{1}{K}\mathbf{X}_{j'i}^{K,U}[Kt] \tag{14}$$

for $t = 0, 1/K, 2/K, \cdots$. For other $t \geq 0$, $\bar{\mathbf{A}}_{j'i}^K[t]$ and $\bar{\mathbf{X}}^{K,U}[t]$ are defined by linear interpolation.

**Definition 2 (Set of demand arrival sample paths)** *Define $\Gamma^{K,T} \subset C^{n^2}[0,T]$ as the set of all scaled demand arrival sample paths of the $K$-th system for $\lfloor TK \rfloor$ periods. Mathematically, any $\bar{\mathbf{A}}[\cdot] \in \Gamma^{K,T}$ satisfies:*

1. $\bar{\mathbf{A}}[0] = \mathbf{0}$.
2. *For any $t = 0, 1/K, 2/K, \cdots, \lfloor TK \rfloor/K$, $\bar{\mathbf{A}}_{i'j}[t] = \bar{\mathbf{A}}_{i'j}[t+1/K]$ for but one pair $(i_0', j_0) = (o[K(t+1)], d[K(t+1)])$. We have $\bar{\mathbf{A}}_{i_0'j_0}[t+1/K] = \bar{\mathbf{A}}_{i_0'j_0}[T] + 1/K$.*
3. *$\bar{\mathbf{A}}[t]$ is defined by linear interpolation for other values of $t$.*

**Definition 3 (Queue-length correspondence of SMW(w))** *For each given demand arrival sample path $\bar{\mathbf{A}}^K[\cdot] \in \Gamma^{K,T}$ and initial state $\bar{\mathbf{X}}^K[0]$, scaled queue length $\bar{\mathbf{X}}^K[\cdot]$ is uniquely (recursively) defined by equation (13)(14). Denote this correspondence by the mapping:*

$$\begin{aligned}\Psi^{\mathbf{w}} : C^{n^2}[0,T] \times \Omega &\to C^n[0,T] \\ (\bar{\mathbf{A}}^K[\cdot], \bar{\mathbf{X}}^K[0]) &\mapsto \bar{\mathbf{X}}^K[\cdot].\end{aligned}$$

To obtain a large deviation result, we need to look at the queue-length process at the fluid scaling. We take the standard approach of *fluid sample paths* (FSP) (see Stolyar 2003, Venkataramanan and Lin 2013).

**Definition 4 (Fluid sample paths)** *We call a pair* $(\bar{\mathbf{A}}[\cdot], \bar{\mathbf{X}}[\cdot])$ *a fluid sample path (under* SMW*(*$\mathbf{w}$*)) if there exists a sequence* $(\bar{\mathbf{A}}^K[\cdot], \bar{\mathbf{X}}^K[0], \Psi^{\mathbf{w}}(\bar{\mathbf{A}}^K[\cdot], \bar{\mathbf{X}}^K[0]))$ *where* $\bar{\mathbf{A}}^K[\cdot] \in \Gamma^{K,T}, \bar{\mathbf{X}}^K[0] \in \Omega$*, such that it has a subsequence which converges to* $(\bar{\mathbf{A}}[\cdot], \bar{\mathbf{X}}[0], \bar{\mathbf{X}}[\cdot])$ *uniformly on* $[0,T]$*.*

**Remark 6** *Since for any such sequence all elements* $(\bar{\mathbf{A}}^K[\cdot], \bar{\mathbf{X}}^K[0], \Psi^{\mathbf{w}}(\bar{\mathbf{A}}^K[\cdot], \bar{\mathbf{X}}^K[0]))$ *are Lipschitz continuous with Lipschitz constant* $1$*, it must have a subsequence that converges uniformly to a limit by Arzelà-Ascoli theorem. Meanwhile, uniform convergence passes the Lipschitz continuity to the limit (see Rudin 1964), hence all FSPs are Lipschitz continuous.*

**Remark 7** *One must take extra care when stating results involving FSP. Although each* $\bar{\mathbf{A}}^K[\cdot]$ *uniquely defines a queue-length sample path* $\bar{\mathbf{X}}^K[\cdot]$*, it is not necessarily true that the* $\bar{\mathbf{A}}[\cdot]$ *component in FSP uniquely defines the* $\bar{\mathbf{X}}[\cdot]$ *component. Theoretically, it is possible that for different* $\bar{\mathbf{X}}_1[\cdot]$ *and* $\bar{\mathbf{X}}_2[\cdot]$*,* $(\bar{\mathbf{A}}[\cdot], \bar{\mathbf{X}}_1[\cdot])$ *and* $(\bar{\mathbf{A}}[\cdot], \bar{\mathbf{X}}_2[\cdot])$ *are both FSPs. Contraction principle (see Dembo and Zeitouni 1998) can rule out such behaviors, but it's technically challenging to prove it for MaxWeight policies (see Subramanian 2010, for its proof under a different setting). As a result, we circumvent this technicality by considering all FSPs in the following proofs.*

In brief, FSPs include both typical and atypical sample paths. Recall Fact 1, which gives the likelihood for any atypical sample path to be realized. The following large deviation analysis in section 5.3 will base on finding the 'most-likely' atypical sample path that leads to demand-drop.

### 5.2. Fluid Limits

The fluid limits of the system characterizes its behavior on the Law of Large Numbers scale, i.e. for typical sample paths. Except for the fact that our queueing network is closed rather than open, the results in this section are similar to Dai and Lin (2005).

For the $K$-th system, denote $\mathbf{A}^K_{j'k}[t]$ as the total number of type $(j',k)$ demands arriving in system during the first $t$ periods, denote the number of times $i$ is assigned to serve type $(j',k)$ demand as $\mathbf{E}^K_{ij'k}[t]$; denote the number of times type $(j',k)$ demand being dropped as $\mathbf{D}^K_{j'k}[t]$. We have the following.

**Proposition 2** *The following holds almost surely. Denote*

$$\bar{\mathbf{X}}^K[t] = \frac{1}{K}\mathbf{X}^K[Kt], \quad \bar{\mathbf{E}}^K[t] = \frac{1}{K}\mathbf{E}^K[Kt], \quad \bar{\mathbf{D}}^K[t] = \frac{1}{K}\mathbf{D}^K[Kt]$$

*for $t = 0, 1/K, 2/K, \cdots$. For other $t \geq 0$, $\bar{\mathbf{X}}^{K,U}[t]$ is defined by linear interpolation. For every sequence of initial conditions $\{\bar{\mathbf{X}}^K[0]\}$ such that $K \to \infty$, there exists a subsequence indexed by $K_r$ such that as $r \to \infty$,*

$$(\bar{\mathbf{X}}^{K_r}[\cdot], \bar{\mathbf{E}}^{K_r}[\cdot], \bar{\mathbf{D}}^{K_r}[\cdot]) \Rightarrow (\bar{\mathbf{X}}[\cdot], \bar{\mathbf{E}}[\cdot], \bar{\mathbf{D}}[\cdot]),$$

*where $\Rightarrow$ is uniform convergence on compact sets on $[0,T]$. All the limits are called fluid limits and are Lipschitz continuous hence almost everywhere differentiable. The differentiable points are called regular points. The fluid limits satisfy the following relations at the regular points:*

$$\begin{aligned}
&\frac{d}{dt}\bar{\mathbf{X}}_i[t] = -\sum_{j' \in \partial(i), k \in V_S} \frac{d}{dt}\bar{\mathbf{E}}_{ij'k}[t] + \sum_{j \in V_S}\sum_{k' \in \partial(j)} \frac{d}{dt}\bar{\mathbf{E}}_{jk'i}[t], \\
&\qquad \forall t \in [0,T], i \in V_S \\
&\bar{\mathbf{X}}_i[t] \geq 0, \quad \forall t \in [0,T], i \in V_S \\
&\sum_i \bar{\mathbf{X}}_i[t] = 1, \quad \forall t \in [0,T] \\
&\sum_{i \in \partial(j')} \frac{d}{dt}\bar{\mathbf{E}}_{ij'k}[t] + \frac{d}{dt}\bar{\mathbf{D}}_{j'k}[t] = \phi_{j'k}, \quad \forall t \in [0,T], j' \in V_D, k \in V_S \\
&\bar{\mathbf{E}}, \bar{\mathbf{D}} \text{ are non-decreasing and } \bar{\mathbf{E}}_{ij'k}(0) = \bar{\mathbf{D}}_{j'k}(0) = 0. \forall j' \in V_D, i, k \in V_S.
\end{aligned}$$

*For non-idling policies, it satisfies additional equations:*

$$\int_0^T \left(\sum_{i \in \partial(j')} \bar{\mathbf{X}}_i(t)\right) \frac{d}{dt}\bar{\mathbf{D}}_{j'k}(t) = 0, \quad \forall j' \in V_D, k \in V_S.$$

The lemma can be proved using an adaptation of proof of Theorem 3.1 in Kumar and Kumar (1996) and Dai and Lin (2005).

### 5.3. Lyapunov Functions for Scaled MaxWeight

Lyapunov functions are useful tool for analyzing complex stochastic systems. In the literature on the MaxWeight policy, the quadratic Lyapunov function is a popular choice to facilitate analysis (Tassiulas and Ephremides 1992, Eryilmaz and Srikant 2012, Maguluri and Srikant 2016, etc.); others have also used piecewise linear Lyapunov functions (Bertsimas et al. 2001, Venkataramanan and Lin 2013, etc.). None of these suffice for our closed queueing network setting, however, and so we need to define a novel family of piecewise linear Lyapunov functions, as follows.

**Definition 5** *For each $\mathbf{w} \in relint(\Omega)$, define Lyapunov function $L_{\mathbf{w}}(\mathbf{x})$ as:*

$$L_{\mathbf{w}}(\mathbf{x}) \triangleq 1 - \min_i \frac{x_i}{w_i}. \tag{15}$$

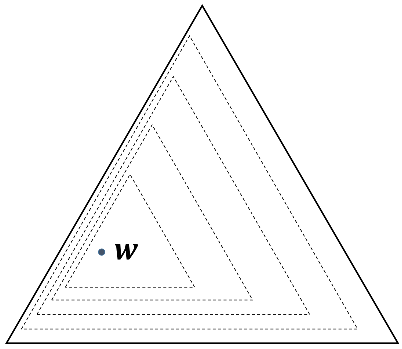


**Figure 3** **Sub-level sets of $L_{\mathbf{w}}$ when $|V_S| = |V_D| = 3$. State space $\Omega$ is a probability simplex in $\mathbb{R}^3$, and its boundary coincides with $\{\mathbf{x} : L_{\mathbf{w}}(\mathbf{x}) \le 1\} \cap h_1$.**

Note that for each $\mathbf{x} \in \Omega$, $L_{\mathbf{w}}(\mathbf{x}) \in [0,1]$. Furthermore, the intersection of sub-level set $\{\mathbf{x} : L_{\mathbf{w}}(\mathbf{x}) \le 1\}$ and hyperplane $h_1 \triangleq \{\mathbf{x} : \mathbf{1}^{\mathrm{T}}\mathbf{x} = 1\}$ is exactly $\Omega$, as is proved in the next lemma. See Figure 3 for illustration.

**Lemma 2 (Sub-level set of Lyapunov functions.)** *For any* $\mathbf{w} \in \mathrm{relint}(\Omega)$*, Lyapunov function* $L_{\mathbf{w}}(\mathbf{x})$ *satisfies:*

$$\{\mathbf{x} : L_{\mathbf{w}}(\mathbf{x}) \le 1\} \cap h_1 = \Omega.$$

To establish system stability using Lyapunov functions, the key step is to show that it has negative drift along fluid limits. In Lemma 3, we explicitly characterize the Lyapunov drift of any fluid sample path under SMW($\mathbf{w}$) when all queue lengths are positive.

**Lemma 3 (Lyapunov drift of fluid sample paths)** *Let* $(\bar{\mathbf{A}}, \bar{\mathbf{X}})$ *be any FSP on* $[0,T]$*,* $\mathbf{w} \in \mathrm{relint}(\Omega)$*. For a regular* $t \in [0,T]$ *such that* $L_{\mathbf{w}}(\bar{\mathbf{X}}[t]) < 1$*, define:*

$$A_1(\bar{\mathbf{X}}[t]) \triangleq \left\{ i \in V_S : i \in \mathrm{argmin} \frac{\bar{\mathbf{X}}_i[t]}{w_i} \right\},$$
$$A_2\left(\bar{\mathbf{X}}[t], \frac{d}{dt}\bar{\mathbf{X}}[t]\right) \triangleq \left\{ i \in A_1(\bar{\mathbf{X}}[t]) : i \in \mathrm{argmin} \frac{1}{w_i}\frac{d}{dt}\bar{\mathbf{X}}_i[t] \right\}.$$

*All the derivatives are well defined since* $t$ *is regular.*

*Denote*

$$c \triangleq \min_{i \in A_1(\bar{\mathbf{X}}[t])} \frac{1}{w_i}\frac{d}{dt}\bar{\mathbf{X}}_i[t].$$

*We have:*

1. $\frac{d}{dt} L_{\mathbf{w}}(\bar{\mathbf{X}}[t]) = -c.$
2. $c = \frac{1}{\mathbf{1}_{A_2}^{\mathrm{T}}\mathbf{w}} \left( \sum_{i' \in V_D, j \in A_2} \frac{d}{dt}\bar{\mathbf{A}}_{i'j}[t] - \sum_{i' : \partial(i') \subseteq A_2, j \in V} \frac{d}{dt}\bar{\mathbf{A}}_{i'j}[t] \right).$

The above lemma is about the Lyapunov drift along fluid sample paths; we now state an analogous lemma bounding the drift for fluid limits. We first need some additional notation.

We denote

$$\lambda_{\min} \triangleq \min_{i \in V_S} \mathbf{1}^T \phi_{(i)} \tag{16}$$

It follows from Assumption 1 that $\lambda_{\min} > 0$. Similarly, we denote

$$\xi \triangleq \min_{J \subsetneq V_D, J \neq \emptyset} \left( \sum_{i \in \partial(J)} \mathbf{1}^T \phi_{(i)} - \sum_{j' \in J} \mathbf{1}^T \phi_{j'} \right), \tag{17}$$

and based on assuming Assumption 2 holds, we have $\xi > 0$.

**Lemma 4 (Lyapunov drift of fluid limits)** *Let $t$ be a regular point, then for any fluid limit $\bar{\mathbf{X}}[\cdot]$:*

$$\frac{d}{dt} L_{\mathbf{w}}(\bar{\mathbf{X}}[t]) \leq -\min\{\xi, \lambda_{\min}\}.$$

Note that the above result holds for any $\bar{\mathbf{X}} \in \Omega$, including when there are empty queues. This is because under SMW policies, the empty queues are replenished 'immediately' in fluid-scale time, see the proof of Lemma 4 for details.

The following lemma is an adaptation of Theorem 5 and Proposition 7 in Venkataramanan and Lin (2013) to our setting. It gives the lower bound of the exponent of $\mathbb{P}(L_{\mathbf{w}}(\bar{\mathbf{X}}^K) \geq \alpha)$ for $\alpha \in (0,1)$, where the system is under SMW($\mathbf{w}$) and in stationarity, given initial state $\mathbf{z} \in \Omega_K$. This event is closely related to demand-drop as $\alpha \to 1$. The key insight from this lemma is that the most likely fluid sample path under SMW($\mathbf{w}$) which leads to the event has a linear structure. See Stolyar and Ramanan (2001), Stolyar (2003) for similar results in a different setting where the achievable exponent is given by a variational problem with piecewise linear solution (called 'simple elements' in Stolyar and Ramanan 2001).

**Lemma 5** *For the system being considered, we have $\forall \alpha \in (0,1)$,*

$$-\limsup_{K \to \infty} \frac{1}{K} \log \left( \max_{\mathbf{z} \in \Omega_K} \mathbb{P}\left( L_{\mathbf{w}}(\bar{\mathbf{X}}_{\mathbf{z}}^{K,\mathrm{SMW}(\mathbf{w})}[\infty]) \geq \alpha \right) \right) \geq \alpha \gamma(\mathbf{w}), \tag{18}$$

*where $\bar{\mathbf{X}}_{\mathbf{z}}^{K,\mathrm{SMW}(\mathbf{w})}[\infty]$ is distributed as the stationary distribution of the $K$-th system under* SMW*($\mathbf{w}$) given initial state $\mathbf{z} \in \Omega_K$. For fixed $T > 0$,*

$$\begin{aligned} \gamma(\mathbf{w}) = & \inf_{v > 0, f, \bar{\mathbf{A}}, \bar{\mathbf{X}}} \frac{1}{v} \Lambda^*(\mathbf{f}), \\ & \textit{s.t. } (\bar{\mathbf{A}}, \bar{\mathbf{X}}) \textit{ is a fluid sample path on } [0,T] \textit{ under } \mathrm{SMW}(\mathbf{w}) \\ & \quad \textit{such that for some regular } t \in [0,T] \\ & \quad \frac{d}{dt}\bar{\mathbf{A}}[t] = \mathbf{f}, \quad L_{\mathbf{w}}(\bar{\mathbf{X}}[t]) = \alpha, \quad \frac{d}{dt} L_{\mathbf{w}}(\bar{\mathbf{X}}[t]) = v. \end{aligned}$$

**Remark 8** *As is discussed in Remark 6, all FSPs* $(\bar{\mathbf{A}}, \bar{\mathbf{X}})$ *are Lipschitz continuous, hence* $L_{\mathbf{w}}(\cdot)$ *is also Lipschitz continuous due to Lemma 10. As a result,* $L_{\mathbf{w}}(\bar{\mathbf{X}}[t])$ *is almost everywhere differentiable w.r.t. t (Rudin 1964).*

The following corollary is a direct consequence of Lemma 5.

**Corollary 1 (Achievability bound of SMW(w))** *Let* $\gamma_p(\mathbf{w})$ *be the pessimistic demand-drop exponent (defined earlier in (5)) under* SMW*(*$\mathbf{w}$*) ,* $\gamma(\mathbf{w})$ *is the value defined in Lemma 5. We have:*

$$\gamma_p(\mathbf{w}) \geq \gamma(\mathbf{w}).$$

The next lemma provides an explicit bound of $\gamma(\mathbf{w})$ (later turns out to be the exact expression of $\gamma(\mathbf{w})$).

**Lemma 6**

$$\gamma(\mathbf{w}) \geq \min_{J \in \mathcal{J}} (\mathbf{1}_{\partial(J)}^{\mathrm{T}} \mathbf{w}) g(\phi, J), \qquad \text{where} \quad g(\phi, J) \triangleq \log \left( \frac{\sum_{i' \notin J} \sum_{j \in \partial(J)} \phi_{i'j}}{\sum_{i' \in J} \sum_{j \notin \partial(J)} \phi_{i'j}} \right)$$

**Proof Idea of Theorem 1 Claim 1:** (See Appendix C.9 for the whole proof.) To prove Theorem 1 Claim 1, it remains to bound $\gamma_o(\mathbf{w})$ from above. For a fix $\mathbf{w} \in \mathrm{relint}(\Omega)$, we first divide the time periods into cycles of length $O(K^2)$. Given any initial state of a cycle, we can show that it goes to a state close to $\mathbf{w}K$ in $O(K)$ by exploiting the negative Lyapunov drift.

Then we explicitly construct a demand arrival sample path that leads to demand drop. Let

$$J^* = \mathrm{argmin}_{J \in \mathcal{J}} (\mathbf{1}_{\partial(J)}^{\mathrm{T}} \mathbf{w}) g(\phi, J),$$

arbitrarily select one if there are multiples minimizers. Define $\mathbf{f}^* \in \mathbb{R}^{n \times n}$ and $T^*$ as follows:

$$\mathbf{f}^*_{j'i} \triangleq \begin{cases} \phi_{j'i} e^{g(\phi, J^*)}, & \text{for } j' \in J^*, i \notin \partial(J^*) \\ \phi_{j'i} e^{-g(\phi, J^*)} & \text{for } j' \notin J^*, i \in \partial(J^*) \\ \phi_{j'i}, & \text{otherwise.} \end{cases},$$

$$T^* \triangleq (\mathbf{1}_{\partial(J^*)}^{\mathrm{T}} \mathbf{w}) \left( \sum_{j' \notin J^*, i \in \partial(J^*)} \phi_{j'i} - \sum_{j' \in J^*, i \notin \partial(J^*)} \phi_{j'i} \right)^{-1}.$$

Roughly speaking, for arbitrary $\delta > 0$, any scaled demand arrival sample path that is within a small enough neighborhood (in terms of sup norm) of $t\mathbf{f}^*(t \in [0, (1+\delta)T^*])$ leads to demand drop. By using Fact 1 and letting $\delta \to 0$, we have:

$$\gamma_o(\mathbf{w}) \leq \min_{J \in \mathcal{J}} (\mathbf{1}_{\partial(J)}^{\mathrm{T}} \mathbf{w}) g(\phi, J).$$

Combined with Corollary 1 and Lemma 6, we conclude the proof.

**Remark 9 (Critical Subset Property)** *In the proof above we explicitly construct a fluid sample path that leads to demand dropping by draining the supply in subset $J^*$. The same proof can be used to show that this is also the most likely sample path that leads to demand drop under any assignment policy given initial normalized state is $\mathbf{w}$. We refer to this as critical subset property: for each normalized state $\mathbf{x} \in \Omega$, there is(are) corresponding critical subset(s):*

$$\mathrm{argmin}_{J \in \mathcal{J}} (\mathbf{1}^{\mathrm{T}}_{\partial(J)} \mathbf{x}) g(\phi, J).$$

### 5.4. Performance of Vanilla MaxWeight

The following theorem shows that SMW with a naive choice of parameter $\mathbf{w} = \frac{1}{n}\mathbf{1}$ achieves an exponent of demand dropping rate that is no smaller than $\frac{1}{n}$ of $\gamma^*$.

**Proposition 3** *Suppose Assumption 2 holds, then we have*

$$\gamma\left(\frac{1}{n}\mathbf{1}\right) \geq \frac{1}{n}\gamma^*.$$

*Main idea:* Simply observe that for each $J \in \mathcal{J}$, the 'buffer supply' vanilla MaxWeight allocates to $\partial(J)$, $\mathbf{1}^{\mathrm{T}}_{\partial(J)}\mathbf{w}$, is at least $\frac{1}{n}$ of that under the optimal SMW.

## 6. Proofs of Converse Bounds

In this section, we will complete the proof of Theorem 1. In addition, we will provide a converse bound for fluid-based policies.

The key observation in converse proofs is that the number of vehicles in a subset of supply locations is upper bounded by a random walk. Under Assumption 2, the best state-dependent policy can provide this random walk with a positive drift, while the best fluid-based policy only leads to zero drift. As a result of this difference, the lower bound on the demand-drop probability decays exponential in $K$ for state-dependent policies, but only polynomially in $K$ for fluid-based policies. Full proofs are included in the appendices.

### 6.1. Proof of Theorem 1 Claim 2: Converse Bound of State-Dependent Policies

We first prove the following lemma that characterizes the probability of a positive drift random walk hitting a negative value with large magnitude.

**Lemma 7** *Let $\{Z_i\}_{i=1}^{\infty}$ be i.i.d. random variables such that $\mathbb{P}(Z_1 = 1) = p$, $\mathbb{P}(Z_1 = -1) = q$, $\mathbb{P}(Z_1 = 0) = 1 - p - q$, where $0 < q < p < 1$, $p + q \leq 1$. Define $S_m \triangleq \sum_{i=1}^{m} Z_i$. Then for any $\epsilon > 0$, there exists $m_0 > 0$ such that for $m > m_0$:*

$$\mathbb{P}(S_m \leq -(p-q)m) \geq e^{-m\epsilon} \left(\frac{p}{q}\right)^{-m(p-q)}.$$

**Remark 10** *Lemma 7 is closely related to the well-known fact that the most-likely path for a stable $M/M/1$ queue to overflow is to reverse the arrival and service rates, see Shwartz and Weiss (1993).*

**Proof Sketch of Theorem 1, Claim 2:** (see Appendix E.2 for the whole proof) Divide the discrete periods into cycles with length $\Theta(K)$. Consider a cycle with initial scaled state $\mathbf{x}$. By Lemma 7, the exponent of the probability that $\partial(J)$ run out of supply in this cycle is $g(\phi, J)$, where

$$g(\phi, J) = \log\left(\frac{\sum_{i'\notin J}\sum_{j\in\partial(J)}\phi_{i'j}}{\sum_{i'\in J}\sum_{j\notin\partial(J)}\phi_{i'j}}\right).$$

Hence the exponent of the probability that some $J \in \mathcal{J}$ gets drained in this cycle is $\gamma(\mathbf{x}) = \min_{J\in\mathcal{J}}(\mathbf{1}^T_{\partial(J)}\mathbf{x})g(\phi, J)$. Taking minimum over all starting state $\mathbf{x}$, the exponent of the probability that demand-drop happens in a cycle is at most $\max_{\mathbf{x}\in\Omega}\gamma(\mathbf{x}) = \gamma^*$. We conclude the proof.

### 6.2. No State-Independent Policy Achieves Exponential Decay

To demonstrate the value of state-dependent controls, we show in the following that under no circumstances can any fluid-based (static) assignment policy lead to exponential decay of demand dropping probability. We first formally define the set of fluid-based policies.

**Definition 6** *We call an assignment policy fluid-based if for any time $t$, $j' \in V_D$, if a demand arrives at $j'$ at $t$, the platform randomly selects an assigned location $i \in \partial(j')$ with probability $u_{(i,j')}[t]$, or drop it with probability $1-\sum_{i\in\partial(j')} u_{(i,j')}[t]$. If there is no supply at the assigned location, the demand is also dropped.*

**Remark 11** *The fluid-based policy defined above includes time-varying policies. It is the same as the randomized policy defined in Ozkan and Ward (2016), definition 2.*

**Proposition 4** *Under any fluid-based assignment policy $\pi$, we have:*

$$P^{K,\pi} = \Omega\left(\frac{1}{K^2}\right).$$

Hence the system converge to fluid limit as $K \to \infty$ much faster under state-dependent controls ($\Theta(e^{-\gamma^*})$) comparing to fluid-based ones ($\Omega(1/K^2)$).

## 7. Conclusions and Future Directions

In this paper we study state-dependent assignment control in a shared transportation system modeled by a closed queueing network. We introduce a family of state-dependent assignment policies called Scaled MaxWeight (SMW) and prove that they have superior performance guarantees in terms of maximizing throughput, comparing with fluid-based policies. In particular, we construct an SMW policy that achieves the optimal large deviation rate of demand-dropping. Our analysis also uncovers the structure of the problem: given system state, demand-dropping is most likely to happen within a state-dependent critical subset of locations. The optimal SMW policy protects all subsets simultaneously. The SMW policies are simple and explicit hence have the potential to influence practice in shared transportation systems.

Before closing, we point out several interesting directions for future research:

1. *Choosing* $\mathbf{w}$ *where* $\phi$ *is imperfectly known.* In practice, a key issue is knowledge of the true arrival rates $\phi$. We remark that if these rates are entirely unknown, but Assumption 2 holds, the platform can pick an SMW policy such as vanilla max weight, and be sure to achieve exponential decay of the demand dropping probability (albeit with a suboptimal exponent). This is already an improvement over the fluid-based/state-independent control policies studied thus far (Ozkan and Ward 2016, Banerjee et al. 2016). If $\phi$ is not precisely known but is known to lie within some set, that setting may lend itself to a natural robust optimization problem of maximizing the demand dropping exponent, worse case over possible $\phi$. We leave this as an open question.

2. *Computational challenges.* A limitation of Theorem 1 is that it does not prescribe how to compute $\mathbf{w}$. It simply specifies a concave maximization problem that must be solved, one where the objective is the minimum over an exponential number of linear functions. We suspect that structural properties of "realistic" $G$ and $\phi$ can be exploited to make this problem tractable; for the moment, we leave it as an open question.

3. *General customer arrival process.* Because our analysis relies mostly on fluid sample paths and large deviations, the result may potentially be generalized to a setting where arrival processes are non-Poisson and spatially correlated. One would expect the specification of arrival process to enter the form of the exponent $\gamma(\mathbf{w})$ through its moment generation function.

4. *Relations to robust queueing.* The large deviation analysis in our paper can be interpreted as a zero-sum game between adversary (nature) and platform: the large deviation rate function is "cost", the adversary tries to force demand dropping while incurring minimum "cost". Another interesting question would be: what's the worse case system performance if the adversary has a finite budget? The latter question is closely related to works on robust queueing.

5. *Bike-sharing systems.* Our model is related to bike-sharing by interpreting the assignment of supply units as suggestion to customers regarding where to pick up a bike among nearby locations. However, bike-sharing systems have some additional features, e.g. limited number of docks. One could consider the flexibility in bike-returning stations, where a natural policy analogous to MaxWeight would suggest returning to a nearby location with most empty docks.

6. *Simulation.* Simulation of closed queueing network instances under proposed policies is ongoing. So far we can conclude from the numerical results that: (1)computation of optimal $\mathbf{w}$ is not an

issue when the number of locations is close to real world situation (when $n \approx 30$, computation takes about 1 hour on PC); (2)with model primitives chosen realistically, the optimal Scaled MaxWeight policy already outperforms vanilla MaxWeight significantly.

## Appendix A: Simplifying the Objective: Proof of Lemma 1

Proof. First, since $\mathbf{1}^{\mathrm{T}}\phi\mathbf{1} = 1$ and $D_{I_{\mathrm{em}}(\bar{\mathbf{X}}^{K,U}[t])} \subseteq D$, we have

$$\mathbf{1}^{\mathrm{T}}_{\{i':\partial(i')\subseteq I_{\mathrm{em}}(\bar{\mathbf{X}}^{K,U}[t])\}}\phi\mathbf{1} \leq \mathbb{1}\left\{\bar{\mathbf{X}}^{K,U}[t] \in D\right\}.$$

On the other hand, whenever $\bar{\mathbf{X}}^{K,U}[t] \in D$, by definition of $D$ there exists $A \subset V_D$ where $\mathbf{1}^{\mathrm{T}}_A\phi\mathbf{1} > 0$ such that

$$\bar{\mathbf{X}}^{K,U}[t] \in D_{\partial(A)},$$

hence

$$\left(\min_{A\subseteq V_D, \mathbf{1}^{\mathrm{T}}_A\phi\mathbf{1}>0} \mathbf{1}^T_A\phi\mathbf{1}\right)\mathbb{1}\left\{\bar{\mathbf{X}}^{K,U}[t] \in D\right\} \leq \mathbf{1}^{\mathrm{T}}_{\{i':\partial(i')\subseteq I_{\mathrm{em}}(\bar{\mathbf{X}}^{K,U}[t])\}}\phi\mathbf{1}.$$

As a result, the ratio between argument of logarithm on both sides of (6) is bounded away from zero and infinity, hence they have the same exponential decay rate. □

## Appendix B: Necessity of CRP condition for Exponential Decay: Proof of Proposition 1

Proof. There are two cases:

1. Some inequality in (9) is strictly reversed, i.e., there exists $J \subsetneq V_D$ s.t. $\sum_{i\in\partial(J)} \mathbf{1}^{\mathrm{T}}\phi_{(i)} < \sum_{j'\in J} \mathbf{1}^{\mathrm{T}}\phi_{j'}$. Consider the following balance equation:

$$\begin{aligned}
\#\{\text{supply in } \partial(J) \text{ at } t\} =& \#\{\text{supply in } \partial(J) \text{ at } 0\} + \#\{\text{supply arrive to } \partial(J) \text{ in } [0,t]\} \\
&- \#\{\text{demand should be served by } \partial(J) \text{ in } [0,t]\} \\
&+ \#\{\text{dropped demand that should be served by } \partial(J) \text{ in } [0,t]\} \\
\leq& \#\{\text{supply in } \partial(J) \text{ at } 0\} + \sum_{s=1}^{t}\mathbb{1}\{d[s]\in\partial(J)\} - \sum_{s=1}^{t}\mathbb{1}\{o[s]\in J\} \\
&+ \#\{\text{dropped demand in } [0,t]\}
\end{aligned}$$

Divide both sides by $K$ and let $K \to \infty$, by SLLN we have:

$$\liminf_{t\to\infty}\{\text{proportion of dropped demand in } [0,t]\} \geq \sum_{j'\in J}\mathbf{1}^{\mathrm{T}}\phi_{j'} - \sum_{i\in\partial(J)}\mathbf{1}^{\mathrm{T}}\phi_{(i)} > 0,$$

hence a positive portion of demand will be dropped, and the second half of the proposition is proved.

2. No inequality in (9) is strictly reversed but there exists $J \in \mathcal{J}$ such that $\sum_{i\in\partial(J)} \mathbf{1}^{\mathrm{T}}\phi_{(i)} = \sum_{j'\in J} \mathbf{1}^{\mathrm{T}}\phi_{j'}$. Divide the discrete time periods into cycles with length $MK^2$, where

$$M \triangleq \frac{6}{\sum_{j'\notin J, i\in\partial(J)}\phi_{j'i}}.$$

Without loss of generality, consider the first cycle $[0, MK^2 - 1]$.

Define random walk $S_J[t]$ with the following dynamics:

- $S_J[0] = \mathbf{1}^{\mathrm{T}}_{\partial(J)} X[0]$.
- $S_J[t+1] = S_J[t] + 1$ if $o[s] \notin J, d[s] \in \partial(J)$.
- $S_J[t+1] = S_J[t] - 1$ if $o[s] \in J, d[s] \notin \partial(J)$.
- $S_J[t+1] = S_J[t]$ if otherwise.

We have:

$$\mathbb{P}\left(\text{some demand is dropped in } [0, MK^2-1]\right) \geq \mathbb{P}\left(S_J[t'] = 0 \text{ for some } t' \in [0, MK^2-1]\right).$$

This is because either (1) some demand is dropped before $t'$, or (2) no demand is dropped before $t'$, then $S_J[t] \geq \mathbf{1}^T_{\partial(J)} \mathbf{X}^{K,U}[t]$ holds for all $t \leq t'$. Either way, the event on RHS implies the event on LHS.
For $J$ of interest, note that $S_J$ is an unbiased random walk, hence a martingale. It's easy to verify that

$$MG_J[t] = S_J^2[t] - 2t \sum_{j' \notin J, i \in \partial(J)} \phi_{j'i}$$

is also a martingale. Apply Optional Stopping Theorem, we have the following result for the first time $S_J$ hit $\{\mathbf{1}^{\mathrm{T}}_{\partial(J)} \mathbf{X}[0] - 2K, \mathbf{1}^{\mathrm{T}}_{\partial(J)} \mathbf{X}[0] + 2K\}$, denoted by $\tau$:

$$\mathbb{E}[\tau] = \frac{2K^2}{\sum_{j' \notin J, i \in \partial(J)} \phi_{j'i}}.$$

Note that $S_J$ hitting $\{\mathbf{1}^{\mathrm{T}}_{\partial(J)} \mathbf{X}[0] - 2K\}$ implies that it hit 0 before, we have

$$\begin{aligned}
&\mathbb{P}\left(S_J[t] = 0 \text{ for some } t' \in [0, MK^2-1]\right) \\
=&1 - \mathbb{P}(\tau > MK^2) - \mathbb{P}(\tau \leq MK^2)\mathbb{P}(\mathbf{X}[\tau] = \mathbf{1}^{\mathrm{T}}_{\partial(J)} \mathbf{X}[0] + 2K | \tau \leq MK^2) \\
\geq&1 - \mathbb{P}(\tau > 3\mathbb{E}[\tau]) - \frac{1}{2} \\
\geq&\frac{1}{6}. \quad \text{(Markov inequality)}
\end{aligned}$$

As a result, for any $U$, we have:

$$P_p^{K,U} \geq P_o^{K,U} = \Omega\left(\frac{1}{K^2}\right),$$

hence $\gamma_p(U) = \gamma_o(U) = 0$ for any $U$.

□

## Appendix C: Achievability Proofs

### C.1. Sub-level Set of Lyapunov Functions: Proof of Lemma 2

Proof.

$$\{\mathbf{x} : L_{\mathbf{w}}(\mathbf{x}) \leq 1, \mathbf{x} \in h_1\} = \left\{\mathbf{x} : \min_{i \in V_S} \frac{x_i}{w_i} \geq 0, \mathbf{1}^T \mathbf{x} = 1\right\} = \left\{\mathbf{x} : \mathbf{x} \geq 0, \mathbf{1}^{\mathrm{T}} \mathbf{x} = 1\right\} = \Omega.$$

□

### C.2. Lyapunov Drift of FSPs: Proof of Lemma 3

Proof. For notation simplicity, we will write $A_1(\bar{\mathbf{X}}[t])$ as $A_1$, $A_2\left(\bar{\mathbf{X}}[t], \frac{d}{dt}\bar{\mathbf{X}}[t]\right)$ as $A_2$ in the following.

1. Denote $b \triangleq \mathrm{argmin}_{i\in V_S} \frac{\bar{\mathbf{X}}_i[t]}{w_i}$. By definition of derivatives, $\forall \delta > 0$, $\exists \epsilon_0 > 0$ such that $\forall \epsilon \in [0, \epsilon_0]$,

$$\frac{\bar{\mathbf{X}}_i(t+\epsilon)}{w_i} \geq b + (c-\delta)\epsilon \quad \forall i \in A_1.$$

Hence

$$\frac{d}{dt}L_{\mathbf{w}}(\bar{\mathbf{X}}[t]) = -\lim_{\epsilon\to 0}\frac{\min_{i\in A_1}\frac{\bar{\mathbf{X}}_i(t+\epsilon)}{w_i} - b}{\epsilon} \leq -\lim_{\epsilon\to 0}\frac{b+(c-\delta)\epsilon - b}{\epsilon} = \delta - c.$$

Since $\delta > 0$ is chosen arbitrarily, we have $\frac{d}{dt}L_{\mathbf{w}}(\bar{\mathbf{X}}[t]) \leq -c$. Similarly, we can show that $\frac{d}{dt}L_{\mathbf{w}}(\bar{\mathbf{X}}[t]) \geq -c$, hence:

$$\frac{d}{dt}L_{\mathbf{w}}(\bar{\mathbf{X}}[t]) = -c.$$

2. Define

$$c' \triangleq \mathrm{argmin}_{i\in A_1\backslash A_2}\frac{1}{w_i}\frac{d}{dt}\bar{\mathbf{X}}_i[t],$$

then $c' > c$. By definition of derivatives, there exists $\epsilon_0 > 0$ such that $\forall \epsilon \in [0, \epsilon_0]$,

$$\frac{\bar{\mathbf{X}}_k(t+\epsilon)}{w_k} \geq b + \left(c' - \frac{c'-c}{3}\right)\epsilon, \quad \forall k \in A_1\backslash A_2 \tag{19}$$

$$\frac{\bar{\mathbf{X}}_l(t+\epsilon)}{w_l} \leq b + \left(c + \frac{c'-c}{3}\right)\epsilon, \quad \forall l \in A_2. \tag{20}$$

Let $(\bar{\mathbf{A}}^K[\cdot], \bar{\mathbf{X}}^K[0], G^{\mathbf{w}}(\bar{\mathbf{A}}^K[\cdot], \bar{\mathbf{X}}^K[0]))$ (where $\bar{\mathbf{A}}^K[\cdot] \in \Gamma^{K,T}$) converges uniformly to a FSP $(\bar{\mathbf{A}}, \bar{\mathbf{X}})$. By definition of uniform convergence, for any $M > 1$ there exists $K_0$ such that for any $K > K_0$ and $t \in [0, T]$, we have $\forall i \in V_S$,

$$\left|\frac{\bar{\mathbf{X}}_i^K[t]}{w_i} - \frac{\bar{\mathbf{X}}_i[t]}{w_i}\right| < \frac{c'-c}{6}\frac{\epsilon_0}{M}. \tag{21}$$

As a result, for any $\epsilon \in \left[\frac{\epsilon_0}{M}, \epsilon_0\right]$ and $K \geq K_0$, let $k \in A_1\backslash A_2$, $l \in A_2$, we have

$$\begin{aligned}
&\frac{\bar{\mathbf{X}}_k^K(t+\epsilon)}{w_k} \\
\geq& \frac{\bar{\mathbf{X}}_k(t+\epsilon)}{w_k} - \frac{c'-c}{6}\frac{\epsilon_0}{M} && \text{(plug in (21))}\\
>& b + \left(c' - \frac{c'-c}{3}\right)\epsilon - \frac{c'-c}{6}\epsilon && \text{(plug in (19))}\\
=& b + \left(c - \frac{c'-c}{2}\right)\epsilon \\
\geq& \frac{\bar{\mathbf{X}}_l(t+\epsilon)}{w_l} + \frac{c'-c}{6}\epsilon && \text{(plug in (20))}\\
\geq& \frac{\bar{\mathbf{X}}_l^K(t+\epsilon)}{w_l} && \text{(plug in (21))}.
\end{aligned}$$

By definition of SMW($\mathbf{w}$), the $K$-th system will not use dispatch within $A_2$ to serve demand outside $\{j' \in V_D : \partial(j') \in A_2\}$ during (scaled) time $[t + \frac{\epsilon_0}{M}, t + \epsilon_0]$. Also, note that since $L_{\mathbf{w}}(\bar{\mathbf{X}}[t]) < 1$, we have $\bar{\mathbf{X}}[t] \in \mathrm{relint}(\Omega)$ by

Lemma 2; by uniform convergence, we know that for large enough $K$ and $s$ close enough to $t$, $U^{\mathbf{w}}(\bar{\mathbf{X}}^K[s],k)$ will never be $\emptyset$. Hence

$$\sum_{i\in A_2}(\bar{\mathbf{X}}_i^K[t+\epsilon_0]-\bar{\mathbf{X}}_i^K[t+\frac{\epsilon_0}{M}]) = -\sum_{j'\in V_D:\partial(j')\subseteq A_2, i\in V_S}(\bar{\mathbf{A}}_{j'i}^K[t+\epsilon_0]-\bar{\mathbf{A}}_{j'i}^K[t+\frac{\epsilon_0}{M}]) + \sum_{j'\in V_D, i\in A_2}(\bar{\mathbf{A}}_{j'i}^K[t+\epsilon_0]-\bar{\mathbf{A}}_{j'i}^K[t+\frac{\epsilon_0}{M}]).$$

Using the Lipschitz property of $\bar{\mathbf{X}}^K[\cdot]$, we know

$$\left|\sum_{i\in A_2}(\bar{\mathbf{X}}_i^K[t+\frac{\epsilon_0}{M}]-\bar{\mathbf{X}}_i^K[t])\right| \le \frac{\epsilon_0}{M}.$$

Combined, we have:

$$\sum_{i\in A_2}(\bar{\mathbf{X}}_i^K[t+\epsilon_0]-\bar{\mathbf{X}}_i^K[t]) \le -\sum_{j'\in V_D:\partial(j')\subseteq A_2, i\in V_S}(\bar{\mathbf{A}}_{j'i}^K[t+\epsilon_0]-\bar{\mathbf{A}}_{j'i}^K[t+\frac{\epsilon_0}{M}]) + \sum_{j'\in V_D, i\in A_2}(\bar{\mathbf{A}}_{j'i}^K[t+\epsilon_0]-\bar{\mathbf{A}}_{j'i}^K[t+\frac{\epsilon_0}{M}]) + \frac{\epsilon_0}{M}.$$

First let $K\to\infty$, then let $M\to\infty$. Using the continuity of $\bar{\mathbf{A}}[\cdot]$, we have:

$$\sum_{k\in A_2}(\bar{\mathbf{X}}_i[t+\epsilon_0]-\bar{\mathbf{X}}_i[t]) \le -\sum_{j'\in V_D:\partial(j')\subseteq A_2, i\in V_S}(\bar{\mathbf{A}}_{j'i}[t+\epsilon_0]-\bar{\mathbf{A}}_{j'i}[t]) + \sum_{j'\in V_D, i\in A_2}(\bar{\mathbf{A}}_{j'i}[t+\epsilon_0]-\bar{\mathbf{A}}_{j'i}[t])$$

Divided by $\epsilon_0$ on both sides and let $\epsilon_0\to 0$. Since $t$ is regular, we have:

$$\sum_{i\in A_2}\frac{d}{dt}\bar{\mathbf{X}}_i[t] \le \left(\sum_{j'\in V_D, i\in A_2}\frac{d}{dt}\bar{\mathbf{A}}_{j'i}[t] - \sum_{j'\in V_D:\partial(j')\subseteq A_2, i\in V_S}\frac{d}{dt}\bar{\mathbf{A}}_{j'i}[t]\right).$$

Similarly, we can show that

$$\sum_{i\in A_2}\frac{d}{dt}\bar{\mathbf{X}}_i[t] \ge \left(\sum_{j'\in V_D, i\in A_2}\frac{d}{dt}\bar{\mathbf{A}}_{j'i}[t] - \sum_{j'\in V_D:\partial(j')\subseteq A_2, i\in V_S}\frac{d}{dt}\bar{\mathbf{A}}_{j'i}[t]\right).$$

Hence

$$\begin{aligned}\frac{d}{dt}L_{\mathbf{w}}(\bar{\mathbf{X}}[t]) &= -\frac{1}{w_i}\frac{d}{dt}\bar{\mathbf{X}}_i[t], \quad \forall i\in A_2\\ &= -\frac{1}{\mathbf{1}_{A_2}^{\mathrm{T}}\mathbf{w}}\sum_{i\in A_2}w_i\frac{1}{w_i}\frac{d}{dt}\bar{\mathbf{X}}_i[t]\\ &= \frac{1}{\mathbf{1}_{A_2}^{\mathrm{T}}\mathbf{w}}\left(\sum_{j'\in V_D:\partial(j')\subseteq A_2, i\in V_S}\frac{d}{dt}\bar{\mathbf{A}}_{j'i}[t] - \sum_{j'\in V_D, i\in A_2}\frac{d}{dt}\bar{\mathbf{A}}_{j'i}[t]\right).\end{aligned}$$

□

### C.3. Lyapunov Drift of Fluid Limits: Proof of Lemma 4

The following lemma says that in the fluid limit, any node will be replenished with supplies 'immediately' once it becomes empty. However, one should be careful when interpreting this result because a fluid-scale time interval with length $\delta$ corresponds to $\lfloor \delta K \rfloor$ time periods in the original system.

**Lemma 8** *Let $\bar{\mathbf{X}}[\cdot]$ be a fluid limit of the system under* SMW(**w**) *policy. Suppose $t=0$ is a regular point (right derivative exists), then for any $i \in V_S$ such that $\bar{\mathbf{X}}_i[0]=0$, we have*

$$\frac{d}{dt}\bar{\mathbf{X}}_i[t]|_{t=0} > 0.$$

Proof. We prove the result by contradiction. The set $V_S$ can be partitioned into three disjoint sets:

$$\begin{aligned}
A_1[t] &= \{i \in V_S : \bar{\mathbf{X}}_i[t] > 0\},\\
A_2[t] &= \{i \in V_S : \bar{\mathbf{X}}_i[t] = 0, \frac{d}{dt}\bar{\mathbf{X}}_i[t] > 0\},\\
A_3[t] &= \{i \in V_S : \bar{\mathbf{X}}_i[t] = 0, \frac{d}{dt}\bar{\mathbf{X}}_i[t] = 0\}.
\end{aligned}$$

Suppose $A_3[t] \neq \emptyset$.

Denote:

$$\begin{aligned}
x^+_{min} &\triangleq \min_{i\in A_1[t]} \frac{\bar{\mathbf{X}}_i[t]}{w_i}, & p^+_{min} &\triangleq \min_{i\in A_1[t]} \frac{d}{dt}\frac{\bar{\mathbf{X}}_i[t]}{w_i},\\
p^0_{max} &\triangleq \max_{i\in A_2[t]} \frac{d}{dt}\frac{\bar{\mathbf{X}}_i[t]}{w_i}, & p^0_{min} &\triangleq \min_{i\in A_2[t]} \frac{d}{dt}\frac{\bar{\mathbf{X}}_i[t]}{w_i}.
\end{aligned}$$

By definition of derivatives, we know that for arbitrary $\delta > 0$, $\exists \epsilon_0 > 0$ such that for any $\epsilon \in [0, \epsilon_0]$, we have:

$$\begin{aligned}
\frac{\bar{\mathbf{X}}_i[t+\epsilon]}{w_i} &\geq \frac{\bar{\mathbf{X}}_i[t]}{w_i} - (p^+_{min} + \delta)\epsilon, && \forall i \in A_1[t]\\
\frac{\bar{\mathbf{X}}_i[t+\epsilon]}{w_i} &\leq (p^0_{max} + \delta)\epsilon, && \forall i \in A_2[t]\\
\frac{\bar{\mathbf{X}}_i[t+\epsilon]}{w_i} &\geq (p^0_{min} - \delta)\epsilon, && \forall i \in A_2[t]\\
\frac{\bar{\mathbf{X}}_i[t+\epsilon]}{w_i} &\leq \delta\epsilon, && \forall i \in A_3[t].
\end{aligned}$$

Choose $\delta = \frac{p^0_{min}}{3}$, $\epsilon_0 \leq \frac{x^+_{min}}{2(p^0_{max} + |p^+_{min}| + 2\delta)}$.

For $M > 1$ and any subsequence $\bar{\mathbf{X}}^{K_r}[\cdot]$ (as in Proposition 2) that uniformly converges to $\bar{\mathbf{X}}[\cdot]$, there exists $K_0 > 0$ such that for any $K_r > K_0$, we have:

$$\sup_{t\in[0,T]} \left| \frac{\bar{\mathbf{X}}_i[t]}{w_i} - \frac{\bar{\mathbf{X}}^{K_r}_i[t]}{w_i} \right| < \frac{\delta\epsilon_0}{6M}, \quad \forall i \in V_S.$$

As a result, for $K_r > N_0$ in the uniformly converging subsequence, the $K_r$-th system satisfies:

$$\frac{\bar{\mathbf{X}}_i[\tau]}{w_i} < \frac{\bar{\mathbf{X}}_j[\tau]}{w_j}, \forall i \in A_3[t], j \in A_1[t] \cup A_2[t]$$

for $\tau \in \{\lfloor K_r(t+\frac{\epsilon_0}{M})\rfloor, \cdots, \lfloor K_r[t+\epsilon_0]\rfloor\}$. Under SMW($\mathbf{w}$), all the demands arriving at $d(A_1[t]\cup A_2[t])$ will be met by supply within $A_1[t]\cup A_2[t]$ during $\{\lfloor K_r(t+\frac{\epsilon_0}{M})\rfloor, \cdots, \lfloor K_r[t+\epsilon_0]\rfloor\}$ for the $K_r$-th system. Moreoever, supplies in $A_1[t]\cup A_2[t]$ will not be depleted during this period. As a result, we have:

$$\begin{aligned}
&\sum_{i\in A_1[t]\cup A_2[t]} \bar{\mathbf{X}}_i(\lfloor K_r[t+\epsilon_0]\rfloor) - \sum_{i\in A_1[t]\cup A_2[t]} \bar{\mathbf{X}}_i(\lfloor K_r t\rfloor) \\
\leq &\lfloor K_r\frac{\epsilon_0}{M}\rfloor + \sum_{\tau=\lfloor K_r(t+\frac{\epsilon_0}{M})\rfloor}^{\lfloor K_r(t+\epsilon_0)\rfloor} \left( \sum_{j'\in V_D, i\in A_1[t]\cup A_2[t]} \mathbb{1}\{o[\tau]=j', d[\tau]=i\} \right. \\
&\left. - \sum_{j'\in\partial(A_1[t]\cup A_2[t]), i\in V_S} \mathbb{1}\{o[\tau]=j', d[\tau]=i\} \right).
\end{aligned}$$

LHS of the above inequality is the net increase of supplies in subset $A_1[t]\cup A_2[t]$, RHS is an upper-bound that let supplies accumulate in $A_1[t]\cup A_2[t]$ as the fastest speed possible (one supply per period) during the first $\frac{1}{M}$ fraction of the considered time interval, and use supplies in $A_1[t]\cup A_2[t]$ to serve all demands in $\partial(A_1[t]\cup A_2[t])$ during the rest of the period. Divide both sides by $K_r$ and let $r\to\infty$. By SLLN we have almost surely,

$$\begin{aligned}
&\sum_{i\in A_1[t]\cup A_2[t]} \bar{\mathbf{X}}_i[t+\epsilon_0] - \sum_{i\in A_1[t]\cup A_2[t]} \bar{\mathbf{X}}_i[t] \\
\leq &\, t\frac{\epsilon_0}{M} + \frac{M-1}{M}\epsilon_0 \left( \sum_{j'\in V_D, i\in A_1[t]\cup A_2[t]} \phi_{j'i} - \sum_{j'\in\partial(A_1[t]\cup A_2[t]), i\in V_S} \phi_{j'i} \right).
\end{aligned}$$

Since we assume that $A_1[t]\cup A_2[t]\subsetneq V$, by Assumption 2 and (17) we have:

$$\sum_{i\in A_1[t]\cup A_2[t]} \bar{\mathbf{X}}_i[t+\epsilon_0] - \sum_{i\in A_1[t]\cup A_2[t]} \bar{\mathbf{X}}_i[t] \leq t\frac{\epsilon_0}{M} - \frac{M-1}{M}\epsilon_0\xi.$$

Since $M$ can be chosen arbitrarily, we choose increasingly large $M$ and have:

$$\sum_{i\in A_1[t]\cup A_2[t]} \bar{\mathbf{X}}_i[t+\epsilon_0] - \sum_{i\in A_1[t]\cup A_2[t]} \bar{\mathbf{X}}_i[t] \leq -\epsilon_0\xi.$$

Since the system is closed, this is equivalent to:

$$\sum_{i\in A_3[t]} \bar{\mathbf{X}}_i[t+\epsilon_0] - \sum_{i\in A_3[t]} \bar{\mathbf{X}}_i[t] \geq \epsilon_0\xi. \tag{22}$$

But we derived above that

$$\frac{\bar{\mathbf{X}}_i[t+\epsilon_0]}{w_i} \leq \frac{\bar{\mathbf{X}}_i[t]}{w_i} + \delta\epsilon_0, \quad \forall i\in A_3[t].$$

Let $\delta < \frac{1}{2}\xi$, then the above implies

$$\sum_{i\in A_3[t]} \bar{\mathbf{X}}_i[t+\epsilon_0] - \sum_{i\in A_3[t]} \bar{\mathbf{X}}_i[t] \leq \frac{1}{2}\left(\sum_{i\in V_S} w_i\right)\epsilon_0\xi = \frac{1}{2}\epsilon_0\xi. \tag{23}$$

(22) and (23) contradicts each other, hence the lemma is proved. □

We now complete the proof of Lemma 4.

**Proof of Lemma 4.** Based on the definition of fluid sample paths and fluid limits, we can see that fluid limits are equivalent to FSP that satisfies:

$$\frac{d}{dt}\bar{\mathbf{A}}_{j'i}[t]=\phi_{j'i},\quad \forall j'\in V_D, i\in V_S, \text{for almost every } t\in[0,T].$$

There are two cases:

- $\bar{\mathbf{X}}[t]>0$. In this case, plug in Lemma 3 and we have

$$\begin{aligned}\frac{d}{dt}L_{\mathbf{w}}(\bar{\mathbf{X}}[t])&=-\frac{1}{\mathbf{1}_{A_2}\mathbf{w}}\left(\sum_{j'\in V_D, i\in A_2}\frac{d}{dt}\bar{\mathbf{A}}_{j'i}[t]-\sum_{j'\in V_D:\partial(j')\subseteq A_2, i\in V_S}\frac{d}{dt}\bar{\mathbf{A}}_{j'i}[t]\right)\\ &\leq-\min_{A_2\subsetneq V_S}\frac{1}{\mathbf{1}_{A_2}\mathbf{w}}\left(\sum_{j'\in V_D, i\in A_2}\phi_{j'i}-\sum_{j'\in V_D:\partial(j')\subseteq A_2, i\in V_S}\phi_{j'i}\right)\\ &\leq-\xi.\end{aligned}$$

- There exists $i\in V$ such that $\bar{\mathbf{X}}_i[t]=0$. Similar to the proof of Lemma 3, we focus on the behavior of the fluid limit during $[t,t+\epsilon_0]$ for small enough $\epsilon_0>0$, and partition it into two time intervals: $[t,t+\frac{\epsilon_0}{M}]$ and $[t+\frac{\epsilon_0}{M},t+\epsilon_0]$ for $M>1$. By Lemma 8, we know that $\frac{d}{dt}\bar{\mathbf{X}}_i[t]>0$ for the nodes who have zero supply at time $t$, hence in the fluid limit the system will have positive supply at each node at $\bar{\mathbf{X}}_i[t+\frac{\epsilon_0}{M}]>0$. We can derive the same Lyapunov drift as in the first case on $[t+\frac{\epsilon_0}{M},t+\epsilon_0]$, and get the desired result by letting $M\to\infty$. The details are omitted here.

□

### C.4. Robustness of Lyapunov Drift

The following lemma shows that we still have negative Lyapunov drift if the arrival process is a small perturbation of fluid limits.

**Lemma 9** *There exists $\epsilon>0$ such that for all fluid sample paths $(\bar{\mathbf{A}},\bar{\mathbf{X}})$ and any $t\in[0,T]$, if $\frac{d}{dt}\bar{\mathbf{A}}[t]\in B(\phi,\epsilon)$, we have*

$$\frac{d}{dt}L_{\mathbf{w}}(\bar{\mathbf{X}}[t])\leq-\frac{1}{2}\min\{\xi,\lambda_{\min}\}.$$

Proof. From the proof of Lemma 4, we can see that the Lyapunov drift depends on the gap $\xi$ defined in (17) and $\lambda_{\min}$ defined in (16). Mathematically, for fluid sample path $(\bar{\mathbf{A}},\bar{\mathbf{X}})$ and regular point $t$, we have:

$$\begin{aligned}\frac{d}{dt}L_{\mathbf{w}}(\bar{\mathbf{X}}[t])&\leq-\min_{A\subsetneq V_S}\left(\sum_{j'\in V_D, i\in A}\frac{d}{dt}\bar{\mathbf{A}}_{j'i}[t]-\sum_{j':\partial(j')\subseteq A, i\in V_S}\frac{d}{dt}\bar{\mathbf{A}}_{j'i}[t]\right)\\ &=-\min\{\xi,\lambda_{\min}\}.\end{aligned}\tag{24}$$

Note that

$$G(\mathbf{f})\triangleq\min_{A\subsetneq V_S}\left(\sum_{j'\in V_D, i\in A}\mathbf{f}_{j'i}-\sum_{j':\partial(j')\subseteq A, i\in V_S}\mathbf{f}_{j'i}\right)$$

is continuous for $\mathbf{f}\in\mathcal{M}\triangleq\{\mathbf{g}\in\mathbb{R}^{n\times n}:\mathbf{g}\geq 0,\mathbf{1}^T\mathbf{g}\mathbf{1}=1\}$. Since $G(\phi)\leq-\min\{\xi,\lambda_{\min}\}<0$, by continuity there must exist $\epsilon$ such that for any $\mathbf{f}\in M$ such that $\mathbf{f}\in B(\phi,\epsilon)$, we have

$$G(\mathbf{f})\leq-\frac{1}{2}\min\{\xi,\lambda_{\min}\}.$$

□

### C.5. Properties of Lyapunov Functions

The Lyapunov functions we defined admit many nice properties. The following lemma is a collection of basic technical facts regarding $L_{\mathbf{w}}(\mathbf{x})$ that are useful for the following proofs. The norm $||\cdot||$ is Euclidean norm if note stated otherwise.

**Lemma 10** *Let $L_{\mathbf{w}}(\mathbf{x})$ be the Lyapunov function defined in (15) ($\mathbf{w}\in\mathrm{relint}(\Omega)$). Note that $L_{\mathbf{w}}(\mathbf{x})$'s domain is $h_1=\{\mathbf{x}:\mathbf{1}^T\mathbf{x}=1\}$. We have:*

1. *$L_{\mathbf{w}}(\mathbf{x})$ is a continuous function of $\mathbf{x}$.*
2. *$L_{\mathbf{w}}(\mathbf{x})\geq 0$ for all $\mathbf{x}\in h_1$, and $L_{\mathbf{w}}(\mathbf{x})=0$ if and only if $\mathbf{x}=\mathbf{w}$.*
3. *There exists $c_{\mathbf{w}}>0, l_{\mathbf{w}}>0, u_{\mathbf{w}}>0$ such that*

$$\min_{\mathbf{x}:||\mathbf{x}-\mathbf{w}||\geq c_{\mathbf{w}},\mathbf{x}\in h_1} L_{\mathbf{w}}(\mathbf{x})\geq l_{\mathbf{w}}, \quad \max_{\mathbf{x}:||\mathbf{x}-\mathbf{w}||\leq c_{\mathbf{w}},\mathbf{x}\in h_1} L_{\mathbf{w}}(\mathbf{x})\leq u_{\mathbf{w}}.$$

4. *$L_{\mathbf{w}}(\mathbf{x})$ is globally Lipschitz on $h_1$, i.e.*

$$|L_{\mathbf{w}}(\mathbf{x}_1)-L_{\mathbf{w}}(\mathbf{x}_2)|\leq \frac{1}{\min_{i\in V_S} w_i}||\mathbf{x}_1-\mathbf{x}_2||.$$

5. *For all fluid sample paths $(\bar{\mathbf{A}},\bar{\mathbf{X}})$, there exists $M>0$ such that:*

$$\frac{d}{dt}L_{\mathbf{w}}(\bar{\mathbf{X}}[t])\leq M.$$

6. *(Centered linear-in-scale property) Assume $c\mathbf{z}+\mathbf{w}$ lies in the domain $h_1$. For any $\mathbf{z}$ such that $\mathbf{1}^{\mathrm{T}}\mathbf{z}=0$ and $c>0$, we have:*

$$L_{\mathbf{w}}(c\mathbf{z}+\mathbf{w})=cL_{\mathbf{w}}(\mathbf{z}+\mathbf{w}).$$

Proof. Property (1)(2)(4)(5)(6) are easy to verify, so we only prove (3) below. For fixed $\mathbf{w}\in\mathrm{relint}(\Omega)$:
(3) Let $c_{\mathbf{w}}\triangleq\min_i w_i$. For $\mathbf{x}\in\{\mathbf{x}:L_{\mathbf{w}}(\mathbf{w})\leq\epsilon\}\cap h_1$, we have:

$$\begin{aligned}||\mathbf{x}-\mathbf{w}|| &\leq\sqrt{n}\cdot\max_{i\in V}|x_i-w_i| \\ &\leq\sqrt{n}\cdot\max_{i\in V}\left(\max\{\epsilon w_i,\epsilon(1-w_i)\}\right).\end{aligned}$$

Hence for $\epsilon_0\triangleq c_{\mathbf{w}}\left(2\sqrt{n}\cdot\max_{i\in V}\left(\max\{w_i,(1-w_i)\}\right)\right)^{-1}>0$, we have $\{\mathbf{x}:L_{\mathbf{w}}(\mathbf{x})\leq\epsilon_0\}\cap h_1\subset\{\mathbf{x}:||\mathbf{x}-\mathbf{w}||<c_{\mathbf{w}}\}\cap h_1$. Let $l_{\mathbf{w}}\triangleq\epsilon_0$, we have

$$\min_{\mathbf{x}:||\mathbf{x}-\mathbf{w}||\geq c_{\mathbf{w}},\mathbf{x}\in h_1} L_{\mathbf{w}}(\mathbf{x})>l_{\mathbf{w}}>0.$$

Also note that for any $\mathbf{x}\in\{\mathbf{x}:||\mathbf{x}-\mathbf{w}||\leq c_{\mathbf{w}}\}\cap h_1$, we have $x_i\geq w_i-c_{\mathbf{w}}\geq 0$, hence

$$\max_{\mathbf{x}:||\mathbf{x}-\mathbf{w}||\leq c_{\mathbf{w}},\mathbf{x}\in h_1} L_{\mathbf{w}}(\mathbf{x})\leq u_{\mathbf{w}}\triangleq 1.$$

□

### C.6. Achievability Bound of SMW Policies: Proof of Lemma 5

For $\mathbf{x} \geq \mathbf{0}$, let $|\mathbf{x}| \triangleq \mathbf{1}^{\mathrm{T}}\mathbf{x}$. Let $\bar{\mathbf{X}}^{\mathbf{x}_l}[\cdot]$ be the scaled process of the $|\mathbf{x}_l|$-th system with initial state $\mathbf{x}_l$. We have the following technical lemma.

**Lemma 11** *For any* $\mathbf{w} \in \mathrm{relint}(\Omega)$*, the following holds:*

$$\lim_{|\mathbf{x}_l| \to \infty} \mathbb{E}\left(L_{\mathbf{w}}\left(\bar{\mathbf{X}}^{\mathbf{x}_l}[\nu]\right)\right) = 0, \qquad \textit{where } \nu = \frac{1}{\min\{\xi, \lambda_{\min}\}}. \tag{25}$$

Proof. The proof is analogous to the proof of Theorem 3.1 in Dai (1995), but there are minor differences because of our closed queueing network setting. Let $\{\mathbf{x}_l\}$ be any sequence of (scaled) initial states such that $|\mathbf{x}_l| \to \infty$, then almost surely there is a subsequence $\{\mathbf{x}_{l_r}\}$ such that $\bar{\mathbf{X}}^{\mathbf{x}_{l_r}}[\cdot]$ converges uniformly on compact sets to a fluid limit $\bar{\mathbf{X}}[\cdot]$ (Proposition 2). We know that any fluid limit $\bar{\mathbf{X}}[\cdot]$ is absolutely continuous, and it satisfies:

$$\frac{d}{dt} L_{\mathbf{w}}(\bar{\mathbf{X}}[t]) \leq -\min\{\xi, \lambda_{\min}\}$$

for $\{\bar{\mathbf{X}}[t] : L_{\mathbf{w}}(\bar{\mathbf{X}}[t]) > 0\}$ at any regular point (Lemma 4). Since $L_{\mathbf{w}}(\mathbf{x}) \leq 1$ for any $\mathbf{x} \in \Omega$ (Lemma 2) we have

$$\lim_{r \to \infty} L_{\mathbf{w}}\left(\bar{\mathbf{X}}^{\mathbf{x}_{l_r}}[\nu]\right) = 0 \quad \text{a.s.}$$

Because of the boundedness of $L_{\mathbf{w}}\left(\bar{\mathbf{X}}^{\mathbf{x}_{l_r}}[\nu]\right)$, the sequence is uniformly integrable hence we have:

$$\lim_{r \to \infty} \mathbb{E}\left(L_{\mathbf{w}}\left(\bar{\mathbf{X}}^{\mathbf{x}_{l_r}}[\nu]\right)\right) = 0.$$

Since the sequence $\{\mathbf{x}_{l_r}\}$ is arbitrary, we have:

$$\lim_{|\mathbf{x}_l| \to \infty} \mathbb{E}\left(L_{\mathbf{w}}\left(\bar{\mathbf{X}}^{\mathbf{x}_l}[\nu]\right)\right) = 0.$$

□

We have the following corollary from Lemma 11.

**Corollary 2** *Fix* $\mathbf{w} \in \mathrm{relint}(\Omega)$*.For any* $\epsilon > 0$*, there exists* $K_0 > 0$ *such that for any* $K > K_0$*, each recurrent class in the* $K$*-th system under* SMW*(*$\mathbf{w}$*) contains at least one element from:*

$$\Pi_{K,\epsilon} \triangleq \{\mathbf{x} \in \Omega_K : L_{\mathbf{w}}(\mathbf{x}/K) \leq \epsilon\}.$$

Proof. By Lemma 11 we have: for any $\epsilon > 0$, there exists $K_0 > 0$ such that for any $K > K_0$, we have

$$\mathbb{E}\left(L_{\mathbf{w}}(\bar{\mathbf{X}}^{\mathbf{x}_l}[\nu])\right) < \frac{\epsilon}{2}. \tag{26}$$

Suppose the corollary doesn't hold, there exists $\mathbf{x}$ where $|\mathbf{x}| > K_0$ and:

$$\mathbb{E}\left(L_{\mathbf{w}}(\bar{\mathbf{X}}^{\mathbf{x}}[t])\right) \geq \epsilon \quad \text{for any } t > 0,$$

which contradicts (26). Hence the corollary must hold.

The above corollary reveals the recurrent class structure of system under SMW($\mathbf{w}$). As a result, we have the following decomposition (see Theorem 6.4.5 in Durrett 2010):

$$\Omega^K = \mathcal{T}^K \cup \left( \cup_{\mathbf{z} \in \Pi_{K,\epsilon}} \mathcal{R}_{\mathbf{z}} \right),$$

where $\mathcal{T}^K$ is the set of transient states, and $\mathcal{R}_{\mathbf{z}}$ is the recurrent class containing at least one state $\mathbf{z}$ in $\Pi_{K,\epsilon}$. Since any possible stationary distribution can be expressed as a convex combination of the unique stationary distribution of each recurrent class (see p.14 in Asmussen 2008), we have

$$\max_{\mathbf{z} \in \Omega_K} \mathbb{P}\left( L_{\mathbf{w}}(\bar{\mathbf{X}}_{\mathbf{z}}^{K,\mathrm{SMW}(\mathbf{w})}[\infty]) \geq \alpha \right) = \max_{\mathbf{z} \in \Pi_{K,\epsilon}} \mathbb{P}\left( L_{\mathbf{w}}(\bar{\mathbf{X}}_{\mathbf{z}}^{K,\mathrm{SMW}(\mathbf{w})}[\infty]) \geq \alpha \right).$$

The following technical lemma bounds the expected time it takes for $L_{\mathbf{w}}(\bar{\mathbf{X}}^K[t])$ to reach any $\delta \in (0,1)$ given starting point $\mathbf{X}^K[0] \in \mathcal{R}_{\mathbf{z}}$ for some $\mathbf{z} \in \Pi_{K,\delta}$.

**Lemma 12** *For $\delta \in (0,1)$, let $\mathbf{X}^K[0] \in \mathcal{R}_{\mathbf{z}}$ for some $\mathbf{z} \in \Pi_{K,\delta}$,*

$$\beta^K \triangleq \frac{\lceil K \inf\{t \geq 0 : L_{\mathbf{w}}(\bar{\mathbf{X}}^K[t]) \leq \delta\} \rceil}{K},$$

*then there exists $K_0 > 0$ such that for any $K > K_0$, we have*

$$\mathbb{E}_{\mathbf{x}}(\beta_1^K) \leq C(\mathbf{w}, \xi, \lambda_{min}) \delta^{-1}$$

*for a finite positive constant $C(\mathbf{w}, \xi, \lambda_{min})$ that doesn't depend on $K$ or initial state $\mathbf{x}$ (as long as $\mathbf{x} \in \mathcal{R}_{\mathbf{z}}$ for some $\mathbf{z} \in \Pi_{K,\delta}$).*

Proof. As a result there exists $K_0 > 0$ such that for any $K > K_0$ we have:

$$\mathbb{E}\left( L_{\mathbf{w}}\left( \bar{\mathbf{X}}^K[\nu] \right) \right) \leq \delta/2.$$

We now consider an arbitrary $K > K_0$ and initial state $\mathbf{x} \in \mathcal{R}_{\mathbf{z}}$ where $\mathbf{z} \in \Pi_{K,\delta}$. For notation simplicity we ignore the integrality of time below. Let

$$S \triangleq \{\mathbf{x} \in \Omega_K : L_{\mathbf{w}}(\mathbf{x}/K) \leq \delta\},$$

then for any $\mathbf{X}_{\mathbf{z}}^K[0] \in \Omega_K \backslash S$,

$$\mathbb{E}\left( L_{\mathbf{w}}(\mathbf{X}_{\mathbf{z}}^K[K\nu]/K) \right) \leq \frac{1}{2} L_{\mathbf{w}}(\mathbf{X}_{\mathbf{z}}^K[0]/K);$$

for $\mathbf{X}_{\mathbf{z}}^K[0] \in S$,

$$\mathbb{E}\left( L_{\mathbf{w}}(\mathbf{X}_{\mathbf{z}}^K[\nu]/K) \right) \leq \frac{1}{2} L_{\mathbf{w}}(\mathbf{X}_{\mathbf{z}}^K[0]/K) + 1$$

because the Lyapunov function is always smaller than 1 for argument in $\Omega$.

Combined, we have:

$$\begin{aligned}
\mathbb{E}\left( K L_{\mathbf{w}}\left( \mathbf{X}_{\mathbf{z}}^K\left[ f(\mathbf{X}_{\mathbf{z}}^K[0]) \right] \right) \right) &\leq \frac{1}{2} K L_{\mathbf{w}}(\mathbf{X}_{\mathbf{z}}^K[0]) + K \mathbb{1}\{\mathbf{X}_{\mathbf{z}}^K[0] \in S\} \\
&\leq K L_{\mathbf{w}}(\mathbf{X}_{\mathbf{z}}^K[0]) - \frac{\delta}{2\nu} f(\mathbf{X}_{\mathbf{z}}^K[0]) + \left( \frac{\delta}{2} + K \right) \mathbb{1}\{\mathbf{X}_{\mathbf{z}}^K[0] \in S\},
\end{aligned}$$

where $f(\mathbf{x}) = \nu K$ if $\mathbf{x} \notin S$, otherwise $f(\mathbf{x}) = \nu$. Since $\mathbf{x} \in \mathcal{R}_{\mathbf{z}}$, the first hitting time of $S$ equals to the first hitting time of $\mathcal{R}_{\mathbf{z}} \cap S$. Because any state in $\mathcal{R}_{\mathbf{z}}$ is positively recurrent, hence $\mathcal{R}_{\mathbf{z}} \cap S$ is a petite set. Proceed exactly the same as in the proof of Theorem 2.1(ii) of Meyn and Tweedie (1994), for each $\mathbf{x} \in \Omega_K$ we have:

$$\mathbb{E}_{\mathbf{x}}(\beta_1^K) \le \frac{2\nu}{\delta} \frac{KL_{\mathbf{w}}(\mathbf{x}/K) + \delta/2 + K}{K} \le \frac{6\nu}{\delta} \triangleq C(\mathbf{w}, \xi, \lambda_{min})\delta^{-1}.$$

The hitting time is divided by $K$ because by definition $\beta_1^K$ is scaled time. Here $C(\mathbf{w}, \xi, \lambda_{min}) \in (0, \infty)$ is constant that doesn't depend on $K$ or $\mathbf{x}$. $\square$

Now we complete the proof of Lemma 5.

**Proof of Lemma 5**. The proof is lengthy, so we divide it into several steps:

*Step 1. Define stopping times and consider the sampling chain.* In this step, we mostly follow the approach in Venkataramanan and Lin (2013) (Freidlin-Wentzell theory) and decompose the desired expression. There are minor differences between our proof and Proof of Theorem 4 in Venkataramanan and Lin (2013) because of our closed queueing network setting, so we will write down each step for completeness.

In the following, let $\mathbf{X}_{\mathbf{z}}^K$ be a random vector distributed as the stationary distribution of recurrent class associated with initial state $\mathbf{z} \in \Omega_K$.

We want to upper bound:

$$\limsup_{K \to \infty} \frac{1}{K} \log \left( \max_{\mathbf{z} \in \Omega_K} \mathbb{P}\left( L_{\mathbf{w}}(\mathbf{X}_{\mathbf{z}}^K / K) \ge \alpha \right) \right).$$

Recall that $\bar{\mathbf{X}}_{\mathbf{z}}^K[t] \triangleq \mathbf{X}_{\mathbf{z}}^K[Kt]/K$ for $t = 0, 1/K, 2/K, \cdots$; for other values of $t$ it's defined via linear interpolation.

Choose positive constants $\delta, \epsilon, \rho$ such that $0 < \delta < \epsilon < \rho < \alpha$. Consider the following stopping times defined on a sample path $\bar{\mathbf{X}}_{\mathbf{z}}^K[\cdot]$:

$$\begin{aligned}
\beta_1^K &\triangleq \frac{\lceil K \inf\{t \ge 0 : L_{\mathbf{w}}(\bar{\mathbf{X}}_{\mathbf{z}}^K[t]) \le \delta\} \rceil}{K}, \\
\eta_i^K &\triangleq \frac{\lceil K \inf\{t \ge \beta_i^K : L_{\mathbf{w}}(\bar{\mathbf{X}}_{\mathbf{z}}^K[t]) \ge \epsilon\} \rceil}{K}, \quad i = 1, 2, \cdots \\
\beta_i^K &\triangleq \frac{\lceil K \inf\{t \ge \eta_{i-1}^K : L_{\mathbf{w}}(\bar{\mathbf{X}}_{\mathbf{z}}^K[t]) \le \delta\} \rceil}{K}, \quad i = 2, 3, \cdots
\end{aligned}$$

Let the Markov chain $\hat{\mathbf{X}}_{\mathbf{z}}^K[i]$ be obtained by sampling $\bar{\mathbf{X}}_{\mathbf{z}}^K[t]$ at the stopping times $\eta_i^K$. Since $\bar{\mathbf{X}}_{\mathbf{z}}^K$ is stationary, there must also exist a stationary distribution for Markov chain $\hat{\mathbf{X}}_{\mathbf{z}}^K[i]$. Let $\Theta_{\mathbf{z}}^K$ denote the state space of the sampled chain, $\hat{\mathbb{P}}_{\mathbf{z}}^K$ is the chain's stationary distribution. The stationary distribution of $\bar{\mathbf{X}}_{\mathbf{z}}^K[\cdot]$ can be expressed as (see Lemma 10.1 in Stolyar 2003)

$$\mathbb{P}\left( L_{\mathbf{w}}(\bar{\mathbf{X}}_{\mathbf{z}}^K) \ge \alpha \right) = \frac{\int_{\Theta_{\mathbf{z}}^K} \hat{\mathbb{P}}_{\mathbf{z}}^K(d\mathbf{x}) \mathbb{E}_{\mathbf{x}} \left( \int_0^{\eta_1^K} \mathbb{I}\left\{ L_{\mathbf{w}}(\bar{\mathbf{X}}_{\mathbf{z}}^K[t]) \ge \alpha \right\} dt \right)}{\int_{\Theta_{\mathbf{z}}^K} \hat{\mathbb{P}}_{\mathbf{z}}^K(d\mathbf{x}) \mathbb{E}_{\mathbf{x}}(\eta_1^K)} \tag{27}$$

Here $\mathbb{E}_{\mathbf{x}}(\cdot)$ denotes the expectation conditioned on the event that $\bar{\mathbf{X}}_{\mathbf{z}}^K[0] = \mathbf{x}$.

As argued above, we are only concerned with the stationary distribution of each recurrent class for each finite system. In the following, let the initial state $\mathbf{z}$ in the $K$-th system belong to $\mathcal{R}_{\mathbf{y}}$ for some $\mathbf{y} \in \Pi_{K,\delta}$.

*Step 2. Bounding the RHS of (27).*

- *Step 2a. Bounding the Denominator.* It's not hard to see that

$$||\mathbf{X}_{\mathbf{z}}^K[t+1]-\mathbf{X}_{\mathbf{z}}^K[t]||_2 \le \sqrt{2}.$$

As a result we have

$$\begin{aligned}||\bar{\mathbf{X}}_{\mathbf{z}}^K[t+1]-\bar{\mathbf{X}}_{\mathbf{z}}^K[t]||_2 &\le \sqrt{2},\\ L_{\mathbf{w}}(\bar{\mathbf{X}}_{\mathbf{z}}^K[\eta_1^K])-L_{\mathbf{w}}(\bar{\mathbf{X}}_{\mathbf{z}}^K[\beta_1^K]) &\le \frac{1}{\min_i\{w_i\}}||\bar{\mathbf{X}}_{\mathbf{z}}^K[\eta_1^K]-\bar{\mathbf{X}}_{\mathbf{z}}^K[\beta_1^K]||_2\\ &\le \frac{\sqrt{2}}{\min_i\{w_i\}}(\eta_1^K-\beta_1^K).\end{aligned}$$

Since $\eta_1^K$ is within $1/K$ (scaled) time unit of the time of $L_{\mathbf{w}}(\bar{\mathbf{X}}_{\mathbf{z}}^K)$ hitting $\epsilon$, we have:

$$L_{\mathbf{w}}(\bar{\mathbf{X}}_{\mathbf{z}}^K[\eta_1^K]) \ge \epsilon - \frac{\sqrt{2}}{\min_i\{w_i\}}\frac{1}{K}.$$

Similarly, we have:

$$L_{\mathbf{w}}(\bar{\mathbf{X}}_{\mathbf{z}}^K[\beta_1^K]) \le \delta + \frac{\sqrt{2}}{\min_i\{w_i\}}\frac{1}{K}.$$

Hence there exists $K_1>0$ such that for any $K>K_1$, the denominator of (27) can be lower bounded by

$$\eta_1^K \ge \frac{\min_i\{w_i\}}{2\sqrt{2}}(\epsilon-\delta). \tag{28}$$

- *Step 2b. Bounding the Numerator.* This part is more complex, and we first decompose the numerator into several terms.

By definition, each $\eta_i^K$ is within $\frac{1}{K}$ (scaled) time unit after $L_{\mathbf{w}}(\bar{\mathbf{X}}_{\mathbf{z}}^K[t])$ just exceeds $\epsilon$. Since $\epsilon<\rho$, using the boundedness of arrival rates and service rates, we can conclude that there exists $K_2>0$ (that depends on $\rho-\epsilon$), such that $L(\bar{\mathbf{X}}_{\mathbf{z}}^K[\eta_i^K]) \le \rho$ for all $N \ge N_2$.

We next define another stopping time:

$$\eta^{K,\uparrow} \triangleq \frac{\lceil \inf\{t\ge 0: L_{\mathbf{w}}(\bar{\mathbf{X}}_{\mathbf{z}}^K[t]) \ge \alpha\}K\rceil}{K}.$$

Then for any $\mathbf{x}\in\Theta_{\mathbf{z}}^K$, we must have:

$$\mathbb{E}_{\mathbf{x}}\left(\int_0^{\eta_1^K}\mathbb{1}\{L_{\mathbf{w}}(\bar{\mathbf{X}}_{\mathbf{z}}^K[t])\ge\alpha\}dt\right) \le \mathbb{E}_{\mathbf{x}}\left(\mathbb{1}\{\eta^{K,\uparrow}\le\beta_1^K\}(\beta_1^K-\eta^{K,\uparrow})\right).$$

The above inequality holds because:

—if $\beta_1^K$ occurs before $\eta^{N,\uparrow}$, then both sides will be zero (because the Lyapunov function will hit $\epsilon$ before $\alpha$);

—if $\beta_1^K$ occurs after $\eta^{N,\uparrow}$, then the amount of time $L_{\mathbf{w}}(\bar{\mathbf{X}}_{\mathbf{z}}^K[t])\ge\alpha$ must be no greater than $\beta_1^K-\eta^{K,\uparrow}$.

Let $\mathbb{P}_{\mathbf{x}}^K$ denote the probability distribution conditioned on $\bar{\mathbf{X}}_{\mathbf{z}}^K[0]=\mathbf{x}$, we then have:

$$\mathbb{E}_{\mathbf{x}}\left(\int_0^{\eta_1^K}\mathbb{1}\{L_{\mathbf{w}}(\bar{\mathbf{X}}_{\mathbf{z}}^K[t])\ge\alpha\}dt\right) \le \mathbb{E}_{\mathbf{x}}\left(\beta_1^K-\eta^{N,\uparrow}\,\middle|\,\eta^{N,\uparrow}\le\beta_1^K\right)\mathbb{P}_{\mathbf{x}}^K\left(\eta^{N,\uparrow}\le\beta_1^K\right).$$

Using the properties of Markov chains and conditional expectation, we have:

$$\mathbb{E}_{\mathbf{x}}\left(\beta_1^K-\eta^{N,\uparrow}\,\middle|\,\eta^{N,\uparrow}\le\beta_1^K\right) = \mathbb{E}_{\mathbf{x}}\left(\mathbb{E}_{\bar{\mathbf{X}}_{\mathbf{z}}^K(\eta^{N,\uparrow})}(\beta_1^K)\,\middle|\,\eta^{N,\uparrow}\le\beta_1^K\right) \tag{29}$$

Let $C \in (\alpha, 1)$, then using the boundedness of arrival rates and service rates again, there exists $K_3 > 0$ such that for $K > K_3$:

$$(29) \leq \sup_{\{\mathbf{y}: L_{\mathbf{w}}(\mathbf{y}) \leq C\}} \mathbb{E}_{\mathbf{y}}(\beta_1^K)$$

Let $T$ be a positive number which will be chosen later. Recall that $L_{\mathbf{w}}(\mathbf{x}) \leq \rho$ for all $\mathbf{x} \in \Theta_{\mathbf{z}}^K$ when $K \geq K_2$. Hence, for any such $\mathbf{x} \in \Theta_{\mathbf{z}}^K$, we have,

$$\begin{aligned}
&\text{numerator of (27)}\\
=&\mathbb{E}_{\mathbf{x}}\left(\beta_1^K - \eta^{N,\uparrow} \,|\eta^{N,\uparrow} \leq \beta_1^K\right)\mathbb{P}_{\mathbf{x}}^K\left(\eta^{N,\uparrow} \leq \beta_1^K\right)\\
\leq&\left(\sup_{\{\mathbf{y}: L_{\mathbf{w}}(\mathbf{y}) \leq C\}} \mathbb{E}_{\mathbf{y}}(\beta_1^K)\right)\left[\mathbb{P}_{\mathbf{x}}^K\left(\eta^{N,\uparrow} \leq T\right) + \mathbb{P}_{\mathbf{x}}^K\left(\beta_1^K \geq T\right)\right]\\
\leq&\underbrace{\left(\sup_{\{\mathbf{y}: L_{\mathbf{w}}(\mathbf{y}) \leq C\}} \mathbb{E}_{\mathbf{y}}(\beta_1^K)\right)}_{\text{(a)}}\left[\underbrace{\sup_{\{\mathbf{x}: L_{\mathbf{w}}(\mathbf{x}) \leq \rho\}} \mathbb{P}_{\mathbf{x}}^K\left(\eta^{N,\uparrow} \leq T\right)}_{\text{(b)}} + \underbrace{\sup_{\{\mathbf{x}: L_{\mathbf{w}}(\mathbf{x}) \leq \rho\}} \mathbb{P}_{\mathbf{x}}^K\left(\beta_1^K \geq T\right)}_{\text{(c)}}\right]
\end{aligned}$$

Note that all the terms are independent of $\mathbf{z}$.

— *Step 2b(i). Bounding term (a).* Apply Lemma 12, there exists $K_4$ such that for $K \geq K_4$, we have

$$(a) \leq C\delta^{-1} \text{ for some } C \in (0, \infty).$$

— *Step 2b(ii). Asymptotics for (b).* For $K \to \infty$, apply Proposition 2 in Venkataramanan and Lin (2013) to $\bar{\mathbf{X}}^K[\cdot]$, which doesn't require $\bar{\mathbf{X}}^K[\cdot]$ to be irreducible. We have:

$$\begin{aligned}
&\limsup_{K\to\infty} \frac{1}{K} \log\left(\sup_{\{\mathbf{x}\in\Omega: L_{\mathbf{w}}(\mathbf{x}) \leq \rho\}} \mathbb{P}_{\mathbf{x}}\left(\eta^{K,\uparrow} \leq T\right)\right)\\
\leq& - \inf_{\bar{\mathbf{A}},\bar{\mathbf{X}}} \int_0^T \Lambda^*\left(\frac{d}{dt}\bar{\mathbf{A}}[t]\right) dt,\\
&\textit{where } \bar{\mathbf{A}}, \bar{\mathbf{X}} \textit{ is any FSP such that } L_{\mathbf{w}}(\bar{\mathbf{X}}[0]) \leq \rho,\ L_{\mathbf{w}}(\bar{\mathbf{X}}[t]) \geq \alpha \textit{ for some } t \in [0, T].
\end{aligned}$$

— *Step 2b(iii). Asymptotics for (c).* Proceed exactly the same as in Proof of Theorem 4, part b(3) in Venkataramanan and Lin (2013), which only uses the properties equivalent to Lemma 4 and Lemma 9 in our paper, we have:

$$\limsup_{K\to\infty} \frac{1}{K} \log\left(\sup_{\{\mathbf{x}: L_{\mathbf{w}}(\mathbf{x}) \leq \rho\}} \mathbb{P}_{\mathbf{x}}\left(\beta_1^K \geq T\right)\right) \leq -\frac{T\min\{\xi, \lambda_{\min}\} + 2(\delta - \rho)}{2/\min_i \mathbf{w}_i + \min\{\xi, \lambda_{\min}\}} J_{\min},$$

where

$$J_{\min} = \min_{\mathbf{f}\notin B(\phi,\epsilon)} \Lambda^*(\mathbf{f})$$

for the $\epsilon$ implicitly specified in Lemma 9. It's not hard to see that $J_{\min} > 0$.

Now combine all the terms and proceed exactly the same as in the proof of Theorem 4, part C in Venkataramanan and Lin (2013), we have:

$$\begin{aligned}
&\limsup_{K\to\infty} \frac{1}{K} \log\left(\max_{\mathbf{z}\in\Omega_K} \mathbb{P}\left(L_{\mathbf{w}}(\mathbf{X}_{\mathbf{z}}^K/K) \geq \alpha\right)\right) \qquad (30)\\
\leq& - \inf_{T>0} \inf_{\bar{\mathbf{A}},\bar{\mathbf{X}}} \int_0^T \Lambda^*\left(\frac{d}{dt}\bar{\mathbf{A}}[t]\right) dt,\\
&\textit{where } \bar{\mathbf{A}}, \bar{\mathbf{X}} \textit{ is any FSP such that } L_{\mathbf{w}}(\bar{\mathbf{X}}[0]) = 0,\ L_{\mathbf{w}}(\bar{\mathbf{X}}[T]) \geq \alpha.
\end{aligned}$$

*Step 3. Reduce (30) to an one-dimensional variation problem.* This part is exactly the same as proof of Theorem 5 in Venkataramanan and Lin (2013), we state the result here and omit the details.

Let

$$\begin{aligned} l_{\mathbf{w},T}(y,v) \triangleq & \inf_{\bar{\mathbf{A}},\bar{\mathbf{X}}} \Lambda^*(\mathbf{f}) \\ & \text{s.t. } (\bar{\mathbf{A}},\bar{\mathbf{X}}) \text{ is an FSP on } [0,T] \text{ such that for some regular } t\in[0,T] \\ & \frac{d}{dt}\bar{\mathbf{A}}[t]=\mathbf{f}, \quad L_{\mathbf{w}}(\bar{\mathbf{X}}[t])=y, \quad \frac{d}{dt}L_{\mathbf{w}}(\bar{\mathbf{X}}[t])=v. \end{aligned}$$

Moreover, define:

$$\begin{aligned} \theta_T \triangleq & \inf_{L[\cdot]} \int_0^T l_{\mathbf{w},T}(L[t],\frac{d}{dt}L[t])dt \\ & \text{s.t. } L[\cdot] \text{ is absolutely continuous and } L[0]=0, \quad L[T]\geq\alpha. \end{aligned}$$

**Claim:**

$$\limsup_{K\to\infty}\frac{1}{K}\log\mathbb{P}\left(\max_{\mathbf{z}\in\Pi_K}\left(L_{\mathbf{w}}(\mathbf{X}_{\mathbf{z}}/K)\geq\alpha\right)\right)\leq-\inf_{T>0}\theta_T. \tag{31}$$

*Step 4. Reduce RHS of (31) to an optimization problem.* It's not hard to see that the RHS of (31) equals to

$$\begin{aligned} -\inf_{T>0}\tilde{\theta}_T, \quad \text{where } \tilde{\theta}_T \triangleq & \inf_{L[\cdot]}\int_0^T l_{\mathbf{w},T}(L[t],\frac{d}{dt}L[t])dt \\ & \text{s.t. } L[\cdot] \text{ is absolutely continuous and } L[0]=0, \quad L[T]=\alpha. \end{aligned}$$

We can provide a lower bound to $l_{\mathbf{w},T}(y,v)$ which is independent of $y$:

$$l_{\mathbf{w},T}(y,v)\geq\tilde{l}_{\mathbf{w}}(v)\triangleq\inf_{y\in(0,\alpha)}l_{\mathbf{w},T}(v)$$

As a result, we have:

$$\begin{aligned} \tilde{\theta}_T \geq & \inf_{L[\cdot]}\int_0^T\tilde{l}_{\mathbf{w}}\left(\frac{d}{dt}L[t]\right)dt \\ & \text{s.t. } L[\cdot] \text{ is absolutely continuous and } L[0]=0, \quad L[T]=\alpha. \end{aligned} \tag{32}$$

Using Lemma C.6 in Shwartz and Weiss (1995), we have the following bound that is independent of $T$:

$$\text{RHS of (32)}\geq\alpha\inf_{m>0}\frac{\tilde{l}_{\mathbf{w}}(m)}{m},$$

where the RHS is exactly $\alpha\gamma(\mathbf{w})$ after plugging in the definition of $\tilde{l}_{\mathbf{w}}$. $\square$

### C.7. Proof of Corollary 1

Proof. Let $\alpha\to1$ in (18), we have:

$$-\limsup_{K\to\infty}\frac{1}{K}\log\left(\max_{\mathbf{z}\in\Omega_K}\mathbb{P}\left(L_{\mathbf{w}}\left(\bar{\mathbf{X}}_{\mathbf{z}}^{K,\mathrm{SMW}(\mathbf{w})}[\infty]\right)\geq1\right)\right)\geq\gamma(\mathbf{w}).$$

By Lemma 2 we have $D\subseteq\{\mathbf{x}:L_{\mathbf{w}}(\mathbf{x})\geq1\}$, hence

$$-\limsup_{K\to\infty}\frac{1}{K}\log\left(\max_{\mathbf{z}\in\Omega_K}\mathbb{P}\left(\bar{\mathbf{X}}_{\mathbf{z}}^{K,\mathrm{SMW}(\mathbf{w})}[\infty]\in D\right)\right)\geq\gamma(\mathbf{w}).$$

Finally, using Lemma 1 we have:

$$\gamma_p(\mathbf{w})=-\limsup_{K\to\infty}\frac{1}{K}\log\left(\max_{\mathbf{z}\in\Omega_K}\mathbb{P}\left(\bar{\mathbf{X}}_{\mathbf{z}}^{K,\mathrm{SMW}(\mathbf{w})}[\infty]\in D\right)\right)\geq\gamma(\mathbf{w}).$$

$\square$

### C.8. Explicit Exponent: Proof of Lemma 6

Proof. Let $t$ be regular and $\mathbf{f} \triangleq \frac{d}{dt}\bar{\mathbf{A}}[t]$. In the following, denote

$$\text{gap}_A(\mathbf{f}) \triangleq \sum_{j':\partial(j')\subseteq A, i\in V_S} \mathbf{f}_{j'i} - \sum_{j'\in V_D, i\in A} \mathbf{f}_{j'i}.$$

Using the result of Lemma 3, we have:

$$\frac{d}{dt}L_{\mathbf{w}}(\bar{\mathbf{X}}[t]) \leq \bar{v}(\mathbf{f}) \triangleq \max_{A\subseteq V_S}\left\{\frac{1}{\mathbf{1}_A^T\mathbf{w}}\text{gap}_A(\mathbf{f})\right\}.$$

Note that in the definition of $\gamma(\mathbf{w})$ in Lemma 5, if we replace any feasible pair $(v, \mathbf{f})$ (where $v > 0$) with $(\bar{v}(\mathbf{f}), \mathbf{f})$, since $v \leq \bar{v}(\mathbf{f})$ and therefore $v > 0 \Rightarrow \bar{v}(\mathbf{f}) > 0$, we have

$$\frac{1}{v}\Lambda^*(\mathbf{f}) \geq \frac{1}{\bar{v}(\mathbf{f})}\Lambda^*(\mathbf{f}).$$

Combine the above observation with Lemma 5, we have

$$\begin{aligned}
\gamma(\mathbf{w}) &\geq \min_{f:\bar{v}(\mathbf{f})>0} \frac{\Lambda^*(\mathbf{f})}{\bar{v}(\mathbf{f})} \\
&\geq \min_{A\subseteq V_S}\left\{\min_{\mathbf{f}:\text{gap}_A(\mathbf{f})>0} \frac{\Lambda^*(\mathbf{f})}{\frac{1}{\mathbf{1}_A^T\mathbf{w}}\text{gap}_A(\mathbf{f})}\right\} \\
&= \min_{J\in\mathcal{J}}\left\{(\mathbf{1}_{\partial(J)}^T\mathbf{w})\left\{\min_{\mathbf{f}:\text{gap}_{\partial(J)}(\mathbf{f})>0} \frac{\Lambda^*(\mathbf{f})}{\text{gap}_{\partial(J)}(\mathbf{f})}\right\}\right\}.
\end{aligned}$$

The last equality holds because for any $A \subseteq V_S$ such that $\{j' \in V_D : \partial(j') \subseteq A, \exists i \notin \partial(j')\ s.t. \phi_{j'i} > 0\} = \emptyset$, we always have $\text{gap}_A(\mathbf{f}) \leq 0$. Denote the optimal value of the inner minimization problem as $g(\phi, J) > 0$, then we have:

$$\min_{\mathbf{f}:\text{gap}_{\partial(J)}(\mathbf{f})>0} \Lambda^*(\mathbf{f}) - g(\phi, J)\left(\sum_{j'\in J, i\in V_S} f_{j'i} - \sum_{j'\in V_D, i\in\partial(J)} f_{j'i}\right) = 0. \tag{33}$$

We can get rid of the constraint on $\mathbf{f}$ since any $\mathbf{f}$ that violates it will not achieve minimum in (33). Then using Legendre transform, we have:

$$\begin{aligned}
&\min_{\mathbf{f}} \Lambda^*(\mathbf{f}) - g(\phi, J)\left(\sum_{j'\in J, i\in V_S} f_{j'i} - \sum_{j'\in V_D, i\in\partial(J)} f_{j'i}\right) \\
&= \min_{\mathbf{f}} \Lambda^*(\mathbf{f}) - \left\langle \mathbf{f}, g(\phi, J)\sum_{j'\in J, i\in V_S} \mathbf{e}_{j'i} - g(\phi, J)\sum_{j'\in V_D, i\in\partial(J)} \mathbf{e}_{j'i}\right\rangle \\
&= -\Lambda\left(g(\phi, J)\sum_{j'\in J, i\in V_S} \mathbf{e}_{j'i} - g(\phi, J)\sum_{j'\in V_D, i\in\partial(J)} \mathbf{e}_{j'i}\right) \\
&= -\log\left(\sum_{j'\in V_D, i\in V_S} \phi_{j'i} e^{g(\phi,J)\mathbb{1}\{j'\in J\} - g(\phi,J)\mathbb{1}\{i\in\partial(J)\}}\right).
\end{aligned}$$

Hence equation (33) reduces to nonlinear equation:

$$\left(\sum_{j'\notin J, i\in\partial(J)} \phi_{j'i}\right)e^{-g(\phi,J)} + \left(\sum_{j'\in J, i\notin\partial(J)} \phi_{j'i}\right)e^{g(\phi,J)} = \sum_{j'\notin J, i\in\partial(J)} \phi_{j'i} + \sum_{j'\in J, i\notin\partial(J)} \phi_{j'i}.$$

Let $y \triangleq e^{g(\phi,J)}$, this becomes a quadratic equation:

$$\left(\sum_{j'\in J, i\notin\partial(J)} \phi_{j'i}\right) y^2 - \left(\sum_{j'\notin J, i\in\partial(J)} \phi_{j'i} + \sum_{j'\in J, i\notin\partial(J)} \phi_{j'i}\right) y + \left(\sum_{j'\notin J, i\in\partial(J)} \phi_{j'i}\right) = 0.$$

Hence

$$y = \frac{\sum_{j'\notin J, i\in\partial(J)} \phi_{j'i}}{\sum_{j'\in J, i\notin\partial(J)} \phi_{j'i}} \text{ or } 1.$$

Since $g(\phi, J) > 0$, we have

$$g(\phi, J) = \log\left(\frac{\sum_{j'\notin J, i\in\partial(J)} \phi_{j'i}}{\sum_{j'\in J, i\notin\partial(J)} \phi_{j'i}}\right).$$

Plug in the original inequality, we have:

$$\gamma(\mathbf{w}) \geq \min_{J\in\mathcal{J}} (\mathbf{1}^T_{\partial(J)}\mathbf{w}) \log\left(\frac{\sum_{j'\notin J, i\in\partial(J)} \phi_{j'i}}{\sum_{j'\in J, i\notin\partial(J)} \phi_{j'i}}\right).$$

□

### C.9. Proof of Theorem 1 Claim 1

Proof. Fix $\mathbf{w} \in \mathrm{relint}(\Omega)$ and consider systems under SMW($\mathbf{w}$). For any $\epsilon \in (0,1)$, define

$$S^K(\epsilon) \triangleq \{\mathbf{x} \in \Omega : L_{\mathbf{w}}(\mathbf{x}) \leq \epsilon\}, \quad \tau^K \triangleq \inf\{t \geq 0 : \mathbf{X}^K[t]/K \in S^K(\epsilon)\}.$$

By Lemma 12, there exists $K_1 > 0$ such that for $K > K_1$ and any $\mathbf{X}^K[0]$ that belongs to some recurrent class in $\Omega_K$, we have:

$$\mathbb{E}(\tau^K) \leq \frac{C}{\epsilon} K.$$

for some universal constant $C > 0$. Consider the $K$-th system where $K > K_1$. Divide the periods into cycles with length $2K^2$. We study the lower bound of demand-drop probability on each cycle. Without loss of generality, consider the first cycle. We have:

$$\begin{aligned}
&\mathbb{P}(\text{demand drop at } t \text{ for some } t \in [1, 2K^2]) && (34)\\
\geq &\mathbb{P}(\tau^K \leq K^2, \text{demand drop at } t \text{ for some } t \in [\tau^K, \tau^K + K^2])\\
= &\mathbb{P}(\tau^K \leq K^2)\mathbb{P}(\text{demand drop at } t \text{ for some } t \in [0, \tau^K + K^2] | \mathbf{X}^K[0]/K \in S^K(\epsilon))\\
\geq &\left(1 - \frac{C}{\epsilon K}\right)\mathbb{P}(\text{demand drop at } t \text{ for some } t \in [0, K^2] | \mathbf{X}^K[0]/K \in S^K(\epsilon)).
\end{aligned}$$

Here the equality follows from strong Markov property, the last inequality follows from Markov inequality.

Now we construct a set of sample paths that will lead to demand drop in $[0, K^2]$. Let

$$J^* = \mathrm{argmin}_{J\in\mathcal{J}} (\mathbf{1}^{\mathrm{T}}_{\partial(J)}\mathbf{w}) g(\phi, J),$$

arbitrarily select one if there are multiples minimizers. Define $\mathbf{f}^* \in \mathbb{R}^{n\times n}$ and $T^*$ as follows:

$$\mathbf{f}^*_{j'i} \triangleq \begin{cases} \phi_{j'i} e^{g(\phi,J^*)}, & \text{for } j' \in J^*, i \notin \partial(J^*) \\ \phi_{j'i} e^{-g(\phi,J^*)} & \text{for } j' \notin J^*, i \in \partial(J^*) \\ \phi_{j'i}, & \text{otherwise.} \end{cases},$$

$$T^* \triangleq (\mathbf{1}^{\mathrm{T}}_{\partial(J^*)}\mathbf{w}) \left(\sum_{j'\notin J^*, i\in\partial(J^*)} \phi_{j'i} - \sum_{j'\in J^*, i\notin\partial(J^*)} \phi_{j'i}\right)^{-1}.$$

For $\delta \triangleq \frac{4n\epsilon}{\mathbf{1}^{\mathrm{T}}_{\partial(J^*)}\mathbf{w}}$, define a neighborhood $B_\delta \subseteq C^{n^2}[0,(1+\delta)T^*]$:

$$B_\delta \triangleq \left\{ \mathbf{f}[\cdot] \in C^{n^2}[0,(1+\delta)T^*] : \sup_{0\le t\le(1+\delta)T^*} ||\mathbf{f}[t] - t\mathbf{f}^*||_\infty \le \frac{\epsilon}{n} \right\},$$

Recall that $\Gamma^{K,T}$ is the set of demand arrival sample paths from Definition 2. It's not hard to verify that there exists $K_\delta > 0$ such that for $K > K_\delta$, $B_\delta \cap \Gamma^{K,(1+\delta)T^*}$ is non-empty; we omit the details here.

Now we carefully count the number of supplies within set $\partial(J^*)$ at time $\lfloor K(1+\delta)T^* \rfloor$ given the initial state satisfies $\mathbf{X}^K[0]/K \in S^K(\epsilon)$, and the scaled demand arrival process falls in $B_\delta$. It's easy to verify that given $\mathbf{X}^K[0]/K \in S^K(\epsilon)$, we have $\max_i \mathbf{X}^K_i[0]/K \le \mathbf{w}_i + 2\epsilon$. For the process, either there are already demand-drops before $\lfloor K(1+\delta)T^* \rfloor$, or we must have:

$$\begin{aligned}
&\mathbf{1}^{\mathrm{T}}_{\partial(J^*)}\mathbf{X}^K[K(1+\delta)T^*]\\
\le&\mathbf{1}^{\mathrm{T}}_{\partial(J^*)}\mathbf{X}^K[0] + K \max_{\mathbf{f}^K[\cdot]\in B_\delta\cap\Gamma^{K,(1+\delta)T^*}} \left( \sum_{j'\notin J^*, i\in\partial(J^*)} \mathbf{f}^K_{j'i}[(1+\delta)T^*] - \sum_{j'\in J^*, i\notin\partial(J^*)} \mathbf{f}^K_{j'i}[(1+\delta)T^*] \right)\\
\le&2n\epsilon K + (\mathbf{1}^{\mathrm{T}}_{\partial(J^*)}\mathbf{w})K + K(1+\delta)T^* \left( \sum_{j'\notin J^*, i\in\partial(J^*)} \mathbf{f}^*_{j'i} - \sum_{j'\in J^*, i\notin\partial(J^*)} \mathbf{f}^*_{j'i} \right) + \frac{\epsilon}{n} n^2 K\\
=&2n\epsilon K - \delta(\mathbf{1}^{\mathrm{T}}_{\partial(J^*)}\mathbf{w})K + n\epsilon K\\
=&-n\epsilon K\\
<&0,
\end{aligned}$$

which cannot happen. As a result, we have:

$$(34) \ge \left(1 - \frac{C}{\epsilon K}\right) \mathbb{P}(\mathbf{f}^K[\cdot] \in B_\delta).$$

Let $K \to \infty$ and consider the exponents on both sides, we have:

$$\begin{aligned}
\gamma_o(\mathbf{w}) &\le -\liminf_{K\to\infty} \frac{1}{K} \log \mathbb{P}(\mathbf{f}^K[\cdot] \in B_\delta) \le \inf_{\mathbf{f}\in B^o_\delta} \int_0^{(1+\delta)T^*} \Lambda^* \left( \frac{d}{dt}\mathbf{f}[t] \right) dt\\
&\le \int_0^{(1+\delta)T^*} \Lambda^* \left( \frac{d}{dt}\mathbf{f}^*[t] \right) dt = (1+\delta)T^*\Lambda^*(\mathbf{f}^*) = (1+\delta)(\mathbf{1}^{\mathrm{T}}_{\partial(J^*)}\mathbf{w})g(\phi, J^*),
\end{aligned}$$

where the first inequality follows from Fact 1. Since $\delta$ can be chosen arbitrarily, we have

$$\gamma_o(\mathbf{w}) \le (\mathbf{1}^{\mathrm{T}}_{\partial(J^*)}\mathbf{w})g(\phi, J^*) = \min_{J\in\mathcal{J}} (\mathbf{1}^{\mathrm{T}}_{\partial(J)}\mathbf{w})g(\phi, J).$$

Combined with Corollary 1 and Lemma 6, we have:

$$\gamma_o(\mathbf{w}) = \gamma(\mathbf{w}) = \gamma_p(\mathbf{w}),$$

which concludes the proof. $\square$

## Appendix D: Performance of Vanilla MW: Proof of Proposition 3

Proof. We have

$$\text{LHS} = \min_{J\in\mathcal{J}} \frac{|\partial(J)|}{n} \log\left(\frac{\sum_{j'\notin J, i\in\partial(J)} \phi_{j'i}}{\sum_{j'\in J, i\notin\partial(J)} \phi_{j'i}}\right), \tag{35}$$

$$\gamma^* = \max_{\mathbf{w}\in\Omega} \min_{J\in\mathcal{J}} (\mathbf{1}_{\partial(J)}^{\mathrm{T}}\mathbf{w}) \log\left(\frac{\sum_{j'\notin J, i\in\partial(J)} \phi_{j'i}}{\sum_{j'\in J, i\notin\partial(J)} \phi_{j'i}}\right). \tag{36}$$

Let $J^* \subsetneq V_D$ be one of the minimizer of (35), $\mathbf{w}^*$ is the maximizer of (36). Hence

$$\begin{aligned}
\gamma^* =& \min_{J\in\mathcal{J}} (\mathbf{1}_{\partial(J)}^{\mathrm{T}}\mathbf{w}^*) \log\left(\frac{\sum_{j'\notin J, i\in\partial(J)} \phi_{j'i}}{\sum_{j'\in J, i\notin\partial(J)} \phi_{j'i}}\right)\\
\le& (\mathbf{1}_{\partial(J^*)}^{\mathrm{T}}\mathbf{w}^*) \log\left(\frac{\sum_{j'\notin J^*, i\in\partial(J^*)} \phi_{j'i}}{\sum_{j'\in J^*, i\notin\partial(J^*)} \phi_{j'i}}\right)\\
\le& |\partial(J^*)| \log\left(\frac{\sum_{j'\notin J^*, i\in\partial(J^*)} \phi_{j'i}}{\sum_{j'\in J^*, i\notin\partial(J^*)} \phi_{j'i}}\right)\\
=& \gamma\left(\frac{1}{n}\mathbf{1}\right).
\end{aligned}$$

□

## Appendix E: Converse Proofs

### E.1. Large Deviation of a Random Walk: Proof of Lemma 7

Proof. Apply Cramér's Theorem (see Theorem 2.2.3 in Dembo and Zeitouni 1998), we have:

$$\lim_{m\to\infty} \frac{1}{m} \log \mathbb{P}(S_m \le -(p-q)m) = -\Lambda_1^*(q-p),$$

where

$$\Lambda_1^*(q-p) = \sup_{x\in\mathbb{R}} \left\{(q-p)x - \log \mathbb{E}e^{xZ_1}\right\} = (p-q)\log\left(\frac{p}{q}\right).$$

As a result, for any $\epsilon > 0$ there exists $m_0 > 0$ such that for any $m > m_0$:

$$\frac{1}{m} \log \mathbb{P}(S_m \le -(p-q)m) \ge -\Lambda_1^*(q-p) - \epsilon.$$

□

### E.2. Proof of Theorem 1 Claim 2

Proof. For any $J \in \mathcal{J}$, define:

$$B_J \triangleq \left\{\mathbf{x}\in\Omega : (\mathbf{1}_{\partial(J)}^T\mathbf{x}) g(\phi, J) = \operatorname{argmin}_{A\in\mathcal{J}} (\mathbf{1}_{\partial(A)}^T\mathbf{x}) g(\phi, A)\right\}.$$

To make $\{B_J\}_{J\in\mathcal{J}}$ a non-overlapping partition, index $J$ by natural numbers and only keep $\mathbf{x}$'s membership in the set with smallest index $J$. Let

$$z_J \triangleq \sup_{\mathbf{x}\in B_J} \mathbf{1}_{\partial(J)}^{\mathrm{T}}\mathbf{x}.$$

Recall the definition $\xi$ in (17), define

$$\xi_J \triangleq \sum_{i'\notin J} \sum_{j\in\partial(J)} \phi_{i'j} - \sum_{i'\in J} \sum_{j\notin\partial(J)} \phi_{i'j}.$$

By Assumption 2, we have:

$$\xi_{\min} \triangleq \min_{J\in\mathcal{J}} \xi_J > 0.$$

For the $K$-th system, we divide the discrete periods into cycles with length $K/\xi$ (ignoring the technicality regarding the integrality of $K/\xi$). We are going to lower bound the probability of demand dropping in any cycle for any state-dependent dispatch policy $U$, even including time-varying policies. Without loss of generality, consider the first cycle $[0, K/\xi - 1]$. Suppose $\bar{\mathbf{X}}^{K,U}[0] = \mathbf{y} \in B_J$.

Consider the random walk $S_J[t]$ with the following dynamics:

- $S_J[0] = \mathbf{1}^{\mathrm{T}}_{\partial(J)}\mathbf{y}$.
- $S_J[t+1] = S_J[t] + 1$ if $o[s] \notin J, d[s] \in \partial(J)$.
- $S_J[t+1] = S_J[t] - 1$ if $o[s] \in J, d[s] \notin \partial(J)$.
- $S_J[t+1] = S_J[t]$ if otherwise.

Observe that:

$$\mathbb{P}\left(\text{some is was dropped in } [0, K/\xi - 1]\right) \geq \mathbb{P}\left(S_J[t] = 0 \text{ for some } t' \in [0, K/\xi]\right). \tag{37}$$

This is because either (1) some demand is dropped before $t'$, or (2) no demand is dropped before $t'$, then $S_J[t] \geq \mathbf{1}^T_{\partial(J)}\bar{\mathbf{X}}^{K,U}[t]$ holds for all $t \leq t'$; $S_J(t') = 0$ implies that there is no supply in $\partial(J)$ at $t'$. Either way, the event on RHS implies the event on LHS.

Fix $\epsilon > 0$. For any $J \in \mathcal{J}$, by Lemma 7 there exists $K_{J>0}$ such that for $K > K_J$ (ignore the integrality of arguments below):

$$\mathbb{P}(S_J[z_J K/\xi_J] \leq -z_J K) \geq e^{-z_J K\epsilon/\xi_J} e^{-g(\phi,J)z_J K}$$

Then for $K > \max_{J\in\mathcal{J}} K_J$, we have:

$$\begin{aligned}
(37) &= \sum_{J\in\mathcal{J}} \mathbb{P}\left(S_J[t] - S_J[0] = -K(\mathbf{1}^T_{\partial(J)}\mathbf{y}) \text{ for some } t' \in [0, K/\xi] | \mathbf{y} \in B_J\right) \mathbb{P}(\mathbf{y} \in B_J) \\
&\geq \min_{J\in\mathcal{J}} \mathbb{P}\left(S_J[z_J K/\xi_J] - S_J[0] \leq -K(\mathbf{1}^T_{\partial(J)}\mathbf{y})\right) \\
&\geq \min_{J\in\mathcal{J}} \mathbb{P}\left(S_J[z_J K/\xi_J] - S_J[0] \leq -K z_J\right) \\
&\geq \min_{J\in\mathcal{J}} e^{-z_J K\epsilon/\xi_J} e^{-g(\phi,J)z_J K}.
\end{aligned}$$

Hence

$$\begin{aligned}
\liminf_{K\to\infty} \frac{1}{K} \log P_o^{K,U} &\geq \liminf_{K\to\infty} \frac{1}{K} \log \frac{\min_{J\in\mathcal{J}} e^{-z_J K\epsilon/\xi_J} e^{-g(\phi,J)z_J K}}{K/\xi} \\
&\geq -\max_{J\in\mathcal{J}} \left(\frac{z_J \epsilon}{\xi_J} + g(\phi,J) z_J\right).
\end{aligned}$$

Since $\epsilon$ can be chosen arbitrarily, we have:

$$\liminf_{K\to\infty} \frac{1}{K} \log P_o^{K,U} \geq -\max_{J\in\mathcal{J}} \left(g(\phi,J)z_J\right) = -\sup_{\mathbf{w}\in\mathrm{relint}(\Omega)} \min_{J\in\mathcal{J}} \left(\mathbf{1}^{\mathrm{T}}_{\partial(J)}\mathbf{w}\right) g(\phi,J) = \gamma(\mathbf{w}).$$

This concludes the proof. $\square$

### E.3. Converse of State-Independent Policies: Proof of Proposition 4

Proof. We first define an 'augmented' policy $\tilde{\pi}$ for any state-independent policy $\pi$. Policy $\tilde{\pi}$ is also state independent, where:

$$\tilde{u}_{(i,j')}[t] = u_{(i,j')}[t] + \frac{1}{|\partial(j')|}\left(1 - \sum_{i\in\partial(j')} u_{(i,j')}[t]\right).$$

Note that $\tilde{\pi}$ is a non-idling policy, and $\tilde{\pi} = \pi$ if $\pi$ is non-idling. In the following analysis, we couple $\pi$ and $\tilde{\pi}$ in such a way that if $\pi$ dispatch from $i$ to serve the $t$-th demand, then $\tilde{\pi}$ will do the same.

We first divide the discrete periods into cycles with length $K^2$. We will lower bound the probability of demand-drop on any cycle. Without loss of generality, consider cycle $[0, K^2-1]$. Suppose $X^{K,\pi}[0] = \mathbf{X}_0$. By Assumption 1, there exists a subset $J \in \mathcal{J}$. Consider the 'virtual' process of change of number of supplies in $\partial(J)$, denoted by $S_J[t]$:

- $S_J[0] = 0$.
- $S_J[t+1] = S_J[t] + 1$ if $d[t] \in \partial(J)$ and policy $\tilde{\pi}$ dispatches a vehicle from $\partial(J)^c$ to serve the $t$-th demand (regardless of whether the demand is fulfilled).
- $S_J[t+1] = S_J[t] - 1$ if $d[t] \in \partial(J)^c$ and policy $\tilde{\pi}$ dispatches a vehicle from $\partial(J)$ to serve the $t$-th demand (regardless of whether the demand is fulfilled).
- $S_J[t+1] = S_J[t]$ if otherwise.

Observe that:

$$\mathbb{P}\left(\text{some demand is dropped in } \in [0, K^2-1]\right) \geq \mathbb{P}\left(S_J(K^2-1) + \mathbf{1}^T_{\partial(J)}\mathbf{X}_0 \geq K \text{ or } \leq 0\right). \tag{38}$$

This is because : (1) if $S_J[t] + \mathbf{1}^T_{\partial(J)}\bar{\mathbf{X}}_0 = \mathbf{1}^T_{\partial(J)}\bar{\mathbf{X}}^{K,\pi}[t]$ for all $t \leq T$, then $S_J(K^2) + \mathbf{1}^T_{\partial(J)}\mathbf{X}_0$ is exactly the number of supplies in $\partial(J)$ at $K^2$, and demand will be dropped if $\partial(J)$ is empty or $\partial(J)^c$ is empty; (2)if $S_J[t] + \mathbf{1}^T_{\partial(J)}\mathbf{X}_0 \neq \mathbf{1}^T_{\partial(J)}\bar{\mathbf{X}}^{K,\pi}[t]$ for some $t < K^2$, it means a unit of demand was already dropped at the smallest such $t$. Either way, RHS implies LHS.

Note that $S_J(K^2)$ is the sum of $K^2$ independent random variables $Z_t$, where $Z_t = S_J[t+1] - S_J[t]$. Here $Z_t$ has support $\{-1, 0, 1\}$ and satisfies:

$$\mathbb{P}(Z_t = -1) \geq \mathbb{P}(o[t] \in J, d[t] \in \partial(J)^c)$$

There are two scenarios:

- If $\mathbb{E}[S_J(K^2)] \leq -\frac{K^2}{2}$, then for $K > 8$,

$$\begin{aligned}
\mathbb{P}\left(S_J(K^2) + \mathbf{1}^T_{\partial(J)}\mathbf{X}_0 \geq K \text{ or } \leq 0\right) &\geq 1 - \mathbb{P}\left(S_J(K^2) \in [-K, K]\right) \\
&\geq 1 - \mathbb{P}\left(S_J(K^2) - \mathbb{E}[S_J(K^2)] \geq -K + \frac{K^2}{2}\right) \\
&\geq 1 - 2\exp\left(-\frac{K^2}{32}\right) \quad \text{(Hoeffding's inequality)} \\
&\geq \frac{1}{2}.
\end{aligned}$$

- If $\mathbb{E}[S_J(K^2)] > -\frac{K^2}{2}$, then:

$$\text{the number of } t\text{'s where } \mathbb{E}[Z_t] \geq -\frac{3}{4} \text{ is at least } \frac{K^2}{7}. \tag{39}$$

Denote the set of these $t$'s as $\mathcal{T}$. Hence

$$\mathrm{Var}(S_J(K^2)) = \sum_{t=1}^{K^2} \mathrm{Var}(\xi_t) \geq \sum_{t\in\mathcal{T}} \mathrm{Var}(\xi_t) \geq \frac{K^2}{7} \cdot \delta \left(1 - \frac{3}{4}\right)^2 = \frac{\delta K^2}{102}. \tag{40}$$

Apply Theorem 7.4.1 in Chung (2001) (Berry-Esseen Theorem), we have:

$$\sup_{x\in\mathbb{R}} \left| \mathbb{P}\left(S_J(K^2) - \mathbb{E}(S_J(K^2)) \leq x\sqrt{\mathrm{Var}(S_J(K^2))}\right) - \Phi(x) \right| \leq \frac{\sum_{t=1}^{K^2} \mathbb{E}|Z_t - \mathbb{E}Z_t|^3}{\mathrm{Var}(S_J(K^2))^{3/2}}, \tag{41}$$

where $\Phi(\cdot)$ is the cumulative distribution function of standard normal distribution. Plug in (39) and (40), we can upper bound the RHS of (41) by $10000\delta^{-3/2}\frac{1}{K}$. For $K > \frac{40000A_0\delta^{-3/2}}{2\bar{\Phi}(50/\sqrt{\delta})}$:

$$\mathbb{P}(S_J(K^2) \in B\left(\mathbb{E}(S_J(K^2)), 4K\right) \in B\left(2\bar{\Phi}\left(\frac{4K}{\mathrm{Var}(S_J(K^2))}\right), \frac{1}{2}\bar{\Phi}\left(\frac{4K}{\mathrm{Var}(S_J(K^2))}\right)\right)$$
$$\mathbb{P}(S_J(K^2) \in B\left(\mathbb{E}(S_J(K^2)), 2K\right) \in B\left(2\bar{\Phi}\left(\frac{2K}{\mathrm{Var}(S_J(K^2))}\right), \frac{1}{2}\bar{\Phi}\left(\frac{4K}{\mathrm{Var}(S_J(K^2))}\right)\right).$$

Here $B(x,a) \triangleq [x-a, x+a]$. Now we are ready to bound demand-drop probability. Note that one of the three cases will happen: $[-K,K] \subset B\left(\mathbb{E}(S_J(K^2)), 4K\right)$, $[-K,K] \cap B\left(\mathbb{E}(S_J(K^2)), 4K\right) = \emptyset$, $[-K,K] \cap B\left(\mathbb{E}(S_J(K^2)), 2K\right) = \emptyset$. In any case, the probability of $S_J(K^2)$ fall into an interval with width $2K$ is uniformly bounded away from 1, denoted by $c < 1$. Hence

$$\mathbb{P}\left(\text{some demand is dropped in } \in [1, K^2]\right) = c > 0.$$

Combine the two cases, we have the desired result. $\square$

# References

Adan I, Weiss G (2012) A loss system with skill-based servers under assign to longest idle server policy. *Probability in the Engineering and Informational Sciences* 26(3):307–321.

Adelman D (2007) Price-directed control of a closed logistics queueing network. *Operations Research* 55(6):1022–1038.

Asmussen S (2008) *Applied Probability and Queues*, volume 51 (Springer Science & Business Media).

Ata B, Kumar S (2005) Heavy traffic analysis of open processing networks with complete resource pooling: asymptotic optimality of discrete review policies. *Annals of Applied Probability* 15(1A):331–391.

Banerjee S, Freund D, Lykouris T (2016) Pricing and optimization in shared vehicle systems: an approximation framework. *CoRR abs/1608.06819* .

Bertsekas DP (1995) *Dynamic Programming and Optimal Control*, volume 1 (Athena scientific Belmont, MA).

Bertsimas D, Gamarnik D, Tsitsiklis JN (2001) Performance of multiclass markovian queueing networks via piecewise linear lyapunov functions. *Annals of Applied Probability* 1384–1428.

Bimpikis K, Candogan O, Saban D (2016) Spatial pricing in ride-sharing networks .

Blanchet J (2013) Optimal sampling of overflow paths in jackson networks. *Mathematics of Operations Research* 38(4):698–719.

Bodas S, Shakkottai S, Ying L, Srikant R (2014) Scheduling in multi-channel wireless networks: rate function optimality in the small-buffer regime. *IEEE Transactions on Information Theory* 60(2):1101–1125.

Braverman A, Dai JG, Liu X, Ying L (2016) Empty-car routing in ridesharing systems. *arXiv preprint arXiv:1609.07219* .

Bušić A, Meyn S (2015) Approximate optimality with bounded regret in dynamic matching models. *ACM SIGMETRICS Performance Evaluation Review* 43(2):75–77.

Cachon GP, Daniels KM, Lobel R (2017) The role of surge pricing on a service platform with self-scheduling capacity. *Manufacturing & Service Operations Management* .

Caldentey R, Kaplan EH, Weiss G (2009) Fcfs infinite bipartite matching of servers and customers. *Advances in Applied Probability* 41(3):695–730.

Chung KL (2001) *A Course in Probability Theory* (Academic press).

Dai JG (1995) On positive harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *Annals of Applied Probability* 49–77.

Dai JG, Lin W (2005) Maximum pressure policies in stochastic processing networks. *Operations Research* 53(2):197–218.

Dai JG, Lin W (2008) Asymptotic optimality of maximum pressure policies in stochastic processing networks. *Annals of Applied Probability* 18(6):2239–2299.

Dembo A, Zeitouni O (1998) *Large Deviations Techniques and Applications*, volume 38 of *Application of Mathematics* (Springer), 2nd edition.

Durrett R (2010) *Probability: Theory and Examples* (Cambridge university press).

Eryilmaz A, Srikant R (2012) Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Systems* 72(3-4):311–359.

Gurvich I, Whitt W (2009) Scheduling flexible servers with convex delay costs in many-server service systems. *Manufacturing & Service Operations Management* 11(2):237–253.

Hall J, Kendrick C, Nosko C (2015) The effects of uber's surge pricing: a case study. *The University of Chicago Booth School of Business* .

Harrison JM, López MJ (1999) Heavy traffic resource pooling in parallel-server systems. *Queueing systems* 33(4):339–368.

Kumar S, Kumar P (1996) Closed queueing networks in heavy traffic: fluid limits and efficiency. *Stochastic Networks*, 41–64 (Springer).

Maguluri ST, Srikant R (2016) Heavy traffic queue length behavior in a switch under the maxweight algorithm. *Stochastic Systems* 6(1):211–250.

Mairesse J, Moyal P (2016) Stability of the stochastic matching model. *Journal of Applied Probability* 53(4):1064–1077.

Majewski K, Ramanan K (2008) How large queue lengths build up in a jackson network .

Meyn S (2009) Stability and asymptotic optimality of generalized maxweight policies. *SIAM Journal on Control and Optimization* 47(6):3259–3294.

Meyn SP, Tweedie R (1994) State-dependent criteria for convergence of markov chains. *Annals of Applied Probability* 149–168.

Ozkan E, Ward AR (2016) Dynamic matching for real-time ridesharing .

Rudin W (1964) *Principles of Mathematical Analysis*, volume 3 (McGraw-hill New York).

Shi C, Wei Y, Zhong Y (2015) Process flexibility for multi-period production systems .

Shwartz A, Weiss A (1993) Induced rare events: analysis via large deviations and time reversal. *Advances in Applied Probability* 25(3):667–689.

Shwartz A, Weiss A (1995) *Large deviations for performance analysis: queues, communication and computing*, volume 5 (CRC Press).

Stolyar AL (2003) Control of end-to-end delay tails in a multiclass network: Lwdf discipline optimality. *Annals of Applied Probability* 1151–1206.

Stolyar AL (2004) Maxweight scheduling in a generalized switch: state space collapse and workload minimization in heavy traffic. *Annals of Applied Probability* 14(1):1–53.

Stolyar AL, Ramanan K (2001) Largest weighted delay first scheduling: Large deviations and optimality. *Annals of Applied Probability* 1–48.

Subramanian VG (2010) Large deviations of max-weight scheduling policies on convex rate regions. *Mathematics of Operations Research* 35(4):881–910.

Tassiulas L, Ephremides A (1992) Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE transactions on automatic control* 37(12):1936–1948.

Venkataramanan V, Lin X (2013) On the queue-overflow probability of wireless systems: a new approach combining large deviations with lyapunov functions. *IEEE Transactions on Information Theory* 59(10):6367–6392.

Waserhole A, Jost V (2016) Pricing in vehicle sharing systems: optimization in queuing networks with product forms. *EURO Journal on Transportation and Logistics* 5(3):293–320.