Reinforcement with Fading Memories

Kuang Xu
Graduate School of Business, Stanford University, Stanford, CA 94305, kuangxu@stanford.edu

Se-Young Yun
Los Alamos National Laboratory, Los Alamos, NM 87545, yunseyoung@gmail.com

We study the effect of imperfect memory on decision making in the context of a stochastic sequential action-reward problem. An agent chooses a sequence of actions which generate discrete rewards at different rates. She is allowed to make new choices at rate $\beta$, while past rewards disappear from her memory at rate $\mu$. We focus on a family of decision rules where the agent makes a new choice by randomly selecting an action with a probability approximately proportional to the amount of past rewards associated with each action in her memory.

We provide closed-form formulae for the agent’s steady-state choice distribution in the regime where the memory span is large ($\mu \rightarrow 0$), and show that the agent’s success critically depends on how quickly she updates her choices relative to the speed of memory decay. If $\beta \gg \mu$, the agent almost always chooses the best action, i.e., the one with the highest reward rate. Conversely, if $\beta \ll \mu$, the agent chooses an action with a probability roughly proportional to its reward rate.*

Key words: memory, stochastic model, $M/M/\infty$ queues, Markov process, fluid model.

1. Introduction

Memories serve as a crucial link between the past and future in dynamic decision-making problems, allowing subsequent decisions to draw on earlier experiences. However, they are hardly perfect in reality: humans routinely rely on faulty memories when making choices, and a firm’s time-varying states, such as an active user base that is prone to unpredictable attrition, could sway its strategic focus. This raises the question: to what extent can we make good decisions despite having an imperfect memory? In this paper, we investigate this question in the context of a stochastic action-reward model. While the subject can be approached from many different angles, we will focus on understanding the interplay between the rate of memory decay and the rate of decision updates; the former a measure of the quality of memory, and the latter a proxy of the decision maker’s temporal adaptivity. Our main results demonstrate that the relative magnitude between the two rates can have profound performance implications.

Let us begin with an informal description of our model (Figure 1). An agent operating in continuous time makes choices among a finite menu of actions. When action $k$ is chosen, the agent accrues discrete rewards according to a Poisson process with rate $\lambda_k$, while a unit of reward “expires” and disappears from her memory after a random amount of time that is exponentially distributed with rate $\mu$. Opportunities for the agent to make new choices arise according to a Poisson process with rate $\beta$. We will restrict our attention to a simple family of decision rules that the agent may use, referred to as the reward matching rule, which reinforces actions with more positive past stimuli in proportional manner: she randomly samples an action such that the probability of choosing action $k$ is proportional to $\max\{Q_k, \alpha\}$, where $Q_k$ is the total units of rewards accrued through action $k$ that have yet to expire, and $\alpha > 0$ an exploration parameter representing the agent’s minimal willingness to experiment with any action. If $\alpha$ is set to 0, then the reward matching rule corresponds to the celebrated Luce’s linear probabilistic choice model (Luce (1959), Erev and Starmer (1996)).

* This version: January 10, 2018.
Roth (1998)), which has been extensively studied in behavioral economics (Erev and Roth (1998), Beggs (2005)) and evolutionary biology (Harley (1981)). We are interested in the probability that the agent chooses a particular action in steady state. We next examine two illustrative examples, which also serve as motivating applications for the more stylized model we study.

**Example 1 - Consumer Choice Modeling.** We may think of the agent as a consumer, the actions as products or service providers, and the opportunities to choose new actions as the times when the consumer is allowed to re-select her membership or subscription. The rewards act as positive experiences or impressions resulting from using a service, and the strength of each impression in the agent’s memory diminishes as time progresses. When an opportunity to renew the subscription arises, the agent chooses a new service offering in a manner that is biased towards the ones associated with more recallable impressions.

**Example 2 - Dynamic Product Offering under Customer Attrition.** In this example, we slightly refine the model by letting the departure rates of rewards vary across actions. Consider a firm who operates a number of service contracts (actions), e.g., mobile phone plans, and needs to decide which contract to offer new customers (rewards) in a series of promotion periods, where one contract is offered for each period. Customers arrive to the system and are subscribe to the service contract of the corresponding promotion period. Those who subscribe to contract type $k$ remain an active customer for an exponentially distributed amount of time with rate $\mu_k$, before departing from the system. The departure rate $\mu_k$ reflects some underlying qualities of the contract $k$. The reward matching rule corresponds to the service provider’s choosing, at the end of each promotion period, a new contract with probabilities roughly proportional to the number of active customers associated with each contract.

**Motivation for studying the reward matching rule.** An implied objective of our model is that the agent would like to maximize the steady-state total amount of rewards in system. If all the parameters were known a priori, it is not difficult to see that this objective could be trivially achieved by choosing at all times the best action, i.e., one with the highest reward rate $\lambda_k$ (Example 1) or the highest weighted reward rate, $\lambda_k/\mu_k$ (Example 2). The reward matching rule, therefore, can be viewed as an intuitive heuristic for approximately solving this optimization problem in the absence of the knowledge of the parameters. While the rule is by no means the only plausible heuristic for this purpose, we believe that it has a number of advantages which merit a theoretical investigation such as the one carried out in this work:

1. Firstly, Luce’s rule, after which the reward matching rule is modeled, has a long history as a fundamental building block in psychology and behavioral economics (see Section 3 for more details). The reward matching rule hence serves a natural starting point for understanding other, potentially more complex, cognitive models.
2. Secondly, the reward matching rule is useful not only for modeling human choice behavior. It could also be a policy deliberately chosen by a decision maker, such as in Example 2. While a conscious decision maker could in principle adopt more sophisticated decision rules or even use external information storage to mitigate memory decay, the reward matching rule offers a simple and intuitive rule of thumb that can be substantially easier for a practitioner to understand and implement, and is applicable in settings where external information storage is not possible.
3. Last but not least, our results show that, under a large number of parameter values, the reward matching rule is in fact near optimal and induces a steady-state choice probabilities that heavily concentrate on the best action. Along with the simplicity and conceptual appeal, the near-optimality therefore further justifies studying the reward-matching rule.

**Preview of main results.** Our main results provide simple, closed-form formulae for the distribution of the agent’s choice in steady state, in the regime where the agent’s memory span is large ($\mu \to 0$). Using these formulae, we show that the agent’s success critically depends on how
fast her choices are updated ($\beta$) relative to the rate of memory decay ($\mu$), when $\alpha$ is relatively small compared to the overall memory size. In particular, if $\beta \gg \mu$, the agent nearly always chooses the best action, i.e., the one with the highest reward rate $\lambda_k$. In contrast, if $\beta \ll \mu$, the agent splits her attention more smoothly among all actions, with frequencies roughly proportional to the respective $\lambda_k$'s.

![Figure 1](image-url)  

**Figure 1.** A snapshot of the sequential decision-making problem, where the choice at time $t$ is action 1, and the agent accrues discrete rewards according to Poisson process of rate $\lambda_1$. Each existing recallable reward departs from the system at rate $\mu$, as depicted by the horizontal arrows.

1.1. The Model  
We now describe our model more formally, as is depicted in Figure 1. The system consists of an agent operating in continuous time, who makes choices by selecting from a set of $K$ actions, $\mathcal{K} = \{1, \ldots, K\}$. We denote by $C(t) \in \mathcal{K}$ the action chosen by the agent at time $t$, and refer to $\{C(t)\}_{t \in \mathbb{R}^+}$ as the choice process. When the $C(t) = k$ for some $k \in \mathcal{K}$, the agent accrues discrete rewards according to a Poisson process with rate $\lambda_k > 0$, where each reward can be treated as a token of unit size. We refer to $\lambda_k$ as the reward rate of action $k$. We assume that there exists one reward rate that strictly dominates the rest, and, without loss of generality, we rank them in a decreasing order, so that $\lambda_1 > \lambda_2 \geq \ldots \geq \lambda_K > 0$.

The agent’s fading memory is modeled by the “expiration” of past rewards. Each reward is associated with a lifespan, an exponential random variable with rate $\mu > 0$ drawn independently from all other aspects of the system, and a reward departs permanently after staying in the system for a time duration equal to its lifespan. We refer to $\mu$ as the rate of memory decay. For simplicity of notation, the majority of our exposition and proofs will focus on the case where $\mu$ is uniform across all actions, but our main results extend to the case with action-dependent memory decay rates in a straightforward manner; such a refinement will be discussed in Section 2.2.

For $k \in \mathcal{K}$, we denote by $Q_k(t)$ the total units of rewards accrued from choosing action $k$ that have yet to depart by time $t$, and we refer to $\{Q(t)\}_{t \in \mathbb{R}^+}$ as the recallable reward process. Note that at time $t$ the recallable rewards associated with action $k$ depart at an aggregate rate of $\mu Q_k(t)$, and arrive at rate $\lambda_k$, if $C(t) = k$, or 0, otherwise. We assume that the agent’s initial recallable rewards at time 0 is a bounded vector in $\mathbb{Z}_+^K$.

The agent has the opportunity to make a new choice at a set of update points scattered across time. The update points are distributed according to an exogenous Poisson process with rate $\beta$, which we refer to as the update rate, so that the time between two adjacent update points is an exponentially distributed random variable with mean $\beta^{-1}$, independent from all other aspects of the system.

---

1 The notation $\beta \gg \mu$ represents $\lim_{\mu \to 0} \beta/\mu = 0$. 
How does the agent make a new choice when an update point arises? In this work, we will focus on a family of choice heuristics referred to as the reward matching rules. We assume that at time $t$ the agent makes a new choice by sampling from the distribution

$$\mathbb{P}(\text{new choice} = k) = \frac{Q_k(t) \vee \alpha}{\sum_{i \in K} (Q_i(t) \vee \alpha)}, \quad k \in K,$$

where $x \vee y \triangleq \max\{x, y\}$. Here, $\alpha$ is a positive constant referred to as the exploration parameter, which captures the agent’s willingness to experiment with an action even if there is currently little or no recallable reward attached to it. The justification and background behind the choice of this decision rule will be further elaborated on in Section 3.

1.2. Performance Metrics and Scaling Regime The main goal of this paper is to study the agent’s long-run behavior in the model described above, and a key quantity of interest is the steady-state distribution of the choice process, $C(\cdot)$, i.e.,

$$\lim_{t \to \infty} \mathbb{P}(C(t) = k), \quad k \in K,$$

Unfortunately, the choice process is non-Markovian as its transitions are influenced by the recallable reward process $Q(\cdot)$ in a non-trivial way, and this makes it difficult to obtain explicit expressions for its steady-state distribution. To circumvent this challenge, we will focus on the regime where the average lifespan of a reward, $\mu^{-1}$, tends to infinity, as follows. Fix the action set, $K$, and the reward rates, $\{\lambda_k\}_{k \in K}$. We consider a sequence of systems, indexed by $m \in \mathbb{N}$, where the average lifespan of a reward in the $m$th system is equal to $m$, i.e.,

$$\mu = 1/m.$$  

We will denote by $\alpha_m$, $\beta_m$, $Q^m(\cdot)$, and $C^m(\cdot)$ the corresponding quantities in the $m$th system. Finally, as the rewards’ average lifespan, $m$, tends to infinity, the total amount of recallable rewards in the system also scales as $\Theta(m)$\footnote{To see this, note that if the choice process were to stay in action 1 all the time, then $Q^m(\cdot)$ would evolve according to an $M/M/\infty$ queue with arrival rate $\lambda_1$ and departure rate $m^{-1}$, leading to a steady-state expected total recallable reward of $\lambda_1m$.} For this reason, we will scale the exploration parameter $\alpha_m$ linearly with respect to $m$, by setting $\alpha_m = \alpha_0 \ast m$, where $\alpha_0$ is a fixed parameter.

Remark. It is natural to ask whether by taking a sequence of systems where the memory decay rate approaches zero, we will end up with the trivial case where there is no memory loss at all ($\mu = 0$), which would have been at odds with our goal of studying imperfect memories. This not the case. Because we always look at the steady state of the system by taking the limit as $t \to \infty$.
for every system in the sequence, the limiting regime is non-trivial and is not equivalent to simply setting \( \mu = 0 \), as is evidenced by the existence of two distinct limiting steady-state distributions in our main theorems (This point is further elaborated on in Appendix C). The limit of \( \mu \to 0 \) is intended as an approximation for a pre-limit system where \( \mu \) is positive but small. Numerical results in Figure 4 and 5 suggest that our theorems provide a reasonably good approximation for \( m \) as small as 200.

1.3. Notation For \( x, y \in \mathbb{R} \), \( x \vee y \) and \( x \wedge y \) denote \( \max\{x, y\} \) and \( \min\{x, y\} \), respectively. For a vector \( x \in \mathbb{R}_+^K \), \( \|x\| \) denotes the \( L_1 \) norm of \( x \), \( \|x\| = \sum_{i=1}^K |x_i| \).

We use the notation \( X(\cdot) \) and \( X[\cdot] \) as a shorthand for processes \{\( X(t) \)\}_{t \in \mathbb{R}_+} \text{ and } \{\( X[n] \)\}_{n \in \mathbb{N}} \text{, respectively. For two random variables } X \text{ and } Y \text{ taking values in } \mathbb{R}, \text{ we use } X \overset{d}{=} Y \text{ to mean that } X \text{ and } Y \text{ have the same distribution. The expression } X \preceq Y \text{ means } X \text{ is stochastically dominated by } Y, \text{ i.e., } \mathbb{P}(X \geq x) \leq \mathbb{P}(Y \geq x) \text{ for all } x \in \mathbb{R}. \text{ For an event } A, \text{ we use } X|A \text{ to denote the random variable with the cumulative distribution function } \mathbb{P}(X \leq x \mid A), \ x \in \mathbb{R}. \text{ For a stochastic process, } \{\( X(t) \)\}_{t \in \mathbb{R}_+}, \text{ we denote by } X(\infty) \text{ the random variable distributed according to the stationary distribution of } \{\( X(t) \)\}_{t \in \mathbb{R}_+}, \text{ if it exists.}

For two functions \( f \) and \( g \), \( f \gg g \) means that \( \lim_{x \to \infty} \frac{f(x)}{g(x)} = \infty \); \( f \ll g \) is defined analogously. When such a limiting notation is involved, the limit is assumed to be as the argument tends to \( \infty \), unless explicitly specified otherwise.

2. Main Results We now present our main results, which provide exact expressions for the steady-state distribution of the choice process, as well as the expected value of the (scaled) recallable reward process, in the limit as \( m \to \infty \). For the remainder of the paper, we will assume that the update rate \( \beta_m \) satisfies either \( \beta_m \gg 1/m \) or \( \beta_m \ll 1/m \), as \( m \to \infty \).

**Theorem 1 (Steady-State Choice Probabilities).** Fix \( \alpha_0 > 0 \). The following limits exist

\[
c_k = \lim_{m \to \infty} \mathbb{P}(C^m(\infty) = k), \quad k \in \mathcal{K}.
\]

Furthermore, their expressions depend on the scaling of \( \beta_m \), as follows.

1. (Memory-abundant regime) Suppose that \( \beta_m \gg 1/m \), as \( m \to \infty \). If \( \alpha_0 \leq \lambda_1/K \), then

\[
\begin{align*}
    c_1 &= 1 - \frac{K - 1}{\lambda_1} \alpha_0, \\
    c_k &= \frac{\alpha_0}{\lambda_1}, \quad k = 2, \ldots, K.
\end{align*}
\]

If \( \alpha_0 > \lambda_1/K \), then

\[
c_k = 1/K, \quad k = 1, \ldots, K.
\]

2. (Memory-deficient regime) Suppose that \( \beta_m \ll 1/m \), as \( m \to \infty \). Then,

\[
c_k = \frac{\lambda_k \vee \alpha_0 + (K - 1)\alpha_0}{\sum_{i \in \mathcal{K}} [\lambda_i \vee \alpha_0 + (K - 1)\alpha_0]}, \quad k = 1, \ldots, K.
\]

The next theorem characterizes the steady-state expected value of the recallable reward process, \( Q^m(\cdot) \). For \( m \in \mathbb{N} \), define the scaled recallable reward process:

\[
\overline{Q}^m(t) = \frac{1}{m} Q^m(mt), \quad t \in \mathbb{R}_+.
\]
Theorem 2 (Steady-State Recallable Rewards). Fix $\alpha_0 > 0$. For all $m \in \mathbb{N}$, the scaled recallable reward process, $\{\overline{Q}^m_t\}_{t \in \mathbb{R}_+}$, admits a unique steady-state distribution, and the following limits exist:

$$q_k = \lim_{m \to \infty} \mathbb{E}(\overline{Q}^m_k(\infty)), \quad k \in K.$$  

(9)

Furthermore, their expressions depend on the scaling of $\beta_m$, as follows.

1. (Memory-abundant regime) Suppose that $\beta_m \gg 1/m$, as $m \to \infty$. If $\alpha_0 \leq \lambda_1/K$, then

$$q_1 = \lambda_1 - (K-1)\alpha_0,$$

$$q_k = (\lambda_k/\lambda_1)\alpha_0, \quad k = 2, \ldots, K,$$

(10)

If $\alpha_0 > \lambda_1/K$, then

$$q_k = \lambda_k/K, \quad k = 1, \ldots, K.$$  

(11)

2. (Memory-deficient regime) Suppose that $\beta_m \ll 1/m$, as $m \to \infty$. Then,

$$q_k = \lambda_k \sum_{i \in K} [\lambda_i \lor \alpha_0 + (K-1)\alpha_0], \quad k = 1, \ldots, K.$$  

(12)

2.1. Implications of the Theorems A prominent feature of Theorems 1 and 2 is that the results critically depend on how quickly the update rate $\beta_m$ scales relative to the rate of memory decay $1/m$, but are otherwise insensitive to the exact expression of $\beta_m$. This leads to a natural dichotomy of the system dynamics into what we called the memory-abundant and memory-deficient regimes, and we discuss below several notable properties of the two regimes.

**Near optimality and winner-takes-all.** Suppose that the constant $\alpha_0$ in the exploration parameter is substantially smaller than $\lambda_1/K$. Then, our results show that the best action attracts nearly all of the agent’s attention, if $\beta_m \ll 1/m$ (memory-abundant regime). Specifically, in the limit as $m \to \infty$, the agent chooses action 1 with a probability of $(1 - K^{-1}\alpha_0)$, which is almost 1 when $\alpha_0$ is very small compared to $\lambda_1/K$. Recall from Section 1 that the optimal strategy for the agent, should she know all parameters of the problem, would be to choose action 1 at all times. Therefore, in this regime, the reward matching rule is able to deliver near optimal performance.

Moreover, the system exhibits an interesting winner-takes-all phenomenon, whereby the degree to which the agent focuses on the best action is independent of the reward rates of all other actions. For instance, if we were to increase the reward rate of the second-best action, $\lambda_2$, the probability that the agent chooses action 1 would stay unchanged (and close to 1), and that of choosing action 2 would still be close to zero no matter how close $\lambda_2$ is to $\lambda_1$. One consequence of this effect is that, if we imagine the actions as being “competitive” service providers vying for a consumer’s attention, then the winner-takes-all phenomenon can be interpreted as a form of extreme competition among the providers, where the best attracts nearly all of the consumer’s budget, even if its superiority compared to the second best is not significant.

**Reward-rate-proportional allocation.** In contrast to the memory-abundant regime, an update rate that scales substantially slower than the memory decay rate will lead to an allocation of choices that is much smoother as a function of the reward rates of all other actions. For instance, if we were to increase the reward rate of the second-best action, $\lambda_2$, the probability that the agent chooses action 1 would stay unchanged (and close to 1), and that of choosing action 2 would still be close to zero no matter how close $\lambda_2$ is to $\lambda_1$. One consequence of this effect is that, if we imagine the actions as being “competitive” service providers vying for a consumer’s attention, then the winner-takes-all phenomenon can be interpreted as a form of extreme competition among the providers, where the best attracts nearly all of the consumer’s budget, even if its superiority compared to the second best is not significant.
Figure 3. An example of the limiting steady-state distribution of the choice process $C_m(\cdot)$ in the two regimes given by Eqs. (5) through (7), with reward rates $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (8, 6, 4, 2)$ and $\alpha_0 = 0.5$. The agent’s attention sharply concentrates on the best action in one regime, and is more spread out among all actions in the other.

The intuition behind this reward-rate-proportional allocation can be roughly explained as follows. When the update rate is slow, by the time a new choice is to be made the agent would have forgotten the rewards associated with all actions other than her most recent choice. Nevertheless, she still exhibits a mild preference over actions with higher reward rates in steady state, because she does retain a good memory of the rewards of her most recent choice, $C(t)$, which can be shown to be proportional to $\lambda_{C(t)}$. The reward rate $\lambda_{C(t)}$ dictates how likely the agent will make the same choices, which in turn leads to the reward rates’ approximately linear appearance in the steady-state distribution of the choice process.

Rapid updates lead to complete oblivion. The theorems also show that, perhaps surprisingly, a slower update rate is not always bad. While it may be tempting to conclude that faster update rates always lead to better outcomes for the agent, this turns out to be true only when the exploration parameter is relatively small. In fact, as soon as $\alpha_0$ grows beyond $\lambda_1/K$, in the memory-abundant regime, the agent becomes completely oblivious and chooses actions uniformly at random (Eq. (6)). In contrast, the agent is always biased towards the best action in the memory-deficient regime, albeit mildly, as long as $\alpha_0 < \lambda_1$ (beyond which point the exploration parameter is so large that it trivially overwhelms even the best action). The true culprit behind the complete oblivion is a lack of patience: when the update rate is fast and the exploration parameter high, the agent tends to switch her choices rapidly, before she is able to collect enough rewards to “learn” the reward rate of any individual action. Consequently, in the long run, not a single component of the recallable reward, not even the best action, can distinguish itself by consistently rising above the exploration parameter, and hence complete oblivion becomes the only possible outcome.

2.2. Generalization with Heterogeneous Memory Decay Rates As was eluded to in Example 2 in the Introduction, we may also consider the case where the memory decay rates are not constant across actions, so that the rewards associated with action $k$ depart at rate $\mu_k$. This generalization is useful for modeling service system where the rate of customer departure or attrition depends on the service type a customer is associated with.

Analogous to the case of uniform decay rate, the scaling of Section 1.2 corresponds to setting $\mu = \mu_0 m^{-1}$ in the $m$th system, where $\{\mu_0^k\}_{k \in K}$ are positive constants. Under this scaling, it is not difficult to verify that both Theorems 1 and 2 extend to this more general case, if we replace every occurrence of $\lambda_k$ with a decay-rate-weighted version, $\lambda_k/\mu_k$. In other words, as far as the limiting system is concerned, the version of the problem with heterogeneous memory decay rates is equivalent to one with uniform decay rates but appropriately weighted reward rates. We should note that this equivalence only holds in the limit as $m \to 0$, and is not exact in a pre-limit system with a finite $m$. 
2.3. Numerical Results  While Theorems 1 and 2 only apply in the limit where the average reward lifespan tends to infinity, the simulation results in Figures 4 and 5 show that our theoretical results also provide fairly accurate predictions, quantitatively and qualitatively, in systems with a moderate, finite lifespan. In Figure 4 we fix $m$, while varying the update rate, $\beta$, and we observe that concentration on the best action intensifies as $\beta$ increases. Figure 5 shows example sample paths of the scaled recallable reward processes, $\bar{Q}^m(\cdot)$. One can discern without difficulty that, even for a modest $m$, the process in the memory-abundant regime (plots (a) through (c)) stays close to a set of smooth “fluid” trajectories, whose invariant state (the flat portion of the trajectories) coincide with the limiting value of $E(\bar{Q}^m(\infty))$ predicted in Eq. (10), while in the memory-deficient regime the process changes more abruptly over time. As will become clear in the sequel, this qualitative discrepancy is a direct consequence of the overall system dynamics being dominated by the evolution of either $\bar{Q}^m(\cdot)$ or $C^m(\cdot)$, depending on the scaling of the update rate $\beta_m$, a key fact that we will exploit in our proof of the main theorems.

![Figure 4](image_url)

**Figure 4.** Time-average scaled recallable rewards (plot (a)) and distribution of choices (plot (b)) as a function of the update rate, $\beta$, with $K = 4$, $(\lambda_1, \lambda_2, \lambda_3, \lambda_4) = (8, 6, 4, 2)$, $\alpha_0 = 1$, and $m = 200$. Each point is averaged over $5 \times 10^7$ to $5 \times 10^5$ time units for $\beta$ ranging from $10^{-4}$ to 1, respectively (longer duration for a slower update rate).

3. Related Literature  The assumption of a reward’s lifespan being exponentially distributed is modeled after the exponential memory decay theory of Ebbinghaus (Ebbinghaus (1964)), which posits that the recall probability of a past event decays exponentially as time passes. The reward matching decision rule (Eq. (1)) can be viewed as a generalization of Luce’s linear probabilistic choice rule (Luce (1959)), originally created to model empirical observations where humans or animals subjects make choices with probabilities proportional to associated amount of past rewards or stimuli. Luce’s rule has been studied extensively in a variety of disciplines, ranging across cognitive science (Herrnstein (1970)), behavioral economics (Erev and Roth (1998)), and evolutionary biology (Harley (1981)). Notably, the celebrated work of Erev and Roth (1998) demonstrates that a discrete-time version of Luce’s rule can be a powerful reinforcement learning model for explaining a player’s behavior in sequential games, and subsequent theoretical analysis showed that the empirical frequency of the best action under Luce’s rule converges to one in single player games (Rustichini (1999), Beggs (2005), and more recently, Mertikopoulos and Sandholm (2016)). These results, however, assume the agent has perfect memory, while models involving memory decay has received relatively little attention. Beggs (2005) analyzes a variation of the model involving “forgetting” as a means to speed up convergence to mixed equilibria, with a crucial feature that the
forgetting be performed in a noiseless fashion by weighing distant past experience with a deterministic discount factor. As such, among other differences, the stochastic memory decay model that we adopt exhibits fundamentally different dynamics than the deterministic discount model; for instance, under our model there is a non-negligible probability that all past rewards could be entirely erased, while the deterministic version never truly forgets the past, no matter how distant.

The impact of limited memory to sequential decision-making has been examined in the statistics literature dating back to the seminal work of Cover and Hellman (1970), Hellman and Cover (1970), which proposed algorithms for solving multi-arm bandit and sequential hypothesis testing problems when the decision maker has access to only a finite number of bits of memory. In the queueing theory literature, Mitzenmacher et al. (2002) and Gamarnik et al. (2016) study load balancing problems with space-bounded routing algorithms. In contrast to the present paper, the memories in these models are assumed to be perfect and immune to random erasures.

The dynamic induced in our model is related to the celebrated Pólya’s urn scheme (cf. Pemantle (2007)). In its basic form, one ball is drawn uniformly at random from an urn containing balls of different colors. The chosen ball is then returned to the urn along with a new ball of the same color, and the procedure repeats. Since the probability that a certain color being chosen is proportional to the number of balls of that color present in the urn, by viewing the balls as rewards and colors as actions, our model can be thought of as the urn scheme with the modifications where (a) each ball disappears from the urn after some random amount of time, corresponding to memory loss, and (b) the number of new balls added is random, whose distribution depends on the color, corresponding to the variations in the reward rates. Note that in Pólya’s urn scheme the total number of balls in the urn eventually tends to infinity as the number of draws increases, while in our model the
recallable rewards always stay a finite random variable as a result of memory loss. As such, our model can be viewed as a stationary variation of the urn scheme, which admits a very different dynamic.

On the technical end, a part of our proof relies on a certain fluid model to approximate the evolution of the original, stochastic recallable reward process. While the use of fluid models is a popular technique in the literature of queueing networks (cf. [Kurtz 1970], [Bramson 1998], [Tsitsiklis and Xu 2012], [Massoulié and Xu 2016]), our work departs from conventional fluid models in a notable way: the recallable reward process is not Markovian, and the analysis hence must take into account the dynamics of the auxiliary choice process. It is worth noting that there has been a large-deviation theory developed for obtaining fluid-like approximation results where the transition rates of a continuous-time jump process are modulated by a finite-state auxiliary chain (cf. Theorem 1 of Mitzenmacher et al. (2002), which is based on Theorem 8.15 of Shwartz and Weiss (1995)). Unfortunately, these results do not apply in our model for two reasons. First, they require the transition rate of the jump process of interest to be bounded, a condition not met here due to the aggregate departure rate of the rewards being unbounded over the state space. Second, the existing results typically have one scaling parameter, , whereby time is sped up by a factor of , and space scaled down by a factor of . We use a similar scaling for the recallable reward process, , while the update rate serves as a second scaling parameter that is not captured by the model in Mitzenmacher et al. (2002), Shwartz and Weiss (1995). The addition of is non-trivial: in fact, if is too small, the fluid approximation does not hold, and a different (memory-deficient) regime arises. Due to these considerations, we will develop our fluid approximation results starting from first principles, by using elementary properties of continuous-time martingales and some arguments developed by Kurtz (1978).

4. Proof Overview The remainder of the paper is devoted to the proof of Theorem 2: the proof of Theorem 1 is a straightforward extension of Theorem 2 and will be given in Section 7. We discuss in this section some high-level ideas before delving into the details. Our proof consists of two main components corresponding to the memory-abundant and memory-deficient regimes, respectively. The system dynamics turn out to be quite different depending on the scaling of , and as a result, so are our proof techniques. Fix , and define the joint reward-choice process

\[
W^m(t) = (Q^m(t), C^m(t)), \quad t \in \mathbb{R}_+.
\]  

(13)

It is not difficult to check that the joint process is Markovian, while, for any fixed , the component processes and depend on one another and are not individually Markovian (see Figure 2). However, a key observation is that one of the two component processes becomes asymptotically Markovian as :  

1. (Memory-abundant) If , a large number of choice updates take place before substantial change occurs in the recallable reward process. As a result, the recallable reward process, , becomes a sufficient Markov representation of the system dynamics, as .  
2. (Memory-deficient) If , the interval between two consecutive update points is large. By the time a new choice is to be made, the agent will have “forgotten” nearly all of the rewards associated with the actions other than the most current choice, and the rewards associated with the current choice will have become quite predictable. Here, the choice process , becomes asymptotically Markovian as .

Our proofs for the two regimes will be tailored to the dichotomy outlined above.  

1. For the memory-abundant regime, our analysis focuses on characterizing the scaled recallable reward process, , and uses a certain fluid model to show that, as , the evolution of is well approximated by the solution to a system of ordinary differential equations...
(ODE), and its stationary distribution converges to the unique invariant state of the ODEs. This portion of the proof will be presented in Section 5.

2. For the memory-deficient regime, we turn instead to the choice process, $C^m(\cdot)$, and use its asymptotic Markovian property to obtain an explicit formula for its steady-state distribution, in the limit as $m \to \infty$. We then leverage this formula to calculate the expected steady-state recallable reward using the stationarity properties of the continuous-time processes. The proof for this portion of the theorem is presented in Section 6.

5. The Memory-abundant Regime

We establish the first claim of Theorem 2 in this section, and we will do so by actually proving a stronger statement, that the steady-state distribution of $Q^m_k(\cdot)$ converges in $L^1$ to the corresponding expressions in Eqs. (10) and (11), as $m \to \infty$. As was alluded to earlier, our main tool relies on certain fluid solutions, defined as the solutions to a set of ordinary differential equations, whose dynamics will be shown to closely approximate that of $Q^m(\cdot)$ when $m$ is large. We begin by introducing the fluid solutions and some of their basic properties.

5.1. Fluid Solutions and Their Basic Properties

Define the function $p : \mathbb{R}^K_+ \to \mathbb{R}^K_+$, where

$$p_k(q) = \frac{q_k \vee \alpha_0}{\sum_{i \in K} q_i \vee \alpha_0}, \quad q \in \mathbb{R}^K_+, k \in K. \quad (14)$$

**Definition 1 (Fluid Solutions).** Fix $q^0 \in \mathbb{R}^K_+$. A continuous function, $q : \mathbb{R}_+ \to \mathbb{R}^K_+$, is called a fluid solution with initial condition $q^0$, if it satisfies the following:

1. $q(0) = q^0$;
2. for almost all $t \in \mathbb{R}_+$, the function is differentiable along all coordinates, with

$$\dot{q}_k(t) = \lambda_k p_k(q(t)) - q_k(t), \quad \forall k \in K, \quad (15)$$

where $p(\cdot)$ is defined in Eq. (14).

We say that $q^f \in \mathbb{R}^K_+$ is an invariant state of the fluid solutions, if by setting $q^0 = q^f$, we have that $q(t) = q^f$ for all $t \in \mathbb{R}_+$.

The first term on the right-hand side of Eq. (15) corresponds to the instantaneous arrival rate of the rewards, where $p_k(q(t))$ is the probability that action $k$ is chosen given the current states of recallable rewards, and $\lambda_k$ the reward rate for that action. The second term corresponds to the instantaneous departure rate of rewards, which is proportional to the amount of recallable rewards associated with action $k$. The following lemma states some basic properties of the fluid solutions. The proof is given in Appendix B.1.

**Lemma 1.** For all $q^0 \in \mathbb{R}^K_+$, there exists a unique fluid solution \{q(q^0, t)\}_{t \in \mathbb{R}_+}$ with initial condition $q^0$. Furthermore, for all $t \in \mathbb{R}_+$, $q(q^0, t)$ is a continuous function with respect to $q^0$.

The next result states that the fluid solutions admit a unique invariant state. Note that the expressions of the invariant state coincide with those in Eqs. (10) and (11) in Theorem 2.

**Theorem 3.** Fix $\alpha_0 > 0$. The fluid solutions admit a unique invariant state, $q^f$, whose expressions depend on the value of $\alpha_0$, as follows.

1. Suppose $\alpha_0 \in (0, \lambda_1/K]$. Then,

$$q^f_k = \begin{cases} \lambda_1 - (K - 1)\alpha_0, & k = 1, \\ (\lambda_k/\lambda_1)\alpha_0, & k = 2, \ldots, K. \end{cases} \quad (16)$$

2. Suppose $\alpha_0 > \lambda_1/K$. Then,

$$q^f_k = \frac{\lambda_k}{K}, \quad k = 1, \ldots, K. \quad (17)$$
Proof. It can be easily verified that the expressions in Eqs. (16) and (17) are valid invariant states in their respective regimes of $\alpha_0$ values. We next show that they are indeed the unique invariant states in both cases. Let $q^i$ be an invariant state of the fluid solutions. Then, $q^i$ must satisfy:

$$\lambda_k \frac{q^i_k \vee \alpha_0}{\sum_{i \in \mathcal{K}} q^i_k \vee \alpha_0} - q^i_k = 0, \quad \forall k \in \mathcal{K}. \quad (18)$$

Define a partition of $\mathcal{K}$ into $\mathcal{K}^+$ and $\mathcal{K}^-$, where:

$$\mathcal{K}^+ = \{k \in \mathcal{K} : q^i_k > \alpha_0\}, \quad \mathcal{K}^- = \{k \in \mathcal{K} : q^i_k \leq \alpha_0\}. \quad (19)$$

Note that by definition,

$$q^i_k \vee \alpha_0 = q^i_k, \quad k \in \mathcal{K}^+, \quad (20)$$

$$q^i_k \vee \alpha_0 = \alpha_0, \quad k \in \mathcal{K}^- \quad (21)$$

Suppose that $\alpha_0 \in (0, \lambda_1/K]$. We next show that in this case $\mathcal{K}^+$ is the singleton set, $\{1\}$. First, suppose that $\mathcal{K}^+$ is empty, that is, $q^i_k \leq \alpha_0$ for all $k \in \mathcal{K}$. We have that

$$q^i_1 = \lambda_1 \frac{\alpha_0}{\sum_{i \in \mathcal{K}^+} q^i_1 + \alpha_0 |\mathcal{K}^-|} = \lambda_1 \frac{\alpha_0}{K \alpha_0} = \frac{\lambda_1}{K} > \alpha_0, \quad (22)$$

where the last inequality follows from the assumption that $\alpha_0 < \lambda_1/K$. This leads to a contradiction, implying that $\mathcal{K}^+ \neq \emptyset$. Fix $k \in \mathcal{K}^+$. Combining Eqs. (18), (20) and (21), we have that

$$\lambda_k \frac{q^i_k}{\sum_{i \in \mathcal{K}^+} q^i_k + \alpha_0 |\mathcal{K}^-|} = q^i_k, \quad (23)$$

which, after rearrangement, can be written as

$$\sum_{i \in \mathcal{K}^+} q^i_k = \lambda_k - \alpha_0 |\mathcal{K}^-|, \quad (24)$$

Note that in Eq. (24) only the term $\lambda_k$ on the right-hand depends on $k$. This implies that, if $\mathcal{K}^+ \neq \emptyset$, then all coordinates in $\mathcal{K}^+$ must have the same arrival rate of rewards. Thus, we can assume that there exists $\bar{\lambda} \in \mathbb{R}_+$, such that $\lambda_k = \bar{\lambda}$, for all $k \in \mathcal{K}^+$, and it suffices to show that $\bar{\lambda} = \lambda_1$.

For the sake of contradiction, suppose that $\bar{\lambda} \neq \lambda_1$, and hence $\lambda_1 > \bar{\lambda}$. This implies that $1 \notin \mathcal{K}^+$, and hence $q^i_1 \leq \alpha_0$. Invoking Eq. (18) again, we have that

$$q^i_1 = \lambda_1 \frac{\alpha_0}{\sum_{i \in \mathcal{K}^+} q^i_1 + \alpha_0 |\mathcal{K}^-|} \overset{(a)}{=} \alpha_0 \frac{\lambda_1}{\bar{\lambda}} \overset{(b)}{>} \alpha_0, \quad (25)$$

where step (a) follows from Eq. (24), and (b) from the fact that $\lambda_1 > \bar{\lambda}$. Clearly, Eq. (25) contradicts with the fact that $q^i_1 \leq \alpha_0$. We thus conclude that $\bar{\lambda} = \lambda_1$, $\mathcal{K}^+ = \{1\}$, and $\mathcal{K}^- = \{2, \ldots, K\}$. By setting $k = 1$ in Eq. (24), we further conclude that

$$q^i_1 = \sum_{i \in \mathcal{K}^+} q^i_1 = \lambda_1 - \alpha_0 |\mathcal{K}^-| = \lambda_1 - (K - 1) \alpha_0. \quad (26)$$

Now, fix $k \in \{2, \ldots, K\}$. We have, from Eq. (18), that

$$q^i_k = \lambda_k \frac{\alpha_0}{\sum_{i \in \mathcal{K}^+} q^i_1 + \alpha_0 |\mathcal{K}^-|} = \alpha_0 \frac{\lambda_k}{\lambda_1}, \quad (27)$$

which completes the proof.
where the last equality follows from the fact that
\[
\left( \sum_{i \in K^+} q_i^f \right) + \alpha_0 |K^-| = q_1^f + (K - 1) \alpha_0 = [\lambda_1 - (K - 1) \alpha_0] + (K - 1) \alpha_0 = \lambda_1.
\] (28)

Combining Eqs. (26) and (27), we have shown that the expressions in Eq. (18) are the unique invariant state of the fluid model, when \( \alpha_0 \in (0, \lambda_1/K] \).

We turn next to the second case of Theorem 3 and assume that \( \alpha_0 > \lambda_1/K \). First, suppose that \( K^+ \neq \emptyset \). By the arguments leading to Eq. (25), which do not depend on the value of \( \alpha_0 \), we know that \( K^+ = \{1\} \), and
\[
q_1^f = \lambda_1 - (K - 1) \alpha_0 \leq \lambda_1 - (1 - 1/K) \lambda_1 = \frac{\lambda_1}{K} \leq \alpha_0,
\] (29)
where both inequalities follow from the assumption that \( \alpha_0 > \lambda_1/K \). This contradicts with the assumption that \( 1 \in K^+ \) and hence \( q_1^f > \alpha_0 \), and we conclude that \( K^+ \) must be empty. This implies that \( q_k^f \leq \alpha_0 \) for all \( k \in K \), and by Eq. (18), we have that
\[
q_k^f = \frac{\alpha_0}{\alpha_0 K} \lambda_k K, \quad k = 1, \ldots, K.
\] (30)

This completes the proof of Theorem 3. Q.E.D.

The following theorem is the main result of this section, which would imply the first claim of Theorem 2.

**Theorem 4.** Fix \( m \in \mathbb{N} \). The process \( Q^m(\cdot) \) admits a unique steady-state distribution, denoted by \( \pi^m \). Suppose that \( \beta_m \gg 1/m \) as \( m \to \infty \). We have that
\[
\pi^m \text{ converges weakly to } \delta_{q^f}, \quad \text{as } m \to \infty,
\] (31)
where \( q^f \) is the unique invariant state of the fluid model, and \( \delta_{q^f} \) is the probability measure with unit mass on \( q^f \). Furthermore, we have that
\[
\lim_{m \to \infty} \mathbb{E}\left( |Q^m_k(\infty) - q_k^f| \right) = 0, \quad \forall k \in K.
\] (32)

The remainder of Section 5 is devoted to showing Theorem 4 and consists of three main parts, as illustrated in Figure 6.

1. **Theorem 5 of Section 5.2** shows that the scaled recallable-reward process converges uniformly to a fluid solution over any finite time horizon, as \( m \to \infty \). This is the more technical portion of the proof, and the key idea is to use a quasi-fluid process, one where the evolution of the reward process is deterministic while the choice process remains random, as an intermediary to bridge the gap between the fluid solution and the original stochastic process.

2. **Theorem 6 of Section 5.3** shows that starting from any bounded initial condition, \( q^0 \), the fluid solution \( q(q^0, t) \) converges exponentially fast to the unique invariant state, as \( t \to \infty \). The proof is centered around the evolution of a simple piece-wise linear potential function, which ensures that \( q(q^0, t) \) rapidly enters a suitable subset of the state space, from which point exponential convergence takes place.

3. Finally, in Section 5.4, by combining Theorems 5 and 6, we show that the sequence of steady-state scaled recallable rewards converges in \( L_1 \) to a unique limiting point concentrated on the invariant state of the fluid solution, completing the proof of Theorem 4.
5.2. Convergence to the Fluid Solution over Finite Horizon  

In this subsection, we show that, with high probability, the scaled recallable reward process converges to the fluid solution uniformly over any finite horizon, as $m \to \infty$, as summarized in the following theorem (See plots (a) through (b) of Figure 5 for a sample-path example of how this convergence takes place.)

**Theorem 5.** Suppose that $\beta_m \gg 1/m$, as $m \to \infty$. Fix $q^0 \in \mathbb{R}^K_+$, and suppose that
\[ \lim_{m \to \infty} P(\|Q^m_0 - q^0\| > \epsilon) = 0, \quad \forall \epsilon > 0. \] (33)

Let $q(q^0, \cdot)$ be the unique fluid solution with initial condition $q^0$, as defined in Lemma 1. We have that, for all $T > 0$,
\[ \lim_{m \to \infty} P\left( \sup_{0 \leq t \leq T} \|Q^m(t) - q(q^0, t)\| > \epsilon \right) = 0, \quad \forall \epsilon > 0. \] (34)

A main challenge in proving Theorem 5 is that the processes $Q^m(\cdot)$ and $C^m(\cdot)$ interact in an intricate way: the recallable reward vector influences how the subsequent choices are chosen, while the current choice in turn dictates the arrival rate of new rewards associated with each action. The main idea of the proof is to “disentangle” their interactions by means of an intermediate, quasi-fluid process, where the dynamics of the arrivals and departures of rewards becomes deterministic and “fluid-like”, while the choice process remains random.

We now construct the quasi-fluid process. Fix $q \in \mathbb{R}^K_+$ and $c \in \mathcal{K}$, and define
\[ G_k(q,c) = \lambda_k \mathbb{I}(c = k) - q_k, \quad k \in \mathcal{K}. \] (35)

In words, $G_k(q,c)$ represents the rate of change in recallable rewards associated with action $k$ when the current recallable reward vector and action are $q$ and $c$, respectively. Define the scaled action process:
\[ \bar{C}^m(t) = C^m(mt), \quad t \in \mathbb{R}_+. \] (36)

Note that, unlike $\bar{Q}^m(\cdot)$, the scaling of $\bar{C}^m(\cdot)$ only occurs in time and not in space. We then construct a quasi-fluid solution, $\{V^m(t)\}_{t \in \mathbb{R}_+}$, by letting
\[ V^m_k(t) = q^0_k + \int_0^t G_k(V^m(s), \bar{C}^m(s)) ds, \quad k \in \mathcal{K}, t \in \mathbb{R}_+. \] (37)

In words, $V^m(\cdot)$ corresponds to a system in which recallable rewards are governed by deterministic arrival and departure processes, as specified by Eq. (36), and where the choice process is scaled in time by a factor of $m$, but remains stochastic.

We note some basic properties of $V^m_k(\cdot)$ that will become useful. First, since $C^m(\cdot)$ is a Markov jump process and the function $G_k(\cdot, \cdot)$ uniformly Lipschitz continuous, analogously to Lemma 1.
it is not difficult to show that $V^m(\cdot)$, expressed as a solution to the above integral equation, is uniquely defined almost surely. Also, it is not difficult to verify that

$$0 \leq V^m_k(t) \leq q^0_k + \lambda_k t, \quad \forall t \in \mathbb{R}_+,$$

where the right-hand side of the second inequality corresponds to the scenario where $C^m_k(s) = k$ for all $s \in [0,t]$ and the negative drift term ($-q_k$) is absent from the expression of $G_k(q,c)$.

The proof of Theorem 5 proceeds in two main steps. We first show that $\{Q^m_t\}_{t \geq 0}$ converges to the quasi-fluid process, $\{V(t)\}_{t \geq 0}$, as $m \to \infty$, as is summarized in the following proposition.

**Proposition 1.** Fix $q^0 \in \mathbb{R}^K_+$, and suppose that

$$\lim_{m \to \infty} \mathbb{P}(\|Q^m(0) - q^0\| > \epsilon) = 0, \quad \forall \epsilon > 0. \tag{39}$$

Then,

$$\lim_{m \to \infty} \mathbb{P}\left(\sup_{0 \leq t \leq T} \left\|Q^m(t) - V^m(t)\right\| > \epsilon\right) = 0, \quad \forall \epsilon > 0. \tag{40}$$

We then show that, when $\beta_m \gg 1/m$, the process $\{V^m(t)\}_{t \geq 0}$ converges to the fluid solution, $\{q(q^0,t)\}_{t \geq 0}$:

**Proposition 2.** Suppose, in addition to the condition of Eq. (39) in Proposition 1, that $m\beta_m \to \infty$, as $m \to \infty$. For all $\epsilon > 0$,

$$\lim_{m \to \infty} \mathbb{P}\left(\sup_{0 \leq t \leq T} \left\|V^m(t) - q(q^0,t)\right\| > \epsilon\right) = 0.$$

Note that Proposition 1 holds independently of the scaling of $\beta_m$, while Proposition 2 requires that the system be in the memory-abundant regime. Together, Propositions 1 and 2 imply Theorem 5.

The proof for Propositions 1 and 2 will be given in Sections 5.2.1 and 5.2.2, respectively.

**Remark.** We believe that there could be an alternative approach to establishing Theorem 5 by analyzing the limiting behavior of a discrete-time embedded Markov chain of the reward process $Q(\cdot)$ using stochastic approximation theory, and yet there are non-trivial difficulties along this direction. In particular, let $S_i$ be the $i$th update point, then the embedded reward process $\{Q[i] \triangleq Q(S_i)\}_{i \in \mathbb{N}}$ can be shown to be a Markov chain, even though the original continuous-time process $Q(t)$ is not Markovian. One may then hope to invoke stochastic approximation theory to demonstrate that the a scaled version of $Q[i]$ converges to a fluid solution, such as the approach taken in Beggs (2005). There are however several difficulties. First, the jumps in $Q[i]$ can be unbounded due to the Poisson reward structure, while most standard results in the stochastic approximate literature requires bounded updates between adjacent steps (e.g., Theorem 1 of Benveniste et al. (2012)); it is likely possible to remove this requirement but it will require a more delicate argument. Second, even assuming such convergence has been established, it would only apply to the update points themselves, and to obtain the type of uniform convergence over an entire continuous time interval in the form of Theorem 5 would require additional procedures to bound the maximal fluctuation of the reward process between event points. In light of these observations, the authors tend to believe that the approach we adopt in this paper is more direct and easier to carry out.

### 5.2.1. Proof of Proposition 1

The proof of Proposition 1 is largely based on maximal inequalities for continuous-time martingales, and the main steps follow the approach developed in Kurtz (1978). Define $X^m(t) \in \mathbb{R}^K_+$, where

$$X^m_k(t) = Q^m_k(t) - \left(\frac{Q^m_k(0)}{K} + \int_0^t G_k(Q^m(s),C^m(s))ds\right), \quad t \in \mathbb{R}_+, k \in K.$$  \tag{41}
The proof proceeds in two stages: we first show, via Gronwall's lemma (Proposition 6 in Appendix A), that bounding the deviation of $\overline{Q}_k^m(\cdot)$ from $V_k^m(\cdot)$ can be reduced to bounding the magnitude of $X_k^m(\cdot)$. We then show that this can be accomplished by writing $X_k^m(\cdot)$ as a sum of two continuous-time martingales. From the definition of $\{V_m(t)\}_{t \geq 0}$, we have that

$$
\| \overline{Q}_k^m(t) - V_m^m(t) \| = \sum_{k \in K} | \overline{Q}_k^m(t) - q_k^0 - \int_0^t G_k(V_m(s), \overline{C}_m(s)) ds |
$$

$$
\leq \sum_{k \in K} | \overline{Q}_k^m(0) - q_k^0 | + \sum_{k \in K} \left| X_k^m(t) + \int_0^t \left( G_k(\overline{Q}_k^m(s), \overline{C}_m(s)) - G_k(V_m(s), \overline{C}_m(s)) \right) ds \right|
$$

$$
\leq \left( \left\| \overline{Q}_k^m(0) - q^0 \right\| + \left\| X_m^m(t) \right\| \right) + \int_0^t \sum_{k \in K} \left| G_k(\overline{Q}_k^m(s), \overline{C}_m(s)) - G_k(V_m(s), \overline{C}_m(s)) \right| ds
$$

$$
\leq \left( \left\| \overline{Q}_k^m(0) - q^0 \right\| + \left\| X_m^m(t) \right\| \right) + \int_0^t \left| \overline{Q}_k^m(s) - V_k^m(s) \right| ds
$$

$$
= \left( \left\| \overline{Q}_k^m(0) - q^0 \right\| + \left\| X_m^m(t) \right\| \right) + \int_0^t \left| \overline{Q}_k^m(s) - V_m^m(s) \right| ds
$$

(42)

For step (a), we observe that for all $c, k \in K$, $G_k(\cdot, c)$ is a 1-Lipschitz continuous function. We now apply Gronwall’s lemma (Proposition 6 in Appendix A) to Eq. (42), with $a(t) = \left\| \overline{Q}_k^m(0) - q^0 \right\| + \left\| X_m^m(t) \right\|$, $b(t) = 1$, and $f(t) = \left\| \overline{Q}_k^m(t) - V_m^m(t) \right\|$, and obtain that

$$
P \left( \sup_{0 \leq t \leq T} \left\| \overline{Q}_k^m(t) - V_m^m(t) \right\| > \epsilon \right) \leq P \left( \sup_{0 \leq t \leq T} \left( \left\| \overline{Q}_k^m(0) - q^0 \right\| + \left\| X_m^m(t) \right\| \right) e^t > \epsilon \right)
$$

$$
\leq P \left( \left\| \overline{Q}_k^m(0) - q^0 \right\| > \epsilon e^{-T} \right) + P \left( \sup_{0 \leq t \leq T} \left\| X_m^m(t) \right\| > \epsilon e^{-T} \right). \quad (43)
$$

Because we have assumed that $\lim_{m \to \infty} P \left( \left\| \overline{Q}_k^m(0) - q^0 \right\| > \epsilon \right) = 0$ for all $\epsilon > 0$, to prove Lemma 1, it therefore suffices to show that

$$
\lim_{m \to \infty} P \left( \sup_{0 \leq t \leq T} \left\| X_m^m(t) \right\| > \epsilon \right) = 0, \quad \forall \epsilon > 0. \quad (44)
$$

We now prove Eq. (44) by expressing $X_k^m(\cdot)$ as the sum of two continuous-time martingales corresponding to the arrivals and departures of rewards, respectively. Fix $t \in \mathbb{R}_+$ and $k \in K$, and denote by $A_k^m(t)$ and $D_k^m(t)$ the amount of rewards associated with action $k$ that have arrived and departed, respectively, during the interval $[0, t]$. In particular, we can write

$$
Q_k^m(t) = Q_k^m(0) + A_k^m(t) - D_k^m(t), \quad t \in \mathbb{R}_+.
$$

(45)

We have that

$$
|X_k^m(t)| = \left| \frac{1}{m} (A_k^m(mt) - D_k^m(mt)) - \int_0^t G_k(\overline{Q}_k^m(s), \overline{C}_m(s)) ds \right|
$$

$$
= \left| \frac{1}{m} \left( A_k^m(mt) - \int_0^{mt} \lambda_k \mathbb{I}(C^m(s) = k) ds \right) - \frac{1}{m} \left( D_k^m(mt) - \int_0^{mt} Q_k^m(s) m^{-1} ds \right) \right|
$$

$$
\leq \left| \frac{1}{m} \left( A_k^m(mt) - \int_0^{mt} \lambda_k \mathbb{I}(C^m(s) = k) ds \right) \right| + \left| \frac{1}{m} \left( D_k^m(mt) - \int_0^{mt} Q_k^m(s) m^{-1} ds \right) \right|. \quad (46)
$$
For the remainder of the proof, we will focus on showing that, for all \( \epsilon > 0 \),
\[
\lim_{m \to \infty} \mathbb{P} \left( \sup_{0 \leq t \leq T} \frac{1}{m} \left( A^m_k(mt) - \int_0^{mt} \lambda_k(C^m(s) = k) \, ds \right) > \epsilon \right) = 0, \quad \text{and} \quad \lim_{m \to \infty} \mathbb{P} \left( \sup_{0 \leq t \leq T} \frac{1}{m} \left( D^m_k(mt) - \int_0^{mt} Q^m_k(s)m^{-1} \, ds \right) > \epsilon \right) = 0. \tag{48}
\]

In light of Eq. (46), the above two equations imply the validity of Eq. (44), which proves Proposition 1. The proofs for both Eqs. (47) and (48) hinge upon the following technical result, which states that the sample path of a certain time-inhomogenous Poisson process stays close to its mean, with high probability. The proof of the lemma is given in Appendix B.2 and uses a similar line of argument as that of Theorem 2.2 in Kurtz (1978).

**Lemma 2.** Fix \( l \in \mathbb{Z}^{K+1} \) and \( T \in \mathbb{R}_+ \). Let \( \{N(t)\}_{t \geq 0} \) be the counting process where \( N(t) \) is the number of times in \([0, t]\) that the process \( W^m(\cdot) \) jumps from state \( w \) to \( w + l \), for some \( w \in \mathbb{Z}^{K+1} \). Denote by \( \psi : \mathbb{Z}^{K+1} \to \mathbb{R}_+ \) the corresponding rate function of \( N(\cdot) \), so that the instantaneous transition rate of \( N(\cdot) \) at time \( t \) is equal to \( \psi(W^m(t)) \). For all \( \epsilon, \phi > 0 \), we have that
\[
\mathbb{P} \left( \sup_{0 \leq t \leq T} \left| N(t) - \int_0^t \mu(W^m(s)) \, ds \right| \geq \epsilon \right) \leq 2e^{-\phi T \cdot h(\epsilon/\phi T)} + \mathbb{P} \left( \sup_{0 \leq t \leq T} \psi(W^m(t)) > \phi \right), \tag{49}
\]
where \( h(x) = (1 + x) \log(1 + x) - x \).

We now prove Eqs. (47) and (48). In the context of Lemma 2, \( A_k(\cdot) \) is the counting process with \( l = e_k \), where \( e_k \) is the vector \( (K + 1) \times 1 \) vector whose \( k \)th coordinate is 1 and all other coordinates zero, corresponding to an arrival to \( Q^m_k(\cdot) \). The rate of \( A_k(\cdot) \) at time \( t \) is equal to \( \lambda_k(C^m(t) = k) \), which is bounded from above by \( \lambda_k \) for all \( t \in \mathbb{R}_+ \). By applying Lemma 2 with \( \psi(W^m(t)) = \lambda_k(C^m(t) = k) \) and \( \phi = \lambda_k \), we have that \( \mathbb{P} \left( \sup_{0 \leq t \leq T} \psi(W^m(t)) > \phi \right) = 0 \), and for all \( \epsilon > 0 \),
\[
\lim_{m \to \infty} \mathbb{P} \left( \sup_{0 \leq t \leq T} \frac{1}{m} \left( A^m_k(mt) - \int_0^{mt} \lambda_k(C^m(s) = k) \, ds \right) \geq \epsilon \right) = \lim_{m \to \infty} \mathbb{P} \left( \sup_{0 \leq t \leq mT} A^m_k(t) - \int_0^t \lambda_k(C^m(s) = k) \, ds \geq m \epsilon \right) \leq \lim_{m \to \infty} 2 \exp \left( - \lambda_k mT \cdot h \left( \frac{\epsilon}{\lambda_k T} \right) \right) = 0. \tag{50}
\]

The proof of Eq. (48) uses essentially the same idea, but the argument needs to be more delicate due to the fact that the rate of the counting process \( D^m_k(\cdot) \), which is equal to \( Q^m_k(\cdot)m^{-1} \), is not bounded over the state space. Therefore, we first derive an upper bound on the tail probabilities of \( Q^m_k(\cdot) \), as follows. Let \( \psi(\cdot) \) be the rate function for \( D_k(\cdot) \) as in Lemma 2 and \( \phi = 2q^0_k + (1 + \epsilon)\lambda_k T \).

Fixing \( \epsilon > 0 \), we have that, for all \( r \in \mathbb{Z}_+ \),
\[
\mathbb{P} \left( \sup_{0 \leq t \leq mT} \psi(W^m(t)) \geq \phi \right) = \mathbb{P} \left( \sup_{0 \leq t \leq mT} Q^m_k(t) m^{-1} \geq \phi \right) \\
\leq \mathbb{P}(Q^m_k(0) \geq 2mq^0_k) + \max_{r = 1, \ldots, 2mq^0_k} \mathbb{P} \left( \sup_{0 \leq t \leq mT} Q^m_k(t) \geq m\phi \mid Q^m(0) = r \right) \\
= \mathbb{P}(Q^m_k(0) \geq 2mq^0_k) + \max_{r = 1, \ldots, 2mq^0_k} \mathbb{P} \left( \sup_{0 \leq t \leq mT} Q^m_k(t) - Q^m_k \geq m\phi - r \mid Q^m(0) = r \right) \\
\leq \mathbb{P}(Q^m_k(0) \geq 2mq^0_k) + \max_{r = 1, \ldots, 2mq^0_k} \mathbb{P} \left( \sup_{0 \leq t \leq mT} A^m_k(t) \geq m\phi - r \mid Q^m(0) = r \right). 
\]
\[
\begin{align*}
\{a\} & \leq \mathbb{P} (Q^m_k(0) \geq 2mq^0_k) + \mathbb{P} \left( \sup_{0 \leq t \leq mT} A^m_k(t) \geq (1 + \epsilon) m\lambda T \right) \\
\leq & \mathbb{P} (Q^m_k(0) \geq 2mq^0_k) + \mathbb{P} \left( \sup_{0 \leq t \leq mT} \left(A^m_k(t) - \int_0^t \lambda_k (C^m(s) = k) ds \right) \geq m[(1 + \epsilon)\lambda T - \lambda_k T] \right) \\
\leq & \mathbb{P} (m^{-1}Q^m_0(0) \geq 2q^0_k) + 2 \exp(-\lambda_k mT \cdot h(\epsilon)),
\end{align*}
\]

where step (a) follows from the fact that \( A^m_k(\cdot) \) is independent of \( Q^m_k(0) \), and hence the maximum in the second term is attained by setting \( T = 2mq^0_k \). Step (b) follows from \( \lambda_k (C^m(s) = k) \leq \lambda_k \) for all \( t \in \mathbb{R}_+ \), and (c) from the inequality in Eq. (50), by replacing \( \epsilon \) with \( \epsilon \lambda_k T \).

We are now ready to establish Eq. (48):

\[
\mathbb{P} \left( \sup_{0 \leq t \leq T} \left| \frac{1}{m} D^m_k(mt) - \int_0^{mt} Q^m_k(s) m^{-1} ds \right| \geq \epsilon \right) = \mathbb{P} \left( \sup_{0 \leq t \leq mT} \left| D^m_k(t) - \int_0^t Q^m_k(s) m^{-1} ds \right| \geq me \right)
\]

\\[\leq 2 \exp \left( -\phi mT \cdot h \left( \frac{\epsilon}{\phi T} \right) \right) + \mathbb{P} \left( \sup_{0 \leq t \leq mT} \psi(W^m(t)) \geq \phi \right)\]

\[
\leq 2 \exp \left( -\phi mT \cdot h \left( \frac{\epsilon}{\phi T} \right) \right) + 2 \exp(-\lambda mT \cdot h(\epsilon)) + \mathbb{P} (m^{-1}Q^m_0(0) \geq 2q^0_k),
\]

where step (a) follows from Lemma [2] and (b) from Eq. (51). Note that \( \phi, \epsilon \) and \( T \) are positive constants, and by our assumption, \( \lim_{m \to \infty} \mathbb{P} \left( |m^{-1}Q^m_0(0) - q^0_k| > \delta \right) = 0 \) for all \( \delta > 0 \). Therefore, Eq. (48) follows by taking the limit in the above inequality as \( m \to \infty \). This completes the proof of Proposition [1].

### 5.2.2. Proof of Proposition 2

For the purpose of this proof, we will use an alternative representation of the fluid solutions using integral equations. Define the drift function, \( F : \mathbb{R}_+^{\mathcal{K}} \to \mathbb{R}_+^{\mathcal{K}} \):

\[
F_k(q) = \lambda_k p_k(q) - q_k, \quad q \in \mathcal{K},
\]

where \( p(\cdot) \) is defined in Eq. (44). It can be verified from the definition of \( p(\cdot) \) that there exists \( l_F \in \mathbb{R}_+ \) such that \( F_k(\cdot) \) is \( l_F \)-Lipschitz continuous for all \( k \in \mathcal{K} \). Fix \( q^0 \in \mathbb{R}_+^{\mathcal{K}} \), let \( \{q(t)\}_{t \in \mathbb{R}_+} \), \( q(t) \in \mathbb{R}_+^{\mathcal{K}} \), be a solution to the following integral equation:

\[
q_k(t) = q^0_k + \int_0^t F_k(q(s)) ds, \quad k \in \mathcal{K}.
\]

Similar to Lemma [1], it is not difficult to show that the function \( q(\cdot) \) defined in Eq. (54) exists, is unique, and coincides with the fluid solution with initial condition \( q^0 \). We have that

\[
\|V^m(t) - q(t)\| = \sum_{k \in \mathcal{K}} \left| \int_0^t G_k(V^m(s), \overline{C}^m(s)) - F_k(q(s)) ds \right|
\]

\[
\leq \sum_{k \in \mathcal{K}} \left| \int_0^t G_k(V^m(s), \overline{C}^m(s)) - F_k(V^m(s)) ds \right| + \left| \int_0^t F(V^m(s)) - F(q(s)) ds \right|
\]

\[
\leq \sum_{k \in \mathcal{K}} \left| \int_0^t G_k(V^m(s), \overline{C}^m(s)) - F_k(V^m(s)) ds \right| + l_F \int_0^t \|V^m(s) - q(s)\| ds,
\]

where the last inequalities come from the fact that \( F_k(\cdot) \) is \( l_F \)-Lipschitz continuous for all \( k \in \mathcal{K} \). In a manner analogous to Eq. (42) from the proof of Proposition [1], by applying Gronwall’s lemma (Proposition [6] in Appendix A), we have that, for all \( \epsilon > 0 \),

\[
\mathbb{P} \left( \sup_{0 \leq t \leq T} \|V^m(t) - q(t)\| > \epsilon \right) \leq \mathbb{P} \left( \sup_{0 \leq t \leq T} \sum_{k \in \mathcal{K}} \left| \int_0^t G_k(V^m(s), \overline{C}^m(s)) - F_k(V^m(s)) ds \right| > \epsilon e^{-l_FT} \right).
\]
Therefore, in order to establish Proposition 2, it suffices to show that

\[
\lim_{m \to \infty} P \left( \sup_{0 \leq t \leq T} \sum_{k \in K} \left| \int_0^t G_k(V^m(s), \overline{C}^m(s)) - F_k(V^m(s)) \, ds \right| > \epsilon \right) = 0, \quad \forall \epsilon > 0. \quad (57)
\]

In what follows, we will show (57) by using the discrete-time embedded process of \( C^m(\cdot) \) and analyzing the system dynamics at the times when new choices are chosen. We begin by introducing some notation. Fixing \( i \in \mathbb{Z}_+ \), we denote by \( S_i \) the \( i \)th update point, i.e., time of the \( i \)th update in \( C^m(\cdot) \), with \( S_0 \triangleq 0 \), and

\[
\tau^m_i = m^{-1}(S_i - S_{i-1}), \quad i \in \mathbb{N}. \quad (58)
\]

Note that \( \{\tau_i\}_{i \in \mathbb{N}} \) are independent exponential random variables with mean \( (m \beta_m)^{-1} \). Let \( \{\overline{Q}^m[i]\}_{i \in \mathbb{N}} \) be the discrete-time process, where \( \overline{Q}^m[i] \) corresponds to the value of \( \overline{Q}^m(\cdot) \) immediately following the \( i \)th update point:

\[
\overline{Q}^m_k[i] = m^{-1}Q^m_k(S_i) = \overline{Q}^m_k(S_i/m), \quad k \in \mathcal{K}, i \in \mathbb{N}. \quad (59)
\]

and \( \{Z^m_k[i]\}_{i \in \mathbb{N}} \) be the process of indicator variables:

\[
Z^m_k[i] = \mathbb{I}(C^m(S_i) = k), \quad k \in \mathcal{K}, i \in \mathbb{N}. \quad (60)
\]

That is, \( Z^m_k[i] = 1 \) if action \( k \) is selected on the \( i \)th update point. Finally, let \( \{\hat{Q}^m(t)\}_{t \geq 0} \) be a right-continuous piece-wise constant process which coincides with \( \overline{Q}^m(\cdot) \) at the points \( \{S_i/m\}_{i \in \mathbb{N}} \):

\[
\hat{Q}^m(t) = \overline{Q}^m[i], \quad \forall t \in [S_i/m, S_{i+1}/m). \quad (61)
\]

Fix \( k \in \mathcal{K} \), and denote by \( I_t \) the number of updates in \( C^m(\cdot) \) by time \( t \), i.e.,

\[
I_t = \max\{i : S_i \leq t\}. \quad (62)
\]

By the triangle inequality, we have that

\[
\begin{align*}
\lambda_k^{-1} \sup_{0 \leq t \leq T} \left| \int_0^t G_k(V^m(s), \overline{C}^m(s)) - F_k(V^m(s)) \, ds \right| \\
= \lambda_k^{-1} \sup_{0 \leq t \leq T} \left| \int_0^t \lambda_k \left( \mathbb{I}(C^m(s) = k) - p_k(V^m(s)) \right) \, ds \right| \\
\leq \sup_{0 \leq t \leq T} \left[ \int_0^T \left( \mathbb{I}(C^m(s) = k) - p_k(\hat{Q}^m(s)) \right) \, ds + \int_0^t \left( p_k(\hat{Q}^m(s)) - p_k(V^m(s)) \right) \, ds \right] \\
= \sup_{0 \leq t \leq T} \left[ \sum_{i=1}^{I_{t+1}} \tau^m_i \left( Z^m_k[i] - p_k(\overline{Q}^m[i]) \right) + \int_0^t \left( p_k(\hat{Q}^m(s)) - p_k(V^m(s)) \right) \, ds \right] \\
\leq \sup_{0 \leq t \leq T} \sum_{i=1}^{I_{t+1}} \tau^m_i \left( Z^m_k[i] - p_k(\overline{Q}^m[i]) \right) + \sup_{0 \leq t \leq T} \int_0^t \left( p_k(\hat{Q}^m(s)) - p_k(V^m(s)) \right) \, ds \quad (63)
\end{align*}
\]

We now derive upper bounds on tail probabilities for each term on the right-hand side of Eq. (63). For the first term, define

\[
\xi^n_m = \sum_{i=1}^{n} \tau^m_i \left( Z^m_k[i] - p_k(\overline{Q}^m[i]) \right), \quad n \in \mathbb{N}. \quad (64)
\]

Fix \( m \in \mathbb{N} \) and \( \epsilon > 0 \). Let

\[
J^\epsilon_m = (1 + \epsilon)m\beta_m T. \quad (65)
\]
(To avoid the excessive use of floors and ceilings, we assume that $J^m$, as well as the number of summands in the subsequent Lemma 3 are positive integers. The results extend easily to the general case.) Define the events

$$\mathcal{A}_\epsilon^m = \{I_m T < J^m\}, \quad \text{and} \quad \mathcal{B}_\epsilon^m = \left\{ \max_{0 \leq n \leq J^m} |\xi_n^m| \leq \epsilon \right\}. \quad (66)$$

We have that

$$\mathbb{P} \left( \sup_{0 \leq t \leq T} \left| \sum_{i=1}^{I_m T} \tau_i^m \left( Z_k^m[i] - p_k(\overline{Q}^m[i]) \right) \right| \geq \epsilon \right) = \mathbb{P} \left( \max_{0 \leq n \leq I_m T} |\xi_n^m| \geq \epsilon \right) \leq 1 - \mathbb{P} (\mathcal{A}_\epsilon^m \cap \mathcal{B}_\epsilon^m) \leq (1 - \mathbb{P}(\mathcal{A}_\epsilon^m)) + (1 - \mathbb{P}(\mathcal{B}_\epsilon^m)). \quad (67)$$

With the above equation in mind, we now proceed to demonstrate that both $\mathbb{P}(\mathcal{A}_\epsilon^m)$ and $\mathbb{P}(\mathcal{B}_\epsilon^m)$ are diminishingly small in the limit as $m \to \infty$. The following lemma is useful, whose proof is a straightforward application of Chebyshev’s inequality and is given in Appendix B.3.

**Lemma 3.** Fix $r, \epsilon$ and $y > 0$. Let $\{X_i\}_{i \in \mathbb{N}}$ be a sequence of independent and identically distributed exponentially variables with mean $r^{-1}$. Then,

$$\mathbb{P} \left( \sum_{i=1}^{(1+r)\epsilon y} X_i \leq y \right) \leq \frac{1+\epsilon}{ry \epsilon^2}. \quad (68)$$

Recall that the $\tau_i^m$’s are i.i.d. exponential random variables with mean $(m \beta_m)^{-1}$. Applying Lemma 3 with $r = m \beta_m$ and $y = T$, we have that

$$\mathbb{P}(\mathcal{A}_\epsilon^m) = \mathbb{P}(I_m T < (1+\epsilon)m \beta_m T) = 1 - \mathbb{P} \left( \sum_{i=1}^{(1+\epsilon)m \beta_m T} \tau_i^m \leq T \right) \geq 1 - \frac{1+\epsilon}{T \epsilon^2 (m \beta_m)^{-1}}. \quad (69)$$

We next turn to the value of $\mathbb{P}(\mathcal{B}_\epsilon^m)$. Recall from the definition of $Z_k^m[i]$ that

$$\mathbb{E}(Z_k^m[i] | \overline{Q}^m[i]) = p_k(\overline{Q}^m[i]). \quad (70)$$

It is therefore not difficult to verify that $\{\xi_n^m\}_{n \in \mathbb{N}}$ is a martingale, and our objective would be to derive an upper bound on its maximum upward excursion over $\{0, \ldots, J^m\}$. Unfortunately, we cannot apply the Azuma-Hoeffding inequality, because the $i$th increment of $\{\xi_n^m\}_{n \in \mathbb{N}}$ involves the term $\tau_i^m$, which does not admit a bounded support. Instead, we will use the following upper bound on the moment generating function of $\xi_n^m$, whose proof is based on Doob’s inequality and is given in Appendix B.4.

**Lemma 4.** Fix $n \in \mathbb{N}$ and $\theta \in (0, m \beta_m)$. We have that

$$\mathbb{E} \left( \exp (\theta \xi_n^m) \right) \leq \exp \left( \frac{\theta^2 n}{4 m \beta_m (m \beta_m - \theta)} \right). \quad (71)$$

We are now ready to establish an upper bound on the quantity $1 - \mathbb{P}(\mathcal{B}_\epsilon^m)$. Recall that $J^m = (1+\epsilon)m \beta_m T$. Fix $\eta \in (0, \min\{T, 1/2\})$, and let $\theta_0 = \eta \frac{\epsilon}{1+\epsilon} T^{-1} m \beta_m$. In particular, $\theta_0 \in (0, m \beta_m/2)$. We have that

$$1 - \mathbb{P}(\mathcal{B}_\epsilon^m) = \mathbb{P} \left( \max_{0 \leq n \leq J^m} \xi_n^m \geq \epsilon \right) \leq \exp \left( - \frac{\theta_0^2 n}{4 m \beta_m (m \beta_m - \theta_0)} \right). \quad (72)$$
It therefore suffices to show that, if $m\beta$ is a submartingale, and $(\theta m^2/4)J_\theta$ is a submartingale, we have that

\[
\begin{align*}
&\leq \inf_{\theta > 0} \exp(-\theta \epsilon) \mathbb{E} \left( \exp \left( \theta \xi^m_{\epsilon^m_{\theta^m}} \right) \right) \\
&\leq \exp \left( -\theta_0 \epsilon + \frac{J_\theta^m \theta_0^2/4}{m\beta_m (m\beta_m - \theta_0)} \right) \\
&\leq \exp \left( -\theta_0 \epsilon + \frac{J_\theta^m \theta_0^2}{2(m\beta_m)^2} \right) \\
&= \exp \left( -\theta_0 \epsilon + \frac{(1 + \epsilon) T \theta_0}{2m\beta_m} \right) \\
&\leq \exp \left( -\theta_0 (\epsilon - \epsilon/2) \right) \\
&= \exp \left( -\eta \frac{\epsilon^2}{2 + 2\epsilon} T^{-1} m\beta_m \right). \quad (72)
\end{align*}
\]

Step (a) follows from Doob’s inequality and the fact that, for any $\theta > 0$, the sequence $\{\exp(\theta \xi^m_n)\}_{n \in \mathbb{N}}$ is a submartingale, and (b) from Lemma 4 with $n = J_\theta$. Step (c) is due to $\theta_0 < m\beta_m/2$, and hence $m\beta_m - \theta_0 > m\beta_m/2$. Finally, step (d) follows from the definition of $\theta_0$ and that $\eta < 1$.

With a derivation analogous to that of Eq. (72), we have that

\[
\mathbb{P} \left( \min_{0 \leq n \leq J_\epsilon} \epsilon^m \eta^m \leq -\epsilon \right) \leq \exp \left( -\eta \frac{\epsilon^2}{2 + 2\epsilon} T^{-1} m\beta_m \right). \quad (73)
\]

Therefore, combining Eq. (67) with Eqs. (69), (72) and (73), we have that

\[
\mathbb{P} \left( \sup_{0 \leq t \leq T} \left| \sum_{i=1}^{l_\eta + 1} \tau_i^m \left( Z_k^m[i] - p_k (Q_k^m[i]) \right) \right| \geq \epsilon \right) \leq \left( 1 - \mathbb{P}(A_\epsilon^m) \right) + (1 - \mathbb{P}(B_\epsilon^m)) \leq \frac{1 + \epsilon}{T \epsilon^2} (m\beta_m)^{-1} + 2 \exp \left( -\eta \frac{\epsilon^2}{2 + 2\epsilon} T^{-1} m\beta_m \right). \quad (74)
\]

Since $m\beta_m \to \infty$ as $m \to \infty$, the above equation further implies that

\[
\lim_{m \to \infty} \mathbb{P} \left( \sup_{0 \leq t \leq T} \left| \sum_{i=1}^{l_\eta + 1} \tau_i^m \left( Z_k^m[i] - p_k (Q_k^m[i]) \right) \right| \geq \epsilon \right) \leq \lim_{m \to \infty} \left( \frac{1 + \epsilon}{T \epsilon^2} (m\beta_m)^{-1} + 2 \exp \left( -\eta \frac{\epsilon^2}{2 + 2\epsilon} T^{-1} m\beta_m \right) \right) = 0. \quad (75)
\]

We now bound the second term in Eq. (63). It is not difficult to verify, from Eq. (14), that there exists $l_p \in \mathbb{R}_+$ such that $p_k(\cdot)$ is $l_p$-Lipschitz continuous for all $k \in \mathcal{K}$. We have that

\[
\mathbb{P} \left( \sup_{0 \leq t \leq T} \left| \int_0^t \left( p_k(Q_k^m(s)) - p_k(V_k^m(s)) \right) ds \right| \geq \epsilon \right) \leq \mathbb{P} \left( \sup_{0 \leq t \leq T} \left| \int_0^t l_p \left( \hat{Q}_k^m(s) - V_k^m(s) \right) ds \right| \geq \epsilon \right) \leq \mathbb{P} \left( \int_0^T l_p \left| \hat{Q}_k^m(s) - V_k^m(s) \right| ds \geq \epsilon \right). \quad (76)
\]

It therefore suffices to show that, if $m\beta_m \to \infty$ as $m \to \infty$, then

\[
\lim_{m \to \infty} \mathbb{P} \left( \int_0^T \left| \hat{Q}_k^m(s) - V_k^m(s) \right| ds \geq \epsilon \right) = 0, \quad \forall \epsilon > 0. \quad (77)
\]

To this end, we have that

\[
\int_0^T \left| \hat{Q}_k^m(s) - V_k^m(s) \right| ds = \sum_{i=0}^{l_\eta + 1} \int_{S_{i+1}/m}^{S_i/m} \left| \hat{Q}_k^m(t) - V_k^m(t) \right| ds
\]
which is isolated in the form of the following lemma. The proof involves an elementary application
(Eq. (38)). It remains to derive an upper bound on the tail probabilities of the term
the definition of
(G
This establishes Eq. (57), which, in light of Eq. (56) completes the proof of Proposition 2.

In step (a) we have invoked the property that \( \hat{Q}^m(t) \) is piece-wise constant. Step (b) follows from
the definition of \( G_k(\cdot, \cdot) \) in Eq. (35), and (c) from the fact that \( |V_k^m(t)| \leq q^0 + \lambda_k t \) for all \( t \in \mathbb{R}_+ \)
(Eq. (38)). It remains to derive an upper bound on the tail probabilities of the term \( \sum_{i=1}^{I_{mT+1}} (\tau_i^m)^2 \),
which is isolated in the form of the following lemma. The proof involves an elementary application
of Markov’s inequality, and is given in Appendix B.5.

**Lemma 5.** Suppose that \( m \beta_m \to \infty \) as \( m \to \infty \). We have that

\[
\lim_{m \to \infty} \mathbb{P} \left( \sum_{i=1}^{I_{mT+1}} (\tau_i^m)^2 \geq \epsilon \right) = 0. \quad \forall \epsilon > 0.
\]

(79)

We are now ready to prove Eq. (77). By Eq. (78), we have that

\[
\lim_{m \to \infty} \mathbb{P} \left( \int_0^T |\hat{Q}^m_{k}(s) - v_k^m(s)| ds \geq \epsilon \right)
\leq \lim_{m \to \infty} \mathbb{P} \left( \sup_{0 \leq s \leq T} |\overline{Q}^m_{k}(s) - V^m_k(s)| \geq \frac{\epsilon}{2T} \right) + \lim_{m \to \infty} \mathbb{P} \left( \sum_{i=1}^{I_{mT+1}} (\tau_i^m)^2 \geq \frac{\epsilon}{2[q_0 + \lambda_k(T + 1)]} \right)
\]

(80)

for all \( \epsilon > 0 \), where the first inequality follows from a union bound, and the last step from Proposition 1 and Lemma 5. Substituting Eqs. (75) and (80) into Eq. (63), we have that

\[
\lim_{m \to \infty} \mathbb{P} \left( \sup_{0 \leq t \leq T} \sum_{k \in K} \left| \int_0^t G_k(V^m(s), \overline{C}^m(s)) - F_k(V^m(s))ds \right| > \epsilon \right)
\leq \lim_{m \to \infty} \mathbb{P} \left( \sup_{0 \leq t \leq T} \sum_{i=1}^{I_{mT+1}} (\tau_i^m) \left| Z_k^m[i] - p_k(\overline{Q}^m[i]) \right| \geq \epsilon/\lambda_k \right)
+ \lim_{m \to \infty} \mathbb{P} \left( \sup_{0 \leq t \leq T} \int_0^t \left| p_k(\overline{Q}^m(s)) - p_k(V^m(s)) \right| ds \geq \epsilon/\lambda_k \right) = 0.
\]

(81)

This establishes Eq. (57), which, in light of Eq. (56) completes the proof of Proposition 2.
5.3. Exponential Convergence of Fluid Solutions to Invariant State  In this section, we show that the fluid solution converges exponentially fast to the unique invariant state, starting from any bounded set of initial conditions.

**Theorem 6.** Fix $h_0 > 0$. There exist $a, b > 0$, such that for all $u \in \mathbb{R}^K_+$, $\|u\| \leq h_0$,

$$
\|q(u, t) - q^I\| \leq a \exp(-bt), \; \forall t \in \mathbb{R}_+.
$$

(82)

**Main idea.** The main difficulty in the proof of Theorem 6 is that, depending on the initial condition, the individual coordinates of $q(\cdot)$ may not converge to their equilibrium state monotonically, as can be seen in the trajectory of $q_2(\cdot)$ in the first three plots of Figure 5. The main idea behind our proof is to obtain a more accurate description of $q(\cdot)$ with the aid of the following piece-wise linear potential function:

$$
g(q) = \sum_{k \in K} q_k \lor \alpha_0, \; q \in \mathbb{R}^K_+.
$$

(83)

By analyzing the the evolution of $g(q(\cdot))$, our proof shows that: (1) for every bounded initial condition of $q(\cdot)$, the fluid solution rapidly enters a desirable subset of the state space, and (2) starting from this subset exponential convergence to the invariant state takes place.

While potential functions are often used to establish convergence results, to the best of our knowledge, the use of the potential function $g(\cdot)$ in Eq. (83) is new. This is largely due to the fact that our potential function is specifically designed to mirror the total “weights” across all actions under the reward matching rule. In addition, our use of the potential function also diverges from their conventional role in proving convergence results: we do not use $g(\cdot)$ directly for characterizing the rate of convergence; the primary role of $g(\cdot)$ is for controlling the position of the trajectory in the state-space. Once we are assured that the trajectory has entered a suitable subset of the state space, the rate of convergence can be more readily obtained via elementary arguments.

**Proof of Theorem 6** Fix $u \in \mathbb{R}^K_+$, such that $\|u\| \leq h_0$. For simplicity of notation, we will omit the dependence of the initial condition and write $q(\cdot)$ in place of $q(u, \cdot)$, whenever doing so does not cause any confusion. The proof of the theorem is divided into two parts, depending on the value of $\alpha_0$. For the first part of the proof, we will assume that $\alpha_0 \leq \lambda_1/K$.

Let $g(\cdot)$ be the potential function defined in Eq. (83). We will first partition $\mathbb{R}^K_+$ into three disjoint subsets, depending on the value of $g(\cdot)$. Let $\epsilon$

$$
\epsilon = \min \left\{ \frac{\lambda_1 - \lambda_2}{2}, \lambda_1 - (K-1)\alpha_0 \right\},
$$

(84)

and define

$$
\mathcal{Q}_1 = \{q \in \mathbb{R}^K_+ : g(q) \leq \lambda_1 - \epsilon\}, \\
\mathcal{Q}_2 = \{q \in \mathbb{R}^K_+ : \lambda_1 - \epsilon < g(q) \leq \lambda_1\}, \\
\mathcal{Q}_3 = \{q \in \mathbb{R}^K_+ : g(q) > \lambda_1\}.
$$

(85)

We observe some useful properties of the fluid solutions.

**Lemma 6.** Fix $t > 0$ so that all coordinates of $q(\cdot)$ are differentiable. The following holds:

1. If $q(t) \in \mathcal{Q}_1 \cup \mathcal{Q}_2$, then

$$
\dot{q}_1(t) \geq 0.
$$

(86)

Furthermore, there exists a constant $d_1 > 0$, such that if $q(t) \in \mathcal{Q}_1$, then

$$
\dot{q}_1(t) \geq d_1.
$$

(87)
2. Denote by $\mathcal{K}^+ (t)$ the coordinates over which $q(t)$ is at least $\alpha_0$:

$$\mathcal{K}^+ (t) = \{ k \in \mathcal{K} : q_k (t) \geq \alpha_0 \}. \quad (88)$$

There exists a constant $d_2 > 0$, such that if $q(t) \in \mathcal{Q}_2 \cup \mathcal{Q}_3$, then for all $k \in \mathcal{K}^+ (t)$, $k \geq 2$,

$$\dot{q}_k (t) \leq -d_2. \quad (89)$$

3. If $q(t) \in \mathcal{Q}_3$, then

$$\frac{d}{dt} g(q(t)) < 0. \quad (90)$$

Proof. Suppose that $q(t) \in \mathcal{Q}_1 \cup \mathcal{Q}_2$. We have that

$$\dot{q}_1 (t) = \frac{\lambda_1 q_1 (t) \lor \alpha_0}{g(q(t))} - q_1 (t)$$

$$\geq \frac{\lambda_1 q_1 (t) \lor \alpha_0}{g(q(t))} - q_1 (t) \lor \alpha_0$$

$$= (q_1 (t) \lor \alpha_0) \left( \frac{\lambda_1}{g(q(t))} - 1 \right)$$

$$\overset{(a)}{\geq} \alpha_0 \left( \frac{\lambda_1}{\lambda_1} - 1 \right) = 0, \quad (91)$$

where step $(a)$ follows from the assumption that $q(t) \in \mathcal{Q}_1 \cup \mathcal{Q}_2$. In a similar fashion, now suppose that $q(t) \in \mathcal{Q}_1$ and hence $g(q(t)) \leq \lambda_1 - \epsilon$. We obtain that

$$\dot{q}_1 (t) \geq \alpha_0 \left( \frac{\lambda_1}{\lambda_1 - \epsilon} - 1 \right) \overset{\triangle}{=} d_1 > 0. \quad (92)$$

This proves the first claim of the lemma.

For the second claim, fix $q(t) \in \mathcal{Q}_2 \cup \mathcal{Q}_3$ and $k \in \mathcal{K}^+ (t)$, $k \geq 2$. We have that

$$\dot{q}_k (t) = \frac{\lambda_k q_k (t) \lor \alpha_0}{g(q(t))} - q_k (t)$$

$$= q_k (t) \left( \frac{\lambda_k}{g(q(t))} - 1 \right)$$

$$\overset{(a)}{\leq} \alpha_0 \left( \frac{\lambda_k}{\lambda_1} - 1 \right)$$

$$\overset{(b)}{\leq} \alpha_0 \left( \frac{\lambda_k}{\lambda_2 + (\lambda_1 - \lambda_2)/2} - 1 \right) < 0, \quad (93)$$

where step $(a)$ follows from the assumption that $q(t) \in \mathcal{Q}_2 \cup \mathcal{Q}_3$, and $(b)$ from the fact that $\lambda_2 \geq \lambda_k$ for all $k \geq 2$, and $\epsilon \leq \frac{\lambda_1 - \lambda_2}{2}$. This proves the second claim.

Finally, for the last claim, note that since $\alpha_0 \leq \lambda_1 / K$, the fact that $q(t) \in \mathcal{Q}_3$ implies that at least one component of $q(t)$ is no less than $\alpha_0$, and hence $\mathcal{K}^+(t) \neq \emptyset$. We have that

$$\frac{d}{dt} g(q(t)) = \frac{d}{dt} \left( \sum_{k \in \mathcal{K}} q_k (t) \lor \alpha_0 \right)$$

$$= \sum_{k \in \mathcal{K}^+ (t)} \dot{q}_k (t)$$

$$= \sum_{k \in \mathcal{K}^+ (t)} q_k (t) \left( \frac{\lambda_k}{g(q(t))} - 1 \right)$$

$$\overset{(a)}{<} \sum_{k \in \mathcal{K}^+ (t)} \alpha_0 \left( \frac{\lambda_k}{\lambda_1} - 1 \right)$$

$$\leq 0. \quad (94)$$
where the strict inequality in step (a) follows from the definition of $Q_3$, the fact that $K^+(t)$ is non-empty, and that $\lambda_k \leq \lambda_1$ for all $k$. This completes the proof of Lemma 6. Q.E.D.

We proceed by considering two cases for the initial condition of $q(\cdot)$.

**Case 1:** $q(0) \in Q_1 \cup Q_2$. Claim 3 in Lemma 6 implies that
\begin{equation}
q(t) \notin Q_3, \quad \forall t \in \mathbb{R}_+.
\end{equation}
By Claim 1 of the same lemma, we hence conclude that $q_1(t)$ is non-decreasing in $t$ for all $t \in \mathbb{R}_+$. By the same claim, we observe that $q_1(t)$ is bounded from below by the constant $d_1$ whenever $q(t)$ lies in $Q_1$. Since $g(q(t)) \geq q_1(t)$, we further conclude that in order for Eq. (95) to be valid, the amount of time $q(t)$ spends in $Q_1$ must satisfy:
\begin{equation}
\int_{t \geq 0} I(q(t) \in Q_1) dt \leq \frac{\lambda_1}{d_1},
\end{equation}
which, combined with Eq. (95), implies that for all $t \geq 0$:
\begin{equation}
\int_0^t I(q(s) \in Q_2) ds \geq t - \frac{\lambda_1}{d_1}.
\end{equation}
Claim 2 of Lemma 6 states that if $q(t) \in Q_2$, then any component of $q(t)$ in $K^+(t) \setminus \{1\}$ will have a negative drift that is bounded away from zero. Fix $t \in \mathbb{R}_+$, and suppose there exists $k' \in K^+(t) \setminus \{1\}$. Since $\dot{q}_k(t) \leq \lambda_1$ for all $k$ and $t$, by Eq. (97), we have that
\begin{equation}
q_{k'}(t) \leq q_{k'}(0) + \lambda_1 \left( \frac{\lambda_1}{d_1} - \left( t - \frac{\lambda_1}{d_1} \right) d_2 \right) \leq h_0 + \frac{\lambda_1}{d_1} - \left( t - \frac{\lambda_1}{d_1} \right) d_2.
\end{equation}
Since the components of $q(t)$ must be non-negative, the above equation implies that for all $k \geq 2$,
\begin{equation}
q_k(t) \leq \alpha_0, \quad \forall t \geq T_0,
\end{equation}
where $T_0 = \frac{h_0 + \lambda_1^2}{d_1^2} + \frac{\lambda_1}{d_1}$.

Eq. (99) will let us strengthen our previous observation. For all $t \geq T_0$, we have that
\begin{equation}
g(q(t)) = \sum_{k \in K} q_k(t) \vee \alpha_0 = q_1(t) \vee \alpha_0 + (K-1)\alpha_0.
\end{equation}
Since $q_1(t)$ is non-decreasing in $t$ by the first claim of Lemma 6, we conclude that $g(q(t))$ is non-decreasing in $t$ after $T_0$. Combined with Eq. (96), this implies that
\begin{equation}
q(t) \in Q_2, \quad \forall t \geq T_1,
\end{equation}
where $T_1 = T_0 + \frac{\lambda_1}{d_1}$. Fix $t > T_1$. Observe that by the definition of $\epsilon$, we have that $q_1(t) \geq \alpha_0$ whenever $q(t) \in Q_2$. We have that
\begin{equation}
g(q(t)) = q_1(t) + (K-1)\alpha_0,
\end{equation}
and hence
\begin{align*}
\frac{d}{dt}(q^t - q_1(t)) &= -\dot{q}_1(t) = -\lambda_1 \frac{q_1(t)}{q_1(t) + (K-1)\alpha_0} + q_1(t) \\
&= -q_1(t) \left( \frac{\lambda_1}{q_1(t) + (K-1)\alpha_0} - 1 \right) \\
&\leq -\alpha_0 \left( \frac{\lambda_1}{q_1(t) + (K-1)\alpha_0} - 1 \right)
\end{align*}
respectively. It can be verified that the solution to the ODE \( \dot{\gamma} \) where
\[
\frac{1}{1 - (\lambda_1 - (K - 1)\alpha_0 - q_1(t))/\lambda_1 - 1}
\]
where step \((a)\) follows from the fact that \( q_1(t) \geq \alpha_0 \). For step \((b)\), note that since \( q(t) \in Q_2 \) and \( g(q(t)) = q_1(t) + (K - 1)\alpha_0 \), we have that
\[
q_1^t - q_1(t) = q_1^t - [g(q(t)) - (K - 1)\alpha_0] \in [0, \epsilon) \subset [0, \lambda_1/2).
\]
The inequality in step \((b)\) thus follows from the fact that \( 1/(1 - x) \geq 1 + x \), for all \( x \in [0, 1) \), where we let \( x = (q_1^t - q_1(t))/\lambda_1 \).

From Eq. (103), we conclude that for all \( s > 0 \),
\[
q_1^t - q_1(T_1 + s) \leq (q_1^t - q_1(T_1)) \exp \left( -\frac{\alpha_0}{\lambda_1} s \right) \leq \epsilon \exp \left( -\frac{\alpha_0}{\lambda_1} s \right),
\]
where the last inequality follows from the fact that \( q(T_1) \in Q_2 \) and \( q_k(T_1) \leq \alpha_0 \) for all \( k \geq 2 \), and hence \( q_1(T_1) \geq \lambda_1 - (K - 1)\alpha_0 - \epsilon = q_1^t - \epsilon \). Let \( a_1 = q_1^t \vee \epsilon \), and \( b_1 = \alpha_0/\lambda_1 \). Since \( q_k(t) \) is non-decreasing by Lemma 6 from Eqs. (101) and (106), we have that if \( q(0) \in Q_1 \cup Q_2 \), then
\[
|q_1(t) - q_1| \leq a_1 \exp(-b_1 t), \quad \forall t \in \mathbb{R}_+,
\]
where \( \lambda \geq 2\lambda_1 \).

We now show the convergence of \( \dot{q}_k(\cdot) \) for \( k \geq 2 \). Fix \( k \in \{2, \ldots, K\} \). Define
\[
\delta_k(t) = q_k(t) - q_1^t, \quad k \in K,
\]
and recall from Theorem 3 that \( q_1^t = \frac{\lambda_1}{\lambda^2} \alpha_0 \), \( k = 2, \ldots, K \). By Eq. (99), we have that \( q_k(t) \leq \alpha_0 \) for all \( t \geq T_0 \). Therefore, for all \( t \geq T_1 \),
\[
\dot{q}_k(t) = \lambda_k - \frac{\alpha_0}{g(q(t)) - q_k(t)} - q_k(t)
\]
\[
= \lambda_k - \frac{\alpha_0}{q_1(t) + (K - 1)\alpha_0} - q_k(t)
\]
\[
= \lambda_k - \frac{\alpha_0}{\lambda_1 + \delta_1(t)} - q_k(t).
\]
Using the Taylor expansion of the function \( f(x) = \frac{x}{\lambda_1^2 + x} \) around \( x = 0 \) and noting that \( \delta_1(t) \leq a_1 \exp(-b_1 t) \) (Eq. (107)), we have that there exists \( T_2 \geq T_1 \), such that for all \( t \geq T_2 \),
\[
\dot{\delta}_k(t) \in -\delta_k(t) \pm \frac{2\lambda_k \alpha_0}{\lambda_1} |\delta_1(t)| \pm \gamma_k \exp(-b_1 t),
\]
where \( \gamma_k \triangleq \frac{2\lambda_k \alpha_0}{\lambda_1} \), and we use the notation \( x \in y \pm z \) to mean \( y - z \leq x \leq y + z \). Given a fixed value of \( \delta_k(T_2) \), Eq. (110) implies that the value of \( \delta_k(T_2 + t) \) must lie between the solutions to the ODEs
\[
\dot{x}(t) = -x(t) + \gamma_k \exp(-b_1 t) \quad \text{and} \quad \dot{x}(t) = -x(t) - \gamma_k \exp(-b_1 t),
\]
with initial condition \( x(0) = \delta_k(T_2) \), respectively. It can be verified that the solution to the ODE
\[
\dot{x}(t) = -x(t) + c_1 \exp(c_2 t),\]
condition \( x(0) = x_0 \), is given by \( x(t) = \frac{c_1}{1 - c_2} \exp(-c_2 x) + \left( x_0 - \frac{c_1}{1 - c_2} \right) \exp(-x) \). Setting \( c_1 = \gamma_k \) and \( c_2 = b_1 \), we thus conclude that there exist \( \bar{a}_k, \bar{b}_k > 0 \), such that

\[
|\delta_k(T_2 + t)| \leq (\delta_k(T_2) + \bar{a}_k) \exp(-\bar{b}_k).
\] (111)

Note that \( \dot{\delta}_k(t) \) is always bounded. We thus conclude that there exists \( a_k > 0 \) such that

\[
|q_k(t) - q_k^1| = |\delta_k(t)| \leq a_k \exp(-b_k t), \quad \forall t \in \mathbb{R}_+.
\] (112)

Since the above equation holds for all \( k \geq 2 \), together with Eq. (107), we have proven Eq. (82) for the first case, by assuming that \( q(0) \in Q_1 \cup Q_2 \).

**Case 2**: \( q(0) \in Q_3 \). We now consider the second case where the initial condition, \( q(0) \), belongs to the set \( Q_3 \). Recall from Eq. (89) of Lemma 6 that when \( q(t) \in Q_2 \cup Q_3 \), all components of \( q(t) \) in \( K^+(t) \) exhibit a negative drift of magnitude at least \( d_2 \). Letting \( T_3 = h_0/d_2 \), we thus conclude that one of the following scenarios must occur:

1. \( q(s') \in Q_1 \cup Q_2 \) for some \( s' \in [0, T_3] \).
2. \( q(t) \in Q_3 \) for all \( t \in [0, T_3] \), and there exists \( s' \in [0, T_3] \) such that

\[
q_k(s') \leq \alpha_0, \quad \forall k \geq 2.
\] (113)

For the first scenario, it is equivalently to “re-initializing” the system at time \( t = s' \) at a point in \( Q_1 \). The proof for Case 1 presented earlier thus applies and Eq. (82) follows. In what follows, we will focus on the second scenario. Without loss of generality, it suffices to show the validity of Eq. (82) by fixing an initial condition:

\[
q(0) \in Q_3, \quad \text{and} \quad \max_{k \geq 2} q_k(0) \leq \alpha_0.
\] (114)

Not that, by the definition of \( Q_3 \), this implies that

\[
q_1(0) > q_1^1 > \alpha_0.
\] (115)

We next show that \( q_1(t) > q_1^1 \) for all \( t \in \mathbb{R}_+ \). Fix \( t > 0 \) such that

\[
q(s) \in Q_3, \quad \forall s \in [0, t].
\] (116)

Such \( t \) exists because the set \( Q_3 \) is open and \( q(\cdot) \) is continuous. From Lemma 3 we know that if \( q(t) \in Q_3 \), then any component in \( K^+(t) \setminus \{1\} \) must have a negative drift. It thus follows that

\[
q_k(t) < \alpha_0, \quad \forall k \geq 2.
\] (117)

We have that, whenever \( q_k^1 < q_1(t) < q_k^1 + 1 \),

\[
\frac{d}{dt} (q_1(t) - q_1^1) = \dot{q}_1(t) \overset{(a)}{=} -q_1(t) \left( 1 - \frac{\lambda_1}{q_1(t) + (K - 1)\alpha_0} \right)
\]

\[
= -q_1(t) \left( 1 - \frac{1}{1 + (q_1(t) - q_1^1)/\lambda_1} \right)
\]

\[
\geq -q_1^1 + 1 \left( 1 - \frac{1}{1 + (q_1(t) - q_1^1)/\lambda_1} \right)
\]

\[
\overset{(b)}{=} -\frac{q_1^1 + 1}{2\lambda_1} (q_1(t) - q_1^1)
\] (118)

where step \( (a) \) follows from the fact that \( q_k(t) \leq \alpha_0 \) for all \( k \geq 2 \), and \( (b) \) from \( 1/(1 + x) \leq 1 - x/2 \) for all \( x \in [0, 1] \). In light of Eq. (103), the above inequality implies if the value of \( q_1(t) - q_1^1 \) is ever
to be below 1, then from that point on it will be bounded from below by an exponential function that is strictly positive. This shows that \( q_1(t) > q_1' \) for all \( t > 0 \).

We now prove the exponential rate of convergence of \( q_1(t) \) to \( q_1' \). Using again Eq. (118), we obtain

\[
\frac{d}{dt} (q_1(t) - q_1') = -q_1(t) \left( 1 - \frac{\lambda_1}{q_1(t) + (K - 1)\alpha_0} \right) \\
\leq -q_1' \left( 1 - \frac{1}{1 + (q_1(t) - q_1')/\lambda_1} \right) \\
\leq -\frac{q_1'}{\lambda_1} (q_1(t) - q_1'),
\]

(119)

where step (a) follows from the fact that \( q_1(t) > q_1' \), and (b) from \( 1/(1 + x) \geq 1 - x \) for all \( x \geq 0 \). Using the reasoning identical to that in Eqs. (103) through (107), we conclude that there exists \( a_1, b_1 > 0 \), such that

\[
|q_1(t) - q_1'| \leq a_1 \exp(-b_1 t), \quad \forall t \in \mathbb{R}_+.
\]

(120)

Finally, the fact that \( q_1(t) > q_1' \) implies that \( g(q(t)) > \lambda_1 \), and hence \( q(t) \in Q_3 \) for all \( t > 0 \). By Eq. (117), this further implies that for any \( k \geq 2 \), \( q_k(t) \leq \alpha_0 \) for all \( t \in \mathbb{R}_+ \). The exponential rate of convergence of \( q_k(t) \) for \( k \in \{2, \ldots, K\} \) thus follows from the same argument as in Eqs. (109) through (112). This completes the proof of Theorem 6, assuming that \( \alpha_0 < \lambda_1/K \).

We now turn to the case where \( \alpha_0 > \lambda_1/K \). We will use a similar proof strategy by partitioning the state space based on the value of \( g(q(t)) \), albeit with a different partitioning. In particular, define the sets

\[
\hat{Q}_1 = \{ q \in \mathbb{R}^K_+ : g(q) = K\alpha_0 \}, \\
\hat{Q}_2 = \{ q \in \mathbb{R}^K_+ : g(q) > K\alpha_0 \}.
\]

(121)

Note that by definition \( g(q) \geq K\alpha_0 \) for all \( q \in \mathbb{R}^K_+ \), so the above sets constitute a partition of \( \mathbb{R}^K_+ \). We have the following lemma.

**Lemma 7.** Fix \( t > 0 \) so that all coordinates of \( q(\cdot) \) are differentialble. If \( q(t) \in \hat{Q}_2 \), then there exists a constant \( d_3 > 0 \), such that \( \frac{d}{dt} g(q(t)) < -d_3 \).

**Proof.** Suppose that \( q(t) \in \hat{Q}_2 \). Since \( g(q(t)) > K\alpha_0 \), we have that \( K^+(t) \neq \emptyset \). We have that

\[
\frac{d}{dt} g(q(t)) = \frac{d}{dt} \left( \sum_{k \in \mathcal{K}} q_k(t) \lor \alpha_0 \right) = \sum_{k \in K^+(t)} \dot{g}_k(t) \\
= \sum_{k \in K^+(t)} q_k(t) \left( \frac{\lambda_k}{g(q(t))} - 1 \right) \\
\leq -\sum_{k \in K^+(t)} \alpha_0 \left( 1 - \frac{\lambda_k}{K\alpha_0} \right) \\
\leq -\alpha_0 \left( 1 - \frac{\lambda_1}{K\alpha_0} \right) \triangleq -d_3,
\]

(122)

where step (a) follows from the fact that \( g(q(t)) > K\alpha_0 \), and (b) from \( K^+(t) \neq \emptyset \). Finally, note that since \( \alpha_0 > \lambda_1/K \), \( d_3 \) is strictly positive. This completes the proof of the lemma. Q.E.D.

Lemma 7 implies that if \( q(0) \in \hat{Q}_2 \), then we must have that

\[
q(t) \in \hat{Q}_1, \quad \forall t \geq h_0/d_3.
\]

(123)
Since the rate of change in every coordinate of $q(t)$ is bounded, in light of the above equation, it suffices to establish Eq. (122) by considering only the scenario where $q(t) \in \mathcal{Q}_1$ for all $t \in \mathbb{R}_+$. Fixing $k \in \mathcal{K}$, we have that

$$
\frac{d}{dt} (q_k(t) - q_k^*) = \lambda_k \frac{q_k(t)}{g(q(t))} - q_k(t) \overset{(a)}{=} \lambda_k \frac{\alpha_0}{g(q(t))} - q_k(t) = \lambda_k / K - q_k(t) \overset{(c)}{=} -(q_k(t) - q_k^*),
$$

(124)

where step (a) follows from the fact that $q_k(t) \leq \alpha_0$ whenever $q(t) \in \mathcal{Q}_1$, (b) from the fact that $g(q(t)) = K \alpha_0$, and (c) from the definition $q_k^* = \lambda_k / K$. In light of Eq. (105), Eq. (124) implies that

$$
|q_k(t) - q_k^*| = |q_k(0) - q_k^*| \exp(-t) \leq (h_0 \vee q_k^*) \exp(-t), \quad t \in \mathbb{R}_+.
$$

(125)

Since the above equation holds for all $k$, we have thus established an exponential rate of convergence of $q(t)$ to $q^*$, as $t \to \infty$. This completes the proof of Theorem 4. Q.E.D.

5.4. Convergence of Steady-State Distributions We complete the proof of Theorem 4 in this subsection. We will begin by establishing a useful distributional upper bound on $Q_k^m(\infty)$. Fix $t \in \mathbb{R}_+$, and denote by $U_k^m(t)$ the number of jobs in system at time $t$ in an $M/M/1$ queue with arrival rate $\lambda$ and departure rate $1/m$, and by $U_k^m$ its steady-state distribution. It is well known that $U_k^m$ is a Poisson random variable with mean $m \lambda$, and it follows that, for all $\lambda > 0$,

$$
m^{-1}U_k^m \to \lambda, \quad \text{almost surely, as } m \to \infty.
$$

(126)

Define $W = \mathbb{R}_+^\mathcal{K} \times \mathcal{K}$, and the process

$$
\mathcal{W}^m(t) = (\mathcal{Q}^m(t), C^m(t)), \quad t \in \mathbb{R}_+.
$$

(127)

We have the following lemma, whose proof is given in Appendix B.6.

**Lemma 8.** The process $\{\mathcal{W}^m(t)\}_{t \in \mathbb{R}_+}$ is positive recurrent, and $Q_k^m(\infty) \leq m^{-1}U_k^m$, for all $k \in \mathcal{K}$.

Denote by $\pi^m$ the probability distribution over $W$ that corresponds to the steady-state distribution of $\mathcal{W}^m(\cdot)$. Lemma 8 can be used to to show that the sequence $\{\pi^m\}_{m \in \mathbb{N}}$ is tight, as formalized in the following result, whose proof is given in Appendix B.7.

**Lemma 9.** For every $x \in \mathbb{R}_+$, define $\bar{Q}_x = \{q \in \mathbb{R}_+^\mathcal{K} : \max_{k \in \mathcal{K}} q_k \leq x\} \times \mathcal{K}$, $x \in \mathbb{R}_+$. Then, for every $\epsilon > 0$, there exists $M \in \mathbb{N}$ such that

$$
\pi^m(\bar{Q}_x) \geq 1 - \epsilon, \quad \forall m \in \mathbb{N},
$$

(128)

We say that $y$ is a limit point of $\{x_i\}_{i \in \mathbb{N}}$ if there exists a sub-sequence of $\{x_i\}$ that converges to $y$. It is not difficult to verify that the space $W$ is separable as a result of both $\mathbb{R}_+^\mathcal{K}$ and $\mathcal{K}$ being separable. By Prohorov’s theorem, the tightness of $\{\pi^m\}_{m \in \mathbb{N}}$ thus implies that any sub-sequence of $\{\pi^m\}_{m \in \mathbb{N}}$ admits a limit point with respect to the topology of weak convergence. Let $\{\pi^m_i\}_{i \in \mathbb{N}}$ be a sub-sequence of $\{\pi^m\}_{m \in \mathbb{N}}$, and $\pi$ a limit point of the sub-sequence. Denote by $\pi$ and $\pi^m_i$ the marginals of $\pi^m$ and $\pi^m_i$ over the first $\mathcal{K}$ coordinates (corresponding to $Q^m(\cdot)$), respectively. In the remainder of the proof, we will show that $\pi$ necessarily concentrates on $q^*$. We first show that $\pi$ is an stationary measure with respect to the deterministic fluid solution. To this end, we will use the method of continuous test functions (cf. Section 4 of Ethier and Kurtz 2005). Let $\mathcal{F}$ be the space of all bounded continuous functions from $\mathbb{R}_+^\mathcal{K}$ to $\mathbb{R}_+$. We will demonstrate that, for all $f \in \mathcal{F},$

$$
|\mathbb{E}_\pi (f(q(t))) - \mathbb{E}_\pi (f(q^0))| = 0, \quad \forall t \in \mathbb{R}_+.
$$

(129)
Define $Q^m = \{x/m, x \in \mathbb{Z}^\infty_+\}$, $m \in \mathbb{N}$. Fix $f \in \mathcal{C}$. Let $W^m(0)$ be distributed according to $\pi^m_W$, and define

$$F^m(q^0, t) = \mathbb{E}\left( f(\overline{Q}^m(t)) \left| \overline{Q}^m(0) = q^0 \right. \right), \quad t \in \mathbb{R}_+, q^0 \in Q^m. \quad (130)$$

Fix $t > 0$. We have that

$$\left| \mathbb{E}_\pi \left( f(q^0, t) \right) - \mathbb{E}_\pi \left( f(q^0) \right) \right| \leq \limsup_{i \to \infty} \left| \mathbb{E}_\pi \left( f(q^0, t) \right) - \mathbb{E}_{\pi^m_i} \left( f(q^0, t) \right) \right| + \limsup_{i \to \infty} \left| \mathbb{E}_{\pi^m_i} \left( f(q^0, t) \right) - \mathbb{E}_{\pi^m_i} \left( F^{m_i}(q^0, t) \right) \right|$$

and therefore,

$$\limsup_{i \to \infty} \left| \mathbb{E}_\pi \left( f(q^0, t) \right) - \mathbb{E}_{\pi^m_i} \left( F^{m_i}(q^0, t) \right) \right| \leq \limsup_{i \to \infty} \left| \mathbb{E}_\pi \left( f(q^0, t) \right) - \mathbb{E}_{\pi^m_i} \left( f(q^0, t) \right) \right| + \limsup_{i \to \infty} \left| \mathbb{E}_{\pi^m_i} \left( f(q^0, t) \right) - \mathbb{E}_{\pi^m_i} \left( F^{m_i}(q^0, t) \right) \right|$$

$$= \limsup_{i \to \infty} \left| \mathbb{E}_\pi \left( f(q^0, t) \right) - \mathbb{E}_{\pi^m_i} \left( f(q^0, t) \right) \right| + \limsup_{i \to \infty} \left| \mathbb{E}_{\pi^m_i} \left( f(q^0, t) \right) - \mathbb{E}_{\pi^m_i} \left( F^{m_i}(q^0, t) \right) \right|$$

Step (a) follows from the triangle inequality, and (b) from the fact that $Q^m(\cdot)$ is stationary when initialized according to $\pi^m$, and therefore,

$$\mathbb{E}_{\pi^m} \left( F^m(q^0, t) \right) = \int_{q^0 \in \mathbb{R}^K_+} \mathbb{E}\left( f(\overline{Q}^m(t)) \left| \overline{Q}^m(0) = q^0 \right. \right) d\pi^m = \mathbb{E}_{\pi^m} \left( f(q^0) \right). \quad (132)$$

For step (c), we note that since $q(q^0, t)$ depends continuously on $q^0$ (Lemma [1]), the function $g(x) = f(q(x, t))$, $x \in \mathbb{R}^K_+$, belongs to $\mathcal{C}$. Therefore, the step follows from the fact that $\pi^m_i$ converges weakly to $\pi$ as $i \to \infty$.

Fix $\epsilon > 0$. There exists $M > 0$ such that the the right-hand side of Eq. (131) satisfies

$$\limsup_{i \to \infty} \left| \mathbb{E}_{\pi^m_i} \left( f(q^0, t) \right) - \mathbb{E}_{\pi^m_i} \left( F^{m_i}(q^0, t) \right) \right| \leq \limsup_{i \to \infty} \left| \int_{q^0 \in \mathbb{R}^K_+} f(q^0, t) d\pi^m_i - \int_{q^0 \in \mathbb{R}^K_+} F^m(q^0, t) d\pi^m_i \right| +$$

$$+ \limsup_{i \to \infty} \left| \int_{q^0 \in \mathbb{R}^K_+ \setminus \mathcal{M}} f(q^0, t) d\pi^m_i - \int_{q^0 \in \mathbb{R}^K_+ \setminus \mathcal{M}} F^m(q^0, t) d\pi^m_i \right|$$

$$\leq \limsup_{i \to \infty} \left| \int_{q^0 \in \mathbb{R}^K_+ \setminus \mathcal{M}} |f(q^0, t) - F^{m_i}(q^0, t)| d\pi^m_i + 2\epsilon \sup_{x \in \mathbb{R}^K_+} f(x) \right|$$

$$= 2\epsilon \sup_{x \in \mathbb{R}^K_+} |f(x)|. \quad (133)$$

where (a) follows from the tightness property of Eq. (128), and step (b) involves an argument that allows for interchanging the order of taking the limit and integration. We isolate step (b) in the following lemma, whose proof leverages Theorem [5].

**Lemma 10.** Fix a compact set $S \subset \mathbb{R}^K_+$ and $f \in \mathcal{C}$, we have that

$$\limsup_{m \to \infty} \int_{q^0 \in S} |f(q^0, t) - F^m(q^0, t)| d\pi^m = 0. \quad (134)$$

**Proof.** We first show the following uniform convergence property:

$$\lim_{m \to \infty} \sup_{x \in S \cap Q^m} P \left( \|q(x, t) - \overline{Q}^m(x, t)\| > \delta \right) = 0, \quad \forall \delta > 0. \quad (135)$$
where $\overline{Q}^m(x, \cdot)$ denotes a process $\overline{Q}^m(\cdot)$ initialized with $\mathbb{P}(\overline{Q}^m(0) = x) = 1$. Suppose, for the sake of contradiction, that there exist $\delta > 0$ and a sequence $\{x_i\}_{i \in \mathbb{N}}$, $x_i \in S \cap \overline{Q}^m$, such that
\[
\limsup_{i \to \infty} \mathbb{P}\left( \|q(x_i, t) - \overline{Q}^m(x_i, t)\| > \delta \right) > 0.
\] (136)
Because $S$ is compact, there exists a sub-sequence $\{x_{i_j}\}_{j \in \mathbb{N}} \subset \{x_i\}_{i \in \mathbb{N}}$ and $x^* \in S$ such that $x_{i_j} \to x^*$ as $j \to \infty$. We have that
\[
\limsup_{j \to \infty} \mathbb{P}\left( \|q(x_{i_j}, t) - \overline{Q}^m(x_{i_j}, t)\| > \delta \right) \leq \limsup_{j \to \infty} \mathbb{P}\left( \|q(x^*, t) - \overline{Q}^m(x^*, t)\| > \delta \right) > 0,
\] (137)
where step (a) follows from the fact that, for all $t \in \mathbb{R}_+$, $q(x, t)$ is continuous with respect to $x$ (Lemma 1). This leads to a contradiction with Theorem 5, and hence proves Eq. (135).

Fix $\delta > 0$. We have that
\[
\lim_{m \to \infty} \int_{q^0 \in S} \left| f(q^0, t) - F^m(q^0, t) \right| d\pi^m \leq \lim_{i \to \infty} \int_{q^0 \in S} \mathbb{E}\left( \left| f(q^0, t) - f(\overline{Q}^m(t)) \right| \right) \left| \overline{Q}^m(0) = q^0 \right) d\pi^m
\]
\[
\leq w_f(S, \delta) + \left( \sup_{x \in \mathbb{R}_+^k} |f(x)| \right) \lim_{i \to \infty} \sup_{x \in S} \mathbb{P}\left( \|q(x, t) - \overline{Q}^m(x, t)\| > \delta \right)
\]
\[
\overset{(b)}{=} w_f(S, \delta),
\] (138)
where $w_f$ is the modulus of continuity of $f$ in $S$: $w_f(S, \delta) = \sup_{x, y \in S, \|x - y\| \leq \delta} |f(x) - f(y)|$. Step (a) follows from the fact that $\pi^m(\overline{Q}^m) = 1$, and (b) from Eq. (135). Because $S$ is compact, $f$ is uniformly continuous in $S$ and hence $\lim_{\delta \to 0} w_f(S, \delta) = 0$. The lemma follows by taking the limit as $\delta \to 0$ in Eq. (138). Q.E.D.

Since Eq. (133) holds for all $\alpha > 0$, combining it with Eq. (131), we conclude that
\[
\left| \mathbb{E}_\pi\left( f(q^0, t) \right) - \mathbb{E}_\pi\left( f(q^0) \right) \right| \leq \limsup_{i \to \infty} \left| \mathbb{E}_{\pi^m_i}\left( f(q^0, t) \right) - \mathbb{E}_{\pi^m_i}\left( F^m(q^0, t) \right) \right| = 0,
\] (139)
and this proves Eq. (129). We now show that Eq. (129) implies that $\pi = \delta_f$, i.e., $\delta_f$ is the unique invariant measure with respect to the dynamics induced by the fluid solution. Define the truncated norm:
\[
\|x\|_M \overset{\Delta}{=} \min\{\|x\|, M\}, \quad M \in \mathbb{R}_+,
\] (140)
and the set
\[
\mathcal{R}_L \overset{\Delta}{=} \{q \in \mathbb{R}_+^k : \|q\| \leq L\}, \quad L \in \mathbb{R}_+.
\] (141)
Let $q^0$ be a random vector in $\mathbb{R}_+^k$ distributed according to $\pi$. Suppose, for the sake of contradiction, that $\mathbb{E}(\|q^0 - q^f\|_M) \overset{\Delta}{=} \mu_0 > 0$ for some $M > 0$, where the expectation is taken with respect to the randomness in $q^0$. Fix $L > 0$ such that $\mathbb{P}(q^0 \notin \mathcal{R}_L) \leq \frac{\mu_0}{4} \cdot \frac{1}{M}$. By Theorem 6, there exists $T > 0$ such that $\sup_{x \in \mathcal{R}_L} \|q(x, T) - q^f\|_M < \mu_0/4$. Therefore,
\[
\mathbb{E}(\|q^0(T) - q^f\|_M) < \mu_0/4 + M \cdot \mathbb{P}(q^0 \notin \mathcal{R}_L) \leq \mu_0/2 = \mathbb{E}(\|q^0 - q^f\|_M)/2,
\] (142)
where the first inequality uses the fact that $\|q\|_M \leq M$ for all $q \in \mathbb{R}_+^k$. Because $\mathbb{E}(\|q^0 - q^f\|_M)$ was assumed to be strictly positive, this is in contradiction with the stationarity of $\pi$ with respect to the fluid solution, which would imply that $\mathbb{E}(\|q^0(T) - q^f\|_M) = \mathbb{E}(\|q^0 - q^f\|_M)$ for all $T \in \mathbb{R}_+$. We thus conclude that
\[
\mathbb{E}(\|q^0 - q^f\|_M) = 0, \quad \forall M \in \mathbb{R}_+,
\] (143)
which in turn implies that $\pi = \delta_f$. This proves Eq. (131).

Finally, to show Eq. (32), note that by Lemma 8, for all $m \in \mathbb{N}$ and $k \in \mathcal{K}$, the random variable $\overline{Q}^m_k(\infty)$ is non-negative and stochastically dominated by $m^{-1}U_k^m$. Using similar steps as those in Eq. (182), it is not difficult to show that $\sup_{m \in \mathbb{N}} \mathbb{E}(\|m^{-1}U_k^m\|_2^2) < \infty$, which implies the uniform integrability of $\{\overline{Q}^m_k\}_{m \in \mathbb{N}}$. In combination with Eq. (131), this proves the convergence in $L^1$ stated in Eq. (32). This completes the proof of Theorem 1, as well as the first claim of Theorem 2.
6. The Memory-Deficient Regime  We now prove the second claim of Theorem 2 concerning the memory-deficient regime, where $\beta_m \ll 1/m$, as $m \to \infty$. The statement is repeated below for easy reference.

**Theorem 7.** Suppose that $\beta_m \ll 1/m$, as $m \to \infty$. Then,

$$q_k = \lambda_k \frac{\lambda_k \vee \alpha_0 + (K - 1)\alpha_0}{\sum_{i \in \mathcal{K}} [\lambda_i \vee \alpha_0 + (K - 1)\alpha_0]}, \quad k = 1, \ldots, K,$$

(144)

where $q_k = \lim_{m \to \infty} \mathbb{E} (m^{-1} Q^m_k(\infty))$.

Fix $m \in \mathbb{N}$. Let $S^m_n$ be the $n$th update point, and set $S^m_0 \triangleq 0$. Denote by

$$C^m[n] = C^m(S^m_n), \quad n \in \mathbb{Z}_+,$$

(145)

the embedded discrete-time process associated with the choice process $\{C^m(t)\}_{t \in \mathbb{R}_+}$. The discrete-time processes $\{Q^m[n]\}_{n \in \mathbb{Z}_+}$ and $\{W^m[n]\}_{n \in \mathbb{Z}_+}$ are defined analogously.

The key to proving Theorem 7 hinges upon the following property: the discrete-time choice process, $\{C^m[n]\}_{n \in \mathbb{N}}$, becomes asymptotically Markovian in the limit as $m$ tends to infinity. Note that $\{C^m[n]\}_{n \in \mathbb{N}}$ is not Markovian for a finite $m$, because the choices are sampled from a distribution that is a function of the current recallable reward vector, which in turn depends on past choices. However, the dependence of the reward vector on past choices is greatly weakened in the memory-deficient regime: because the life span of a unit reward is so small compared to the time between two adjacent choices, by the time a new choice is to be made, the agent will have essentially forgotten all of the past rewards associated with any action other than the one she currently uses. Therefore, when $m$ is large, knowing the choice $C^m[n]$ for some $n \in \mathbb{N}$, we can predict with high accuracy the recallable reward vector at the $n + 1$ update point, rendering the choice process approximately Markovian. The approximate Markovian property will allow us to explicitly characterize the steady-state distribution of $C^m[\cdot]$ in the limit as $m \to \infty$, which we then leverage to analyze the steady-state distribution of the recallable reward process, $\{Q^m(t)\}_{t \in \mathbb{R}_+}$. Following this line of thinking, our proof will consist of three main parts:

1. (Proposition 3) Fix $n \in \mathbb{N}$. We first show that, in the limit as $m \to \infty$, the value of the scaled recallable reward process at the $(n + 1)$-th update point converges in probability to a deterministic vector that only depends on $C^m[n]$.
2. (Proposition 4) Using the convergence of the recallable reward vector, we derive an explicit expression for the steady-state distribution of the discrete choice process, $\{C^m[n]\}_{n \in \mathbb{N}}$, in the limit as $m \to \infty$.
3. Using the stationarity of the original recallable reward process, we combine the above two results to arrive at a characterization of the steady-state distribution of $\{Q^m(t)\}_{t \in \mathbb{R}_+}$, thus completing the proof of Theorem 7.

We begin by stating some basic properties of $Q^m[\cdot]$ and $C^m[\cdot]$, which will allow us to relate the steady-state behavior of these processes to their continuous-time counterparts; the proof of the following lemma is given in Appendix B.8.

**Lemma 11.** Fix $m \in \mathbb{N}$. The discrete-time process $\{W^m[n]\}_{n \in \mathbb{N}}$ is positive recurrent, whose steady-state distribution satisfies

$$Q^m_k[\infty] \preceq U^m_{\lambda_1}, \quad \forall k \in \mathcal{K},$$

(146)

and

$$C^m[\infty] \overset{d}{=} C^m(\infty), \quad \text{and} \quad Q^m[\infty] \overset{d}{=} Q^m(\infty).$$

(147)
For the remainder of the proof, we will fix $h : \mathbb{Z}_+ \to \mathbb{Z}_+$ to be an increasing function, such that $1/\beta_m \gg h(m) \gg m$, and, in particular,

$$
\lim_{m \to \infty} h(m)/m = \infty, \quad \text{and} \quad \lim_{m \to \infty} h(m)\beta_m = 0. \quad (148)
$$

Define the scaled discrete-time recallable reward process: $\overline{Q}^m[n] = m^{-1}Q^m(S^m_n) = m^{-1}Q^m[n]$, $n \in \mathbb{N}$. The following proposition states that when $m$ is large, the value of $\overline{Q}^m[1]$ converges to a deterministic value that depends solely on $C^m[0]$: $\overline{Q}^m_k[1]$ is equal to $\lambda_k$ for $k = C^m[0]$, and zero, otherwise. In other words, the agent will have forgotten all past rewards other than those associated with her current choice. The proof is given in Appendix B.9.

**Proposition 3.** Define the $K \times K$ matrix $q^*$, where

$$
q^*_{k,i} = \mathbb{I}(k = i)\lambda_i, \quad k, i \in \mathcal{K}, \quad (149)
$$

Suppose that the continuous-time process $\{W^m(t)\}_{t \in \mathbb{N}}$ is initialized at $t = 0$ according to its steady-state distribution. Then, for all $i, k \in \mathcal{K},$

$$
\lim_{m \to \infty} P\left(\left|\overline{Q}^m[1] - q^*_{k,i}\right| \leq \epsilon \mid C^m[0] = k\right) = 1, \quad \forall \epsilon > 0, \quad (150)
$$

and

$$
\lim_{m \to \infty} E\left(\left|m^{-1}Q^m(h(m)) - q^*_{k,i}\right| \mid C^m[0] = k\right) = 0. \quad (151)
$$

Using Proposition 3, we will be able to obtain an explicit expression for the limit $\lim_{m \to \infty} P(C^m[\infty] = k)$, as is stated in the following proposition.

**Proposition 4.** Define $p^m_k = P(C^m[\infty] = k)$, $k \in \mathcal{K}$. We have that

$$
\lim_{m \to \infty} p^m_k = \frac{\lambda_k \lor \alpha_0 + (K - 1)\alpha_0}{z_0}, \quad k \in \mathcal{K}, \quad (152)
$$

where $z_0 = \sum_{i \in \mathcal{K}} [\lambda_i \lor \alpha_0 + (K - 1)\alpha_0]$, is a normalizing constant.

**Proof.** Suppose that $\{W^m[n]\}_{n \in \mathbb{N}}$ is initialized at $n = 0$ in its steady-state distribution. We have that, for all $i \in \mathcal{K},$

$$
P(C^m[1] = i) = \sum_{k \in \mathcal{K}} P(C^m[0] = k) \sum_{u \in \mathcal{Q}^m} \left(\frac{u \lor \alpha_0}{\sum_{j \in \mathcal{K}} (u \lor \alpha_0)} P\left(\overline{Q}^m[1] = u \mid C^m[0] = k\right)\right), \quad (153)
$$

where $\mathcal{Q}^m \triangleq \frac{1}{m} \mathbb{Z}^\mathcal{K}_+$. By Eq. (150) in Proposition 3, Eq. (153) implies that, there exists a sequence of $K \times 1$ vectors $\{\delta^m\}_{m \in \mathbb{N}}$ with $\lim_{m \to \infty} \max_{i \in \mathcal{K}} |\delta^m_i| = 0$, such that for all $i \in \mathcal{K}$ and $m \in \mathbb{N},$

$$
P(C^m[1] = i) = \delta^m_i + \sum_{k \in \mathcal{K}} P(C^m[0] = k) \frac{q^*_{k,i} \lor \alpha_0}{\sum_{j \in \mathcal{K}} (q^*_{k,j} \lor \alpha_0)}, \quad (154)
$$

where $q^*_{k,i} = \mathbb{I}(k = i)\lambda_k$. Note that by the stationarity of $C^m[\cdot]$, we have that $C^m[0] \overset{d}{=} C^m[1] \overset{d}{=} C^m[\infty]$. Therefore, by treating $p^m = (p^m_1, p^m_2, \ldots, p^m_K)^T$ as a column vector, and defining the $K \times K$ matrix, $R$, where

$$
R_{k,i} = \frac{q^*_{k,i} \lor \alpha_0}{\sum_{j \in \mathcal{K}} (q^*_{k,j} \lor \alpha_0)}, \quad \forall i, k \in \mathcal{K}, \quad (155)
$$

Eq. (154) can be written more compactly as

$$
p^m_i = \delta^m_i + (Rp^m)_i, \quad \forall i \in \mathcal{K}. \quad (156)\]
It is not difficult to verify that $R$ is row-stochastic and corresponds to the transition kernel of a discrete-time, irreducible Markov chain over $\mathcal{K}$. Therefore, the equation $x = Rx$ admits a unique solution, $x^*$, in the $K$-dimensional simplex. Furthermore, it is not difficult to check that the following choice of $x^*$ satisfies $x^* = Rx^*$, and is hence the unique solution:

$$
    x_k^* = \frac{\lambda_k \vee \alpha_0 + (K-1)\alpha_0}{\sum_{i \in \mathcal{K}} \lambda_i \vee \alpha_0 + (K-1)\alpha_0}, \quad k = 1, 2, \ldots, K.
$$

(157)

Since $\lim_{m \to \infty} \max_{i \in \mathcal{K}} |\delta_i^m| = 0$, combining Eqs. (156) and (157), we conclude that $p^m$ converges to $x^*$ coordinate-wise, as $m \to \infty$. This completes the proof of Theorem 7.

7. Proof of Theorem 1

We are now ready to complete the proof of Proposition 3. Fix $m \in \mathbb{N}$. Suppose that we initiate the continuous-time process $\{W^m(t)\}_{t \in \mathbb{N}}$ in its steady-state distribution so that it is stationary. The stationarity implies that for all $k \in \mathcal{K}$,

$$
    \mathbb{E}(m^{-1}Q_k^m(0)) = \mathbb{E}(m^{-1}Q_k^m(h(m))) = \sum_{i \in \mathcal{K}} \mathbb{P}(C^m(0) = i) \mathbb{E}(m^{-1}Q_k^m(h(m)) \mid C^m(0) = i)
$$

(158)

(159)

where $p^m_0$ is as defined in Proposition 4 and step (a) follows from the fact that $C^m[\infty] = C^m(\infty)$ (Lemma 11). Taking the limit as $m \to \infty$, we have that

$$
    q_k = \lim_{m \to \infty} \mathbb{E}(m^{-1}Q_k^m(0)) = \lim_{m \to \infty} \sum_{i \in \mathcal{K}} p_i^m \mathbb{E}(m^{-1}Q_k^m(h(m)) \mid C^m[0] = i)
$$

(158)

where step (a) follows from Eq. (151) in Proposition 3 and Eq. (152) in Proposition 4. This completes the proof of Theorem 7.

7. Proof of Theorem 1

We prove Theorem 1 in this section. Let $W^m[.] = (C^m[.], Q^m[.])$ be the discrete-time embedded Markov chain defined in Eq. (145). By Lemma 11, $C^m[\infty]$ and $C^m(\infty)$ admit the same distribution, and it hence suffices to characterize the former. Because the limiting expressions for $C^m[\infty]$ in the memory deficient regime have already been established in Proposition 4, we will focus on the memory abundant regime, where $\beta_m \gg 1/m$. By Theorem 4, for all $\epsilon > 0$

$$
    \lim_{m \to \infty} \mathbb{P}(\|m^{-1}Q^m[\infty] - q'\| > \epsilon) = \lim_{m \to \infty} \mathbb{P}(\|m^{-1}Q^m(\infty) - q'\| > \epsilon) = 0.
$$

(159)

where the first step follows from Eq. (147) in Lemma 11. Fix $k \in \mathcal{K}$. We have that

$$
    \lim_{m \to \infty} \mathbb{P}(C^m[\infty] = k) = \lim_{m \to \infty} \sum_{q \in \mathbb{R}_+} q_k \vee (m\alpha_0) \mathbb{P}(Q^m[\infty] = q) = \frac{q_k \vee \alpha_0}{\sum_{i} q_i \vee \alpha_0}.
$$

(160)

where the last equality follows from Eq. (159) and the fact that $\frac{q_k \vee (m\alpha_0)}{\sum_{i} q_i \vee (m\alpha_0)}$ is uniformly bounded over $q \in \mathbb{R}_+$. Eq. (160), along with the expression for $q'$ in Theorem 3, leads to Eqs. (5) and (6), depending on the value of $k$ and $\alpha_0$. This completes the proof of Theorem 1.
8. Discussion We have proposed in this paper a stochastic action-reward model for studying the impact of imperfect memory in dynamic decision making. In the limiting regimes where the agent’s memory span is large, our main results provide exact expressions for the steady-state choice probabilities of an agent following the so-called reward-matching rule, and demonstrate that these probabilities are highly sensitive to the relative scaling between two parameters: the rate of memory decay, \( \mu \), and the rate at which the agent makes new choices, \( \beta \).

There is a number of potentially interesting extensions of our formulation. In the present limiting regime where \( \mu \to 0 \), our main results have left out the case where \( \beta = \Theta(\mu) \) as \( \mu \to 0 \). The complication here is that neither \( Q(\cdot) \) nor \( C(\cdot) \) becomes Markovian as \( \mu \) diminishes, and our current proof technique, which exploits the asymptotic Markovian property, does not extend easily. Alternatively, one may also look at an opposite limiting regime where \( \mu \) tends to infinity, i.e., the agent becomes extremely forgetful and her memory span approaches zero. Note that since the total amount of recallable rewards in system is on the order of \( \mu^{-1} \) (regardless of the value of \( \beta \)), in this limit we should see very few recallable rewards at any point in time and the process of recallable rewards frequently hitting the zero state. It is therefore reasonable to assume that the reward-matching rule will perform poorly when \( \mu \) is very large, with the agent choosing all actions essentially uniformly at random. Nevertheless, it may still be interesting to understand the system dynamics when the magnitude of \( \mu \) is moderate. Unfortunately, we do not yet have the tools needed to characterize this regime, since the fluid approximations developed in this paper heavily rely on mean-field-type evolution of a large amount of recallable rewards and are unlikely to be applicable when \( \mu \) is large.

The reward matching rule is a special case of a broader family of choice heuristics, in which action \( k \) is chosen with a probability proportional to \( w(Q_k(t)) \lor \alpha \), where \( w: \mathbb{R}_+ \to \mathbb{R}_+ \) is a non-decreasing weight function. For instance, when \( w(x) = \exp(cx) \) for some constant \( c > 0 \), the sampling procedure becomes reminiscent of the celebrated logit model in choice theory (cf. Chapter 3, Train (2009)), as well as the exponential-weight algorithm in the multi-arm bandit literature (cf. Auer et al. (2002)). We believe that it is possible to extend our results to this broader family of heuristics with additional work. For instance, in the memory-deficient regime, the limiting probability of choosing action \( k \) would become proportional to \( w(\lambda_k) \lor \alpha_0 + (K-1)\alpha_0 \), i.e., \( \lambda_k \) is replaced by \( w(\lambda_k) \). Note that if we had the freedom of choosing \( w(\cdot) \), then applying a highly skewed weight function, e.g., \( w(x) = 10^{20x} \), would induce the agent to almost always choose the best action in steady state, even in the memory-deficient regime. However, the downside is that with a skewed weight function, the agent is highly unlikely to give up the current choice at an update point, regardless of its reward rate. It will thus take a long time for the agent’s time-average behavior to approach that of the steady state, resulting in bad transient performance. It would be interesting to better understand the tradeoff between the agent’s steady-state versus transient performance as the weight function varies.

While the present paper focuses largely on the theory, there appears to be several interesting potential applications of our model. First, the reward matching rule can be viewed as a simplified cognitive model to capture humans’ unconscious reinforcement behavior, and it will be interesting to see if the theory could be used to make predictions on a consumer’s long-run choice behavior when switching among similar products or services (eg., Example 1 of Section 1), or to investigate the effect of memory loss on a player’s learning and performance in repeated games similar to those studied by Erev and Roth (1998). Second, memory loss can be interpreted in a more metaphorical sense, and be used to model the departure of customers (eg., Example 2 of Section 1). Cast in this light, the reward matching policy may act as a simple and intuitive heuristic in a manager’s algorithmic toolbox for dynamically choosing product offerings or demographic targets for advertisements.

While memory loss is regarded as a given constraint in the current work, one could also ask whether it would be beneficial to artificially induce memory loss even when perfect recall is feasible.
For instance, it has been observed in the multi-arm bandit literature that policies with built-in regularization that penalizes distant past experience can achieve better regret with time-varying, or adversarially chosen, parameters (cf. Garivier and Moulines (2008), Van Erven and Kotl (2014), Besbes et al. (2015), Keskin and Zeevi (2016)). This is because these policies tend to be more adaptive to the environment and not “weighed down” by past experience. While our theorems do not directly apply to time-varying reward rates, Theorem shows that the fluid solution converges exponentially quickly to the invariant state from any bounded initial condition, suggesting that the reward matching rule could be an attractive algorithm for dynamic learning in non-stationary environments.

Finally, at a higher level, there appear to be other interesting ramifications of “memory loss” in organizations to be explored. For instance, departures of employees could lead to the loss of skills and expertise (cf. Benkard (2000)), and such organizational forgetting could potentially affect a firm’s efficiency in a significant way.

References
Appendix A: Technical Preliminaries

**Proposition 5 (Doob’s inequality (cf. Section 12.6, Grimmett and Stirzaker (2001))).** Let \( \{X(t)\}_{t \geq 0} \) be a discrete- or continuous-time non-negative submartingale. Fix \( T \geq 0 \). We have that
\[
\mathbb{P} \left( \sup_{0 \leq t \leq T} X(t) \geq c \right) \leq \frac{\mathbb{E}(X(T))}{c}, \quad \forall c > 0.
\]

**Proposition 6 (Gronwall’s lemma (cf. Section 1.3, Ames and Pachpatte (1997))).** Let \( f, b : \mathbb{R}_+ \rightarrow \mathbb{R}_+ \) be continuous and non-negative functions, and let \( a : \mathbb{R}_+ \rightarrow \mathbb{R}_+ \) be a continuous, positive and non-decreasing function. If \( f(t) \leq a(t) + \int_0^t b(s)f(s)ds \), for all \( t \in [0, T] \), then
\[
f(T) \leq a(T) \exp \left( \int_0^T b(t) dt \right).
\]

Appendix B: Proofs

**B.1. Proof of Lemma**

The existence and uniqueness of the fluid solution follow from Picard’s existence theorem (Section 2, Chapter 1, Coddington and Levinson (1955)) by verifying that the right-hand side of Eq. (15) is uniformly Lipschitz-continuous in \( q(t) \) over \( \mathbb{R}_+^K \), and (trivially) continuous in \( t \). To show the fluid solution’s continuous dependence on initial condition, note that because of the Lipschitz...
a representation of the Markov process \( W \) of arguments similar to that used in the proof of Theorem 2.2 of Kurtz (1978). First, we introduce

\[
\|q(x, t) - q(y, t)\| \leq \|x - y\| + \int_0^t \| (p(q(x, s)) - q(x, s)) - (p(q(y, s)) - q(y, s)) \| ds
\]

\[
\leq \|x - y\| + l \int_0^t \| q(x, s) - q(y, s) \| ds
\]

\[
\leq \|x - y\| e^{lt}, \tag{161}
\]

where the last inequality follows from Gronwall’s lemma (Proposition 6). Therefore, \( \lim_{x \to y} q(x, t) = q(y, t) \) for all \( y \in \mathbb{R}_+^k \). This completes the proof of the lemma. Q.E.D.

**B.2. Proof of Lemma 2**

Proof. Define

\[
\gamma(t) = N(t) - \int_0^t \psi(W^m(s)) ds, \quad t \in \mathbb{R}_+.
\tag{162}
\]

Let \( \{F_t\}_{t \in \mathbb{R}_+} \) be the natural filtration associated with \( W^m(\cdot) \). It is not difficult to show, by the definition of \( N(\cdot) \), that \( \{\gamma(t)\}_{t \in \mathbb{R}_+} \) is martingale with respect to \( \{F_t\}_{t \in \mathbb{R}_+} \). Define the stopping time

\[
T_s = \inf\{t : \psi(W^m(t)) \geq \phi\}. \tag{163}
\]

Let \( \{\tilde{N}(t)\}_{t \in \mathbb{N}} \) be a counting process defined as:

\[
\tilde{N}(t) = N(t \wedge T_s), \quad t \in \mathbb{R}_+.
\tag{164}
\]

That is, \( \tilde{N}(t) \) coincides with \( N(t) \) up until \( t = T_s \), and stays constant afterwards. Let

\[
\tilde{\psi}(t) \overset{\Delta}{=} \psi(W^m(t)) I(t \leq T_s). \tag{165}
\]

Then, it is not difficult to show that \( \tilde{N}(\cdot) \) is a counting process whose instantaneous rate at time \( t \) is \( \tilde{\psi}(t) \), and the process

\[
\tilde{\gamma}(t) = \tilde{N}(t) - \int_0^t \tilde{\psi}(s) ds, \quad t \in \mathbb{R}_+,
\]

is a martingale with respect to \( \{F_t\}_{t \in \mathbb{R}_+} \). From the definition of \( \tilde{\gamma}(\cdot) \) and \( \gamma(\cdot) \), we have that, for all \( \epsilon > 0 \),

\[
\mathbb{P}\left( \sup_{0 \leq t \leq T} |\gamma(t)| \geq \epsilon \right) = \mathbb{P}\left( \sup_{0 \leq t \leq T} |\gamma(t)| \geq \epsilon, T_s > T \right) + \mathbb{P}\left( \sup_{0 \leq t \leq T} |\gamma(t)| \geq \epsilon, T_s \leq T \right)
\]

\[
\leq \mathbb{P}\left( \sup_{0 \leq t \leq T} |\tilde{\gamma}(t)| \geq \epsilon \right) + \mathbb{P}(T_s \leq T)
\]

\[
= \mathbb{P}\left( \sup_{0 \leq t \leq T} |\tilde{\gamma}(t)| \geq \epsilon \right) + \mathbb{P}\left( \sup_{0 \leq t \leq T} \psi(W^m(t)) \geq \phi \right). \tag{166}
\]

To complete the proof, therefore, it suffices to show that

\[
\mathbb{P}\left( \sup_{0 \leq t \leq T} |\tilde{\gamma}(t)| \geq \epsilon \right) \leq 2e^{-\phi T - h(\epsilon/\phi) T}. \tag{167}
\]

We now show Eq. (167) using Doob’s inequality (Proposition 5 in Appendix A), by following a line of arguments similar to that used in the proof of Theorem 2.2 of Kurtz (1978). First, we introduce a representation of the Markov process \( W^m(\cdot) \) using Poisson processes. Let \( \{\Xi_y(\cdot)\}_{y \in \mathbb{Z}^{k+1}} \) a family
of mutually independent unit-rate Poisson counting processes, indexed by \( \mathbb{Z}^{K+1} \). For every \( y \in \Xi_y \), let \( \psi(y) \) be the rate function of \( W^m(\cdot) \) for the jump with value \( y \), i.e., \( \psi(y) \) is the instantaneous rate at which \( W^m(\cdot) \) jumps to state \( w + y \) when in state \( w \in \mathbb{Z}_+^{K+1} \). Then, the process \( W^m(\cdot) \) can be expressed as a solution to the following integral equation:

\[
W^m(t) = W^m(0) + \sum_{y \in \Xi_y} y \Xi_y \left( \int_0^t \psi_y(W^m(s)) ds \right), \quad t \in \mathbb{R}_+.
\]  

(168)

In this representation, \( \Xi_y \left( \int_0^t \psi_y(W^m(s)) ds \right) \) counts the number of jumps of value \( y \) over the interval \([0, t]\). We thus have that, by setting \( y = l \),

\[
N(t) = \Xi_l \left( \int_0^t \psi(W^m(s)) ds \right), \quad t \in \mathbb{R}_+,
\]

(169)

and

\[
\hat{\gamma}(t) = \tilde{N}(t) - \int_0^t \tilde{\psi}(s) ds = \Xi_l \left( \int_0^t \tilde{\psi}(s) ds \right) - \int_0^t \tilde{\psi}(s) ds, \quad t \in \mathbb{R}_+.
\]

(170)

Fix \( \theta > 0 \). Since \( \hat{\gamma}(\cdot) \) is a martingale and \( \exp(\cdot) \) a positive convex function, \( \{\exp(\theta\hat{\gamma}(t))\}_{t \in \mathbb{R}_+} \) is a submartingale. From Eq. (170), we have that

\[
\mathbb{E}(\exp(\hat{\gamma}(T))) = \mathbb{E}(\exp(\Xi_l(\tau_T) - \tau_T)),
\]

(171)

where \( \tau_t = \int_0^t \tilde{\psi}(s) ds \). Note that \( \tau_T \) is a stopping time with respect to \( \Xi_l(\cdot) \). Since by definition \( \tilde{\psi}(t) \leq \phi \) for all \( t \), we have that \( \tau_T \leq \phi T \). Applying the optional sampling theorem for submartingales indexed by partially ordered sets (cf. [Washburn and Willsky 1981]) to \( \{\exp(\theta\hat{\gamma}(t))\}_{t \in \mathbb{R}_+} \), we have that

\[
\mathbb{E}(\exp(\theta\hat{\gamma}(T))) \leq \mathbb{E}(\exp(\theta\Xi_l(\phi T) - \theta \phi T)) \leq \exp\left(\left(e^\theta - \theta - 1\right) \phi T\right),
\]

(172)

where the last inequality follows from the fact that \( \Xi_l(\phi T) \) is a Poisson random variable with mean \( \phi T \) whose moment generating function is given by \( \mathbb{E}(\exp(a\Xi_l(\phi T))) = \exp(\phi T(e^a - 1)) \), \( a \in \mathbb{R} \). Analogously, we can show that

\[
\mathbb{E}(\exp(-\theta\hat{\gamma}(T))) \leq \exp\left(\left(e^{-\theta} + \theta - 1\right) \phi T\right) \leq \exp\left(\left(e^\theta - \theta - 1\right) \phi T\right),
\]

(173)

where the last inequality follows from the fact that \( e^\theta - e^{-\theta} \geq 2\theta \) for all \( \theta \geq 0 \).

Note that for any \( \theta > 0 \), both \( \{\exp(\theta\hat{\gamma}(t))\}_{t \geq 0} \) and \( \{\exp(-\theta\hat{\gamma}(t))\}_{t \geq 0} \) are non-negative submartingales. Using Doob’s inequality and Eqs. (172) and (173), we have that, for all \( \epsilon, \theta > 0 \),

\[
\mathbb{P}\left( \sup_{0 \leq t \leq T} \hat{\gamma}(t) \geq \epsilon \right) + \mathbb{P}\left( \sup_{0 \leq t \leq T} -\hat{\gamma}(t) \geq \epsilon \right) \leq \mathbb{E}(\exp(\theta\hat{\gamma}(T))) \geq \exp(\theta\epsilon) + \mathbb{P}(\exp(-\theta\hat{\gamma}(T)) \geq \exp(\theta\epsilon)) \leq \frac{\mathbb{E}(\exp(\theta\hat{\gamma}(T)))}{\exp(\theta\epsilon)} + \frac{\mathbb{E}(\exp(-\theta\hat{\gamma}(T)))}{\exp(\theta\epsilon)} \leq 2\exp(\phi T(\epsilon^\theta - \theta - 1) - \theta\epsilon).
\]

(174)

By setting \( \theta = \log(1 + \epsilon/\phi T) \) in Eq. (174), we conclude that

\[
\mathbb{P}\left( \sup_{0 \leq t \leq T} |\hat{\gamma}(t)| \geq \epsilon \right) \leq 2 \exp(-\phi T \cdot h(\epsilon/\phi T)),
\]

where \( h(x) \equiv (1 + x) \log(1 + x) - x \). This completes the proof. Q.E.D.
B.3. Proof of Lemma 3

Proof. Define \( Z = \frac{1}{(1+\epsilon)r_y} \sum_{i=1}^{(1+\epsilon)r_y} X_i \). In particular, \( \mathbb{E}(Z) = r^{-1} \) and \( \text{STD}(Z) = \frac{1}{\sqrt{(1+\epsilon)r_y}} r^{-1} \), where STD(\( Z \)) stands for the standard deviation of Z. We have that

\[
\mathbb{P}\left( \sum_{i=1}^{(1+\epsilon)r_y} X_i \leq y \right) \leq \mathbb{P}\left( |Z - r^{-1}| \geq \frac{\epsilon}{(1+\epsilon)r} \right) = \mathbb{P}\left( |Z - r^{-1}| \geq \epsilon \sqrt{\frac{r_y}{1+\epsilon}} \text{STD}(Z) \right) \leq \frac{1 + \epsilon}{r \epsilon^2}, \tag{175}
\]

where the last step follows from Chebyshev’s inequality. Q.E.D.

B.4. Proof of Lemma 4

Proof. We show the result by induction. For the base case, we extend the definition of \( \{\xi_i^m\}_{i \in \mathbb{N}} \) by letting \( \xi_0^m = 0 \), and it is not difficult to see that the inequality holds when \( n = 0 \).

Fix \( i \in \{0, \ldots, n - 1\} \), and suppose that

\[
\mathbb{E}\left( \exp\left( \theta \xi_i^m \right) \right) \leq \exp\left( \frac{i \theta^2 / 4}{m \beta_m (m \beta_m - \theta)} \right). \tag{176}
\]

In light of the base case, it then suffices to show that the above equation implies

\[
\mathbb{E}\left( \exp\left( \theta \xi_i^{m+1} \right) \right) \leq \exp\left( \frac{(i + 1) \theta^2 / 4}{m \beta_m (m \beta_m - \theta)} \right). \tag{177}
\]

Let \( \{C_l\}_{l \in \mathbb{Z}_+} \) be the natural filtration induced by \( \{\tau_i^m, Z_k^m[l], \overline{Q}^m[l]\}_{l \in \mathbb{N}} \), with \( C_0 \triangleq \emptyset \). We have that

\[
\mathbb{E}\left( \exp\left( \theta \xi_i^{m+1} | C_i \right) \right) = \mathbb{E}\left( \exp\left( \theta \xi_i^m \right) \exp\left( \theta \tau_i^{m+1} \left( Z_k^m[i+1] - p_k(\overline{Q}^m[i+1]) \right) \right) | C_i \right)
\]

\[= \left( \frac{\exp\left( \theta \tau_i^{m+1} \left( Z_k^m[i+1] - p_k(\overline{Q}^m[i+1]) \right) \right)}{\exp\left( \theta \tau_i^{m+1} \left( Z_k^m[i+1] - p_k(\overline{Q}^m[i+1]) \right) \right)} \right) | C_i \right), \tag{178}
\]

where step (a) follows from \( \xi_i^m \) being \( C_i \)-measurable. We now develop an upper bound for the second term on the right-hand side of Eq. (178), as follows.

\[
\mathbb{E}\left( \exp\left( \theta \tau_i^{m+1} \left( Z_k^m[i+1] - p_k(\overline{Q}^m[i+1]) \right) \right) | C_i \right) = \mathbb{E}\left( \exp\left( \theta \tau_i^{m+1} \right) \left( Z_k^m[i+1] - p_k(\overline{Q}^m[i+1]) \right) \right) | C_i \right)
\]

\[= \left( \frac{\exp\left( \theta \tau_i^{m+1} \right)}{\exp\left( \theta \tau_i^{m+1} \right)} \right) | C_i \right), \tag{179}
\]
Step (a) follows from the fact that, for a given value of $Q^m[i + 1]$, $Z^m[i + 1]$ is a Bernoulli random variable with $P(Z^m[i + 1] = 1) = p_k(Q^m[i + 1])$, and is independent from $\tau_{i+1}^m$. For step (b), note that $\tau_{i+1}$ is an exponential random variable with mean $(m\beta_m)^{-1}$, independent from $C_i$ and $Q^m[i + 1]$. Its moment generating function is given by $E(e^{\theta \tau_{i+1}}) = \frac{\beta_m}{\beta_m - h}$ for all $h \leq \beta_m$, where $h$, in our case, corresponds to $\theta(1 - p_k(Q^m[i + 1]))$ and $\theta p_k(Q^m[i + 1])$, for the two terms respectively. Step (c) stems from the fact that $1 + x \leq e^x$ for all $x \in \mathbb{R}_+$. Finally, for step (d) we have used the fact that $p_k(Q^n[i + 1]) \in [0, 1]$ by definition, and that $\theta < m\beta_m$; the exponent is hence bounded from above by setting $p_k(Q^n[i + 1])$ to $1/2$ and 0 in the numerator and denominator, respectively.

Substituting Eq. (179) into Eq. (178), and invoking the induction hypothesis of Eq. (176), we have that

$$E(\exp(\theta \tau_{i+1}^m)) = E\left(E(\exp(\theta \tau_{i+1}^m) \mid C_i)\right)$$

$$= E\left(\exp(\theta \tau_{i+1}^m) E(\exp(\theta \tau_{i+1}^m (Z^m[i + 1] - p_k(Q^m[i + 1]))) \mid C_i)\right)$$

$$\leq \exp\left(\frac{\theta^2}{4}\frac{m\beta_m(m\beta_m - \theta)}{(i + 1)\theta^2/4 + m\beta_m(m\beta_m - \theta)}\right)$$

$$\leq \exp\left(\frac{\theta^2}{4}\frac{m\beta_m(m\beta_m - \theta)}{m\beta_m(m\beta_m - \theta)}\right).$$

This proves our claim. Q.E.D.

**B.5. Proof of Lemma 5**

**Proof.** Fix $\epsilon > 0$ and $m \in \mathbb{N}$. We have that

$$P\left(\sum_{i=1}^{I_mT+1} (\tau_{i}^m)^2 \geq \epsilon\right) \leq P\left(\sum_{i=1}^{I_mT+1} (\tau_{i}^m)^2 \geq \epsilon, I_mT < (1 + \epsilon)m\beta_mT\right) + P(I_mT \geq (1 + \epsilon)m\beta_mT)$$

$$\leq P\left(\sum_{i=1}^{(1+\epsilon)m\beta_mT} (\tau_{i}^m)^2 \geq \epsilon\right) + \frac{1 + \epsilon}{Te^2} (m\beta_m)^{-1}$$

$$\leq P\left(\sum_{i=1}^{(1+\epsilon)m\beta_mT - \mathbb{E}((\tau_{i}^m)^2)} (\tau_{i}^m)^2 \geq \epsilon\right) + \frac{1 + \epsilon}{Te^2} (m\beta_m)^{-1}$$

$$= 2(1 + \epsilon)T(em\beta_m)^{-1} + \frac{1 + \epsilon}{Te^2} (m\beta_m)^{-1}.$$  

where step (a) follows from Eq. (69), (b) from the Markov’s inequality, and (c) from $\tau_{i}^m$ being an exponentially distributed random variable with mean $(m\beta_m)^{-1}$ and hence $\mathbb{E}((\tau_{i}^m)^2) = 2(m\beta_m)^{-2}$. Because $m\beta_m \to \infty$ as $m \to \infty$, the claim follows. Q.E.D.

**B.6. Proof of Lemma 8**

**Proof.** We will use a simple coupling argument as follows. Fixing $k \in K$, the evolution of the process $U^m_k(\cdot)$ corresponds to that of $Q^m_k(\cdot)$ if action $k$ was selected for all $t \in \mathbb{R}_+$, and it follows, given the same initial condition, that $Q^m_k(t)$ is stochastically dominated by $U^m_k(t)$ for all $t \in \mathbb{R}_+$. Since $U^m_k(\cdot)$ is positive recurrent for all $k \in K$, we know that $Q^m(\cdot)$ is also positive recurrent, which in turn implies the positive recurrence of $W^m(\cdot)$, because the evolution of $C^m(\cdot)$ is derived by sampling from $K$ based solely on the value of $Q^m(\cdot)$. Lemma 8 follows from the above-mentioned stochastic dominance, the fact that under any finite initial condition, $U^m_k(t)$ converges in distribution to $U^m_{\lambda}$ as $t \to \infty$, and the observation that $U^m_{\lambda} \leq U^m_{\lambda'}$ whenever $\lambda \geq \lambda' \geq 0$. Q.E.D.

**B.7. Proof of Lemma 9**
Proof. To show Eq. (128), observe that
\[
\limsup_{x \to \infty} \sup_{m \in \mathbb{N}} (1 - \pi^m_W(\mathcal{Q}_x)) = \limsup_{x \to \infty} \sup_{m \in \mathbb{N}} \mathbb{P}(\max_{k \in \mathcal{K}} Q^m_k(\infty) > x) \leq \limsup_{x \to \infty} K \mathbb{P}(U^m_{\lambda_1}/m > x) \\
= \limsup_{x \to \infty} K e^{-m \lambda_1} \sum_{i=m \bar{x}}^{\infty} \frac{(m \lambda_i)^i}{i!} \leq \limsup_{x \to \infty} K \sum_{i=m \bar{x}}^{\infty} \left( \frac{e \lambda_1}{x} \right)^i \\
\leq \limsup_{x \to \infty} K 2^{-i(x-1)} = 0,
\]
(182)
where step (a) follows from Lemma 8 and the union bound, and (b) from the elementary inequality \( il \geq (i/e)^i \). Eq. (182) thus shows that for all \( \epsilon, \inf_{m \in \mathbb{N}} \pi^m_W(\mathcal{Q}_x) > 1 - \epsilon \) for all sufficiently large \( x \), which proves Eq. (128). Q.E.D.

B.8. Proof of Lemma 11

Proof. It is not difficult verify that \( \{W^m[n]\}_{n \in \mathbb{Z}_+} \) is a time-homogeneous, aperiodic, and irreducible Markov chain. Because the continuous-time process, \( \{W^m(t)\}_{t \in \mathbb{R}_+} \), is positive recurrent, so is its discrete-time counterpart, \( \{W^m[n]\}_{n \in \mathbb{N}} \), and \( W^m[n] \) converges to its steady-state distribution, \( W^m[\infty] \), as \( n \to \infty \). Eq. (146) follows from the same argument as that of Lemma 8. We now show Eq. (147). Denote by \( N^m(t) \) the index of the last update point by time \( t \):
\[
N^m(t) = \sup\{n : S^m_n \leq t\},
\]
(183)
and by \( T^m(t) \) its value
\[
T^m(t) = S^m_{N^m(t)}.
\]
(184)
Recall that the update points are generated according to a Poisson process, and it not difficult to show that, almost surely, \( N^m(t) \to \infty \) and \( T^m(t) \to \infty \), as \( t \to \infty \). We thus have that, for all \( k \in \mathcal{K} \),
\[
\mathbb{P}(C^m[\infty] = k) = \lim_{n \to \infty} \mathbb{P}(C^m[n] = k) = \lim_{t \to \infty} \mathbb{P}(C^m(T^m(t)) = k) = \mathbb{P}(C^m(\infty) = k).
\]
(185)
The same argument applies for \( Q^m[\infty] \) versus \( Q^m(\cdot) \). This completes the proof of Lemma 11. Q.E.D.

B.9. Proof of Proposition 3

Proof. We begin by showing the following, strengthened version of Lemma 8 which states that a similar stochastic dominance property holds for \( Q^m(\cdot) \) even when conditioning on \( C^m(\cdot) \) being of a specific value.

Lemma 12. Let \( W^m = (Q^m, C^m) \) be a random vector drawn from the steady-state distribution, \( W^m(\infty) \). There exist constants \( m_0 \) and \( \gamma > 0 \), such that for all \( m \geq m_0 \),
\[
Q^m_i \mid \{C^m = k\} \geq \gamma U^m_{\lambda_1}, \quad \forall i, k \in \mathcal{K}.
\]
(186)
Proof. Let \( Q^m = \{x/m : x \in \mathbb{Z}_+\} \). We have that, for all \( k \in \mathcal{K}, n \in \mathbb{N} \),
\[
\mathbb{P}(C^m[n] = k) = \sum_{u \in Q^m} \frac{u_k \vee \alpha_0}{\sum_{i \in \mathcal{K}} (u_i \vee \alpha_0)} \mathbb{P} (\mathcal{Q}^m[n] = u) \geq \sum_{u \in Q^m} K \alpha_0 + \sum_{i \in \mathcal{K}} u_i \mathbb{P} (\mathcal{Q}^m[n] = u)
\]
(187)
By Eq. (147) of Lemma 11 the above equation implies that
\[
P(C^m = k) = \lim_{n \to \infty} P(C^m[n] = k) \geq \sum_{u \in \Omega^m} \frac{\alpha_0}{K\alpha_0 + \sum_{i \in K} u_i} P(Q^m[\infty] = u).  
\] (188)

where the last step follows from the fact that \(\frac{\alpha_0}{K\alpha_0 + \sum_{i \in K} u_i}\) is always bounded from above by \(1/K\).

By Eq. (126), we have that \(U_{\lambda_1}^m / m \to \lambda_1\) almost surely as \(m \to \infty\). Therefore, there exist \(m_0\) and \(y > 0\), such that
\[
P \left( \frac{1}{m} U_{\lambda_1}^m \geq \lambda_1 + y \right) \leq \frac{1}{2K^2}, \quad \forall m \geq m_0.  
\] (189)

Combining Eq. (146) in Lemma 11 with Eqs. (188) and (189), we have that, for all \(m \geq m_0\),
\[
P(C^m = k) \geq \sum_{u \in \Omega^m} \frac{\alpha_0}{K\alpha_0 + \sum_{i \in K} u_i} P(Q^m[\infty] = u) \\
\geq P \left( \max_{i \in K} Q_i^m[\infty] \leq \lambda_1 + y \right) \frac{\alpha_0}{K\alpha_0 + K(\lambda_1 + y)} + P \left( \max_{i \in K} Q_i^m[\infty] > \lambda_1 + y \right) \cdot 0 \\
\geq \left( 1 - K P \left( \frac{1}{m} U_{\lambda_1}^m \leq \lambda_1 + y \right) \right) \frac{\alpha_0}{K\alpha_0 + K(\lambda_1 + y)} \\
= \gamma^{-1},  
\] (190)

where \(\gamma \triangleq \frac{K\alpha_0 + K(\lambda_1 + y)}{\alpha_0 (1 - 1/2K)}\). Step (a) follows from Eq. (188), (b) from Lemma 8 and a union bound, and (c) from Eq. (189).

Fix \(x \in \mathbb{Z}_+\). We have that, for all \(m \geq m_0\),
\[
P \left( Q_i^m \geq x \mid C^m = k \right) = \frac{P \left( Q_i^m \geq x, C^m = k \right)}{P(C^m = k)} \leq \frac{P \left( Q_i^m \geq x \right)}{P(C^m = k)} \\
\leq \frac{P \left( U_{\lambda_1}^m \geq x \right)}{P(C^m = k)} \\
\leq \gamma P \left( U_{\lambda_1}^m \geq x \right),  
\] (191)

for all \(i, k \in K\), where step (a) follows from Lemma 8 and (b) from Eq. (190). Since the above inequality holds for all \(x \in \mathbb{Z}_+\), this completes the proof of Lemma 12. Q.E.D.

We now prove the convergence in Eq. (150). Recall that the first update point, \(S_i^m\), is exponentially distributed with mean \(1/\beta_m\), and independent from \(W^m(0)\). Define the event \(\mathcal{E}^m = \{ S_i^m \geq h(m) \}\). We have that
\[
P(\mathcal{E}^m) = P(S_i^m \geq h(m)) = \exp(-\beta_m h(m)) \to 1, \quad \text{as} \ m \to \infty.  
\] (192)

Fix \(i, k \in K\). Recall from Eq. (126) that \(U_{\lambda_1}^m / m\) converges to \(\lambda_1\) almost surely as \(m \to \infty\). This implies that there exists \(v > 0\), independent of \(m\), such that
\[
\limsup_{m \to \infty} P \left( \overline{Q_i^m}[0] > v \mid C^m[0] = k \right) \overset{(a)}{=} \limsup_{m \to \infty} P (U_{\lambda_1}^m / m > \gamma^{-1} v) = 0,  
\] (193)

where step (a) follows from Lemma 12.

Fix \(x \in \mathbb{R}_+\). We have that
\[
\limsup_{m \to \infty} P \left( \overline{Q_i^m}[1] \geq x \mid C^m[0] = k \right) \overset{(a)}{=} \limsup_{m \to \infty} \left[ P \left( \overline{Q_i^m}[1] \geq x \mid C^m[0] = k, \mathcal{E}^m \right) + (1 - P(\mathcal{E}^m) | C^m[0] = k) \right] \\
\overset{(b)}{=} \limsup_{m \to \infty} \left[ P \left( \overline{Q_i^m}[1] \geq x \mid C^m[0] = k, \mathcal{E}^m \right) + (1 - P(\mathcal{E}^m)) \right].  
\] (194)
where step (a) follows from the independence between $E^m$ and $C^m[0]$, and (b) from Eq. (192).

We now bound the term on the right-hand side of Eq. (194). Denote by $B(n, p)$ a binomial random variable with $n$ trials and a success probability of $p$ per trial, and by $U^m(t)$ the number of jobs in system at time $t$ in an initially empty $M/M/\infty$ queue with arrival rate $\lambda$ and departure rate $1/m$. We have that

$$P \left( \{ \mathcal{Q}_i^m \geq x \} \mid C^m[0] = k, E^m \right) \overset{(a)}{=} P \left( \frac{1}{m} U_{\lambda_k}^m (S_1^m) \mathbb{1} (i = k) + B \left( Q_i^m (0), e^{-S_1^m/m} \right) \geq x \mid C^m[0] = k, E^m \right)$$

$$\leq P \left( \frac{1}{m} U_{\lambda_k}^m (S_1^m) \mathbb{1} (i = k) + B \left( Q_i^m (0), e^{-h(m)/m} \right) \geq x \mid C^m[0] = k \right)$$

$$\leq P \left( \frac{1}{m} U_{\lambda_k}^m (S_1^m) \mathbb{1} (i = k) + B \left( Q_i^m (0), e^{-h(m)/m} \right) \geq x \right) + P \left( \mathcal{Q}_i^m (0) > v \mid C^m[0] = k \right). \quad (195)$$

For step (a), note that each unit of reward initially present at time $t = 0$ has probability of $\exp(-S_1^m/m)$ or remaining in the system by $t = S_1^m$. Therefore, the rewards in site $i$ at time $S_1^m$ satisfy the following decomposition:

$$Q_i^m[1] \overset{d}{=} U_{\lambda_k}^m (S_1^m) \mathbb{1} (i = k) + B \left( Q_i^m (0), e^{-S_1^m/m} \right). \quad (196)$$

The first term corresponds to the units of rewards at $t = S_1^m$ that had arrived during the interval $(0, S_1^m)$, and hence is non-zero only if $i = k$. The second term corresponds to those individuals initially present at $t = 0$ who remained in the system by $t = S_1^m$. Step (b) follows from the definition of $E^m$, and (c) from the well-known fact that the number of jobs in system in an initially empty $M/M/\infty$ queue at any time is always stochastically dominated by its steady-state distribution.

Because $h(m)/m \to 0$ as $m \to \infty$, we have that $\lim_{m \to \infty} E \left( \frac{1}{m} B \left( vm, e^{-h(m)/m} \right) \right) = 0$. Applying Markov’s inequality, we obtain that

$$\frac{1}{m} B \left( vm, e^{-h(m)/m} \right) \overset{P}{\to} 0, \quad \text{as } m \to \infty, \quad (197)$$

where $\overset{P}{\to}$ denotes convergence in probability. Recall from Eq. (126) that, almost surely,

$$\frac{1}{m} U_{\lambda_k}^m \mathbb{1} (i = k) \to \lambda_k \mathbb{1} (i = k) = q_{k,i}. \quad (198)$$

Fix $\epsilon > 0$, and substitute Eq. (195) into Eq. (194). We have that

$$\limsup_{m \to \infty} P \left( \mathcal{Q}_i^m [1] \geq q_{k,i}^* + \epsilon \mid C^m[0] = k \right) = \limsup_{m \to \infty} P \left( \mathcal{Q}_i^m [1] \geq q_{k,i}^* + \epsilon \mid C^m[0] = k, E^m \right)$$

$$\leq \limsup_{m \to \infty} \left\{ \frac{1}{m} U_{\lambda_k}^m (i = k) + B \left( Q_i^m (0), e^{-h(m)/m} \right) \geq q_{k,i}^* + \epsilon \right\} + \limsup_{m \to \infty} P \left( \mathcal{Q}_i^m (0) > v \mid C^m[0] = k \right)$$

$$\overset{(a)}{=} \limsup_{m \to \infty} \left\{ \frac{1}{m} U_{\lambda_k}^m (i = k) + B \left( Q_i^m (0), e^{-h(m)/m} \right) \geq q_{k,i}^* + \epsilon \right\}$$

$$\overset{(b)}{=} 0, \quad (199)$$
where step (a) follows from Eq. (193), and (b) from Eqs. (197) and (198). Using the same line of arguments as that in Eqs. (194) through (199), we can show that
\[
\limsup_{m \to \infty} P\left( Q_{i}^{m}[1] \leq q_{k,i}^{*} - \epsilon \mid C^{m}[0] = k \right) = 0,
\]
which, along with Eq. (199), yields that
\[
\limsup_{m \to \infty} P\left( |Q_{i}^{m}[1] - q_{k,i}^{*}| > \epsilon \mid C^{m}[0] = k \right) = 0. \tag{201}
\]
Since the above equation holds for all \(i, k \in K\), this proves Eq. (150) in Proposition 3

We now turn to Eq. (151). Fix \(i, k \in K\). Using essentially identical arguments as those for Eq. (201), we can show that
\[
\limsup_{m \to \infty} P\left( |m^{-1}Q_{i}^{m}(h(m)) - q_{k,i}^{*}| > \epsilon \mid C^{m}[0] = k \right) = 0, \quad \forall \epsilon > 0. \tag{202}
\]
We have that,
\[
Q_{i}^{m}(h(m))|\{C^{m}[0] = k\} \overset{(a)}{\leq} U_{\lambda_{1}}^{m}(h(m)) + Q_{i}^{m}(0)\{C^{m}[0] = k\} \overset{(b)}{\leq} U_{\lambda_{1}}^{m}(h(m)) + \gamma U_{\lambda_{1}}^{m} \leq (\gamma + 1)U_{\lambda_{1}}^{m}, \tag{203}
\]
Step (a) follows from a decomposition similar to Eq. (196), by the writing the recallable rewards at time \(h(m)\) as those who arrived after \(t = 0\), which is dominated by \(U_{\lambda_{1}}^{m}(h(m))\), and those who were in the system at \(t = 0\), which is dominated by \(Q_{i}^{m}(0)\{C^{m}[0] = k\}\). Step (b) follows from Lemma 8. Since \(U_{\lambda_{1}}^{m}\) is a Poisson distribution with mean \(m\lambda_{1}\), it is not difficult show that there exists a random variable \(Y \in \mathbb{R}_{+}\), such that
\[
m^{-1}Q_{i}^{m}(h(m))|\{C^{m}[0] = k\} \overset{(a)}{\leq} \frac{1}{m}U_{\lambda_{1}}^{m}(\gamma + 1) \leq Y, \quad \forall m \in \mathbb{N}. \tag{204}
\]
Combining Eqs. (150) and (204), the dominated convergence theorem implies that, for all \(i, k \in K\),
\[
\lim_{m \to \infty} E\left(|m^{-1}Q_{i}^{m}(h(m)) - q_{k,i}^{*}| \mid C^{m}[0] = k\right) = 0. \tag{205}
\]
This shows Eq. (151), and thus completes the proof of Proposition 3. Q.E.D.

Appendix C: Comparison to Perfect Memory

We discuss in this appendix what could happen if there were no memory decay in our model, i.e., if \(\mu = 0\), and why it is different from the limiting regime considered in this paper, with \(\mu \to 0\).

When \(\mu = 0\), all recallable rewards remain in the system indefinitely. If we view the recallable rewards sampled at the update points as a discrete-time process, and in addition set the exploration parameter \(\alpha \to 0\), then our model becomes essentially the same as the choice process analyzed by Beggs (2005), where it is shown that the probability of choosing the best action converges to one as \(t \to \infty\) under Luce’s rule, as long as each action is associated with some strictly positive initial rewards. If \(\alpha\) is a positive constant, because there is no memory decay, the effect of \(\alpha\) disappears as soon as the rewards for all actions exceed \(\alpha\), and the same conclusion should hold.

Therefore, one would expect that when \(\mu = 0\) and \(\alpha \geq 0\), the choice probability under the reward-matching rule will concentrate on the best action as \(t \to \infty\), regardless of the update rate, \(\beta\). This is however different from the conclusion of Theorems 4, which shows the existence of two distinct limiting steady-state probabilities, one of which does not exhibit concentration on the best action. These observations thus suggest that our scaling regime do capture unique effects of imperfect memory. This is perhaps not too surprising in hindsight: if we had set \(\mu\) to zero, any positive update rate \(\beta\) would become, by definition, significantly greater than \(\mu\), and hence the second (memory-deficient) regime in Theorems 4 and 2 could not have appeared when \(\mu = 0\).