Multi-Armed Bandits with Endogenous Learning Curves: Applications to Split Liver Transplantation and Personalized Marketing

Yanhan (Savannah) Tang, Andrew Li, Alan Scheller-Wolf, Sridhar Tayur
Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA 15213, {yanhanta, aali1, awolf, stayur}@andrew.cmu.edu

Proficiency in many sophisticated tasks is attained through experience. For example, transplant surgeons need to practice difficult surgeries to master the skills required. Learning through experience is common and has broad implications in marketing. For example, consumers' knowledge of new products or services is often attained through exposure and user experience. Repeated exposure and increased knowledge of a product or advertisement can lead to increased liking, preference, or positive attitudes; this is also known as the *mere exposure effect*. In the above examples, one may leverage the inevitable phenomenon of learning through experience by exposing the transplant centers (TCs) to complex surgeries frequently and by running consistently personalized advertisements to nudge consumers to make purchases. However, learning through experience requires resources and can be costly; thus, it is crucial to identify inexperienced TCs with (initially unknown) high potentials quickly, and efficiently pick up the individual-level styles and patterns of advertisements that most likely lead to purchases. To rapidly identify high-potential TCs and effective styles of advertisements, we formulate a novel multi-armed bandit (MAB) model, in which parametric reward curves are embedded in the reward functions to capture endogenous, learning through experience (including customer learning) and the mere exposure effect. In addition, our model includes provisions ensuring that the choices of arms are subject to fairness constraints to guarantee equity in liver transplantation or to preserve variety in personalized marketing. To solve our MAB problem, we propose the L-UCB and FL-UCB algorithm families, variants of the upper confidence bound (UCB) algorithm, and we prove that they attain the optimal $O(\log t)$ regret. We demonstrate our model and algorithms on the split liver transplantation (SLT) allocation and the personalized marketing problems, showing that our algorithms have superior numerical performance compared to standard MAB algorithms settings where learning through experience/the mere exposure effect and fairness/variety-seeking concerns exist. From a methodological point of view, our proposed MAB model and the L-UCB and FL-UCB algorithm families are generic and have broad application prospects. From an application standpoint, our algorithms could be applied to help evaluate strategies to increase the proliferation of SLT. Marketers may apply our algorithms to harness the mere exposure effect in personalized marketing to drive increased sales and brand loyalty.

Key words: Multi-armed bandit, upper confidence bound algorithms, endogenous learning curves, non-stationary reward curves, split liver transplantation, personalized marketing, fairness, variety
History:

1. Introduction

Learning through experience is everywhere. For example, transplant centers (TCs) need to learn difficult medical procedures by performing them, staff in a call center need to handle customer calls to improve their ability to resolve customer issues efficiently and courteously, new franchisees learn to operate smoothly over time. Learning through experience is not limited to gaining surgical proficiency or streamlining business processes; it also has broad implications in marketing. For example, when introducing new products or services to customers, video tutorials in personalized styles are highly effective in elucidating how to use the product/service, addressing barriers to sales, and provoking interest in purchasing.

Consumers' knowledge of new products or services is often attained through exposure and user experience. Repeated exposure and increased knowledge of a stimulus (e.g., a product, service, brand, or advertisement) can lead to increased liking, preference, or positive attitudes; this is also known as the *mere exposure effect*. Potential customers watch short video advertisements and can get a glimpse into the experience. When the viewers encounter the product/service multiple times, they tend to be more familiar with it and are more likely to make purchase decisions. These video ads and tutorials are especially compelling when personalized for the individual customer. Therefore, marketers may design and produce multiple video tutorials of different styles and generate a multitude of video content catering to various tastes and preferences.

Marketers can leverage the mere exposure effect for customer acquisition and retention by running frequent advertisements and maintaining a consistent and strong online presence for targeted customers. However, these marketing campaigns can be costly as they involve multiple touchpoints and consistency. In many real-world scenarios, customers' true tastes and tendencies to purchase are unknown; therefore, marketers need to make decisions on the fly and learn about customers in real-time. To obtain desirable outcomes such as sales and customer loyalty, marketers must identify the most promising video ads for personalization; however, this is challenging because marketers can only observe (noisy) proxies/indicators based on which they estimate customers' actual likelihood of purchasing. Meanwhile, customers' tendencies to purchase are influenced by the personalized marketing strategy as they learn about the product/service, get familiarized, and grow interested in purchasing. Marketers need to balance the *exploration versus exploitation* and identify optimal options quickly: They need to decide between spending the budgeted marketing resources to explore more potential styles of video ads/tutorials and focusing on using the currently best-yielding ones for their personalized marketing campaigns. Further complicating the problem, supervisors may also seek to incorporate other customer-centered metrics into their decision-making, such as quality, fairness, variety, or market breadth. An effective marketing strategy involves two crucial steps: promptly identifying customers' preferred styles for personalized purposes and predominantly leveraging the most effective options to capitalize on the mere exposure effect (unless a certain degree of variety is desired).

The exploration versus exploitation tradeoff also arises in other applications characterized by initially unknown potentials and learning through experience. Learning through experience, while necessary and important in the long term, may come with a short-term cost and affect other stakeholders. For example, an inexperienced surgeon may only be able to perform certain surgeries in their learning phase, and may have a lower success rate. As a result, certain patients requiring a more intricate surgery may not be eligible for surgery with the TC, and expected outcomes may be worse even if a patient is eligible. At the same time inexperienced TCs learn by doing, the organ allocation organization learn about their aptitude/potential and may adjust strategies or allocate resources to achieve larger objectives. Importantly, the organ allocation authority needs to identify and nurture enough young, promising surgeons to learn sophisticated surgeries to treat patients nationwide into the future. Crucial to efficiently learning the potentials of these surgeons is to evaluate the results of learning through experience in its early stages.

We develop a methodology to solve these sorts of dynamic learning problems, focusing on one of the four scenarios described above for illustration purposes: allocating livers for expanded utilization of split liver transplantation (SLT) in the US, and provide numerical results for the personalized marketing example. We describe the SLT problem in greater detail below.

SLT is a procedure that potentially saves two lives using one donor liver, by splitting the donor liver into two partial livers and transplanting each of them into a size-appropriate recipient (Emre and Umman 2011). This is in contrast to traditional liver transplantation, also known as whole liver transplantation (WLT), which transplants one whole liver into a single recipient. SLT thus provides a unique opportunity to save more lives with the existing pool of livers—despite years of effort to increase organ donation in the United States, there remains a grievous shortage: As of June 2022, there were 11247 candidates on the liver waiting list; and the median waiting time on the list before receiving a liver was 1026.7 days. In the year 2021 alone, there were 13440 new additions to the waiting list, while only 9236 liver transplants were performed. However, despite the acute liver shortage, less than 2% of medically splittable livers are actually used for SLT in the US (Perito et al. 2019a,b).

Besides increasing the total number of transplants, SLT can potentially reduce treatment disparities based on size among patients with end-stage liver diseases (ESLD), whose only chance to survive is liver transplantation. Generally speaking, ESLD patients of smaller physical size face longer wait times and overall lower access to transplants, because there are fewer size-appropriate donated livers. (In WLT liver-recipient size matching, a recipient may receive an organ of the same size or a slightly smaller size.) In fact, in the US, SLT is currently used primarily to increase liver availability for pediatric patients; resulting outcomes of SLT for children are comparable to those of WLT (Hackl et al. 2018). Liver transplant allocation in the US is managed by a central planner—the Organ Procurement and Transplantation Network (OPTN). Their allocation rule can be essentially described as "sickest first": ESLD patients' health conditions evolve dynamically, and their position on the national waitlist, and consequent probability of getting a transplant, change over time (Emre and Umman 2011, Akan et al. 2012).

Surgeons' aptitudes for performing different types of transplant surgeries are affected by the coordination and expertise of their whole medical team, or even the entire transplant center. In fact, it is not uncommon for the expertise levels of surgeons and medical teams on complicated surgeries to be significantly influenced by the transplant centers (TCs) they belong to. TCs report transplant outcomes on the aggregate TC level, not on the individual surgeon level. This is because new surgeons are usually matched to TCs based on overall surgical skills, and usually, surgeries are performed by medical teams which involve supporting staff that may assist multiple surgeons in a TC. Moreover, new surgeons likely have the chance to observe, assist, and learn from experienced surgeons of their TC during actual SLT surgeries, and these hands-on experiences are crucial in achieving their full potential. Therefore, we focus on transplant center level proficiency: We divide transplant centers into several classes based on their features, such as surgical experience, performance history, geographical regions, etc. Henceforth, we use "surgeon," "transplant center," "center," and "medical teams" interchangeably.

Like many complicated and potentially risky medical procedures, gaining SLT expertise requires a learning process. This process is not only arduous but also may involve a lower initial transplant success rate, as the medical team acquires skills (Perito et al. 2019b). To encourage more transplant teams to learn SLT, and reduce surgical risks, policymakers might consider accommodating their learning, for example by allocating them high-quality organs. To evaluate the benefits of helping surgeons acquire the skills needed to perform SLTs while identifying the most suitable medical teams to specialize in SLT, we model organ allocation in a centralized transplantation network using a *multi-armed bandit* (MAB) model. We then develop novel variants of the *upper confidence bound* (UCB) algorithm to find allocation policies that balance aptitude exploitation and learning exploration.

Within our MAB model we explicitly incorporate the following features:

- **Endogenous learning curves**: Transplant centers' SLT expertise increases as they accumulate experience. In our MAB formulation, arms' rewards are parametric functions (with unknowm scalar or vector parameters) of previous arm choices, capturing increasing proficiency with practice.

- **Fairness**: A UNOS public comment proposal (OPTN and UNOS 2016) states that "...increased utilization of split liver transplantation could increase access to transplants," and "The Committee affirms that optimal allocation policies involving whole livers or split liver allografts should reflect a balance between the principles of equity and utility." We propose two fairness notions: best-K probabilistic fairness (BK-fairness) and arbitrary arm fairness (AA-fairness). These notions seek to expand the number of facilities equipped with SLT capabilities, and/or address equity concerns by imposing rules that diminish disparities in access to transplants.

The incorporation of these model features significantly complicates the MAB model; nevertheless, we propose L-UCB and FL-UCB frameworks, which solve a broad class of MAB problems where learning and fairness constraints exist. We prove that our L-UCB and FL-UCB algorithms achieve the optimal $O(\log t)$ regret, where t denotes the number of transplants, under benign conditions.

We note that our problem could also be captured using a reinforcement learning (RL) model. We choose not to formulate a general RL model because of our problem's special structure: the more SLTs performed by a medical team, the more experienced the medical team becomes. By exploiting this structure, we can use an enhanced MAB with learning curves embedded in its non-stationary rewards to fully characterize the structured RL problem, while maintaining parsimony and tractability.

Our methodology could potentially be applied to help evaluate strategies to increase the proliferation of SLT and other medically-difficult procedures, for example how to effectively and fairly develop a base of skilled practitioners. Moreover, our model and algorithms can be applied to any resource allocation problem where learning exists, including the call center and franchisee examples mentioned previously.

This paper is organized as follows: Section 2 discusses the literatures relevant to our work. Section 3 introduces the SLT learning problem and the MAB model formulation with learning curves embedded in the arm reward functions. Section 4 describes the L-UCB algorithm and analyzes its regret bound for the MAB models. Section 5 introduces our novel fairness notions, and describes our FL-UCB algorithm with its $O(\log t)$ regret bound. Section 6 discusses extensions to our MAB model and summarizes relevant findings. Section 7 presents the results of numerical experiments based on real-world SLT data. Section 8 summarizes the conclusions and contributions of this paper and discusses the limitations and potential directions for future work. An appendix, containing more details about the extensions, simulations, and all proofs, can be found in a supplemental file.

2. Literature Review

This work is closely related to seven streams of literature: a) exploration and exploitation trade-off; b) dynamic learning; c) organ transplantation; d) MAB with delayed feedback; e) experience-based learning; f) fairness; and g) the mere exposure effect and personalized marketing.

Exploration and exploitation trade-off. A classical model for the exploration-exploitation dilemma used in statistics, artificial intelligence (AI), and MS/OR is the multi-armed bandit (MAB), first introduced by Thompson (1933) for clinical trials. In this paper, we formulate a *stochastic bandit* with parameterized, endogenously non-stationary reward functions. Researchers have also studied contextual bandits, adversarial bandits, and linear bandits extensively (Lattimore and Szepesvári 2020).

In the vanilla stochastic MAB problem, arm rewards are stationary; however, to model endogenous experience-based learning and its resulting improved proficiency as experience accumulates, we embed a learning curve in each arm's reward function. Specifically, we consider parametric learning curve functions with unknown scalar or vector parameters. Nonstationary rewards, i.e., a reward distribution that can evolve over time, in MABs have primarily been studied when nonstationarity comes from the exogenous environment (Besbes et al. 2019, Cheung et al. 2020, Garivier and Moulines 2011), making arm rewards independent of policy history. Cheung et al. (2020) also studied endogenous reward nonstationarity using a discrete-time Markov decision process (MDP), where both the discrete reward and discrete state-transition distributions depend (solely) on the current state and action. We consider an infinite-horizon, continuous-time formulation where nonstationarity can be fully characterized by a parametric learning curve. Our formulation of parametric nonstationary rewards is significantly different from existing work, and has advantages in terms of parsimony and extending the upper confidence bound algorithm class.

Dynamic learning. Dynamic learning problems in an endogenously or exogenous changing environment have been studied in different contexts, e.g., online search and consumer lending. In endogenously changing environments, den Boer and Keskin (2022) studied a dynamic pricing problem where demand is influenced by the current selling price and also by customers' hidden reference prices that may endogenously evolve over time. The seller needs to learn customers' true reference price through price exploration and balance the tradeoff between demand learning and earning. In exogenously changing environments, Keskin and Zeevi (2017) studied a dynamic pricing problem where a seller faces an unknown demand model that can exogenously change subject to some finite variation "budget"; their variation metric allows for a broad spectrum of temporal behavior. Keskin and Li (2021) considered heterogeneous customers and exogenous Markovian market transitions and analyzed a firm's optimal pricing policy and its structural properties. In this paper, we study a MAB variant where the reward curves following parametric functions and the expected rewards of arms change endogenously as a function of historical pulls. This parsimoniously captures our motivating applications.

Our work is also relevant to recent work on feature-based rewards and high dimensionality in dynamic learning problems (see Section 6.2). Ban and Keskin (2021) proved bounds for expected regret in a personalized demand model with customers' characteristics encoded as a potentially high-dimensional feature vector, where a seller learns the relationship between customer features and product demand through sales observation. Keskin et al. (2022) considers an electric utility company serving retail electricity customers over a discrete time horizon, where the company observes customers' consumption, high-dimensional customer characteristics, and exogenous factors, and dynamically adjusts price at the customer level. They jointly optimized spectral clustering and feature-based pricing and show their proposed policy achieves near-optimal performance.

Organ transplantation. While much work has been done on kidney allocation (Zenios et al. 2003), fewer papers have addressed the allocation problem for livers (Akan et al. 2012, Bertsimas et al. 2020), and those have only studied whole liver allocation. Akan et al. (2012) analytically modeled the liver allocation problem as a fluid model with utilitarian objectives incorporating patients' dynamically-changing MELD/PELD scores. Their work did not consider medical learning, the practice of SLT, or any fairness concerns. Bertsimas et al. (2020) proposed a novel continuous distribution model that balances efficiency and fairness in liver allocation. None of these papers considered SLT or experience-based learning. Our paper concentrates on the selection of transplant centers, surgical techniques, and livers for specialized procedures in their initial phases of expanding uses, specifically, SLTs.

In the transplantation community, most SLT papers are retrospective reviews, in which transplant centers share their SLT experiences. Other topics covered include ethics (Vulchev et al. 2004), statistical analysis using open data (Perito et al. 2019a), and policy guidelines (OPTN and UNOS 2016). Recent studies show that the outcomes of SLT can be as good as WLT in big TCs, for example, the transplant center at the University of California, San Francisco (UCSF).

MAB with delayed feedback. In many healthcare applications, including organ transplantation, the outcomes may only be observed after some delay (Anderer et al. 2022, Kantidakis et al. 2020). For example, 90-day survival labels are only obtained 90 days after the surgery, and quality-adjusted life years (QALY) may not be fully observed until many years later, but some transplant objectives can be observed right away or within days after transplant, e.g., postoperative outcomes, including graft function/dysfunction/failure and cellular rejection. There is a stream of research specifically discussing using such early-on intermediate indicators or surrogate outcomes for medical decision-making. For example, Anderer et al. (2022) study algorithms that use surrogate and true outcomes to improve decision-making within a late-phase clinical trial. We adopt a similar approach to extend our base algorithm to accommodate a delay in observing true rewards: We consider using estimated (expected) rewards (based on demographic features and clinical metrics) available immediately after surgery as temporary/surrogate outcomes. When true rewards are observed, the estimates are replaced with the true outcomes. We show in Section 6.1 that we obtain the same $O(\log t)$ regret for our problems under mild assumptions.

Delayed feedback arises in multi-armed bandit applications beyond healthcare, such as searching over fast-charging policies for electrochemical batteries to maximize battery lifetime (Joulani et al. 2013, Grover et al. 2018). Joulani et al. (2013) found that the delayed feedback inflates the regret in an additive fashion in stochastic MAB problems, and developed modifications of UCB algorithms. Grover et al. (2018) considered a setting where partial feedback is available (analogous to our surrogate outcomes) and proposed an extension where an agent can control a batch of arms. Compared with these works, our contributions differ in the following ways: We show that in an expanded class of MAB problems where the expected rewards are endogenous and nonstationary, if we have temporary estimates of the delayed rewards that satisfy mild conditions, we obtain the same optimal regret upper bound scale: $O(\log t)$.

Experience-based learning. For transplant surgeons, many procedures involve the same repetitive tasks; thus it is appropriate to use learning curves with a focus on increases in success rate to represent learning and improvement in performance over time. Several functional forms have been used in the literature to capture human learning, such as S-curves (Sigmoid curves), diminishing-returns curves, and increasing-returns curves. We primarily use the Sigmoid functional form for our SLT numerical study, which nicely captures the features of learning complicated surgeries, such as a slow learning rate at the beginning and stable long-term performance (Pusic et al. 2015, Le Morvan and Stock 2005). Discussions on learning in other applications are deferred to EC.8.

Fairness. Concerns about the fairness of access to medical care and resources have existed for centuries. We study a fairness notion that is not based entirely on a meritocratic basis but on some protected features (e.g., patient physical size, age, geographical region, etc.). Similar to Schumann et al. (2019), we define our notion of fairness probablistically. However, instead of equal or proportional group probability, we use max-min group probability, a notion adapted from Rawls (2001), where the arms within the group (corresponding, for example, a specific group of patients) are guaranteed to be selected with no less than certain probabilities. To characterize the efficiency-fairness trade-off, we adapt the *price of fairness* definition from Bertsimas et al. (2011), that is, the ratio of total reward loss to the optimal total rewards.

The mere exposure effect and personalized marketing. Psychology and consumer behavior researchers have found that mere exposure to a brand or product can encourage consumers to hold a more favorable attitude toward the brand or product (Janiszewski 1993, Montoya et al. 2017). Personalization in marketing refers to the practice of tailoring marketing content and experiences to individual consumers based on their specific preferences. Successful personalized marketing creates customized and engaging interactions to enhance customer satisfaction and overall experience (Chandra et al. 2022). However, little attention has been paid to leveraging the mere exposure effect in personalized marketing on social media; we fill this gap in this paper. On the individual level, customers have their own preferences and tastes that might be unknown to marketers initially. Marketers need to find out what options work best for each customer and commit to the best ones to employ the advantages of exposure repetition. A solid marketing strategy comprises two essential steps: swiftly identifying customers' preferred styles for personalized purposes and predominantly utilizing the most effective options to leverage the mere exposure effect unless a certain level of variety is desired.

3. Problem Formulation and Model Setup

In this paper we focus on the SLT problem where medical teams need to learn SLT by actually performing SLT surgeries on patients. Meanwhile, a central planner learns which combinations of surgical teams, liver types, and recipient types have the highest long-term rewards (i.e., 1-year graft survival). We aim to develop a novel bandit algorithm to accelerate learning the highest full-potential combinations under stochastic (and potentially delayed) observations.

We explicitly model each type of surgery as an "arm" in the vocabulary of bandit problems: Each arm, or surgery, incorporates information about the features of the transplant center(s), the liver(s) to be transplanted, and the patient(s) associated with the surgery.

3.1. SLT Learning Problem Formulation

Consider a discrete-time horizon $\mathcal{T} := \{1, 2, \ldots, T\}$. We group transplant centers into classes. Let $\mathcal{D}$ denote the set of transplant center classes comprising centers (with no, little, or some prior SLT experience) yet who are willing to learn and practice SLT. Throughout the planning horizon, each transplant center of class d is capable of learning and performing SLTs, provided there is a medically appropriate patient and liver pair. We assume that there are significantly more patients than the number of transplant centers and livers. In other words, at least one or one pair of ESLD patients of each defined patient class is always available so that any transplant center can perform any surgery in each period. This assumption is reasonable because the liver waitlists are overloaded (patient arrival rates are greater than liver arrival rates) and a large transplant center typically consists of more than six surgeons and tens of supporting staff so at least one medical team is on duty at any time.

The set of liver types is $\mathcal{L}$, where a liver type is determined by its quality, the geographical location of the donor, and compatibility requirements; let $L = |\mathcal{L}|$. A fixed portion of all deceased-donor livers are eligible for splitting. To focus on the SLT problem, we consider only livers that are medically splittable, assuming non-splittable livers are assigned by another process. Moreover, we assume that all information on transplant centers' experiences prior to the planning horizon, which are publicly available (UNOS 2020b), are summarized in the shapes and structures (intercept, slopes, etc.) of their SLT learning curves. Patients who are medically compatible with a liver of type ℓ might have different health conditions, e.g., some might be critically sick while others are healthier; we denote the set of these patient classes as $\mathcal{P}^{\ell}$.

When a splittable liver is split, the two partial grafts can be allocated to two recipients in different transplant centers at different times. At each time stamp $t \in \mathcal{T}$ there is exactly one (partial) liver arrival: ℓ_t, which can be transplanted into one recipient. Let P_t be a potential recipient, i.e. $P_t \in \mathcal{P}$. The action space at time t is to choose an allocation, defined as an eligible center-recipient(s)-type pair $(d, P_t) \in \mathcal{D} \times \mathcal{P}$.

We define the arm set of the MAB problem $\mathcal{A}_t$. For presentation clarity, we focus on the case where livers are homogeneous (i.e., $\mathcal{L} = \{\ell\}$, and $\mathcal{A}_t = \mathcal{A} := \mathcal{D} \times \mathcal{P}$); the heterogeneous-liver case is a direct extension and is discussed in Sections EC.5.2. Henceforth, we use the term "arm" and "surgery" interchangeably. Each arm is associated with a known accumulated experience level $s_{a,t}, a \in \mathcal{A}$, sometimes written as $s_a(t)$, for $t \in \{1, \ldots, T\}$, and with an unknown aptitude/full potential $\alpha_a \in \mathcal{U}$, i.e., the highest possible mastery level. The experience level indicates the efforts and experience of the surgeon or medical team, corresponding to a value on the x-axis of the learning curves. Depending on $s_{a,t}$, and the learning curve structure, we obtain a $\theta_a(s)$ value, denoting the current mastery or proficiency level of the arm.

Figure 1 illustrates three learning curves. The variable s on the x-axis represents learning efforts, or the number of attempts, while θ on the y-axis represents the proficiency or mastery of a specific arm. A higher proficiency level is associated with a higher survival probability or a better expected outcome of a surgery. All curves are Sigmoid curves (the "S"-curves), i.e., $\theta_i(s) := \frac{\alpha_i}{1+\exp(-s+\omega_i)}$, $i = 1, 2, 3$, with $\alpha_1 = 0.5$, $\omega_1 = 1$ (blue), and $\alpha_2 = 0.85$, $\omega_2 = -6$ (orange), and $\alpha_3 = 0.8$, $\omega_2 = -1$ (green). We assume that we know the structures and form of the arms' learning curves (e.g., a Sigmoid curve parameterizing a Bernoulli variable, which represent the values of surgical outcomes). Still, we do not know the parameter α of the curve. This assumption can be relaxed; please refer to Section 4.6 for more detail about learning multiple unknown parameters. Specifically, $\omega_i, i = 1, 2, 3$, can be known parameters (describing the existing experience) or unknown in which case they need to be learned along with $\alpha_i, i = 1, 2, 3$ (see Section 4.6).

Let $T_{a,t-1}$, sometimes written as $T_a(t-1)$, denote the number of times that arm a (or a's corresponding surgery) has been chosen (practiced) up to and including time $t-1$ (this may be different from $s_{a,t-1}$ in models with arm correlation, see Section EC.5). We define the state of the SLT learning problem as (ℓ_t, S_t) where $\ell_t \in \mathcal{L}$ and $S_t := (T_{a,t-1}, s_{a,t-1})_{a \in \mathcal{A}} \in (\mathbb{N}_+ \times \mathbb{R}_+)^{|A|}$. Let σ_t be the decision rule at time t, i.e. $a_t = \sigma_t(S_t)$. A policy π ($\pi(t)$) is a series of decision rules, i.e. $\pi = \{\sigma_\tau\}_{\tau=1}^{T}$ ($\pi(t) = \{\sigma_\tau\}_{\tau=1}^{t}$). Given $a_t \in \mathcal{A}_{\ell_t}$, we obtain a random reward $r(\ell_t, a_t, S_t)$, e.g., 1-year graft survival. When $\mathcal{L} = \{\ell\}$, for simplicity, we denote $r_{a,s_{a,t}} := r(\ell, a_t, S_t)$. We assume that the reward is a discrete Bernoulli variable, with the mean being the hidden expertise or mastery level

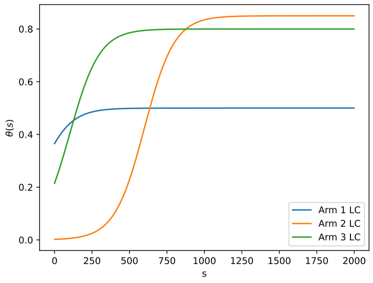


Figure 1 **An example with three learning curves. All are Sigmoid functions with different full potentials, shape parameters, and starting experience levels.**

of the participating medical team(s) for a certain type of surgery, i.e. $\theta_a(s_{a,t-1})$, where $a \in \mathcal{A}_{\ell_t}$ is the surgery type/arm and $s_{a,t-1}$ is a's experience level prior t.

The objective of the SLT learning problem is to find the policy which maximizes the objective function, e.g., the 1-year graft survival, for large T:

$$\max_{\pi} \mathbb{E} \sum_{t=1}^{T} r(\ell_t, a_t^{\pi}, S_t^{\pi}). \tag{1}$$

3.2. The Multi-Armed Bandit Model

Here, we summarize important notation and explain how we map the SLT liver allocation problem to a MAB model with endogenously nonstationary reward curves. At each time t, we choose an arm $a_t \in \mathcal{A}_\ell$ (i.e., allocate a liver for an SLT surgery) and receive a random reward $r(\ell, a_t, S_t)$, that is, the outcome of the SLT surgery. A strategy π is a series of allocation actions or choices of arms; $\pi(t)$ denotes the series of actions from time 1 up to t. We call $\pi(t)$ the policy history at t. Let $\theta_a^{\pi(t-1)}$ be an SLT arm a's unknown SLT performance level at time t, under a policy history $\pi(t-1)$, and $T_{a,t}$ or $T_a(t)$ denotes the number of times that arm $a \in \mathcal{A}_\ell$ was chosen prior to and including time t. Recall that $s_{a,t-1}$ or $s_a(t-1)$ denotes the experience level of arm a prior to time t. As we assume that arms are independent, $s_{a,t-1} = T_{a,t-1}$, and the hidden performance level of an SLT arm can be rewritten as $\theta_a(T_{a,t-1})$. The outcome of action a is a Bernoulli random variable with mean $\theta_a(T_{a,t-1})$.

3.3. Regret

We define the offline policy/the optimal full-information policy π_t^* which achieves the highest cumulative rewards, i.e., $\pi_t^* := \arg\max_\pi \sum_{\tau=1}^{t} r(\ell_\tau, a_\tau^\pi, S_\tau^\pi)$. To evaluate the utility loss due to lack of information on TC aptitudes where learning curves exist, we use a common objective in the bandit literature — minimizing total expected regret, that is, the expected deficit suffered relative to the optimal full-information policy. For any fixed turn $t \in \mathcal{T}$, the regret is defined as

$$R_t = \sum_{\tau=1}^{t} r(\ell_\tau, a_\tau^{\pi_t^*}, S_\tau^{\pi_t^*}) - \sum_{\tau=1}^{t} r(\ell_\tau, a_\tau^{\pi}, S_\tau^{\pi}). \tag{2}$$

When arms are independent of each other and arms' aptitude parameters $\alpha_a, \forall a \in \mathcal{A}$ are known, the optimal policy as t grows large is trivial, always selecting the arm with the highest long-term aptitude; in other words, it always chooses $a^* := \arg\max_{a\in\mathcal{A}} \alpha_a$. For any given t, full-information dynamic programming can solve the offline policy, which may not be trivial for small t.

4. L-UCB Algorithm and Regret Bounds

In this section we study the MAB problem with learning curves embedded in the reward functions, as shown in Figure 2. Each transplant center has an unknown true aptitude that determines its hidden (unobservable) expertise or proficiency level $\theta_a(s_{a,t-1})$ when a center's experience level is $s_{a,t-1}$. The observable variables are its experience level $s_{a,t-1}$ at time t, and past and current outcome variables $r^t := r(\ell, a_t, S_t)$ (e.g., 1-year graft survival). Note that the outcomes are also affected by environmental variables $e_t \in \mathcal{P} \times \mathcal{L}$ (patient health condition, liver quality, etc.); these environmental variables are taken into account in our formulation of MAB.

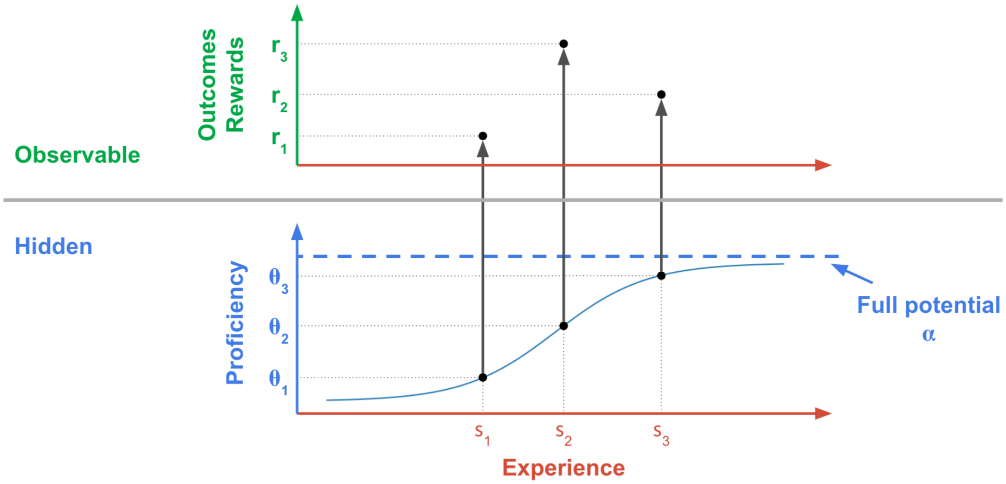


Figure 2 **A graphical representation of the SLT learning MAB problem. The observable outcome of surgery, r, is a random function of the hidden proficiency level θ. For a specific arm, when its experience with a certain type of SLT surgery is s and s', its hidden proficiency levels would be θ and θ', while the observable (stochastic) outcomes are r and r', respectively.**

In Section 4.1, we introduce the classical, vanilla UCB algorithm; in Section 4.2, we present the L-UCB algorithm; in Section 4.3, we prove L-UCB's regret bounds; in Section 4.4, we provide a generic method of moments (MoM) approach to construct unbiased L-UCB estimators; in

Section 4.5, we discuss the use of biased estimators in L-UCB; in Section 4.6, we present a general approach using maximum likelihood estimation (MLE) and maximum a posteriori probability (MAP) to construct estimators for α and/or multiple unknown parameters; finally, in Section 4.7, we study the scenarios where the parametric form of learning curves may be unknown, and propose nonparametric algorithms within the L-UCB framework.

4.1. Upper Confidence Bound (UCB) Algorithms

The *upper confidence bound* (UCB) method is a class of algorithms for the MAB that give an asymptotically optimal solution achieving an $O(\log t)$ regret (Lattimore and Szepesvári 2020). To illustrate this method, we start with a simplified scenario where $\theta_a(s) = \alpha_a, \forall s, a$; i.e., there is no learning present.

The standard, or vanilla UCB algorithm uses Hoeffding's inequality to derive upper confidence bounds on the unknown aptitudes; these bounds are greater than their *de facto* values with high probabilities. It then selects the arm with the maximal upper bound. For any surgery a with unknown aptitude α_a that has been chosen n times and yielded random rewards $r_a^{(1)}, \ldots, r_a^{(n)}$, the vanilla UCB uses $\hat{\alpha}_{a,n} := \frac{1}{n}\sum_{i=1}^n r_a^{(i)}$ as the estimator of α_a, the empirical or sample mean. Recall that $T_a(t-1)$ denotes the number of times surgery a has been chosen prior to time t. Define the upper bound for the estimate of α_a as

$$B_{a,t,T_a(t-1)} := \hat{\alpha}_{a,T_a(t-1)} + \delta_{a,t,T_a(t-1)}, \quad \text{where} \quad \delta_{a,t,n} := \sqrt{\frac{2\log\eta(t)}{n}}.$$

We choose $\eta(t) = t$ in the vanilla UCB; then the algorithm is formally defined as

$$a_t = \begin{cases} \arg\max_a B_{a,t,T_a(t-1)} & \text{if } t > |\mathcal{A}| \\ t & \text{if } t \leq |\mathcal{A}| \end{cases} \tag{3}$$

4.2. The L-UCB Algorithm

Now we describe the L-UCB algorithm for MAB problems with learning curves embedded in the reward functions. Similar to the notation used in the vanilla UCB, we denote by $\hat{\alpha}_{a,n}$ the estimator of α_a after arm a has been chosen n times. But instead of being restricted to the empirical mean, the estimator $\hat{\alpha}_{a,n}$ can be any function of n random rewards $(r_a^{(\tau)})_{\tau=1}^n$ and the corresponding n experience levels $(s_{a,\tau})_{\tau=1}^n$ to the estimate of the value of α_a, where $r_a^{(\tau)}$ denotes the random reward obtained the τth time arm a was chosen, when the experience level was $s_{a,\tau}$. Thus in the L-UCB algorithm $\hat{\alpha}_{a,n}$ can utilize a broad class of mapping functions and estimators, including the empirical mean. Some other potential estimators are method of moments (MoM) estimators, maximum likelihood estimation (MLE) estimators, and maximum a posteriori probability (MAP) estimators, which we discuss more in Section 4.4 $\sim$ 4.6. The estimator $\hat{\alpha}_{a,n}$ in L-UCB takes n additional arguments—the experience levels—compared to the estimator used in the vanilla UCB.

In this section, the experience level $s_{a,\tau} = \tau - 1$, for all $\tau = 1, 2, \ldots, n$, i.e., the experience level is equivalent to the number of historical pulls. In Section 6.2 we discuss an extension in which arms are correlated; in such scenarios, $s_{a,\tau-1}$ and $T_a(\tau - 1)$ may not be equivalent. Note that the estimator $\hat{\alpha}_{a,n}$ can incorporate the parametric forms of learning curves, if such information is available. Alternatively, $\hat{\alpha}_{a,n}$ can be chosen without any knowledge of the learning curves' parametric form; we discuss nonparametric methods in Section 4.7.

Because $\hat{\alpha}$ might be a more complicated function than the empirical mean, we define the following properties over this function class.

Definition 1 (Bias of an estimator)**.** *The bias of an estimator $\hat{\alpha}$ of parameter α is the difference between the expected value of the estimator and the true value of α; that is, $\mathbb{E}[\hat{\alpha}] - \alpha$.*

Definition 2 (Unbiased estimator)**.** *Estimator $\hat{\alpha}$ is unbiased if its bias is zero.*

It should be noted that many widely used estimators are biased; for example, the MLE estimator of the Gaussian variance is biased.

Definition 3 (Gap of a sub-optimal arm a)**.** *The gap of arm a is $\Delta_a := \max_{a' \in \mathcal{A}} \alpha_{a'} - \alpha_a$.*

Remark: In our SLT learning problem, because all α_a's take values from a bounded set $\mathcal{U} := [0, 1]$, the sub-optimal gaps are also bounded throughout the planning horizon.

Definition 4 (Per-coordinate difference bound)**.** *Suppose $\mathcal{X}$ is a sample space and $\varphi : \mathcal{X}^n \to \mathbb{R}$. If there exists $w_1, \ldots, w_n \geq 0$ such that*

$$\sup_{x_1,\ldots,x_n,x_i^{'} \in \mathcal{X}} |\varphi(x_1, \ldots, x_{i-1}, x_i, x_{i+1}, \ldots, x_n) - \varphi(x_1, \ldots, x_{i-1}, x_i^{'}, x_{i+1}, \ldots, x_n)| \leq w_i \tag{4}$$

for $i \in \{1, 2, \ldots, n\}$, then $(w_i)_{i=1}^n$ is said to be a per-coordinate difference bound for φ.

Equation (4) states that any modification of the value of the ith coordinate changes the value of φ by at most w_i whatever values the other coordinates take. This φ can be any function including any aforementioned estimator $\hat{\alpha}_{a,n}$. For any φ, the per-coordinate difference bound doesn't have an upper bound (as $w_i = \infty$ satisfies (4)) but does have an infimum which varies with φ. When φ is independent of the ith coordinate, that is, changing the value of the ith coordinate solely never changes the value of φ, the infimum of w_i is zero. In this case, the ith coordinate is obsolete.

Definition 5 (Per-coordinate difference bound parameter)**.** *If mapping function $\varphi : \mathcal{X}^n \to \mathbb{R}$ has per-coordinate difference bound $w_1, \ldots, w_n$, then we say φ has a per-coordinate difference bound with parameter $C_n^w := \frac{1}{n \sum_{i=1}^n w_i^2}$.*

Remark: For any function φ, C_n^w is not unique and doesn't have a positive lower bound, i.e., C_n^w can be arbitrarily small, because $w_1 = w_2 = \cdots = w_n = +\infty$ is a per-coordinate difference bound of φ with parameter $C_n^w = 0$. However, as w_i has an infimum, C_n^w does have a supremum, which depends on the nature of φ.

Now we present pseudo-code of the L-UCB algorithm. Assume $\hat{\alpha}_{a,n}$, the estimator of α after arm a has been chosen n times, has a per-coordinate upper bound with parameter $C^w_{a,n}(>0)$. Similar to the vanilla UCB, we define

$$\delta_{a,t,n} := \sqrt{\frac{2\log t}{nC^w_{a,n}}} \quad \text{and} \quad B_{a,t,T_a(t-1)} := \hat{\alpha}_{a,T_a(t-1)} + \delta_{a,t,T_a(t-1)}.$$

For simplicity, $B_{a,T_a(t-1)} := B_{a,t,T_a(t-1)}$. Denote by $b_{a,n}$ the bias of $\hat{\alpha}_{a,n}$, and assume there exists an $m_a \in \mathcal{T}$ for arm a such that $|b_{a,n}| \leq \frac{1}{10}\sqrt{\frac{2\log n}{nC^w_{a,n}}}$ for all $n \geq m_a$ (we discuss m_a below).

Algorithm 1: L-UCB Algorithm Pseudo Code

```
1: Initialization: Select each arm $a$ $m_a$ times
2: Update statistic: $B_{a,T_a(t-1)} \leftarrow \hat{\alpha}_{a,T_a(t-1)} + \sqrt{\frac{2\log t}{C^w_{a,T_a(t-1)} T_a(t-1)}}$,   $\forall a \in \mathcal{A}$
3: Select arm: $a_t \leftarrow \arg\max_a B_{a,t,T_a(t-1)}$, and update $T_{a_t,t}$
4: Increment $t$ and Go to Step 2
```

Now, we discuss the behaviors of m_a when the bias $|b_{a,n}|$ shrinks at different rates and when $C^w_{a,n}$ has a positive lower bound, i.e. $C^w_a := \inf_{n\in\mathbb{N}_+} C^w_{a,n} > 0$. When estimator $\hat{\alpha}_{a,n}$ is unbiased for all n, we select arm a exactly once in the initialization, just as the vanilla UCB. When the bias $|b_{a,n}|$ decays at $O\left(\sqrt{\frac{1}{n}}\right)$ rate, i.e. there exists a constant K^b_a such that $|b_{a,n}| \leq K^b_a n^{-1/2}$ for any n, we can set $m_a = \lceil \exp(50(K^b_a)^2 C^w_a) \rceil$. If the bias doesn't decay we need to choose an alternative estimator with zero or decaying bias, unless the bias is known and can be corrected.

4.3. L-UCB Regret Bounds

In this subsection we derive the upper bound on the regret of the L-UCB algorithm.

Proposition 1 (Upper and lower bounds of the supremum of the per-coordinate difference bound)**.** *Suppose $\mathcal{X}$ is a sample space, $\varphi : \mathcal{X}^n \to \mathbb{R}$ is a function whose image set is $[0,1]$, and $\{\omega_i\}_{i=1}^n$ is any per-coordinate bound defined over $\mathcal{X}$. Then the supremum of the per-coordinate difference bound of φ over all possible values of $\omega_1, \ldots, \omega_n$, denoted by $C^*_n := \sup_{\omega_1,\ldots,\omega_n} C^w_n$, satisfies $\frac{1}{n^2} \leq C^*_n \leq 1$.*

To prove the left inequality we note that $w_1 = 1, \ldots, w_n = 1$ is a per-coordinate difference bound of φ with parameter $C^w_n = \frac{1}{n^2}$. Because C^*_n is the supremum of all feasible C^w_n, we know $C^* \geq \frac{1}{n^2}$. The proof of the right inequality uses Chebyshev's sum inequality (Hardy et al. 1952). Please refer to Section EC.1.2 for proof details.

Lemma 1 (Bounded Difference Inequality)**.** *Suppose $\mathcal{X}$ is a sample space, and function $\varphi : \mathcal{X}^n \mapsto \mathbb{R}$ has per-coordinate difference bound $w_1, \ldots, w_n$ with parameter C^w_n, i.e. $w_1, \ldots, w_n > 0$ satisfy $C^w_n = \frac{1}{n\sum_{i=1}^n w_i^2}$ and*

$$\sup_{x_1,\ldots,x_n,x_i^{'}} |\varphi(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n) - \varphi(x_1,\ldots,x_{i-1},x_i^{'},x_{i+1},\ldots,x_n)| \leq w_i \tag{5}$$

for $i \in \{1, \ldots, n\}$*. Then,*

$$P\left(\varphi - \mathbb{E}[\varphi] > \varepsilon\right) \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n w_i^2}\right) = \exp\left(-2nC_n^w \varepsilon^2\right), \tag{6}$$

$$P\left(\varphi - \mathbb{E}[\varphi] < -\varepsilon\right) \leq \exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n w_i^2}\right) = \exp\left(-2nC_n^w \varepsilon^2\right). \tag{7}$$

For proof details, readers are referred to McDiarmid (1998) (see their Theorems 1 and 2 and references therein). Lemma 1 states that the probability of the value of φ being close to its expectation is higher when φ is less sensitive to its arguments, i.e., the upper bounds of the above probabilities are smaller if the w_i's are smaller and C_n^w is larger.

When φ is the empirical mean used in the vanilla UCB, we have $w_a^{(i)} = \frac{1}{n}$ for any i. Because the empirical mean achieves the maximum $C_n^w = 1$, we take it as the standard and compare other functions' behaviors with it. In this sense, nC_n^w can be thought of as a reduced number of samples: For the empirical mean, nC_n^w is precisely n, which is the number of samples governing the rate of decay of the bound. When the estimator φ has larger $w_a^{(i)}$s, we have C_n^w less than 1, and then the probabilistic bounds on $\varphi - \mathbb{E}[\varphi]$ are as tight as the corresponding bounds of the empirical mean with $nC_n^w < n$ samples, i.e., the estimator with C_n^w achieves the same accuracy with fewer samples compared to the empirical mean estimator, as $C_n^w < 1$.

For example, suppose $r_t \sim \text{Bernoulli}(\frac{\alpha t}{t+1})$ for $t \in \{1, \ldots, T\}$, the higher the aptitude α and/or the proficiency level $s_t := t$, the more likely that a surgery is successful. The value of α is hidden, but t is known in each round. The estimator $\hat{\alpha} := \frac{1}{T}\sum_{t=1}^T \frac{t+1}{t} r_t$ is an unbiased estimator of α. This $\hat{\alpha}$ can be thought of as a weighted empirical mean (although the weights don't sum to one), so, similar to the empirical mean, this $\hat{\alpha}$ has per-coordinate difference bound $w_1 = \frac{1}{T}, \ldots, w_t = \frac{t+1}{tT}, \ldots, w_T = \frac{T+1}{T^2}$ with parameter $C_T^w = \frac{T}{T + 2\sum_{t=1}^T t^{-1} + \sum_{t=1}^T t^{-2}}$; this C_T^w is less than 1 at any finite T and approaches 1 as T approaches infinity.

Theorem 1. *Denote the reward of choosing arm* a *for the* n*th time by* $r_a^{(n)}$*. Suppose* $r_a^{(1)}, r_a^{(2)}, \ldots$ *are independent of each other conditioned on the latent aptitude* α_a*. For each* $n \in \mathcal{T}$*, suppose estimator* $\hat{\alpha}_{a,n}$ *has a per-coordinate difference bound* $w_{a,n}^{(1)}, \ldots, w_{a,n}^{(n)}$ *with parameter* $C_{a,n}^w$*, i.e.* $w_{a,n}^{(1)}, \ldots, w_{a,n}^{(n)} \in \mathbb{R}_+$ *satisfy* $C_{a,n}^w := \frac{1}{n\sum_{i=1}^n (w_{a,n}^{(i)})^2}$ *and*

$$\sup_{r_a^{(1)}, \ldots, r_a^{(n)}, r'} |\varphi(r_a^{(1)}, r_a^{(2)}, \ldots, r_a^{(i-1)}, r_a^{(i)}, r_a^{(i+1)}, \ldots, r_a^{(n)}) \\ - \varphi(r_a^{(1)}, r_a^{(2)}, \ldots, r_a^{(i-1)}, r', r_a^{(i+1)}, \ldots, r_a^{(n)})| \leq w_{a,n}^{(i)}, \quad \forall a \in \mathcal{A}, i \in \{1, \ldots, n\} \tag{8}$$

Let $t \in \mathcal{T}$ *be any timestamp. When* $C_{a,n}^w$ *has a positive lower bound, i.e.* $C_a^w := \inf_{n \in \mathbb{N}_+} C_{a,n}^w > 0$*, and when the bias of* $\hat{\alpha}_{a,n}$ *satisfies* $|b_{a,n}| \leq \frac{1}{10}\sqrt{\frac{2\log n}{nC_a^w}}$*, each sub-optimal arm is pulled in expectation at most*

$$\mathbb{E}[T_a(t)] \leq \frac{8 \log t}{C_a^w \Delta_a^2} + 2\zeta(1.24) \tag{9}$$

times, where $\zeta(1.24) \approx 4.76$*, and* $\zeta(s)$ *is the Riemann zeta function, i.e.* $\zeta(s) = \sum_{i=1}^{\infty} i^{-s}$*.*

The expected cumulative regret of the L-UCB algorithm is bounded by

$$\mathbb{E}[R(t)] \leq \sum_{a \neq a^*} (\bar{r}_{a^*} - \underline{r}_a) \left(\frac{8 \log t}{C_a^{\omega} \Delta_a^2} + 2\zeta(1.24) \right). \tag{10}$$

Our proof adapts some techniques from the proof of bounds for vanilla UCB algorithms, but our results are applicable to a more general class of bandits. The primary differences/improvements of our result are: a) our regret bounds apply in a broader class of UCB algorithms that use any estimators φ that satisfy certain benign criteria in the L-UCB algorithms; b) our proof allows these estimators to be biased, up to $\frac{1}{10}\sqrt{\frac{\log n C_{a,n}^w}{n}}$, where n is the sample size and is adequately large. Taken together, these innovations significantly expand the scope of MAB regret bounds, including those with embedded learning curves and a broad class of estimators; examples and discussions in Section 4.4 illustrate these benefits.

4.4. A Generic Method of Moment (MoM) Estimator: An Explicit Formula

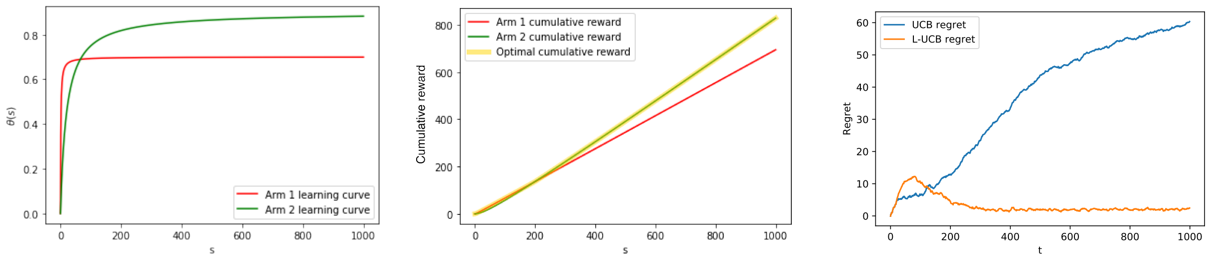
For learning curves that satisfy the following $\theta(s) = \alpha g_\omega(s) + f(s)$, e.g., the learning curves in Example 1 and Section 7, a generic Method of Moments (MoM) estimator can be $\hat{\alpha}_n^{MoM} = \frac{1}{n}\sum_{s=1, g_\omega(s)\neq 0}^{n} \frac{r^{(s)} - f(s)}{g_\omega(s)}$. (We assume not all $g_\omega(s) = 0$; if so, $\alpha_n^{MoM} = 0$.) Note that α^{MoM} is unbiased, because $\mathbb{E}\alpha_n^{MoM} = \mathbb{E}\frac{1}{n}\sum_{s=1}^{n} \frac{r^{(s)} - f(s)}{g_\omega(s)} = \frac{1}{n}\sum_{s=1}^{n} \frac{\mathbb{E}r^{(s)} - f(s)}{g_\omega(s)} = \frac{1}{n}\sum_{s=1}^{n} \frac{\alpha g_\omega(s)}{g_\omega(s)} = \alpha$, assuming $g_\omega(s) \neq 0$. (If for certain s', $g_\omega(s') = 0$, we drop s' when taking the average.) $C_n^{w,MoM} = \frac{1}{n\sum_{s=1}^{n}(g_\omega(s))^2}$. α^{MoM}, as $r^{(s)} \in \{0,1\}$. In cases where ω is unknown (see Example 3 in Section 4.6), we can use estimates $\hat{\omega}_s$ to replace ω, i.e. $C_n^{w,MoM} = \frac{1}{n\sum_{s=1}^{n}\left(1/g_{\hat{\omega}_s}(s)\right)^2}$.

Example 1 illustrates constructing an MoM estimator and applying L-UCB.

Example 1 (Incorporating information about the learning curve)**.** *Consider a bandit with two independent arms, whose reward curves are illustrated in Figure 3a. Arm 1 has a learning curve* $\theta_1(\alpha_1, s) = \alpha_1 \frac{s}{s+1}$ *where* α_1*'s unknown true value is* 0.7*, while arm 2's learning curve is* $\theta_2(\alpha_2, s) = \alpha_2 \frac{s}{s+20}$ *while* α_2*'s unknown true value is* 0.9*. Suppose the random outcome* $r_{a,s}^{(i)}$ *is a Bernoulli variable with parameter* $\theta_a(\alpha_a, s)$*. We use the unbiased MLE estimators* $\hat{\alpha}_{1,n} = \frac{1}{n}\sum_{i=1}^{n} \frac{i+1}{i} r_{1,i}^{(i)}$ *and* $\hat{\alpha}_{2,n} = \frac{1}{n}\sum_{i=1}^{n} \frac{i+20}{i} r_{2,i}^{(i)}$ *in the L-UCB algorithm. The estimator for the vanilla UCB is* $\hat{\alpha}_{1,n} = \frac{1}{n}\sum_{i=1}^{n} r_{1,i}^{(i)}$*.*

In Figure 3c, it is clear that the L-UCB algorithm has significantly lower numerical regret when t is large enough (i.e., $t > 194$) because L-UCB incorporates the learning curve information in the early stages to identify the "best" arm in the long-term more efficiently.

It might be counterintuitive that the L-UCB's regret curve increases before decreasing; however, this is possible when the offline policy (i.e., the optimal policy solved by a clairvoyant who knows all parameters) is nonstationary in t. In Example 1, the "best" arm to play is dependent on the time horizon: If we only consider $t \leq 194$, the "optimal" strategy is always pulling arm 1, but in the



(a) Learning curves (b) Expected cum. rewards (c) Regret comparison

Figure 3 Illustrations for Example 1. The regret results shown are averaged over 20 instances.

long term ($t \geq 195$), the “optimal” policy is always pulling arm 2. The optimal offline policy may not be trivial for any given t, and in general, can be tricky to solve and sometimes intractable. The non-asymptotic regime is important for general MAB problems (Garivier et al. 2019). Nevertheless, in Example 1 and the SLT problem, we focus on identifying the best long-term arm, which is a special scenario in dynamic learning in which constantly pulling the arm with the highest full potential (α) is the optimal policy asymptotically.

For smaller t the vanilla UCB does the “correct” thing by playing the (temporally) more valuable arm 1. The “optimal” cumulative reward for each t in Figure 3b is the optimal cumulative reward obtained by a clairvoyant who knows the true parameters (including α_1, α_2) of all arms’ learning curves. Because our L-UCB policy is designed to identify the best long-term arm (i.e., arm 2), in the short term ($t \leq 194$) where the short-term “optimal” policy is pulling arm 1, L-UCB may incur more temporal regret, as seen in Figure 3c.

It is possible in applications where learning is present that the vanilla UCB, without information on learning curves, may still result in $O(\log t)$ regret. However, this is typically not the case in general dynamic learning problems. Performances of the UCB and L-UCB are more extensively compared numerically in Sections 4.6 $\sim$ 4.7 and in Section 7.

4.5. Biased Estimators

We further illustrate the benefit of Theorem 1 by allowing the estimators to be biased.

Example 2 (Estimator bias)**.** *Consider a bandit problem where rewards* $(r_a^{(i)})_{i=1}^n$ *are i.i.d. Bernoulli random variables with parameter* p*. To reduce the estimator’s variance, people often use a MAP estimator with a Beta prior. Mathematically, the MAP estimator with prior* $\mathrm{Beta}(\alpha,\beta)$ *is* $\hat{\alpha}_n := \frac{h+\alpha-1}{n+\alpha+\beta-2}$*, where* h *is the number of ones in rewards. With* n *samples, the bias of this MAP estimator is* $\frac{(1-p)(\alpha-1)-p(\beta-1)}{n+\alpha+\beta-2}$*, which is nonzero for most combinations of* α*,* β*, and* p*. As we obtain more samples, i.e., as* n *increases, the bias decays at* $O(\frac{1}{n})$ *rate. Hence, Theorem 1 is applicable to this case and guarantees an* $O(\log t)$ *regret. In contrast, the vanilla UCB cannot be guaranteed to work with this estimator.*

This flexibility with respect to estimators yields one further advantage of the L-UCB algorithm: The vanilla UCB cannot guarantee identifying the "best" arm when the empirical mean is not an appropriate estimator for the metric of interest, for example, the variance or standard deviation of a random variable. In contrast, in the L-UCB algorithm, we can use any estimator, and if the premises of Theorem 1 are met, we immediately have that the regret is bounded by $O(\log t)$, thus providing much greater freedom in the choice of metric.

In cases where the bias is initially large but decays quickly, i.e., $|b_{a,n}| \leq C_a^b \sqrt{\frac{1}{n}}$ for some large constant $C_a^b \in \mathbb{R}_+$, the bounds on $T_a(t)$ and $R(t)$ in the theorem may not hold for small t, because the bias condition $|b_{a,n}| \leq \frac{1}{10}\sqrt{\frac{2\log n}{nC_a^w}}$ may not hold for $n = t$ in this case. Nevertheless, these bounds hold for any adequately large t. Because Theorem 1 is intended to show the scale of how many times we choose sub-optimal arms and the scale of the regret, we omit discussions of issues around these cases, such as the minimum t where the bounds hold for a given C_a^b.

In Theorem 1's premise, we assumed $|b_{a,n}| \leq \frac{1}{10}\sqrt{\frac{2\log n}{nC_a^w}}$. The biases of many standard estimators, e.g., the MLE estimator of logistic regression, is $O(\frac{1}{n})$. Thus, the premise of Theorem 1 holds for most common estimators (Lehmann and Casella 2006). Note that in rare cases, verifying the bias conditions for some algorithm instances within our proposed L-UCB framework may be nontrivial; one may skip verifying the bias conditions, apply the L-UCB algorithms and see if they have logarithmic regrets empirically. See Section 4.6 for a detailed discussion about ways to verify the bias conditions for L-UCB with MLE and MAP.

4.6. MLE and MAP for Estimating Unknown Vector Parameters

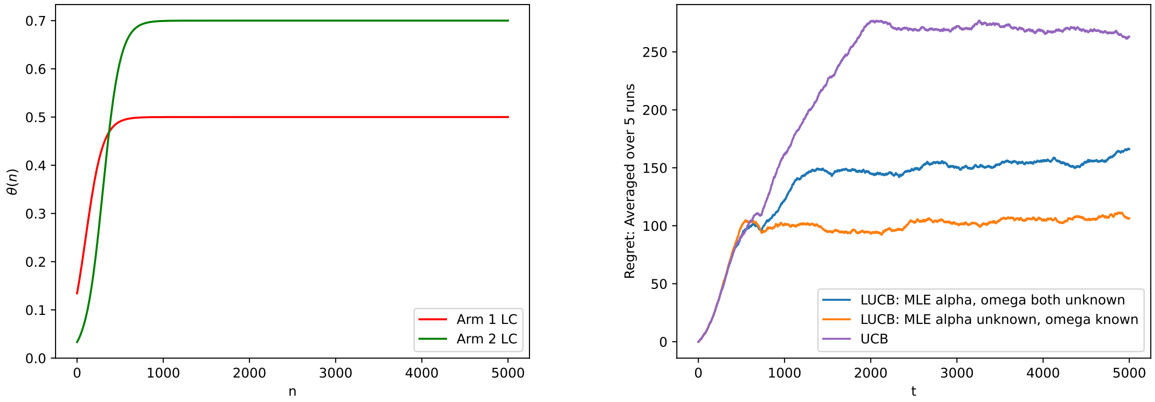
When the parametric forms of the learning curves are known, but one or more parameters are unknown, a systematic approach for finding unknown parameters is to use the MLE or MAP. So far, we have illustrated several examples where we obtain explicit formulas for $\hat{\alpha}^{MoM}$. More generally, one may write down the likelihood function (for MLE) or posterior probability (for MAP) and apply optimization algorithms, such as gradient descent (Goodfellow et al. 2016), to get point estimates.

For general parametric learning curves, standard results for the MLE imply that it will satisfy the bias condition (assuming typical identification and regulatory conditions); see section 6.5 of Lehmann and Casella (2006). We may also numerically verify the bias condition by taking the log scale in both the number of pulls (n) and the absolute values of empirical biases $|b_\alpha|$ and $|b_\omega|$ and fit the curves using linear regression, assuming the empirical biases are observable. The bias condition in Theorem 1, i.e., $|b_{a,n}| \leq \frac{1}{10}\sqrt{\frac{2\log n}{nC_a^w}}$, is equivalent to $\log|b_{a,n}| \leq -0.5\log n + \log\frac{1}{10}\sqrt{\frac{2\log n}{C_a^w}}$. A sufficient condition of the bias condition is that the slope of the fitted line is lower than -0.5. Note that the result of empirical verification of the bias conditions is instance specific.

In Example 3, we show the steps of estimating α and ω simultaneously using MLE and compare: (i) the regret of L-UCB with MLE estimators for both α and ω, (ii) UCB, and (iii) L-UCB with an MLE estimator for α and *known* ω. Comparing L-UCB estimating two unknown parameters with L-UCB with only α unknown shows that knowing ω reduces regret.

Example 3 (Estimate multiple unknown parameters.)**.** *Consider a 2-armed bandit with two "S"-shaped learning curves embedded in the reward function, respectively. Both learning curves share the parametric form:* $\theta_{\alpha,\omega}(n) = \frac{\alpha}{1+\exp(-0.01n-\omega)}$*. The true parameters for this example are* $\alpha_1 = 0.5$, $\omega_1 = -2$ *and* $\alpha_2 = 0.7$, $\omega_2 = -4$.

In Figure 5 and Figure EC.1, we empirically verify that the MLE estimators' biases of α_n^{MLE} *and* ω_n^{MLE} *are both* $o\left(\sqrt{\frac{\log n}{n}}\right)$*. Specifically, the biases of* $\alpha_{1,n}^{MLE}$ *and* $\omega_{1,n}^{MLE}$ *are* $O(n^{-0.97})$ *and* $O(n^{-1.28})$*, and the biases of* $\alpha_{2,n}^{MLE}$ *and* $\omega_{2,n}^{MLE}$ *are* $O(n^{-0.67})$ *and* $O(n^{-1.80})$*. Therefore, MLE estimators of the parameters meet the bias conditions in Theorem 1. We show the empirical verification figures for* $\hat{\alpha}_{i,n}^{MLE}$, $i = 1, 2$ *in Figure 5; figures illustrating bias scales for* $\hat{\omega}_{i,n}^{MLE}$ *are deferred to the appendix.*

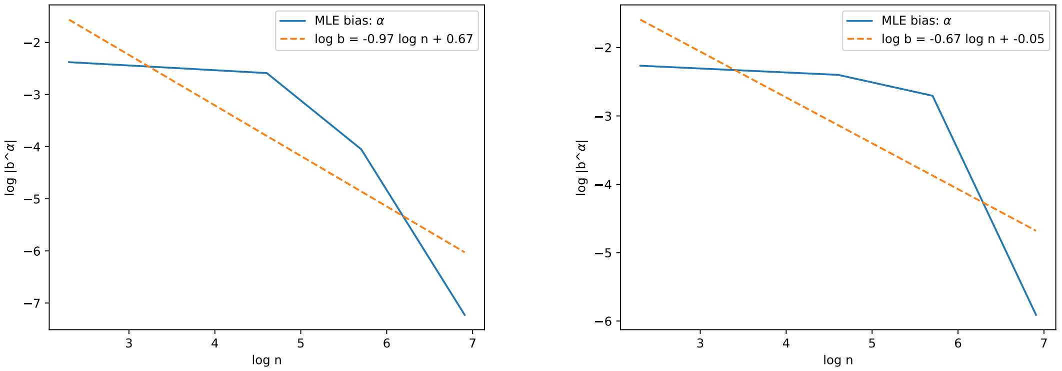


(a) Learning curves

(b) Regret comparison

Figure 4 **The two learning curves (left). On the right, we compare the regrets (averaged over five runs) of the L-UCB with MLE estimators where both parameters are unknown (blue), L-UCB with $\hat{\alpha}^{MLE}$ where ω is known (orange), and vanilla UCB (purple). The L-UCB: MLE with only α unknown plateaus to regret of approximately 100 after about 1000 trials, with α and ω unknown plateaus to regret of approximately 150 after 1200 trials, and the vanilla UCB plateaus to approximately 250 after about 2000 trials. Thus we see the benefit of utilizing the learning curves' parametric forms and knowing ω, respectively.**

Our L-UCB algorithm for the SLT problem only uses $\hat{\alpha}_{a,n}$'s for arm selection, while estimates of ω help construct the $\hat{\alpha}_{a,n}$'s and $C_{a,n}^{w}$'s and verify $\hat{\alpha}_{a,n}$'s bias conditions for Theorem 1 to hold. In general, dynamic learning algorithms may use vector parameter estimates for arm selection; in such cases, our Theorem 1 may be applied element-wise to obtain theoretical bounds. In rare scenarios



(a) The bias of $\alpha_{1,n}^{MLE}$ is $O(n^{-0.97})$ (b) The bias of $\alpha_{2,n}^{MLE}$ is $O(n^{-0.67})$

Figure 5 **Verifying the bias scales of $\alpha_{1,n}^{MLE}$ and $\alpha_{2,n}^{MLE}$, MLE estimators for arm 1 and arm 2's learning curves. The bias scales are both $o\left(\sqrt{\frac{\log n}{n}}\right)$, satisfying the bias condition in Theorem 1.**

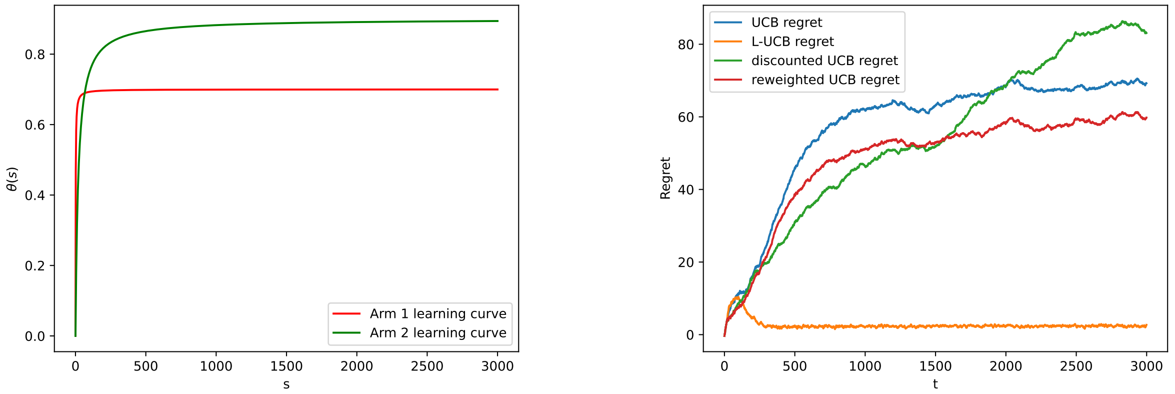
when verifying Theorem 1's premises is tricky, L-UCB with MLE or MAP is still applicable and can provide guidance for finding the optimal online policy.

4.7. L-UCB with Unknown Learning Curves

In order to apply MLE and MAP estimators, we assumed the parametric forms of the learning curves are known. When such information is not available, our L-UCB method can still be applied with an estimator $\hat{\alpha}$ that does not exploit the parametric form. For example, the vanilla UCB (a special case of L-UCB) is a nonparametric method, as it does not use any knowledge about the parametric form (or even that the expected rewards are non-stationary). There are other nonparametric estimators that adapt more quickly to nonstationary expected rewards: Instead of using an empirical mean where all historical observations are given the same weight $\frac{1}{n}$, we can assign more recent observations greater weight. For example, we may discount past observations, i.e., $\hat{\alpha}_{a,n}^{disc} := (\sum_{s=1}^{n} \delta_a^{n-s} r_a^{(s)}) / \frac{1-\delta_a^n}{1-\delta_a}$, where $\delta_a \in (0,1)$. We may also make the weights a function of n, e.g., $\hat{\alpha}_{a,n}^{rew} := (\sum_{s=1}^{n} s r_a^{(s)}) / \frac{n(n+1)}{2}$. Similar ideas have been studied in Garivier and Moulines (2011), see their D-UCB and SW-UCB, for example. In Section EC.8 we discuss the differentiation.

Figure 6b shows a comparison between L-UCB with MLE, vanilla UCB, discounted UCB ($\hat{\alpha}_{a,n}^{disc} := (\sum_{s=1}^{n} \delta_a^{n-s} r_a^{(s)}) / \frac{1-\delta_a^n}{1-\delta_a}$ and $\delta = 0.9$ in this example), and reweighted UCB ($\hat{\alpha}_{a,n}^{rew} := (\sum_{s=1}^{n} s r_a^{(s)}) / \frac{n(n+1)}{2}$); the latter two are special cases of L-UCB where the nonparametric estimators $\hat{\alpha}_{a,n}^{i}$ $(i = \{disc, rew\})$ do not utilize the parametric form of learning curves. The numerical setup is the same as those in Example 1 except that in Figure 6, we show a wider range of t.

Discounted UCB and reweighted UCB may be particularly useful when we know the expected rewards are nonstationary because they put more weight on recent observations. Reweighted UCB assigns recent observations more weight than UCB, therefore may respond more quickly to the



(a) Two learning curves

(b) Regret: L-UCB vs. nonparametrics

Figure 6 Same numerical setup as Example 1. The regret results shown are averaged over 20 instances. L-UCB obtains the lowest long-term regret, and reweighted UCB performs better than UCB in this instance.

nonstationary environment. Interestingly, discounted UCB has lower regret than reweighted UCB initially ($t \leq 1500$), but accumulates more regret than all other three algorithms once $t > 1500$ before it shows signs of potentially converging around $t = 3000$. This could result from assigning too much, constant weight to recent observations and discounting the past too aggressively, therefore, not efficiently using all the historical data (e.g., the weight assigned to the oldest observation is δ^{n-1}, which decays fast and can become very close to 0). Reweighted UCB is more sample efficient because all the sample weights are between $\left[\frac{2}{n(n+1)}, \frac{2}{n+1}\right]$. All of the nonparametric UCB algorithms perform much worse than the L-UCB, again because the latter exploits the parametric forms of the reward curves.

5. Fairness and the FL-UCB Algorithm

In this section we introduce probabilistic fairness definitions within the SLT context. We add constraints that require specific arms to be chosen with no less than certain predefined probabilities; one can interpret probabilistic fairness as long-term average max-min fairness. By properly configuring a set of fair probabilities, we guarantee that donor livers are equitably distributed to a broader range of recipients, rather than being offered continuously to a very narrow group.

Our first type of probabilistic fairness we call *best-K θ-fairness*, or *BK-fairness*, where we require the best $K(\leq |\mathcal{A}|)$ arms be chosen with probability greater than or equal to a vector of predefined levels, θ^{BK}. Note that the choices of θ^{BK} are constrained to render the BK-fairness concept well-defined, i.e. $|\theta^{BK}|_1 \leq 1$ and $\theta^{BK} \geq 0$, where $|X|_1$ is the $L1$ norm of vector X.

Definition 6 (Best-K θ-Fairness/BK-fairness)**.** *Any a in the set of the best-K arms, $\mathcal{A}^{BK}$, has to be chosen with probability no less than $\theta_a^{BK} - \epsilon$ when $t \to \infty$, for any $\epsilon \in (0, \min_{a \in \mathcal{A}_{BK}}\{\theta_a^{BK}\})$, where $\sum_{a \in \mathcal{A}_{BK}} \theta_a^{BK} \leq 1$ and $\theta^{BK} \geq 0$.*

We are interested in BK-fairness because more widespread use of SLT could bring benefits in practice. For example, if more transplant centers are capable of performing SLT, it could be easier to schedule surgeries and potentially facilitate logistics and reduce organ wastage.

The second type of fairness we define is *arbitrary arm fairness*, or *AA-Fairness*, which prioritizes a set of arbitrary arms, independent of surgeons' aptitudes or expertise. This could ensure that certain populations have access to organs, even if their outcomes are not among the K-best.

Definition 7 (Arbitrary-Arm θ-Fairness)**.** *For a set of arbitrarily-selected arms, $\mathcal{A}_A$, the vector of probabilities of being chosen is no less than $\theta^A \in [0,1]^{|\mathcal{A}_A|}$, where $|\theta^A|_1 \leq 1$.*

5.1. The FL-UCB Algorithm

We define a linear optimization program, FL-LP, as follows ($\mathcal{A}, \mathcal{A}_{BK}, \theta^{BK}, \mathcal{A}_A, \theta^A$ are inputs):

$$\max \quad \sum_{a \in \mathcal{A}} \alpha_a z_a \tag{11}$$

$$s.t. \quad z_a \geq \theta_a^A \qquad \forall a \in \mathcal{A}_A \tag{12}$$

$$z_a \geq \theta_a^{BK} \qquad \forall a \in \mathcal{A}_{BK} \tag{13}$$

$$\sum_{a \in \mathcal{A}} z_a = 1 \tag{14}$$

$$z_a \geq 0 \qquad \forall a \in \mathcal{A} \tag{15}$$

The solution to LP (11) - (15), z^*, gives the true optimal fair policy in an offline setting. In an online setting where α is not known, we use $B_{a,t,T_a(t-1)}$ instead of α_a, replacing (11) with

$$\max \quad \sum_{a \in \mathcal{A}} B_{a,T_a(t-1)} z_a. \tag{16}$$

Algorithm 2: The FL-UCB Algorithm Pseudo Code

```
1: **Initialization:** Select each arm $m_a$ times
2: **Update statistic:** $B_{a,T_a(t-1)} \leftarrow \hat{\alpha}_{a,T_a(t-1)} + \sqrt{\frac{2\log t}{C_a^w T_a(t-1)}}, \quad \forall a \in \mathcal{A}$
3: **Select arm:**
4: Sort $\{B_{a,T_a(t-1)}\}_{a=1}^{|\mathcal{A}|} : B_{(1),T_{(1)}(t-1)} \geq B_{(2),T_{(2)}(t-1)} \geq \ldots, B_{(K),T_{(K)}(t-1)}, \ldots, B_{(|\mathcal{A}|),T_{(|\mathcal{A}|)}(t-1)}$
5: $\mathcal{A}_{BK} \leftarrow \{B_{(1),T_{(1)}(t-1)}, B_{(2),T_{(2)}(t-1)}, \ldots, B_{(K),T_{(K)}(t-1)}\}$
                                        ▷ Construct a set ($\mathcal{A}_{BK}$) that contains the top-K indexes
6: $z^* \leftarrow$ SolveFLLP($\mathcal{A}, \mathcal{A}_{BK}, \theta^{BK}, \mathcal{A}_A, \theta^A$)
                        ▷ The objective of SolveFLLP is (16); the solution satisfies BK- and AA-fairness
7: Choose arm $a \in \mathcal{A}$ with probability $z_a^*$
8: **Increment** $t$ and **Go to Step 2**
```

Above is the pseudo code for our proposed FL-UCB algorithm, where the optimization of FL-LP is called as a subroutine. The objective function of SolveFLLP in step 6 is (16).

5.2. The FL-UCB Regret Bounds

When $\theta^A \neq \mathbf{0}$, the difference between the offline (optimal) policy without fairness constraints and an optimal fair policy is, in general, $O(t)$, by the definition of BK-fairness and AA-fairness. (Only when $\mathcal{A}_{BK}$ and $\mathcal{A}_A$ contain only the optimal arm does this fail to hold.) We therefore define the *price of fairness* in the SLT context.

Definition 8. *The price of fairness, or* PoF*, is the gap between the total reward of the optimal policy and the optimal fair policy.*

We thus define the difference between the objective value of the optimal fair policy and a given fair policy, which we call the *F-regret*; it is incurred solely due to a lack of information about the arm parameters. Specifically, in the original definition of *regret*, π_t^* is defined as the offline policy over all possible policies; now, we are restricting the feasibility set by imposing fairness constraints. Since the price of fairness causes an inevitable linear loss, we focus on lowering the additional loss by efficiently using information, i.e., controlling the *F-regret*. When appropriate, we may alternatively use the terms F-regret and regret without ambiguity.

Next, we analyze the regret upper bound for the proposed FL-UCB algorithm. The regret lower bound for FL-UCB is $O(\log t)$ because the vanilla bandit is a special case, and its regret lower bound is $O(\log t)$ (Lai and Robbins 1985). For convenience, we denote $a_{(i)}, i \in [|\mathcal{A}|]$, and $\alpha_{(i)}$ as the i-th best arm and its aptitude parameter, respectively, and let $\Delta_{a_{(i)},a_{(j)}} := \alpha_{(i)} - \alpha_{(j)}$, $i,j \in [|\mathcal{A}|]$. Recall that $r(\ell, a, (T_{a,t-1}, s_{a,t-1}))$ is the random reward of pulling arm $a \in \mathcal{A}$ with experience level $s_{a,t-1}$ (when arms are mutually independent, $s_{a,t-1} = T_{a,t-1}$). We further define $\bar{r}_a = \sup_t r_a^{(t)}$ and $\underline{r}_a = \inf_t r_a^{(t)}$. Theorem 2 establishes bounds on the FL-UCB regrets.

Theorem 2. *When LP* (11) $\sim$ (15) *has a unique solution:*

(a) The expected number of times that the (non-degenerate) solution of the LP with objective (16) *is different from that of* (11) $\sim$ (15), *satisfies*

$$\sum_{a \neq a^*} \mathbb{E}[T_a] \leq \left(\sum_{a \neq a^*} \frac{8 \log t}{C_a^w \Delta_a^2} + \sum_{k=2}^{K} \sum_{i=1}^{|\mathcal{A}|-K} \frac{8 \log t}{C_a^w \Delta^2_{a_{(k)},a_{(K+i)}}} + \frac{8 \log t}{C_a^w \Delta^2_{a_{(k-1)},a_{(k)}}} \right) + (2|\mathcal{A}| - K)(K+1)\zeta(1.24)$$

(b) The F-regret is bounded by

$$\mathbb{E}[R(t)] \leq \sum_{k=1}^{K} \sum_{i=1}^{|\mathcal{A}|-k} \left(\bar{r}_{a^*} - \underline{r}_{a_{(i)}} \right) \left(\frac{8 \log t}{C_a^w \Delta^2_{a_{(k)},a_{(k+i)}}} + 2\zeta(1.24) \right)$$

6. Extensions

This section discusses extensions of our model to incorporate delayed feedback and arm correlation.

6.1. Delayed Feedback

Some SLT outcomes are not immediately observed after the surgery, e.g., 1-month and 1-year survival. When the true outcome $r_a^{(i)}$ is only observed after some delay, we may use perioperative data and clinical metrics to provide an initial outcome estimate, $\hat{r}_a^{(i)}$, and replace it with the true outcome $r_a^{(i)}$ when it becomes available.

Corollary 1. *Let $k_a := O(1)$ be the maximum number of true rewards that haven't been revealed yet for arm a. Assume $\exists\, n_e > 0$, and the estimated outcome $\hat{r}_a^{(i)}$ and estimator function ϕ satisfy the property:*

$$e_{a,n} := \phi(\hat{r}_a^{(n)}, \dots, \hat{r}_a^{(n-k_a+1)}, r_a^{(n-k_a)}, \dots, r_a^{(1)}) - \phi(r_a^{(n)}, \dots, r_a^{(1)}) \le \frac{1}{40}\sqrt{\frac{\log n}{nC_a^w}}, \quad \forall n > n_e \tag{17}$$

When all premises in Theorem 1 hold; then, even when feedback is delayed by k_a for each arm a, each sub-optimal arm is pulled in expectation at most

$$\mathbb{E}[T_a(t)] \le \frac{8\log t}{C_a^w \Delta_a^2} + 2\zeta(1.063) \tag{18}$$

times, where $\zeta(1.063) \approx 16.45$, and $\zeta(s)$ is the Riemann zeta function, i.e. $\zeta(s) = \sum_{i=1}^{\infty} i^{-s}$. The expected cumulative regret of the L-UCB algorithm when feedback may be delayed is bounded by

$$\mathbb{E}[R(t)] \le \sum_{a \neq a^*} (\bar{r}_{a^*} - \underline{r}_a)\left(\frac{8\log t}{C_a^\omega \Delta_a^2} + 2\zeta(1.063)\right). \tag{19}$$

The proof is shown in Section EC.3.

Note that the estimator error bound in (17) is relatively mild: It is much looser than the error decay rate of taking a sample average while having k_a delayed, unobserved outcomes, which is at the scale of $O\left(\frac{1}{n}\right)$. For learning curves of the form: $\theta = \alpha g_\omega(s)$ and an MoM estimator $\hat{\alpha}_n^{MoM} = \frac{1}{n}\sum_{s=1, g_\omega(s)\neq 0}^{n} \frac{r^{(s)}}{g_\omega(s)}$, the decay rate of the MoM estimator's $e_{a,n}$ is also $O\left(\frac{1}{n}\right)$ which satisfies (17). Our assumption that k_a does not scale with n (the number of arm pulls of arm a) is reasonable in the SLT application because, in practice, only a finite number of livers become available within any fixed period; that is, there are a finite number of arm pulls during the survival period (e.g., one-year).

6.2. Incorporating Feature-Based Rewards and Arm Correlation

Each transplant surgery outcome/bandit reward is determined by the surgical team's proficiency (that is unknown and needs to be learned), the patient's clinical (e.g., serum bilirubin, creatinine, and the international normalized ratio) and demographic information (e.g., age, BMI), and the donated liver's compatibility (e.g., size matching, ABO compatibility) and quality (e.g., donor age and health, cold ischemia time.) Thus, a natural extension of our MAB model is to formulate feature-based rewards for each transplant surgery: Each arm is fully characterized by a potentially high-dimensional vector consisting of known patient and liver attributes, and the central planner learns the relationship between surgical teams' experience and transplant outcomes. Moreover, we can further decompose surgical proficiency to capture overlaps in required skills across surgeries. Feature-based rewards and high dimensionality in dynamic learning problems have been studied in revenue management contexts (Ban and Keskin 2021, Keskin et al. 2022). Exploring the salient surgery features and surgical teams' experience could be a promising direction and facilitate detailed characterization of likely correlated expected rewards for different arms.

In the appendix, we discuss a special type of arm correlation: Linear correlation, and discuss its impact on the optimal policy compared to a clairvoyant policy, and L-UCB/FL-UCB performances.

7. Numerical Study

We run numerical experiments based on real-world data to test the performance of our proposed algorithms. In Section 7.1 we describe the simulation setup for the SLT problem. In Section 7.2 we show that our proposed algorithms converge rapidly and demonstrate asymptotic advantages. The problem of showing theoretical bounds for small-t scenarios is beyond the scope of this paper, as the offline policy may change as t grows, as we illustrated in Example 1. Nevertheless, in Section 7.3, we illustrate a small-t scenario for personalized marketing problems with limited budget or a narrow opportunity window.

7.1. Simulation Setup for the SLT Problem

We consider the training and selection of medical teams as part of an SLT expansion effort coordinated by the OPTN, the central planner overseeing organ allocation in the US. We estimate parameters and generate outcomes based on Standard Transplant Analysis Research (STAR) files and Potential Transplant Recipient (PTR) dataset provided by Organ Procurement and Transplant Networks (OPTN) to capture current SLT practice. The "true" optimal cumulative reward for t is $\mathbb{E}\sum_{\tau=1}^{t} r(\ell_\tau, a_\tau^{\pi_t^*}, S_\tau^{\pi_t^*})$, where $\pi_t^* := \arg\max_\pi \sum_{\tau=1}^{t} r(\ell_\tau, a_\tau^{\pi}, S_\tau^{\pi})$.

The time horizon in this experiment is $\{1, 2, \ldots, 3600\}$; each time step corresponds to an arrival of a split liver graft and marks the beginning of a matching run (that may last for hours after the donor dies and donates their liver). Recall that more than 10% of all deceased-donor livers in the US are medically safe to split; there were 14905 deceased donors in total during 2022. We consider a geographical region that includes OPTN regions 2, 9, 10, 11, and Wisconsin and Illinois (see Section EC.6 for details about allocating heterogeneous livers as parallel MABs). Around 8000 livers are donated annually, and 10 large transplant centers locate in the 500NM Circle. While in theory, all medically-safe livers can be split, in conversations with UCSF transplant surgeons, they suggested it would be helpful to consider a more gradual rollout in the initial phase of SLT expansion to accommodate surgical learning. Thus, if we assume that 150 or $\sim 2\%$ of the total deceased-donor livers will be split for 300 surgeries in a year for the first two years, and 500 or $\sim 6\%$ livers to be used for 1000 SLT surgeries annually in the third to fifth years, the time horizon $\{1, 2, \ldots, 3600\}$ would be around five years at the typical deceased liver donation level in the US. We focus on identifying the arm(s) with the highest aptitude(s) such that they would perform the best in the long term, i.e., expanding the base of SLT among transplant teams with the highest potential. We investigate how we can accelerate bandit learning by incorporating information about the learning curve structure via our proposed FL-UCB algorithms.

SLT has been primarily practiced in a few big TCs in the US since its development in the 1980s (Ge et al. 2020, Duke 2021); thus, there is no historical data for widespread SLT learning. Nevertheless, based on findings from existing studies on medical learning and the nature of SLT surgeries (involving repetitive tasks such as dividing and connecting blood vessels), the medical teams' learning curves likely follow an 'S'-shaped structure (Pusic et al. 2015, Le Morvan and Stock 2005). The bandit rewards or SLT outcomes are 1-year graft survivals, primarily dependent on surgeon proficiency and experience; 1-year graft survival may also correlate with donor and recipient age and the recipients' health conditions. These factors and surgical expertise collectively determine the MAB asymptotic rewards or the arm's full potential/aptitude. There has been ongoing research investigating how survival outcomes depend on TC expertise, high-dimensional demographics, and perioperative clinical metrics; unfortunately, there are no exact mappings from these factors to the outcomes. As discussed in Section 6.2, high-dimensional feature-based dynamic learning in SLT is a potential research direction. Here, we simulate the arm parameters and outcome distributions based on historical data without specifying the exact feature mapping.

Specifically, we formulate the reward functions of arms following the Sigmoid curve; each with hidden aptitude parameter, where the bounds of the range are estimated directly from the STAR files. Except for few big TCs that already perform SLTs regularly, most TCs need to learn SLT with limited existing experience/initial proficiency and overcome barriers in the initial phase of surgical learning (characterized by ω). Since we do not have a direct data source, as SLT has not been widely learned or practiced, we consulted UCSF surgeons, and they believe the range $\omega \in [1, 14]$ is realistic: Typically, after performing $6 \sim 15$ SLT surgeries, a medical team can be considered sufficiently experienced and the proficiency starts to stabilize. Factoring in existing experience and skills transferred from similar surgeries, we arrive at $\omega \in [1, 14]$. We assume that the ranges of α's that we draw from are the same as the ranges of long-term expected outcomes in historical data containing mostly traditional WLT and limited SLT surgeries, $(0.3, 0.95)$; recent findings show that SLT outcomes can be comparable to those of WLT (Hackl et al. 2018). The parameters are specified in Table 1, where the learning curves follow (20):

$$\theta(\alpha, s) = \frac{\alpha}{1 + \exp(-s + \omega)} \tag{20}$$

Parameter	Value	Comment
Number of arms	50	Each arm has a learning curve
BK-Fairness	$(K, \theta^{BK}) = (1, 0.05)$	$K = 1$ implies no BK constraint
AA-Fairness	$\theta^{A} = 0.001$	Uniform for every arm; $\mathcal{A}_A = \mathcal{A}$
Aptitude/Full potential	$\alpha \in (0.3, 0.95)$	α unknown
Initial setup cost	$\omega \in [1, 14]$	Known, existing skills

Table 1 Experiment parameters. More details can be found in Section EC.6.

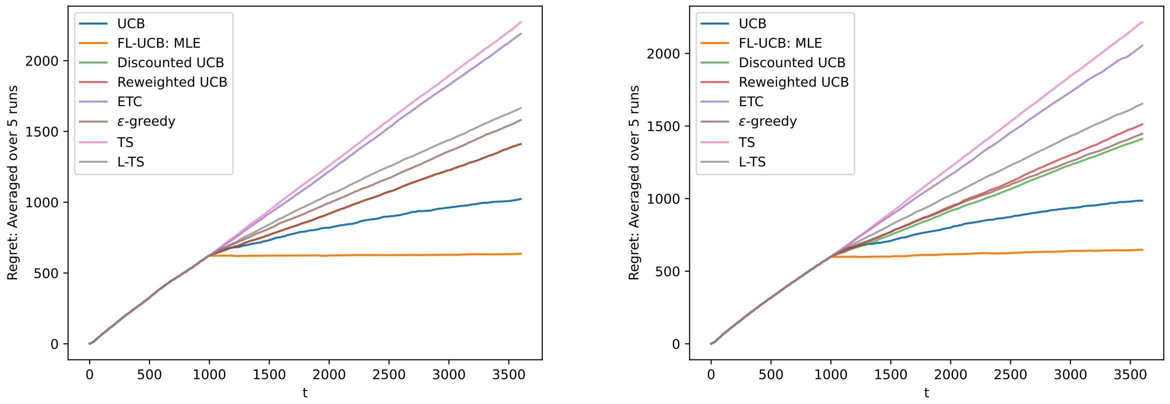
In one set of the simulation, we assume the true rewards, 1-year graft survivals, are not observed immediately after the surgery. Since the time horizon is $\{1, \ldots, 3600\}$ and we consider an approximately five-year, gradual rollout of SLT expansion. The delay in observing our true rewards are 300, 300, 1000, 1000, and 1000, for SLT surgeries performed in the 1st to 5th year, respectively. In other words, the true rewards, i.e., 1-year graft survival, is delayed 300-time steps if an arm pull takes place in the first two years and 1000-time steps in the third to fifth years. Before the true rewards are observed, we have 1-year survival estimates based on perioperative clinical metrics and demographic information. The prediction accuracy for 1-year graft survival can be as good as around 85% (Kantidakis et al. 2020, Nitski et al. 2021); we choose to be more conservative in this simulation and assume the accuracy for the estimates is 0.6. See Section EC.6 for more details.

Given the experimental configurations above, we compare the performances (i.e., regrets) of the FL-UCB algorithm with MLE against seven other bandit algorithms — vanilla UCB, discounted UCB, reweighted UCB, ϵ-greedy, explore-then-commit (ETC), vanilla Thompson sampling (TS) (Lattimore and Szepesvári 2020), and learning-enhanced Thompson sampling (L-TS) where we infuse learning-curve information into the TS posterior updating function.

Vanilla UCB, discounted UCB, and reweighted UCB can be viewed as special cases of FL-UCB; they do not assume knowledge of the parametric form of learning curves, see Section 4.7. In our setting, the parametric forms of medical learning are known; FL-UCB with MLE can leverage this knowledge and accelerate bandit learning and achieve faster convergence compared to these nonparametric methods. Nevertheless, in scenarios where either the existence of endogenous learning or the parametric form is unknown, FL-UCB with nonparametric estimators can usually generalize well. As our paper focuses on FL-UCB with MLE and nonparametric estimators we do not elaborate on L-TS; nevertheless, numerical results show that L-TS performs consistently better than canonical TS when endogenous learning is present and is second only to FL-UCB when the true feedback is delayed. For all bandit algorithms, we start with 20 round robins and report regrets/rewards that are averages over five runs. We also provide numerical results on how the offline reward (i.e., optimal cumulative reward) and PoF (i.e., price of fairness) grow as functions of t.

7.2. Numerical Results for the SLT Problem

Figures 7 and 8 show the total regret of each algorithm as a function of t, the total number of surgeries performed; while figure 9 illustrates the optimal cumulative reward, PoF, and PoF percentage as a function of t. Specifically, Figure 7 demonstrates that when surgical learning occurs, FL-UCB outperforms the benchmarks as it has the lowest regrets and converges rapidly, whether the fairness constraints are imposed or not. Figure 8 shows that when the true rewards are delayed and 60%-accurate reward estimates are available, the advantage of FL-UCB is preserved: FL-UCB with the MLE estimator still outperforms other bandit algorithms and achieves similar regrets, while UCB and nonparametric L-UCB variants incur greater regrets compared to the no-delay simulations, regardless of the presence of fairness constraints.

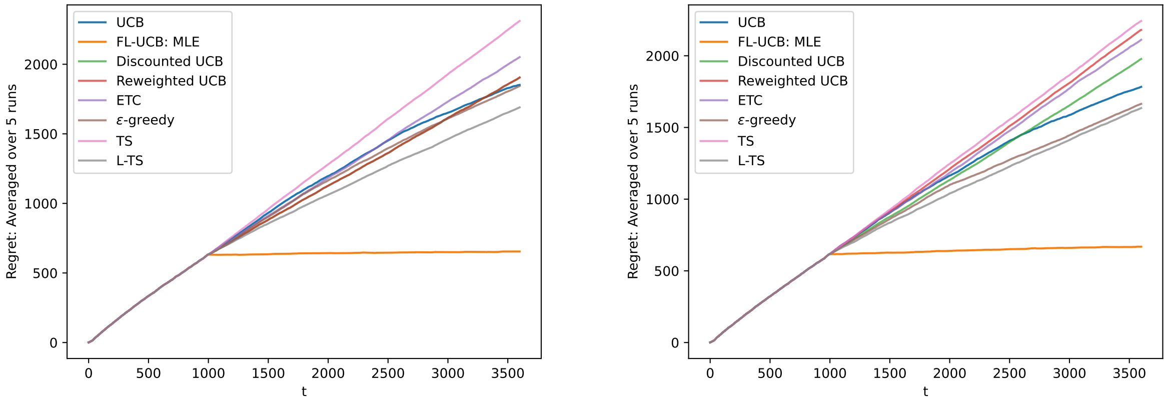


(a) Regret w/o fairness

(b) F-regret w/ fairness

Figure 7 **Comparing FL-UCB regret against benchmarks when medical learning exists and assuming no delay in observing true rewards. FL-UCB with MLE estimation has the lowest regrets and converges rapidly.**

Figure 9 shows that the PoF, the loss in utility in optimal fair solutions relative to optimal solutions without fairness constraints, is small (although it is still $O(t)$, as the PoF / Optimal cumulative reward ratio remains constant as t grows). In figure 10 we illustrate the breakdown of data points available at time t, $t \in \{1, \ldots, 3600\}$. For the first 300 time steps, all available information comes from outcome estimates, while in later stages, only a dwindling proportion of cumulative rewards is delayed and requires prediction. Specifically, the number of delayed rewards is t for $t \leq 300$, 300 for $t \in [301, 900]$, and $\min\{1000, t-600\}$ when $t \geq 900$.



(a) Regret w/o fairness

(b) F-regret w/ fairness

Figure 8 **Comparing FL-UCB regret against benchmarks when medical learning exists and rewards (i.e., 1-year graft survival) are delayed. We assume estimates based on demographics and perioperative clinical metrics are available and are 60% accurate. FL-UCB with MLE estimation learns efficiently in the initial round-robin exploration phase (where each arm observes 12 true outcomes and 8 estimated outcomes) and still has the lowest regret and converges fast. Meanwhile, UCB regrets are much higher when the true feedback is delayed.**

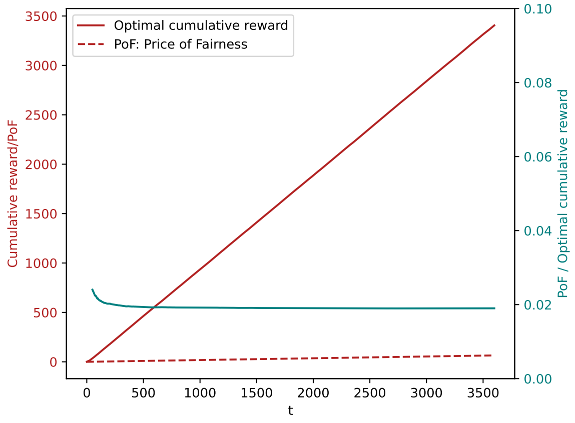


Figure 9 **[PoF / Optimal cumulative reward] is constant, i.e., PoF is $O(t)$; when $t < 200$, the ratio could be subject to numerical instability.**

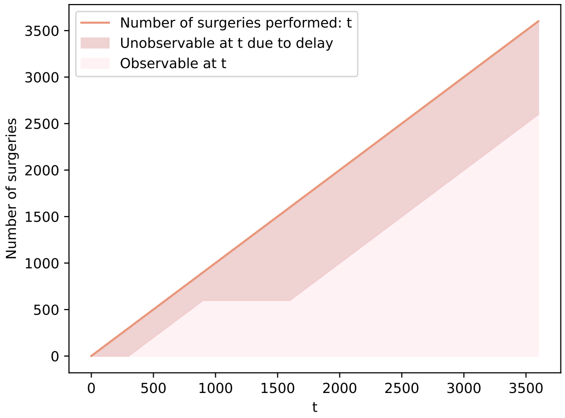


Figure 10 **The delay in observing rewards. For each t, the true rewards are not revealed until after some delay and can only be estimated using a 60% accurate surrogate.**

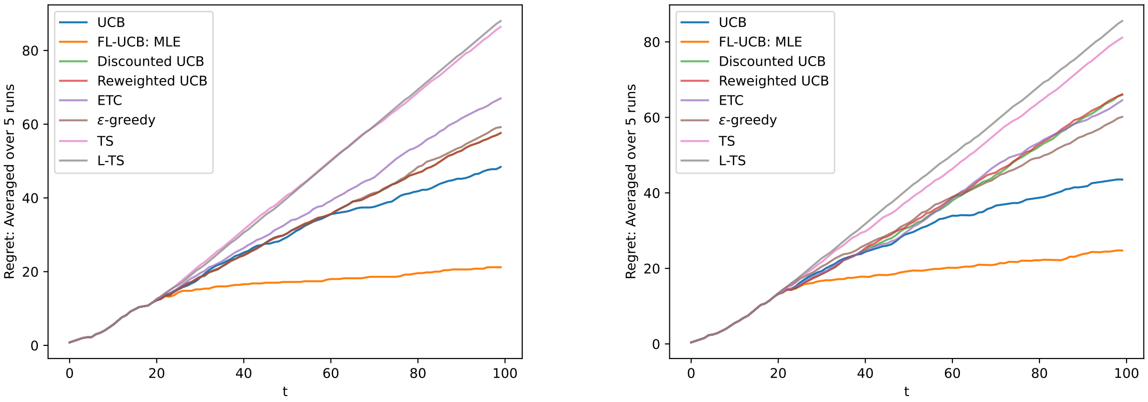
7.3. Simulation Setup and Results for the Personalized Marketing Problem

We simulate a personalized marketing problem where marketers intend to leverage the mere exposure effect. For this numerical experiment, we are particularly interested in the performance of FL-UCB with MLE and how it compares to other standard MAB algorithms in a short time horizon: $\{1,...,100\}$. A small-t scenario may arise when a new company is building its customer base on a budget and must present satisfactory outcomes to investors for earning renewed support. Specifically, we consider a total of 100 encounters possible with a specific customer. The company may have produced 10 video ads of different styles featuring the same or similar products for marketing the brand on social media, e.g., Instagram or YouTube. Suppose the company has targeted a specific group of customers and employs a recommendation algorithm to determine which video ads to be sent to the customers. We consider the same algorithms used in the previous numerical simulation. Instead of fairness constraints, the company might want to ensure a certain level of variety in its interaction with (potential) customers, thus we set the "fairness level" $\theta^A = 0.01$. Customers may have existing exposure to the product, which could expedite the marketing campaign and lead to purchases sooner, or inversely impact the personalized marketing effort due to negative prior experience; we capture existing exposure levels using ω. In addition we consider the possibility that customers may not immediately place an order even they have decided to make a purchase. We assume that 60%-accurate estimates based on customer activities on the social platform (including the time spent viewing the video ads, click streams, website visits, etc.) are immediately available to the marketers. We consider several possible reward curves to characterize the exposure effects, including the S-shaped curves, diminishing reward curves. Marketers do not observe the parametric forms of the curves or the parameters. Table 2 summarizes the parameters and values used for this personalized marketing problem.

Figure 11a and Figure 11b show that FL-UCB has the lowest regrets when the reward feedback is immediate, i.e., when customers make the decision to purchase, they place the orders and the company receives them immediately. In figure 11b, imposing variety requirements increases the regrets of FL-UCB, discounted UCB, and reweighted UCB, while other algorithms are not negatively affected. Figure 12a and 12b demonstrate that FL-UCB's consistent advantage over other MAB algorithms: It has the lowest regrets even when reward feedback is delayed. Nevertheless, compared to the no-delay scenario, the regrets of FL-UCB is significantly higher and there is no sign of converging soon.

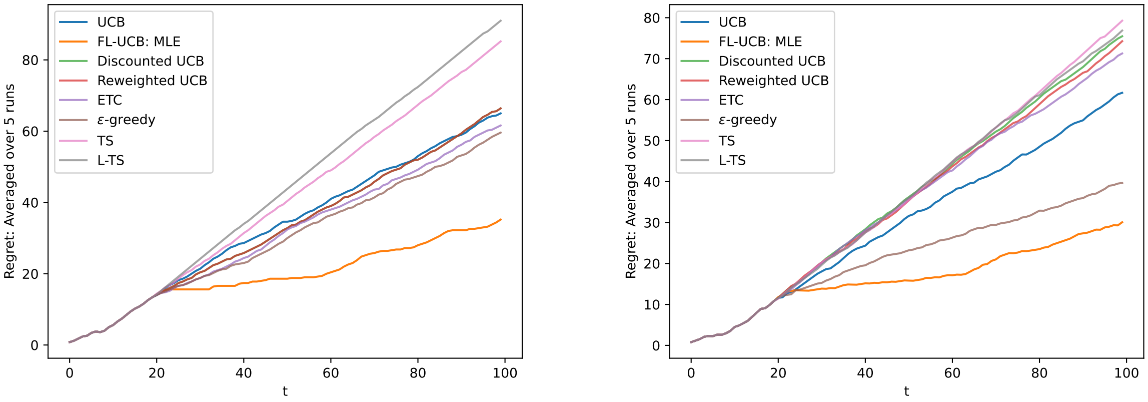
Parameter	Value	Comment
Time horizon	$\{1,...,100\}$	Reward curves reflect exposure effects
Number of arms/styles	10	Reward curves reflect exposure effects
BK-Fairness	$(K, \theta^{BK}) = (1, 0.05)$	Selecting the most effective style
AA-Fairness	$\theta^A = 0.01$	Enforced variety; $\mathcal{A}_A = \mathcal{A}$
Aptitude/Full potential	$\alpha \in (0.01, 0.99)$	α unknown, ultimate likelihood of purchase
Initial setup cost	$\omega \in [-5, 5]$	The amount of existing exposure

Table 2 Experiment parameters for the personalized marketing example.



(a) Regret w/o variety requirements (b) F-regret w/ variety requirements

Figure 11 **Comparing FL-UCB regret against benchmarks when the exposure effect exists and there is no delay in purchasing products/services. FL-UCB with MLE estimation learns efficiently in the initial round-robin exploration phase and has the lowest regret in the short horizon, with or without variety requirements. Imposing variety requirements slow down the FL-UCB algorithm in its efficient identification and leads to higher regrets (not including the price of fairness).**



(a) Regret w/o variety requirements (b) F-regret w/ variety requirements

Figure 12 **Comparing FL-UCB regret against benchmarks when the exposure effect exists and rewards are delayed (e.g., customers may view the ads, decide to buy immediately, but place orders later). We assume estimates based on customer profiles and their activities are available and are 60% accurate. FL-UCB with MLE estimation learns efficiently in the initial round-robin exploration phase (where each arm has two estimated outcomes) and still has the lowest regret, though it shows no sign of converging in the 100th round. Compared to the no-delay scenario, all MAB algorithms incur much higher regrets when the observations of purchases are delayed. Except that ϵ-greedy seems to be relatively robust, maintaining approximately the same amount of regrets compared to the no-delay scenario.**

8. Concluding Remarks

To address the trade-off between exploration versus exploitation in SLT allocation, we formulated an MAB variant with endogenous learning curves embedded in the arm reward functions. Our model also enables the incorporation of learning curves, fairness constraints, and nonparmetrics (i.e., UCB variants that do not assume knowing the parametric forms). We propose UCB variants—the L-UCB and FL-UCB algorithms, which converge to the optimal offline policy incurring optimal total regret: $O(\log t)$ scale. Application of our model can potentially shed light to strategies (e.g. liver allocation, transplant center/surgery selection) to expand SLT use in the US, as well as in other settings characterized by decision-making with learning through experience or the mere exposure effect, such as personalized marketing. Methodologically, our formulation and proposed UCB variants significantly extend the canonical UCB to bandit problems where the estimates of the unknown parameters can be different from the empirical mean.

There are several potential directions for future work that will generalize and deepen the conclusions of this paper. Methodologically, we developed non-parametric methods, discounted UCB and reweighted UCB, and compared them against parametric L-UCB; results show that we can leverage the parametric forms of learning curves through MoM or MLE or MAP estimators in L-UCB to achieve lesser regrets. In scenarios where the parametric form is unknown, reweighted UCB and discounted UCB and vanilla UCB can be used, and the former two may sometimes perform better than vanilla UCB, but not always. An intriguing future direction is to study the selection of good nonparametric estimators under the FL-UCB framework and estimators' robustness in various applications. Another promising direction is to incorporate feature-based rewards that explore the correlation between the expected rewards of different arms, please refer to Section 6.2 for more details. An important follow-up work is to study the short-term performance of L-UCB and FL-UCB in MAB with nonstationary reward curves, and potentially prove theoretical bounds or propose novel algorithms with theoretical guarantees. Evaluating the performance of adaptive UCB variants including L-UCB in MABs with many arms and endogenous reward curves is another promising direction. On the application side, an important extension is to incorporate consumer behavior research findings to quantify the mere exposure effect in greater detail and elaborately characterize customers' variety-seeking behaviors (for enhancing the performance of FL-UCB). Explicit and implicit costs (e.g., explicit payment to platforms, implicit fatigue) associated with exploring customer tastes can also be factored into the definition of the bandit rewards and subroutines of the L-UCB algorithm. In addition, both the substitution effect and complementary effect might play a role in the effectiveness of personalized marketing; one may capture these subtle interplay by allowing arms to be correlated and estimate the correlation structure based on relevant consumer behavior research findings. Finally, despite consumer learning and the mere exposure effect, consumers may become overwhelmed by the volume of advertising or disengaged over time. Therefore, a better understanding of consumer learning, customer fatigue, and customer journeys is essential for determining marketing frequencies, patterns, and sequences. To achieve this, again, one may look into relevant marketing research describing and quantifying consumer behaviors.

References

Akan M, Alagoz O, Ata B, Erenay FS, Said A (2012) A Broader View of Designing the Liver Allocation System. *Operations Research* 60(4):757–770, ISSN 0030-364X, URL `http://dx.doi.org/10.1287/opre.1120.1064`.

Anderer A, Bastani H, Silberholz J (2022) Adaptive clinical trial designs with surrogates: When should we bother? *Management Science* 68(3):1982–2002.

Ban GY, Keskin NB (2021) Personalized dynamic pricing with machine learning: High-dimensional features and heterogeneous elasticity. *Management Science* 67(9):5549–5568.

Bertsimas D, Farias VF, Trichakis N (2011) The price of fairness. *Operations research* 59(1):17–31.

Bertsimas D, Papalexopoulos T, Trichakis N, Wang Y, Hirose R, Vagefi PA (2020) Balancing efficiency and fairness in liver transplant access: tradeoff curves for the assessment of organ distribution policies. *Transplantation* 104(5):981–987.

Besbes O, Gur Y, Zeevi A (2019) Optimal exploration–exploitation in a multi-armed bandit problem with non-stationary rewards. *Stochastic Systems* 9(4):319–337.

Chandra S, Verma S, Lim WM, Kumar S, Donthu N (2022) Personalization in personalized marketing: Trends and ways forward. *Psychology & Marketing* 39(8):1529–1562.

Cheung WC, Simchi-Levi D, Zhu R (2020) Reinforcement learning for non-stationary markov decision processes: The blessing of (more) optimism. *International Conference on Machine Learning*, 1843–1854 (PMLR).

den Boer AV, Keskin NB (2022) Dynamic pricing with demand learning and reference effects. *Management Science* 68(10):7112–7130.

Duke H (2021) Duke health blog. `https://www.dukehealth.org/blog/split-liver-transplant-saves-two-lives-one-donor-liver#:~:text=In\%20a\%20split\%20liver\%20transplant,baby\%20or\%20a\%20small\%20child.`, accessed: 2023-03-20.

Emre S, Umman V (2011) Split liver transplantation: An overview. *Transplantation Proceedings*, volume 43, 884–887, ISSN 00411345, URL `http://dx.doi.org/10.1016/j.transproceed.2011.02.036`.

Garivier A, Ménard P, Stoltz G (2019) Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research* 44(2):377–399.

Garivier A, Moulines E (2011) On upper-confidence bound policies for switching bandit problems. *Algorithmic Learning Theory: 22nd International Conference, ALT 2011, Espoo, Finland, October 5-7, 2011. Proceedings 22*, 174–188 (Springer).

Ge J, Perito ER, Bucuvalas J, Gilroy R, Hsu EK, Roberts JP, Lai JC (2020) Split liver transplantation is utilized infrequently and concentrated at few transplant centers in the united states. *American Journal of Transplantation* 20(4):1116–1124.

Goodfellow I, Bengio Y, Courville A (2016) *Deep learning* (MIT press).

Grover A, Markov T, Attia P, Jin N, Perkins N, Cheong B, Chen M, Yang Z, Harris S, Chueh W, et al. (2018) Best arm identification in multi-armed bandits with delayed feedback. *International Conference on Artificial Intelligence and Statistics*, 833–842 (PMLR).

Hackl C, Schmidt KM, Süsal C, Döhler B, Zidek M, Schlitt HJ (2018) Split liver transplantation: current developments. *World Journal of Gastroenterology* 24(47):5312.

Hardy G, Littlewood J, Polya G (1952) Inequalities cambridge univ. *Press, Cambridge* (1988).

Janiszewski C (1993) Preattentive mere exposure effects. *Journal of Consumer research* 20(3):376–392.

Joulani P, Gyorgy A, Szepesvári C (2013) Online learning under delayed feedback. *International Conference on Machine Learning*, 1453–1461 (PMLR).

Kantidakis G, Putter H, Lancia C, Boer Jd, Braat AE, Fiocco M (2020) Survival prediction models since liver transplantation-comparisons between cox models and machine learning techniques. *BMC Medical Research Methodology* 20:1–14.

Keskin NB, Li M (2021) Selling quality-differentiated products in a markovian market with unknown transition probabilities. *Available at SSRN 3526568* .

Keskin NB, Li Y, Song JS (2022) Data-driven dynamic pricing and ordering with perishable inventory in a changing environment. *Management Science* 68(3):1938–1958.

Keskin NB, Zeevi A (2017) Chasing demand: Learning and earning in a changing environment. *Mathematics of Operations Research* 42(2):277–307.

Lai TL, Robbins H (1985) Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics* 6(1):4–22.

Lattimore T, Szepesvári C (2020) *Bandit algorithms* (Cambridge University Press).

Le Morvan P, Stock B (2005) Medical learning curves and the kantian ideal. *Journal of medical ethics* 31(9):513–518.

Lehmann EL, Casella G (2006) *Theory of point estimation* (Springer Science & Business Media).

McDiarmid C (1989) On the method of bounded differences. *Surveys in combinatorics* 141(1):148–188.

McDiarmid C (1998) Concentration. *Probabilistic methods for algorithmic discrete mathematics*, 195–248 (Springer).

Montoya RM, Horton RS, Vevea JL, Citkowicz M, Lauber EA (2017) A re-examination of the mere exposure effect: The influence of repeated exposure on recognition, familiarity, and liking. *Psychological bulletin* 143(5):459.

Nitski O, Azhie A, Qazi-Arisar FA, Wang X, Ma S, Lilly L, Watt KD, Levitsky J, Asrani SK, Lee DS, et al. (2021) Long-term mortality risk stratification of liver transplant recipients: real-time application of deep learning algorithms on longitudinal data. *The Lancet Digital Health* 3(5):e295–e305.

OPTN, UNOS (2016) Split Versus Whole Liver Transplantation OPTN/UNOS Ethics Committee. Technical report.

Perito ER, Roll G, Dodge JL, Rhee S, Roberts JP (2019a) Split liver transplantation and pediatric waitlist mortality in the united states: potential for improvement. *Transplantation* 103(3):552–557.

Perito ER, Roll G, Dodge JL, Rhee S, Roberts JP (2019b) Split Liver Transplantation and Pediatric Waitlist Mortality in the United States: Potential for Improvement. *Transplantation* 103(3):552–557, ISSN 00411337, URL `http://dx.doi.org/10.1097/TP.0000000000002249`.

Pusic MV, Boutis K, Hatala R, Cook DA (2015) Learning curves in health professions education. *Academic Medicine* 90(8):1034–1042.

Rawls J (2001) *Justice as fairness: A restatement* (Harvard University Press).

Schumann C, Lang Z, Mattei N, Dickerson JP (2019) Group fairness in bandit arm selection. *arXiv preprint arXiv:1912.03802* .

Thompson WR (1933) On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika* 25(3/4):285–294.

UNOS (2020a) Liver and intestinal organ distribution based on acuity circles to be implemented feb. 4, 2020. URL `https://unos.org/news/pre-imp-notice-liver-intestinal-dist-acuity-circles-feb-4-2020/#:~:text=The\%20acuity\%20circle\%20policy\%20replaces,donor\%20hospital\%20and\%20transplant\%20hospital`.

UNOS (2020b) United networks of organ sharing data. `https://optn.transplant.hrsa.gov/data/view-data-reports/national-data/`.

Vulchev A, Roberts JP, Stock PG (2004) Ethical issues in split versus whole liver transplantation. URL `http://dx.doi.org/10.1111/j.1600-6143.2004.00630.x`.

Zenios SA, Chertow GM, Wein LM (2003) Dynamic Allocation of Kidneys to Candidates on the Transplant Waiting List. *Operations Research* 48(4):549–569, ISSN 0030-364X, URL `http://dx.doi.org/10.1287/opre.48.4.549.12418`.

Acknowledgments

The authors gratefully appreciate the valuable input and advice from Dr. John P. Roberts (MD, MS) and Dr. Emily R. Perito (MD, MS) from UCSF School of Medicine, who helped us understand the complex SLT logistics and current SLT practice in the US.

Electronic Companions

EC.1. Proofs for Theoretical Results in Section 4

EC.1.1. Alternative Statement of Theorem 1 and Proof

While Theorem 1 is a canonical statement of the regret upper bounds, Theorem EC.1 is a stronger statement mathematically.

Theorem EC.1. *Let $\hat{\alpha}_{a,n}$ be the estimator for the aptitude of arm a, i.e. α_a, after it has been chosen n times. Suppose $\hat{\alpha}_{a,n}$ has a per-coordinate difference bound with parameter $C^w_{a,n}$ and bias $b_{a,n}$. Define $\delta_{a,\tau,n} := \sqrt{\frac{2\log\tau}{nC^w_{a,n}}}$ For any sub-optimal arm a, if there exists a $u_{a,t} \in [1,t]$ such that $\Delta_a \geq 2\delta_{a,\tau,n}$ and $|b_{a,n}| \leq \frac{1}{10}\delta_{a,\tau,n}$ hold for any $t \geq \tau \geq n \geq u_{a,t}$, then arm a is pulled on average at most*

$$\mathbb{E}[T_a(t)] \leq u_{a,t} + 2\zeta(1.24)$$

times, where ζ is the Riemann zeta function, i.e. $\zeta(s) = \sum_{n=1}^{+\infty} n^{-s}$, and $\zeta(1.24)$ is approximately 4.76. If such a $u_{a,t}$ exists for any sub-optimal arm, then the expected cumulative regret is bounded by

$$\mathbb{E}[R_t] \leq \sum_{a \neq a^*} \left(\bar{r}_{a^*} - \underline{r}_a\right)\left(u_{a,t} + 2\zeta(1.24)\right)$$

Remark: Before we prove this theorem, we show its application in some simple cases.

First, when $\hat{\alpha}_{a,n}$ is the empirical mean of n independent Bernoulli random variables or any random variables on $[0,1]$, we have $C^w_{a,n} = 1$ and $b_{a,n} = 0$. We may choose $u_{a,t} = \frac{8\log t}{\Delta_a^2}$, indicating that this theorem recovers the bound of the vanilla UCB yet with the larger constant $2\zeta(1.24) \approx 9.52$ compared to $\frac{\pi^2}{3} \approx 3.29$. The larger constant results from the loose inequality dealing with the bias, i.e. as we decrease the $\frac{1}{10}$ in $|b_{a,n}| \leq \frac{1}{10}\delta_{a,\tau,n}$ towards 0, the constant will approach $\frac{\pi^2}{3}$.

Second, if we scale the value of $\hat{\alpha}_{a,n}$ and the sub-optimal gap by ℓ, then $C^w_{a,n}$ becomes $\frac{1}{\ell^2}$, and thus $u_{a,n}$ is unchanged. This indicates the bound is scale-free.

Third, when $\hat{\alpha}_{a,n}$s have smaller and/or different $C^w_{a,n}$s and still zero bias, and when $C^w_a := \inf_n C^w_{a,n} > 0$, i.e. when $C^w_{a,n}$ is uniformly bounded by a constant from below, we know the minimal $u_{a,t}$ is at most $\frac{2\log t}{C^w_a \Delta_a^2}$ (because we proved $\frac{2\log t}{C^w_a \Delta_a^2}$ is a valid choice for $u_{a,t}$ in Theorem 1) and therefore $\mathbb{E}[T_a(t)]$ is still in $O(\log t)$ scale, although the coefficient is larger.

Fourth, when $C^w_a := \inf_n C^w_{a,n} > 0$ and $C^b_a := \sup_n \sqrt{n}|b_{a,n}| < +\infty$, i.e. $|b_{a,n}| = O\left(\frac{1}{\sqrt{n}}\right)$, we may let $u_{a,t} = \max\left\{\exp\left(C^w_a (C^b_a)^2/200\right), \frac{2\log t}{C^w_a \Delta_a^2}\right\}$, and then $\mathbb{E}[T_a(t-1)]$ is still in $O(\log t)$ scale.

Fifth, similarly, as long as $C^w_a := \lim_{n\to+\infty} \frac{nC^w_{a,n}}{\log n} \geq \frac{8}{\Delta_a^2}$, i.e. either $C^w_{a,n} = \Omega(\frac{\log n}{n})$ or $C^w_{a,n} = \Theta(\frac{\log n}{n})$ but $C^w_{a,n} \leq \frac{8\log n}{n\Delta_a^2}, \forall n$, such a $u_{a,t}$ exists, but $u_{a,t}$ might be in $\Omega(\log t)$. Again, the exact value of $u_{a,t}$

is beyond our concern, because we aim to provide a bound for a general scenario. When $C^w_{a,n} \to 0$, $|b_{a,n}| = O\left(\sqrt{\frac{\log n}{n}}\right)$ is a sufficient condition of the existence of such a $u_{a,t}$.

Sixth, in contrast, when $C^w_{a,n}$ diminishes too fast, i.e. $C^w_{a,n} = o(\frac{\log n}{n})$, $\delta_{a,t,n}$ is no longer a decreasing function of n. This implies $\delta_{a,t,t}$ might be greater than Δ_a for any arbitrarily large t. Hence, no feasible $u_{a,t}$ exists for large t and this theorem is not applicable to these cases. Again, the exact or approximate threshold of the arbitrarily large value of t is not related to this theorem which focuses on what we can bound for an estimator with good properties, i.e. large $C^w_{a,n}$ and small $|b_{a,n}|$.

Below, we show the full proof for Theorem EC.1 and Theorem 1.

Proof: Let $a \in \mathcal{A}$, $\tau \in \mathcal{T}$, and $n := T_a(\tau - 1)$. And we derive probabilistic bounds for $\hat{\alpha}_{a,n}$,

$$\begin{aligned}
P\left(\hat{\alpha}_{a,n} - \alpha_a \geq \varepsilon\right) &= P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] + \mathbb{E}[\hat{\alpha}_{a,n}] - \alpha_a \geq \varepsilon\right) \\
&\leq P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] + |\mathbb{E}[\hat{\alpha}_{a,n}] - \alpha_a| \geq \varepsilon\right) \\
&= P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] \geq \varepsilon - |b_{a,n}|\right) \\
P\left(\hat{\alpha}_{a,n} - \alpha_a \leq -\varepsilon\right) &= P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] + \mathbb{E}[\hat{\alpha}_{a,n}] - \alpha_a \leq -\varepsilon\right) \\
&\leq P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] - |\mathbb{E}[\hat{\alpha}_{a,n}] - \alpha_a| \leq -\varepsilon\right) \\
&= P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] \leq -\varepsilon + |b_{a,n}|\right)
\end{aligned}$$

Let $\bar{\varepsilon} := \varepsilon - |b_{a,n}|$. When $\bar{\varepsilon} > 0$, using the bounded difference inequality McDiarmid (1989), we have

$$\begin{aligned}
P(\hat{\alpha}_{a,n} - \alpha_a \geq \varepsilon) &\leq P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] \geq \bar{\varepsilon}\right) \leq \exp\left(-2n\bar{\varepsilon}^2 C^w_{a,n}\right) \\
P(\hat{\alpha}_{a,n} - \alpha_a \leq -\varepsilon) &\leq P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] \leq -\bar{\varepsilon}\right) \leq \exp\left(-2n\bar{\varepsilon}^2 C^w_{a,n}\right)
\end{aligned}$$

Set $\varepsilon = \delta_{a,\tau,n} = \sqrt{\frac{2\log\tau}{nC^w_{a,n}}}$, and thus $\bar{\varepsilon} = \sqrt{\frac{2\log\tau}{nC^w_{a,n}}} - |b_{a,n}|$. When $\bar{\varepsilon} > 0$, the above two inequalities can be rewritten as

$$P\left(\hat{\alpha}_{a,n} - \alpha_a \geq \sqrt{\frac{2\log\tau}{nC^w_{a,n}}}\right) \leq \exp\left(-2\left(\sqrt{2\log\tau} - |b_{a,n}|\sqrt{nC^w_{a,n}}\right)^2\right) \tag{EC.1}$$

$$P\left(\hat{\alpha}_{a,n} - \alpha_a \leq -\sqrt{\frac{2\log\tau}{nC^w_{a,n}}}\right) \leq \exp\left(-2\left(\sqrt{2\log\tau} - |b_{a,n}|\sqrt{nC^w_{a,n}}\right)^2\right) \tag{EC.2}$$

If a sub-optimal arm a is pulled at time τ, i.e. $\sigma_\tau = a$, we know that $B_{a,\tau,T_a(\tau-1)} \geq B_{a^*,\tau,T_{a^*}(\tau-1)}$, where a^* denotes the arm with maximum aptitude. This indicates either $B_{a,\tau,T_a(\tau-1)}$ is at least α_a or $B_{a^*,\tau,T_{a^*}(\tau-1)}$ underestimates α_{a^*} (or both), i.e. either $B_{a,\tau,T_a(\tau-1)} \geq \alpha_a$ or $B_{a^*,\tau,T_{a^*}(\tau-1)} \leq \alpha_{a^*}$ (or both). If arm a has been chosen at least $u_{a,t}$ times prior to this time, i.e. $T_a(\tau-1) \geq u_{a,t} = \frac{8\log\tau}{C^w_a \Delta^2_a}$, then $\Delta_a \geq 2\delta_{a,\tau,T_a(\tau-1)}$, which implies, if $B_{a,\tau,T_a(\tau-1)} \geq \alpha_a$, then $\hat{\alpha}_{a,\tau} - \delta_{a,\tau,T_a(\tau-1)} \geq \alpha_a$, i.e. even the 'lower bound' of arm a overestimates α_a. Therefore, if $\sigma_\tau = a$ and $T_a(\tau-1) \geq u_{a,t}$ for some τ, at least one of the following two inequalities holds

$$\hat{\alpha}_{a,T_a(\tau-1)} - \delta_{a,\tau,T_a(\tau-1)} \geq \alpha_a$$
$$\hat{\alpha}_{a^*,T_{a^*}(\tau-1)} + \delta_{a^*,\tau,T_{a^*}(\tau-1)} \leq \alpha_{a^*}$$

Now, by definition and the above results, the following inequalities hold for any real number $u > 1$

$$\begin{aligned}
T_a(t) &\leq u + \sum_{\tau=\lfloor u \rfloor+1}^{t} \mathbb{1}\left\{\sigma_\tau = a \wedge T_a(\tau-1) \geq u\right\} \\
&\leq u + \sum_{\tau=\lfloor u \rfloor+1}^{t} \mathbb{1}\left\{B_{a,\tau,T_a(\tau-1)} \geq B_{a^*,\tau,T_{a^*}(\tau-1)} \wedge T_a(\tau-1) \geq u\right\} \\
&\leq u + \sum_{\tau=\lfloor u \rfloor+1}^{t} \mathbb{1}\left\{\exists v \in \{\lfloor u \rfloor, \ldots, \tau-1\}, v^* \in \{1, \ldots, \tau-1\} : B_{a,\tau,v} \geq B_{a^*,\tau,v^*}\right\} \\
&\leq u + \sum_{\tau=\lfloor u \rfloor+1}^{t} \sum_{v=\lfloor u \rfloor}^{\tau-1} \sum_{v^*=1}^{\tau-1} \mathbb{1}\left\{B_{a,\tau,v} \geq B_{a^*,\tau,v^*}\right\} \\
&\leq u + \sum_{\tau=\lfloor u \rfloor+1}^{t} \sum_{v=\lfloor u \rfloor}^{\tau-1} \sum_{v^*=1}^{\tau-1} \mathbb{1}\left\{\hat{\alpha}_{a,v} - \delta_{a,\tau,v} \geq \alpha_a \vee \hat{\alpha}_{a^*,v^*} + \delta_{a^*,\tau,v^*} \leq \alpha_{a^*}\right\} \\
&\leq u + \sum_{\tau=\lfloor u \rfloor+1}^{t} \sum_{v=\lfloor u \rfloor}^{\tau-1} \sum_{v^*=1}^{\tau-1} \left(\mathbb{1}\left\{\hat{\alpha}_{a,v} - \delta_{a,\tau,v} \geq \alpha_a\right\} + \mathbb{1}\left\{\hat{\alpha}_{a^*,v^*} + \delta_{a^*,\tau,v^*} \leq \alpha_{a^*}\right\}\right)
\end{aligned}$$

Set $u = u_{a,t}$, take the expectation on both side and we have

$$\begin{aligned}
\mathbb{E}[T_a(t)] &\leq u_{a,t} + \sum_{\tau=\lfloor u_{a,t} \rfloor+1}^{t} \sum_{v=\lfloor u_{a,t} \rfloor}^{\tau-1} \sum_{v^*=1}^{\tau-1} \left(P\left(\hat{\alpha}_{a,v} - \delta_{a,\tau,v} \geq \alpha_a\right) + P\left(\hat{\alpha}_{a^*,v^*} + \delta_{a^*,\tau,v^*} \leq \alpha_{a^*}\right)\right) \\
&\leq u_{a,t} + \sum_{\tau=\lfloor u_{a,t} \rfloor+1}^{t} \sum_{v=\lfloor u_{a,t} \rfloor}^{\tau-1} \sum_{v^*=1}^{\tau-1} \left(\exp\left(-2\left(\sqrt{2\log\tau} - |b_{a,v}|\sqrt{vC^w_{a,v}}\right)^2\right)\right. \\
&\qquad \left. + \exp\left(-2\left(\sqrt{2\log\tau} - |b_{a^*,v^*}|\sqrt{v^*C^w_{a^*,v^*}}\right)^2\right)\right) \\
&\leq u_{a,t} + \sum_{\tau=\lfloor u_{a,t} \rfloor+1}^{t} \sum_{v=\lfloor u_{a,t} \rfloor}^{\tau-1} \sum_{v^*=1}^{\tau-1} 2\exp\left(-2\left(\frac{9}{10}\sqrt{2\log\tau}\right)^2\right) \\
&\leq u_{a,t} + \sum_{\tau=\lfloor u_{a,t} \rfloor+1}^{t} 2\tau^2 \exp\left(-\frac{324}{100}\log\tau\right) \\
&\leq u_{a,t} + 2\sum_{\tau=1}^{+\infty} \tau^{-\frac{124}{100}} \\
&= u_{a,t} + 2\zeta(1.24)
\end{aligned}$$

The third inequality holds because $b_{a^*,v^*} \leq \frac{1}{10}\sqrt{\frac{2\log\tau}{v^* C^{\omega}_{a^*,v^*}}}$.

Once we have the bounds of $\mathbb{E}[T_a(t-1)]$, we can directly derive the bounds for total regret. Let $\bar{r}_a := \sup_s r_{a,s}$ and $\underline{r}_a := \inf_s r_{a,s}$, then

$$\mathbb{E}[R(t)] \leq \sum_{a\neq a^*} \mathbb{E}[(\bar{r}_{a^*} - \underline{r}_a) T_a(t-1)] \leq \sum_{a\neq a^*} (\bar{r}_{a^*} - \underline{r}_a)\left(u_{a,t} + 2\zeta(1.24)\right) \tag{EC.3}$$

EC.1.2. Proof of Proposition 1 (Upper and lower bounds of the infimum of the per-coordinate difference bound)

Proof Let $\omega_i^* = \inf w_i$. By definition, we know $\omega_i^* \leq 1$, as the image set of φ is $[0,1]$, thus $C_n^* = \frac{1}{n\sum_{i=1}^n \omega_i^{*2}} \geq \frac{1}{n\sum_{i=1}^n 1^2} = \frac{1}{n^2}$, proving the left inequality. Before we proceed to prove the right inequality, we briefly introduce Chebyshev's sum inequality (Hardy et al. 1952):

Lemma EC.1 (Chebyshev's sum inequality). *Suppose* $c_1,\ldots,c_n,b_1,\ldots,b_n \in \mathbb{R}$ *such that* $c_1 \geq c_2 \geq \ldots c_n$ *and* $b_1 \geq b_2 \geq \ldots b_n$*, and then* $\frac{1}{n}\sum_{i=1}^n c_i b_i \geq \left(\frac{1}{n}\sum_{i=1}^n c_i\right)\left(\frac{1}{n}\sum_{i=1}^n b_i\right)$.

By Chebyshev's sum inequality, $\sum_{i=1}^n w_i^* \leq \sqrt{n\sum_{i=1}^n w_i^{*2}} = \sqrt{\frac{1}{C_n^*}}$. Suppose by contradiction that $C_n^* > 1$, that is $\sum_{i=1}^n w_i^* \leq \sqrt{n\sum_{i=1}^n w_i^{*2}} = \sqrt{\frac{1}{C_n^*}} < 1$ Using Chebyshev's sum inequality, for any two points $x, x' \in \mathcal{X}^n$, $|\varphi(x) - \varphi(x')| \leq \sum_{i=1}^n w_i^* \leq \sqrt{\frac{1}{C_n^*}} < 1$. This indicates that the image set of φ has a length at most C^* that is strictly less than 1, which contradicts the assumption that φ has an image set of length 1. Thus, $C_n^* \leq 1$, the right inequality holds.

EC.2. More on Bias Conditions in Example 3

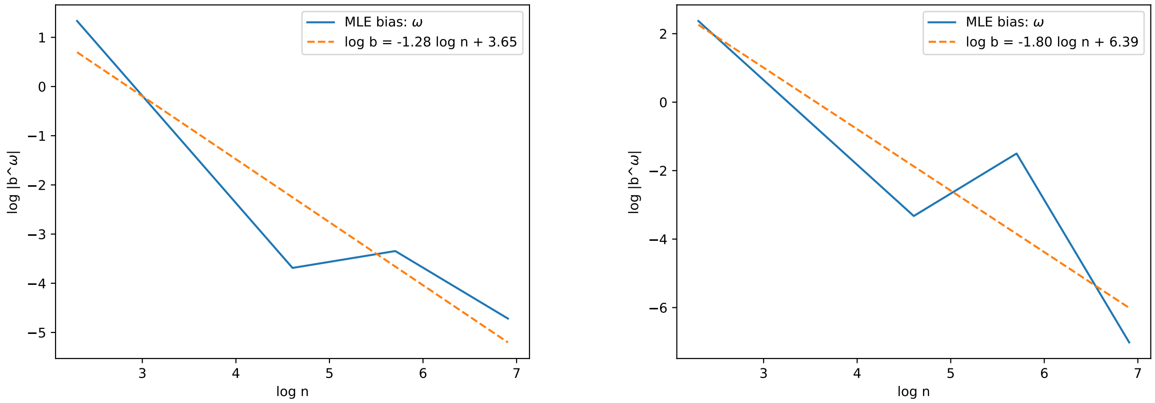
Figure EC.1 shows the bias decay rates of $\omega_{1,n}^{MLE}$ and $\omega_{2,n}^{MLE}$. We might be interested in $\omega_{1,n}^{MLE}$ and $\omega_{2,n}^{MLE}$'s bias decay rates for general dynamic learning problems with unknown vector parameters. For Theorem 1 to hold in the SLT problem that focuses on the long-term, full potentials of arms, we only need to verify the bias conditions for $\alpha_{1,n}^{MLE}$ and $\alpha_{2,n}^{MLE}$.

EC.3. Proof of Bandits with Delayed Feedback in Section 6.1

Let $\hat{\hat{\alpha}}_{a,n}$ denote our point estimate of α_a when $T_a(t) = n$ and up to k_a true rewards have not been revealed but reward estimates are available.

Proof: Let $a \in \mathcal{A}$, $\tau \in \mathcal{T}$, and $n := T_a(\tau - 1)$. Below we derive probabilistic bounds for $\hat{\hat{\alpha}}_{a,n}$,

$$\begin{aligned}
P\left(\hat{\hat{\alpha}}_{a,n} - \alpha_a \geq \varepsilon\right) &= P\left(\hat{\hat{\alpha}}_{a,n} - \hat{\alpha}_{a,n} + \hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] + \mathbb{E}[\hat{\alpha}_{a,n}] - \alpha_a \geq \varepsilon\right)\\
&\leq P\left(|\hat{\hat{\alpha}}_{a,n} - \hat{\alpha}_{a,n}| + (\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}]) + |\mathbb{E}[\hat{\alpha}_{a,n}] - \alpha_a| \geq \varepsilon\right)\\
&= P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] \geq \varepsilon - |b_{a,n}| - |e_{a,n}|\right)\\
P\left(\hat{\alpha}_{a,n} - \alpha_a \leq -\varepsilon\right) &= P\left(\hat{\hat{\alpha}}_{a,n} - \hat{\alpha}_{a,n} + \hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] + \mathbb{E}[\hat{\alpha}_{a,n}] - \alpha_a \leq -\varepsilon\right)\\
&\leq P\left(-|\hat{\hat{\alpha}}_{a,n} - \hat{\alpha}_{a,n}| + (\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}]) - |\mathbb{E}[\hat{\alpha}_{a,n}] - \alpha_a| \leq -\varepsilon\right)\\
&= P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] \leq -\varepsilon + |b_{a,n}| + |e_{a,n}|\right)
\end{aligned}$$



(a) The bias of $\omega_{1,n}^{MLE}$ is $O(n^{-1.28})$ (b) The bias of $\omega_{2,n}^{MLE}$ is $O(n^{-1.80})$

Figure EC.1 Verifying bias scales of $\omega_{1,n}^{MLE}$ and $\omega_{2,n}^{MLE}$. The bias scales are both $o\left(\sqrt{\frac{\log n}{n}}\right)$; although not needed, we can see that the bias decay rates of $\omega_{1,n}^{MLE}$ and $\omega_{2,n}^{MLE}$ satisfy the bias condition in Theorem 1.

Let $\bar{\varepsilon} := \varepsilon - |b_{a,n}| - |e_{a,n}|$. When $\bar{\varepsilon} > 0$, using the bounded difference inequality McDiarmid (1989), we have

$$P(\hat{\alpha}_{a,n} - \alpha_a \geq \varepsilon) \leq P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] \geq \bar{\varepsilon}\right) \leq \exp\left(-2n\bar{\varepsilon}^2 C_{a,n}^w\right)$$
$$P(\hat{\alpha}_{a,n} - \alpha_a \leq -\varepsilon) \leq P\left(\hat{\alpha}_{a,n} - \mathbb{E}[\hat{\alpha}_{a,n}] \leq -\bar{\varepsilon}\right) \leq \exp\left(-2n\bar{\varepsilon}^2 C_{a,n}^w\right)$$

Set $\varepsilon = \delta_{a,\tau,n} = \sqrt{\frac{2\log\tau}{nC_{a,n}^w}}$, and thus $\bar{\varepsilon} = \sqrt{\frac{2\log\tau}{nC_{a,n}^w}} - |b_{a,n}| - |e_{a,n}|$. When $\bar{\varepsilon} > 0$, the above two inequalities can be rewritten as

$$P\left(\hat{\alpha}_{a,n} - \alpha_a \geq \sqrt{\frac{2\log\tau}{nC_{a,n}^w}}\right) \leq \exp\left(-2\left(\sqrt{2\log\tau} - |b_{a,n}|\sqrt{nC_{a,n}^w} - |e_{a,n}|\sqrt{nC_{a,n}^w}\right)^2\right) \quad \text{(EC.4)}$$
$$P\left(\hat{\alpha}_{a,n} - \alpha_a \leq -\sqrt{\frac{2\log\tau}{nC_{a,n}^w}}\right) \leq \exp\left(-2\left(\sqrt{2\log\tau} - |b_{a,n}|\sqrt{nC_{a,n}^w} - |e_{a,n}|\sqrt{nC_{a,n}^w}\right)^2\right) \quad \text{(EC.5)}$$

The rest of the proof follows that of Theorem 1 in Section EC.1.1. The only minor changes are needed after we set $u = u_{a,t}$ and take the expectation on both side; we have

$$\begin{aligned}
\mathbb{E}[T_a(t)] &\leq u_{a,t} + \sum_{\tau=\lfloor u_{a,t}\rfloor+1}^{t} \sum_{v=\lfloor u_{a,t}\rfloor}^{\tau-1} \sum_{v^*=1}^{\tau-1} \left(P\left(\hat{\alpha}_{a,v} - \delta_{a,\tau,v} \geq \alpha_a\right) + P\left(\hat{\alpha}_{a^*,v^*} + \delta_{a^*,\tau,v^*} \leq \alpha_{a^*}\right)\right) \\
&\leq u_{a,t} + \sum_{\tau=\lfloor u_{a,t}\rfloor+1}^{t} \sum_{v=\lfloor u_{a,t}\rfloor}^{\tau-1} \sum_{v^*=1}^{\tau-1} \left(\exp\left(-2\left(\sqrt{2\log\tau} - |b_{a,v}|\sqrt{vC_{a,v}^w} - |e_{a,v}|\sqrt{vC_{a,v}^w}\right)^2\right)\right. \\
&\qquad \left.+ \exp\left(-2\left(\sqrt{2\log\tau} - |b_{a^*,v^*}|\sqrt{v^*C_{a^*,v^*}^w} - |e_{a^*,v^*}|\sqrt{v^*C_{a^*,v^*}^w}\right)^2\right)\right) \\
&\leq u_{a,t} + \sum_{\tau=\lfloor u_{a,t}\rfloor+1}^{t} \sum_{v=\lfloor u_{a,t}\rfloor}^{\tau-1} \sum_{v^*=1}^{\tau-1} 2\exp\left(-2\left(\left(1 - \frac{1}{10} - \frac{1}{40}\right)\sqrt{2\log\tau}\right)^2\right)
\end{aligned}$$

$$\begin{aligned}
&\leq u_{a,t} + \sum_{\tau=\lfloor u_{a,t}\rfloor+1}^{t} 2\tau^2 \exp\left(-\frac{1225}{400}\log\tau\right) \\
&\leq u_{a,t} + 2\sum_{\tau=1}^{+\infty} \tau^{-\frac{425}{400}} \\
&= u_{a,t} + 2\zeta(1.063)
\end{aligned}$$

The third inequality holds because $b_{a,v} \leq \frac{1}{10}\sqrt{\frac{2\log v}{vC^{\omega}_{a,v}}} \leq \frac{1}{10}\sqrt{\frac{2\log\tau}{vC^{\omega}_{a,v}}}$ and $b_{a^*,v^*} \leq \frac{1}{10}\sqrt{\frac{2\log v^*}{v^*C^{\omega}_{a^*,v^*}}} \leq \frac{1}{10}\sqrt{\frac{2\log\tau}{v^*C^{\omega}_{a^*,v^*}}}$. Similarly, $e_{a,v} \leq \frac{1}{40}\sqrt{\frac{2\log v}{vC^{\omega}_{a,v}}} \leq \frac{1}{40}\sqrt{\frac{2\log\tau}{vC^{\omega}_{a,v}}}$ and $e_{a^*,v^*} \leq \frac{1}{40}\sqrt{\frac{2\log v^*}{v^*C^{\omega}_{a^*,v^*}}} \leq \frac{1}{40}\sqrt{\frac{2\log\tau}{v^*C^{\omega}_{a^*,v^*}}}$.

Once we have the bounds of $\mathbb{E}[T_a(t-1)]$, we can directly derive the bounds for total regret. Let $\bar{r}_a := \sup_s r_{a,s}$ and $\underline{r}_a := \inf_s r_{a,s}$, then

$$\mathbb{E}[R(t)] \leq \sum_{a\neq a^*} \mathbb{E}[(\bar{r}_{a^*} - \underline{r}_a)T_a(t-1)] \leq \sum_{a\neq a^*} (\bar{r}_{a^*} - \underline{r}_a)\left(u_{a,t} + 2\zeta(1.063)\right) \tag{EC.6}$$

EC.4. Proof of Theorem 2: FL-UCB regret bounds

Proof of FL-UCB regret bounds: First, we consider the LP defined by (11) $\sim$ (15). For the sake of notational simplicity and generality, we write it in the standard form

$$\max_{z} \quad f(z) \tag{EC.7}$$

$$s.t. \quad z \in C_{\text{set}} \tag{EC.8}$$

where $C_{\text{set}} \in \mathbb{R}^{|\mathcal{A}|}$ is a nonempty convex set and $f : \mathbb{R}^{|\mathcal{A}|} \mapsto \mathbb{R}^{|\mathcal{A}|}$ is a convex function. A point z^* is optimal for the convex optimization problem (EC.7) $\sim$ (EC.8) if

$$\exists \xi \in \mathbb{R}^{|\mathcal{A}|}\backslash\{\mathbf{0}\} \qquad s.t. \quad \xi(z - z^*) \leq 0, \quad \forall z \in C_{\text{set}} \tag{EC.9}$$

Before we further analyze the optimality criterion (EC.9), we define the concept of *normal cones.*

Definition EC.1. *The normal cone of a closed, convex set $C_{set} \in \mathbb{R}^n$ is*

$$N_C(z^*) = \begin{cases} \{\xi \in \mathbb{R}^n | (\forall z \in C_{set})\xi^T(z - z^*) \leq 0\} & \textit{if } z^* \in C_{set} \\ \emptyset & \textit{if } z^* \notin C_{set} \end{cases} \tag{EC.10}$$

(EC.9) is equivalent to requiring that $\xi \in N_C(z^*)\backslash\{\mathbf{0}\}$. To find the normal cone of the feasible region defined in (12) $\sim$ (15), we need to use the following lemma

Lemma EC.2. *Let $A \in \mathbb{R}^{m\times n}$ and let $b \in \mathbb{R}^m$. Consider the polyhedron $Q(A,b) = \{x \mid Ax \leq b\}$. Suppose $x \in Q(A,b)$, then $N_{Q(A,b)}(x) = \{A^Ty | y \in \mathbb{R}^m$ such that $y \geq 0$ and $y^T(b - Ax) = 0\}$.*

The optimal solution to (11) $\sim$ (15) is:

$$z^*_a = \begin{cases} \theta^A_a, & a \in \mathcal{A}_A \setminus \mathcal{A}_{BK} \\ \theta^{BK}_a, & a \in \mathcal{A}_{BK} \setminus \mathcal{A}_A \setminus \{a^*\} \\ \max\{\theta^{BK}_a, \theta^A_a\} & a \in \mathcal{A}_{BK} \cap \mathcal{A}_A \setminus \{a^*\} \\ 0, & a \notin \mathcal{A}_{BK} \cup \mathcal{A}_A \\ 1 - \sum_{a\in\mathcal{A}\setminus\{a^*\}} z^*_a, & a = a^* \end{cases} \tag{EC.11}$$

Therefore, the normal cone at an optimal solution z^* for the convex set Q_{FUCB} as defined by (12) $\sim$ (15) is a convex cone defined by the following inequalities, assuming $\mathcal{A}_{BK}$ is known (we will estimate it later):

$$B_{a^*,T_{a^*}(t-1)} \geq B_{a,T_a(t-1)} \qquad \forall a \in \mathcal{A} \tag{EC.12}$$

$$B_{a,T_a(t-1)} \geq 0 \qquad \forall a \in \mathcal{A} \tag{EC.13}$$

If we replace α with $B_{s_{t-1}} := B_{\mathcal{A},T_{\mathcal{A}}(t-1),s_{\mathcal{A},t-1}} = [B_{a,T_a(t-1),s_{a,t-1}}]_{a=1}^{|\mathcal{A}|}$, as long as $B_{s_{t-1}} \in N_C(z^*)$, the optimal basis stays optimal for the new LP problem with the objective defined by (16). (EC.12) $\sim$ (EC.13) show that as long as we are able to identify a^* (the unique solution of the new LP is the optimal) and other top-K arns using the UCB indexes soon enough and only explore the other arms rarely afterwards.

Moreover, we need to bound the regret incurred while estimating the members of $\mathcal{A}_{BK}$ and the ordering. Specifically, we want to distinguish the difference between the k-th best arm $a^{(k)}$ and the $k+i$-th best arm $a^{(k+i)}$, $\forall i \in \{1,\ldots,|\mathcal{A}|-k\}$. The proof of regret bound in distinguishing the k-th and the $(k+i)$-th best arm is analogous to that of proving L-UCB regret upper bounds: The difference is that we are not only interested in a^* or $a^{(1)}$, but also $a^{(k)}$ for $k \in \{2,\ldots,K\}$. Specifically, to compute the expected number of pulls of $a^{(k+i)}$ when we actually want to pull $a^{(k)}$, we choose

$$u_{a^{(k+i)}}^{a^{(k)}} = \bar{T} := \frac{8\log t}{C_a^w \Delta^2_{a^{(k)},a^{(k+i)}}}$$

where $\Delta_{a^{(i)},a^{(j)}} = \alpha_{(i)} - \alpha_{(j)}$, $\forall i,j \in \{1,\ldots,|\mathcal{A}|\}$. The number of times that an arm $a^{(k+i)}$ is mistaken in the $\mathcal{A}_{BK}$ as $a^{(k)}$ set is bounded by

$$\frac{8\log t}{C_a^w \Delta^2_{a^{(k)},a^{(k+i)}}} + 2\zeta(1.24)$$

Therefore, the expected number of times we pull "worse" arms (whose true parameters are worse than those of the ones we intend to pull) when imposing BK-fairness, is bounded by

$$\sum_{k=1}^{K} \sum_{i=1}^{|\mathcal{A}|-k} \left(\frac{8\log t}{C_a^w \Delta^2_{a^{(k)},a^{(k+i)}}} + 2\zeta(1.24) \right)$$

And thus the expected regret when imposing BK-fairness is bounded by

$$\begin{aligned}
\mathbb{E}[R^{BK}(t)] &= \sum_t r_{a^*,T_{a^*},T_{a^*}(t-1)} - r_{a_t,T_a(t-1)} \leq \sum_a (\bar{r}_{a^*} - \underline{r}_a)\, \mathbb{E}[T_a] \\
&\leq \sum_{k=1}^{K} \sum_{i=1}^{|\mathcal{A}|-k} (\bar{r}_{a^*} - \underline{r}_{a^{(i)}}) \left(\frac{8\log t}{C_a^w \Delta^2_{a^{(k)},a^{(k+i)}}} + 2\zeta(1.24) \right)
\end{aligned}$$

Note that imposing BK- or AA-fairness incurs linear PoF as long as $\theta^A + \theta^{BK} \neq \mathbf{0}$, which is not counted as part of *F-regret* or regret. Moreover, $\mathcal{A}_A$ is assumed to be known based on inherent, known arm features and thus does not require estimation.

EC.5. Extension: Arm Correlation

In this section we study bandit problems where the learning processes of arms are correlated. Specifically, we study bandits where arm experience is correlated in a linear fashion: Linear correlation is among the most common dependence patterns in the literature, and its mathematical simplicity enables us to derive clean analytical results that shed light on the influence of arm dependence on bandits. We have proved that the regret bounds of bandits with mutually independent learning arms are $O(\log t)$. We will show that similar results hold when arms are correlated.

In our SLT problem setting, the arms are patient-TC-surgery tuples. Arm correlation may arise from the fact that the skill sets required to perform successful SLT surgeries of various types typically overlap. This translates to a bandit problem where an arm's hidden parameter might change along with its the learning curve, even if that particular arm is not chosen.

EC.5.1. Experience-Correlated Bandits

As discussed in Section 6, the skills learned from different surgeries could be partially transferable as the skill sets required for similar surgeries may overlap. We consider linear correlation based on experience in bandit contexts; we explicitly define bandits with this particular form of arm dependence.

Definition EC.2. *(Experience-Correlated Bandit) A bandit problem is experience-correlated if the experience score $s_{a,t-1}$ of an arm $a \in \mathcal{A}$ can be written as*

$$s_a(t) = s_a(t-1) + \sum_{j \in \mathcal{A}^a} \beta_{a,j,t} \mathbb{1}(a_t = j) \qquad \beta_{a,j,t} \geq 0, \forall t \geq 1, \tag{EC.14}$$

where $\mathcal{A}^a \neq \emptyset$ is the set of arms that are correlated with a.

When arms are uncorrelated/independent, $\mathcal{A}^a = \{a\}$ and $\beta_{a,a,t} = 1, \forall t$. When $\exists j \neq a, j \in \mathcal{A}^a$, s.t. $\beta_{a,j,t} > 0$, we say arm a is dependent on arm j. In this case, $s_a(t) \geq T_a(t)$, with this inequality being strict for at least one a: $T_a(t)$ is the number of times that an arm has been pulled (affecting the total regret), while $s_a(t)$ affects the current proficiency parameter, $\theta_a(\alpha_a, s_{a,t})$. Here, correlation affects learning. And we demonstrate through the examples below that such problems can be challenging, in general.

Unlike the vanilla bandit problem, the optimal policy of an experience-correlated bandit is not a straightforward stationary policy, i.e., always pulling the arm with the highest aptitude α^* may turn out to be a sub-optimal strategy in both large-t the long-term regime and small-t the short-term regime. Consider the following example:

Example EC.1. *Consider an experience-correlated bandit with two arms: arm 1 and arm 2. The learning curves and correlations are explicitly known:*

$$\theta_{1,t} := l_1(s_1(t)) = 0.5 + \min\left\{\frac{0.5 s_1}{100}, 0.5\right\} \quad t \geq 1 \tag{EC.15}$$

$$\theta_{2,t} := l_2(s_2(t)) = \min\left\{\frac{s_2}{100}, 0.9\right\} \quad t \geq 1 \tag{EC.16}$$

$$s_1(t) = s_1(t-1) + \mathbb{1}(a_t = 1) + 100 \cdot \mathbb{1}(a_t = 2) \quad t \geq 1 \tag{EC.17}$$

$$s_2(t) = s_2(t-1) + \mathbb{1}(a_t = 2) \quad t \geq 1 \tag{EC.18}$$

where $s_1(0) = s_2(0) = 0$.

The optimal policy is to pull arm 2 at $t = 1$, *and choose arm 1 when* $t \geq 2$.

Proof: Under π^*, the total expected reward is $\mathbb{E}\left[\sum_t r_{a_t, s_{a_t,t-1}}\right] = 0 + 1 + \cdots + 1 = T - 1$.

First, we show that if arm 2 has been pulled once, we should always pull arm 1 in later rounds. Because pulling arm 2 once will guarantee that $s_1(t) \geq 100$ and $\theta_1(t) = 1$; thus, the expected reward of pulling arm 1 in a later time will yield the highest possible expected single period reward ($=1$), while pulling arm 2 will give no more than 0.9 expected reward. As a result, in an optimal policy, once arm 2 is pulled, arm 1 should always be chosen in later rounds.

Now, we show that we will pull arm 2 at least once. If we never pull arm 2, then we always pull arm 1; the expected total reward under this policy is $0.5 + 0.505 + 0.51 + 0.515 + \cdots + 0.995 + 1 * \times (T - 99) < T - 1$. Therefore, π^* has higher expected total rewards compared to the policy that pulls arm 1 throughout the time horizon.

Finally, we show that we will pull arm 2 precisely at $t = 1$. If we follow a policy π' that first pulls arm 2 at $t = k, k \in \{2, \cdots, t\}$, then the expected reward at $t = 1$, $\mathbb{E}r_{1,0}(1) = 0.5$. The expected total reward of the policy π', $\mathbb{E}\left[\sum_t r^{\pi'}(t)\right] \leq r^{\pi'}(1) + r^{\pi'}(k) + 1 \times (T-2) = 0.5 + 0 + T - 2 = T - 1.5 < T - 1$. Therefore, π^* has higher expected total reward compared to any policy that dictates pulling arm 2 at time $t = 1$.

The arguments above show that π^* is the optimal policy.

In many applications the solution to the offline experience-based bandits can be found using dynamic programming. The following example shows that the optimal policy for our problem may require switching arms more than once and revisiting an arm.

Example EC.2. *Consider an experience-correlated bandit with two arms, arm* 1 *and arm* 2. *The learning curves and correlations are explicitly known:*

$$\theta_1(t) := l_1(s_1(t-1)) = 0.5 + \min\{\frac{0.5 s_1}{100}, 0.5\} \quad t \geq 1 \tag{EC.19}$$

$$\theta_2(t) := l_2(s_2(t-1)) = \min\{\frac{s_2}{100}, 0.9\} \quad t \geq 1 \tag{EC.20}$$

$$s_1(t) = s_1(t-1) + \mathbb{1}(a_t = 1) + 100 \cdot \mathbb{1}(a_t = 2) \quad t \geq 1 \tag{EC.21}$$

$$s_2(t) = s_2(t-1) + 100 \cdot \mathbb{1}(a_t = 1) + \mathbb{1}(a_t = 2) \quad t \geq 1 \tag{EC.22}$$

where $s_1(0) = s_2(0) = 0$.

The optimal policy is to choose arm 1 at $t = 1$, *choose arm 2 at* $t = 2$, *and then choose arm 1 when* $t \geq 3$.

Proof: First, we prove that in the optimal policy, we pull arm 1 and 2 each at least once. The expected reward of always choosing arm 1 is $0.5+0.505+\cdots+0.995+1\times(T-100)<T-1$; similarly, the expected reward of always choosing arm 2 is $0+0.01+\cdots+0.89+0.9\times(T-90)<T-1$. However, the expected total reward of π^* is $0.5+0.9+1\times(T-2)=T-0.6>T-1$. Therefore, both stationary policies cannot be optimal.

Next, we show that we pull arm 2 at most once. Now we know that both arms are chosen at least once in the optimal policy. Suppose we have pulled arm 2 at time t', then pulling arm 2 at any time $t''>t'$ will not increase $\theta_1(t'')$, but will yield a lower immediate reward; thus, the marginal benefit of choosing arm 2 at t'' is strictly negative. As a result, we choose arm 2 exactly once.

Finally, we prove that we choose arm 2 at $t=2$. If we choose arm 2 at $t=1$, then the expected total reward is $0+1\times(T-1)=T-1<T-0.6$, thus choosing arm 2 at $t=2$ is better than pulling arm 2 at $t=1$. If a policy π' pulls arm 2 at $t=k, k\in\{3,\cdots,T\}$, then the expected total reward $\mathbb{E}[\sum_t r^{\pi'}(t)]<r^{\pi'}(1)+r^{\pi'}(2)+r^{\pi'}(k)+1\times(T-3)\leq 0.5+0.505+0.9+T-3=T-1.05<T-0.6$. In summary, the optimal policy is to pull arm 2 at $t=2$.

Example EC.2 shows that in the optimal policy, a low immediate-reward, high contributed-experience arm (arm 2) may be chosen after higher-reward, low experience arms in an optimal strategy. An explanation for the fluidity and complexities is that the hidden parameters of arms may change when any arms are pulled, and depending on the specific structure of learning curves and correlation patterns, an arm that is useless at one time may be incredibly useful in later rounds. This being said, imposing conditions on the correlation would potentially yield structural results on the optimal policy. Such an exploration is deferred for future work.

EC.5.2. Heterogeneous livers

We can incorporate liver heterogeneity in two ways. The most straightforward approach is to formulate parallel MABs, one for each liver type; within each MAB, livers are viewed as homogeneous. This formulation is practical and realistic, as livers are often allocated to TCs and patients within the geographical region in which they are acquired (see Section EC.7 for more information). Surgical experience and expertise at one TC rarely transfer to another that is geographically remote.

Alternatively, we can incorporate liver heterogeneity in one MAB, i.e., $|\mathcal{L}|>1$, implying that for each $\ell\in\mathcal{L}$, all arms in $\mathcal{A}$ can be pulled, i.e., a liver of type ℓ can be allocated to any TC and any patient type. This formulation can be helpful when we are interested in a more granular classification of liver types within the same geographical region, as experience gained from operating with different liver types is carried forward. Our FL-UCB algorithms apply to the case $|\mathcal{L}|>1$, except that we estimate $\hat{\alpha}^{\ell}_{a,n}$ for each $\ell\in\mathcal{L}$. All theoretical regret bounds hold (the upper bounds for heterogeneous livers are $|\mathcal{L}|$ times the original bounds for the homogeneous case). The actual regrets might be much lower, as surgical experience transfers and accumulates faster.

EC.6. More Details about the Numerical Study in Section 7

EC.6.1. Details about the SLT simulation setup

Below we detail how we estimate α's from the STAR files. For each medically-splittable liver, it can save two patients' lives. In current SLT practice, the smaller left lobe is usually allocated to a sick child. The other half, depending on its size and the patient waitlists, may be allocated to a small adult/big child or a medium adult. There is a liver-splitting technique that allows a more even splitting of a donor's liver and thus can save two small or medium adults' lives. The two partial livers can be used for two recipients at two different transplant centers; thus, we view each partial liver arrival as an independent time step. A partial liver may be shared across a large geographical area; see UNOS (2020a) for detail about the acuity circles policy.

Currently, a splittable liver may be shared across a large geographical area; see the acuity circles policy UNOS (2020a) for detail. We consider a 500 nautical mile circle that includes OPTN regions 2, 9, 10, 11, and Wisconsin and Illinois (URL: `https://optn.transplant.hrsa.gov/about/regions/`). In 2022, there were around 8000 donated livers and 10 big transplant centers in the 500NM Circle. (See `https://optn.transplant.hrsa.gov/data/view-data-reports/regional-data/` for more detail.)

Each (partial) liver graft can be allocated to a patient within one of the five health condition groups. Patients' health conditions are described by the Model for End-Stage Liver Disease (MELD) score (for adults) and Pediatric End-Stage Liver Disease (PELD) score (for children), which are indicators of medical urgency. MELD and PELD scores take integer values in $[6, 40]$; for critically sick patients, there are 1A, 1B, 2A, and 2B special urgent categories. We divide the patients into five score buckets: ≥ 40 (including MELD/PELD $= 40$ and critically sick patients), $35 \sim 39$, $30 \sim 34$, $20 \sim 29$, $6 \sim 19$. The current OPTN system allocates (whole) livers preferentially to eligible patients with the highest scores (the sickest patients) (Emre and Umman 2011); SLT surgeries are rarely performed, but the current SLT patient matching does not strictly follow the "sickest-first" rule, due to lack of policy clarity in matching the secondary recipient, and the primary recipient is often a child. Since SLT is a challenging medical procedure and saves twice as many lives, it makes sense to consider allocating partial livers to healthier patients to maximize overall survival and welfare.

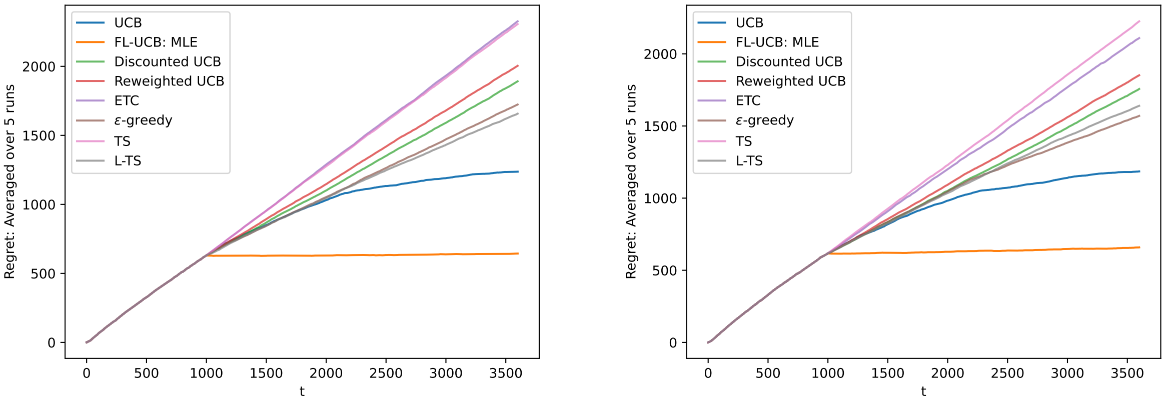
Therefore, in total, we have $10 \times 5 = 50$ arms for the livers splittable in the geographical region of interest. Recall that 10% of livers are medically safe to split (OPTN and UNOS 2016), so at least 800 livers can be used for SLT a year in the 500NM Circle, with each liver supporting two SLT surgeries. A total of 1600 SLT surgeries are possible. Livers are heterogeneous; among the medically safe livers, it is estimated that $\sim 63\%$ (Perito et al. 2019a), or around 1008 of them, satisfy the strictest medical criteria and thus are of the highest quality. In our simulation, we consider allocating these high-quality livers to patients and TCs in the 500NM geographical circle.

See Section EC.6 for more details about the allocation of high-quality livers acquired in different geographical regions (i.e., heterogeneity) and please refer to Section EC.7 for more facts about current SLT practice in the US.

The α 's are drawn from $(0.3, 0.95)$, where the upper and lower bounds of the range are estimated directly from the STAR files: We compute the 1-year graft survival for different surgery technique types in each geographical region; these statistics are then used for simulate the distribution and range of SLT's 1-year survival outcomes. These statistics of past surgeries (WLTs and a small number of SLTs) show that 1-year graft survival range from 0.33 to 1. Retrospective reviews and anecdotal accounts report that SLT outcomes can be comparable and as good as WLT outcomes in few, proficient TCs that have gained SLT mastery through a good amount of experience (Hackl et al. 2018, Duke 2021). Since SLT is a more complex surgery by nature, we adjust the lower limit of the 1-year graft survival rate to 0.3.

EC.6.2. Outcome prediction accuracy and uncertainty quantification

In Section 7 we assume the prediction accuracy in SLT is 60%; Figure EC.2 shows results assuming the prediction accuracy is 85%.



(a) Regret w/o fairness

(b) F-regret w/ fairness

Figure EC.2 Comparing FL-UCB regret against benchmarks when medical learning exists and assuming there is a 1-year delay in observing true rewards (the rollout policy is described in Section 7.1). Estimates based on demographics and perioperative clinical metrics are available and are 85% accurate.

Similar to the case where the prediction accuracy is 60%, FL-UCB with MLE estimation has the lowest regrets and converges fast when an 85%-accurate estimate is available. However, with a higher accuracy level, the UCB performance is significantly improved and is second only to FL-UCB; its regrets also show signs of convergence at $t = 3600$.

EC.6.3. More about the bandit algorithms used for comparison

Algorithm 3: L-TS Algorithm Pseudo Code

```
1: **Initialization:** Choose prior distributions $Beta(\tilde{\alpha}_{a,0}, \tilde{\beta}_{a,0})$, $\forall a \in \mathcal{A}$. Select each arm $a$ $m_a$ times. Update posterior distribution as in Step 3.
2: **Select arm:** Sample $a_t \sim Beta(\tilde{\alpha}_{a,n}, \tilde{\beta}_{a,n})$
3: **Update distribution:** $\tilde{\alpha}_{a_t,T_{a_t}(t-1)}, \tilde{\beta}_{a_t,T_{a_t}(t-1)}) \leftarrow$
   $\Big((\tilde{\alpha}_{a,T_a(t-1)} + r^t(1+\exp(\omega_{a_t} - T_{a,t-1})), \tilde{\beta}_{a,n-1} + (1 - r^t(1+\exp(\omega_{a_t} - T_{a,t-1}))\Big)$
4: **Increment** $t$**,** $T_{a_t,t} = T_{a_t,t-1} + 1$ and **Go to Step 2**
```

In TS, we update the posterior distribution using $(\tilde{\alpha}_{a_t,T_{a_t}(t-1)}, \tilde{\beta}_{a_t,T_{a_t}(t-1)}) \leftarrow (\tilde{\alpha}_{a,T_a(t-1)} + r^t, \tilde{\beta}_{a,n-1} + (1-r^t))$. Recall that r^t is the random reward (or the estimated reward) at time t. In our numerical study, we choose $m_a = 20$ and $(\tilde{\alpha}_{a,0}, \tilde{\beta}_{a,0}) = (2,2)$ for all a in both L-TS and TS. For ETC algorithm implemented, the exploitation starts once the 500 rounds of round robin conclude. In ϵ-greedy, with probability 0.95 we greedily choose the arm with the highest estimated reward (breaking ties arbitrarily) and we explore arms with equal probability when not exploiting. In our discounted UCB, we use $\delta = 0.9$.

EC.7. Current SLT practice in the US

To better understand how SLTs are practiced, we consulted with senior surgeons from the globally renowned transplant center affiliated with the University of California, San Francisco (UCSF). In the US, after graduating from medical schools and having chosen their specialization areas, surgeons complete their residency programs to obtain an unrestricted license to practice medicine and a board certificate for their chosen surgical specialty, in our case, the liver transplant. It is during residency that prospective surgeons may learn SLTs at selected TCs, such as the one at UCSF. Such residency programs involve assisting with actual SLT surgeries. Besides graduation requirements that enable some residents to learn SLT, young physicians may be intrinsically interested in saving more lives, expanding their skill sets, and mastering the techniques to perform complex surgeries; extrinsic motivations such as recognition from the surgical community and income increase brought by more transplants can also incentivize medical learning and overcome risk aversion.

In practice, successful SLTs involve a complicated process, including registration, procurement, allocation, logistics, surgical operations, and post-surgery recovery. To start, eligible ESLD patients choose transplant centers and register for the national liver transplant waitlists. When a deceased-donor liver becomes available and is being evaluated to determine whether it is medically splittable (based on donor age, body mass index, size, etc.), OPTN (which is administered by United Networks for Organ Sharing, short for UNOS) generates a ranked list (known as the *match-run*), based on computerized algorithms. The organ is offered to the match-run candidates sequentially until a candidate/candidate pair accepts it. The longer it takes between the removal of blood supply from the deceased-donor organ and the transplantation into the recipient(s) (the *cold ischemia time*), the more the organ's quality deteriorates. An organ is discarded if the cold ischemia time is determined to be too long (exceeding 12 - 18 hours). From our discussions with UCSF transplant physicians and transplant software professionals, we learned that transplant centers could also make provisional offers to more than one patient to reduce organ waste and maximize societal welfare. Therefore, practically speaking, UNOS/OPTN essentially assigns livers to transplant centers and certain patient health groups; at the center level, the medical teams on call perform the surgery with the assigned recipients or make adjustments under OPTN guidelines when necessary.

Once a liver is accepted, the organ is harvested (and split if to be used in SLTs) by a trained team at the donor hospital. The matching of transplant surgeons and candidates is finalized after a candidate has accepted an offer and right before the surgery. After being harvested, the procured split liver grafts are transported to the SLT recipient TCs, where the two transplant teams perform the SLT surgeries. After that, patient recoveries occur. Currently, most SLTs are performed in few major transplant centers; thus, the primary recipients (usually children) and secondary recipients are usually within the same TC. However, because of UNOS's new acuity circles policy that took effect in 2019 (UNOS 2020a) , patients from different TCs within the acuity circles may receive halves of the same donor liver more frequently. The acuity circles policy aims to enable broader organ sharing but has been controversial due to challenging logistics, incentive misalignment, and increased organ waste. There has also been debate over the transplant objective itself: Should we allocate livers to the sickest patient(s) within the 500NM circle? Is the current "sickest-first" principle simply preventing more immediate deaths but not optimizing societal welfare (e.g., survival outcomes, quality-adjust life years, equity)?

EC.8. More on Related Work

UCB variants for nonstationary environments: Garivier and Moulines (2011) consider abrupt changing environments where the reward distributions may remain constant for epochs and change at unknown breakpoints. The authors proposed D-UCB and SW-UCB policies to overcome environment nonstationarity. Similar to the idea of our discounted UCB, D-UCB averages past rewards with a discount factor which gives more weight to recent observations. The difference between our discounted UCB and D-UCB is in the padding function (i.e., the term added to our estimate of the arm parameter). Their SW-UCB relies on a local empirical average of the last few plays, leaving out earlier observations. Alternatively, we proposed reweighted UCB to discount the past observations less aggressively, as in our SLT problem the reward distribution changes gradually.

Bandits and queueing: Our MAB formulation can further be enhanced by adding a constant (computed by a separate queueing module) to each arm's full potential to capture the endogenizing queueing effects (e.g., number of patient deaths while waiting for transplants) while keeping the problem stateless. Previous attempts to incorporate queueing or non-stationarity into bandit problems have utilized stateful formulations, for instance, *restless bandits* (Bertsimas and Niño-Mora 2000, Whittle 1988, Jacko 2010, Krishnasamy et al. 2016). Such explicit modeling would likely render our SLT learning problem intractable; thus, analyzing queueing dynamics via a separate module sounds more viable: We may incorporate them into a subroutine of our proposed algorithm to maintain a stateless bandit. Compared to these previous works, the objective in the SLT problem is also markedly different: For example, Krishnasamy et al. (2016) combined MAB with queueing by designating the MAB's rewards as queue lengths; while Whittle (1988) and Bertsimas and Niño-Mora (2000) did not incorporate queueing behaviors in their analysis. In our problem, the reward functions can be written as convex combinations of immediate rewards and queueing metrics, both of which are functions of the bandit decisions.

Experience-based learning: Human learning describes how human individuals acquire and possess knowledge or skills under cognitive and environmental influences, taking into account prior experience (Lefrancois 2019, Illeris 2002, Jarvis 2006).

In the call center literature, Arlotto et al. (2014) studied the hiring and retention of heterogeneous agents who learn over time and formulated it as an infinite-armed bandit with switching costs. However, they did not explicitly model a broad class of parametric learning curves and stochastic observations, or the queueing dynamics associated with learning and scheduling. For the new franchisee problem, Darr et al. (1995) studied the transfer of knowledge acquired through learning by doing empirically—they found evidence of learning based on weekly data collected from 36 pizza stores. To our best knowledge, there has not been analytical modeling work that studies both experience-based learning and queueing dynamics.

Reference for the Appendix

Arlotto A, Chick SE, Gans N (2014) Optimal hiring and retention policies for heterogeneous workers who learn. *Management Science* 60(1):110–129.

Bertsimas D, Niño-Mora J (2000) Restless bandits, linear programming relaxations, and a primal-dual index heuristic. *Operations Research* 48(1):80–90.

Darr ED, Argote L, Epple D (1995) The acquisition, transfer, and depreciation of knowledge in service organizations: Productivity in franchises. *Management science* 41(11):1750–1762.

Duke H (2021) Duke health blog. `https://www.dukehealth.org/blog/split-liver-transplant-saves-two-lives-one-donor-liver#:~:text=In\%20a\%20split\%20liver\%20transplant,baby\%20or\%20a\%20small\%20child.`, accessed: 2023-03-20.

Emre S, Umman V (2011) Split liver transplantation: An overview. *Transplantation Proceedings*, volume 43, 884–887, ISSN 00411345, URL `http://dx.doi.org/10.1016/j.transproceed.2011.02.036`.

Hackl C, Schmidt KM, Süsal C, Döhler B, Zidek M, Schlitt HJ (2018) Split liver transplantation: current developments. *World Journal of Gastroenterology* 24(47):5312.

Illeris K (2002) The three dimensions of learning .

Jacko P (2010) Restless bandits approach to the job scheduling problem and its extensions. *Modern trends in controlled stochastic processes: theory and applications* 248–267.

Jarvis P (2006) *Towards a comprehensive theory of human learning*, volume 1 (Psychology Press).

Krishnasamy S, Sen R, Johari R, Shakkottai S (2016) Regret of queueing bandits. *Advances in Neural Information Processing Systems* 29:1669–1677.

Lefrancois GR (2019) *Theories of human learning* (Cambridge University Press).

OPTN, UNOS (2016) Split Versus Whole Liver Transplantation OPTN/UNOS Ethics Committee. Technical report.

Perito ER, Roll G, Dodge JL, Rhee S, Roberts JP (2019) Split liver transplantation and pediatric waitlist mortality in the united states: potential for improvement. *Transplantation* 103(3):552–557.

UNOS (2020) Liver and intestinal organ distribution based on acuity circles to be implemented feb. 4, 2020. URL `https://unos.org/news/pre-imp-notice-liver-intestinal-dist-acuity-circles-feb-4-2020/#:~:text=The\%20acuity\%20circle\%20policy\%20replaces,donor\%20hospital\%20and\%20transplant\%20hospital.`

Whittle P (1988) Restless bandits: Activity allocation in a changing world. *Journal of applied probability* 25(A):287–298.