Electronic Companions for Multi-Armed Bandits with Endogenous Learning Curves: An Application to Split Liver Transplantation

Yanhan (Savannah) Tang, Andrew Li, Alan Scheller-Wolf, Sridhar Tayur

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 $b_{a,n}$, and that Assumption 2 holds. Define $\delta_{a,\tau,n} := \sqrt{\frac{2\log\tau}{nC^w_{a,n}}}$. For any sub-optimal arm a, if there exists $u_{a,t} \in [1,t]$ 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 $u_{a,t}$ exists for any sub-optimal arm, then the expected cumulative regret is bounded by

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

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, 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. $T_a(\tau-1) \geq u_{a,t} = \frac{8\log\tau}{C^w_a \Delta_a^2}$, 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

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

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

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

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

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

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

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

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

EC.1.2. Proof of Proposition 1 (Upper and Lower Bounds of the Supremum of the Per-Coordinate Difference Bound Parameter)

Proof Let $\omega_i^* = \inf w_i$. By definition, we know $\omega_i^* \le 1$, as the image set of φ is $[0,1]$, thus $C_n^* = \frac{1}{n\sum_{i=1}^n \omega_i^{*2}} \ge \frac{1}{n\sum_{i=1}^n 1^2} = \frac{1}{n^2}$, 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* $c_1,\dots,c_n,b_1,\dots,b_n \in \mathbb{R}$ *such that* $c_1 \ge c_2 \ge \dots c_n$ *and* $b_1 \ge b_2 \ge \dots b_n$, *and then* $\frac{1}{n}\sum_{i=1}^n c_i b_i \ge \left(\frac{1}{n}\sum_{i=1}^n c_i\right)\left(\frac{1}{n}\sum_{i=1}^n b_i\right)$.

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

EC.2. Extensions: Delayed Feedback and Feature-Based Rewards

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

EC.2.1. 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 EC.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:*

$$e_{a,n} := \phi(\hat{r}_a^{(n)},\dots,\hat{r}_a^{(n-k_a+1)}, r_a^{(n-k_a)},\dots,r_a^{(1)}) - \phi(r_a^{(n)},\dots,r_a^{(1)}) \le \frac{1}{40}\sqrt{\frac{\log n}{nC_{a,n}^w}}, \quad \forall n > n_e. \tag{EC.4}$$

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*

$$\mathbb{E}[T_a(t)] \le \frac{8\log t}{C_a^w \Delta_a^2} + 2\zeta(1.063) \tag{EC.5}$$

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*

$$\mathbb{E}[R(t)] \le \sum_{a\ne a^*} (\bar{r}_{a^*} - \underline{r}_a)\left(\frac{8\log t}{C_a^w \Delta_a^2} + 2\zeta(1.063)\right). \tag{EC.6}$$

Proof: Let $\hat{\hat{\alpha}}_{a,n}$ denote our point estimate of α_a when $n = T_a(\tau-1)$ and up to k_a true rewards have not been revealed but reward estimates are available. Below we derive probabilistic bounds for $\hat{\hat{\alpha}}_{a,n}$,

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

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

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

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

$$P\left(\hat{\alpha}_{a,n} - \alpha_a \geq \sqrt{\frac{2\log\tau}{nC_{a,n}^w}}\right) \leq \exp\left(-2\left(\sqrt{2\log\tau} - |b_{a,n}|\sqrt{nC_{a,n}^w} - |e_{a,n}|\sqrt{nC_{a,n}^w}\right)^2\right), \quad \text{(EC.7)}$$

$$P\left(\hat{\alpha}_{a,n} - \alpha_a \leq -\sqrt{\frac{2\log\tau}{nC_{a,n}^w}}\right) \leq \exp\left(-2\left(\sqrt{2\log\tau} - |b_{a,n}|\sqrt{nC_{a,n}^w} - |e_{a,n}|\sqrt{nC_{a,n}^w}\right)^2\right). \quad \text{(EC.8)}$$

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

$$\begin{aligned}
\mathbb{E}[T_a(t)] &\leq u_{a,t} + \sum_{\tau=\lfloor u_{a,t}\rfloor+1}^{t} \sum_{v=\lfloor u_{a,t}\rfloor}^{\tau-1} \sum_{v^*=1}^{\tau-1} \left(P\left(\hat{\alpha}_{a,v} - \delta_{a,\tau,v} \geq \alpha_a\right) + P\left(\hat{\alpha}_{a^*,v^*} + \delta_{a^*,\tau,v^*} \leq \alpha_{a^*}\right)\right) \\
&\leq u_{a,t} + \sum_{\tau=\lfloor u_{a,t}\rfloor+1}^{t} \sum_{v=\lfloor u_{a,t}\rfloor}^{\tau-1} \sum_{v^*=1}^{\tau-1} \left(\exp\left(-2\left(\sqrt{2\log\tau} - |b_{a,v}|\sqrt{vC_{a,v}^w} - |e_{a,v}|\sqrt{vC_{a,v}^w}\right)^2\right)\right. \\
&\qquad \left. + \exp\left(-2\left(\sqrt{2\log\tau} - |b_{a^*,v^*}|\sqrt{v^*C_{a^*,v^*}^w} - |e_{a^*,v^*}|\sqrt{v^*C_{a^*,v^*}^w}\right)^2\right)\right) \\
&\leq u_{a,t} + \sum_{\tau=\lfloor u_{a,t}\rfloor+1}^{t} \sum_{v=\lfloor u_{a,t}\rfloor}^{\tau-1} \sum_{v^*=1}^{\tau-1} 2\exp\left(-2\left(\left(1 - \frac{1}{10} - \frac{1}{40}\right)\sqrt{2\log\tau}\right)^2\right) \\
&\leq u_{a,t} + \sum_{\tau=\lfloor u_{a,t}\rfloor+1}^{t} 2\tau^2 \exp\left(-\frac{1225}{400}\log\tau\right) \leq u_{a,t} + 2\sum_{\tau=1}^{+\infty} \tau^{-\frac{425}{400}} = u_{a,t} + 2\zeta(1.063).
\end{aligned}$$

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

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

$$\mathbb{E}[R(t)] \le \sum_{a\neq a^*} \mathbb{E}[(\bar{r}_{a^*} - \underline{r}_a)T_a(t-1)] \le \sum_{a\neq a^*} (\bar{r}_{a^*} - \underline{r}_a)\left(u_{a,t} + 2\zeta(1.063)\right). \tag{EC.9}$$

Note that the estimator error bound in (EC.4) 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 $\hat{\alpha}^{MoM}_n = \frac{1}{n}\sum_{s=1,g_\omega(s)\neq 0}^{n} \frac{r^{(s)}}{g_\omega(s)}$, the decay rate of the MoM estimator's $e_{a,n}$ is also $O\left(\frac{1}{n}\right)$ which satisfies (EC.4). 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).

EC.2.2. 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. 2024). 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.

EC.3. FL-UCB Regret Bounds and Proof

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 EC.1. 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 $\underline{r}_a = \inf_t r_a^{(t)}$, for $a \in \mathcal{A}$. Theorem EC.2 establishes bounds on the FL-UCB regrets.

THEOREM EC.2. *When LP* (10) $\sim$ (14) *has a unique solution:*
(a) The expected number of times that the (non-degenerate) solution of the LP with objective (15) *is different from that of* (10) $\sim$ (14)*, satisfies*

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

(b) The F-regret is bounded by

$$\mathbb{E}[R(t)] \leq \sum_{k=1}^{K} \sum_{i=1}^{|\mathcal{A}|-k} \left(\frac{8 \log t}{C^w_{a_{(k+i)}} \Delta^2_{a_{(k)},a_{(k+i)}}} + 2\zeta(1.24) \right).$$

Proof: First, we consider the LP defined by (10) $\sim$ (14). For the sake of notational simplicity and generality, we write it in the standard form:

$$\max_{z} \quad f(z) \tag{EC.10}$$

$$s.t. \quad z \in C_{\text{set}} \tag{EC.11}$$

where $C_{\text{set}} \subset \mathbb{R}^{|\mathcal{A}|}$ is a nonempty convex set and $f : \mathbb{R}^{|\mathcal{A}|} \to \mathbb{R}^{|\mathcal{A}|}$ is a convex function. We define the concept of *normal cones*.

DEFINITION EC.2. The normal cone of a closed, convex set $C_{\text{set}} \in \mathbb{R}^n$ is

$$N_{C_{\text{set}}}(z^*) = \begin{cases} \{\xi \in \mathbb{R}^n | (\forall z \in C_{\text{set}}) \xi^T (z - z^*) \leq 0\} & \text{if } z^* \in C_{\text{set}} \\ \emptyset & \text{if } z^* \notin C_{\text{set}} \end{cases} \tag{EC.12}$$

To find the normal cone of the feasible region defined in (11) $\sim$ (14), 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 $N_{Q(A,b)}(x) = \{A^T y | y \in \mathbb{R}^m$ such that $y \geq 0$ and $y^T(b - Ax) = 0\}$.*

The optimal solution to (10) $\sim$ (14) is:

$$z_a^* = \begin{cases} \theta_a^A, & a \in \mathcal{A}_A \setminus \mathcal{A}_{BK} \\ \theta_a^{BK}, & a \in \mathcal{A}_{BK} \setminus \mathcal{A}_A \setminus \{a^*\} \\ \max\{\theta_a^{BK}, \theta_a^A\} & a \in \mathcal{A}_{BK} \cap \mathcal{A}_A \setminus \{a^*\} \\ 0, & a \notin \mathcal{A}_{BK} \cup \mathcal{A}_A \\ 1 - \sum_{a \in \mathcal{A} \setminus \{a^*\}} z_a^*, & a = a^*. \end{cases} \tag{EC.13}$$

Therefore, the normal cone at an optimal solution z^* for the convex set Q_{FUCB} as defined by (11) $\sim$ (14) is a convex cone defined by the following inequalities, where $\mathcal{A}_{BK}$ is assumed to be correctly estimated through line 5 in Algorithm 2:

$$B_{a^*, T_{a^*}(t-1)} \geq B_{a, T_a(t-1)}, \qquad \forall a \in \mathcal{A}. \tag{EC.14}$$

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 (15).

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

$$u_{a_{(k+i)}}^{a_{(k)}} = \frac{8 \log t}{C_{a_{(k+i)}}^w \Delta_{a_{(k)}, a_{(k+i)}}^2},$$

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

$$\frac{8 \log t}{C_{a_{(k+i)}}^w \Delta_{a_{(k)}, a_{(k+i)}}^2} + 2\zeta(1.24).$$

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

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

And thus the expected F-regret, or simply regret, is bounded by

$$\begin{aligned}\mathbb{E}[R(t)] &\le \sum_{a\neq a^*}(\bar{r}_{a^*}-\underline{r}_a)\,\mathbb{E}[T_a] \le \left(\bar{r}_{a^*}-\underline{r}_{a_{(|\mathcal{A}|)}}\right)\sum_{a\neq a^*}\mathbb{E}[T_a]\\ &\le (1-0)\sum_{a\neq a^*}\mathbb{E}[T_a] \le \sum_{k=1}^{K}\sum_{i=1}^{|\mathcal{A}|-k}\left(\frac{8\log t}{C^w_{a_{(k+i)}}\Delta^2_{a_{(k)},a_{(k+i)}}}+2\zeta(1.24)\right).\end{aligned}$$

Note that imposing BK- or AA-fairness incurs linear PoF as long as $\theta^A+\theta^{BK}\neq\mathbf{0}$ and a^* is not the only element included in both $\mathcal{A}^{BK}$ and $\mathcal{A}^A$, 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.4. Extension: Non-Parametric Estimators

EC.4.1. Vanilla UCB

Recall that the estimator in vanilla UCB is defined as $\hat{\alpha}^{\text{UCB}}_{a,n}:=\frac{1}{n}\sum_{i=1}^{n}r_a^{(i)}$. For example, if $r_a^{(i)}\sim\text{Bern}\left(\frac{i\alpha}{1+i}\right)$, then the bias of the empirical mean estimator is: $\mathbb{E}\hat{\alpha}^{\text{UCB}}_{a,n}-\alpha=\mathbb{E}\frac{1}{n}\sum_{i=1}^{n}r_a^{(i)}-\alpha=\mathbb{E}\frac{1}{n}\sum_{i=1}^{n}\frac{i\alpha}{1+i}-\alpha=O\left(\frac{\log n}{n}\right)$. If the empirical mean estimator is unbiased, or the bias decay rate is no slower than $\sqrt{\frac{2\log n}{nC^w_{a,n}}}$, e.g., $O\left(\frac{\log n}{n}\right)=o\left(\sqrt{\frac{2\log n}{n}}\right)$, then it meets the bias condition (i.e., Assumption 1) required to apply Theorem 1. Because of its linearity, the tightest per-coordinate difference bounds are $w_i=1/n$ and therefore the estimator $\hat{\alpha}^{\text{disc}}_{a,n}$ has a per-coordinate difference bound with parameter $C^w_{a,n}=\frac{1}{n\sum_{i=1}^{n}w_i^2}=1$, satisfying the condition for per-coordinate difference bounds (i.e., Assumption 2) for Theorem 1.

EC.4.2. Discounted Non-Parametric Estimator

Recall that we define the discounted estimator as

$$\hat{\alpha}^{\text{disc}}_{a,n}:=\frac{1-\delta_{a,n}}{1-\delta^n_{a,n}}\sum_{i=1}^{n}\delta^{n-i}_{a,n}r_a^{(i)},$$

where $\delta_{a,n}\in(0,1)$. The steps for verifying the bias condition are analogous to those for vanilla UCB. For example, if $r_a^{(i)}\sim\text{Bern}\left(\frac{i\alpha}{1+i}\right)$, then the bias of $\hat{\alpha}^{\text{disc}}_{a,n}$ is: $\mathbb{E}\left(\frac{1-\delta_{a,n}}{1-\delta^n_{a,n}}\sum_{i=1}^{n}\delta^{n-i}_{a,n}r_a^{(i)}\right)-\alpha=\frac{1-\delta_{a,n}}{1-\delta^n_{a,n}}\sum_{i=1}^{n}\delta^{n-i}_{a,n}\frac{i\alpha}{1+i}-\alpha=-\frac{1-\delta_{a,n}}{1-\delta^n_{a,n}}\sum_{i=1}^{n}\delta^{n-i}_{a,n}\frac{\alpha}{1+i}=O\left(\frac{\log n}{n}\right)$. Thus, in this example Assumption 1 is satisfied.

Due to the linearity, the tightest per-coordinate difference bounds are:

$$w_i=\frac{1-\delta_{a,n}}{1-\delta^n_{a,n}}\delta^{n-i}_{a,n},$$

and therefore the estimator $\hat{\alpha}^{\text{disc}}_{a,n}$ has a per-coordinate difference bound with parameter:

$$C^w_{a,n} = \frac{1}{n\sum_{i=1}^n w_i^2} = \left(\frac{1-\delta_{a,n}^n}{1-\delta_{a,n}}\right)^2 \cdot \frac{1}{n} \cdot \frac{1}{\sum_{i=0}^{n-1}\delta_{a,n}^{2i}} = \left(\frac{1-\delta_{a,n}^n}{1-\delta_{a,n}}\right)^2 \cdot \frac{1}{n} \cdot \frac{1-\delta_{a,n}^2}{1-\delta_{a,n}^{2n}} = \frac{1}{n} \cdot \frac{(1-\delta_{a,n}^n)^2(1+\delta_{a,n})}{(1-\delta_{a,n})(1-\delta_{a,n}^{2n})}.$$

If $\delta_{a,n} = \delta_a$ and δ_a is a fixed number, in the asymptotic case when $n \to +\infty$, we have: $\lim_{n\to+\infty} C^w_{a,n} = \lim_{n\to+\infty} \frac{1}{n} \cdot \frac{1\cdot(1+\delta_a)}{(1-\delta_a)\cdot 1} = \lim_{n\to+\infty} \frac{1}{n} \cdot \frac{1+\delta_a}{(1-\delta_a)} = 0$. In this case, this discounted UCB with fixed δ_a does not satisfy the per-coordinate-difference-bound condition (i.e., Assumption 2) for Theorem 1. However, we may allow $\delta_{a,n}$ to depend on n, such as $\delta_{a,n} = 1 - 1/n$, and then we have

$$\begin{aligned}\lim_{n\to+\infty} C^w_{a,n} &= \lim_{n\to+\infty} \frac{1}{n} \cdot \frac{\left(1-\left(1-\frac{1}{n}\right)^n\right)^2\left(1+\left(1-\frac{1}{n}\right)\right)}{\frac{1}{n}\cdot\left(1-\left(1-\frac{1}{n}\right)^{2n}\right)} = \lim_{n\to+\infty} \frac{\left(1-\left(1-\frac{1}{n}\right)^n\right)^2\cdot 2}{\left(1-\left(1-\frac{1}{n}\right)^{2n}\right)} \\ &= 2\frac{\left(1-\lim_{n\to+\infty}\left(1-\frac{1}{n}\right)^n\right)^2}{\left(1-\lim_{n\to+\infty}\left(1-\frac{1}{n}\right)^{2n}\right)} = 2\frac{\left(1-\frac{1}{e}\right)^2}{1-\frac{1}{e^2}} \approx 0.92. \end{aligned} \tag{EC.15}$$

Because $\lim_{n\to+\infty} C^w_{a,n} > 0$ and $C^w_{a,n}$ is monotonically decreasing in n, Assumption 2 is satisfied.

EC.4.3. Reweighted Non-Parametric Estimator

Recall that we defined the estimator as: $\hat{\alpha}^{\text{rew}}_{a,n} := \frac{2}{n(n+1)}\sum_{i=1}^n i r_a^{(i)}$. For example, if $r_a^{(i)} \sim \text{Bern}\left(\frac{i\alpha}{1+i}\right)$, then the bias of $\hat{\alpha}^{\text{rew}}_{a,n}$ is: $\mathbb{E}\frac{2}{n(n+1)}\sum_{i=1}^n i r_a^{(i)} - \alpha = \frac{2}{n(n+1)}\sum_{i=1}^n i^2\frac{\alpha}{1+i} - \alpha = -\frac{2\alpha}{n(n+1)}\sum_{i=1}^n \frac{i^2}{1+i} - \alpha = O\left(\frac{1}{n}\right)$. Thus, in this example, Assumption 1 is satisfied.

Because of the linearity, the tightest per-coordinate difference bounds are $w_i = \frac{2i}{n(n+1)}$ and therefore the estimator $\hat{\alpha}^{\text{rew}}_{a,n}$ has a per-coordinate difference bound with a parameter:

$$C^w_{a,n} = \frac{1}{n\sum_{i=1}^n w_i^2} = \frac{n(n+1)^2}{4\sum_{i=1}^n i^2} = \frac{6n(n+1)^2}{4n(n+1)(2n+1)} = \frac{3(n+1)}{2(2n+1)}. \tag{EC.16}$$

Note that the right hand side is decreasing in n. As $n \to +\infty$, we have $C^w_{a,n} \to \frac{3}{4} > 0$, satisfying Assumption 2.

EC.5. Extension: MAP Estimators

In this section, we provide additional details about the MAP estimators for the settings where both α and ω are unknown. We assume α is in a closed interval within $(0,1)$ (e.g., $[0.01, 0.99]$) and the feasible region of ω is bounded. For simplicity, we omit the sub/superscript about arms.

EC.5.1. Likelihood and Its Derivatives

We first introduce a new notation: $\sigma_i := 1/(1+\exp(-\gamma s_i + \omega))$ for each i in $\mathcal{T}$. When the value of ω is specified to $\hat{\omega}$, we define $\hat{\sigma}_i$ accordingly. For any generic $s \in \mathbb{R}$, we also define $\sigma := 1/(1+\exp(-\gamma s+\omega))$. Its derivatives can be written as $\frac{\partial\sigma}{\partial\omega} = -\sigma(1-\sigma)$ and $\frac{\partial\sigma}{\partial s} = \gamma\sigma(1-\sigma)$. Asymptotically, as $n \to +\infty$, we have $s_n \to +\infty$ and $\sigma_n \to 1$.

The posterior probability, including a Gaussian prior on ω with zero-mean and variance $\sigma_\omega^2 = \gamma^{-1}$, is equal to

$$\ell(\alpha, \omega; r_1, \ldots, r_n) := -\frac{1}{\sqrt{2\pi\sigma_\omega^2}} - \frac{\omega^2}{2\sigma_\omega^2} + \sum_{i=1}^{n} r_i(\log\alpha + \log\sigma_i) + (1-r_i)\log(1-\alpha\sigma_i). \tag{EC.17}$$

The first-order derivatives w.r.t. α and ω are:

$$\frac{\partial\ell}{\partial\alpha} = \sum_{i=1}^{n} \frac{r_i}{\alpha} - \frac{(1-r_i)\sigma_i}{1-\alpha\sigma_i}, \qquad \frac{\partial\ell}{\partial\omega} = -\frac{\omega}{\sigma_\omega^2} + \sum_{i=1}^{n} -r_i(1-\sigma_i) + \frac{(1-r_i)\alpha\sigma_i(1-\sigma_i)}{1-\alpha\sigma_i}. \tag{EC.18}$$

The second-order derivatives are:

$$\frac{\partial^2\ell}{\partial r_i\partial\alpha} = \frac{1}{\alpha(1-\alpha\sigma_i)}, \qquad \frac{\partial^2\ell}{\partial r_i\partial\omega} = -\frac{1-\sigma_i}{1-\alpha\sigma_i}, \tag{EC.19}$$

$$\frac{\partial^2\ell}{\partial\alpha^2} = \sum_{i=1}^{n} -\frac{r_i}{\alpha^2} - \frac{(1-r_i)\sigma_i^2}{(1-\alpha\sigma_i)^2}, \qquad \frac{\partial^2\ell}{\partial\alpha\partial\omega} = \frac{\partial^2\ell}{\partial\omega\partial\alpha} = \sum_{i=1}^{n} \frac{(1-r_i)\sigma_i(1-\sigma_i)}{(1-\alpha\sigma_i)^2}, \tag{EC.20}$$

$$\frac{\partial^2\ell}{\partial\omega^2} = -\frac{1}{\sigma_\omega^2} - \sum_{i=1}^{n} \sigma_i(1-\sigma_i)\left(1 - \frac{(1-r_i)(1-\alpha)}{(1-\alpha\sigma_i)^2}\right). \tag{EC.21}$$

EC.5.2. Bias Condition of MAP Estimators

We first derive an analytic form of $\hat{\alpha}$ as a function of $\boldsymbol{r}$ and $\hat{\omega}$ from (EC.18). Since the gradients are zero at $(\hat{\alpha}, \hat{\omega})$, we have

$$0 = \hat{\alpha}(1-\hat{\alpha})\left.\frac{\partial\ell}{\partial\alpha}\right|_{\hat{\alpha},\hat{\omega}} - \hat{\alpha}\left.\frac{\partial\ell}{\partial\omega}\right|_{\hat{\alpha},\hat{\omega}} = \frac{\hat{\alpha}\hat{\omega}}{\sigma_\omega^2} + \sum_{i=1}^{n} r_i - \hat{\alpha}\hat{\sigma}_i \quad \Longrightarrow \quad \hat{\alpha} = \frac{\frac{1}{n}\sum_{i=1}^{n} r_i}{\frac{1}{n}\left(-\hat{\omega}/\sigma_\omega^2 + \sum_{i=1}^{n}\hat{\sigma}_i\right)}. \tag{EC.22}$$

As $n \to +\infty$, the denominator converges to 1. By the central limit theorem, the numerator has an asymptotic distribution: $\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} r_i - \alpha^*\right) = \mathcal{N}(0, \alpha^*(1-\alpha^*))$, where α^* is the true value. Therefore, the bias of $\hat{\alpha}$ is $o(1/\sqrt{n})$, satisfying the bias condition for Theorem 1.

EC.5.3. Asymptotic Properties of the Second-Order Derivatives

Since $\alpha, \sigma_i \in (0,1)$ and $r_i \in [0,1]$, the individual terms in (EC.19) $\sim$ (EC.20) are negative in $\partial^2\ell/\partial\alpha^2$ and non-negative in others. Because $1-\sigma_n \le \exp(-\gamma n + \omega)$, the terms in $\partial^2\ell/\partial\alpha\partial\omega$ can be bounded by an exponential series converging to zero, and hence the cumulative sum can be bounded by a constant. Therefore, as $n \to +\infty$,

$$-\frac{\partial^2\ell}{\partial\alpha^2} = \Theta(n), \quad \frac{\partial^2\ell}{\partial\alpha\partial\omega} = \frac{\partial^2\ell}{\partial\omega\partial\alpha} = O(1), \quad \frac{\partial^2\ell}{\partial r_i\partial\alpha} = \Theta(1), \quad \frac{\partial^2\ell}{\partial r_i\partial\omega} = \Theta(1). \tag{EC.23}$$

Although $\partial^2\ell/\partial\omega^2$ contains positive and negative terms, the positive ones can be bounded by 1. Specifically,

$$\frac{\partial^2\ell}{\partial\omega^2} = -\frac{1}{\sigma_\omega^2} - \sum_{i=1}^{n}\left(\sigma_i(1-\sigma_i) + \frac{(1-\alpha)\sigma_i(1-\sigma_i)r_i}{(1-\alpha\sigma_i)^2}\right) + \sum_{i=1}^{n}\frac{(1-\alpha)\sigma_i(1-\sigma_i)}{(1-\alpha\sigma_i)^2} \tag{EC.24}$$

$$\leq -\frac{1}{\sigma_\omega^2} - \sum_{i=1}^{n}\left(\sigma_i(1-\sigma_i) + \frac{(1-\alpha)\sigma_i(1-\sigma_i)r_i}{(1-\alpha\sigma_i)^2}\right) + \sum_{s=-\infty}^{+\infty}\frac{(1-\alpha)\sigma_i(1-\sigma_i)}{(1-\alpha\sigma_i)^2} \tag{EC.25}$$

$$\approx -\frac{1}{\sigma_\omega^2} - \sum_{i=1}^{n}\left(\sigma_i(1-\sigma_i) + \frac{(1-\alpha)\sigma_i(1-\sigma_i)r_i}{(1-\alpha\sigma_i)^2}\right) + \int_{-\infty}^{+\infty}\frac{(1-\alpha)\sigma_i(1-\sigma_i)}{(1-\alpha\sigma_i)^2}\,\mathrm{d}s \tag{EC.26}$$

$$= -\frac{1}{\sigma_\omega^2} - \sum_{i=1}^{n}\left(\sigma_i(1-\sigma_i) + \frac{(1-\alpha)\sigma_i(1-\sigma_i)r_i}{(1-\alpha\sigma_i)^2}\right) + \frac{(1-\alpha)\gamma}{\alpha}\left(\frac{1}{1-\alpha\sigma}\right)\Bigg|_{s=-\infty}^{+\infty} \tag{EC.27}$$

$$= -\frac{1}{\sigma_\omega^2} - \sum_{i=1}^{n}\left(\sigma_i(1-\sigma_i) + \frac{(1-\alpha)\sigma_i(1-\sigma_i)r_i}{(1-\alpha\sigma_i)^2}\right) + \gamma \tag{EC.28}$$

$$= -\sum_{i=1}^{n}\left(\sigma_i(1-\sigma_i) + \frac{(1-\alpha)\sigma_i(1-\sigma_i)r_i}{(1-\alpha\sigma_i)^2}\right) < 0. \tag{EC.29}$$

Similar to the argument about $\partial^2\ell/\partial\alpha\partial\omega$, we can show $-\partial^2\ell/\partial\omega^2 = \Theta(1)$.

The determinant of the Hessian matrix is then:

$$\det(H) = \det\left(\begin{bmatrix}\partial^2\ell/\partial\alpha^2 & \partial^2\ell/\partial\alpha\partial\omega\\ \partial^2\ell/\partial\alpha\partial\omega & \partial^2\ell/\partial\omega^2\end{bmatrix}\right) = \frac{\partial^2\ell}{\partial\alpha^2}\frac{\partial^2\ell}{\partial\omega^2} - \left(\frac{\partial^2\ell}{\partial\alpha\partial\omega}\right)^2 = \Theta(n). \tag{EC.30}$$

The positive determinant and negative trace indicate that the Hessian matrix of the maximization problem is negative-definite asymptotically. Besides, the positive determinant guarantees the Hessian matrix invertible, and its inverse is

$$H^{-1} = \det(H)^{-1}\begin{bmatrix}\partial^2\ell/\partial\omega^2 & -\partial^2\ell/\partial\alpha\partial\omega\\ -\partial^2\ell/\partial\alpha\partial\omega & \partial^2\ell/\partial\alpha^2\end{bmatrix} = \begin{bmatrix}\Theta(1/n) & O(1/n)\\ O(1/n) & \Theta(1)\end{bmatrix}. \tag{EC.31}$$

EC.5.4. Theoretical Per-Coordinate Difference Bound

It is easy to see that a sufficient condition of w_i being a feasible per-coordinate difference bound is that $\frac{\mathrm{d}\alpha}{\mathrm{d}r_i}\Big|_{\hat\alpha,\hat\omega}$ is bounded by w_i from both sides for any $r_i \in [0,1]$. To derive $\frac{\mathrm{d}\alpha}{\mathrm{d}r_i}\Big|_{\hat\alpha,\hat\omega}$, we use the first-order Taylor expansion of (EC.18) at $(\hat\alpha,\hat\omega,r_i)$:

$$\left[\begin{matrix}\frac{\partial\ell}{\partial\alpha}\\ \frac{\partial\ell}{\partial\omega}\end{matrix}\right]\Bigg|_{\hat\alpha',\hat\omega',r_i'} = \left[\begin{matrix}\frac{\partial\ell}{\partial\alpha}\\ \frac{\partial\ell}{\partial\omega}\end{matrix}\right]\Bigg|_{\hat\alpha,\hat\omega,r_i} + H|_{\hat\alpha,\hat\omega,r_i}\begin{bmatrix}\hat\alpha'-\hat\alpha\\ \hat\omega'-\hat\omega\end{bmatrix} + \left[\begin{matrix}\frac{\partial\ell^2}{\partial r_i\partial\alpha}\\ \frac{\partial\ell^2}{\partial r_i\partial\omega}\end{matrix}\right]\Bigg|_{\hat\alpha,\hat\omega,r_i}(r_i'-r_i). \tag{EC.32}$$

Observing that the first-order derivatives of ℓ are zero at $(\hat\alpha,\hat\omega,r_i)$ and $(\hat\alpha',\hat\omega',r_i')$, and that H is invertible, and taking the limit of $r_i' \to r_i$, the above equation can be rearranged to

$$\begin{bmatrix}\mathrm{d}\hat\alpha\\ \mathrm{d}\hat\omega\end{bmatrix} = -\left(H|_{\hat\alpha,\hat\omega,r_i}\right)^{-1}\left[\begin{matrix}\frac{\partial\ell^2}{\partial r_i\partial\alpha}\\ \frac{\partial\ell^2}{\partial r_i\partial\omega}\end{matrix}\right]\Bigg|_{\hat\alpha,\hat\omega,r_i}\mathrm{d}r_i. \tag{EC.33}$$

Plugging (EC.31) above yields

$$\frac{\mathrm{d}\hat\alpha}{\mathrm{d}r_i} = -\det(H)^{-1}\left(\frac{\partial^2\ell}{\partial\omega^2}\frac{\partial\ell^2}{\partial r_i\partial\alpha} - \frac{\partial^2\ell}{\partial\alpha\partial\omega}\frac{\partial\ell^2}{\partial r_i\partial\omega}\right) = \Theta(1/n). \tag{EC.34}$$

Hence, a feasible per-coordinate difference bound can be defined as

$$w_i := \max_{\hat\alpha,\hat\omega,\boldsymbol{r}}\left|\frac{\mathrm{d}\hat\alpha}{\mathrm{d}r_i}\right|. \tag{EC.35}$$

Since $\hat{\alpha}$ and $\hat{\omega}$ are bounded, we have $w_i = \Theta(1/n)$ and $C_n^w = \Theta(1)$, meaning that Assumption 2 is also satisfied. Thus, we may apply Theorem 1, which guarantees that the regret of our L-UCB algorithm with MAP estimators is logarithmic in T.

EC.5.5. Practical Values of the Per-Coordinate Difference Bound Parameters

The last subsections established the theoretical sufficient conditions for our MAP estimator to satisfy the premises of Theorem 1. This subsection discusses practical procedures to find the values of the per-coordinate difference bounds. The upper bound of (EC.35) can be estimated empirically. We first do a grid search over the feasible region of α and ω. For each pair of α and ω, we randomly sample one sequence of rewards (or multiple replicates) parameterized by them, and find $\hat{\alpha}$ and $\hat{\omega}$ by the MAP estimator. We then directly calculate the value of (EC.34). Lastly, we find the maximum among all pairs of α and ω and use that for the value of w_i.

Our numerical experiments show that random sampling is a reliable approach to estimate w_i efficiently. Specifically, using only 5 samples for each pair of α and ω gives very similar results as having 100 samples per pair.

EC.6. Details of the Numerical Study in Section 6

EC.6.1. Details of the SLT Simulation Setup

Below we detail how we estimate α's from the STAR files and open data sources. 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 (500NM) 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.

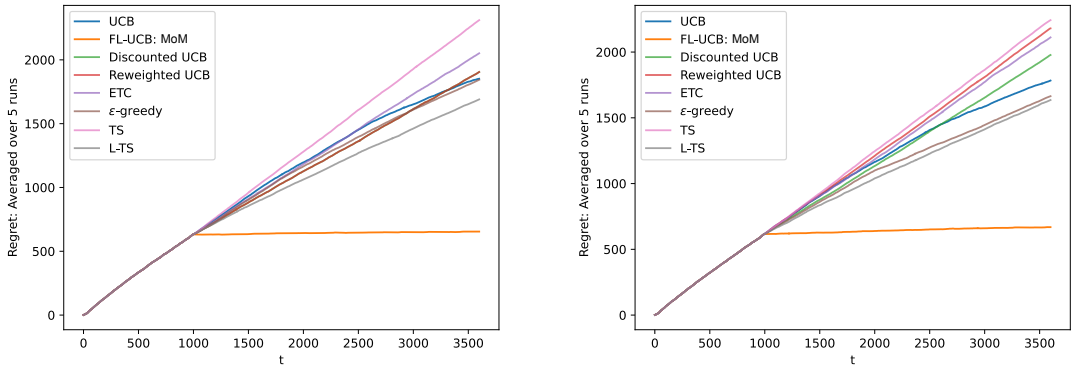
Therefore, in total, we have $10 \times 5 = 50$ arms for the livers splittable in the geographical region of interest. Recall that 10% of livers are medically safe to split (OPTN and UNOS 2016), so at least 800 livers can be used for SLT a year in the 500NM Circle, with each liver supporting two SLT 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. 2019), 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.

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. Price of Fairness and Delayed Rewards

Figure EC.2 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 EC.3 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 $t \geq 900$.

In Figure EC.1, we assume the prediction accuracy in SLT is 60%; Figure EC.4 shows results assuming the prediction accuracy is 85%.



(a) Regret w/o fairness (b) F-regret w/ fairness

Figure EC.1 **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.**

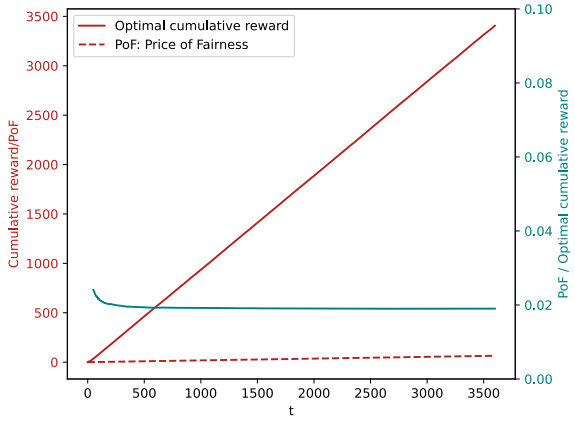


Figure EC.2 **[PoF / Optimal cumulative reward] is constant, i.e., PoF is $O(t)$; when $t < 200$, the ratio could be subject to numerical instability.**

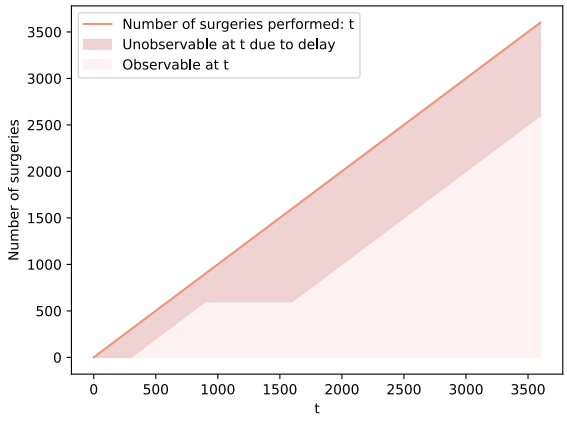
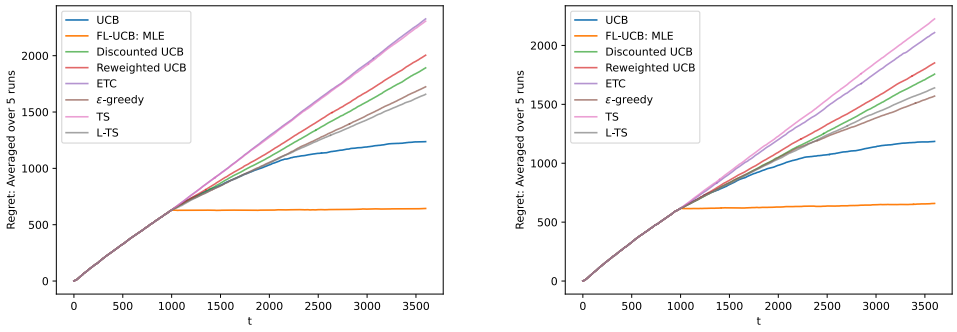


Figure EC.3 **The delay in observing rewards. For each t, the true rewards are not revealed until after some delay and can only be estimated using a 60% accurate surrogate.**

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 $t = 3600$.



(a) Regret w/o fairness (b) F-regret w/ fairness

Figure EC.4 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 6.1). Estimates based on demographics and perioperative clinical metrics are available and are 85% accurate.

Reference for the Appendix

Duke H (2021) Duke health blog. `https://www.dukehealth.org/blog/split-liver-transplant-saves-two-lives-one-donor-liver`, accessed: 2023-03-20.

Emre S, Umman V (2011) Split liver transplantation: an overview. *Transplantation proceedings*, volume 43, 884–887 (Elsevier).

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.

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/`.