A Concentration results

A.1 Poissonization of the KL indices

We require variants of Lemma 9 and Theorem 10 in [5] adapted to our setting.

Lemma 13. *For $\theta \in [0,1]$, $(\tau_i)_{1\leq i\leq L} \in [0,1]$, let $(X_{i,j})_{1\leq i\leq L, j\geq 1}$ be a collection of independent Bernoulli random variables such that $\mathbb{E}(X_{i,j}) = \tau_i\theta$ and $(\epsilon_{i,j}) \in \{0,1\}$ associated deterministic indicators. For $1 \leq i \leq L$, denote by $n_i = \sum_{j=1}^{\infty} \epsilon_{i,j}$ and whe shall assume that all n_i are finite an that at least one of them is non-zero. Let $X = \sum_{i=1}^{L}\sum_{j=1}^{\infty} \epsilon_{i,j}X_{i,j}$ and denote by $\phi(\lambda) = \log \mathbb{E}\left[\exp(\lambda X)\right]$ its log-Laplace transform and by $\phi^*(x) = \sup_\lambda x\lambda - \phi(\lambda)$ the associated convex conjugate (Fenchel–Legendre transform).*

Then, for all $\lambda \in \mathbb{R}$,

$$\phi(\lambda) \leq \left(\sum_{i=1}^{L} \tau_i n_i\right) \theta \left(e^\lambda - 1\right), \tag{7}$$

and, for all $x \geq 0$,

$$\phi^*(x) \geq \left(\sum_{i=1}^{L} \tau_i n_i\right) d_{\mathrm{Pois}}\left(\frac{x}{\sum_{i=1}^{L} \tau_i n_i}, \theta\right), \tag{8}$$

where $d_{\mathrm{Pois}}(p,q) = p\log p/q + q - p$ denotes the Poisson Kullback-Leibler divergence.

Proof By direct calculation,

$$\phi(\lambda) = \sum_{i=1}^{L} n_i \log\left(1 - \tau_i\theta + \tau_i\theta e^\lambda\right).$$

The function $\tau_i \to \log\left(1 - \tau_i\theta + \tau_i\theta e^\lambda\right)$ is a strictly concave function on $[0,1]$ and we upper bound it by its tangent in 0, that is,

$$\log\left(1 - \tau_i\theta + \tau_i\theta e^\lambda\right) \leq \tau_i\theta(e^\lambda - 1),$$

which yields (7) upon summing on i.

The r.h.s. of (7) is easilly recognized as the log-Laplace transform of the Poisson distribution with expectation $\left(\sum_{i=1}^{L} \tau_i n_i\right)\theta$. To obtain (8), we use the observations that $x\lambda - a\left(e^\lambda - 1\right)$ is maximized for $\lambda = \log(x/a)$ where it is equal to $d_{\mathrm{Pois}}(x,a)$ as well as the fact that $d_{\mathrm{Pois}}(\tau x, \tau a) = \tau d_{\mathrm{Pois}}(x,a)$. ■

Lemma 13 bounds the log-Laplace transform of the Bernoulli distribution with that of the Poisson distribution with the same mean and uses the stability of the Poisson distribution. Using $d_{\mathrm{Pois}}(p,q)$ instead of $d(p,q)$ –where $d(p,q) = p\log p/q + (1-p)\log[1-p)/(1-q)$ denotes the Bernoulli Kullback-Leibler divergence– will of course induce a performance gap, which is however not significant for low values of the probabilities as shown by the Lemma 8, which we recall below.

Lemma 8. *For $0 < p < q < 1$,*

$$(1-q)d(p,q) \leq d_{\mathrm{Pois}}(p,q) \leq d(p,q).$$

Proof For the upper bound,

$$d(p,q) - d_{\mathrm{Pois}}(p,q) = (1-p)\log\frac{1-p}{1-q} - (q-p) = -(1-p)\log(1 + \frac{p-q}{1-p}) + (p-q) \geq 0,$$

using $-\log(1+x) \geq -x$.

For the lower bound,

$$d_{\mathrm{Pois}}(p,q) - (1-q)d(p,q) = qp\log\frac{p}{q} + q - p - (1-q)(1-p)\log\left(\frac{1-p}{1-q}\right). \tag{9}$$

One has $d_{\mathrm{Pois}}(q,q) = d(q,q) = 0$ and the derivative of (9) wrt. p is equal to

$$q \log \frac{p}{q} + (1-q) \log \left(\frac{1-p}{1-q} \right) = -d(q,p) \leq 0.$$

Hence, $d_{\mathrm{Pois}}(p,q) - (1-q)d(p,q)$ is positive when $p \leq q$. ■

We can now prove the concentration result stated in Lemma 7 that we recall below for readability purpose.

Lemma 7. *Assume that the sequence of pulls is fixed beforehand and let k be an arm in $\{1,...,K\}$. Then for any $\delta > 0$ and for all $t > 0$,*

$$\mathbb{P}\left(\left\{ \hat{\theta}_k(t) < \theta_k \right\} \cap \left\{ \tilde{N}_k(t) d_{\mathrm{Pois}}(\hat{\theta}_k(t), \theta_k) > \delta \right\} \right) < e^{-\delta}.$$

where $d_{\mathrm{Pois}}(p,q) = p \log p/q + q - p$ denotes the Poisson Kullback-Leibler divergence.

Proof To bound $\mathbb{P}(\hat{\theta}_k(t) < x) = \mathbb{P}(S_k(t) < \tilde{N}_k(t)x)$, for $0 < x < \theta_k$, apply Chernoff's method using the result of Lemma 13 to obtain

$$\mathbb{P}(\hat{\theta}_k(t) < x) \leq e^{-\tilde{N}_k(t) d_{\mathrm{Pois}}(x, \theta_k)}.$$

Using that $x \mapsto d_{\mathrm{Pois}}(x, \theta_k)$ is decreasing on $[0, \theta_k]$, we can apply it on both side of the inequality on the left-hand side to obtain

$$\mathbb{P}\left(\left\{ \hat{\theta}_k(t) < \theta_k \right\} \cap \left\{ \tilde{N}_k(t) d_{\mathrm{Pois}}(\hat{\theta}_k(t), \theta_k) > \tilde{N}_k(t) d_{\mathrm{Pois}}(x, \theta_k) \right\} \right) \leq e^{-\tilde{N}_k(t) d_{\mathrm{Pois}}(x, \theta_k)}.$$

Denoting $\delta = \tilde{N}_k(t) d_{\mathrm{Pois}}(x, \theta_k)$ yields the desired result. ■

Theorem 14. *Consider $(\tau_i)_{1 \leq i \leq L} \in [0,1]$, $\theta \in (0,1)$, and independent sequences $(X_i(s))_{s \geq 1}$ of independent Bernoulli random variables such that $\mathbb{E}X_i(s) = \tau_i \theta$. Let $\mathcal{F}_t$ denote an increasing sequence of sigma-fields, such that for each t and all i, $\sigma(X_i(1), \ldots, X_i(t)) \subset \mathcal{F}_t$. Also consider a predictable sequence of indicator variables $\epsilon_i(s) \in \{0,1\}$, that is, such that $\sigma(\epsilon_1(t+1), \ldots, \epsilon_L(t+1)) \subset \mathcal{F}_t$.*

Define

$$S_i(t) = \sum_{s=1}^{t} \epsilon_i(s) X_i(s), \qquad N_i(t) = \sum_{s=1}^{t} \epsilon_i(s);$$

and the pooled quantities

$$S(t) = \sum_{i=1}^{L} S_i(t), \qquad N(t) = \sum_{i=1}^{L} N_i(t), \qquad \tilde{N}(t) = \sum_{i=1}^{L} \tau_i N_i(t), \qquad \hat{\theta}(t) = \frac{S(t)}{\tilde{N}(t)}.$$

The KLUCB index, defined as,

$$U^{\mathrm{KL}}(n) = \max \left\{ q \in \left[\hat{\theta}(n), \theta_M \right] \,:\, \tilde{N}(n) d_{\mathrm{Pois}}(\hat{\theta}(n), q) \leq \delta \right\}.$$

satisfies

$$\mathbb{P}\left(U(n) \leq \theta \right) \leq e \lceil \delta \log(n) \rceil e^{-\delta}.$$

Proof The proof is analogous to that of Theorem 10 of [5] and we only detail the step that differs, namely, the identification of the supermartingale W_t^{λ}.

Define, $W_0^{\lambda} = 1$ and, for $t \geq 1$,

$$W_t^{\lambda} = \exp\left(\lambda S(t) - \tilde{N}(t) \theta \left(e^{\lambda} - 1 \right) \right).$$

$$\mathbb{E}\left[\exp(\lambda(S(t+1)-S(t)))\,|\mathcal{F}_t\right] = \mathbb{E}\left[\exp\left(\lambda\sum_{i=1}^{L}\epsilon_i(t+1)X_i(t+1)\right)\middle|\mathcal{F}_t\right]$$
$$\leq \exp\left(\left(\sum_{i=1}^{L}\tau_i\epsilon_i(t+1)\right)\theta\left(e^{\lambda}-1\right)\right)$$
$$= \exp\left(\left(\tilde{N}(t+1)-\tilde{N}(t)\right)\theta\left(\theta e^{\lambda}-1\right)\right),$$

where we have used (7) and the definition of $\tilde{N}(t)$. Multiplying both sides of the inequality by $\exp\left(\lambda S(t)-\tilde{N}(t+1)\theta\left(e^{\lambda}-1\right)\right)$ show that $\mathbb{E}W_{t+1}^{\lambda}\leq\mathbb{E}W_t^{\lambda}$ and hence that W_t^{λ} is a supermartingale.

The rest of the proof is as in [5] replacing $N(t)$ by $\tilde{N}(t)$ and $\phi_\mu(\lambda)$ by $\theta\left(e^{\lambda}-1\right)$. ■

B Details on the Lower Bound Results

We provide here the details of the proof of Theorems 4 and 3. The key result that we use is a lower bound on the log-likelihood ratio under two alternative bandit models θ and θ' that do not have the same best arm. Namely, according to Lemma 1 of [13], we have

$$\liminf_{T\to\infty}\frac{\mathbb{E}[\ell_T]}{\log(T)}\geq 1.$$

Now, considering specific changes of measures θ' that only modify the distribution of one single suboptimal arm, we are going to obtain lower bounds on each expected number of pulls $\mathbb{E}[N_k(T)]$ for $k\neq 1$ as in Appendix B of [13].

Uncensored Setting: As argued in Section 4, in the uncensored setting the likelihood of the observations is

$$\mathbb{E}_\theta\left[\ell_T\right]=\sum_{s=1}^{T}d(\theta_{A_s}\tau_{T-s},\theta'_{A_s}\tau_{T-s}).$$

Now, fix arm $k\neq 1$ and for $\epsilon>0$, consider $\theta'=(\theta_1,\ldots,\theta_{k-1},\theta_1+\epsilon,\ldots,\theta_K)$. For this change of measure, the expected log-likelihood only contains the terms involving arm k:

$$\mathbb{E}_\theta\left[\ell_T\right]=\sum_{s=1}^{T}\mathbb{1}\{A_s=k\}d(\theta_k\tau_{T-s},(\theta_1+\epsilon)\tau_{T-s}).$$

Now, in order to obtain an expression that involves $\mathbb{E}[N_k(T)]$, we need to bound from above this sum using Lemma 5 of the Appendix B of [11], which we recall here for completeness.

Lemma 15. *Let p,q be any fixed real numbers in $(0,1)$. The function $f:\alpha\mapsto d(\alpha p,\alpha q)$ is convex and increasing on $(0,1)$. As a consequence, for any $\alpha<1$, $d(\alpha p,\alpha q)<d(p,q)$.*

Thus, for each $s\geq 1$ we have $\tau_{T-s}\leq 1$ and according to the above result,

$$d(\theta_k,(\theta_1+\epsilon))\geq d(\theta_k\tau_{T-s},(\theta_1+\epsilon)\tau_{T-s})$$

and

$$\mathbb{E}[N_k(T)]d(\theta_k,\theta_1+\epsilon)\geq\sum_{s=1}^{T}\mathbb{1}\{A_s=k\}d(\theta_k\tau_{T-s},(\theta_1+\epsilon)\tau_{T-s}).$$

We obtain

$$\liminf_{T\to\infty}\frac{\mathbb{E}[N_k(T)]d(\theta_k,\theta_1+\epsilon)}{\log(T)}\geq\liminf_{T\to\infty}\frac{\mathbb{E}[\ell_T]}{\log(T)}\geq 1.$$

Letting $\epsilon\to 0$ yields

$$\liminf_{T\to\infty}\frac{\mathbb{E}[N_k(T)]}{\log(T)}\geq\frac{1}{d(\theta_k,\theta_1)}.$$

In order to bound the expected regret L_T, we use the inequality (2) from Lemma 1:

$$L_T \geq \sum_{k=2}^{K} (\theta_1 - \theta_k) \left(\mathbb{E}[N_k(t)] - \mu \right),$$

where $\mu = \mathbb{E}[D_s]$. We now lower bound each $\mathbb{E}[N_k(t)]$ and under the assumption that $\mathbb{E}[D_s] < \infty$ and we use that $\liminf_{T\to\infty} \mu / \log(T) = 0$ to obtain

$$\liminf_{T\to\infty} \frac{L_T}{\log(T)} \geq \liminf_{T\to\infty} \frac{\sum_{k=2}^{K} (\theta_1 - \theta_k) \left(\mathbb{E}[N_k(t)] - \mu \right)}{\log(T)} \geq \sum_{k=2}^{K} \frac{(\theta_1 - \theta_k)}{d(\theta_k, \theta_1)}.$$

Censored Setting: The proof in the Censored Setting follows the same step as the proof above expect for the fact that we do not require Lemma 5 of [11] in order to bound the log-likelihood ratio. We directly have

$$\mathbb{E}_\theta \left[\ell_T \right] = \sum_{s=1}^{T-m} d(\theta_{A_s} \tau_m, \theta'_{A_s} \tau_m) + \sum_{s=T-m}^{T} d(\theta_{A_s} \tau_{T-s}, \theta'_{A_s} \tau_{T-s}).$$

Proceeding as above and considering the adequate change of measure involving only one suboptimal arm k and taking $\epsilon \to 0$, we obtain

$$\liminf_{T\to\infty} \frac{\mathbb{E}[N_k(T)] d(\tau_m \theta_k, \tau_m \theta_1) + \sum_{s=T-m}^{T} d(\theta_k \tau_{T-s}, \theta_1 \tau_{T-s})}{\log(T)} = \liminf_{T\to\infty} \frac{\mathbb{E}[N_k(T)] d(\tau_m \theta_k, \tau_m \theta_1)}{\log(T)} \geq 1,$$

where we used the fact that the second term of the sum in the left-hand side is finite. The end of the proof is similar to the uncensored setting case treated above where we can simply bound the regret according to Eq. (3) as

$$L_T \geq \sum_{k=2}^{k} \tau_m (\theta_1 - \theta_k) \mathbb{E}[N_k(T-m)] + \sum_{s=T-m+1}^{T} \tau_{T-s} (\theta_1 - \theta_{A_s})$$

in order to obtain the asymptotic lower bound.

C Analysis of DelayedUCB and DelayedKLUCB

In order to control the empirical averages of the rewards of each arm for different values of $N_k(t)$, we introduce the notation $\hat{\theta}_{k,s} := \sum_{u=1}^{s} X_{k,u}/s$ for the mean over the first s pulls of k.

C.1 DelayedUCB

In this section, we provide the complete proof of Theorem 9.

We decompose the regret after bounding by 1 the first m losses of the policy :

$$L_{\text{UCB}}(T) \leq m + \sum_{k>1} \tau_m \Delta_k \mathbb{E} \left[\sum_{t>m}^{T} \mathbb{1}\{A_t = k\} \right].$$

Hence we only need to bound the number of suboptimal pulls, as in the seminal proof by [1]. For any suboptimal $k > 1$, we have:

$$\begin{aligned} \mathbb{E}[N_k(T)] \leq 1 &+ \sum_{t=K+1}^{T} \mathbb{P}\left(U_1^{\text{UCB}}(t) < \theta_1 \right) \\ &+ \sum_{t=K+1}^{T} \mathbb{P}\left(A_{t+1} = k, U_k^{\text{UCB}}(t) \geq \theta_1 \right). \end{aligned}$$

While the first term is simply handled by Proposition 6 and is $O(1/\epsilon^3) = o(\log(T)$, the second one must be controlled as in the original proof of UCB1 by [1] using the fact that for all $t > m$,

$$\frac{N_k(t)}{\tilde{N}_k(t)} \leq \frac{N_k(t-m)+m}{\tilde{N}_k(t)} \leq \frac{1}{\tau_m} + \frac{m}{\tau_m N_k(t-m)}$$

that allows us to upper-bound the optimistic indices as

$$\hat{\theta}_k(t) + \left(\frac{1}{\tau_m} + \frac{m}{\tau_m N_k(t-m)}\right)\sqrt{\frac{\beta_\epsilon(t)}{2N_k(t)}} \geq U_k^{\text{UCB}}(t).$$

Then, we use this upper bound on the indices in order to bound the relevant sum of probabilities.

$$\begin{aligned}
&\sum_{t=m+1}^{T} \mathbb{P}\left(A_{t+1} = k, U_k^{\text{UCB}}(t) \geq \theta_1\right) \\
&\leq \mathbb{E}\left[\sum_{s\geq 1} \mathbb{1}\left\{\hat{\theta}_{i,s} + \left(\frac{1}{\tau_m} + \frac{m}{\tau_m s}\right)\sqrt{\frac{\beta_\epsilon(t)}{2s}} \geq \theta_i + \Delta_i\right\}\right].
\end{aligned}$$

In order to upper-bound this expectation, we first introduce the quantity $\underline{s}_i > 0$ defined by

$$\left(\frac{1}{\tau_m} + \frac{m}{\tau_m \underline{s}_i}\right)\sqrt{\frac{\beta_\epsilon(t)}{2\underline{s}_i}} = \Delta_i,$$

that we rewrite, with the introduction of $\gamma_i > 0$, as

$$\underline{s}_i = \frac{\beta_\epsilon(t)}{2\tau_m^2\Delta_i^2}(1+\gamma_i)^2 \quad \text{ so that we get } \quad \left(1 + \frac{m}{\underline{s}_i}\right)\frac{1}{1+\gamma_i} = 1.$$

Simple computations finally leads to, if $\gamma_i \leq 1$,

$$\frac{2m\tau_m^2\Delta_i^2}{\beta_\epsilon(t)} = \gamma_i(1+\gamma_i)^2 \leq 4\gamma_i.$$

As a consequence, if T is big enough (so that the left hand side is smaller than 4), we get that

$$\underline{s}_i \leq \frac{(1+\varepsilon)\log(T)}{2\tau_m^2\Delta_i^2}(1+\gamma_i)^2 \leq \frac{(1+\varepsilon)\log(T)}{2\tau_m^2\Delta_i^2}(1+3\gamma_i) \leq \frac{(1+\varepsilon)\log(T)}{2\tau_m^2\Delta_i^2} + m.$$

We now focus on the sum to upper-bound:

$$\begin{aligned}
\mathbb{E}\left[\sum_{s\geq 1} \mathbb{1}\left\{\hat{\theta}_{i,s} + \left(\frac{1}{\tau_m} + \frac{m}{\tau_m s}\right)\sqrt{\frac{\beta_\epsilon(t)}{2s}} \geq \theta_i + \Delta_i\right\}\right] &\leq \lceil \underline{s}_i \rceil + 1 + \sum_{s > \lceil \underline{s}_i \rceil} e^{-2s\left(\Delta_i - \left(\frac{1}{\tau_m} + \frac{m}{\tau_m s}\right)\sqrt{\frac{\beta_\epsilon(t)}{2s}}\right)^2} \\
&\leq \underline{s}_i + 2 + \sum_{s > \lceil \underline{s}_i \rceil} e^{-2\left(\sqrt{s}\Delta_i - \left(1 + \frac{m}{\underline{s}_i}\right)\sqrt{\frac{\beta_\epsilon(t)}{2\tau_m^2}}\right)^2},
\end{aligned}$$

where we used the Chernoff's inequality for bounded random variables.

Standard computations (comparisons between sums and integrals) give the following

$$\begin{aligned}\sum_{s>\lceil \underline{s}_i \rceil} e^{-2\left(\sqrt{s}\Delta_i - \left(1+\frac{m}{\underline{s}_i}\right)\sqrt{\frac{\beta_\epsilon(t)}{2\tau_m^2}}\right)^2} &\leq \int_{\underline{s}_i}^{\infty} e^{-2\left(\sqrt{s}\Delta_i - \left(1+\frac{m}{\underline{s}_i}\right)\sqrt{\frac{\beta_\epsilon(t)}{2\tau_m^2}}\right)^2} ds \\ &\leq \frac{1}{2\Delta_i^2}\left(1+\frac{\sqrt{2\pi}}{4}\left(1+\frac{m}{\underline{s}_i}\right)\sqrt{\frac{\beta_\epsilon(t)}{2\tau_m^2}}\right) \\ &\leq \frac{1}{2\Delta_i^2}\left(1+\frac{\sqrt{2\pi}}{4}\sqrt{\underline{s}_i}\Delta_i\right) \\ &\leq \frac{1}{2\Delta_i^2}\left(1+\frac{\sqrt{\pi}}{4}\sqrt{\frac{(1+\epsilon)\log(T)}{\tau_m^2}}+\frac{\sqrt{\pi}}{4}\sqrt{m}\right).\end{aligned}$$

As a consequence, we have just proved that

$$\sum_{t=m+1}^{T} \mathbb{P}\left(A_{t+1}=k, U_k^{\mathrm{UCB}}(t)\geq\theta_1\right) \leq \frac{(1+\epsilon)\log(T)}{2\tau_m^2\Delta_i^2} + o(\log(T))\,.$$

More precisely, combining all our claims yields that

$$L_{\mathrm{UCB}}(T) \leq \frac{(1+\epsilon)\log(T)}{2\tau_m^2\Delta_i} + O\left(\frac{1}{\Delta_i}\sqrt{\frac{(1+\epsilon)\log(T)}{2\tau_m^2}}\right) + O\left(\frac{1}{\Delta_i}\frac{1}{\epsilon^3}\right) + O\left(\frac{\sqrt{m}}{\Delta_i}+m\right),$$

and the result follows.

C.2 DelayedKLUCB

We follow the steps of [5] and decompose the regret as

$$\begin{aligned}\mathbb{E}[N_k(T)] \leq& 1+m-K\sum_{t=m+1}^{T}\mathbb{P}\left(U_1^{\mathrm{KL}}(t)<\theta_1\right)+\sum_{t=m+1}^{T}\mathbb{P}\left(A_{t+1}=k, U_k^{\mathrm{KL}}(t)\geq\theta_1\right) \\ &\leq 1+m-K+\sum_{t=m+1}^{T}\mathbb{P}\left(U_1^{\mathrm{KL}}(t)<\theta_1\right)+\sum_{t=m+1}^{T}\mathbb{P}\left(A_{t+1}=k, U_k^{\mathrm{KL}}(t)\geq\theta_1\right).\end{aligned}$$

The first term of the above sum is handled by Theorem 14 that shows that it is $o(\log(T))$. We must now bound the second sum corresponding to the cases when suboptimal indices reach the optimal mean θ_1. To proceed, we simply notice that for all t, $\tilde{N}_k(t) \geq \tau_m N_k(t-m)$. We define an alternative optimistic index that upper bounds $U_k^{\mathrm{KL}}(t)$ for $t > m$:

$$U_k^{\mathrm{KL}}(t) \leq \underset{q\in[\hat{\theta}_k,1]}{\arg\max}\{q|\tau_m N_k(t-m)d_{\mathrm{Pois}}(\hat{\theta}_k(t),q)\leq\beta_\epsilon(t)\} := U_k^{\mathrm{KL}+}(t).$$

Now we can finish the proof following the steps of the proof of Theorem 2 in [5]. First, we denote

$$K_k(T) = \frac{(1+\eta)\beta_\epsilon(t)}{d(\tau_m\theta_k, \tau_m\theta_1)}$$

and we decompose the second sum after bounding the first $K_k(T)$ terms by 1 and bounding the remaining terms in a

similar way as in Lemma 11 of [9]:

$$\begin{aligned}
\sum_{t=m+1}^{T} \mathbb{P}\left(A_{t+1}=k, U_k^{\mathrm{KL}+}(t) \geq \theta_1\right) &\leq K_k(T) + \sum_{t \geq K_k(T)+m+1} \mathbb{P}\left(A_{t+1}=k, U_k^{\mathrm{KL}+}(t) \geq \theta_1\right) \\
&\leq K_k(T) + \mathbb{E}\left[\sum_{t \geq K_k(T)+m+1} \sum_{s=1}^{t} \mathbb{1}\left\{A_{t+1}=k, N_k(t-m)=s, \tau_m s d_{\mathrm{Pois}}(\hat{\theta}_{k,s}, \theta_1) \leq \beta_\epsilon(t)\right\}\right] \\
&\leq K_k(T) + \mathbb{E}\left[\sum_{s=K_k(T)}^{T} \mathbb{1}\left\{\tau_m s d_{\mathrm{Pois}}(\hat{\theta}_{k,s}, \theta_1) \leq \beta_\epsilon(t)\right\} \sum_{t=s}^{T} \mathbb{1}\left\{A_{t+1}=k, N_k(t-m)=s\right\}\right] \\
&\leq K_k(T) + m \sum_{s \geq K_k(T)} \mathbb{P}\left(\tau_m s d_{\mathrm{Pois}}(\hat{\theta}_{k,s}, \theta_1) \leq \beta_\epsilon(t)\right) \\
&\leq K_k(T) + \frac{m C_2(\eta)}{T^{f(\eta)}},
\end{aligned}$$

where the last inequality comes from the fact that for all $s \in \{1, \dots, T\}$, $\sum_{t=1}^{T} \mathbb{1}\left\{A_{t+1}=k, N_k(t-m)=s\right\} \leq m$ and from the proof of Fact 2 for exponential family bandits in [2] that proves the existence of the constants $C_2(\eta)$ and $f(\eta)$ that achieve the bound.

We can now upper bound the regret thanks to the decomposition provided by equation 3:

$$L_{\text{KLUCB}}(T) \leq m + \sum_{k>1} \tau_m \Delta_k \mathbb{E}\left[N_k(T)\right] \leq (1+\eta)\beta_\epsilon(t) \sum_{k=2}^{K} \frac{\tau_m \Delta_k}{d_{\mathrm{Pois}}(\tau_m \theta_k, \tau_m \theta_1)} + o(\log(T)).$$

To obtain the final result, we use Lemma 8 that shows that for $\theta_k < \theta_1$, $d_{\mathrm{Pois}}(\tau_m \theta_k, \tau_m \theta_1) > (1 - \tau_m \theta_1) d(\tau_m \theta_k, \tau_m \theta_1)$. Thus,

$$L_{\text{KLUCB}}(T) \leq m + (1+\eta) \frac{\beta_\epsilon(t)}{1 - \tau_m \theta_1} \sum_{k=2}^{K} \frac{\tau_m \Delta_k}{d(\tau_m \theta_k, \tau_m \theta_1)} + o(\log(T)).$$

D Additionnal experiments on delay agnostic policies

As a last additional contribution to this work, we suggest a distribution-agnostic heuristic that estimates the CDF parameters $(\tau_d)_{d \geq 0}$ in an online fashion. Indeed, as the delay distribution is assumed to be shared between actions, each observed reward provides an information on the delays that can be exploited to estimate the CDF without having to deal with the exploration-exploitation dilemma.

Uncensored setting. In the Uncensored setting and under the geometric assumption on the distribution of the delays, the entire CDF can be retrieved using an estimate of the unique parameter $\lambda = 1/\mu$. To this aim, we build an estimate the expected delay at round t, $\hat{\mu}(t)$, using a stochastic approximation process with decreasing weights $\alpha_t = 1/t^\gamma$ for $1 \geq \gamma \geq 0.5$. When an observation D_t arrives we update

$$\hat{\mu}(t) \leftarrow (1 - \alpha_t)\hat{\mu}(t) + \alpha_t D_t.$$

Then we use this estimator as a plug-in quantity to compute $O_k(t)$ defined in (6) for all k.

Censored setting. In the Censored setting however, no observation comes after the threshold m and this does not allow us to directly estimate the expected delay μ as the longest observations are censored. To circumvent this problem, we choose to estimate biased parameters for $\tau_1, \dots, \tau_m$. Concretely, we initialize counts for the observed delay values $\delta_0 = (0, \dots, 0) \in \mathbb{N}^{m+1}$ (delay can be null). Then, after each observation D_t, we increment all the counts δ_s for $s \geq D_t$. Then, the biased empirical CDF is obtained by normalizing those counts by the total number of observations received up to time t, $n_d(t)$. We emphasize that the obtained estimators are biased: For each

$s \in \{0, \dots, m\}, \mathbb{E}[\delta_s(t)/n_d(t)] = \tau_s/\tau_m$ as all observed delays are smaller or equal to m. Thus, plugging those estimates in $\tilde{N}_k(t)$ actually allows to have an estimate of $\tilde{N}_k(t)/\tau_m$ instead of $\tilde{N}_k(t)$ and consequently an estimate of $\tau_m\theta_k$ instead of θ_k.

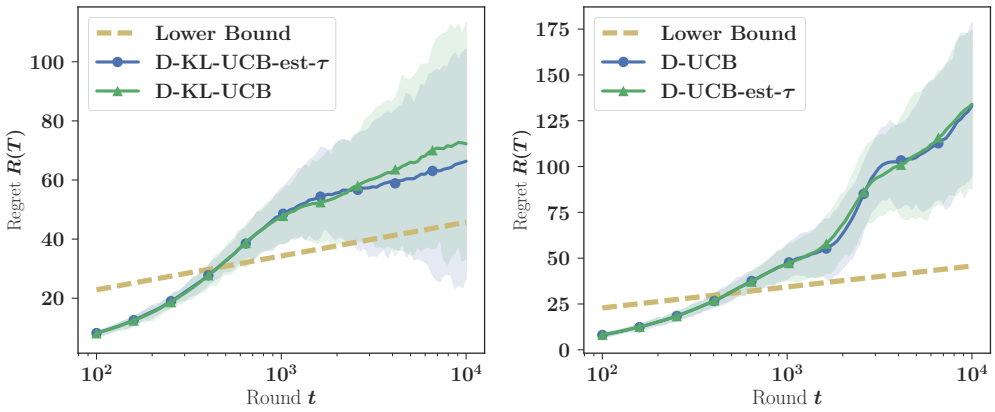

Figure 3: Expected regret of DelayedKLUCB with and without online estimation of the CDF in both the censored and uncensored setting. Results are averaged over 100 independent runs.

Figure 3 compares both our policies to its equivalent, delay-agnostic heuristic using the same confidence intervals with plug-in estimates of the $(\tau_d)_{d\geq 0}$. It is clear from these experiments that using delay parameters estimated on-the-go does not hurt the cumulated regret overall.