Stat 260/CS 294-102. Learning in Sequential Decision Problems.

Peter Bartlett

1. Multi-armed bandit algorithms.
    - Exponential families.
        - Cumulant generating function.
        - KL-divergence.
    - KL-UCB for an exponential family.
    - KL vs c.g.f. bounds.
        - Bounded rewards: Bernoulli and Hoeffding.
    - Empirical KL-UCB.

See (Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos and Gilles Stoltz, 2013)

Recall: Concentration inequalities.

Definition: Cumulant-generating function:

$$\Gamma_X(\lambda) = \log \mathbb{E} \exp(\lambda X),$$

We consider upper bounds $\psi : \mathbb{R} \to \mathbb{R}$, satisfying $\psi(\lambda) \geq \Gamma_X(\lambda)$.
The *Legendre transform (convex conjugate)* of ψ is

$$\psi^*(\epsilon) = \sup_{\lambda \in \mathbb{R}} \left(\lambda\epsilon - \psi(\lambda)\right).$$

Theorem: For $\epsilon \geq 0$, $\mathbb{P}\left(X - \mathbb{E}X \geq \epsilon\right) \leq \exp\left(-\psi^*_{X-\mathbb{E}X}(\epsilon)\right)$.

Recall: Concentration Inequalities.

Theorem: If $X_1, X_2, \ldots, X_n$ are mean zero, i.i.d. with cgf upper bound ψ, then $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ satisfies

$$\mathbb{P}\left(\bar{X}_n \geq \epsilon\right) \leq \exp\left(-n\psi^*(\epsilon)\right),$$

And the exponent can't be improved.

Theorem: **[Cramér-Chernoff]** If $X_1, X_2, \ldots, X_n$ are iid and mean zero, and have cgf Γ, then for $\epsilon > 0$ and $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left(\bar{X}_n \geq \epsilon\right) = -\Gamma^*(\epsilon).$$

(Γ^* sometimes called *Cramér function*. Lower bound is a change-of-measure argument plus central limit theorem.)

Outline.

For an exponential family, we can compute the c.g.f. exactly. Its convex conjugate corresponds to a KL-divergence. For reward distributions from the exponential family, concentration inequalities involving the KL-divergence define an upper confidence bound strategy: KL-UCB.

If the reward distributions are bounded, the c.g.f. of a particular exponential family (a scaled, shifted Bernoulli) gives a bound on the c.g.f. And we can bound this, in turn, with a quadratic (like Hoeffding's inequality), which corresponds to another exponential family (a Gaussian). KL-UCB for Bernoulli improves on KL-UCB for Gaussian. (KL-UCB for Gaussian corresponds to the original UCB strategy.)

There's also a non-parametric version of KL-UCB (called empirical KL-UCB) for bounded rewards. It works with the set of distributions with finite support.

Exponential families.

Definition: Canonical exponential family defined wrt measure P:

$$
\begin{aligned}
\frac{dP_\theta}{dP}(x) &= \exp\left(\theta x - A(\theta)\right), \\
A(\theta) &= \log\left(\int \exp\left(\theta x\right) dP(x)\right), \\
\theta \in \Theta &= \{\theta : A(\theta) < \infty\}.
\end{aligned}
$$

Exponential families.

$$\mu(\theta) := \mathbb{E}_\theta X = A'(\theta).$$

$\theta(\mu)$ defined on $\mu(\Theta)$. (one-to-one because $\mathrm{Var}_\theta X = A''(\theta) > 0$).

$$\Gamma_\theta(\lambda) = A(\theta + \lambda) - A(\theta),$$

$$\Gamma^*_{\theta_1}(\mu(\theta_2)) = A(\theta_1) - A(\theta_2) + \mu(\theta_2)(\theta_2 - \theta_1),$$

$$D_{KL}(P_{\theta_1}, P_{\theta_2}) = \Gamma^*_{\theta_2}(\mu(\theta_1)).$$

Exponential families.

$$A'(\theta) = \frac{\int x \exp(\theta x)\, dP(x)}{\exp(A(\theta))}$$

$$= \mathbb{E}_\theta X.$$

$$\Gamma_\theta(\lambda) = \log\left(\int \exp(\lambda x + \theta x - A(\theta)) dP(x)\right)$$

$$= \log\left(\int \exp\left((\lambda + \theta)x\right) dP(x)\right) - A(\theta)$$

$$= A(\theta + \lambda) - A(\theta).$$

Exponential families.

$$\Gamma^*_{\theta_1}(\mu(\theta_2)) = \sup_\lambda \left(\lambda\mu(\theta_2) - (A(\theta_1 + \lambda) - A(\theta_1))\right)$$

Maximum has $\mu(\theta_2) = \mu(\theta_1 + \lambda),$

that is, $\lambda = \theta_2 - \theta_1,$

so $\Gamma^*_{\theta_1}(\mu(\theta_2)) = (\theta_2 - \theta_1)\mu(\theta_2) + A(\theta_1) - A(\theta_2).$

Exponential families.

$$
\begin{aligned}
D_{KL}(P_{\theta_1}, P_{\theta_2}) &= \int \log \frac{dP_{\theta_1}}{dP_{\theta_2}} dP_{\theta_1} \\
&= \int ((\theta_1 - \theta_2)x) \exp(\theta_1 x - A(\theta_1))\, dP(x) \\
&\quad + A(\theta_2) - A(\theta_1) \\
&= \mu(\theta_1)(\theta_1 - \theta_2) + A(\theta_2) - A(\theta_1) \\
&= \Gamma^*_{\theta_2}(\mu(\theta_1))
\end{aligned}
$$

Exponential families.

Example: Bernoulli:

$$P_\theta(x) = \exp\left(\theta x - A(\theta)\right), \qquad A(\theta) = \log\left(1 + e^\theta\right),$$

$$\mu(\theta) = P_\theta(1) = \frac{e^\theta}{1 + e^\theta}, \qquad \theta = \log\frac{\mu}{1-\mu},$$

$$\Gamma_\theta(\lambda) = \log\left(1 - \mu(\theta) + \mu(\theta)e^\lambda\right), \qquad \Theta = \mathbb{R}.$$

Exponential families.

Example: Bernoulli:

$$\Gamma^*_{\theta_1}(\mu_2) = \sup_\lambda \left(\lambda\mu_2 - \log\left(1 - \mu_1 + \mu_1 e^\lambda\right)\right)$$

Maximum has $$\mu_2 = \frac{\mu_1 e^\lambda}{1 - \mu_1 + \mu_1 e^\lambda},$$

that is, $$\lambda = \log \frac{\mu_2(1-\mu_1)}{\mu_1(1-\mu_2)}$$

so $$\begin{aligned}\Gamma^*_{\theta_1}(\mu_2) &= \mu_2 \log \frac{\mu_2}{\mu_1} + (1-\mu_2)\log\frac{1-\mu_2}{1-\mu_1} \\ &= D_{KL}(P_{\theta_2}, P_{\theta_1}).\end{aligned}$$

KL-UCB for exponential families: Use $\psi = \Gamma$

Define the sample averages

$$\hat{\mu}_j(t) = \frac{1}{T_j(t)} \sum_{s=1}^{t} X_{I_s,s} 1[I_s = j], \qquad \hat{\mu}_{j,t} = \frac{1}{t} \sum_{s=1}^{t} X_{j,s}.$$

If $X_{j,s}$ has mean μ and c.g.f. Γ_μ, and $a < \mu$,

$$\Pr\left(\hat{\mu}_{j,n} \le a\right) \le e^{-n\Gamma^*(a)},$$

that is,

$$\Pr\left(\hat{\mu}_{j,n} < \mu \text{ and } \Gamma_\mu^*\left(\hat{\mu}_{j,n}\right) \ge \frac{f(n)}{n}\right) \le e^{-f(n)},$$

or

$$\Pr\left(\hat{\mu}_{j,n} < \mu \text{ and } D_{KL}\left(P_{\hat{\mu}_{j,n}}, P_\mu\right) \ge \frac{f(n)}{n}\right) \le e^{-f(n)}.$$

(Note that P_μ denotes $P_{\theta(\mu)}$.)

KL-UCB for exponential families.

KL-UCB Strategy for an exponential family (P_μ denotes $P_{\theta(\mu)}$):

$$I_t = t \qquad \text{for } t = 1, \ldots, k,$$

$$I_t = \arg\max_{1 \le j \le k} \sup \Big\{ \mu(\theta) : \theta \in \Theta \text{ and } D_{KL}\left(P_{\hat{\mu}_j(t-1)}, P_\mu\right) \le \frac{f(t)}{T_j(t-1)} \Big\},$$

where $f(t) = \log t + 3 \log\log(t)$.

- Equivalent to UCB with $\psi = \Gamma_\mu$.

KL-UCB for exponential families.

We can think of $D_{KL}\left(P_{\hat{\mu}_{j,t-1}}, P_\mu\right)$ as a divergence defined in terms of means: for any $\hat{\mu}, \mu \in \mu(\Theta)$,

$$d(\hat{\mu}, \mu) = D_{KL}(P_{\hat{\mu}}, P_\mu) = (\theta(\hat{\mu}) - \theta(\mu))\,\hat{\mu} - A(\theta(\hat{\mu})) + A(\theta(\mu)).$$

Then $d(\hat{\mu}, \mu) = 0$ iff $\hat{\mu} = \mu$, d is strictly convex and differentiable. We can extend it to the closure of $\mu(\Theta)$, by taking limits, allowing infinite values, and setting $d(\mu, \mu) = 0$ at boundaries. (Consider, for example, $\hat{\mu} = 0$ for a Bernoulli.)

KL-UCB for exponential families.

Theorem: KL-UCB for an exponential family satisfies:

$$\mathbb{E}T_j(n) \leq \frac{\log n}{D_{KL}\left(P_{\mu_j}, P_{\mu^*}\right)} + O\left(\sqrt{\log n}\right).$$

And the leading term is optimal (including the constant).

KL-UCB for bounded rewards.

Theorem: For $X \in [0,1]$ with $\mathbb{E}X = \mu$, define $Y \sim \text{Bernoulli}(\mu)$. Then

$$\Gamma_X(\lambda) \le \Gamma_Y(\lambda).$$

Notice that this gives a c.g.f. bound ψ_{X_μ} for X satisfying:

$$\psi^*_{X_\mu}(\mu') = \mu' \log \frac{\mu'}{\mu} + (1-\mu') \log \frac{1-\mu'}{1-\mu}.$$

KL-UCB for bounded rewards.

Proof: For $x \in [0,1]$, $\exp(\lambda x)$ lies below the line from $(0, e^0)$ to $(1, e^\lambda)$:

$$\exp(\lambda x) \leq x\left(e^\lambda - e^0\right) + e^0,$$

so

$$\begin{aligned} \mathbb{E}\exp(\lambda X) &\leq \mu\left(e^\lambda - 1\right) + 1 \\ &= \mathbb{E}\exp(\lambda Y). \end{aligned}$$

KL-UCB-Bernoulli for bounded rewards.

KL-UCB-Bernoulli Strategy For the Bernoulli family P_μ:

$$I_t = t \qquad \text{for } t = 1, \ldots, k,$$

$$I_t = \arg\max_{1 \le j \le k} \sup \Big\{ \mu \in (0,1) : d\left(\hat{\mu}_j(t-1), \mu\right) \le \frac{f(t)}{T_j(t-1)} \Big\},$$

where $d(\mu_1, \mu_2) = \mu_1 \log \frac{\mu_1}{\mu_2} + (1-\mu_1) \log \frac{1-\mu_1}{1-\mu_2}$ and $f(t) = \log t + 3 \log\log(t)$.

KL-UCB-Bernoulli for bounded rewards.

Theorem: KL-UCB-Bernoulli satisfies:

$$\mathbb{E}T_j(n) \leq \frac{\log n}{d(\mu_j, \mu^*)} + O\left(\sqrt{\log n}\right),$$

where $d(\mu_1, \mu_2) = \mu_1 \log \frac{\mu_1}{\mu_2} + (1 - \mu_1) \log \frac{1-\mu_1}{1-\mu_2}$.

The leading term is optimal for Bernoulli rewards, but might not be optimal, for example, if the variance is lower than $\mu(1 - \mu)$.

KL-UCB: More concentration inequalities.

Now, Pinsker's inequality gives

$$\psi^*_{X_\mu}(\mu') = D_{KL}(\mu', \mu) = \mu' \log \frac{\mu'}{\mu} + (1 - \mu') \log \frac{1 - \mu'}{1 - \mu}$$
$$\geq 2(\mu' - \mu)^2.$$

which shows this is at least as good as Hoeffding's inequality:

$$\mathbb{P}\left(\bar{X}_n \geq \mu'\right) \leq \exp\left(-2n(\mu' - \mu)^2\right)$$
$$\mathbb{P}\left(\bar{X}_n \geq \mu + \epsilon\right) \leq \exp\left(-2n\epsilon^2\right).$$

Exponential families.

Example: Gaussian:

$$
\begin{aligned}
p_\theta(x) &= \frac{\exp(-x^2/(2\sigma^2))}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2}\right), \\
\theta &= \frac{\mu}{\sigma^2}, \\
\mu(\theta) &= \sigma^2\theta, \\
A(\theta) &= \frac{\sigma^2\theta^2}{2}, \\
\Gamma_\theta(\lambda) &= \frac{\sigma^2}{2}(\lambda+\theta)^2 - \frac{\theta^2\sigma^2}{2}, \\
\Gamma^*_{\theta_1}(\mu_2) &= \frac{1}{2\sigma^2}(\mu_2-\mu_1)^2.
\end{aligned}
$$

Exponential families.

With $\sigma^2 = 1/4$, Pinsker's inequality corresponds to Hoeffding's inequality.

So we can view the UCB strategy (based on Hoeffding's inequality), as a special case of KL-UCB, modeling the reward distributions from $[0, 1]$ as $\mathcal{N}(\mu, 1/4)$.

KL-UCB-Gaussian for bounded rewards.

KL-UCB-Gaussian Strategy For the Gaussian family P_μ:

$$I_t = t \qquad \text{for } t = 1, \ldots, k,$$

$$I_t = \arg\max_{1 \le j \le k} \sup \left\{ \mu \in (0,1) : \right.$$
$$\left. d\left(\hat{\mu}_j(t-1), \mu\right) \le \frac{f(t)}{T_j(t-1)} \right\},$$

where $f(t) = \log t + 3 \log\log(t)$ and $d(\mu_1, \mu_2) = 2(\mu_1 - \mu_2)^2$.

This is equivalent to the UCB strategy (based on Hoeffding) that we saw last time.

UCB for bounded rewards.

Theorem: UCB satisfies:

$$\mathbb{E}T_j(n) \leq \frac{\log n}{d(\mu_j, \mu^*)} + O\left(\sqrt{\log n}\right),$$

where $d(\mu_1, \mu_2) = 2(\mu_1 - \mu_2)^2$.

This result is weaker (because of Pinsker's inequality) than the result for KL-UCB-Bernoulli.

KL-UCB regret bounds: upper versus lower.

Denote the canonical exponential family defined wrt a measure m by $\mathcal{E}_m$:

$$\mathcal{E}_m = \left\{ P : \frac{dP}{dm}(x) = \exp\left(\theta x - A(\theta)\right), \text{ and } A(\theta) < \infty \right\},$$

where $$A(\theta) = \log\left(\int \exp\left(\theta x\right) dm(x) \right).$$

Write $P_{m,\theta}$ for the element of $\mathcal{E}_m$ with parameter θ, and $P_{m,\mu}$ for the element of $\mathcal{E}_m$ with mean μ (and there's a one-to-one map between θ and μ, so it's well-defined.) And define for $\mathcal{E}_m$ the relevant divergence as a function of expectations:

$$d_m(\mu, \mu') := D_{KL}\left(P_{m,\mu}, P_{m,\mu'}\right).$$

KL-UCB regret bounds: upper versus lower.

We have derived bounds on Γ_{P_j} in terms of $\Gamma_{P_{m,\mu_j}}$, for some exponential families $\mathcal{E}_m$. For instance, if we let $\mathcal{P}$ denote the set of distributions on $[0,1]$, and consider two exponential families, the Bernoulli (call it $\mathcal{E}_B$) and the Gaussian with variance $1/4$ (call it $\mathcal{E}_G$), then we have:
For all $P \in \mathcal{P}$ with $PX = \mu$, and all λ,

$$\Gamma_P(\lambda) \leq \Gamma_{P_{B,\mu}}(\lambda) \leq \Gamma_{P_{G,\mu}}(\lambda).$$

And this is equivalent to: for all μ',

$$\Gamma^*_P(\mu') \geq \Gamma^*_{P_{B,\mu}}(\mu') \geq \Gamma^*_{P_{G,\mu}}(\mu'),$$

that is,

$$\Gamma^*_P(\mu') \geq d_B(\mu', \mu) \geq d_G(\mu', \mu).$$

KL-UCB regret bounds: upper versus lower.

We have seen upper bounds on regret based on these inequalities of the form

$$\bar{R}_n \leq \sum_{j:\Delta_j>0} \Delta_j \left(\frac{\log n}{d_m(\mu_j, \mu^*)} + O\left(\sqrt{\log n}\right)\right).$$

And we've seen lower bounds that are (roughly) of the form

$$\bar{R}_n \geq \sum_{j:\Delta_j>0} \Delta_j \left(\frac{\log n}{D_{KL}\left(P_j, P_{j^*}\right)} + o(1)\right).$$

To understand the gap between the upper bounds and the lower bounds, we can consider the I-projection of $P_{j^*} \in \mathcal{E}_{P_{j^*}}$ on to $\{P : PX = \mu_j\}$.

KL-UCB regret bounds: upper versus lower.

Theorem: Fix a measure m and an exponential family $\mathcal{E}_m$. For all $Q \in \mathcal{E}_m$ and P with $PX = \mu$,

$$D_{KL}(P, Q) = D_{KL}(P, P_{m,\mu}) + D_{KL}(P_{m,\mu}, Q).$$

In particular,

$$\inf \{D_{KL}(P, Q) : PX = \mu\} = D_{KL}(P_{m,\mu}, Q).$$

We say that $P_{m,\mu}$ is the I-projection of $Q \in \mathcal{E}_m$ onto $\{P : PX = \mu\}$.

KL-UCB regret bounds: upper versus lower.

The negative KL-divergence

$$
\begin{aligned}
-D_{KL}(P,Q) &= -\int \log \frac{dP}{dQ} dP \\
&= \int \frac{dP}{dQ} \log \frac{dQ}{dP} dQ
\end{aligned}
$$

is also called the entropy of P (defined with respect to Q), $H_Q(P)$. So the result says that among all distributions satisfying the mean constraint $PX = \mu$, the one with maximum entropy (wrt any Q in $\mathcal{E}_m$) is $P_{m,\mu}$ in the exponential family $\mathcal{E}_m$.

KL-UCB regret bounds: upper versus lower.

Using this fact, we can see that

$$\begin{aligned} D_{KL}\left(P_j, P_{j^*}\right) &\geq \inf\left\{D_{KL}\left(P, P_{j^*}\right) : PX = \mu_j\right\} \\ &= D_{KL}\left(P_{P_{j^*},\mu_j}, P_{j^*}\right) \\ &= \Gamma^*_{P_{j^*}}\left(\mu_j\right) \qquad \text{(both distributions are in } \mathcal{E}_{P_{j^*}}\text{)} \\ &\geq \Gamma^*_{P_{m,\mu^*}}\left(\mu_j\right) \\ &= d_m(\mu_j, \mu^*), \end{aligned}$$

where $\mathcal{E}_m$ is one of the exponential families that give the upper bounds (Bernoulli or Gaussian).

KL-UCB regret bounds: upper versus lower.

So the upper bound might be loose because P_j is further from P_{j^*} than the I-projection of P_{j^*} on to $\{PX = \mu_j\}$ (i.e., because P_j is not in $\mathcal{E}_{P_{j^*}}$), or because $\Gamma_{P_{m,\mu^*}}$ is a loose upper bound on $\Gamma_{P_{j^*}}$ (i.e., because P_{j^*} is not in $\mathcal{E}_m$).

KL-UCB: Regret bounds.

The KL-UCB strategies choose $I_1 = 1, \ldots, I_k = k$, and then

$$I_{t+1} = \arg\max_{1 \le j \le k} U_j(t),$$

$$\text{where} \qquad U_j(t) = \sup\left\{\mu \in \mu(\Theta) \text{ s.t. } d(\hat{\mu}_j(t), \mu) \le \frac{f(t)}{T_j(t)}\right\}.$$

For a suboptimal arm j, we want to bound

$$\mathbb{E}T_j(n) = 1 + \sum_{t=k}^{n} \mathbb{P}\{I_{t+1} = j\}.$$

We might have $I_{t+1} = j$ if either $U_{j^*}(t)$ is not an upper bound on μ^* (for a suitable choice for $f(t)$, this has negligible probability), or it is an upper bound, but $U_j(t)$ is bigger (and so exceeds μ^*; this can't happen too often).

KL-UCB: Regret bounds.

$$\begin{aligned}
&\{I_{t+1} = j\} \\
&\subseteq \{\mu^* \geq U_{j^*}(t)\} \cup \{I_{t+1} = j \text{ and } \mu^* < U_{j^*}(t) \leq U_j(t)\} \\
&\subseteq \{\mu^* \geq U_{j^*}(t)\} \cup \{I_{t+1} = j \text{ and } \mu^* < U_j(t)\}.
\end{aligned}$$

Also,

$$\begin{aligned}
\{\mu^* < U_j(t)\} &= \left\{\mu^* < \sup\left\{\mu \in \mu(\Theta) \text{ s.t. } d(\hat{\mu}_j(t), \mu) \leq \frac{f(t)}{T_j(t)}\right\}\right\} \\
&\subseteq \left\{\hat{\mu}_j(t) \geq \mu^*_{f(t)/T_j(t)}\right\}, \\
&\subseteq \left\{\hat{\mu}_j(t) \geq \mu^*_{f(n)/T_j(t)}\right\},
\end{aligned}$$

where $\quad \mu^*_{f(n)/T_j(t)} := \min\left\{\mu : d(\mu, \mu^*) \leq \dfrac{f(n)}{T_j(t)}\right\}.$

KL-UCB: Regret bounds.

$$\mathbb{E}T_j(n) = 1 + \sum_{t=k}^{n-1} \mathbb{P}\{I_{t+1} = j\}.$$

And

$$\underbrace{\sum_{t=k}^{n-1} \mathbb{P}\{\mu^* \geq U_{j^*}(t)\}}_{\text{times upper bound violated}} \leq \cdots \leq 3 + 4e \log \log n.$$

KL-UCB: Regret bounds.

$$\sum_{t=k}^{n-1} \mathbb{P}\left\{ I_{t+1} = j \text{ and } \hat{\mu}_j(t) \geq \mu^*_{f(n)/T_j(t)} \right\}$$

$$= \sum_{t=k}^{n-1} \sum_{m=2}^{n-k+1} \mathbb{P}\left\{ \hat{\mu}_{j,m-1} \geq \mu^*_{f(n)/(m-1)} \text{ and } m\text{th } j \text{ at } t+1 \right\}$$

$$\leq \sum_{m=1}^{n-k} \mathbb{P}\left\{ \hat{\mu}_{j,m} \geq \mu^*_{f(n)/m} \right\}$$

$$\leq M + \sum_{m=M+1}^{n-k} \mathbb{P}\left\{ \hat{\mu}_{j,m} \geq \mu^*_{f(n)/m} \right\},$$

for $M = f(n)/d(\mu_j, \mu^*)$.

KL-UCB: Regret bounds.

$$\sum_{m=M+1}^{n-k} \mathbb{P}\left\{\hat{\mu}_{j,m} \geq \mu^*_{f(n)/m}\right\} \leq \sum_{m=M+1}^{n-k} \exp\left(-md\left(\mu^*_{f(n)/m}, \mu_j\right)\right)$$

$$\vdots$$

$$= O\left(\sqrt{f(n)}\right).$$

(Relate $d\left(\mu^*_{f(n)/m}, \mu_j\right)$ to $d(\mu_j, \mu^*)$, bound by integral, use Laplace's method.)

Empirical KL-UCB for rewards in $[0, 1]$.

Empirical KL-UCB Strategy:

$$I_t = t \qquad \text{for } t = 1, \ldots, k,$$

$$I_t = \arg\max_{1 \le j \le k} \sup \Big\{ \mathbb{E}_P X : |\operatorname{supp}(P)| < \infty, \\ D_{KL}\left(\hat{P}_j(t-1), P\right) \le \frac{f(t)}{T_j(t-1)} \Big\},$$

where $\hat{P}_j(t-1)$ is the empirical distribution of the $T_j(t-1)$ pulls of arm j up to time $t-1$, and $f(t) = \log t + 3 \log\log(t)$.

Empirical KL-UCB for rewards in $[0, 1]$.

It turns out that it's always a finite convex optimization:

$$\sup\left\{\mathbb{E}_P X : |\operatorname{supp}(P)| < \infty, D_{KL}\left(\hat{P}_j(t-1), P\right) \leq \gamma\right\}$$

$$= \sup\Big\{\mathbb{E}_P X : \operatorname{supp}(P) \subseteq \operatorname{supp}(\hat{P}_j(t-1)) \cup \{1\},$$

$$D_{KL}\left(\hat{P}_j(t-1), P\right) \leq \gamma\Big\}.$$

Empirical KL-UCB for rewards in $[0,1]$.

Empirical KL-UCB Strategy:

$$I_t = t \qquad \text{for } t = 1, \ldots, k,$$

$$I_t = \arg\max_{1 \le j \le k} \sup \Big\{ \mathbb{E}_P X : \operatorname{supp}(P) \subseteq \operatorname{supp}(\hat{P}_j(t-1)) \cup \{1\}, \\ D_{KL}\left(\hat{P}_j(t-1), P\right) \le \frac{f(t)}{T_j(t-1)} \Big\},$$

where $\hat{P}_j(t-1)$ is the empirical distribution of the $T_j(t-1)$ pulls of arm j up to time $t-1$, and $f(t) = \log t + 3 \log\log(t)$.

Empirical KL-UCB for rewards in $[0, 1]$.

Theorem: Empirical KL-UCB for rewards in $[0, 1]$ satisfies:

$$\mathbb{E}T_j(n) \leq \frac{\log n}{\inf\{D_{KL}(P_j, P) : PX > \mu^*\}} + O\left(\log^{4/5} n \log\log n\right),$$

provided $\mu_j > 0$ and $\mu^* < 1$.

The leading term is optimal (including the constant). But the remainder term is worse than in the parametric case.