The KL-UCB Algorithm for Bounded Stochastic Bandits and Beyond

Aurélien Garivier, joint work with Olivier Cappé

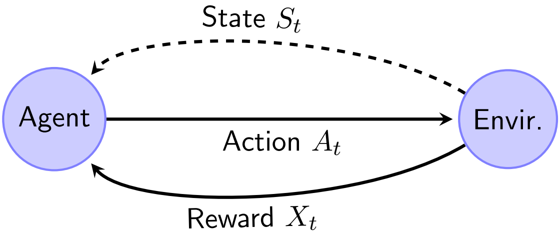
CNRS & Telecom ParisTech

July 11, 2011

Roadmap

1. The Model
2. Lower Bound for the Regret
3. Optimistic Algorithms
4. An Optimistic Algorithm based on Kullback-Leibler Divergence
5. Beyond Bounded Rewards
6. Conclusion

The Framework: Reinforcement Learning



Exploration
|
Exploitation

$\neq$ statistical learning (maximizing the reward is the main goal)
$\neq$ game theory (the environment is stochastic)

The (stochastic) Bernoulli Multi-Armed Bandit Model

Environment K arms with parameters $\theta = (\theta_1, \dots, \theta_K)$ such that for any possible choice of arm $a_t \in \{1, \dots, K\}$ at time t, one receives the reward

$$X_t = X_{a_t,t}$$

where, for any $1 \leq a \leq K$ and $s \geq 1$, $X_{a,s} \sim P_{\theta_a}$, and the $(X_{a,s})_{a,s}$ are independent.

Reward disbtributions can form a parametric family, or not. In this talk, we consider either general bounded rewards, or a canonical exponential family

Example Bernoulli rewards: $\theta \in [0,1]^K$, $X_{a,s} \sim \mathcal{B}(\theta_a)$

Strategy The agent's actions follow a dynamical strategy $\pi = (\pi_1, \pi_2, \dots)$ such that

$$A_t = \pi_t(X_1, \dots, X_{t-1})$$

Performance Evaluation, Regret

Cumulated Reward $S_n = \sum_{t=1}^{n} X_t$

Our goal Choose π so as to maximize

$$E\left[S_n\right] = \sum_{t=1}^{n}\sum_{a=1}^{K} \mathbb{E}\left[\mathbb{E}\left[X_t \mathbb{1}\{A_t = a\} | X_1, \dots, X_{t-1}\right]\right]$$
$$= \sum_{a=1}^{K} \mu_a \mathbb{E}\left[N_n^\pi(a)\right]$$

where $N_t^\pi(a) = \sum_{s \le t} \mathbb{1}\{A_s = a\}$ is the number of draws of arm a up to time n, and $\mu_a = \mathbb{E}\left[P_{\theta_a}\right]$.

Regret Minimization equivalent to minimizing

$$R_n(\theta) = n\mu^* - E\left[S_n\right] = \sum_{a:\mu_a<\mu^*} (\mu^* - \mu_a)\mathbb{E}\left[N_n^\pi(a)\right]$$

where $\mu^* \in \max\{\mu_a : 1 \le a \le K\}$

Roadmap

1. The Model
2. Lower Bound for the Regret
3. Optimistic Algorithms
4. An Optimistic Algorithm based on Kullback-Leibler Divergence
5. Beyond Bounded Rewards
6. Conclusion

Asymptotically Optimal Strategies

- A strategy π is said to be consistent if, for any $\theta \in [0,1]^K$,

$$\frac{1}{n}\mathbb{E}[S_n] \to \theta^*$$

- The strategy is efficient if for all $\theta \in [0,1]^K$ and all $\alpha > 0$,

$$R_n(\theta) = o(n^\alpha)$$

- There are efficient strategies and we consider the best achievable asymptotic performance among efficient strategies

The Bound of Lai & Robbins

Theorem [Lai&Robbins, '85]

If π is an efficient strategy, then , for any $\theta \in [0,1]^K$,

$$\liminf_{n\to\infty} \frac{R_n(\theta)}{\log(n)} \geq \sum_{a:\mu_a<\mu^*} \frac{\mu^* - \mu_a}{\mathrm{KL}(\theta_a, \theta^*)}$$

where $\mathrm{KL}(\theta, \theta')$ denotes the Kullback-Leibler divergence between the distributions P_θ and $P_{\theta'}$.

For example, in the Bernoulli case:

$$\mathrm{kl}(p,q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q}$$

The bound was generalized by Burnetas and Katehakis (1996)

Roadmap

1. The Model
2. Lower Bound for the Regret
3. Optimistic Algorithms
4. An Optimistic Algorithm based on Kullback-Leibler Divergence
5. Beyond Bounded Rewards
6. Conclusion

Optimism in the Face of Uncertainty

Optimism in an heuristic principle popularized by [Lai&Robins '85; Agrawal '95] which consists in letting the agent

> play as if the environment was the most favorable
> among all environments that are sufficiently likely
> given the observations accumulated so far

Surprisingly, this simple heuristic principle can be instantiated into algorithms that are robust, efficient and easy to implement in many scenarios pertaining to reinforcement learning

Upper Confidence Bound Strategies

UCB [Lai&Robins ’85; Auer&al ’02]

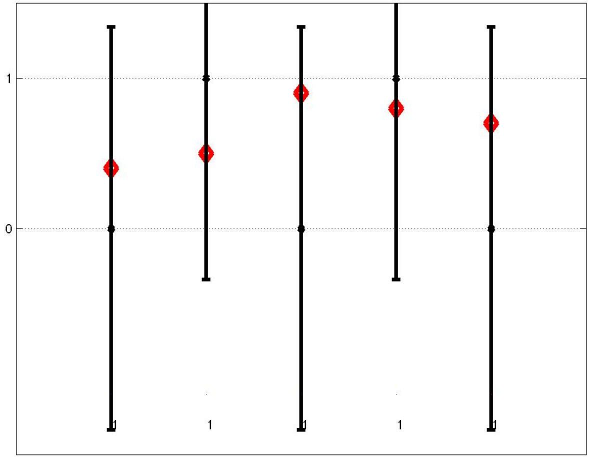
- Construct an upper confidence bound for the expected reward of each arm:

$$\underbrace{\frac{S_t(a)}{N_t(a)}}_{\text{estimated reward}} + \underbrace{\sqrt{\frac{\log(t)}{2N_t(a)}}}_{\text{exploration bonus}}$$

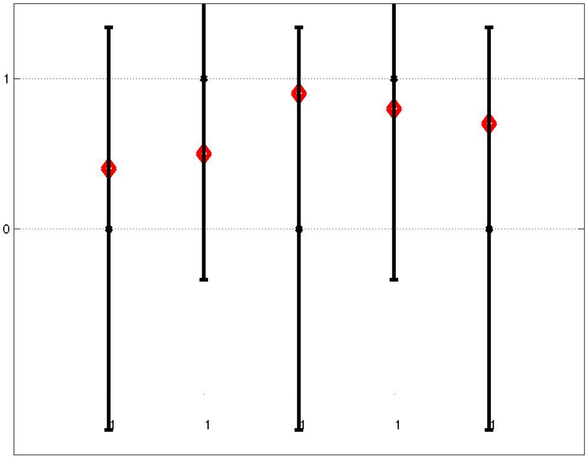
- Choose the arm with the highest UCB

- It is and *index strategy* [Gittins ’79]
- Its behavior is easily interpretable and intuitively appealing

UCB in Action



UCB in Action



Performance of UCB

For rewards in $[0, 1]$, the regret of UCB is upper-bounded as

$$E[R_n] = O(\log(n))$$

(finite-time regret bound) and

$$\limsup_{n\to\infty} \frac{\mathbb{E}[R_n]}{\log(n)} \leq \sum_{a:\mu_a<\mu^*} \frac{1}{2(\mu^* - \mu_a)}$$

Yet, in the case of Bernoulli variables, the rhs. is greater than suggested by the bound by Lai & Robbins

Many variants have been suggested to incorporate an estimate of the variance in the exploration bonus (e.g., [Audibert&al '07])

Roadmap

1 The Model

2 Lower Bound for the Regret

3 Optimistic Algorithms

4 An Optimistic Algorithm based on Kullback-Leibler Divergence

5 Beyond Bounded Rewards

6 Conclusion

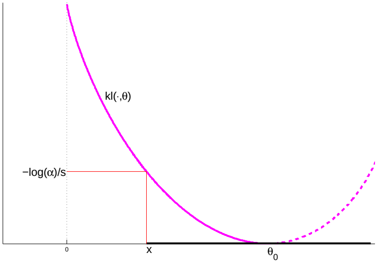
KL-UCB

```
Require: $n$ (horizon), $K$ (number of arms), REWARD (reward function, bounded in $[0,1]$)
for $t = 1$ to $K$ do
    $N[t] \leftarrow 1$
    $S[t] \leftarrow \text{REWARD}(\text{arm} = t)$
end for
for $t = K + 1$ to $n$ do
    $a \leftarrow \underset{1 \leq a \leq K}{\arg\max} \max_{q} \left\{ q \in \left[ \frac{S[a]}{N[a]}, 1 \right] : \mathrm{kl}\left( \frac{S[a]}{N[a]}, q \right) \leq \frac{\log(t)}{N[a]} \right\}$
    $r \leftarrow \text{REWARD}(\text{arm} = a)$
    $N[a] \leftarrow N[a] + 1$
    $S[a] \leftarrow S[a] + r$
end for
```

The KL Upper Confidence Bound in Picture

If $Z_1, \dots, Z_s \overset{iid}{\sim} \mathcal{B}(\theta_0)$, $x < \theta_0$ and if $\hat{p}_s = (Z_1 + \dots + Z_s)/s$, then

$$\mathbb{P}_{\theta_0}\left(\hat{p}_s \le x\right) \le \exp\left(-s\,\mathrm{kl}(x, \theta_0)\right)$$



In other words, if $\alpha = \exp\left(-s\,\mathrm{kl}(x, \theta_0)\right)$:

$$\mathbb{P}_{\theta_0}\left(\hat{p}_s \le x\right) = \mathbb{P}_{\theta_0}\left(\mathrm{kl}(\hat{p}_s, \theta_0) \le -\frac{\log(\alpha)}{s},\ \hat{p}_s < \theta\right) \le \alpha$$

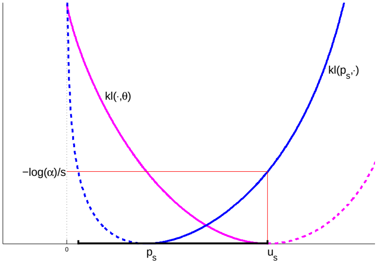
$\Longrightarrow$ upper confidence bound for p at risk α :

$$u_s = \sup\left\{\theta > \hat{p}_s : \mathrm{kl}(\hat{p}_s, \theta) \le -\frac{\log(\alpha)}{s}\right\}$$

The KL Upper Confidence Bound in Picture

If $Z_1, \ldots, Z_s \overset{iid}{\sim} \mathcal{B}(\theta_0)$, $x < \theta_0$ and if $\hat{p}_s = (Z_1 + \cdots + Z_s)/s$, then

$$\mathbb{P}_{\theta_0}\left(\hat{p}_s \leq x\right) \leq \exp\left(-s\,\mathrm{kl}(x, \theta_0)\right)$$



In other words, if $\alpha = \exp\left(-s\,\mathrm{kl}(x, \theta_0)\right)$:

$$\mathbb{P}_{\theta_0}\left(\hat{p}_s \leq x\right) = \mathbb{P}_{\theta_0}\left(\mathrm{kl}(\hat{p}_s, \theta_0) \leq -\frac{\log(\alpha)}{s},\ \hat{p}_s < \theta\right) \leq \alpha$$

$\Longrightarrow$ upper confidence bound for p at risk α :

$$u_s = \sup\left\{\theta > \hat{p}_s : \mathrm{kl}(\hat{p}_s, \theta) \leq -\frac{\log(\alpha)}{s}\right\}$$

Why focus on Bernoulli variables?

$\longrightarrow$ because they maximize deviations among bounded variables with given expectation:

Lemma

Let X denote a random variable such that $0 \leq X \leq 1$ and denote by $\mu = \mathbb{E}[X]$ its mean. Then, for any $\lambda \in \mathbb{R}$,

$$E\left[\exp(\lambda X)\right] \leq 1 - \mu + \mu \exp(\lambda) \ .$$

This fact is well-known for the variance, but also true for all exponential moments and thus for Cramer-type deviation bounds

Regret Bound for KL-UCB

Theorem

For all $\epsilon > 0$, there exist $C_1, C_2(\epsilon)$ and $\beta(\epsilon)$ such that

$$\mathbb{E}[N_n^{\text{KL-UCB}}(a)] \leq \frac{\log(n)}{\text{kl}(\mu_a, \mu^*)}(1+\epsilon) + C_1 \log(\log(n)) + \frac{C_2(\epsilon)}{n^{\beta(\epsilon)}}$$

Corollary

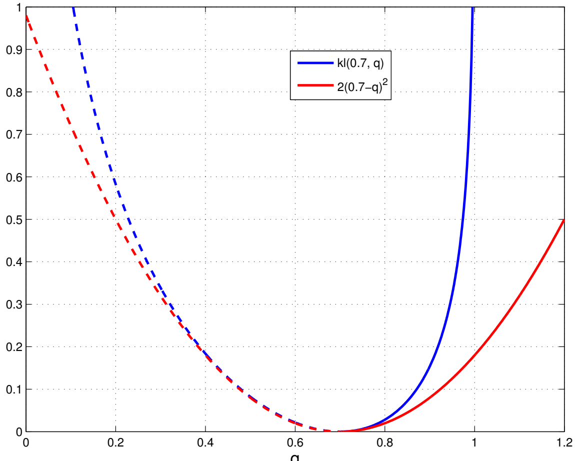
$$\limsup_{n\to\infty} \frac{R_n(\theta)}{\log(n)} \leq \sum_{a:\mu_a<\mu^*} \frac{\mu^* - \mu_a}{\text{kl}(\mu_a, \mu^*)}$$

KL-UCB satisfies an improved logarithmic finite-time regret bound
Besides, it is asymptotically optimal in the Bernoulli case

Comparison to UCB

KL-UCB addresses exactly the same problem as UCB, with the same generality, but it has always a smaller regret as can be seen from Pinsker's inequality

$$\mathrm{kl}(\mu_1, \mu_2) \geq 2(\mu_1 - \mu_2)^2$$



Main Tool: Deviation Inequality for Self-Normalized Sums

Theorem Let $(X_t)_t \geq 1$ be a sequence of independent random variables bounded in $[0, 1]$ and let $(\mathcal{F}_t)_{t>1}$ a collection of increasing sigma-fields such that $\forall t, \sigma(X_1 \ldots, X_t) \subset \mathcal{F}_t$ and, for $s > t$, X_s is independent of $\mathcal{F}_t$. Further assume that $(\epsilon_t)_{t\geq 1}$ is a $(\mathcal{F}_t)$-predictable sequence of Bernoulli random variables. Define, for $\delta > 0$,

$$S(n) = \sum_{s=1}^{n} \epsilon_s X_s \ , \qquad N(n) = \sum_{s=1}^{n} \epsilon_s \ , \qquad \hat{\theta}(n) = \frac{S(n)}{N(n)} \ ,$$

$$u(n) = \max_q \left\{ q > \hat{\theta}_n : N(n) \, \mathrm{kl}\left(\hat{\theta}(n), q \right) \leq \delta \right\} \ .$$

Then

$$\mathbb{P}\left(u(n) < \theta\right) \leq e \left\lceil \delta \log(n) \right\rceil \exp(-\delta)$$

$$\mathbb{P}\Big(N(n) \, \mathrm{kl}(\hat{\theta}(n), \theta) > \delta\Big) \leq 2e \left\lceil \delta \log(n) \right\rceil \exp(-\delta)$$

Results: Two-Arm Scenario

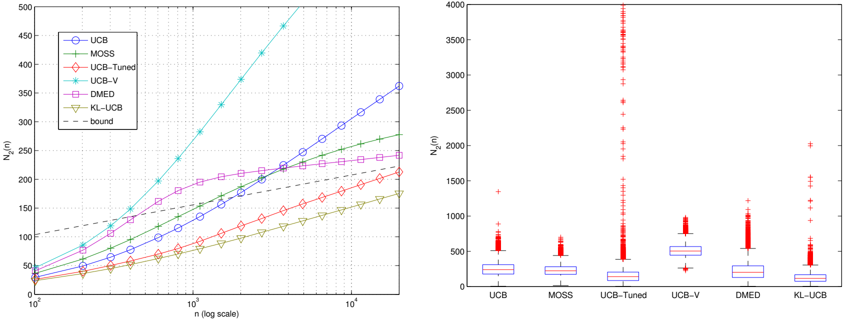


Figure: Performance of various algorithms when $\theta = (0.9, 0.8)$. Left: average number of draws of the sub-optimal arm as a function of time. Right: box-and-whiskers plot for the number of draws of the sub-optimal arm at time $n = 5,000$. Results based on $50,000$ independent replications

Results: Ten-Arm Scenario with Low Rewards

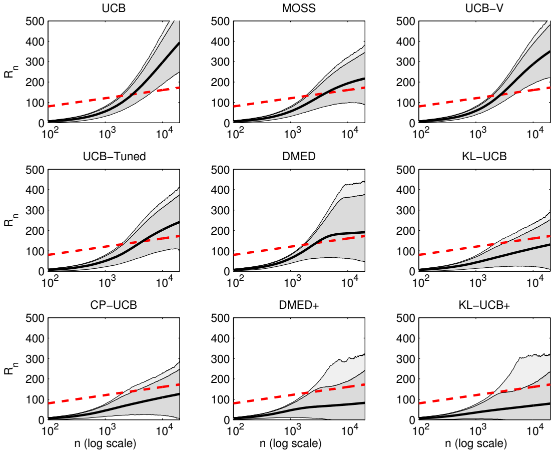


Figure: Average regret as a function of time when $\theta = (0.1, 0.05, 0.05, 0.05, 0.02, 0.02, 0.02, 0.01, 0.01, 0.01)$. Red line: Lai & Robbins lower bound; thick line: average regret; shaded regions: central 99% region an upper 99.95% quantile

Results: Truncated Exponentials

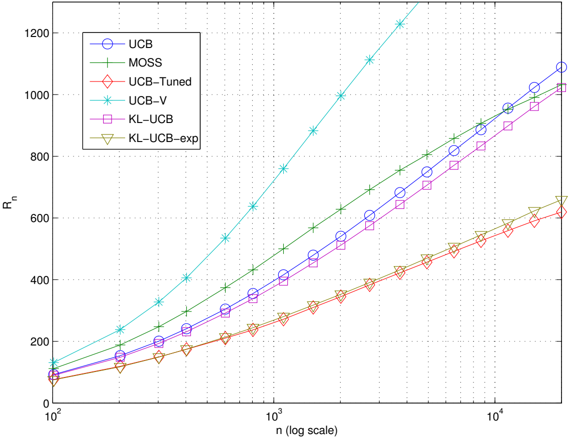


Figure: Average regret as a function of time for 5 exponentially distributed arms (parameters: $1/5$, $1/4$, $1/3$, $1/2$, 1) truncated at $X_{\max} = 10$.

Roadmap

1 The Model

2 Lower Bound for the Regret

3 Optimistic Algorithms

4 An Optimistic Algorithm based on Kullback-Leibler Divergence

5 Beyond Bounded Rewards

6 Conclusion

Exponential Family Rewards

- The method can be directly adapted for reward distributions that belongs to a *canonical exponential family*, i.e. such that the pdf of the rewards is given by

  $$p_{\theta_a}(x) = \exp\big(x\theta_a - b(\theta_a) + c(x)\big), \quad 1 \leq a \leq K$$

  for a parameter $\theta \in \mathbb{R}^K$
- The algorithm is the same: only use KL instead of kl.
- For instance, for exponential rewards $p_{\theta_a}(x) = \theta_a \mathrm{e}^{-\theta_a x}$:

  $$\mathrm{KL}(u, v) = u - v + u \log \frac{u}{v}$$
- Incorporating this change, one obtains analog deviation and regret bounds (with an identical proof)

Results: Exponential Rewards

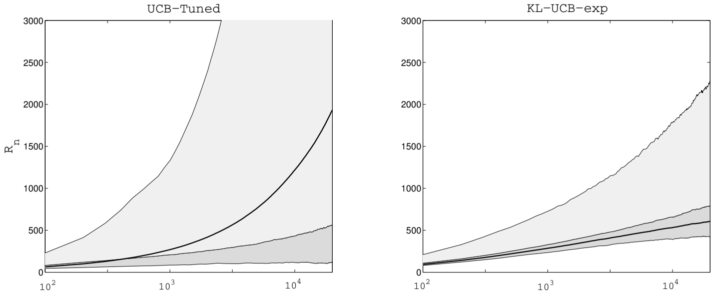


Figure: Average regret as a function of time for 5 exponentially distributed arms. Solid bold curve: mean regret; dark and light shaded regions: central 20% region and upper 1% quantile, respectively

Roadmap

1 The Model

2 Lower Bound for the Regret

3 Optimistic Algorithms

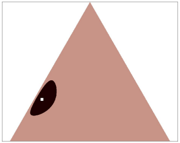
4 An Optimistic Algorithm based on Kullback-Leibler Divergence

5 Beyond Bounded Rewards

6 Conclusion

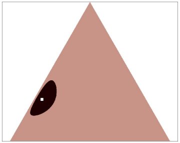
1. Use KL-UCB rather than UCB-1 or UCB-2
2. Can the method be adapted to other families of reward distributions? Listen to the next talk!
3. Can the KL-based deviation bounds be useful in other settings?

Used for model-based RL by Filippi *et al.*, *Optimism in Reinforcement Learning and Kullback-Leibler Divergence*, Allerton Conference, 2010



1. Use KL-UCB rather than UCB-1 or UCB-2
2. Can the method be adapted to other families of reward distributions? Listen to the next talk!
3. Can the KL-based deviation bounds be useful in other settings?

Used for model-based RL by Filippi *et al.*, *Optimism in Reinforcement Learning and Kullback-Leibler Divergence*, Allerton Conference, 2010



Thank you for your attention!