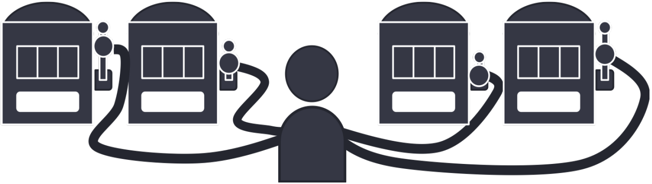


Balancing Optimism and Pessimism in Offline-to-Online learning

F. Sentenac, with I. Lee and C. Szepesvari

Leading question

Online Learning

Optimism principle dominates

+

Offline Learning

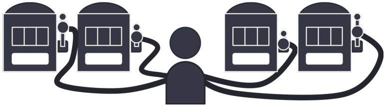
Pessimism principle dominates

How should we combine those two paradigms?

Multi-Armed Bandits: Setting

$$\text{Arms: } i \in [K], \quad r_i \sim \mathcal{B}(\mu_i), \quad \mu^* = \max_i \mu_i, \quad \Delta_i = \mu^* - \mu_i$$

$$\text{Horizon: } T, \quad a_t \in [K], \quad r_t \sim \mathcal{B}(\mu_{a_t})$$



- Regret: $R(T) = \sum_{t=1}^{T}(\mu^* - \mu_{a_t})$
- Minimax: $\Theta(\sqrt{KT})$, instance-dependent: $\Theta\left(\sum_{i:\Delta_i>0} \frac{\log T}{\Delta_i}\right)$

Algorithmic Families

- ϵ-Greedy — fixed or decaying ϵ
- Thompson Sampling — Bayesian posterior sampling (Agrawal & Goyal, 2012)
- Optimism in the Face of Uncertainty — exploration bonus (Auer et al., 2002; Auer & Ortner, 2010)

Optimism Principle

Select $i_t = \arg\max_i [\widehat{\mu}_i(t) + \text{bonus}_i(t)]$

E.g. $\text{bonus}_i(t) = \sqrt{\frac{\log(1/\delta)}{T_i(t)}}$

Offline Learning

Key Challenge: Data coverage—does the dataset sufficiently cover optimal or near-optimal policies?

- **Expert Data**: Generated by near-optimal policies; imitation learning achieves good performances(Ross et al., 2011; Rajaraman et al., 2020, Rashidinejad et al., 2023).
- **Uniform Data**: Covers policies broadly but requires algorithms to adapt to limited coverage (Cheng et al., 2022; Yin et al., 2020).

Pessimism Principle

Avoid under-explored areas

In Multi-Armed Bandits

- Offline sample size for arm i is m_i
- Total offline sample size is m

Algorithm 1: Lower Confidence Bound (LCB)

for $t = 1$ **do**

  Compute lower bound for reward of each arm i, $\hat{\mu}_i - \sqrt{\frac{\log(1/\delta)}{m_i}}$;

  Choose arm with highest lower bound;

end

	LCB	UCB
Minimax regret	$\sqrt{\frac{1}{\min_i m_i}}$ Optimal (ignoring poly-log factors)	

Regret wrt the logging policy

Define the reward of the logging policy:

$$\mu_0 = \frac{1}{m} \sum_i m_i \mu_i .$$

Regret wrt the logging policy for trajectory $(I(t))_{t=1}^{T}$: $R(T) = \sum_{t=1}^{T} \mu_0 - \mu_{I(t)}$.

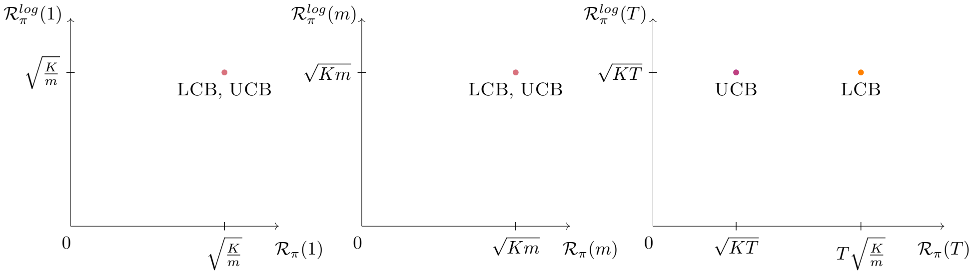
	LCB	UCB
Regret against logging policy	$\frac{\sum_i \sqrt{m_i}}{m}$ (UB)	$\sqrt{\frac{1}{\min_i m_i}}$ (LB)

What about offline-to-online learning?

Literature review

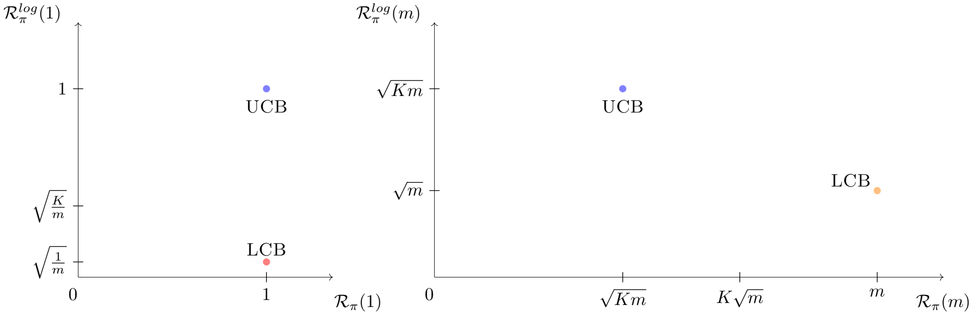
- How does enriching online methods with offline data impact the regret?
  - In MAB, a logarithmic amount of data is enough to get constant regret (Shivaswamy et al., 2012)
  - Results of a similar flavor in more general settings (Gur et al., 2020, Bu et al., 2021)
- How to reduce sample or computational complexity in Hybrid RL? (Song et al., 2022; Xie et al., 2022b; Ball et al., 2023; Wagenmaker and Pacchiano, 2023; Li et al., 2023, 2024; Zhou et al., 2023)

Pessimism vs. Optimism in Offline-to-Online learning



Evolution of the performance of the algorithms when the offline data is perfectly balanced between the arms

Pessimism vs. Optimism in Offline-to-Online learning



Evolution of the performance of the algorithms when the offline data is highly skewed (only 2 arms sampled)

Algoritm OTO

Algorithm Design

At each round, the algorithm computes an exploration budget.

- If the exploration budget is high enough, play UCB.
- If the exploration budget is not high enough, play safe option, i.e., LCB.

The algorithm design is inspired by *conservative bandits* (Wu et al.,2016).

Exploration Budget Computation

A few definitions:

- The benchmark:

$$\gamma = \underline{\mu}_{L(0)}(0) - \alpha\beta,$$

  where $\beta = \frac{\sum_i \sqrt{m_i}}{m}\sqrt{2\log\left(\frac{K}{\delta}\right)}$ and α is a tunable parameter.
- $T_i^U(t)$: Number of times arm i was played by UCB.
- $T^L(t)$: Total number of times LCB has been played up to time t.

Exploration Budget:

$$B_T(t) = \sum_{i=1}^{K} T_i^U(t-1)(\underline{\mu}_i(t) - \gamma) + \underline{\mu}_{U(t)}(t) - \gamma + (T^L(t-1) + T - t)\alpha\beta.$$

Breakdown of the Exploration Budget

Exploration Budget:

$$B_T(t) = \sum_{i=1}^{K} T_i^U(t-1)(\underline{\mu_i}(t) - \gamma) + \underline{\mu_{U(t)}}(t) - \gamma + (T^L(t-1) + T - t)\alpha\beta$$

- First term: lower bound on reward cumulated above benchmark by UCB steps.
- Second term: lower bound on reward above benchmark UCB could get at iteration t.
- Last term: when LCB is played, the reward exceeds the benchmark by at least $\alpha\beta$.

Breakdown of the Exploration Budget

Exploration Budget:

$$B_T(t) = \sum_{i=1}^{K} T_i^U(t-1)(\underline{\mu_i}(t) - \gamma) + \underline{\mu_{U(t)}}(t) - \gamma + (T^L(t-1) + T - t)\alpha\beta$$

- First term: lower bound on reward cumulated above benchmark by UCB steps.
- Second term: lower bound on reward above benchmark UCB could get at iteration t.
- Last term: when LCB is played, the reward exceeds the benchmark by at least $\alpha\beta$.

Breakdown of the Exploration Budget

Exploration Budget:

$$B_T(t) = \sum_{i=1}^{K} T_i^U(t-1)(\underline{\mu_i}(t) - \gamma) + \underline{\mu_{U(t)}}(t) - \gamma + (T^L(t-1) + T - t)\alpha\beta$$

- First term: lower bound on reward cumulated above benchmark by UCB steps.
- Second term: lower bound on reward above benchmark UCB could get at iteration t.
- Last term: when LCB is played, the reward exceeds the benchmark by at least $\alpha\beta$.

Breakdown of the Exploration Budget

Exploration Budget:

$$B_T(t) = \sum_{i=1}^{K} T_i^U(t-1)(\underline{\mu_i}(t) - \gamma) + \underline{\mu_{U(t)}}(t) - \gamma + (T^L(t-1) + T - t)\alpha\beta$$

- First term: lower bound on reward cumulated above benchmark by UCB steps.
- Second term: lower bound on reward above benchmark UCB could get at iteration t.
- Last term: when LCB is played, the reward exceeds the benchmark by at least $\alpha\beta$.
- Last part of the last term: lower bound how much budget you would obtain by playing it safe at every iteration.

Regret Bounds

Theorem

On any instance, with δ the parameter for the confidence intervals, with probability at least $1 - 2T\delta$, OTO *has:*

$$R^{log}(T) \leq T\,(1 + \alpha)\,\beta$$

Elements of proof:

- By design, the budget is positive at the end of the horizon,
- A positive budget implies that the total cumulated reward exceeds the benchmark,
- The benchmark is a discounted UB on the reward of the logging policy.

Regret Bounds

Theorem

On any instance, with probability at least $1 - 2T\delta$*:*

$$R(T) \leq \sum_{i=1}^{K} \Delta_i \left(\frac{4 \log(K/\delta)}{\Delta_i^2} - m_i \right)_+ + \frac{12K \log(K/\delta)}{\alpha\beta} + K.$$

We also have:

$$\mathcal{R}(T) \leq \max_{J \subseteq [K]} 2T \sqrt{\frac{2|J| \log(K/\delta)}{T + \sum_{j \in J} m_j}} + |J| + \frac{12K \log(K/\delta)}{\alpha\beta} + 2T^2\delta.$$

Elements of proof

Regret is split in two parts:

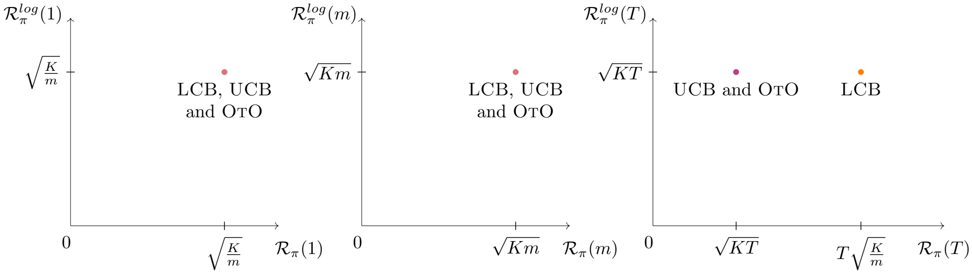
Regret of steps where UCB is played:

- The UB of suboptimal arms exceed UB of optimal arm for a limited number of iteration,
- Gives first part of the regret, *exactly* the same as the UB we have for the regret of UCB.

Regret of steps where LCB is played:

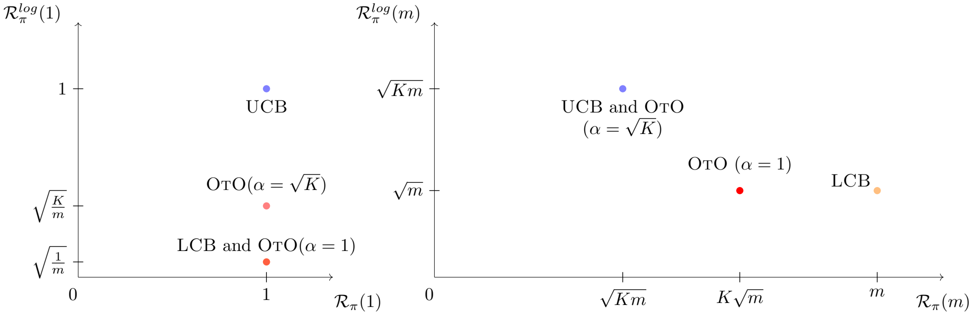
- The budget becomes negative only when suboptimal arms have been pulled by UCB,
- By proof on the left, we can bound the cost of those pulls,
- Each play LCB augments budget by $\alpha\beta$,
- This gives an upper bound on the total number of plays of LCB.

Comparison with LCB and UCB



Evolution of the performance of the algorithms when the offline data is perfectly balanced between the arms

Pessimism vs. Optimism in Offline-to-Online learning



Evolution of the performance of the algorithms when the offline data is highly skewed (only 2 arms sampled)

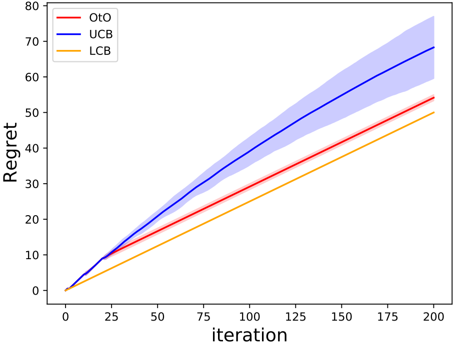
Dealing with unknown horizon

- Usual tricks for confidence interval construction
- Use a doubling horizon for the last part of the last term

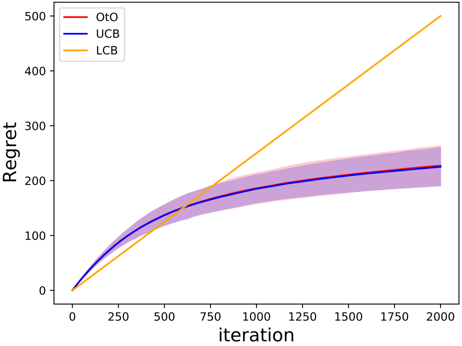
$$\sum_{i=1}^{K} T_i^U(t-1)(\underline{\mu_i}(t)-\gamma)+\underline{\mu_{U(t)}}(t)-\gamma+(T^L(t-1)+T-t)\alpha\beta$$

Experiments

Optimal arm not sampled in the offline data

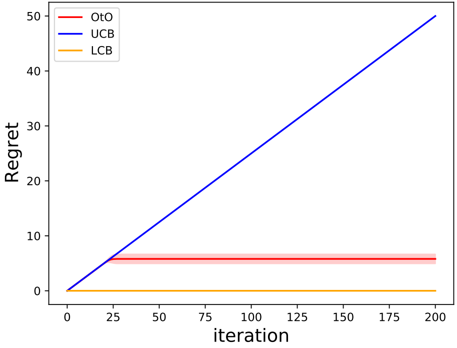


(a) $T = 200$

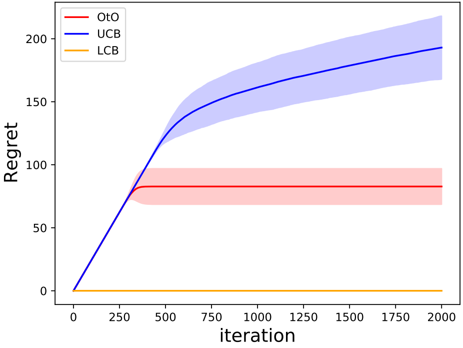


(b) $T = 2000$

Optimal arm sampled in the offline data

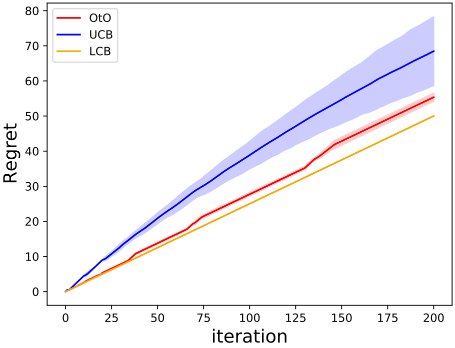


(a) $T = 200$

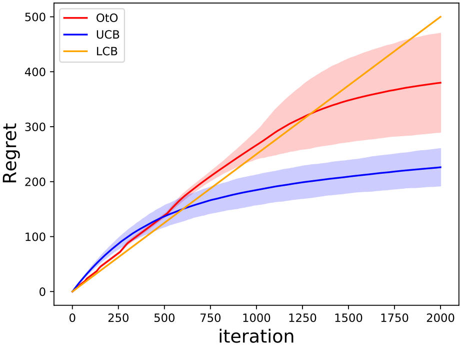


(b) $T = 2000$

Unknown horizon, Optimal arm not sampled in the offline data

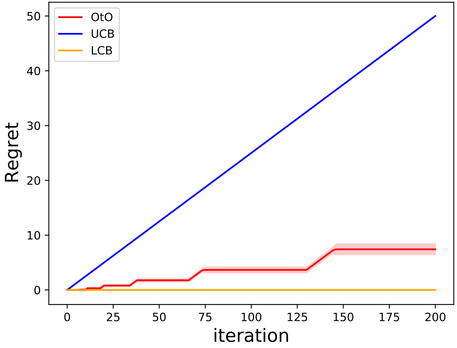


(a) $T = 200$

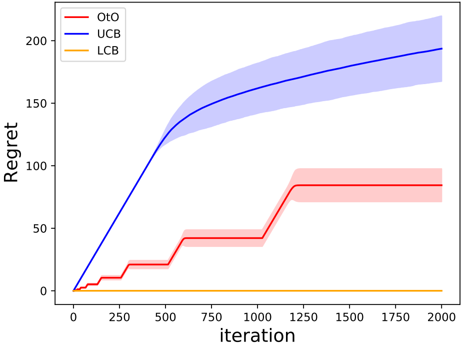


(b) $T = 2000$

Unknown horizon, Optimal arm sampled in the offline data



(a) $T = 200$



(b) $T = 2000$

Analysis of UCB and LCB in the Intermediate Setting

Minimax Regret

T	$T = 1$	$T = m$	$T \gg m$
$\mathcal{R}_{\mathrm{UCB}}(T)$	$\sqrt{\frac{1}{\min_i m_i}}$	$m\sqrt{\frac{K}{\sum_i m_i}}$	$\sqrt{KT}$
$\mathcal{R}_{\mathrm{LCB}}(T)$	$\sqrt{\frac{1}{\min_i m_i}}$	$m\sqrt{\frac{1}{\min_i m_i}}$	$T\sqrt{\frac{1}{\min_i m_i}}$

Table 1: Evolution of the pseudo regret of LCB and UCB as T grows (ignoring poly log terms, exact expressions in the Lemmas)

Lower bound on the Minimax Regret of any algorithm

Theorem

For any $T \geq 1$ and for any strategy π, we have:

$$\mathcal{R}_\pi(T) \geq \frac{1}{31} T \sqrt{\max_{J \subseteq [K]} \frac{|J|}{T - 1 + \sum_{j \in J} m_j}}.$$

The above bound may be hard to interpret. Notice it implies the two following looser bounds for any $T \geq 1$ and any strategy π:

$$\mathcal{R}_\pi(T) \geq \frac{1}{31} T \sqrt{\frac{(K-1)}{T - 1 + m - \max_i m_i}}, \text{ and } \mathcal{R}_\pi(T) \geq \frac{1}{31} T \sqrt{\frac{1}{T - 1 + \min_i m_i}}.$$

Regret of UCB

Theorem (UCB's upper bound on the minimax regret)

For any $T \geq 1$ and any $\theta \in \Theta$, with probability at leat $1 - 2T^2\delta$:

$$R(T) \leq \sum_{i=1}^{K} \Delta_i \left(\frac{2}{\Delta_i^2} \log(K/\delta) - m_i \right)_+ + \sum_{i=1}^{K} \Delta_i .$$

Also, we have the following instance-independent bound:

$$\mathcal{R}_{\mathrm{UCB}}(T) \leq \min \left(\max_{J \subseteq [K]} 2T \sqrt{\frac{2|J| \log(K/\delta)}{T + \sum_{j \in J} m_j}} + |J|; T \sqrt{\frac{2 \log(K/\delta)}{\min_i m_i}} \right) + 2T^2\delta .$$

Regret of LCB

Proposition

For $T \geq 1$, we have:

$$\min\left(0.07T, 0.15T\sqrt{\frac{1}{\min_i m_i}}\right) \leq \mathcal{R}_{\mathrm{LCB}}(T) \leq T\sqrt{\frac{2\log(K/\delta)}{\min_i m_i}} + 2T^2\delta.$$

Regret wrt the logging policy

T	1		$T = m$		$T \gg m$	
	LB	UB	LB	UB	LB	UB
$\mathcal{R}^{\mathrm{log}}_{\mathrm{UCB}}(T)$	$\sqrt{\frac{1}{\min_i m_i}}$		$\sum_{i=1}^{K} \left(\frac{m}{K} - m_i\right) \rho_i$	$\sqrt{KT}$	0	$\sqrt{KT}$
$\mathcal{R}^{\mathrm{log}}_{\mathrm{LCB}}(T)$	$\frac{\sqrt{m_2}}{m}$	$\frac{\sum_i \sqrt{m_i}}{m}$	$\sqrt{m_2}$	$\sum_{i=1}^{m} \sqrt{m_i}$	$T\frac{\sqrt{m_2}}{m}$	$T\frac{\sum_i \sqrt{m_i}}{m}$

Table 2: Evolution of the regrets against the logging policy as T grows (ignoring poly log terms), assuming wlog $m_1 \geq m_2 \geq \ldots \geq m_K$, and with $\rho_i = \left[\sqrt{\frac{1}{m_i + \frac{m}{K}}} - \sqrt{\frac{1}{m_1 + \frac{m}{K}}}\right]$.

Regret wrt the logging policy of LCB

Proposition

We have:

$$\mathcal{R}^{\log}_{\mathrm{LCB}}(T) \leq T\frac{\sum_i \sqrt{m_i}}{\sum_i m_i}\sqrt{2\log\left(\frac{K}{\delta}\right)} + 2T^2\delta.$$

If $m_1 = m$ and $m_i = 0$ for any $i > 1$, we obtain:

$$\mathcal{R}^{\log}_{\mathrm{LCB}}(T) \leq T\sqrt{\frac{2\log\left(\frac{K}{\delta}\right)}{m}} + 2T^2\delta.$$

If $m_i = \frac{m}{K}$ for all $i \in [K]$, we get:

$$\mathcal{R}^{\log}_{\mathrm{LCB}}(T) \leq T\sqrt{\frac{2K\log\left(\frac{K}{\delta}\right)}{m}} + 2T^2\delta.$$

Regret wrt the logging policy of UCB

Proposition

For any $T > 0$, $\frac{T}{K} \in \mathbb{N}$, we have

$$\mathcal{R}^{\log}_{\mathrm{UCB}}(T) \geq T \sum_{i=1}^{K} \left(\frac{1}{K} - \frac{m_i}{m} \right) \left[\sqrt{\frac{1}{2(m_i + \frac{T}{K})}} - \sqrt{\frac{1}{2(\max_{j \in [K]} m_j + \frac{T}{K})}} \right].$$

If $m_1 = m$ and $m_i = 0$ for any $i > 1$, we obtain:

$$\mathcal{R}^{\log}_{\mathrm{UCB}}(T) \geq \frac{1}{10} \sqrt{KT}.$$

If $m_i = \frac{m}{K}$ for all $i \in [K]$, we get:

$$\mathcal{R}^{\log}_{\mathrm{UCB}}(T) \geq 0.$$

Thank you for listening !