The Sliding Regret in Stochastic Bandits

Victor Boone

A Workshop in Grenoble

November, 2023

Multi-Armed Bandits

Sequential decision maker: $\{A_t : t \geq 1\}$

Pill a?
Pill b?

Mathematical model:

- Actions $\mathcal{A}$, reward distributions F_a with means μ_a;
- At time t, pick A_t and observe $R_t \sim \mathrm{F}_{A_t}$;
- Independance assumptions.

A performance metric: the regret

Optimal and suboptimal actions.

- Optimal arm achieving $\mu^* := \max_a \mu_a$;
- Optimal *offline* performance: $T\mu^*$.

Principle of regret.

Compare my performance to optimal offline performance:

$$\mathrm{Reg}(T) := T\mu^* - \sum_{t=1}^{T} \mu_{A_t}.$$

(*Remark:* $\sum_{t=1}^{T} \mu(A_t)$ *is a conditional expectation of* $\sum_{t=1}^{T} R_t$.)

Achievable Asymptotical Regret

Assumption. Rewards are Bernoulli, i.e., $\mathrm{F}_a \equiv \mathrm{B}(\mu_a)$.

Theorem (Lai and Robbins, 1985)
If an algorithm satisfies $\mathbf{E}_{\mathrm{F}}[\mathrm{Reg}(T)] = o(T^\epsilon)$ *for all* F *and* $\epsilon > 0$, *then:*

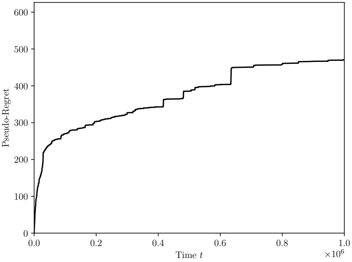
$$\forall \mathrm{F}, \quad \liminf_{T\to\infty} \frac{\mathbf{E}_{\mathrm{F}}[\mathrm{Reg}(T)]}{\log(T)} \geq \sum_{a:\mu_a<\mu^*} \frac{\mu^* - \mu_a}{\mathrm{kl}(\mu_a, \mu^*)}$$

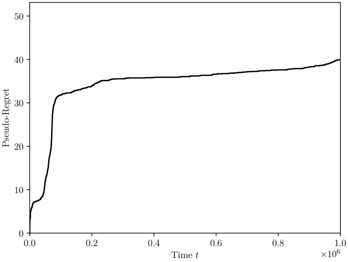
where $\mathrm{kl}(p,q) := p\log(\frac{p}{q}) + (1-p)\log(\frac{1-p}{1-q})$.

Many algorithms achieve this lower bound! Thompson Sampling, KL-UCB, IMED, MED, subsampling algorithms, and more.

Two Performance Portraits

- Fix $\mathrm{F} = (\mathrm{B}(0.85), \mathrm{B}(0.8))$;
- Run **once** every algorithm.

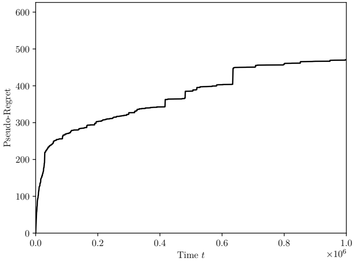


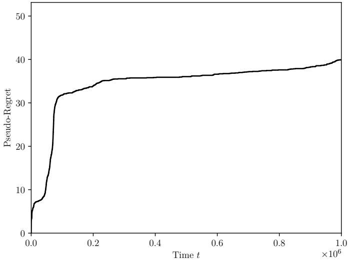


- Two families of regret trajectories.

Two Performance Portraits

- Fix $\mathrm{F} = (\mathrm{B}(0.85), \mathrm{B}(0.8))$;
- Run **once** every algorithm.





- Two families of regret trajectories.

$\boxed{\textbf{Can we explain this?}}$

The Sliding Regret

Definition. The *sliding regret* is

$$\mathrm{SliReg}(T) := \limsup_{t \to \infty} \left(T\mu^* - \sum_{i=0}^{T-1} \mu_{A_{t+i}} \right)$$

The Sliding Regret

Definition. The *sliding regret* is

$$\mathrm{SliReg}(T) := \limsup_{t\to\infty} \left(T\mu^* - \sum\nolimits_{i=0}^{T-1} \mu_{A_{t+i}} \right)$$

Index policies compute at time t, out of observations, an index $I_a(t)$ for every arm, and pick the arm maximizing the index.

Take away:

- Deterministic-index policies have linear sliding regret.
- Random-index policies have sub-linear sliding regret.

(*of course reality is subtler*)

Examples of Index Policies

- Number of visits, $N_a(t) := \sum_{i=1}^{t-1} \mathbf{1}(A_t = a)$
- Number of successes, $S_a(t) := \sum_{i=1}^{t-1} \mathbf{1}(A_t = a) R_t$
- Empirical means, $\hat{\mu}_a(t) := S_a(t)/N_a(t)$

UCB (optimism)

Use the deterministic index:

$$I_a(t) := \hat{\mu}_a(t) + \sqrt{\frac{2\log(t)}{N_a(t)}}$$

Worst $\mathrm{SliReg}(T) = (\mu_1 - \mu_2)T$,

Thompson Sampling

Use the randomized index:

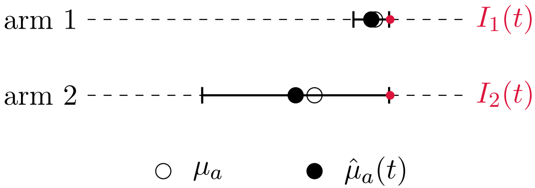
$$I_a(t) \sim \mathrm{Beta}(1+S_a(t), 1+U_a(t))$$

where $U_a(t) := N_a(t) - S_a(t)$.

Optimal $\mathrm{SliReg}(T) = \mu_1 - \mu_2$.

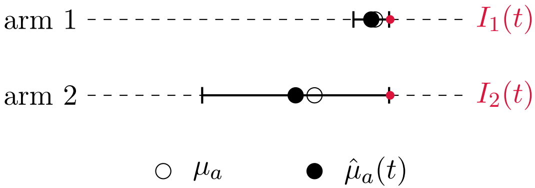
Why is the Sliding Regret of UCB Linear?

- This has to do with how UCB behaves at infinity.
- UCB index: $I_a(t) := \hat{\mu}_a(t) + \sqrt{\frac{2\log(t)}{N_a(t)}}$.



Why is the Sliding Regret of UCB Linear?

- This has to do with how UCB behaves at infinity.
- UCB index: $I_a(t) := \hat{\mu}_a(t) + \sqrt{\frac{2\log(t)}{N_a(t)}}$.



- Solving $I_2(t) = \mu_1$, we get: $\boxed{N_2(t) \approx \frac{2\log(t)}{(\mu_1-\mu_2)^2}}$ when $t \to \infty$.

Why is the Sliding Regret of UCB Linear?

- $I_a(t) := \hat{\mu}_a(t) + \sqrt{\frac{2\log(t)}{N_a(t)}}$ and $N_a(t) \approx \frac{2\log(t)}{(\mu^* - \mu_a)^2}$.
- **Scenario:** Say $A_t = a \neq a^*$ and $R_t = 1$.

$$\begin{aligned} I_a(t+1) &= I_a(t) + \mathrm{d}\left(\hat{\mu}_a(t)\right) + \mathrm{d}\left(\sqrt{\frac{2\log(t)}{N_a(t)}}\right) \\ &\approx I_a(t) + \frac{1-\mu_a}{N_a(t)} - \frac{\mu^* - \mu_a}{2N_a(t)} = I_a(t) + \frac{1 - \frac{\mu^* - \mu_a}{2}}{N_a(t)}. \end{aligned}$$

- **Conclusion:** I_a increases by $\Theta(\frac{1}{N_a(t)})$, I_{a^*} increases by $\mathrm{O}(\frac{1}{t})$.

[redacted]

Why is the Sliding Regret of UCB Linear?

- $I_a(t) := \hat{\mu}_a(t) + \sqrt{\frac{2\log(t)}{N_a(t)}}$ and $N_a(t) \approx \frac{2\log(t)}{(\mu^* - \mu_a)^2}$.
- **Scenario:** Say $A_t = a \neq a^*$ and $R_t = 1$.

$$I_a(t+1) = I_a(t) + \mathrm{d}\left(\hat{\mu}_a(t)\right) + \mathrm{d}\left(\sqrt{\frac{2\log(t)}{N_a(t)}}\right)$$

$$\approx I_a(t) + \frac{1-\mu_a}{N_a(t)} - \frac{\mu^* - \mu_a}{2N_a(t)} = I_a(t) + \frac{1 - \frac{\mu^* - \mu_a}{2}}{N_a(t)}.$$

- **Conclusion:** I_a increases by $\Theta(\frac{1}{N_a(t)})$, I_{a^*} increases by $\mathrm{O}(\frac{1}{t})$.

$\boxed{\textbf{So } a \textbf{ will be picked in the next round!}}$

Why is the Sliding Regret of UCB Linear?

Conlusion: When $t \gg T$,

$$\begin{aligned}\mathbf{P}^{\mathrm{UCB}}\left(\forall i < T : A_{t+i} = a \mid A_t = a\right) &\geq \prod_{i=0}^{T-1} \mathbf{P}(R_{t+i} = 1 \mid A_{t+i} = a) \\ &= \mu_a^T .\end{aligned}$$

Remark: This is not the case for Thompson Sampling, for which

$$\mathbf{P}^{\mathrm{TS}}\left(\forall i < T : A_{t+i} = a \mid A_t = a\right) \xrightarrow[t \to \infty]{} 0.$$

A General Principle

Say that the learner has picked a suboptimal arm recently, and that it gave unexpectedly good rewards. Does the learner significatively increase the probability of picking it?

> If **yes**, then **high** sliding regret.
> If **no**, then **small** sliding regret.

⇒ Small sliding regret is about a robustness to local histories.
⇒ KL-UCB, IMED, UCB, UCB-V, MOSS have provably linear sliding regret.

Is Small Sliding Regret Useful?

$\Rightarrow$ Optimal regret does not imply sublinear sliding regret;

$\Rightarrow$ Sliding regret is sometimes a secondary issue;

$\Rightarrow$ Sublinear sliding regret is sometimes **important**, and measures how predictable is suboptimal play over a single run.

[redacted]

Thank you!