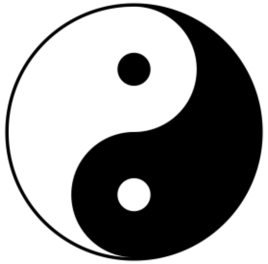
Multi-Armed Bandits:
Bayesian vs. Frequentist

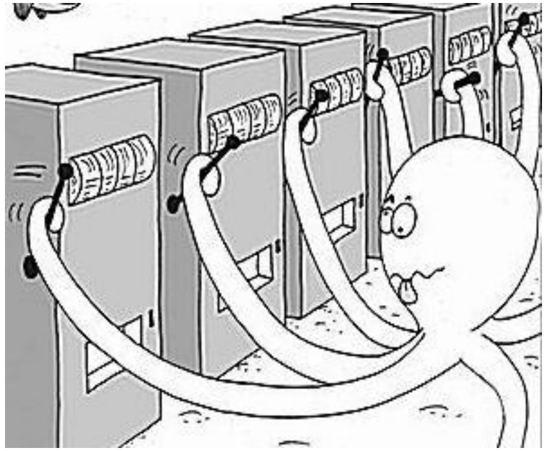


Lorenzo Maggi

Nokia Bell Labs

Basic scenario

- K “arms”
- Arm a = r.v. with distribution ν_a and mean μ_a
- ν_a and μ_a are unknown
- test the arms by obtaining *i.i.d.* samples $\sim \nu_a, \ \forall a$
- **goal**: maximize the sum of rewards (quickly identify $a^* = \mathrm{argmax}_a \, \mu_a$)



Exploration/exploitation dilemma

- K "arms"
- Arm a = r.v. with distribution ν_a and mean μ_a
- ν_a and μ_a are unknown
- **goal**: maximize the sum of rewards ($a^* = \text{argmax}_a \, \mu_a$)
- How? "test" the arms by obtaining *i.i.d.* samples $\sim \nu_a, \ \forall a$

- At time t we have sampled arms and built estimates $\hat{\mu}_{a,t} \approx \mu_a, \ \forall a$
- **Dilemma**:
  - (exploitation) settle for our current estimates and greedily choose what seems to be the best arm ($\hat{a}_t = \text{argmax}_a \, \hat{\mu}_{a,t}$)
  - (exploration) keep sampling the "bad" arms to make sure they're really bad

A simple example

- Arm 1: fixed reward $Y_{1,t}$= 0.25 → $\nu_1 = \delta_{0.25}, \mu_1 = 0.25$
- Arm 2: $Y_{2,t}$ =$\begin{cases} 0 \ w.p. 0.3 \\ 1 \ w.p. 0.7 \end{cases} \rightarrow \ \mu_2 = 0.7$
- **<u>Oracle policy</u>**: always pick arm 2 → unbeatable but not implementable
- **<u>Greedy policy</u>**: choose the arm with highest estimated avg. reward
  → with probability 0.3, we choose the bad arm **forever**! (linear regret)
  - (exploration) time 1: arm 1, reward 0.25 →$\hat{\mu}_1 = .25$
  - (exploration) time 2: arm 2, reward 0 w.p. 0.3 → $\hat{\mu}_2 = 0$
  - (exploitation) time 3: greedily choose arm 1 → $\hat{\mu}_1 = .25$
  - (exploitation) time 3: greedily choose arm 1 → $\hat{\mu}_1 = .25$
  - *… forever and ever…*
- What else…?

Applications

- Clinical trials (which drug should the doctor prescribe?)
- Rate control (at which rate r_i should the BS transmit to user i to maximize throughput $r_i\theta_i$, where θ_i=probability of correct reception)
- Advertising (which ad should the banner display to maximize the revenue?)

… and beyond (restless bandits, not covered here):

- Channel selection in wireless
- Shortest path routing
- Queue control

(formal) goal:

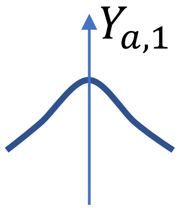
Regret minimization

Rewards have always the same distribution, that is unknown	Rewards are distributed according to *our belief*, that changes over time
Frequentist model	**Bayesian model**
$\mu_1, \dots, \mu_K$ unknown parameters (arm exp. values)	$\mu_1, \dots, \mu_K$ drawn from a prior distribution: $\mu_a \sim \pi_a$
Reward arm a: $\left(Y_{a,t}\right)_t \sim^{i.i.d.} \nu^{\mu_a}$	Reward arm a: $\left(Y_{a,t}\right)_t \mid \boldsymbol{\mu} \sim^{i.i.d.} \nu^{\mu_a}$ $\left(Y_{a,t}\right)_t$ are **not** *i.i.d.* since our belief π_a is updated as: • $Y_{a,1} \sim \nu^{\mu_a},\ \mu_a \sim \pi_a$ • $Y_{a,2} \sim \nu^{\mu'_a},\ \mu'_a \sim \pi'_a = \frac{\Pr(Y_{a,1} \mid \mu_a)\pi_a}{\Pr(Y_{a,1})}$ • …
Regret of algorithm $\mathcal{A}$ (choosing arm A_t at time t)	
$R_T(\mathcal{A}, \boldsymbol{\mu}) = \mathbb{E}\left[\sum_{1 \le t \le T} (\mu^* - Y_{A_t,t})\right]$	$\mathcal{R}_T(\mathcal{A}) = \int R_T(\mathcal{A}, \boldsymbol{\mu}) d\pi(\boldsymbol{\mu})$
"**Good**" algorithm = **sublinear** regret for **all** (unknown) $\boldsymbol{\mu}$	**Optimal** algorithm = **minimum** Bayesian regret given prior π

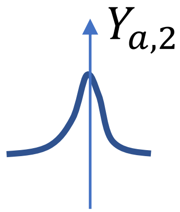
Belief update

on the “goodness” of arms

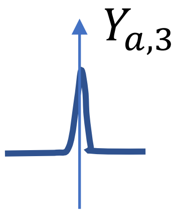
prior π_a on arm a’s avg. reward μ_a



posterior $\pi_{a|Y_{a,1}}$



posterior $\pi_{a|Y_{a,1},Y_{a,2}}$



posterior $\pi_{a|Y_{a,1},Y_{a,2},Y_{a,3}}$

Prior $\pi_a(.) = \Pr(\mu_a =.)$

Posterior $\pi_{a|Y}(.) = \Pr(\mu_a =. |Y) = \frac{\Pr(Y|\mu_a =.)\pi_a(.)}{\Pr(Y)}$

Main intuition: the way we sample the arms has an impact on

- the reward we collect
- the belief we have about the goodness of the arms (*only the sampled arms are observed*!)

Bayesian or Frequentist?

You can update your belief in both cases (no one forbids you!) **but**: subtle difference…

- Bayesian: the *belief* defines your regret (see next)
- Frequentist: the *reward* defines the regret, the belief is just a tool to take better decisions

Bayesian model

Remark: an arm "freezes" when it is not played (pb is weakly decoupled).
If not, refer to Restless Multi-Armed Bandit

Bayesian model

Simpler (but important) case:

- Reward of arm a is Bernoulli(μ_a): $\Pr(Y_a|\mu_a) = \begin{cases} 1\ w.p.\ \mu_a \\ 0\ w.p. (1-\mu_a) \end{cases}$
- Prior on μ_a: $\pi_a = \mathrm{Beta}(n, m)$
  → draw $Y = \{0,1\}$
  posterior is also Beta (conjugate prior!):
  $\pi_{a|Y} = \mathrm{Beta}(n+Y, m+(1-Y))$

Goal: max discounted reward = $\mathbb{E}[\sum_t \beta^t Y_{A_t,t} | \text{belief at time } t]$

Equivalently, solve the following MDP:

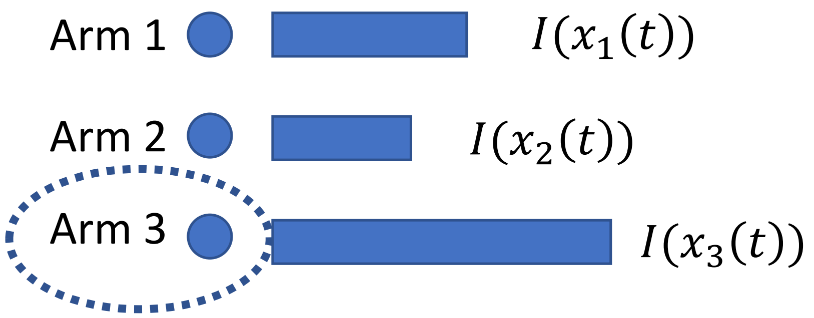
- **state**: belief $\{(n_a, m_a)\}_a = \{(\#1's,\ \#0's, \text{for arm } a)\}_{\text{arm } a}$ ←→ current belief $\pi_{a|Y}$
- **action**: arm A that you pick
- **expected reward** (given state and action): $\frac{n_A}{n_A+m_A}$
- **state transitions** to $\begin{cases} \{(n_A+1, m_A) \cup \{n_a, m_a\}_{a\neq A}\},\ w.p.\ \frac{n_A}{n_A+m_A} \\ \{(n_A, m_A+1) \cup \{n_a, m_a\}_{a\neq A}\},\ w.p.\ 1-\frac{n_A}{n_A+m_A} \end{cases}$

Index policy

- Solving an MDP is conceptually easy ("just" solve an LP)
- **BUT**: curse of dimensionality, the # of states generally explodes!

→ look for an index policy $I: \mathcal{S} \to \mathbb{R}$ such that:

  - (optimality) playing arm with highest index is optimal
  - (decoupling) computing $I(x_a)$ is "easy" since it only depends on arm a

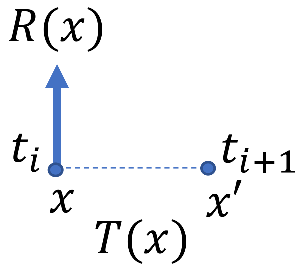


Spoiler: there exists an optimal index policy (see next)

Let's prove the optimality of Gittins index

Semi-Markov Decision Process

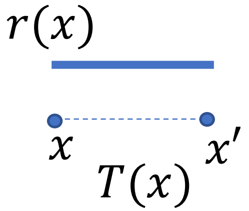
- Each arm a is a semi-Markov process with finite state space $\mathcal{S}_a$
- Arm a is in state $x_a \in \mathcal{S}_a$ and it is *played*. Then,
  - a random reward $R(x_a)$ is received
  - the arm remains "active" over a random time period $T(x_a)$
  - after time $T(x_a)$, the arm moves to a random new state x_a'



- At time t_i a new play starts and we maximize

$$\mathbb{E}\left[\sum_i R_i e^{-\beta t_i}\right]$$

- <u>Equivalent "constant-rate" formulation</u>:
  - reward is received at constant rate $r(x_a) = \frac{\mathbb{E}[R(x_a)]}{\mathbb{E}[\int_{t=0}^{T(x_a)} e^{-\beta t} dt]}$
  - and we maximize: $\mathbb{E}[\int_t r(x(t)) e^{-\beta t} dt]$



Gittins index policy

- *Remember*: we seek for the policy that samples the arms so as to maximize $\mathbb{E}[\int_t r(x(t))e^{-\beta t}dt]$
- Let $x^* = \mathrm{argmax}_x\, r(x)$. Let a^* be the “lucky” arm: $x^* \in \mathcal{S}_{a^*}$
- (auxiliary and intuitive) **Lemma**: ∃ an optimal policy the obeys the rule: If the lucky arm a^* is in state x^*, then play it!
  Proof by contradiction (see [1])

Theorem: The (Gittins) index policy computed as follows is optimal:

$$I(x_a) = \sup_{\tau>0} \frac{\mathbb{E} \int_{t=0}^{\tau} r(t)e^{-\beta t}dt}{\mathbb{E} \int_{t=0}^{\tau} e^{-\beta t}dt} \mid x(0) = x_a, \qquad \tau \text{ stopping time}$$

Proof: see next and [1]

Maximum achievable reward rate [rew/sec] from state x_a

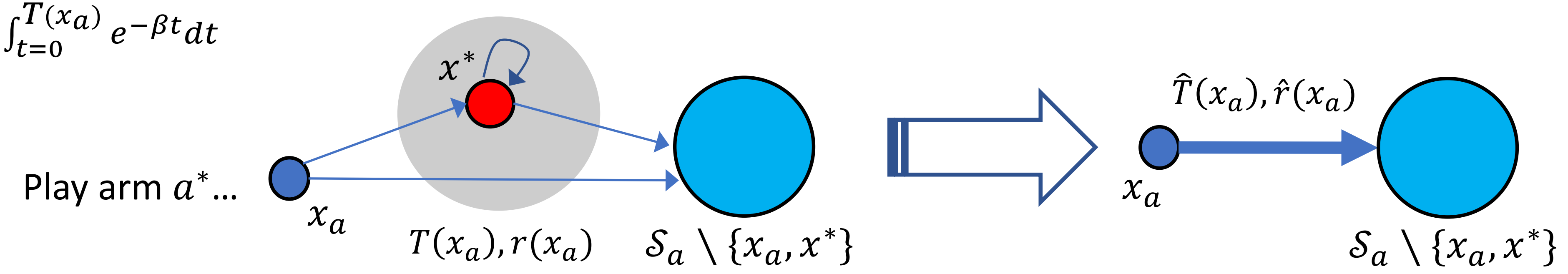
[1] Tsitsiklis, J. N. (1994). A short proof of the Gittins index theorem. *The Annals of Applied Probability*, 194-199.
[2] J. C. Gittins, Bandit Processes and Dynamic Allocation Indices, Journal of the Royal Statistical Society (1979)

Gittins index policy

sketch of the proof in [1] Tsitsiklis, J. N. (1994). A short proof of the Gittins index theorem. The Annals of Applied Probability, 194-199.

- Prove by induction on the # states N that $\exists$ **an** optimal index policy
- For $N = 1$, it is trivially true (just one bandit, always sample it)
- Assume that index policy is optimal for $N = M$, show it for $N = M + 1$

- **Reduce** arm a^* by removing best state $x^* = \text{argmax}_x\, r(x)$:
  - Assume the arm a^* is in state $x_a \neq x^*$
  - Modify reward $\hat{r}(x_a)$ and dwelling time $\hat{T}(x_a)$ by accounting for the fact that when arm a^* in state $x^* = \text{argmax}_x\, r(x)$ then we **must** play it (see before)
    - $\hat{T}(x_a) =$ first time at which state of arm a^* is different from x_a and x^*
    - $\hat{r}(x_a) = \dfrac{\mathbb{E}\int_{t=0}^{\hat{T}(x_a)} r(t) e^{-\beta t} dt}{\mathbb{E}\int_{t=0}^{\hat{T}(x_a)} e^{-\beta t} dt}$



Gittins index policy (cont'd)

sketch of the proof [1]

- After the reduction, state x^* has disappeared from $\mathcal{S}_{a^*}$
- We end up with a MAB with $N = M$ states
- By induction hypothesis, ∃ an optimal index policy for M states!
  → We proved that
  - ∃ an optimal index policy
  - by construction, the optimal index $I(x)$ is as follows:
    - (a) Set $I(x^*) = r(x^*) := \max_x r(x)$. Let a^* be the corresponding arm
    - (b) If set of states $|S_{a^*}| = 1$, then remove arm a^*
    - (c) Else, reduce arm a^* by removing state x^* and go to (a)
  - the index of state x_a only depends on arm a (*curse of dimensionality is broken!*) Complexity is linear in the # arms: $O(\Sigma_a |\mathcal{S}_a|^2)$

Gittins index

Further intuitions

- $I(x_a) = \sup_{\tau>0} \frac{\mathbb{E}\int_{t=0}^{\tau} r(t)e^{-\beta t}dt}{\mathbb{E}\int_{t=0}^{\tau} e^{-\beta t}dt} \mid x(0) = x_a$, τ stopping time
  $=$ highest avg. reward rate (reward/second) achievable from state x_a

- **Further intuition**: Imagine you have 2 arms: $\begin{cases} \text{arm 1: arm } a \\ \text{arm 2: constant reward } \nu \end{cases}$

  Optimal policy: sample arm a **until** a stopping time τ, **then** sample arm 2 **forever** (easy, by contradiction)

P1 *Optimal reward*: $\sup_{\tau\geq 0} \{\mathbb{E}\int_{t=0}^{\tau} r(t)e^{-\beta t}dt + \mathbb{E}\int_{t=\tau}^{\infty} \nu\, e^{-\beta t}\, dt\}$

P2 *Sampling arm 2 forever* gives reward: $\int_{t=0}^{\infty} \nu\, e^{-\beta t}\, dt$

- sup{ν: **P1** better than **P2**} $= \sup_{\nu} \left\{ \sup_{\tau>0} \{\int_{t=0}^{\tau} r(t)e^{-\beta t}dt + \mathbb{E}\int_{t=\tau}^{\infty} \nu\, e^{-\beta t}\, dt\} > \mathbb{E}\int_{t=0}^{\infty} \nu\, e^{-\beta t}\, dt \right\}$

$$= \sup_{\nu} \left\{ \sup_{\tau>0} \{\int_{t=0}^{\tau} r(t)e^{-\beta t}dt > \nu\, \mathbb{E}\int_{t=0}^{\tau} e^{-\beta t}\, dt\} \right\}$$

$$= \sup_{\tau>0} \left\{ \sup_{\nu} \frac{\mathbb{E}\int_{t=0}^{\tau} r(t)e^{-\beta t}dt}{\mathbb{E}\int_{t=0}^{\tau} e^{-\beta t}dt} > \nu \right\} = \sup_{\tau\geq 0} \left\{ \frac{\mathbb{E}\int_{t=0}^{\tau} r(t)e^{-\beta t}dt}{\mathbb{E}\int_{t=0}^{\tau} e^{-\beta t}dt} \right\}$$

➡ **Gittins index is the maximum fixed reward you are ready to give up on to play an arm**

Frequentist model

Frequentist model

- K arms
- Arm a has expected reward μ_a
- *Main differences* w.r.t. Bayesian model:
  - Regret must be low w.r.t. **any** value of $\{\mu_a\}_a$:

$$\min_{A \in \mathcal{A}} R_{\mathcal{A}}(T) = \mathbb{E}\left[\Sigma_{t=1}^{T} (\mu^* - Y_{A_t,t})\right] = \Sigma_a (\mu^* - \mu_a) \mathbb{E}\, n_{a,T}$$

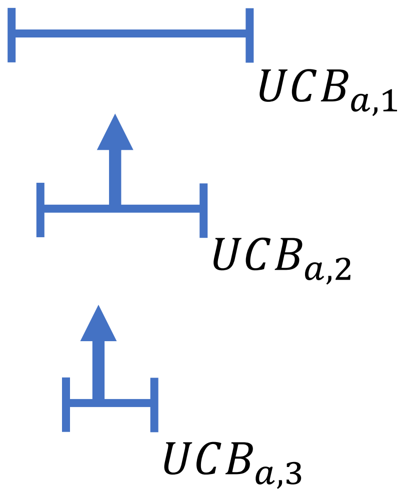
\# times suboptimal arm a is sampled up to time T

  - Regret does **not** depend on the a priori distribution π_a on μ_a
  - Tools: MDPs are no longer useful. Plenty of concentration inequalities instead
- Beware: we may still have a prior π_a! No one forbids us…

A famous frequentist algorithm:

Upper Confidence Bound (UCB)

- While sampling arms, compute the **confidence interval** of the expected reward μ_a, for all arms a and take its **upper bound UCB**
- Always **choose** the arm with the **highest UCB**
- **<u>Intuition</u>:** high UCB <-> *high expected reward* and/or *seldom sampled*



$n_{a,t}$ = # times arm a is sampled up to time t
$\hat{\mu}_{a,t}$ = sampled mean of arm a up to time t

UCB Algorithm:

1. At round $t = 1, \ldots, K$ sample arm t
2. At round $t > K$
   - compute $\mathrm{UCB}_{a,t-1} = \hat{\mu}_{a,t-1} + \sqrt{\ln t - 1 / n_{a,t-1}}$
   - sample arm $A_t = \mathrm{argmax}_a \mathrm{UCB}_{a,t-1}$

Recall: $\mathrm{UCB}_{a,t} = \hat{\mu}_{a,t} + \sqrt{\ln t / n_{a,t}}$

Upper Confidence Bound (UCB)

- Recall: Chernoff bound

$$\Pr\left(\left|\hat{\mu}_{a,t} - \mu_a\right| > \delta\right) \le 2e^{-2n_{a,t}\delta^2}$$

- Use $\delta = \sqrt{\ln t / n_{a,t}}$:
  - $\left|\hat{\mu}_{a,t} - \mu_a\right| > \sqrt{\ln t / n_{a,t}}$ with probability $\ge 1 - 2t^{-2}$

(*) • (UCB is an upper bound w.h.p.) $\mathrm{UCB}_{a,t} \ge \mu_a$ w.p. $\ge 1 - 2t^{-2}$

(**) • $(\hat{\mu}_{a,t} \approx \mu_a)$ $\hat{\mu}_{a,t} < \mu_a + \frac{\mu^* - \mu_a}{2}$ with # samples $n_{a,t} \ge \frac{4 \ln t}{(\mu^* - \mu_a)^2}$ w.p. $\ge 1 - 2t^{-2}$

Upper Confidence Bound (UCB)

Lemma: If at any time t the suboptimal arm a has been played $n_{a,t} \geq \frac{4 \ln t}{(\mu^* - \mu_a)^2}$ times, then $\Pr(A_t = a) \leq 4t^{-2}$.

the more you sampled a suboptimal arm in the past, the less you'll do in the future…

Proof: Show that $\mathrm{UCB}_{a,t} \leq \mathrm{UCB}_{a^*,t}$ w.h.p.:

$$
\begin{aligned}
\mathrm{UCB}_{a,t} &= \hat{\mu}_{a,t} + \sqrt{\ln t \,/ n_{a,t}} \\
&\leq \hat{\mu}_{a,t} + (\mu^* - \mu_a)/2 && \text{since } n_{a,t} \geq \frac{4 \ln t}{(\mu^* - \mu_a)^2} \\
&\leq \left(\mu_a + \frac{(\mu^* - \mu_a)}{2}\right) + \frac{(\mu^* - \mu_a)}{2} && \text{w.p.} \geq 1 - 2t^{-2}, \text{ see } (**) \\
&= \mu^* \\
&\leq \mathrm{UCB}_{a^*,t} && \text{w.p.} \geq 1 - 2t^{-2}, \text{ see } (*)
\end{aligned}
$$

→ $\Pr\left(\mathrm{UCB}_{a,t} \geq \mathrm{UCB}_{a^*,t}\right) \leq 4t^{-2}$ → $\Pr(A_t = a) \leq 4t^{-2}$

$n_{a,t} =$ # times arm a is sampled up to time t

Upper Confidence Bound (UCB)

Lemma: For any suboptimal arm a ($\mu_a < \mu^*$),

$$\mathbb{E}[n_{a,t}] \leq \frac{4 \ln T}{(\mu^* - \mu_a)^2} + 8$$

…and you end up sampling less and less often each suboptimal arm!

Proof: $\mathbb{E}[n_{a,t}] = 1 + \mathbb{E}\, \Sigma_{t=K}^{T} 1(A_{t+1} = a)$

$$= 1 + \mathbb{E}\, \Sigma_{t=K}^{T} 1\left(A_{t+1} = a, n_{a,t} < \frac{4 \ln t}{(\mu^* - \mu_a)^2}\right) + \mathbb{E}\, \Sigma_{t=K}^{T} 1\left(A_{t+1} = a, n_{a,t} \geq \frac{4 \ln t}{(\mu^* - \mu_a)^2}\right)$$

$$\leq \frac{4 \ln T}{(\mu^* - \mu_a)^2} + \Sigma_{t=K}^{T} \Pr\left(A_{t+1} = a, n_{a,t} \geq \frac{4 \ln t}{(\mu^* - \mu_a)^2}\right)$$ *by contradiction*

$$= \frac{4 \ln T}{(\mu^* - \mu_a)^2} + \Sigma_{t=K}^{T} \Pr\left(A_{t+1} = a \mid n_{a,t} \geq \frac{4 \ln t}{(\mu^* - \mu_a)^2}\right) . \Pr\left(n_{a,t} \geq \frac{4 \ln t}{(\mu^* - \mu_a)^2}\right)$$

$$\leq \frac{4 \ln T}{(\mu^* - \mu_a)^2} + \Sigma_{t=K}^{T} 4t^{-2}$$ *by previous slide and* $\Pr \leq 1$

$$\leq \frac{4 \ln T}{(\mu^* - \mu_a)^2} + 8$$

Upper Confidence Bound (UCB)

Remark: the regret holds for any values of μ! (cfr. Bayesian)

Theorem: The regret of UCB algorithm is bounded by:

$$R_T(\text{UCB}) = \mathbb{E}\left[\Sigma_{t=1}^{T}(\mu^* - Y_{A_t,t})\right] \leq \Sigma_{a \neq a^*} \frac{4 \ln T}{(\mu^* - \mu_a)} + 8(\mu^* - \mu_a)$$

Proof: $\mathbb{E}\left[\Sigma_{t=1}^{T}(\mu^* - Y_{A_t,t})\right] = \Sigma_{a \neq a^*} (\mu^* - \mu_a) \mathbb{E}\left[n_{a,T}\right]$

$$\leq \Sigma_{a \neq a^*} (\mu^* - \mu_a) \left(\frac{4 \ln T}{(\mu^* - \mu_a)^2} + 8 \right)$$

$$= \Sigma_{a \neq a^*} \frac{4 \ln T}{(\mu^* - \mu_a)} + 8(\mu^* - \mu_a)$$

and finally…

How “good” is a frequentist MAB algorithm?

Theorem [2]: For any algorithm $\mathcal{A}$,

$$\liminf_{T} \frac{R_T(\mathcal{A})}{\log T} \geq \Sigma_{a \neq a^*} \frac{\mu_a^* - \mu_a}{D_{KL}(\nu_a, \nu_{a^*})}$$

where $D_{KL}(\nu_a, \nu_{a^*}) = \int \nu_a \log \frac{\nu_a}{\nu_{a^*}}$ measures the “distance” between distributions ν_a and ν_{a^*}

[3] Lai, T.L.; Robbins, H. (1985). "Asymptotically efficient adaptive allocation rules". *Advances in Applied Mathematics*. **6** (1): 4–22.

Some more references

- Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games*. Cambridge university press.
- Auer, P.; Cesa-Bianchi, N.; Fischer, P. (2002). "Finite-time Analysis of the Multiarmed Bandit Problem". Machine Learning. 47 (2/3): 235–256.
- Gittins, J. C. (1989), Multi-armed bandit allocation indices, Wiley-Interscience Series in Systems and Optimization., Chichester: John Wiley & Sons, Ltd
- T. Lattimore and C. Szepesvari, “Bandit Algorithms”. Available at http://downloads.tor-lattimore.com/book.pdf


The End