Upper confidence bound algorithms

Christos Dimitrakakis

EPFL

November 6, 2013

The stochastic bandit problem

- A set of K bandits, actions $\mathcal{A} = \{1, \dots, K\}$
- Expected reward of the i-th bandit: $\mu_i \triangleq \mathbb{E}(r_t \mid a_t = i)$.
- Maximise:

$$\sum_{t=1}^{T} r_t, \tag{2.1}$$

where T is arbitrary.

What is a good heuristic strategy?

Definition (Regret)

The (total) regret of a policy π relative to the optimal policy is:

$$L_T(\pi) \triangleq \sum_{t=1}^{T} r_t^* - r_t^\pi \tag{2.2}$$

Empirical average

$$\hat{\mu}_{t,i} \triangleq \frac{1}{n_{t,i}} \sum_{k=1}^{t} r_{k,i} \, \mathbb{I}\left\{a_k = i\right\}, \qquad n_{t,i} \triangleq \sum_{k=1}^{t} \mathbb{I}\left\{a_k = i\right\}.$$

Algorithm 1 Optimistic initial values

Input $\mathcal{A}$, $\mathcal{R}$

$r_{\max} \triangleq \max \mathcal{R}$

for $t = 1, \ldots$ **do**

  $u_{t,i} = \frac{n_{t-1,i}\hat{\mu}_{t-1,i} + r_{\max}}{n_{t-1,i}+1}$

  $a_t = \arg\max_{i \in \mathcal{A}} u_{t,i}$

end for

A simple analysis in the deterministic case

Consider the case where $r_{t,i} = \mu_{t,i}$ for all bandits.

- Then $u_{t,i} \geq \mu_i$ for all t, i.

A simple analysis in the deterministic case

Consider the case where $r_{t,i} = \mu_{t,i}$ for all bandits.

- Then $u_{t,i} \geq \mu_i$ for all t, i.
- At time t, we play i if $u_{t,i} \geq u_{t,j}$ for all j.

A simple analysis in the deterministic case

Consider the case where $r_{t,i} = \mu_{t,i}$ for all bandits.

- Then $u_{t,i} \geq \mu_i$ for all t, i.
- At time t, we play i if $u_{t,i} \geq u_{t,j}$ for all j.
- But $u_{t,j} \geq \mu_j$

A simple analysis in the deterministic case

Consider the case where $r_{t,i} = \mu_{t,i}$ for all bandits.

- Then $u_{t,i} \geq \mu_i$ for all t, i.
- At time t, we play i if $u_{t,i} \geq u_{t,j}$ for all j.
- But $u_{t,j} \geq \mu_j$
- If $\mu^* \triangleq \max_j \mu_j$, we play i at most

$$n_{t,i} \leq \frac{r_{\max}}{\Delta_i}$$

times, where $\Delta_i = \mu^* - \mu_i$.

A simple analysis in the deterministic case

Consider the case where $r_{t,i} = \mu_{t,i}$ for all bandits.

- Then $u_{t,i} \geq \mu_i$ for all t, i.
- At time t, we play i if $u_{t,i} \geq u_{t,j}$ for all j.
- But $u_{t,j} \geq \mu_j$
- If $\mu^* \triangleq \max_j \mu_j$, we play i at most

$$n_{t,i} \leq \frac{r_{\max}}{\Delta_i}$$

  times, where $\Delta_i = \mu^* - \mu_i$.
- Since every time we play i we lose Δ_i, the regret is

$$L_T \leq \sum_{i \neq j} \Delta_i \frac{r_{\max} - \mu^*}{\Delta_i} = (K - 1)(r_{\max} - \mu^*)$$

Algorithm 2 UCB1

Input $\mathcal{A}$, $\mathcal{R}$

$\hat{\mu}_{0,i} = r_{\max}$, $\forall i$.

for $t = 1, \ldots$ **do**

$u_{t,i} = \hat{\mu}_{t-1}, i + \sqrt{2 \frac{\ln t}{n_{t-1,i}}}$.

$a_t = \arg\max_{i \in \mathcal{A}} u_{t,i}$

end for

Theorem (Auer et al [?])

The expected regret of UCB1 after T rounds is at most

$$c_1 \sum_{i:\mu_i<\mu^*} \left(\frac{\ln T}{\Delta_i}\right) + c_2 \sum_{j=1}^{K} \Delta_j$$

Proof.

First we prove that

$$\mathbb{E}\, n_{t,i} \leq O\left(\frac{\ln T}{\Delta_i^2}\right)$$

Then we note that the expected regret can be written as

$$\sum_{i:\mu_i<\mu^*} \Delta_i\, \mathbb{E}\, n_{t,i}$$

due to Wald's identity. □

Let $B_{t,s} = \sqrt{(2 \ln t)/s}$. Then we can prove $\forall c \in \mathbb{Z}$:

$$
\begin{aligned}
n_{T,i} &= 1 + \sum_{t=K+1}^{T} \mathbb{I}\left\{a_t = i\right\} \\
&\leq c + \sum_{t=K+1}^{T} \mathbb{I}\left\{a_t = i \wedge n_{t-1,i} \geq c\right\} \\
&\leq c + \sum_{t=K+1}^{T} \mathbb{I}\left\{\hat{\mu}^*_{n^*_{t-1}} + B_{t-1,n^*_{t-1}} \leq \max \hat{\mu}_{n_i(t-1),i} + B_{t-1,n_i(t-1)}\right\} \\
&\leq c + \sum_{t=K+1}^{T} \mathbb{I}\left\{\min_{0<s<t} \hat{\mu}^*_s + B_{t-1,s} \leq \max_{c \leq s_i < t} \hat{\mu}_{s_i,i} + B_{t-1,s_i}\right\} \\
&\leq c + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_i=c}^{t-1} \mathbb{I}\left\{\hat{\mu}^*_s + B_{t-1,s} \leq \hat{\mu}_{s_i,i} + B_{t-1,s_i}\right\}
\end{aligned}
$$

Let $B_{t,s} = \sqrt{(2 \ln t)/s}$. Then we can prove $\forall c \in \mathbb{Z}$:

$$n_{T,i} \leq c + \sum_{t=1}^{\infty} \sum_{s=1}^{t-1} \sum_{s_i=c}^{t-1} \mathbb{I}\left\{\hat{\mu}_s^* + B_{t-1,s} \leq \hat{\mu}_{s_i,i} + B_{t-1,s_i}\right\}$$

When the indicator function is true one of the following holds:

$$\hat{\mu}_s^* \leq \mu^* - B_{t,s} \tag{2.3}$$

$$\hat{\mu}_{s_i,i} \geq \mu_i + B_{t,s_i} \tag{2.4}$$

$$\mu^* < \mu_i + 2B_{t,s_i} \tag{2.5}$$

Proof idea

- Bound the probability of the first two events.
- Choose c to bound the last term.

From Hoeffding bound:

$$\mathbb{P}(\hat{\mu}_s^* \leq \mu^* - B_{t,s}) \leq e^{-4 \ln t} = t^{-4} \tag{2.6}$$

$$\mathbb{P}(\hat{\mu}_{s_i,i} \geq \mu_i + B_{t,s_i}) \leq e^{-4 \ln t} = t^{-4} \tag{2.7}$$

Setting $c = \lceil (8 \ln n)/\Delta_i^2 \rceil$ makes the last event false as $s_i \geq c$.

$$\mu^* - \mu_i - 2B_{t,s_i} = \mu^* - \mu_i - 2\sqrt{(2 \ln t)/s_i} \geq \mu^* - \mu_i - \Delta_i = 0.$$

Summing up all the terms completes the proof.

Bandits and optimisation

- Continuous stochastic functions[**? ? ?**]
- Constrained deterministic distributed functions[**?**]

First idea[?]

Solve a sequence of discrete bandit problems.

At epoch i, we have some interval A_i

- Split the interval A_i in k regions $A_{i,j}$
- Run UCB on the k-armed bandit problem.
- When a region is sub-optimal with high probability, remove it!

Tree bandits [?]

Create a tree of coverings, with (h, i) being the i-th node at depth h. $\mathcal{D}$ are the descendants and $\mathcal{C}$ the children of a node.
At time t we pick node H_t, I_t. Each node is picked at most once.

$$n_{h,i}(T) \triangleq \sum_{t=1}^{T} \mathbb{I}\left\{(H_t, I_t) \in \mathcal{D}(h, i)\right\} \qquad \text{(visits of } (h, i))$$

$$\widehat{\mu}_{h,i}(T) \triangleq \frac{1}{n_{h,i}(T)} \sum_{t=1}^{T} r_t \, \mathbb{I}\left\{(H_t, I_t) \in \mathcal{C}(h, i)\right\} \qquad \text{(reward from } (h, i))$$

(child bound)

Tree bandits [?]

Create a tree of coverings, with (h, i) being the i-th node at depth h. $\mathcal{D}$ are the descendants and $\mathcal{C}$ the children of a node.
At time t we pick node H_t, I_t. Each node is picked at most once.

$$n_{h,i}(T) \triangleq \sum_{t=1}^{T} \mathbb{I}\left\{(H_t, I_t) \in \mathcal{D}(h, i)\right\} \qquad \text{(visits of } (h, i))$$

$$\widehat{\mu}_{h,i}(T) \triangleq \frac{1}{n_{h,i}(T)} \sum_{t=1}^{T} r_t \, \mathbb{I}\left\{(H_t, I_t) \in \mathcal{C}(h, i)\right\} \qquad \text{(reward from } (h, i))$$

$$C_{h,i}(T) \triangleq \widehat{\mu}_{h,i}(T) + \sqrt{\frac{2 \ln T}{n_{h,i}(T)}} + nu_1 \rho^h \qquad \text{(confidence bound)}$$

$$B_{h,i}(T) \triangleq \min\left\{C_{h,i}(T), \max_{(h+1,j)\in\mathcal{C}(h,i)} B_{h+1,j}\right\} \qquad \text{(child bound)}$$

Infinite horizon, discounted

Discount factor γ such that

$$U_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k} \tag{4.1}$$

Geometric horizon, undiscounted

At each step t, the process terminates with probability $1 - \gamma$:

$$U_t^T = \sum_{k=0}^{T-t} r_{t+k}, \qquad T \sim \mathcal{G}eom(1 - \gamma) \tag{4.2}$$

$$V_\gamma^\pi(s) \triangleq \mathbb{E}(U_t \mid s_t = s)$$

Infinite horizon, discounted

Discount factor γ such that

$$U_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k} \qquad \Rightarrow \mathbb{E}\, U_t = \sum_{k=0}^{\infty} \gamma^k \, \mathbb{E}\, r_{t+k} \tag{4.1}$$

Geometric horizon, undiscounted

At each step t, the process terminates with probability $1 - \gamma$:

$$U_t^T = \sum_{k=0}^{T-t} r_{t+k}, \quad T \sim \mathcal{Geom}(1-\gamma) \qquad \Rightarrow \mathbb{E}\, U_t = \sum_{k=0}^{\infty} \gamma^k \, \mathbb{E}\, r_{t+k} \tag{4.2}$$

$$V_\gamma^\pi(s) \triangleq \mathbb{E}(U_t \mid s_t = s)$$

The expected total reward criterion

$$V_t^{\pi,T} \triangleq \mathbb{E}_\pi U_t^T, \qquad V^\pi \triangleq \lim_{T\to\infty} V^{\pi,T} \tag{4.3}$$

Dealing with the limit

- Consider μ s.t. the limit exists $\forall \pi$.

The expected total reward criterion

$$V_t^{\pi,T} \triangleq \mathbb{E}_\pi U_t^T, \qquad V^\pi \triangleq \lim_{T\to\infty} V^{\pi,T} \tag{4.3}$$

Dealing with the limit

- Consider μ s.t. the limit exists $\forall \pi$.

$$V_+^\pi(s) \triangleq \mathbb{E}_\pi\left(\sum_{t=1}^\infty r_t^+ \,\middle|\, s_t = s\right), \quad V_-^\pi(s) \triangleq \mathbb{E}_\pi\left(\sum_{t=1}^\infty r_t^- \,\middle|\, s_t = s\right) \tag{4.4}$$

$$r_t^+ \triangleq \max\{-r, 0\}, \qquad r_t^- \triangleq \max\{r, 0\}. \tag{4.5}$$

The expected total reward criterion

$$V_t^{\pi,T} \triangleq \mathbb{E}_\pi \, U_t^T, \qquad V^\pi \triangleq \lim_{T\to\infty} V^{\pi,T} \tag{4.3}$$

Dealing with the limit

- Consider μ s.t. the limit exists $\forall \pi$.
- Consider μ s.t. $\exists \pi^*$ for which V^{π^*} exists and

$$\lim_{T\to\infty} V^{\pi^*,T} = V^{\pi^*} \geq \limsup_{T\to\infty} V^{\pi,T}.$$

The expected total reward criterion

$$V_t^{\pi,T} \triangleq \mathbb{E}_\pi \, U_t^T, \qquad\qquad V^\pi \triangleq \lim_{T\to\infty} V^{\pi,T} \tag{4.3}$$

Dealing with the limit

- Consider μ s.t. the limit exists $\forall\pi$.
- Consider μ s.t. $\exists\pi^*$ for which V^{π^*} exists and

$$\lim_{T\to\infty} V^{\pi^*,T} = V^{\pi^*} \geq \limsup_{T\to\infty} V^{\pi,T}.$$

- Use optimality criteria sensitive to the divergence rate.

The average reward (gain) criterion

The gain g

$$g^{\pi}(s) \triangleq \lim_{T \to \infty} \frac{1}{T} V^{\pi,T}(s) \tag{4.4}$$

$$g_{+}^{\pi}(s) \triangleq \limsup_{T \to \infty} \frac{1}{T} V^{\pi,T}(s), \qquad g_{-}^{\pi}(s) \triangleq \liminf_{T \to \infty} \frac{1}{T} V^{\pi,T}(s) \tag{4.5}$$

If $\lim_{T \to \infty} \mathbb{E}(r_T \mid s_0 = s)$ exists then it equals $g^{\pi}(s)$.

Let Π be the set of all history-dependent, randomised policies.
π^* is total reward optimal if

$$V^{\pi^*}(s) \geq V^{\pi}(s) \qquad \forall s \in \mathcal{S}, \pi \in \Pi.$$

π^* is discount optimal for $\gamma \in [0, 1)$ if

$$V_{\gamma}^{\pi^*}(s) \geq V_{\gamma}^{\pi}(s) \qquad \forall s \in \mathcal{S}, \pi \in \Pi.$$

π^* is gain optimal if

$$g^{\pi^*}(s) \geq g^{\pi}(s) \qquad \forall s \in \mathcal{S}, \pi \in \Pi.$$

Overtaking optimality

π^* is overtaking optimal if

$$\liminf_{T\to\infty} \left[V^{\pi^*,T}(s) - V^{\pi,T}(s)\right] \geq 0 \qquad \forall s \in \mathcal{S}, \pi \in \mathit{\Pi}.$$

However, no overtaking optimal policy may exist.

π^* is average-overtaking optimal if

$$\liminf_{T\to\infty} \frac{1}{T} \left[V^{\pi^*,T}(s) - V^{\pi}_{+}(s)\right] \geq 0 \qquad \forall s \in \mathcal{S}, \pi \in \mathit{\Pi}.$$

Sensitive discount optimality

π^* is n-discount optimal for $n \in \{-1, 0, 1, \ldots\}$ if

$$\liminf_{\gamma \uparrow 1} (1-\gamma)^{-n} \left[V_\gamma^{\pi^*}(s) - V_\gamma^{\pi}(s) \right] \geq 0 \qquad \forall s \in \mathcal{S}, \pi \in \Pi.$$

A policy is Blackwell optimal if $\forall s$, $\exists \gamma^*(s)$ such that

$$V_\gamma^{\pi^*}(s) - V_\gamma^{\pi}(s) \geq 0, \qquad \forall \pi \in \Pi, \gamma^*(s) \gamma \gamma < 1.$$

Lemma

If a policy is m-discount optimal then it is n-discount optimal for all $n \leq m$.

Lemma

Gain optimality is equivalent to -1-discount optimality.

An upper-confidence bound algorithm

Confidence region M_t such that

$$\mathbb{P}(\mu \notin M_t) < \delta \tag{4.6}$$

Optimistic value for policy π:

$$V_{+}^{\pi}(M_t) \triangleq \max\left\{ V_{\mu}^{\pi} \;\middle|\; \mu \in M_t \right\} \tag{4.7}$$

An upper-confidence bound algorithm

Confidence region M_t such that

$$\mathbb{P}(\mu \notin M_t) < \delta \tag{4.6}$$

Optimistic value for policy π:

$$V_{+}^{\pi}(M_t) \triangleq \max\left\{V_{\mu}^{\pi} \;\middle|\; \mu \in M_t\right\} \tag{4.7}$$

UCRL [?] outline

An upper-confidence bound algorithm

Confidence region M_t such that

$$\mathbb{P}(\mu \notin M_t) < \delta \tag{4.6}$$

Optimistic value for policy π:

$$V_{+}^{\pi}(M_t) \triangleq \max \left\{ V_{\mu}^{\pi} \;\middle|\; \mu \in M_t \right\} \tag{4.7}$$

UCRL [?] outline

- At round k, start time t_k, calculate M_{t_k}.

An upper-confidence bound algorithm

Confidence region M_t such that

$$\mathbb{P}(\mu \notin M_t) < \delta \tag{4.6}$$

Optimistic value for policy π:

$$V_{+}^{\pi}(M_t) \triangleq \max \left\{ V_{\mu}^{\pi} \;\middle|\; \mu \in M_t \right\} \tag{4.7}$$

UCRL [?] outline

- At round k, start time t_k, calculate M_{t_k}.
- Choose $\pi_k \in \arg\max_{\pi} V_{+}^{\pi}(M_{t_k})$.

An upper-confidence bound algorithm

Confidence region M_t such that

$$\mathbb{P}(\mu \notin M_t) < \delta \tag{4.6}$$

Optimistic value for policy π:

$$V_{+}^{\pi}(M_t) \triangleq \max\left\{ V_{\mu}^{\pi} \;\middle|\; \mu \in M_t \right\} \tag{4.7}$$

UCRL [?] outline

- At round k, start time t_k, calculate M_{t_k}.
- Choose $\pi_k \in \arg\max_{\pi} V_{+}^{\pi}(M_{t_k})$.
- Execute π_k, observe rewards and update model until t_{k+1}.

The confidence region

Let M_t be a set of plausible MDPs for time t with transitions τ s.t.:

$$\left\| \boldsymbol{P}(\cdot \mid s, a) - \hat{\boldsymbol{P}}_t(\cdot \mid s, a) \right\|_1 \leq \sqrt{\frac{n \ln T}{N_t(s, a)}}, \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}, \tag{4.8}$$

where $\hat{\boldsymbol{P}}_t(\cdot \mid s, a)$ is the empirical transition probability.
Then $\mathbb{P}(\mu \in M_t) > 1 - nkT^{-2}$, via a bound due to Weissman [**?**].

The confidence region

Let M_t be a set of plausible MDPs for time t with transitions τ s.t.:

$$\left\| \boldsymbol{P}(\cdot \mid s, a) - \hat{\boldsymbol{P}}_t(\cdot \mid s, a) \right\|_1 \leq \sqrt{\frac{n \ln T}{N_t(s, a)}}, \qquad \forall s \in \mathcal{S}, a \in \mathcal{A}, \tag{4.8}$$

where $\hat{\boldsymbol{P}}_t(\cdot \mid s, a)$ is the empirical transition probability.
Then $\mathbb{P}(\mu \in M_t) > 1 - nkT^{-2}$, via a bound due to Weissman [**?**].

Changing set of plausible MDPs

- This implies that we may have to switch policies.
- We do so when $N_t(s, a)$ doubles for some s, a .

Calculating the upper bound

In effect, create an augmented MDP

$$Q_t(s, a) = r(s, a) + \max \left\{ \sum_{s' \in \mathcal{S}} \boldsymbol{P}(s' \mid s, a) V_{t+1}(s') \;\middle|\; \|\boldsymbol{P} - \hat{\boldsymbol{P}}\|_1 \leq \epsilon \right\} \tag{4.9}$$

$$V_t(s) = \max_{a \in \mathcal{A}} Q_t(s, a) \tag{4.10}$$

Comparison with Bayesian upper bound

High-probability value function bound

$$V_+^* = \max \left\{ V_\mu^* \;\middle|\; \mu \in M_t \right\}, \qquad \mathbb{P}(\mu^* \in M_t) \geq 1 - \delta.$$

Highly credible value function bound

$$V_+^* = \max \left\{ V_\mu^* \;\middle|\; \mu \in M_t \right\}, \qquad \xi_t(M_t) \geq 1 - \delta.$$

Bayesian value function bound (e.g. [**?**])

$$V_+^* = \int_{\mathcal{M}} V_\mu^* \, \mathrm{d}\xi_t(\mu) \qquad \xi_t = \xi_0(\cdot \mid s_t, r_t, \dots)$$