Learning Theory of Optimal Decision Making

Part I: On-line learning in stochastic environments

Csaba Szepesvári[1]

[1]Department of Computing Science
University of Alberta

Machine Learning Summer School, Ile de Re, France, 2008

with thanks to: RLAI group, SZTAKI group, Jean-Yves Audibert, Remi Munos

Outline

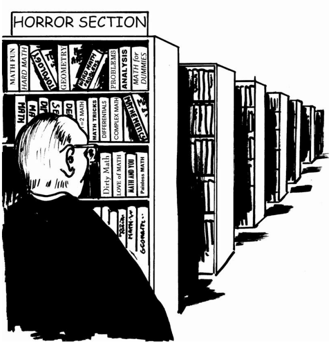
RL
AI

MOTTO

”A theory is something nobody believes, except the person who made it. An experiment is something everybody believes, except the person who made it.”
(Einstein)

HIGH LEVEL OVERVIEW OF THE TALKS

- Day 1: **Online learning in stochastic environments**
- Day 2: Batch learning in Markovian Decision Processes
- Day 3: Online learning in adversarial environments



REFRESHER

- Union bound: $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$
- Inclusion: $A \subset B \Rightarrow \mathbb{P}(A) \leq \mathbb{P}(B)$
- Inversion: If $\mathbb{P}(X > t) \leq F(t)$ holds for all t then for any $0 < \delta < 1$, w.p. $1 - \delta$, $X \leq F^{-1}(\delta)$
- Expectation: If $X \geq 0$, $\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X \geq t)\, dt$

WHAT IS IT?

PROTOCOL OF LEARNING

Concepts: Agent, Environment, sensations, actions, rewards
Time: $t = 1, 2, \ldots$
Perception-action loop:

1. Agent senses Y_t coming from the Environment
2. Agent sends action A_t to the Environment
3. Agent receives reward R_t from the Environment
4. $t := t + 1$, go to Step 1

Goal: $\sum_{t=1}^{T} R_t \to \max$

RELATED PROBLEMS

- Cost-sensitive learning – actions are predictions
- Passive (batch) learning – no influence on training data
- Active learning – learn fast (instead of cheaply)

WHY SHOULD WE CARE?

- Online
  - Opportunity to keep improving
  - Can learn with fewer data points
  - Learning should be cheap
- learning – you know!
- in stochastic environments – convenience, ..

THE CHALLENGE: EXPLORE OR EXPLOIT?

CLINICAL TRIALS

Drugs: $i \in I \stackrel{\text{def}}{=} \{1, 2, \ldots, k\}$

Protocol:

1. Choose drug $I_t \in I$ for the next patient
2. Observe response $R_t(I_t) \in \{0, 1\}$ of the patient
3. $t := t + 1$, go to Step 1

Estimated response of drug i after n steps:

$$Q_n(i) = \frac{\sum_{t=1}^{n} \mathbb{I}_{\{I_t = i\}} R_t(i)}{\sum_{t=1}^{n} \mathbb{I}_{\{I_t = i\}}}$$

WHICH DRUG TO CHOOSE?

- Best response so far? (greedy choice; exploit; optimize return)
- Least explored? (explore; collect information)
- ???

BANDIT PROBLEM [ROBBINS, 1952]

Choices: $i \in I \stackrel{\text{def}}{=} \{1, 2, \ldots, k\}$
Protocol:

1. Choose an option $I_t \in I$
2. Observe response $R_t(I_t) \in \{0, 1\}$
3. $t := t + 1$, go to Step 1

Terminology: option = arm = action

ASSUMPTION

$R_t(i) \sim P_i(\cdot)$, independence
within and between the arms

Examples

Examples

- Clinical trials
- Web advertising
- Job shop scheduling
- ⋮

SOME DEFINITIONS

- Expected payoffs: $Q(i) = \mathbb{E}[R_1(i)]$
- Optimal arm: $Q(i^*) = Q^* \stackrel{\mathrm{def}}{=} \max_i Q(i)$
- Set of optimal arms: $I^* = \{ i \mid Q(i) = Q^* \}$
- Payoff loss ("gap"): $\Delta_i = Q^* - Q(i)$
- Total reward up to time n: $R_{1:n} = \sum_{t=1}^{n} R_t(I_t)$
- A bandit problem instance: B
- Class of bandit problems: $\mathcal{B}$

DEFINING THE GOALS

DEFINITION (CONSISTENCY)

A bandit algorithm $\mathcal{A}$ is strongly consistent on $\mathcal{B}$ if

$$\lim_{t\to\infty} \mathbb{P}\left(I_t \in I^*\right) \to 1$$

holds when $\mathcal{A}$ is run on any instance from $\mathcal{B}$.

DEFINITION (HANNAN-CONSISTENCY ("NO-REGRET"))

A bandit algorithm is Hannan-consistent on $\mathcal{B}$ if its expected regret L_n is sublinear over time:

$$L_n \stackrel{\text{def}}{=} n\,Q^* - \mathbb{E}\left[R_{1:n}\right] = o(n)$$

when $\mathcal{A}$ is run on any instance from $\mathcal{B}$.

Strong vs. Hannan-consistency

Proposition

If an algorithm is strongly consistent then it is also Hannan-consistent.

Proof.

Let $a_t = \mathbb{E}\left[Q^* - R_t(I_t)\right]$. Then

$$\begin{aligned} a_t &= \sum_i \mathbb{E}\left[Q^* - R_t(i) | I_t = i\right] \mathbb{P}\left(I_t = i\right) \\ &= \sum_{i:\Delta_i > 0} \Delta_i \mathbb{P}\left(I_t = i\right). \end{aligned}$$

Hence $a_t \to 0$. Cesaro: $n^{-1} \sum_{t=1}^n a_t \to 0$. □

Proposition

The reverse does not hold.

Exploration is necessary

Definition

A bandit problem is non-trivial if for i^* optimal, i suboptimal, with **positive** probability $R_t(i^*) < R_t(i)$ holds.

Definition

An algorithm **stops exploring** on problem B if there exists a time n such that after time step n the algorithm only chooses exploiting actions: $Q_n(I_t) = \max_i Q_n(i)$. Time n may depend on B.

Proposition

Let $\mathcal{B}$ contain a non-trivial bandit problem. If an algorithm stops exploring it cannot be consistent on $\mathcal{B}$.

METHODS FOR ACHIEVING CONSISTENCY

- Forcing
  - Fixed schedule
  - ϵ-greedy
- Softmax
- "Optimism in the Face of Uncertainty"

Forcing

Idea

Exploration is necessary $\Rightarrow$ make sure that every arm is selected infinitely often

$T_n(i) = \sum_{t=1}^{n} \mathbb{I}_{\{I_t = i\}}$:
Number of times arm i is chosen up to time n.

Forcing Algorithm($(f_t)_{t\in\mathbb{N}}$)

At time t do:

1. $i_0 := \mathrm{argmin}_i T_t(i)$
2. if $T_t(i_0) < f_t$ then $I_t := i_0$ else $I_t := \mathrm{argmax}_i Q_t(i)$.

Comment

Other possibility: Periodic forcing (forcing with a fixed timing)

Regret of the Forcing Algorithm

Proposition

Let $f_t \geq 0$ and $\lim_{t\to\infty} f_t = \infty$. Then the Forcing Algorithm is Hannan-consistent, i.e., $L_n/n \to 0$
(The Forcing Algorithm is not strongly consistent.)

.. but what is the growth rate of L_n?

Central limit theorem

$$\mathbb{P}\left(\sqrt{\frac{n}{\sigma^2}}\left(\overline{X}_n - \mathbb{E}\left[X_1\right]\right) \geq y\right) \to 1 - \Phi(y) \approx \frac{1}{\sqrt{2\pi}} \frac{e^{-y^2/2}}{y},$$

hence

$$\begin{aligned}
\mathbb{P}\left(\overline{X}_n - \mathbb{E}\left[X_1\right] \geq \epsilon\right) &= \mathbb{P}\left(\sqrt{\frac{n}{\sigma^2}}\left(\overline{X}_n - \mathbb{E}\left[X_1\right]\right) \geq \sqrt{\frac{n}{\sigma^2}}\epsilon\right) \\
&\approx e^{-n\epsilon^2/(2\sigma^2)}\sqrt{\frac{\sigma^2}{n\epsilon^2}} \approx e^{-n\epsilon^2/(2\sigma^2)}.
\end{aligned}$$

Theorem (Hoeffding's inequality)

Let $X_i \in [0,1]$ *i.i.d.,* $\mu = \mathbb{E}[X_1]$, $\overline{X}_n = 1/n \sum_{t=1}^{n} X_t$. *Then*

$$\mathbb{P}\left(\overline{X}_n \geq \mu + \epsilon\right) \quad \leq \quad e^{-2n\epsilon^2}$$
$$\mathbb{P}\left(\overline{X}_n \leq \mu - \epsilon\right) \quad \leq \quad e^{-2n\epsilon^2}.$$

Corollary (Hoeffding's bound in deviation form)

Let $\delta > 0$. *With probability* $1 - \delta$,

$$\overline{X}_n \leq \mu + \sqrt{\frac{\log(1/\delta)}{2n}}.$$

Similarly, with probability $1 - \delta$,

$$\overline{X}_n \geq \mu - \sqrt{\frac{\log(1/\delta)}{2n}}.$$

HEURISTIC ANALYSIS

Assume that $R_t(i) \in [0, 1]$.
Then (handwaving) w.p. $1 - \delta_t$, for all $i \in \{1, \ldots, k\}$

$$|Q_t(i) - Q(i)| \leq c_t \stackrel{\text{def}}{=} \sqrt{\frac{\log(2k/\delta_t)}{2f_t}}$$

since we explored i at least f_t times.
Counting mistakes:

1. When forcing: total contribution to L_n is $k\, f_n$
2. When the sample is “atypical”: $\sum_{t=1}^{n} \delta_t$
3. When the sample is “typical”:

   no mistake when $c_t < \Delta^*/2 \stackrel{\text{def}}{=} \min_{j:\Delta_j > 0} \Delta_j/2 \Leftrightarrow$

   (*) $\quad f_t \geq 2\log(2k/\delta_t)/(\Delta^*)^2 \Rightarrow$

$$L_n \leq f_n + \sum_{t=1}^{n} \delta_t + O(1) = O(\log n)$$

if $\delta_t = 1/t$ and $f_t = c' \log(t)$ with $c' = \Omega((\Delta^*)^{-2})$.

Theorem

Assume that the payoffs are bounded. If $f_t = c' \log(t)$ *with* $c' = \Omega((\Delta^*)^{-2})$ *then the regret of the Forcing Algorithm grows logarithmically:*

$$L_n = O(\log n).$$

THE ϵ-GREEDY ALGORITHM

ϵ-GREEDY$((\epsilon_t)_{t\in\mathbb{N}})$

At time t do:

1. Draw U_t in $[0,1]$ uniformly at random
2. if $U_t < \epsilon_t$ then

   Pick I_t randomly from $\{1,2,\ldots,k\}$
3. else

   $I_t := \mathrm{argmax}_i Q_t(i)$.

THEOREM (REGRET OF ϵ-GREEDY [AUER ET AL., 2002])

Assume that the payoffs are bounded. If $\epsilon_t = c'/t$ with $c' = \Omega(k/(\Delta^*)^2)$ then

$$\mathbb{P}(I_t \notin I^*) \le O(\frac{c'}{n}) \quad \text{and} \quad L_n = O(c' \log n).$$

Softmax algorithms

Problem with the previous algorithms:
When exploring they are indifferent about $Q_t(i)$.

Boltzmann($(\tau_t)_{t\in\mathbb{N}}$)

At time t do:

1. Let $w_t(i) = \exp(Q_t(i)/\tau_t)$, $i \in \{1, 2, \ldots, k\}$
2. Let $p_t(i) = \frac{w_t(i)}{\sum_j w_t(j)}$
3. Draw I_t from $p_t(\cdot)$

Comments

- $\tau_t \to 0$; "computational temperature"
- aka exponential weights algorithm, Gibbs-selection
- Plays a big role in adversarial environments

OFU: Optimism in the face of uncertainty

Idea [Lai and Robbins, 1985]

Choose the arm with the best potential

- Assumption: $R_t(i) \sim p_{\theta_i}(\cdot)$, $\theta_i \in \mathbb{R}$ unknown, p known
- Mean payoff: $Q(\theta) = \int r\, p_\theta(r) dr$
- Uncertainty set:

$$U_{i,t} = \{\theta \,|\, \log \mathbb{P}\left(R_1(i), \ldots, X_{T_i(t)}(i) \,|\theta\right) \text{ is ``large'' } \}$$

... "large" depends on t, $T_i(t)$.

- 'Upper confidence index for arm i:

$$\mathrm{UCI}_t(i) = \max_{\theta \in U_{i,t}} Q(\theta)$$

- Algorithm UCI: $I_t := \mathrm{argmax}_i \mathrm{UCI}_t(i)$.
- Two reasons we select an arm: (i) Associated uncertainty is big, (ii) the arm looks good. Is this a good algorithm??

REGRET BOUND FOR UCI [LAI AND ROBBINS, 1985]

THEOREM

For any suboptimal arm i,

$$\mathbb{E}\left[T_n(i)\right] \leq \left(\frac{1}{D(p_i \| p^*)} + o(1)\right) \log(n),$$

where

$$D(p_i \| p^*) = \int p_i(x) \log \frac{p_i(x)}{p^*(x)} dx,$$

$p_i = p_{\theta_i}$ *and* p^* *is the distribution of an optimal arm.*

COROLLARY

$$L_n \leq \left(\sum_{i:\Delta_i > 0} \Delta_i \left\{\frac{1}{D(p_i \| p^*)} + o(1)\right\}\right) \log n.$$

A lower bound [Lai and Robbins, 1985]

$B = (p_1, \ldots, p_K)$; $L_n^{\mathcal{A}}(B)$: regret of $\mathcal{A}$ when run on problem B.

Definition (Uniformly good algorithms)

Algorithm $\mathcal{A}$ is uniformly good if $L_n^{\mathcal{A}}(B) = o(n^a)$ holds for all $a > 0$ and reward distributions $B = (p_1, \ldots, p_K) \in \mathcal{B}$.

This is a minimum requirement!

Theorem (Lower bound)

If $\mathcal{A}$ is uniformly good then for any $B = (p_1, \ldots, p_K) \in \mathcal{B}$ and any i,

$$\mathbb{E}\left[T_i(n)\right] \geq \left(\frac{1}{D(p_i \| p^*)} - o(1)\right) \log n.$$

Corollary

UCI algorithms are asymptotically efficient.

UCB: A nonparametric implementation of OFU

- Goal: Avoid parametric distributions!
- How to implement the OFU principle?!

  What is the "potential" of an arm?

  $\Rightarrow$ Need upper estimates of its mean payoff!
- [Agrawal, 1995] Large-deviation theory $\Rightarrow$ asymptotic results
- [Auer et al., 2002] Avoid asymptotics, use Hoeffding's inequality!
- Hoeffding:

$$Q(i) \leq Q_t(i) + \sqrt{\frac{\log(k/\delta_t)}{2T_t(i)}}$$

- $\mathrm{UCI}_t(i) := Q_t(i) + \sqrt{p\,\frac{\log(k/\delta_t)}{2T_t(i)}}$

  $p > 1$, $\delta_t \to 0$ tuning parameters

ALGORITHM UCB1

UCB1(p)

At time t do:

1. $\mathrm{UCI}_t(i) := Q_t(i) + \sqrt{\frac{p\log(t)}{2T_t(i)}}$
2. $I_t := \mathrm{argmax}_i \mathrm{UCI}_t(i)$

THEOREM (UCB1 REGRET)

Let $0 \leq R_t(i) \leq 1$*, i.i.d., independent between the arms. Then the regret of UCB1 satisfies*

$$\mathbb{E}\left[L_n\right] \leq 2p \left(\sum_{i:\Delta_i>0} \frac{1}{\Delta_i} \right) \log(n) + \left(3 + \frac{2}{p-2}\right) \sum_{i=1}^{K} \Delta_i .$$

- (Slightly) improved bound compared to [Auer et al., 2002]
- One can show that $1/D(p_j \| p^*) \leq 1/(2\Delta_j^2)$
- Still far from the best possible constant $(1/2)$.

UCBTuned: Motivation

Assume that rewards are in $[a, a+b]$.
UCB1 needs to be adjusted:

$$\mathrm{UCI}_t(i) := Q_t(i) + b\sqrt{\frac{p\log(t)}{2T_t(i)}}.$$

Regret bound:

$$R_n \leq 2p\left(\sum_{i:\Delta_i>0}\frac{b^2}{\Delta_i}\right)\log(n) + \left(3+\frac{2}{p-2}\right)b\sum_{i=1}^{K}\Delta_i.$$

Problem: outrageously big when $\mathrm{Var}\left[R_t(i)\right] \ll b^2$!

UCBTuned: Algorithm

Idea: Estimate variance and use it to define the upper index!

Theorem (Empirical Bernstein bound [Audibert et al., 2007])

Let $a \leq X_t \leq a + b$ be ii.i.d., $t > 2$. Let $\overline{X}_t$ be the empirical mean of $X_1, \ldots, X_t$, $V_t = 1/t \sum_{s=1}^{t} (X_s - \overline{X}_t)^2$ be the empirical variance estimate. Then for any $0 < \delta < 1$, w.p. at least $1 - \delta$,

$$|\overline{X}_t - \mathbb{E}[X_1]| \leq \sqrt{\frac{2V_t x_\delta}{t}} + \frac{3bx_\delta}{t},$$

where $x_\delta = \log(3/\delta)$.

UCBTuned(p)

At time t do:

1. $\mathrm{UCI}_t(i) := Q_t(i) + \sqrt{\frac{2V_t(i)p\log t}{T_t(i)}} + \frac{3bp\log t}{T_t(i)}$
2. $I_t := \mathrm{argmax}_i \mathrm{UCI}_t(i)$

UCBTuned: Regret bound

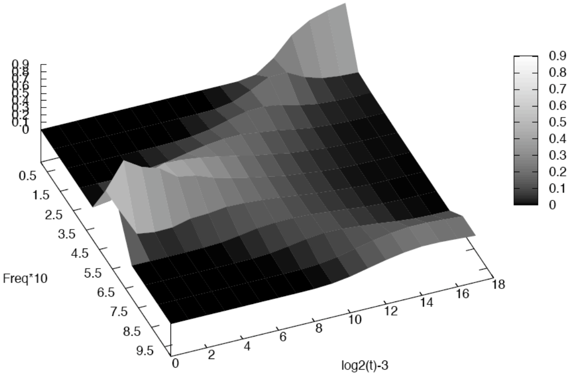
Theorem (UCBTuned Regret [Audibert et al., 2007])

Let $0 \leq R_t(i) \leq b$ be i.i.d., independent between the arms, $p > 1$, $\sigma_i^2 = \mathrm{Var}\left[R_t(i)\right]$. Then the regret of UCBTuned(p) satisfies

$$\mathbb{E}\left[L_n\right] \leq c_p \sum_{i:\Delta_i>0} \left(\frac{\sigma_i^2}{\Delta_i} + 2b\right) \log n.$$

In particular, when $p = 1.2$, $c_p = 10$.

WHAT REALLY HAPPENS..



Distribution of $T_t(2)/t$ plotted against time.

Bandit: $R_t(1) \sim \mathrm{Ber}(0.5)$, $R_t(2) = 0.495$.

EXTENSIONS

- Switching costs [Agrawal et al., 1988]
- Dependent rewards [Lai and Yakowitz, 1995]
- Continuous action spaces
  - Discretization [Kleinberg, 2004, Auer et al., 2007b]
  - Tree-based methods [Kocsis and Szepesvári, 2006, Bubeck et al., 2008]
  - Linear mean-payoff $Q(\cdot)$ [Dani et al., 2008]
- Drifts [Garivier and Moulines, 2008]
- Side information $\Rightarrow$ see below
- Feedback $\Rightarrow$ MDPs; see below

SUMMARY

- Forcing, ϵ-greedy:
  - Exploration schedule provides a lower bound on regret
  - Requires loss tolerance $\Leftrightarrow$ tuning
- OFU-based algs:
  - No fixed schedule for exploration $\Rightarrow$ “adaptivity”
  - Advantage: Minimal tuning
- You can mix these ideas!

Bandits with side information ($\equiv$ associative bandits)

Protocol of learning

Perception-action loop:

1. Agent senses Y_t coming from the Environment
2. Agent sends action A_t to the Environment
3. Agent receives reward R_t from the Environment
4. $t := t + 1$, go to Step 1

Assumption

Sensations are not influenced by the agent's decisions

Variants

- Y_t is from a m-element set $\mathcal{Y} \Rightarrow m$ independent bandit problems
- Dependent payoffs; e.g., $Q : \mathcal{Y} \times \mathcal{A} \to \mathbb{R}$, $Q(\cdot, a)$ is "smooth" [Yang and Zhu, 2002], linear [Auer, 2003]
- Continuous action sets
- Delayed feedback [György et al., 2007]

Online learning in MDPs

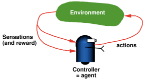
Assumption: Environment is a finite MDP $(\mathcal{X}, \mathcal{A}, p, r)$.

Protocol of learning

Perception-action loop:

1. Agent senses state $X_t \sim p(\cdot|X_{t-1}, A_{t-1})$ of the Environment
2. Agent sends action A_t to the Environment
3. Agent receives reward R_t from the Environment ($\mathbb{E}\left[R_t | X_t = x, A_t = a\right] = r(x, a)$).
4. $t := t + 1$, go to Step 1

Goal: Minimize regret $L_n = n\lambda^* - \sum_{t=1}^{n} R_t$,
where λ^* is the best possible reward per time step



Average reward MDPs

Assumption

Irreducibility of the Markov chains under stationary policies

Average reward Bellman optimality equations (BOE)

Find $\lambda^* \in \mathbb{R}$, $h^* : \mathcal{X} \to \mathbb{R}$,

$$\begin{aligned} \lambda^* + h^*(x) &= \max_{a \in \mathcal{A}(x)} \left[r(x,a) + \langle\, p_x(a)\,,\, h^* \,\rangle \right] \\ &= \max_{a \in \mathcal{A}(x)} Q^*(x,a), \quad x \in \mathcal{X}. \end{aligned}$$

Restricted problem

$D(x) \subset A(x) \Rightarrow h^*_D$, Q^*_D.

OPTIMISTIC LINEAR PROGRAMMING [TEWARI AND BARTLETT, 2007]

ALGORITHM OLP

1. Update $\hat{p}$, the estimate of transition probabilities
2. $D(x) := \{a \in A(x) \mid T_t(x,a) \geq \log^2 T_t(x)\},\ x \in \mathcal{X}$
   // keep "well-sampled" actions only
3. $(\lambda, h, Q) := BOE(D, \hat{p})$ // solve the Bellman equations
4. $\mathrm{UCI}(a) :=$
   $\sup\left\{ r(X_t, a) + \langle\, q\,,\, h \,\rangle \mid q \in \Delta_1, \|\hat{p}_{X_t}(a) - q\|_1 \leq \sqrt{\frac{2\log t}{T_t(X_t,a)}} \right\}$
5. $\Gamma := \left\{ a \in A(X_t) \,\middle|\, Q(X_t, a) = \lambda + h(X_t), T_t(X_t, a) < \log^2(T_t(X_t) + 1) \right\}$
6. if $A(X_t) = \Gamma$ /* all actions are in danger */ then

   let A_t be some element of Γ

   else

   $A_t := \mathrm{argmax}_{a \in A(X_t)} \mathrm{UCI}(a)$.

Regret bounds for OLP

$$
\begin{aligned}
\Delta^*(x,a) &= h^*(x) + \lambda^* - Q^*(x,a) \\
Z_\epsilon(x,a) &= \{q \in \Delta_1 \mid r(x,a) + \langle\, q\,,\, h^* \,\rangle \geq h^*(x) - \epsilon\} \\
C &= \{(x,a) \mid Q^*(x,a) < \lambda^* + h^*(x), \forall \epsilon > 0, Z_\epsilon(x,a) \neq \emptyset\} \\
J_{x,a}(p;\epsilon) &= \inf\{\|p - q\|_1 \mid q \in Z_\epsilon(x,a)\} \\
K(x,a) &= \lim_{\epsilon \longrightarrow 0} J_{x,a}(p_x(a);\epsilon) \\
H &= \sum_{(x,a)\in C} \frac{2\Delta^*(x,a)}{K(x,a)}
\end{aligned}
$$

Theorem ([Tewari and Bartlett, 2007])

Let the MDP M be irreducible. If L_n is the regret of OLP after n steps then

$$
\limsup_{n\to\infty} \frac{L_n}{\log n} \leq H.
$$

RL
AI

INTERPRETATION

PROPOSITION

Assume $A(x) = \mathcal{A}$ for all $x \in \mathcal{X}$. Let $\Delta^* = \min_{(x,a)\in C} \Delta^*(x, a)$. Then

$$H \le \frac{2|\mathcal{X}|\,|\mathcal{A}|\,\|h^*\|_1^2}{\Delta^*}.$$

COMMENT

The algorithm closely follows that of [Burnetas and Katehakis, 1997] who proved asymptotic efficiency for their policy.

The UCRL2 Algorithm [Auer et al., 2007a]

Motivation

- Explore OFU in MDPs
- High probability bounds for finite horizon problems

Algorithm UCRL2(δ, n)

- Phase initialization:
  1. Estimate mean model $\hat{p}_t$ using ML
  2. $\mathcal{U} := \{ p \,|\, \|p(\cdot|x,a) - \hat{p}_t(\cdot|x,a)\|_1 \leq c|\mathcal{X}| \frac{\log(|\mathcal{A}|n/\delta)}{T_t(x,a)} \}$
  3. $p' := \operatorname{argmax}_{p \in \mathcal{U}} \lambda^*(p), \pi := \pi^*(p')$
  4. $C(x,a) := T_t(x,a)$
- Execution of a phase
  1. Execute π until some (x,a) gets visited more than $C(x,a)$ times.

UCRL2 Results

Definition (Diameter)

Let M be an MDP. Then the diameter of M is

$$D(M) = \max_{x,y} \min_{\pi} \mathbb{E}\left[T(x \to y; \pi)\right].$$

Results:

- Lower bound:

$$\mathbb{E}\left[L_n\right] = \Omega(\sqrt{D|\mathcal{X}|\,|\mathcal{A}|\,n})$$

- Upper bounds:
  - w.p. $1 - \delta/n$,

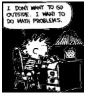
$$L_n \leq O\left(D\,|\mathcal{X}|\,\sqrt{|\mathcal{A}|\,n \log(|\mathcal{A}|n/\delta)}\right)$$

  - w.p. $1 - \delta$,

$$L_n \leq O\left(D^2\,|\mathcal{X}|^2\,|\mathcal{A}|\frac{\log(|\mathcal{A}|n/\delta)}{\Delta^*}\right),$$

where $\Delta^* = \min_{\pi:\lambda^\pi < \lambda^*} \lambda^* - \lambda^\pi$.

CONCLUSIONS



- Exploration is necessary
- .. but should be controlled in a wise manner
- Tools:
  - Forced exploration
  - Softmax
  - Optimism in the face of uncertainty
- After 50 years still lots of work to be done!
  - Scaling up
  - Non-stationary environments
  - Dependencies

REFERENCES I

Agrawal, R. (1995).
Sample mean based index policies with $O(logn)$ regret for the multi-armed bandit problem.
Advances in Applied Probability, 27:1054–1078.

Agrawal, R., Hedge, M., and Teneketzis, D. (1988).
Asymptotically efficient adaptive allocation rules for the multiarmed bandit problem with switching cost.
IEEE Transactions on Automatic Control, 33(10):899–906.

Audibert, J.-Y., Munos, R., and Szepesvári, C. (2007).
Tuning bandit algorithms in stochastic environments.
In *Algorithmic Learning Theory - 18th International Conference, ALT 2007*, pages 150–165. Springer.

Auer, P. (2003).
Using confidence bounds for exploitation-exploration trade-offs.
Journal of Machine Learning Research, 3:397–422.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002).
Finite time analysis of the multiarmed bandit problem.
Machine Learning, 47(2-3):235–256.

Auer, P., Jaksch, T., and Ortner, R. (2007a).
Near-optimal regret bounds for reinforcement learning.
Technical report, University of Leoben.

Auer, P., Ortner, R., and Szepesvári, C. (2007b).
Improved rates for the stochastic continuum-armed bandit problem.
In *Proceedings of the 20th Annual Conference on Learning Theory (COLT-07)*, pages 454–468.

References II

Bubeck, S., Munos, R., and Szepesvári, C. (2008).
Online optimization in X-armed bandits.
In *NIPS-21*.

Burnetas, A. and Katehakis, M. (1997).
Optimal adaptive policies for Markov Decision Processes.
Mathematics of Operations Research, 22(1):222–255.

Dani, V., Hayes, T., and Kakade, S. (2008).
Stochastic linear optimization under bandit feedback.
COLT-2008.

Garivier, A. and Moulines, E. (2008).
On upper-confidence bound policies for non-stationary bandit problems.
Technical report, LTCI.

György, A., Kocsis, L., Szabó, I., and Szepesvári, C. (2007).
Continuous time associative bandit problems.
In *IJCAI-07*, pages 830–835.

Kleinberg, R. (2004).
Nearly tight bounds for the continuum-armed bandit problem.
In *NIPS-2004*.

Kocsis, L. and Szepesvári, C. (2006).
Bandit based Monte-Carlo planning.
In *Proceedings of the 17th European Conference on Machine Learning (ECML-2006)*, pages 282–293.

Lai, T. and Yakowitz, S. (1995).
Machine learning and nonparametric bandit theory.
IEEE Transactions on Automatic Control, 40:1199–1209.

References III

Lai, T. L. and Robbins, H. (1985).
Asymptotically efficient adaptive allocation rules.
Advances in Applied Mathematics, 6:4–22.

Robbins, H. (1952).
Some aspects of the sequential design of experiments.
Bulletin of the American Mathematics Society, 58:527–535.

Tewari, A. and Bartlett, P. (2007).
Optimistic linear programming gives logarithmic regret for irreducible mdps.
NIPS-20.

Yang, Y. and Zhu, D. (2002).
Randomized allocation with nonparametric estimation for a multi-armed bandit problem with covariates.
Annals of Statistics, 30(1):100–121.