Learning Theory of Optimal Decision Making

Part I: On-line Learning in Stochastic Environments

Csaba Szepesvári

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
1 High level overview of the talks

2 Motivation
   - What is it?
   - Why should we care?
   - The challenge

3 Bandits
   - Forcing
   - $\epsilon$-greedy
   - Softmax
   - “Optimism in the Face of Uncertainty”

4 Bandits with side information

5 Online learning in MDPs

6 Conclusions
"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)
Day 1: Online learning in stochastic environments
Day 2: Batch learning in Markovian Decision Processes
Day 3: Online learning in adversarial environments
Union bound: $\Pr(A \cup B) \leq \Pr(A) + \Pr(B)$

Inclusion: $A \subset B \Rightarrow \Pr(A) \leq \Pr(B)$

Inversion: If $\Pr(X > t) \leq F(t)$ holds for all $t$ then for any $0 < \delta < 1$, w.p. $1 - \delta$, $X \leq F^{-1}(\delta)$

Expectation: If $X \geq 0$, $\mathbb{E}[X] = \int_0^\infty \Pr(X \geq t) \, dt$
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: \( \sum_{t=1}^{T} R_t \rightarrow \text{max} \)

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: \( i \in I \overset{\text{def}}{=} \{1, 2, \ldots, k\} \)

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 Problems

Bandit Problem [Robbins, 1952]

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

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

- Clinical trials
- Web advertising
- Job shop scheduling

:
**Some Definitions**

- **Expected payoffs**: $Q(i) = \mathbb{E}[R_1(i)]$
- **Optimal arm**: $Q(i^*) = Q^* \overset{\text{def}}{=} \max_i Q(i)$
- **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

$$\lim_{t \to \infty} \mathbb{P}(l_t \in l^*) \to 1$$

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:

$$L_n \overset{\text{def}}{=} n Q^* - \mathbb{E}[R_1:n] = o(n)$$

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

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

**Proof.**

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

$$a_t = \sum_i \mathbb{E} [Q^* - R_t(i)|l_t = i] \mathbb{P} (l_t = i)$$

$$= \sum_{i: \Delta_i > 0} \Delta_i \mathbb{P} (l_t = i).$$

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 $B$ contain a non-trivial bandit problem. If an algorithm stops exploring it cannot be consistent on $B$.
Methods for achieving consistency

- Forcing
  - Fixed schedule
  - $\epsilon$-greedy
- Softmax
- “Optimism in the Face of Uncertainty”
**FORCING**

**Idea**
Exploration is necessary ⇒ make sure that every arm is selected infinitely often

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

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

At time \( t \) do:

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

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

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

hence

\[
P \left( \bar{X}_n - \mathbb{E} [X_1] \geq \epsilon \right) = P \left( \sqrt{\frac{n}{\sigma^2}} \left( \bar{X}_n - \mathbb{E} [X_1] \right) \geq \sqrt{\frac{n}{\sigma^2}} \epsilon \right)
\]

\[
\approx e^{-n\epsilon^2/(2\sigma^2)} \frac{\sigma^2}{n\epsilon^2} \approx e^{-n\epsilon^2/(2\sigma^2)}.
\]
**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

\[
\begin{align*}
\mathbb{P}(\overline{X}_n \geq \mu + \epsilon) & \leq e^{-2n\epsilon^2} \\
\mathbb{P}(\overline{X}_n \leq \mu - \epsilon) & \leq e^{-2n\epsilon^2}.
\end{align*}
\]

**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 $i \in \{1, \ldots, k\}$

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

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 $c_t < \Delta^*/2 \overset{\text{def}}{=} \min_{j: \Delta_j > 0} \Delta_j/2 \iff$
   
   (*) $f_t \geq 2 \log(2k/\delta_t)/(\Delta^*)^2 \Rightarrow$

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

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((ϵ_t)_{t∈N})

At time t do:

1. Draw U_t in [0, 1] uniformly at random
2. if U_t < ϵ_t then
   Pick l_t randomly from \{1, 2, \ldots, k\}
3. else
   \[ l_t := \text{argmax}_i Q_t(i). \]

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

Assume that the payoffs are bounded. If \( ϵ_t = c'/t \) with \( c' = Ω(k/(Δ^*)^2) \) then

\[ \mathbb{P}(l_t \notin l^*) \leq O\left(\frac{c'}{n}\right) \quad \text{and} \quad L_n = O(c' \log n). \]
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 $l_t$ from $p_t(\cdot)$

COMMENTS

- $\tau_t \to 0$; “computational temperature”
- aka exponential weights algorithm, Gibbs-selection
- Plays a big role in adversarial environments
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:**
  \[
  U_{i,t} = \{ \theta \mid \log \mathbb{P} (R_1(i), \ldots, X_{T_i(t)}(i) \mid \theta) \text{ is “large”} \}
  \]
  ... “large” depends on \( t, T_i(t) \).
- ‘**Upper confidence index**’ for arm \( i \):
  \[
  UCI_t(i) = \max_{\theta \in U_{i,t}} Q(\theta)
  \]
- **Algorithm UCI:** \( l_t := \arg\max_i UCI_t(i) \).
- 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 $i$,

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

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]

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

**Definition (Uniformly good algorithms)**

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

This is a minimum requirement!

**Theorem (Lower bound)**

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

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

**Corollary**

UCI algorithms are asymptotically efficient.
Goal: Avoid parametric distributions!
How to implement the OFU principle?!

What is the “potential” of an arm?
⇒ Need upper estimates of its mean payoff!

[Agrawal, 1995] Large-deviation theory ⇒ asymptotic results

[Auer et al., 2002] Avoid asymptotics, use Hoeffding’s inequality!

Hoeffding:

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

\[ UCI_t(i) := Q_t(i) + \sqrt{p \frac{\log(k/\delta_t)}{2T_t(i)}} \]

\[ p > 1, \delta_t \to 0 \text{ tuning parameters} \]
**Algorithm UCB1**

**UCB1** ($p$)

At time $t$ do:

1. $UCI_t(i) := Q_t(i) + \sqrt{\frac{p \log(t)}{2T_t(i)}}$
2. $l_t := \text{argmax}_i UCI_t(i)$

**Theorem (UCB1 Regret)**

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

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

- (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 $(1/2)$.
Assume that rewards are in $[a, a + b]$. UCB1 needs to be adjusted:

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

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 $\text{Var}[R_t(i)] \ll b^2$!
**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(\( \rho \))**

At time \( t \) do:

1. \( \text{UCI}_t(i) := Q_t(i) + \sqrt{\frac{2V_t(i)p \log t}{T_t(i)}} + \frac{3bp \log t}{T_t(i)} \)
2. \( l_t := \text{argmax}_i \text{UCI}_t(i) \)
**Theorem (UCBTuned Regret [Audibert et al., 2007])**

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

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

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

WHAT REALLY HAPPENS..

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

Bandit: $R_t(1) \sim \text{Ber}(0.5)$, $R_t(2) = 0.495$. 
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
Forcing, $\epsilon$-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 (≡ 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., $Q : \mathcal{Y} \times A \rightarrow \mathbb{R}$, $Q(\cdot, a)$ is “smooth” [Yang and Zhu, 2002], linear [Auer, 2003]
- Continuous action sets
- Delayed feedback [György et al., 2007]
**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
   \((\mathbb{E}[R_t | X_t = x, A_t = a] = r(x, a))\).
4. \(t := t + 1\), go to Step 1

**Goal:** Minimize regret \(L_n = n\lambda^* - \sum_{t=1}^{n} R_t\), where \(\lambda^*\) 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 $\lambda^* \in \mathbb{R}$, $h^* : \mathcal{X} \rightarrow \mathbb{R}$,

\[
\lambda^* + h^*(x) = \max_{a \in A(x)} [r(x, a) + \langle p_x(a), h^* \rangle]
\]

\[
= \max_{a \in A(x)} Q^*(x, a), \quad x \in \mathcal{X}.
\]

**Restricted problem**
\[D(x) \subset A(x) \Rightarrow h^*_D, Q^*_D.\]
Algorithm OLP

1. Update $\hat{\rho}$, 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. $(\lambda, h, Q) := BOE(D, \hat{\rho})$ // solve the Bellman equations
4. $UCI(a) :=$
   \[
   \sup \left\{ r(X_t, a) + \langle q, h \rangle \mid q \in \Delta_1, \| \hat{\rho}_{X_t}(a) - q \|_1 \leq \sqrt{\frac{2 \log t}{T_t(X_t, a)}} \right\}
   \]
5. $\Gamma := \{ a \in A(X_t) \mid Q(X_t, a) = \lambda + h(X_t), T_t(X_t, a) < \log^2(T_t(X_t) + 1) \}$
6. if $A(X_t) = \Gamma$ /* all actions are in danger */ then
   let $A_t$ be some element of $\Gamma$
   else
   $A_t := \arg\max_{a \in A(X_t)} UCI(a)$. 
\[
\begin{align*}
\Delta^*(x, a) &= h^*(x) + \lambda^* - Q^*(x, a) \\
Z_\epsilon(x, a) &= \{ q \in \Delta_1 \mid r(x, a) + \langle q, h^* \rangle \geq h^*(x) - \epsilon \} \\
C &= \{(x, a) \mid Q^*(x, a) < \lambda^* + h^*(x), \forall \epsilon > 0, Z_\epsilon(x, a) \neq \emptyset \} \\
J_{x,a}(p; \epsilon) &= \inf\{\|p - q\|_1\mid q \in Z_\epsilon(x, a)\} \\
K(x, a) &= \lim_{\epsilon \to 0} J_{x,a}(p_x(a); \epsilon) \\
H &= \sum_{(x,a) \in C} \frac{2\Delta^*(x, a)}{K(x, a)}
\end{align*}
\]

**Theorem ( [Tewari and Bartlett, 2007])**

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

\[
\limsup_{n \to \infty} \frac{L_n}{\log n} \leq H.
\]
**PROPOSITION**

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

$$H \leq \frac{2|\mathcal{X}| |\mathcal{A}| \left\| h^* \right\|_1^2}{\Delta^*}.$$ 

**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. \( \mathcal{U} := \{ p | \| p(\cdot| x, a) - \hat{p}_t(\cdot| x, a) \|_1 \leq c |\mathcal{X}| \log(|A|n/\delta) / T_t(x, a) \} \)
3. \( p' := \arg\max_{p \in \mathcal{U}} \lambda^*(p), \pi := \pi^*(p') \)
4. \( C(x, a) := T_t(x, a) \)

Execution of a phase

1. Execute \( \pi \) 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

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

**Results:**

- **Lower bound:**
  $$\mathbb{E} [L_n] = \Omega(\sqrt{D|X||A|n})$$

- **Upper bounds:**
  - w.p. $1 - \delta/n$,
    $$L_n \leq O \left( D|X| \sqrt{|A|n \log(|A|n/\delta)} \right)$$
  - w.p. $1 - \delta$,
    $$L_n \leq O \left( D^2 |X|^2 |A| \frac{\log(|A|n/\delta)}{\Delta^*} \right),$$

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


Online optimization in X-armed bandits.
In NIPS-21.

Optimal adaptive policies for Markov Decision Processes.

Stochastic linear optimization under bandit feedback.
COLT-2008.

On upper-confidence bound policies for non-stationary bandit problems.
Technical report, LTCI.

Continuous time associative bandit problems.
In IJCAI-07, pages 830–835.

Nearly tight bounds for the continuum-armed bandit problem.
In NIPS-2004.

Bandit based Monte-Carlo planning.

Machine learning and nonparametric bandit theory.

