Optimal adaptive control in discrete systems

Alexandre Proutiere

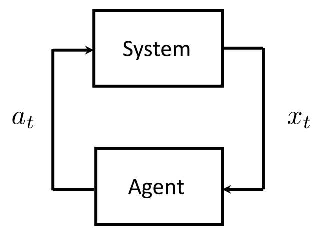
Outline

What can we say about adaptive control of systems with finite state and action spaces?

1. Learning problems
2. Information-theoretical limits
3. Algorithms

Adaptive control (Reinforcement learning)

Learning an optimal control strategy under **unknown** system dynamics and cost function



Finite state and action spaces
Dynamics: $x_{t+1} \sim p(\cdot|x_t, a_t)$
Cost: $(c(x_t, a_t) + \xi_t) \sim q(\cdot|x_t, a_t)$
p and q are initially unknown

Objectives

Learn as fast as possible a policy $\pi^\star : \mathcal{S} \to \mathcal{A}$ maximizing over all possible π

(Average cost) $\liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_x^\pi[c(X_t, A_t)]$

Performance metrics. Sample complexity (PAC framework) or regret

$$\text{Regret of } \pi : R_T^\pi(x) = \mathbb{E}_x^\pi[\sum_{t=1}^{T} c(X_t, A_t)] - \mathbb{E}_x^{\pi^\star}[\sum_{t=1}^{T} c(X_t, A_t)]$$

Structured RL

The *structure* ties the system dynamics and costs at the various (state, action) pairs together. This may speed up the exploration process.

> *Observations at a given (state, action) pair provides useful side-information at other pairs.*

Examples of structure:

- The Q-function belongs to a parametrized set (Deep RL)
- (p, q) are smooth, convex, unimodal, ...
- Linear system, quadratic cost
- ...

Structures

The decision maker knows that $\phi = (p_\phi, q_\phi) \in \Phi$.
Φ encodes the structure.

Examples.

1. Unstructured MDPs. $\phi \in \Phi$ iff for all (x, a), $p_\phi(\cdot|x, a) \in \mathcal{P}(\mathcal{S})$ and $q_\phi(\cdot|x, a) \in \mathcal{P}([0, 1])$.

2. Lipschitz MDPs. $\phi \in \Phi$ iff $p_\phi(\cdot|x, a)$ and $c_\phi(x, a)$ are Lipschitz-continuous:

$$(L1) \qquad \|p_\phi(\cdot|x, a) - p_\phi(\cdot|x', a')\|_1 \le Ld(x, x')^\alpha + L'd(a, a')^{\alpha'}$$

$$(L2) \qquad |c_\phi(x, a) - c_\phi(x', a')| \le Ld(x, x')^\alpha + L'd(a, a')^{\alpha'}$$

Unstructured discounted RL: State-of-the-art

- Minimax lower bound for sample complexity and Q-sample complexity: $\Omega\left(\frac{SA}{(1-\lambda)^3\varepsilon^2}\log\delta^{-1}\right)$.
  (no problem-specific lower bound so far)

- Algorithms:
  - MBIE (**Strehl-Li-Littman**'05): $SC = \mathcal{O}\left(\frac{S^2A}{(1-\lambda)^6\varepsilon^3}\log\delta^{-1}\right)$
  - MoRmax (**Szita-Szepesvari**'10): $SC = \widetilde{\mathcal{O}}\left(\frac{SA}{(1-\lambda)^6\varepsilon^2}\log\delta^{-1}\right)$
  - Q-learning[1]: $QSC = \widetilde{\mathcal{O}}\left(\frac{SA}{(1-\lambda)^5\varepsilon^{5/2}}\mathrm{polylog}\delta^{-1}\right)$
  - Speedy Q-Learning (**Azar et al.**'11): $\widetilde{\mathcal{O}}\left(\frac{SA}{(1-\lambda)^4\varepsilon^2}\mathrm{polylog}\delta^{-1}\right)$
  - (**Sidford et al.**'18): $\widetilde{\mathcal{O}}\left(\frac{SA}{(1-\lambda)^3\varepsilon^2}\log\delta^{-1}\right)$

[1] with optimized learning rate $\alpha_t = 1/(t+1)^{4/5}$

Unstructured average-reward RL: State-of-the-art

- Regret lower bounds
  - Problem-specific with known costs (**Burnetas-Katehakis**'97): $c_\phi \log(T)$
  - Minimax: $\Omega(\sqrt{DSAT})$ (D: diameter)
- No lower bound on sample complexity
- Algorithms:
  - Asymptotically optimal algorithm (**Burnetas-Katehakis**'97)
  - UCRL2 (**Auer-Jaksch-Ortner**'10): $\mathcal{O}\left(\frac{D^2 S^2 A}{\Delta}\log(T)\right)$ and $\widetilde{\mathcal{O}}\left(DS\sqrt{AT}\right)$
  - AJ (**Agrawal-Jia**'17): $\widetilde{\mathcal{O}}\left(D\sqrt{SAT}\right)$
  - Adversarial MDPs (changing every round). **Abbasi-Yadkori**'13: $\widetilde{\mathcal{O}}\left(D\sqrt{SAT}\right)$
- All aforementioned results are for unstructured systems!

Outline

1. Learning problems and performance metrics
2. **Information-theoretical limits**
3. Algorithms

Fundamental limits

Let $\phi \in \Phi$ be an unknown MDP with known structure.

Are there fundamental limits when learning an optimal policy?
Can we derive sample complexity and regret lower bounds?

A generic method to derive problem-specific limits, illustrated on the regret minimization problem

Information constraints

How much must a *uniformly good* algorithm explore (state, action) pair (x, a) in its learning process?

Uniformly good $=$ adaptive, i.e., has reasonable performance on all systems.

<u>Example (regret in average-cost RL):</u> regret $o(T^{\alpha})$ for all $\alpha > 0$ and all $\phi \in \Phi$.

Information constraints

Data processing inequality: Let O_t be the observations up to round t. For all systems $\phi, \psi \in \Phi$, for any event $E \in \sigma(O_t)$,

$$\mathbb{E}_\phi\big[\log \frac{\mathbb{P}_\phi[O_t]}{\mathbb{P}_\psi[O_t]}\big] \geq kl(\mathbb{P}_\phi(E), \mathbb{P}_\psi(E)).$$

A constraint is effective only for ψ such that:

- $\phi \ll \psi$ so that the l.h.s. not infinite
- $\Pi^\star(\phi) \cap \Pi^\star(\psi) = \emptyset$ so that the r.h.s. is as large as possible
  <u>Example (regret in ergodic RL):</u> for uniformly good algorithms r.h.s. $\sim \log(t)$ for the best choice of $E = \{$opt. actions for ϕ taken often$\}$

Towards regret lower bound in average-cost RL

- Information constraints: for all $\psi \in \Lambda_\Phi(\phi)$,

$$\mathbb{E}_\phi\big[\log \frac{\mathbb{P}_\phi[O_t]}{\mathbb{P}_\psi[O_t]}\big] = \sum_{(x,a)} \mathbb{E}_\phi^\pi[N_t(x,a)] KL_{\phi|\psi}(x,a) \geq \log(t)(1+o(1)),$$

where

$$\begin{cases} \Lambda_\Phi(\phi) = \{\psi \in \Phi : \phi \ll \psi, \Pi^\star(\phi) \cap \Pi^\star(\psi) = \emptyset\} \\ KL_{\phi|\psi}(x,a) = KL(p_\phi(\cdot|x,a), p_\psi(\cdot|x,a)) + KL(q_\phi(\cdot|x,a), q_\psi(\cdot|x,a)) \end{cases}$$

- Objective function: $\delta^\star(x,a;\phi)$ is the regret induced by action a in x:

$$\delta^*(x,a;\phi) = (\mathbf{B}_\phi^\star h_\phi^\star)(x) - (\mathbf{B}_\phi^a h_\phi^\star)(x)$$

where $(\mathbf{B}_\phi^a h)(x) = c_\phi(x,a) + \sum_y p_\phi(y|x,a)h(y)$ (Bellman operator)

Regret lower bound

Theorem *Any uniformly good algorithm exhibits a regret asymptotically greater than* $K_\Phi(\phi)\log(T)$ *where* $K_\Phi(\phi)$ *solves:*

$$\begin{aligned} \min_{\eta \geq 0} \quad & \sum_{x,a} \eta(x,a)\delta^\star(x,a;\phi) \\ \text{s.t.} \quad & \sum_{x,a} \eta(x,a) KL_{\phi|\psi}(x,a) \geq 1, \quad \forall \psi \in \Lambda_\Phi(\phi) \end{aligned}$$

- $\eta(x,a)\log(T)$ to be interpreted as the required number of times (x,a) should be explored.
- Valid for any given structure Φ (through $\Lambda_\Phi(\phi)$).

Impact of the structure on feasible regret

The lower bound is given by the solution of a *semi-infinite* LP:

$$P(\phi, \mathcal{F}_\Phi(\phi)) : \min_{\eta \in \mathcal{F}_\Phi(\phi)} \sum_{x,a} \eta(x,a)\delta^\star(x,a;\phi)$$

Simplifying the constraint set $\mathcal{F}_\Phi(\phi)$, we can conclude that:

- Unstructured RL problems: the best regret scales as $\frac{H^2}{\delta_{\min}} SA\log(T)$ where
  - H: span of the bias function (finite for fast mixing systems)
  - $\delta_{\min}$: minimal (state, action) suboptimal gap
- Lipschitz RL problems: the best regret scales as $f(H, \delta_{\min})\log(T)$ (independent of S and A)

An other example of application

Sample complexity in LTI system identification

Uncontrolled system. $x_{t+1} = Ax_t + w_t$.
Sample complexity. τ_A minimum time t to get $\mathbb{P}_A[\|\hat{A}_t - A\|_F \geq \epsilon] \leq \delta$.
Uniform goodness. (ϵ, δ)-locally stable at A, i.e., there exists a finite time τ such that for all $A' \in B(A, 3\epsilon)$, $\mathbb{P}_{A'}[\|\hat{A}_t - A'\|_F \geq \epsilon] \leq \delta$.

Under any (ϵ, δ)-locally stable algorithm at A, we have:

$$\lambda_{\min}\Big(\sum_{s=1}^{\tau_A - 1} \Gamma_{s-1}(A)\Big) \geq \frac{1}{2\epsilon^2} \log(\frac{1}{2.4\delta})$$

where $\Gamma_s(A) = \sum_{k=0}^{s} A^k (A^k)^\top$.

Outline

Algorithm design

Towards low regret in average-cost RL:

1. Optimism in front of uncertainty: UCRL

   "Maintain a confidence ball for $\phi = (p, q)$, solve the MDP with the best system in this ball"

   Complex and sub-optimal in the case of structured problems

2. Posterior sampling: Thompson sampling

   "Maintain a posterior for $\phi = (p, q)$, sample from it"

   Impossible to implement in the case of structured problems

3. Directed Exploration Learning: exploiting regret lower bounds

   "Maintain an estimate ϕ_t of the system ϕ; solve the regret LB optimization problem and explore according to its solution"

Directed Exploration Learning

DEL: An algorithm that targets the exploration rates predicted by the lower bound optimization problem.

In round t:

1. Estimate the system: ϕ_t
2. Solve $\min_{\eta \in \mathcal{F}_\Phi(\phi_t)} \sum_{x,a} \eta(x,a)\delta^\star(x,a;\phi_t)$ or a simplified problem, solution $(\eta_t(x,a))_{x,a}$
3. Select an action:
   - If $N_t(X_t,a) \geq \eta_t(X_t,a)$ for all a, exploit: pick the best action seen so far
   - Else explore: pick an action such that $N_t(X_t,a) < \eta_t(X_t,a)$

Asymptotically optimal and flexible (may tune the regret-complexity trade-off by selecting simplified LB optimization problems)

Analysis via concentration inequalities

- Basic inequalities (e.g. Hoeffding): $\mathbb{P}(\|\phi_t - \phi\| \geq \gamma) \leq c_1 e^{-c_2 \gamma^2 t}$
- Efficient algorithms exploit quantities of the form $\sum_{x,a} N_t(x,a) KL_{\phi_t|\psi}(x,a)$.

  Their regret analysis requires multi-dimensional concentrations of self-normalized averages:

$$\mathbb{P}\left[\sum_{x,a} N_t(x,a) KL_{\phi_t|\phi}(x,a) \geq \gamma\right] \leq e^{-\gamma} \left(\frac{(\gamma)^2 \log t}{SA}\right)^{SA} e^{SA+1}.$$

Conclusions and challenges

- Critical to exploit the structure in large-scale RL (empirically successful algorithms do it!)
- A generic two-step method:
  1. Identify problem-specific fundamental performance limits satisfied by any RL algorithm
  2. Devise algorithms that approach the optimal exploration rates dictated by these limits
- Challenges:
  - Characterizing the performance-complexity trade-off
  - Deriving tight finite-time performance guarantees
  - Letting the state and action space grow – up to being continous

References

- Exploration in structured RL, **Ok-Proutiere-Tranos**, NeurIPS, 2018
- Optimal Adaptive Policies for Markov Decision Processes, **Burnetas-Katehakis**, Maths of OR, 1997
- Near-optimal Regret Bounds for Reinforcement Learning, **Auer-Jaksh-Ortner**, NIPS, 2009
- Optimistic posterior sampling for reinforcement learning: worst-case regret bounds, **Agrawal-Jia**, NIPS, 2017
- Sample complexity lower bounds for linear system identification, **Jedra-Proutiere**, CDC, 2019
- Online Learning in Markov Decision Processes with Adversarially Chosen Transition Probability Distributions, **Abbasi-Yadkori et al.**, NIPS, 2013