Curiosity, | unobserved rewards | and ~~neural networks~~ in RL
function approximation

Csaba Szepesvári
DeepMind & University of Alberta

New Directions in RL and Control
Princeton
2019

Part I: Curiosity

"One of the striking differences between current reinforcement learning algorithms and early human learning is that animals and infants appear to explore their environments with autonomous purpose, in a manner appropriate to their current level of skills."

Models for Autonomously Motivated Exploration in Reinforcement Learning*

Peter Auer[1], Shiau Hong Lim[1], and Chris Watkins[2]

ALT 2011, invited talk by Peter

Modeling Curiosity (ALw model)

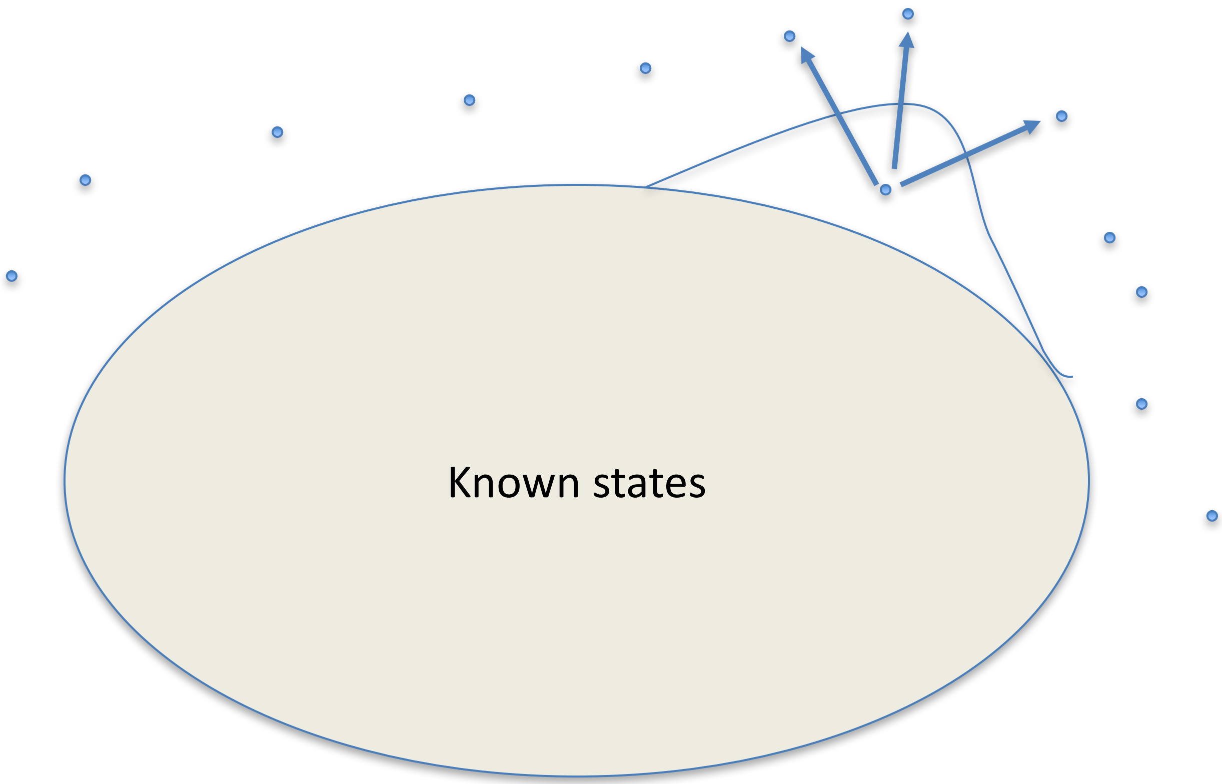
- Controlled process
- Stochasticity: Makes things more interesting/realistic
- Countably many states, they are observed
  - Simplifying assumption
  - Hope: some of the principles/algorithms transfer to the general case
  - You have to start somewhere
- Reset to an initial state
  - Necessary
  - Engineer the environment to make this happen (robot moms!)

- Goal: Extend the set of reliably reachable states as quickly as possible

Performance metric

- # Reliably reachable states/time
- Fix an arbitrary partial order, $\prec$, on states
  - Not known to learner..
- Fix $L > 0$. Define $\mathcal{S}_L^{\prec}$ as follows:
  - $s_0 \in \mathcal{S}_L^{\prec}$
  - $s \in \mathcal{S}_L^{\prec}$ if $\exists \pi$ on $\{s' \prec s : s' \in \mathcal{S}_L^{\prec}\}$ s.t. $\tau(s|\pi) \leq L$
- Define: $\mathcal{S}_L^{\rightarrow} = \cup_{\prec} \mathcal{S}_L^{\prec}$.
- Note: Simpler definitions don't work (counterexamples).
- <u>Prop</u>: $\exists \prec$ s.t. $\mathcal{S}_L^{\rightarrow} = \mathcal{S}_L^{\prec}$ and $\mathcal{S}_L^{\rightarrow}$ is finite.

UCBExplore



1. Discover
2. Propose
3. Verify

Lim and Auer (COLT 2012)

Main result

Theorem 8 *When algorithm* UcbExplore *is run with inputs* s_0, $\mathcal{A}$, $L \geq 1$, $\varepsilon > 0$, *and* $\delta \in (0,1)$, *then with probability* $1-\delta$

- *it terminates after* $O\left(\frac{SAL^3}{\varepsilon^3}\left(\log \frac{SAL}{\varepsilon\delta}\right)^3\right)$ *exploration steps,*
- *discovers a set of states* $\mathcal{K} \supseteq \mathcal{S}_L^{\rightarrow}$,
- *and for each* $s \in \mathcal{K}$ *outputs a policy* π_s *with* $\tau(s|\pi_s) \leq (1+\varepsilon)L$,

where $S = |\mathcal{K}| \leq |\mathcal{S}_{(1+\varepsilon)L}^{\rightarrow}|$.

Anytime, continual learning version:

Corollary 9 *If* UcbExplore *is run with* $L_k = (1+\varepsilon)^k$ *and* $\delta_k = \frac{\delta}{2(k+1)^2}$ *for* $k = 0, 1, 2, \ldots$, *then with probability* $1-\delta$, *for any* $L \geq 1$ *and any* $s \in \mathcal{S}_L^{\rightarrow}$, *the algorithm will discover a policy* π_s *with* $\tau(s|\pi_s) \leq (1+\varepsilon)^2 L$ *after* $O\left(\frac{SAL^3}{\varepsilon^4}\left(\log \frac{SAL}{\varepsilon\delta}\right)^3\right)$ *exploration steps where* $S = |\mathcal{S}_{(1+\varepsilon)^2 L}^{\rightarrow}|$.

Lim and Auer (COLT 2012)

Nonstationarity

w. Auer, Gajane,Ortner (2019)

Performance metric

- F: number of times the transition probabilities change
  - (t=1: always a change)
- $W(L)$ time steps to find all L-reachable states in a single MCP $\Rightarrow F\ W(L)$ time steps when there are F changes
- Classification of time steps: Alg has correct knowledge of what is reachable; or not. Alg is **competent** vs **incompetent**
- Goal: Minimize the # time steps when Alg is incompetent
- Difficulty: The location and number of changes is unknown

Main ideas

- Two phases:
  - Build set of reachable states $\mathcal{K}$ (UCBExplore)
  - Repeat: Check for new reachable states (UCBExplore) or disappearing states (as in verification phase of UCBExplore) – break out when UCBExplore often takes too long compared to predicted runtime
- Checking starts when building is done
- Issues with building:
  - How can the alg know whether a change happened while building? E.g. new state was found reachable. Before change, after change?
- Solution: Staggered start of many parallel building processes. Quit building when any of the processes finishes.

w. Auer, Gajane,Ortner (2019)

Result

- **Theorem**: Up to lower order terms and log factors, the total number of steps when the alg is incompetent is at most $(W(L)F)^2$ irrespective of when the changes happen.

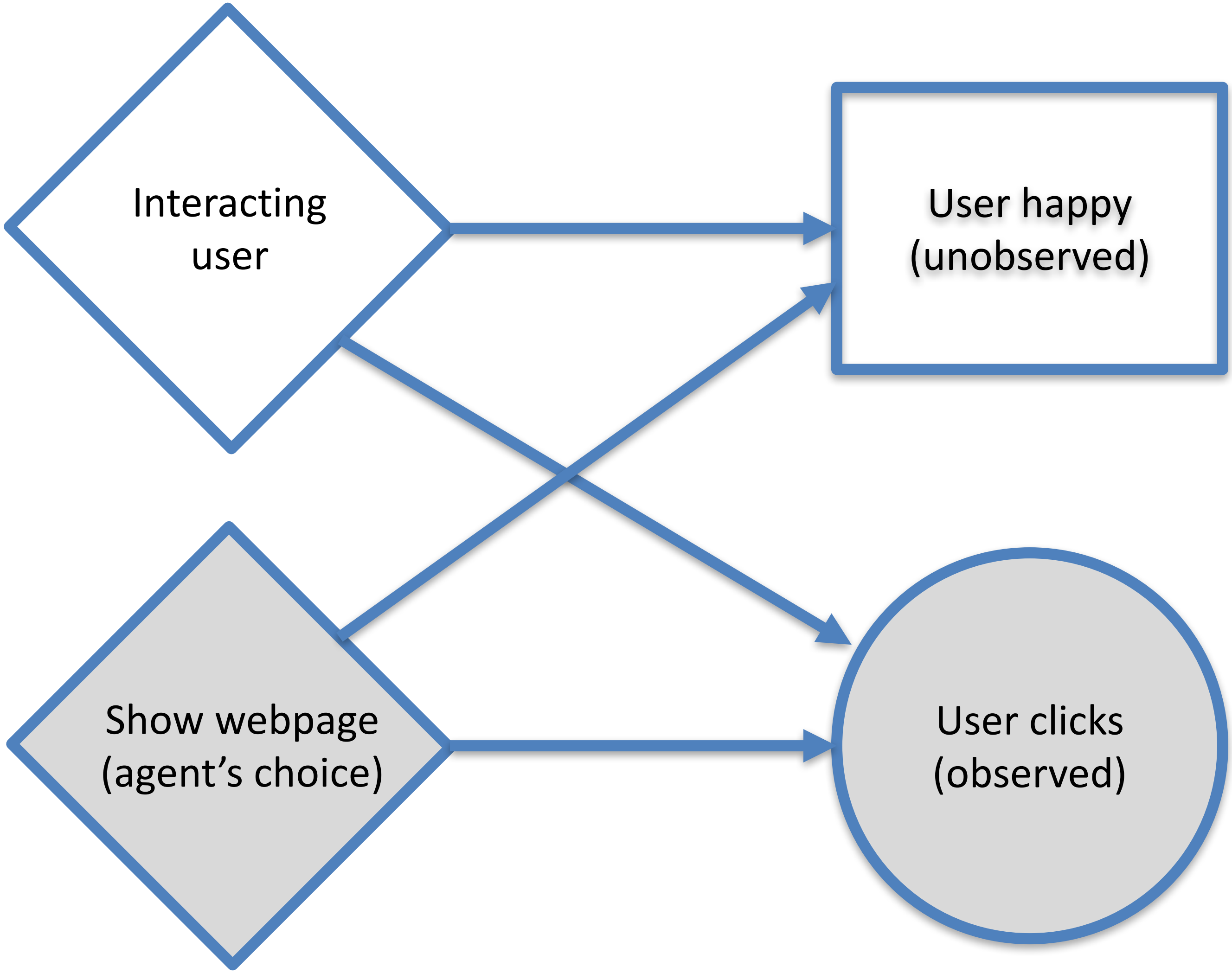
- Questions:
  - Is $W(L)$ cost necessary without changes?
  - Is the quadratic dependence above necessary?
  - Nontabular?

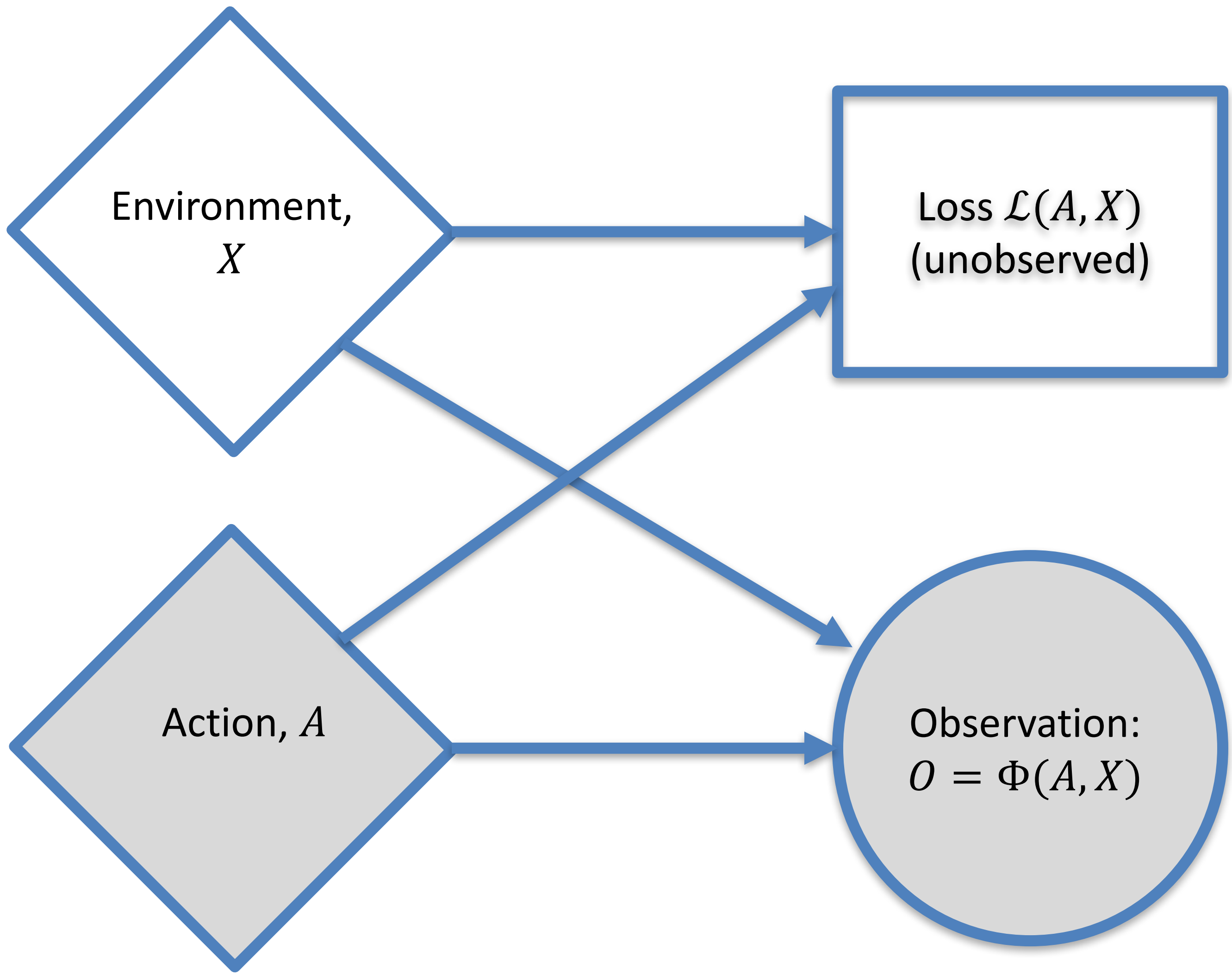
w. Auer, Gajane,Ortner (2019)

Part II: Unobserved rewards

- RL: rewards are always observed
  - internally computed
  - externally provided
- Is this reasonable?
- Is the environment state observable?
- What happens when rewards are not observable?
- Consequences for:
  - Planning
  - Learning ⇒ exploration; which will need planning!
- Bandits: MPDs w. iid state
- Partial monitoring: POMPD^{-r} w. iid state

w. Tor Lattimore (2019)


Interacting user
User happy (unobserved)
Show webpage (agent's choice)
User clicks (observed)


Environment, X
Loss $\mathcal{L}(A, X)$ (unobserved)
Action, A
Observation: $O = \Phi(A, X)$

Partial Monitoring

Learner is given maps $\mathcal{L}, \Phi$

For rounds $t = 1,2, \ldots, n$:

1. Environment chooses $X_t \in \mathcal{X}$
2. Learner chooses $A_t \in \mathcal{A}$
3. Learners suffers loss $\mathcal{L}(A_t, X_t)$ – which remains hidden!
4. Learner observes feedback $\Phi(A_t, X_t)$

Regret: $R_n = \max_a \sum_{t=1}^{n} \mathcal{L}(A_t, X_t) - \mathcal{L}(a, X_t)$

[Rustichini, 1999]

Why great?

- Informal examples of PM problems:
  - Dynamic pricing
  - Altruistic agents
  - Statistical testing (balancing power and cost)
  - Delayed rewards/surrogates
- Subsumes classic frameworks:
  - finite-armed bandits
  - prediction with expert advice
  - bandits with graph feedback
  - linear bandits
  - dueling bandits
  - …

Partial Monitoring – Classification Theorem

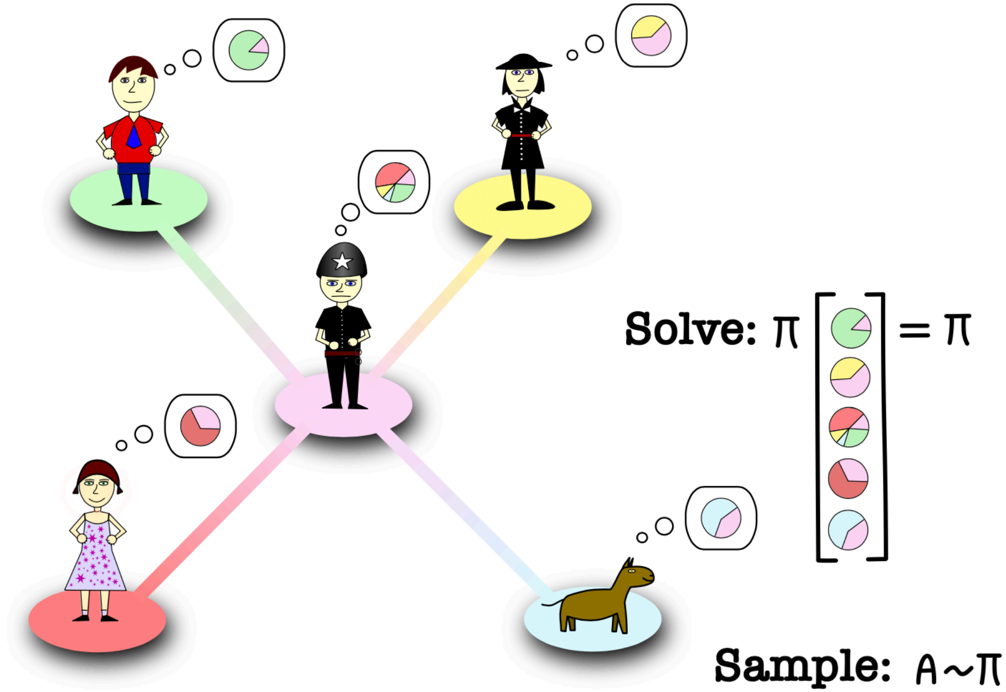
Theorem: Let $\mathcal{A}, \mathcal{X}$ be finite. Let $R_n^*(G)$ be the minimax regret on PM problem $\mathrm{G} = (\mathcal{L}, \Phi)$. Then:

$$R_n^*(G) = \begin{cases} 0 & \text{if } G \text{ has no nb actions} \\ \Theta(\sqrt{n}) & \text{if } G \text{ is L.O. and has nb actions} \\ \Theta(n^{2/3}) & \text{if } G \text{ is G.O. but not L.O.} \\ \Omega(n) & \text{otherwise} \end{cases}$$

[Cesa-Bianchi, Lugosi, Stoltz, 2006; Bartók, Pál, Sz., 2011; Foster and Rakhlin, 2012; Antos, Bartók, Pál and Sz., 2013; Bartók, Foster, Pál, Rakhlin, Sz., 2014; Lattimore and Sz., 2019a].

Algorithms?

- Classical approaches fail in partial monitoring
  - Optimism/Thompson-sampling/exponential weights
- Complicated algorithms exist; none are good!



Exploration by Optimisation

(1) $Q_{ta} = \dfrac{\exp\left(-\eta \sum_{s=1}^{t-1} \hat{\ell}_{sa}\right)}{\sum_{b=1}^{k} \exp\left(-\eta \sum_{s=1}^{t-1} \hat{\ell}_{sb}\right)}$

k actions, learning rate η

$\hat{\ell}_s \in \mathbb{R}^k$ is a loss estimator

$\Psi_q(z) = \langle q, \exp(-z) + z - 1 \rangle$

(2) Find P_t and unbiased $g_t : \mathrm{Actions} \times \mathrm{Obs.} \to \mathbb{R}^k$ minimising

$$\max_{x \in \mathcal{X}} \Bigg[\underbrace{\sum_{a=1}^{k} (P_{ta} - Q_{ta}) \mathcal{L}(a, x)}_{\text{Loss for playing } P_t \text{ not } Q_t} + \underbrace{\frac{1}{\eta} \sum_{a=1}^{k} P_{ta} \Psi_{Q_t}\left(\frac{\eta\, g_t(a, \Phi(a, x))}{P_{ta}}\right)}_{\text{Stability of exponential weights}} \Bigg]$$

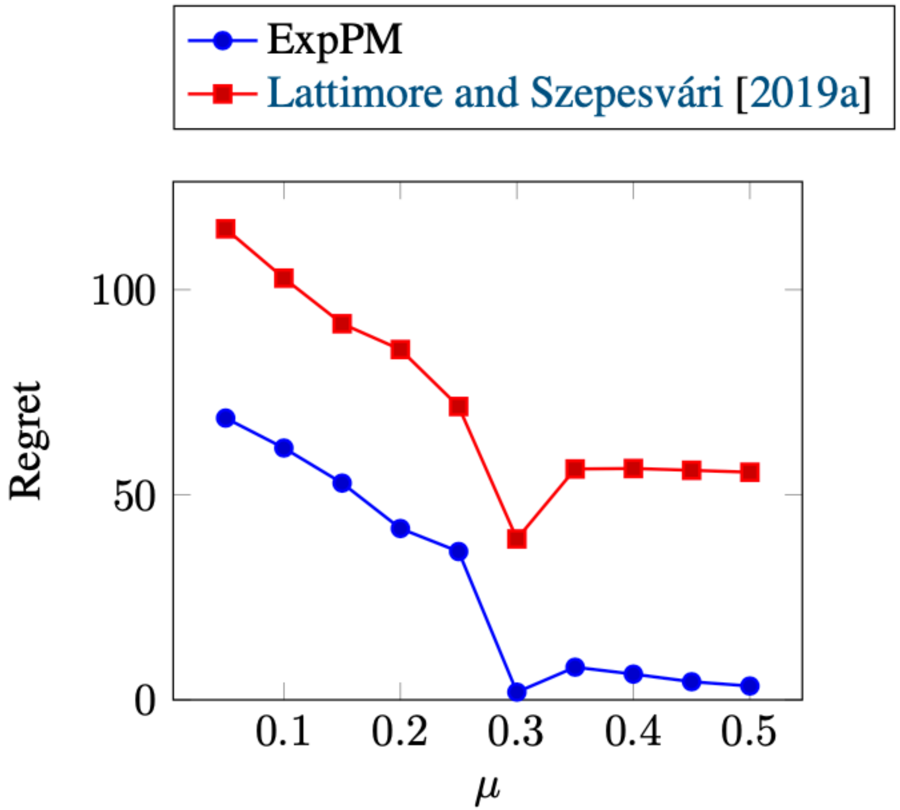
(3) Sample $A_t \sim P_t$ and observe O_t

(4) Set $\hat{\ell}_t = g_t(A_t, O_t)$

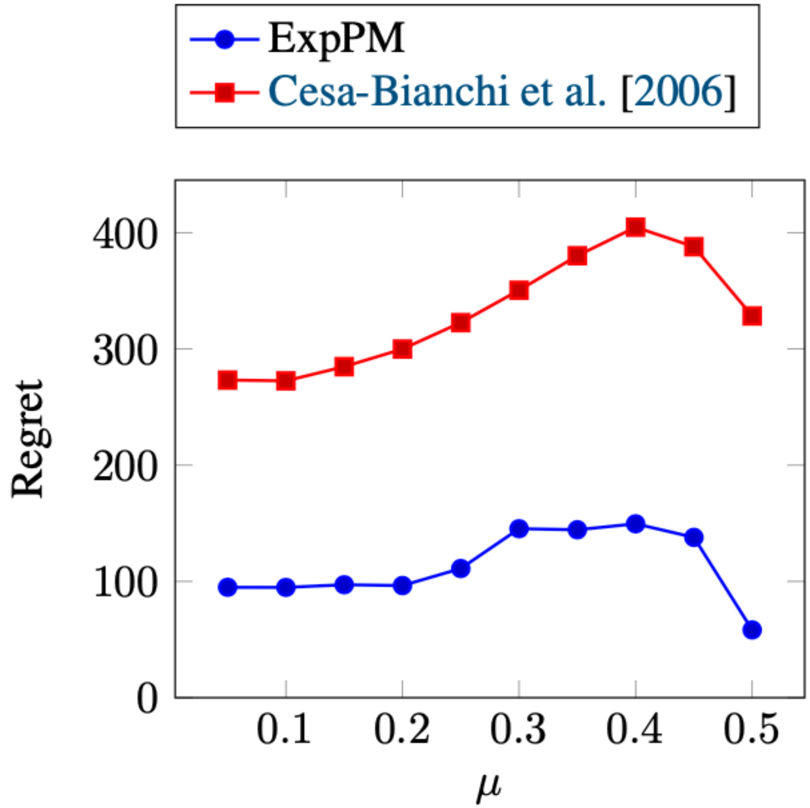
Theory

- **Single** algorithm works in all ‘learnable’ finite games
- Near-optimal for bandits, full information, graph feedback
- Best known bounds in general case
- Essentially no tuning; learning rate tuned online

Experiments



Experiments



Conlusions/Future plans

- It is sometimes good to be ambitious!

- More experiments needed
- How to solve the optimization problem? It is convex! But cost is not $O(k)$ ..
- What happens when $\mathcal{X}$ is large or infinite?
- Generalizations?
  - Add state/context! Use “explore by optimization” beyond PM?
- Find more applications?

Part III: RL & generalization

- The world is big
- Need approximate models
- Minimal assumptions to make RL + Gen work?
- policy error =f(approximation error of “model”)

3 results:

Generative model access/planning by solving a reduced order model
Model-based RL: factored linear models – a convenient model class
Model-free RL

LRA: Linearly Relaxed ALP

$$\min_{r \in \mathbb{R}^k} c^\top \Phi r \text{ s.t.}$$
$$\sum_a W_a^\top \Phi r \geq \sum_a W_a^\top (g_a + \alpha P_a \Phi r)$$

$$c \geq 0, 1^\top c = 1$$
$$W_a \in [0,\infty)^{S\times m}, \psi \in [0,\infty)^S$$
$$\|J\|_{\infty,\psi} = \max_s \frac{|J(s)|}{\psi(s)}$$
$$\beta_\psi := \alpha \max_a \|P_a \psi\|_{\infty,\psi} < 1$$
$$\psi \in \text{span}(\Phi)$$

<u>**Theorem**</u>: Let $\epsilon = \inf_{r \in \mathbb{R}^k} \|J^* - \Phi r\|_{\infty,\psi}, J_{\text{LRA}} = \Phi r_{\text{LRA}}$, where r_{LRA} is the solution to the above LP. Then, under the said assumptions,

$$\|J^* - J_{\text{LRA}}\|_{1,c} \leq \frac{2c^\top \psi}{1-\beta_\psi} (3\epsilon + \|J^*_{\text{ALP}} - J^*_{\text{LRA}}\|_{\infty,\psi})$$

$$J^*_{\text{ALP}}(s) = \min\{ r^\top \phi(s) : \Phi r \geq J^*, r \in \mathbb{R}^k\}$$
$$J^*_{\text{LRA}}(s) = \min\{ r^\top \phi(s) : W^\top E\, \Phi r \geq W^\top E\, J^*, r \in \mathbb{R}^k \}$$

P. J. Schweitzer and A. Seidmann, "Generalized polynomial approximations in Markovian decision processes," *Journal of Mathematical Analysis and Applications*, vol. 110, pp. 568–582, 1985.

D. P. de Farias and B. Van Roy, "The linear programming approach to approximate dynamic programming," *Operations Research*, vol. 51, pp. 850–865, 2003.

——, "On constraint sampling in the linear programming approach to approximate dynamic programming," *Mathematics of Operations Research*, vol. 29, pp. 462–478, 2004.

IEEE TAC, 2018
w. Chandrashekar L. & Sh. Bhatnagar

Model-based RL

Theorem 7 (Baseline bound on MBRL policy error) *Consider some transition probability kernel $\widetilde{\mathcal{P}}$ for the state and action spaces $\mathcal{X}$ and $\mathcal{A}$. Let $\widetilde{V}$ be the fixed point of $MT_{\widetilde{\mathcal{P}}}$, and $\tilde{\pi} = GT_{\widetilde{\mathcal{P}}}\widetilde{V}$. Then*

$$\left\|V^* - V^{\tilde{\pi}}\right\|_\infty \leq \frac{2\gamma}{1-\gamma}\left\|(\mathcal{P} - \widetilde{\mathcal{P}})\widetilde{V}\right\|_\infty .$$

This result is essentially contained in the works of Whitt (1978, Corollary to Theorem 3.1), Singh and Yee (1994, Corollary 2)[2], Bertsekas (2012, Proposition 3.1), and Grünewälder et al. (2011, Lemma 1.1).

Good? Bad?

Bonus question: Can $\|V^* - V^{\tilde{\pi}}\|$ be controlled via controlling $\|\tilde{V} - V^*\|$?

Can we do better? Perhaps using extra structure?

w. Bernardo Ávila Pires COLT 2016

Structure: Factored linear models

$\mathcal{P}(dx'|x,a) \approx \xi(dx')^\top \psi(x,a)$

$\mathcal{P}$: VFUN $\to$ AVFUN

$(\mathcal{P}V)(x,a) = \int V(x')\mathcal{P}(dx'|x,a)$

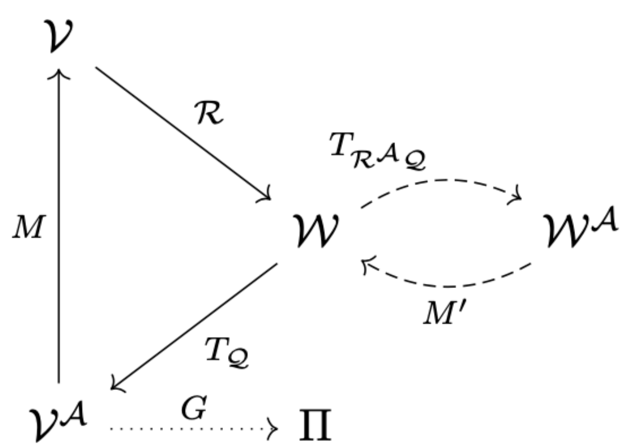
$\mathcal{R}$: VFUN $\to \mathbb{R}^d$

$\mathcal{R}V = \int V(x')\xi(dx')(= w) \in \mathbb{R}^d$

$\mathcal{Q}$: $\mathbb{R}^d \to$ AVFUN

$(\mathcal{Q}w)(x,a) = w^\top \psi(x,a)$

$\mathcal{P} \approx \mathcal{Q}\mathcal{R}$



Legend:
$\mathcal{V}$ = VFUN
$\mathcal{W} = \mathbb{R}^{\mathcal{J}}$ = CVFUN
$\mathcal{V}^{\mathcal{A}}$ = AVFUN

Special cases:
- Tabular
- Linear MDP
- KME
- Stoch. Fact.
- KBRL
- ..

w. Bernardo ÁVILA PIRES COLT 2016/semi/linear/PhD Th.

Policy error in factored linear models

Theorem 8 (Supremum-norm bound) *Let $\hat{\pi}$ be the policy derived from the factored linear model defined using* (1) *and* (2)*. If Assumptions 3 and 5 hold, then*

$$\left\|V^* - V^{\hat{\pi}}\right\|_\infty \leq \varepsilon(V^*) + \varepsilon(V^{\hat{\pi}}), \tag{3}$$

where $\varepsilon(V) = \min(\varepsilon_1(V), \varepsilon_2)$*, and*

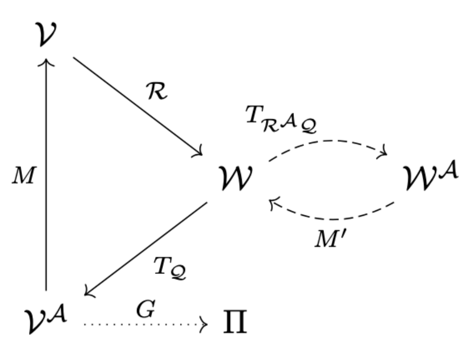
$$\varepsilon_1(V) = \gamma \left\|(\mathcal{P} - \mathcal{Q}\mathcal{R})V\right\|_\infty + \frac{B\gamma^2}{1-\gamma}\left\|\mathcal{R}(\mathcal{P} - \mathcal{Q}\mathcal{R})V\right\|_\infty,$$

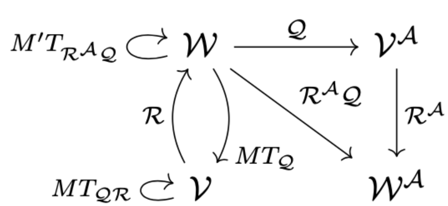
$$\varepsilon_2 = \frac{\gamma}{1-\gamma}\left\|(\mathcal{P} - \mathcal{Q}\mathcal{R})U^*\right\|_\infty.$$

$$U^* = MT_{\mathcal{Q}}u^*$$
$$u^* = M'T_{\mathcal{R}^{\mathcal{A}}\mathcal{Q}}u^*$$
$$\hat{\pi} = GT_{\mathcal{Q}}u^*$$

Assumption 3 *The following hold for $\mathcal{Q}$ and $\mathcal{R}^{\mathcal{A}}$:* $\|\mathcal{R}^{\mathcal{A}}\mathcal{Q}\| \leq 1$.

Assumption 5 *We have that* $B \doteq \|\mathcal{Q}\| < \infty$.





Questions:

- Is the bound tight?
- Time/action abstraction?
- Efficient learning? What specific models to use?

w. Bernardo ÁVILA PIRES COLT 2016/semi/linear/PhD Th.

Online, model-free RL w. neural nets

- Continuing RL; $\bar{R}_T$: pseudo regret; let $Q_t \coloneqq Q_{\pi^{(t)}}$.
- Key identity:

$$\bar{R}_T = \sum_x \nu_{\pi^*}(x) \sum_{t=1}^{T} \langle Q_t(x,\cdot), \pi^{(t)}(\cdot\,|x)\rangle - \langle Q_t(x,\cdot), \pi^*(\cdot\,|x)\rangle$$

- Then..

$$\langle Q_t(x,\cdot), \pi^{(t)}(\cdot\,|x)\rangle - \langle Q_t(x,\cdot), \pi^*(\cdot\,|x)\rangle =$$

$$\langle \hat{Q}_t(x,\cdot), \pi^{(t)}(\cdot\,|x)\rangle - \langle \hat{Q}_t(x,\cdot), \pi^*(\cdot\,|x)\rangle \quad \Rightarrow \text{Control w. OLP}$$

$$+\langle Q_t(x,\cdot), \pi^{(t)}(\cdot\,|x)\rangle - \langle \hat{Q}_t(x,\cdot), \pi^{(t)}(\cdot\,|x)\rangle \quad \Rightarrow \text{A: } L^1(\nu_{\pi^*} \otimes \pi^{(t)})$$

$$+\langle \hat{Q}_t(x,\cdot), \pi^*(\cdot\,|x)\rangle - \langle Q_t(x,\cdot), \pi^*(\cdot\,|x)\rangle \quad \Rightarrow \text{A: } L^1(\nu_{\pi^*} \otimes \pi^*)$$

w. N. Lazic, Y. Abbasi-Yadkori, G. Weisz, ICML'19 + arXiv

Politex

Input: phase length $\tau > 0$, initial state x_0
Set $\widehat{Q}_0(x,a) = 0 \ \forall x, a$
for $i := 1, 2, \ldots,$ **do**

Policy iteration: $\pi_i(\cdot|x) = \underset{u \in \Delta}{\operatorname{argmin}} \langle u, \widehat{Q}_{i-1}(x,\cdot)\rangle$

$$\textsc{Politex}: \quad \pi_i(\cdot|x) = \underset{u \in \Delta}{\operatorname{argmin}} \langle u, \sum_{j=0}^{i-1} \widehat{Q}_j(x,\cdot)\rangle - \eta^{-1}\mathcal{H}(u)$$

$$\propto \exp\left(-\eta \sum_{j=0}^{i-1} \widehat{Q}_j(x,\cdot)\right)$$

Execute π_i for τ time steps and collect dataset $\mathcal{Z}_i$
Estimate $\widehat{Q}_i$ from $\mathcal{Z}_1, \ldots, \mathcal{Z}_i, \pi_1, \ldots, \pi_i$
end for

w. N. Lazic, Y. Abbasi-Yadkori, G. Weisz, ICML'19 + arXiv

Regret bounds

Theorem

Assume that for any policy π, after following π for n steps, a black-box function approximator produces an action-value function whose error is $\epsilon_0 + 1/\sqrt{n}$ up to some universal constant.

Then the average pseudo-regret of Politex after T steps is $\epsilon_0 + T^{-\frac{3}{4}}$.

w. N. Lazic, Y. Abbasi-Yadkori, G. Weisz, ICML'19 + arXiv

Refinements

- Problem: How to get the $\epsilon_0 + \frac{1}{\sqrt{n}}$ error?
  - E.g. linear VFA? LSPE! ϵ_0: limiting error of LSPE could be $\gg$ best error.
- Refinement 1:
  - Use on-policy state value function-approximator
  - add extra action-dithering per state
  - assume all policies excite state-features
- Refinement 2:
  - Assume access to an "exploration policy" that excites features
  - Interleave exploration steps with policy steps
  - Use off-policy(!) VFA (which one?)
  - $\Rightarrow$ Regret degrades a bit
- Questions:
  - Can we do better with other OL methods? Is averaging really necessary?
  - Better value-function learners?

w. N. Lazic, Y. Abbasi-Yadkori, G. Weisz, ICML'19 + arXiv

Summary

Part I: Curiosity

curiosity | kjʊərɪˈɒsɪti |

noun (plural **curiosities**)

1 *[mass noun]* a strong desire to know or learn something: *filled with curiosity, she peered through the window* | ***curiosity got the better of*** *me, so I called him*.

"One of the striking differences between current reinforcement learning algorithms and early human learning is that animals and infants appear to explore their environments with autonomous purpose, in a manner appropriate to their current level of skills."

Models for Autonomously Motivated Exploration in Reinforcement Learning*

Peter Auer[1], Shiau Hong Lim[1], and Chris Watkins[2]

ALT 2011, invited talk by Peter

Add robot vs. dog/child exploring its environment

Exploration by Optimisation

k actions, learning rate η

$\hat{\ell}_s \in \mathbb{R}^k$ is a loss estimator

$\Psi_q(z) = \langle q, \exp(-z) + z - 1 \rangle$

(1) $Q_{ta} = \dfrac{\exp\left(-\eta \sum_{s=1}^{t-1} \hat{\ell}_{sa}\right)}{\sum_{b=1}^{k} \exp\left(-\eta \sum_{s=1}^{t-1} \hat{\ell}_{sb}\right)}$

(2) Find P_t and unbiased $g_t : \text{Actions} \times \text{Obs.} \to \mathbb{R}^k$ minimising

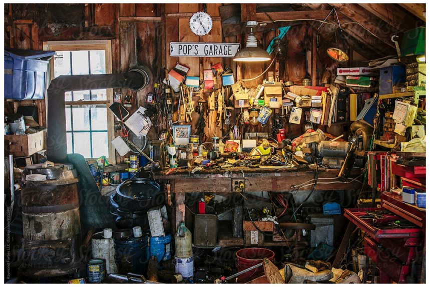
$$\max_{x \in \mathcal{X}} \left[\underbrace{\sum_{a=1}^{k} (P_{ta} - Q_{ta}) \mathcal{L}(a, x)}_{\text{Loss for playing } P_t \text{ not } Q_t} + \underbrace{\frac{1}{\eta} \sum_{a=1}^{k} P_{ta} \Psi_{Q_t} \left(\frac{\eta\, g_t(a, \Phi(a, x))}{P_{ta}} \right)}_{\text{Stability of exponential weights}} \right]$$

(3) Sample $A_t \sim P_t$ and observe O_t

(4) Set $\hat{\ell}_t = g_t(A_t, O_t)$

Part III: RL & generalization

- The world is big
- Need approximate models
- Minimal assumptions to make RL + Gen work?
- policy error =f(approximation error of "model")



3 results:

Generative model access/planning by solving a reduced order model
Model-based RL: factored linear models – a convenient model class
Model-free RL