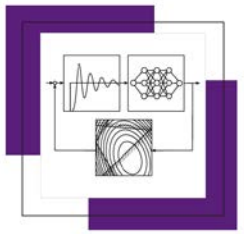
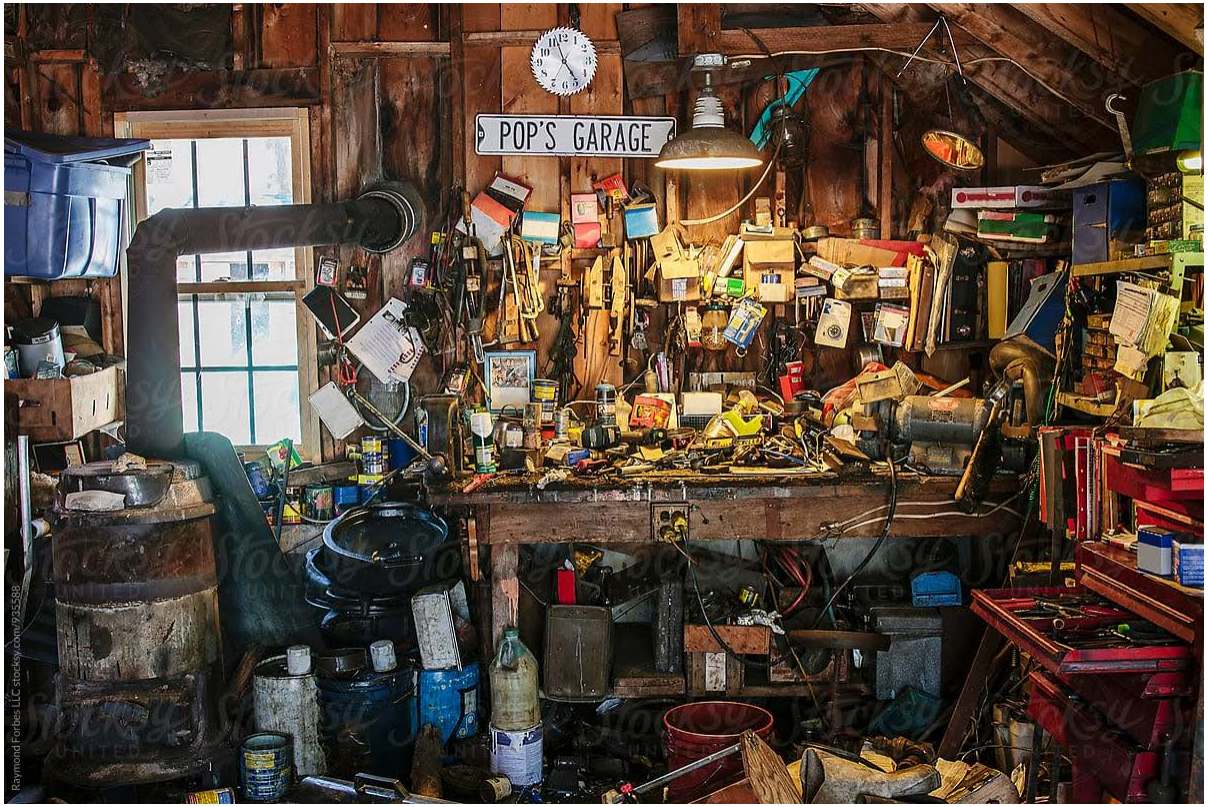
Model misspecification in reinforcement learning

Csaba Szepesvári

DeepMind & University of Alberta

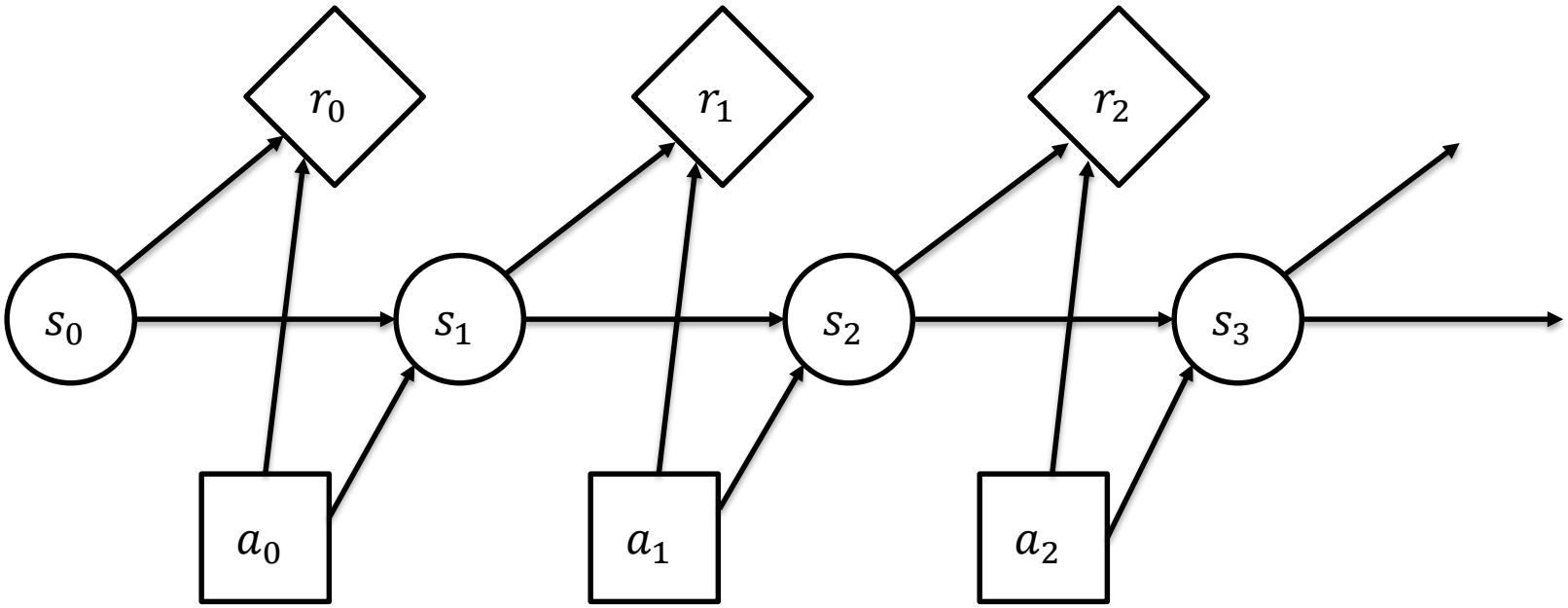
IPAM Workshop on
“Intersections between Control, Learning and Optimization”
LCO 2020





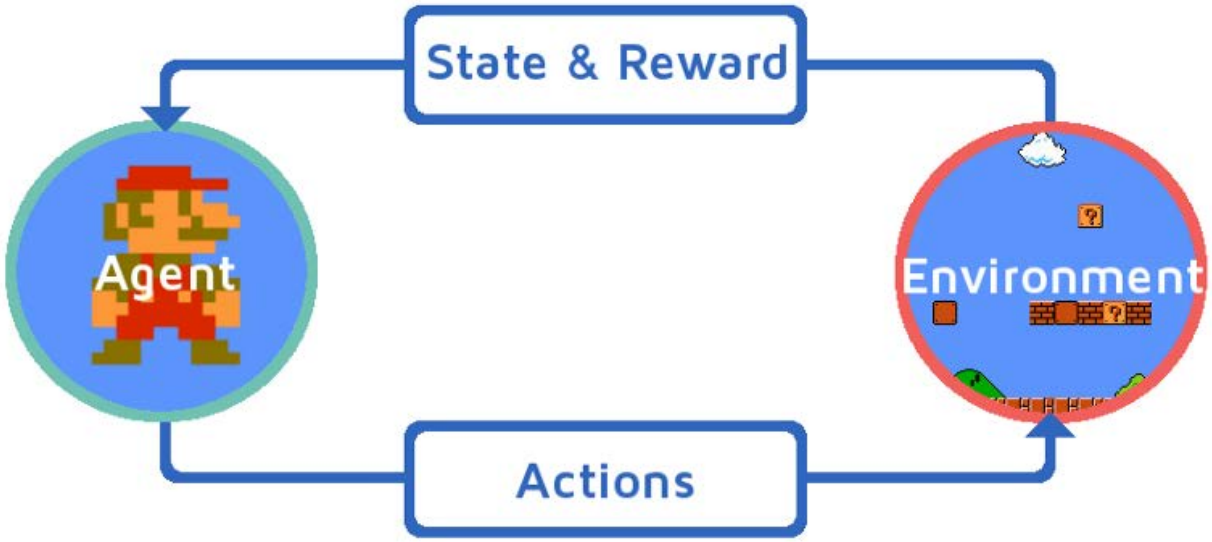
- The world is big
- Need approximate models (Q,V, π, P) ➔ model misspecification
- What is the price? How to keep the price low?

- Markov Decision Processes
- $M = (\mathcal{S}, \mathcal{A}, P = (P_a)_{a \in \mathcal{A}}, r = (r_a)_{a \in \mathcal{A}})$



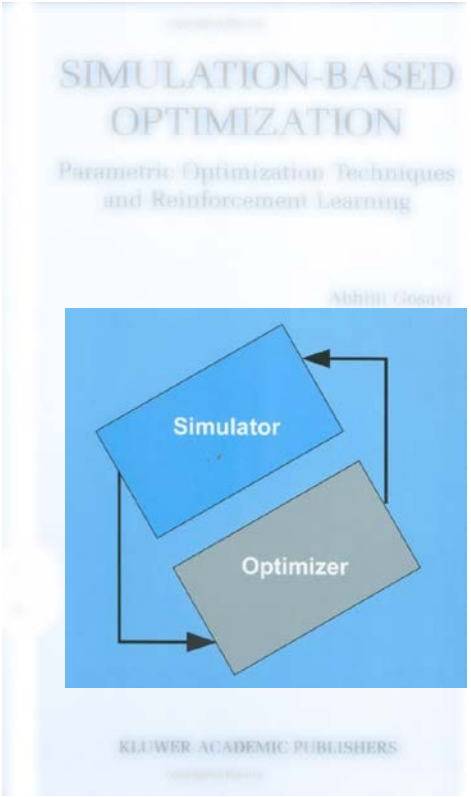
Online RL

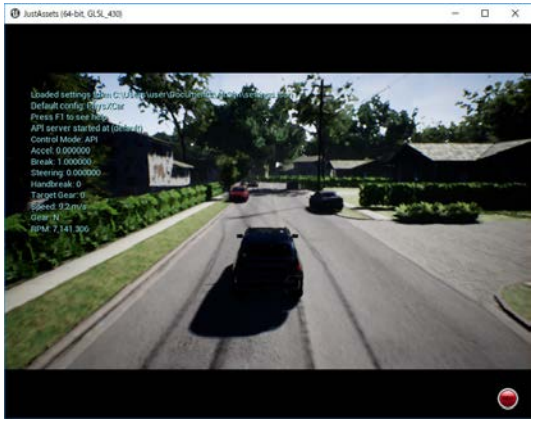
- Given: sequential access to M
- Goal: Take actions to
  maximize expected return (=total reward)

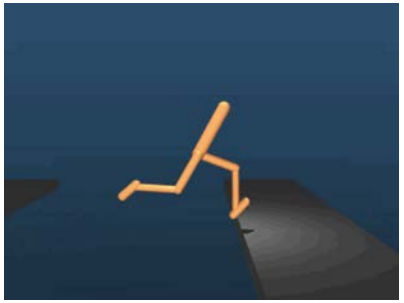


Planning with a simulator (+ reset)

- Given: stochastic simulator of M with reset
- Goal: find a policy with high expected return with a few queries

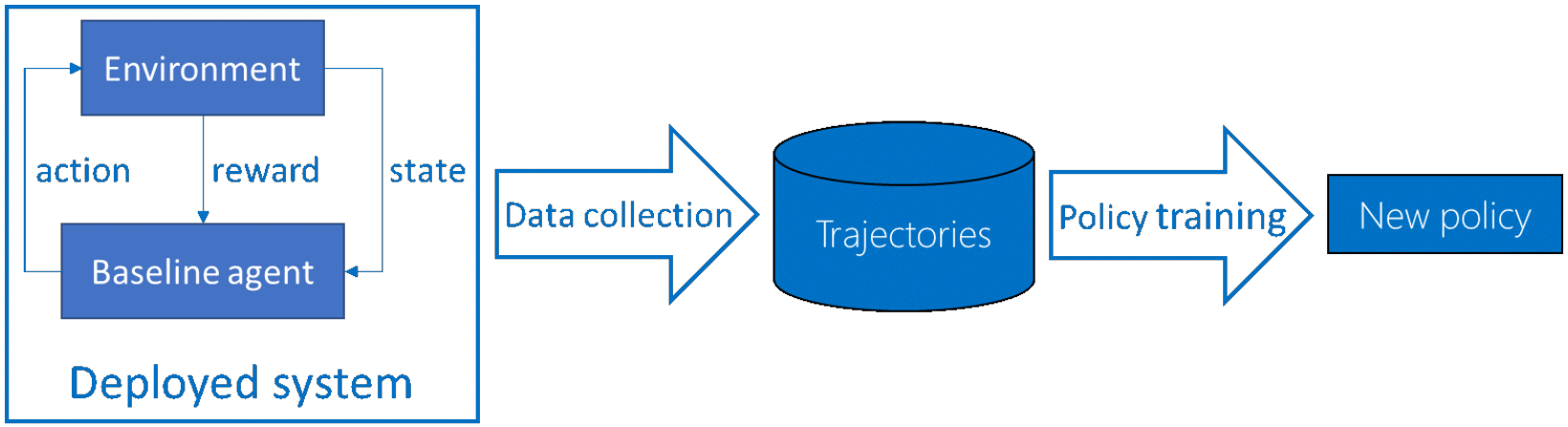




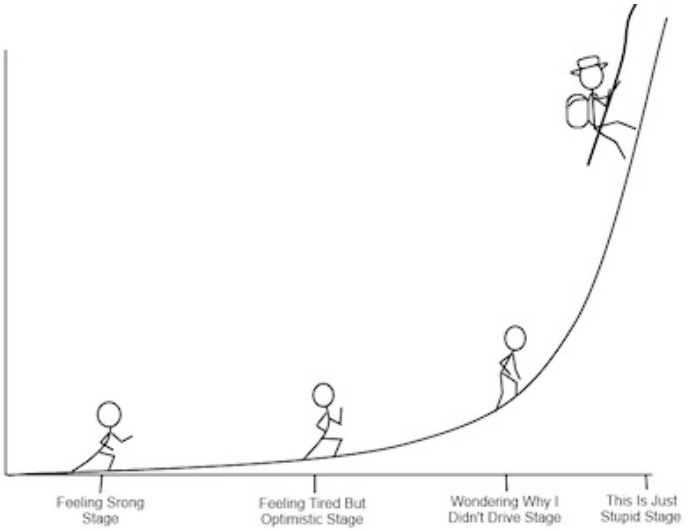


Batch RL

- Given: data from past interaction with M
- Goal: find a policy with high expected return



- Challenges
  - $\mathcal{S}$ is huge
  - $\mathcal{A}$ is huge



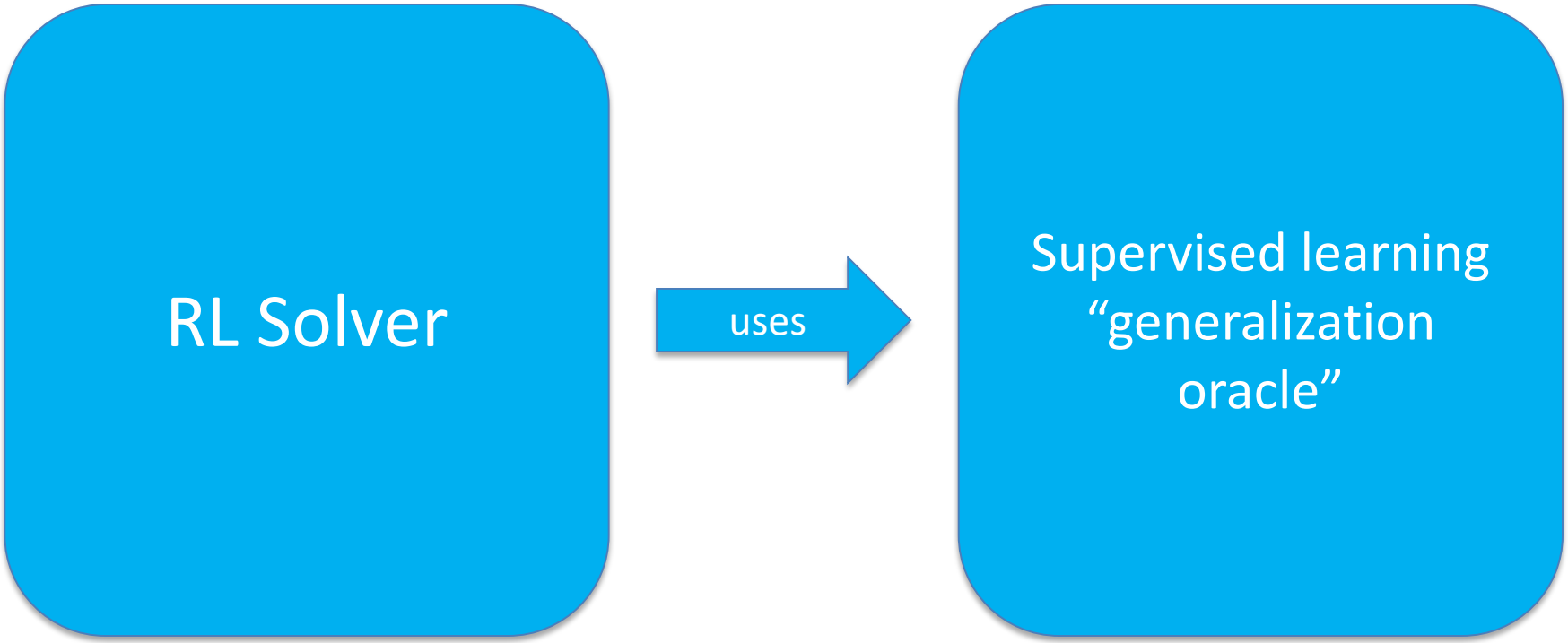
- **Theorem**: Computing π^* is P-complete

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 12, No 3, August 1987
Printed in U.S.A.

THE COMPLEXITY OF MARKOV DECISION PROCESSES* †

CHRISTOS H. PAPADIMITRIOU‡ AND JOHN N. TSITSIKLIS§

We investigate the complexity of the classical problem of optimal policy computation in Markov decision processes. All three variants of the problem (finite horizon, infinite horizon discounted, and infinite horizon average cost) were known to be solvable in polynomial time by dynamic programming (finite horizon problems), linear programming, or successive ap-



Can we design

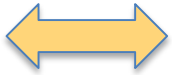
- efficient methods (no scaling with S, A)
- RL-error $\leq$ C * SL-error

???

How???

SL Oracle

Oracle-using solver



Environment or Simulator

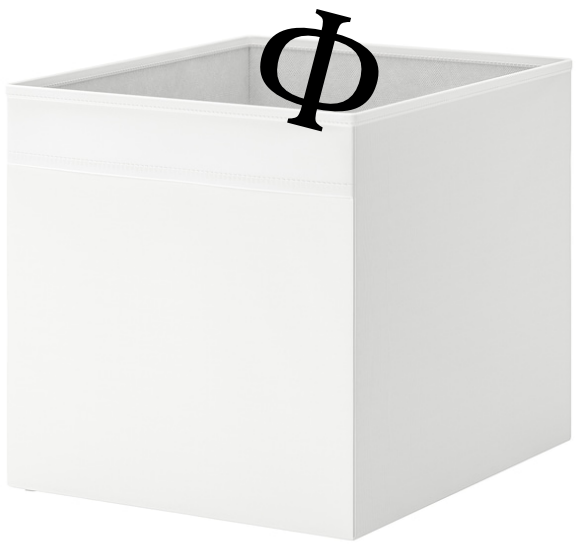
Modeling: Q, V, P, π, …

Black box, supervised learning oracles

- Internals unknown/hidden/..
- Input: Data
- Output: Function
  - Linear function approximation
  - Neural nets
  - Nonparametric techniques
  - ....

White box oracles

- Known internals
- Linear function approximation
  - Action-value approximation: $Q^{\pi} \approx \Phi\theta^{\pi}$
  - Access to $\Phi \in \mathbb{R}^{SA \times d}$



Approximate Dynamic Programming

1960-1980s


DYNAMIC PROGRAMMING
AND
ITS APPLICATIONS

1978

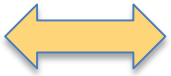
edited by

Martin L. Puterman

FACULTY OF COMMERCE AND BUSINESS ADMINISTRATION
THE UNIVERSITY OF BRITISH COLUMBIA
VANCOUVER, B.C., CANADA

Idealized
SL Oracle

Iterative DP
solver
(VI,PI)



Environment
or
Simulator

- Gains from compressed representation
- No learning, no randomization
- Unconventional/unrealistic oracles

Methods: Value-, policy-iteration, LP

$$V_{k+1} = \hat{T}_k V_k = TV_k + \epsilon_k$$

Oracle: Linear function approximation,
model compression, ..

$\pi_k \in \Gamma\, \mathrm{T}V_k$ (greedy one-step lookahead)

$V^{\pi_k} \geq V^* - \delta 1$ with some $\delta > 0$?

Discounting: $0 \leq \gamma < 1$.

Misspecification error: $\|\epsilon_k\|_\infty \leq \epsilon \; \forall k > 0$

Theorem: For k "large" enough,

$$\delta \leq \frac{C\,\epsilon}{(1-\gamma)^2}.$$

$\epsilon \ll \frac{1}{H} \coloneqq 1 - \gamma$ gives nontrivial results.

Unimprovable!

How do we control ϵ_k?

Machine learning approach (1990s-today):

Learn $T\ V_k$!

Sampling => random training set

Set-up and solve regression

Problem: $\|\epsilon_k\|_\infty$ vs. $\|\epsilon_k\|_{L^2(\mu)}$

Theorem (Sz., Munos, 2005):

Distribution mismatch price

$$\|V^* - V^{\pi_K}\|_{p,\rho} \le \frac{2\gamma}{(1-\gamma)^2}\left\{C(\rho,\mu)^{1/p}\,\epsilon_1 + \epsilon_2\right\}$$

Approximation error

Estimation error

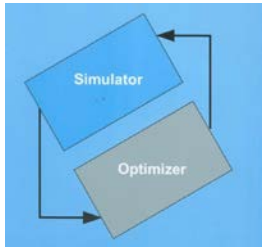
$$\epsilon_1 = d_{L^p(\mu)}(T\mathcal{F},\mathcal{F}) + \text{poly}\left(\frac{\log(N)}{N}, \frac{\log(N|\mathcal{A}|)}{M}, \log(K), \dim(\mathcal{F})\right)$$

Iteration price

$$\epsilon_2 = \text{const} \times \gamma^K$$

Number of queries: $NMK|\mathcal{A}|$

no dependence on $|\mathcal{S}|$!



Can we control $C(\rho, \mu)$?

White box:

Use linear function approximation!

Can achieve δ error with $\mathrm{poly}(\frac{1}{\epsilon}, |\mathcal{A}|, d)$ queries when $d_\infty\left(Q^\Pi, \Phi\right) \leq \epsilon$?

Du et al. 2019: Nope when $\delta \leq \epsilon$.

Weisz, Lattimore, Sz.: Yes, when $\delta \geq C\sqrt{d}\epsilon$.

$\delta \leq \epsilon$: Exp complexity

$\delta \geq \sqrt{d}\epsilon$: Poly complexity

Why?

Insight from bandits! $\mathcal{S} = \{1\}, \mathcal{A}$ is large!

$$r = \Phi\theta_* + \text{e}$$

How many queries of $r(a)$ are needed to find $\text{argmax}_a r(a)$?

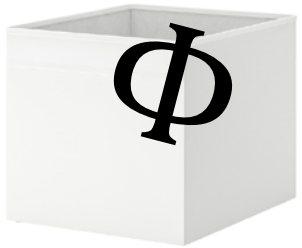
$$r = \Phi\theta_* + \text{e}$$

Φ known, others unknown. Find $\text{argmax}_a r(a)$.

Return $\text{argmax}_a\, \phi(a)^\top\, \hat{\theta}$

- $\hat{\theta} = G^{-1}(\rho) \sum_a \rho(a) r(a) \phi(a)$: least squares estimate of θ_*
- ρ: distribution over $\mathcal{A}$, support $q \ll |\mathcal{A}|$

$$G(\rho) = \sum_a \rho(a) \phi(a)\, \phi_a^\top$$



Theorem (WLSz):

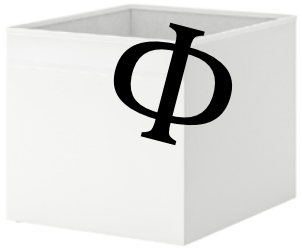
Linearly many queries are enough to get $\delta = \left(2\sqrt{d} + 1\right)\epsilon$.

For any $\Phi \in \mathbb{R}^{k \times d}$ there exists ρ

with $q = |\mathrm{supp}(\rho)| = \tilde{O}(d)$

such that for any $r \in \mathbb{R}^k$:

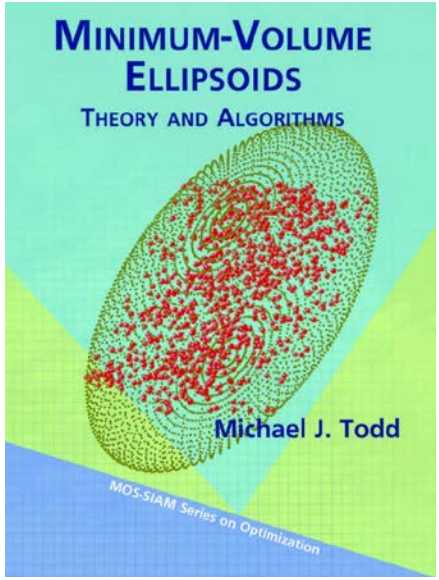
$$\left\| r - \Phi\, \hat{\theta} \right\|_\infty \leq (2\sqrt{d} + 1) \inf_\theta \|r - \Phi\theta\|_\infty.$$



Kiefer-Wolfowitz Theorem (1960):

$$g(\rho) = \max_a \|\phi(a)\|^2_{G^{-1}(\rho)}$$

$$G(\rho) = \sum_a \rho(a)\phi(a)\phi^\top(a)$$



The following are equivalent:

1. ρ^* is a minimizer of g
2. ρ^* is a maximizer of $\log\det G(\rho)$
3. $g(\rho^*) = d$

Todd: $\exists\rho$ s.t. $|\mathrm{supp}(\rho)| = \tilde{O}(d)$ and $g(\rho) \le 2d$

Good:

- Query complexity is small

Bad:

- Errors blow up by a factor of $\sqrt{d}$
- Overall computation scales with k

Can we do it with noise? Yes

Can we do it online? Bandits.. Yes.

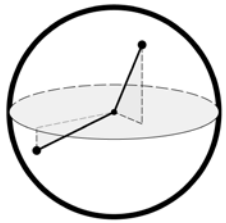
Can we avoid the blow-up? Not in worst-case (Du et al. 2019)

Let $\mathcal{H} = \{\Phi\theta + \text{e}: \|\text{e}\|_\infty \leq \epsilon\}$

(U) If $e_1, \ldots, e_k \in \mathcal{H}$, then at least k queries will be needed to get $\delta \leq 1$.

If $\phi(a)$ are unit length, $|\langle \phi(a), \phi(b) \rangle| \leq \epsilon$ for $a \neq b$ then (U) holds:

with $\theta = \phi(a)$, $\|\Phi\theta - e_a\|_\infty \leq \epsilon$.

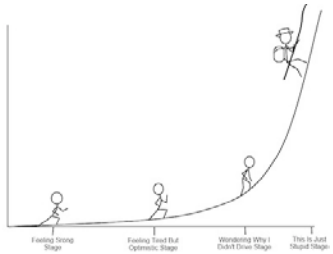


JL Lemma: Such Φ exists if $d \geq \frac{\log(k)}{\epsilon^2}$.

Corollary (WLSz):

For any $\delta > \epsilon$ there exists $\Phi \in \mathbb{R}^{k \times d}$ with k large enough s.t. any method that finds a δ-optimal action for any $r \in \mathcal{H}$ needs at least

$$\frac{1}{2}\exp\left(\frac{d-1}{8}\left(\frac{\epsilon}{\delta}\right)^2\right)$$



queries.

Exponential in d for $\delta = O(\epsilon)$ and

vacuous for $\delta = \Omega\left(\sqrt{d}\epsilon\right)$.

Bandit Algorithm

$$r = \Phi\theta_* + \mathbb{e}$$

with $\|\mathbb{e}\|_\infty = \inf_\theta \|r - \Phi\theta\|_\infty$.

Set $m = \tilde{O}(d)$.

1. Find ρ over $\mathcal{A}$ s.t.
   $|\mathrm{supp}(\rho)| = \tilde{O}(d)$ and $g(\rho) \le 2d$.
2. Play each $a \in \mathcal{A}$, $\lceil m\rho(a) \rceil$ times,
3. Estimate $\hat{\theta}$, let $r^* = \max_{a\in\mathcal{A}} \phi^\top(a)\hat{\theta}$.
4. Let $\mathcal{A} = \{a \in \mathcal{A} : \phi^\top(a)\hat{\theta} > r^* - C(d,m)\}$
5. Double m, goto 1

$$C(d,m) = 2\sqrt{\frac{4d}{m}\log(kn)}$$

Theorem (WLSz): With some $C > 0$ universal, the expected regret, R_n, satisfies

$$R_n \leq C \left(\sqrt{dn \log(nk)} + \epsilon n \sqrt{d} \log(n) \right)$$

Comments on the Du-Kakade-Wang-Yang Lower Bounds

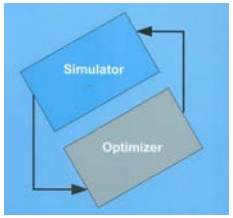
Benjamin Van Roy, Shi Dong

- First, unrealizable stochastic bandit result
- ϵ was not needed by the algorithm
- If it is known, $\log(n)$ can be removed from the second term
- For $k \approx n$, the bound is tight

- Back to RL, planning with a simulator
- $\epsilon := \sup_{\pi} \inf_{\theta} \|Q^{\pi} - \Phi\theta\|_{\infty}$

Also:
Yang and Wang'19

- Can we find a $\delta = C\epsilon\sqrt{d}/(1-\gamma)^2$-optimal policy with $\mathrm{poly}(\frac{1}{\epsilon}, \frac{1}{1-\gamma}, d)$ queries?
- Answer (LWSz): Yes.
  - API, with least-squares regression, to predict action-values using rollouts from a "core set" obtained by optimizing $g(\rho)$.



Details

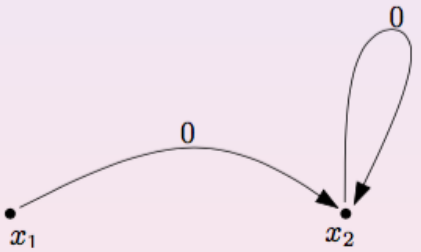
- #iterations: $k = \frac{\log(\dots)}{1-\gamma}$
- #rollouts: $m = \frac{\log(\dots)}{2\epsilon^2(1-\gamma)^2}$
- #steps in rollout: $n = \frac{\log(\dots)}{1-\gamma}$

- Queries: $kmn|\mathcal{C}| = \tilde{O}\left(\frac{d}{\epsilon^2(1-\gamma)^4}\right)$

- "API, with least-squares regression, to predict action-values using rollouts from a "core set" obtained by optimizing $g(\rho)$."
- Why this way?
  - Rollouts, core set: Obvious choice given work on bandits
  - For AVI we would need to deal with the max in setting up the regression problem (YW'19)
  - No max ➔ better query complexity

- Question: What if only $Q^* \approx \Phi\theta$?
- AVI/API is doomed!

- Tsitsiklis & Van Roy (1996)
- State space: $\mathcal{X} = \{x_1, x_2\}$
- Dynamics:



- Bellman operator:

$$(TV)(x_1) = 0 + \gamma V(x_2)$$
$$(TV)(x_2) = 0 + \gamma V(x_2).$$

- Function-space:
  $\mathcal{F} = \{\theta\phi \mid \theta \in \mathbb{R}\}$,

$$\phi(x_1) = 1, \ \phi(x_2) = 2.$$

Iteration:

$$\theta_{t+1} = \operatorname{argmin}_\theta \|\theta\phi - T(\theta_t\phi)\|_2$$
$$= \operatorname{argmin}_\theta (\theta - \gamma 2\theta_t)^2 + (2\theta - \gamma 2\theta_t)^2 = (6/5\gamma)\theta_t \to +\infty$$

μ is the uniform distribution

Alternative approach: Approximate Linear Programming

IEEE TRANSACTIONS ON AUTOMATIC CONTROL, VOL. 63, NO. 4, APRIL 2018

A Linearly Relaxed Approximate Linear Program for Markov Decision Processes

Chandrashekar Lakshminarayanan, Shalabh Bhatnagar, and Csaba Szepesvári

Method

1. Reduce # variables using $V = \Phi\theta$.
2. Reduce # constraints by keeping constraints at a core set $\mathcal{C} \subset \mathcal{S}$
3. Choose $\mathcal{C}$ such that $\forall s \in \mathcal{S}$, $\phi(s) = \sum_{s' \in \mathcal{C}} \lambda(s')\phi(s')$ with $\lambda \geq 0$.

Theorem (LBSz'18):

Let $\hat{\theta}$ be the output. Then:

$$\left\|V^* - \Phi\hat{\theta}\right\|_{1,c} \leq C \frac{\inf_{\theta} \|\mathrm{V}^* - \Phi\theta\|_\infty}{1-\gamma}$$

with a universal constant $C > 0$.

Theorem (w. Roshan Shariff, in prep.):

Query complexity to get an

$\delta = C\epsilon\sqrt{d}/(1-\gamma)^2$-optimal action

at some state is

$$\mathrm{poly}(\tfrac{1}{\epsilon}, \tfrac{1}{1-\gamma}, d, |\mathcal{C}|).$$

Beyond worst-case?

Yes: $\sqrt{d}\epsilon$ can be reduced if Φ is "nice"

What makes a feature map nice?

Blow-up factor:

Allow algorithms q queries.

How much does the approximation error blow up? $\lambda_q(\Phi)$

Proposition:

Zanette et al'19

$$\lambda_q(\Phi) = \min_{\substack{C \subset [k] \\ |C|=q}} \max_i \|\phi(i)\|_{\Phi_C}$$

$$\|x\|_A := \inf\{\textstyle\sum_j |u_j| : x = \sum_j u_j a_j , u_i \in \mathbb{R}\}$$

"gauge fn. of x w.r.t $\mathrm{co}\{a_1, \dots, a_q, -a_1, \dots, -a_q\}$"

Note: $\|x\|_A \leq \min_\rho \|x\|_{G_A^{-1}(\rho)}$.

Refines Zanette et al. 19

Other bandit results

- Infinite action sets:
  - Covering argument, replace $\log(k)$ with d.
- Ghosh et al'17: Cheap linearity test:

$$R_n \leq \min\left(d, \sqrt{k}\right)\sqrt{n}$$

- Contextual case; LinUCB?
  - Gopalan et al'16 $\|\mathbb{e}\|_2 \approx \epsilon$.
  - Needs modification! Needs knowledge of ϵ. Refinement of Jin et al.'19

Approximately linear MDPs ($Q^\pi \approx \Phi\theta^\pi$)

- Avila-Pires, Sz'16:
  - $\mathcal{P} \approx \mathcal{R}Q$; solve compressed model $\Rightarrow U^*$
  - errors at V^*, $V^{\widehat{\pi}}$, U^* matter only!
- Yang and Wang'19a: discounting, "feature regularity" (anchors), incomparable result
- Yang and Wang'19b: online RL, finite horizon
- Jin et al.'19: online RL, finite horizon (LSVI)

With $Q^* \approx \Phi\theta$

- Zanette et al.'19
  Finite-horizon backward computation; multiplicative error propagation!

- Du et al.'20: Deterministic MDPs.
  Find optimal policy when approximation error is very small

Summary

- Models arise because of the need for “compression”
  - Not optional
  - Long history
- Good model ➔ good performance? How good?
- Many modeling assumptions; comparable vs. incomparable
- Even under strong assumptions graceful degradation is not trivial to obtain
  - To control query complexity, we had to open the “black box”!
- Linear function approximation
  - Core sets/good extrapolation are key: How to get them? Implications for feature learning?
  - Not all feature maps are born equal; $\sqrt{d}$ is upper bound
- Do we need to know the level of misspecification? Oh no!
- Beyond linear function approximation?
- Modeling policies?

Details

- #iterations: $k = \frac{\log\left(\frac{1}{\epsilon\sqrt{d}}\right)}{1-\gamma}$
- #rollouts: $m = \frac{\log\left(\frac{2k|\mathcal{C}|}{\alpha}\right)}{2\epsilon^2(1-\gamma)^2}$
- #steps in rollout: $n = \frac{\log\left(\frac{1}{\epsilon(1-\gamma)}\right)}{1-\gamma}$
- Queries: $kmn|\mathcal{C}| = \tilde{O}\left(\frac{d}{\epsilon^2(1-\gamma)^4}\right)$

Wang et al 20:

FAPP closed under optimistic improvement.

No misspecification

- Ancient history
- Fox, B. L. (1973). Discretizing Dynamic Programs. J. Optimization Theory Appl.
- Hinderer, K. (1978). On Approximate Solutions of Finite-Stage Dynamic Programs. Proceedings of the International Conference on Dynamic Programming
- Morin, T. L. (1978). Computational Advances and Reduction of Dimensionality in Dynamic Programming: A Survey. Proceedings of the International Conference on Dynamic P
- Ward Whitt. Approximations of dynamic programs, I. Mathematics of Operations Research, 3(3):231–243, 1978.

- Schweitzer & Seidmann, JMAA, 1985