Simple Reinforcement Learning Algorithms for Continuous State and Action Space Systems

Rahul Jain

Stochastic Systems & Learning Laboratory

Electrical Engineering and Computer Science* Departments

University of Southern California

Simons Institute Berkeley ~ June 2019

(*by courtesy)

Acknowledgements

Current & Former Students

Hiteshi Sharma (USC)

Dileep Kalathil (Texas A&M)

Former Postdocs

Abhishek Gupta (Ohio State)

William Haskell (Purdue)

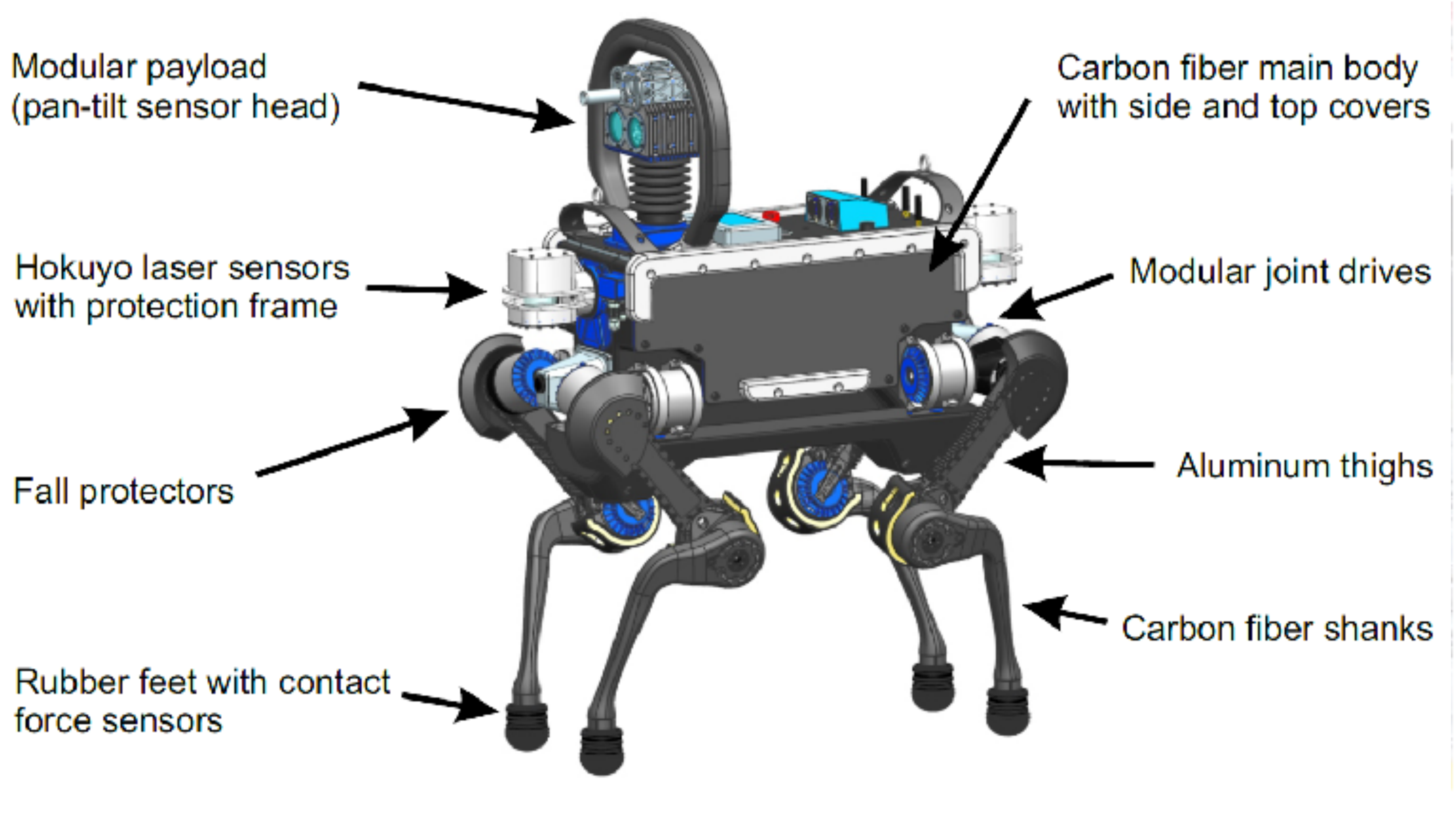
Other Collaborators

Vivek Borkar (IITB)

Peter Glynn (Stanford)

The *successes* of Deep RL

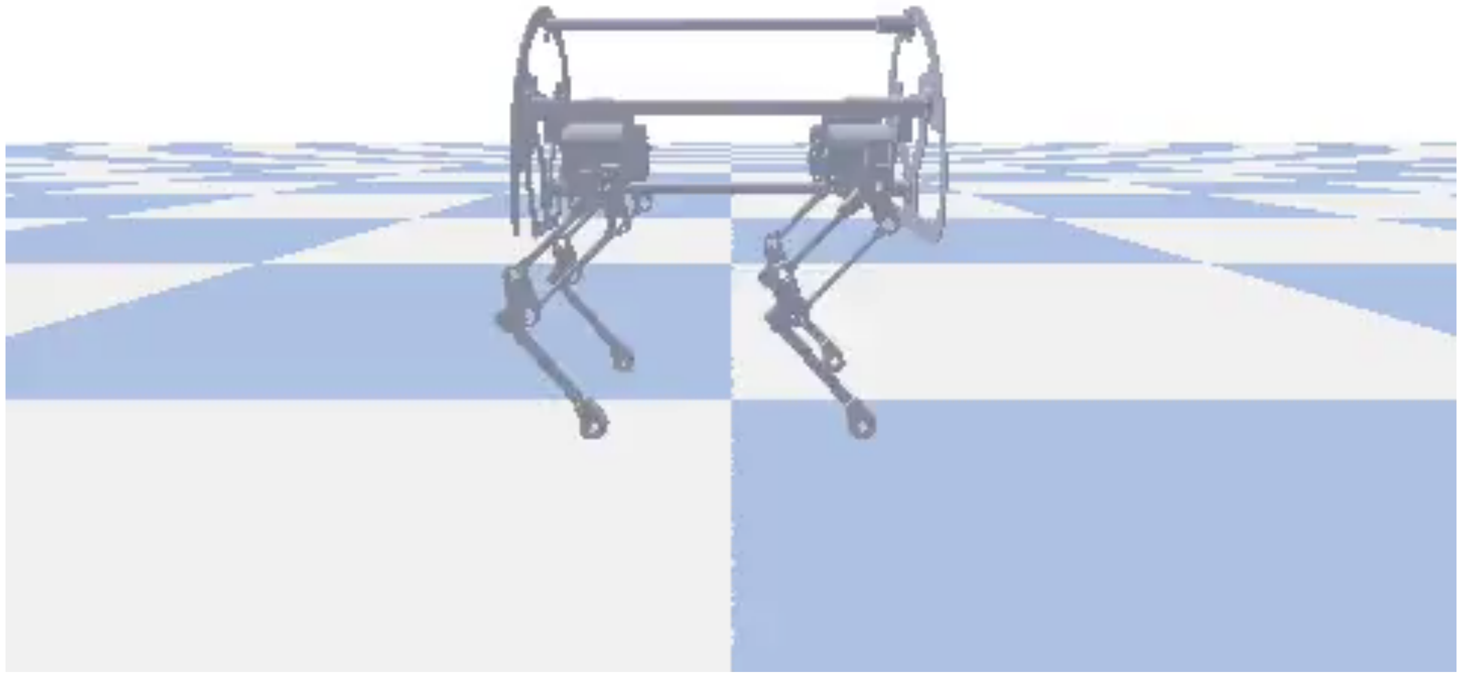
A simple mobile robotics problem



- *Robotic applications:* Continuous state and action spaces

Model-free approaches *near* impossible?

PPO, DDPG

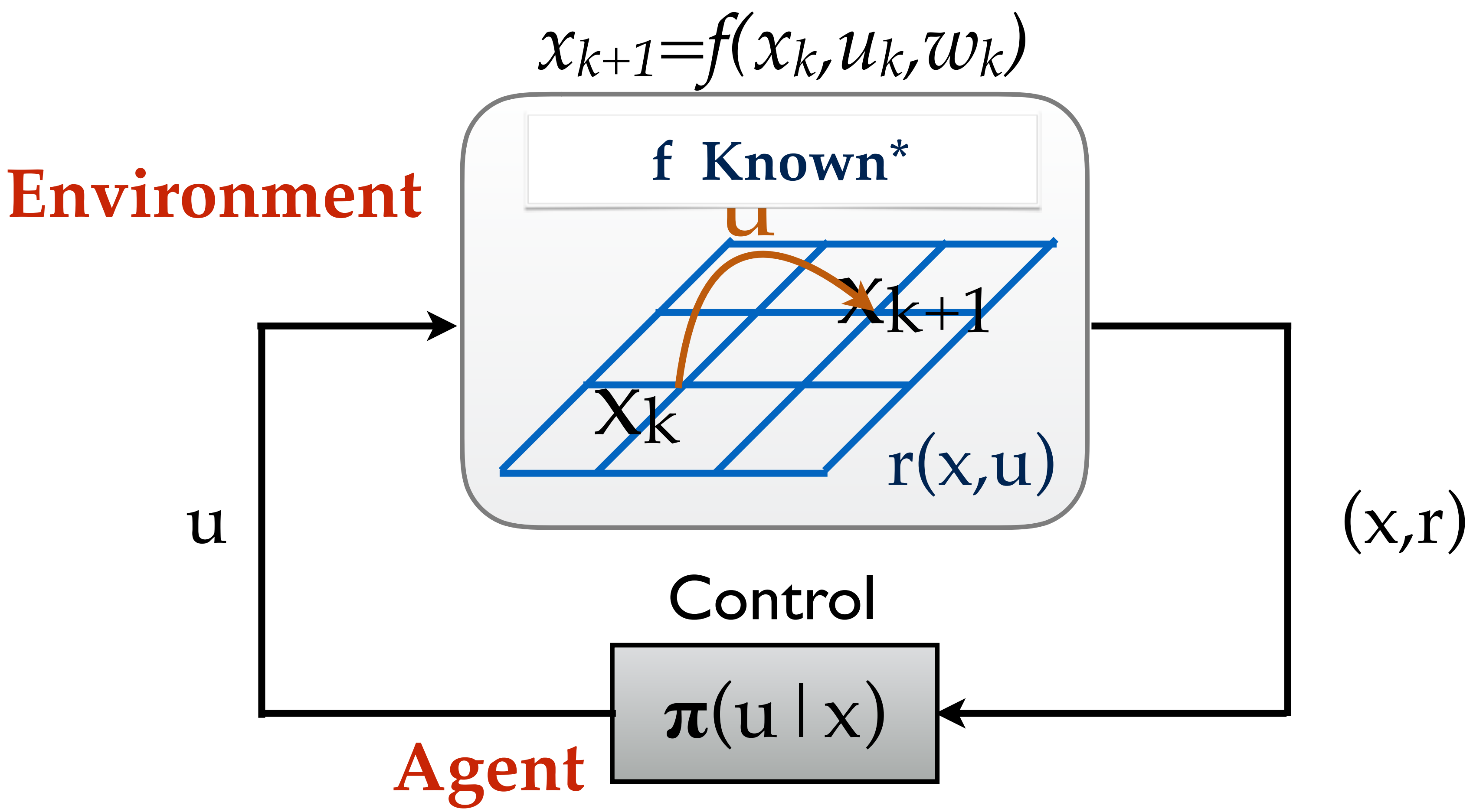


Courtesy: Bosch Center for CPS @ IISc, Bangalore

- ★ Deep RL: long training times, tuning hyper-parameters, no guarantees, *random search...?*
- ★ Train algorithms in simulation using a *generative model*

H. Mania, A. Guy, and B. Recht. Simple random search provides a competitive approach to reinforcement learning. *arXiv preprint arXiv: 1803.07055*, 2018.

The problem of Reinforcement Learning



MDP *Continuous* State space **X** *Continuous* Action space **U**

*Samples from a generative model available

- ★ Value of policy, $V_\pi(x) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r(x_t, u_t)]$ $0<\gamma<1$
- ★ Objective: $V^*(x) = \boldsymbol{sup_\pi}\, V_\pi(x)$

Bellman's Principle of Optimality

- ★ The dynamic programming equation

$$V^*(x) = [TV^*](x) = \sup_u \{r(x,u) + \gamma \textstyle\sum_y V^*(y)\theta(y|x,u)\}$$

E[V*(y) | x,u]

- ★ Bellman operator T is a contraction operator

$$||TV_1 - TV_2|| < ||V_1 - V_2||$$

- ★ Value Iteration: $V_{k+1} = TV_k = T^{k+1}V_0$

$$V_{k+1}(x) = [TV_k](x) = \sup_u \{r(x,u) + \gamma \mathbb{E}_\omega [V_k(\psi(x,u,\omega))\}$$

next state
from a
generative model

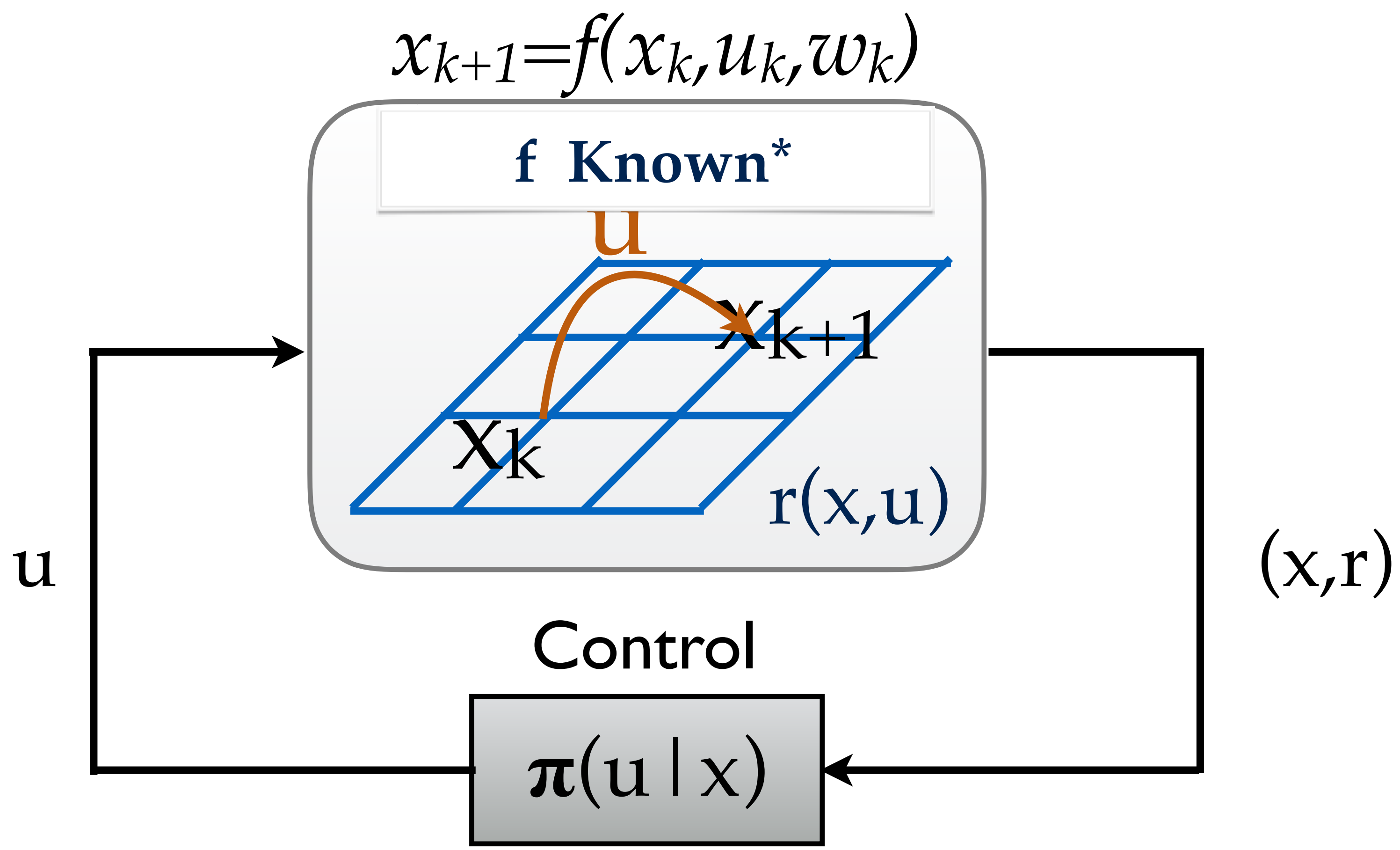
  - $V_k \to V^*$ *a.s.*

Outline

1. **A `Quasi-Model-free' RL Algorithm for finite MDPs**
2. Continuous state MDPs
3. Continuous state-action MDPs
4. `Online' RL for Continuous state MDPs

The Probabilistic Contraction Analysis Framework

Finite MDPs



MDP *Finite* State space **X** *Finite* Action space **U**

*Samples from a generative model available

- ★ Value of policy, $V_\pi(x) = \mathbb{E}_\pi[\sum_{t=0}^{\infty} \gamma^t r(x_t, u_t)]$ $0<\gamma<1$
- ★ Objective: $V^*(x) = \boldsymbol{sup_\pi}\ V_\pi(x)$
- ★ $V_{k+1}(x) = [TV_k](x) = \sup_u \{r(x,u) + \gamma \mathbb{E}[V_k(y)|x,u]\}$

Empirical Value Learning

Value Iteration *by simulation*

- ★ EVL:

$$\begin{aligned}\hat{V}_{k+1}(x) &= [\hat{T}\hat{V}_k](x) \\ &:= \sup_u \{r(x,u) + \gamma \hat{\mathbb{E}}^n [\hat{V}_k(\psi(x,u,\omega)]\} \\ &:= \sup_u \{r(x,u) + \frac{\gamma}{n}\sum_{i=1}^n \hat{V}_k(\psi(x,u,\omega_{k+1,i})]\}\end{aligned}$$

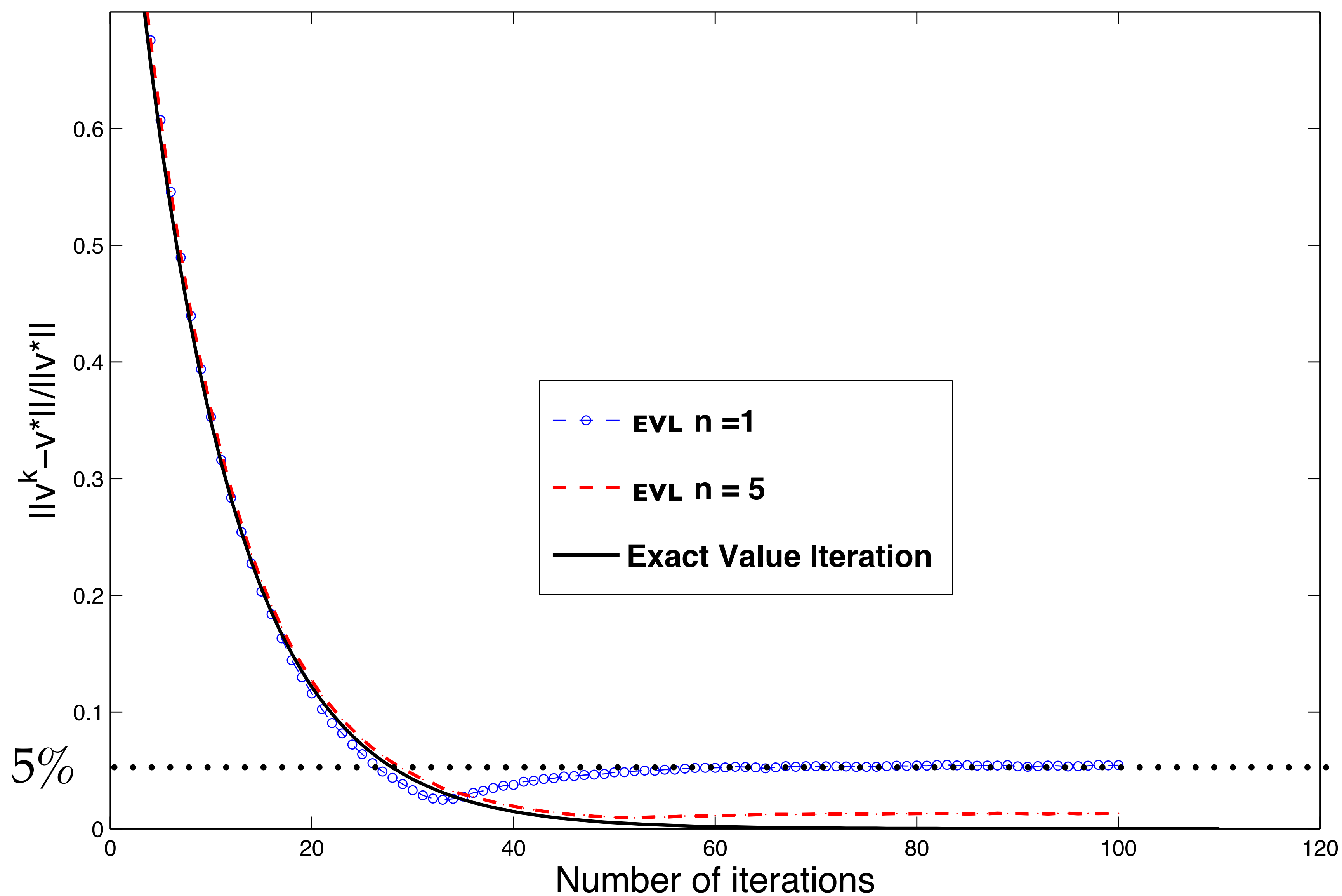
  where ω's are i.i.d. noise RVs

- ★ $V_0, \hat{V}_1, \hat{V}_2, \cdots$ *is a random sequence*
- ★ $\hat{T}$ is a *random operator,* $\mathbb{E}[\widehat{T}(V)] \neq T(V)$
- ★ *Non-incremental updates*

Does EVL Converge?

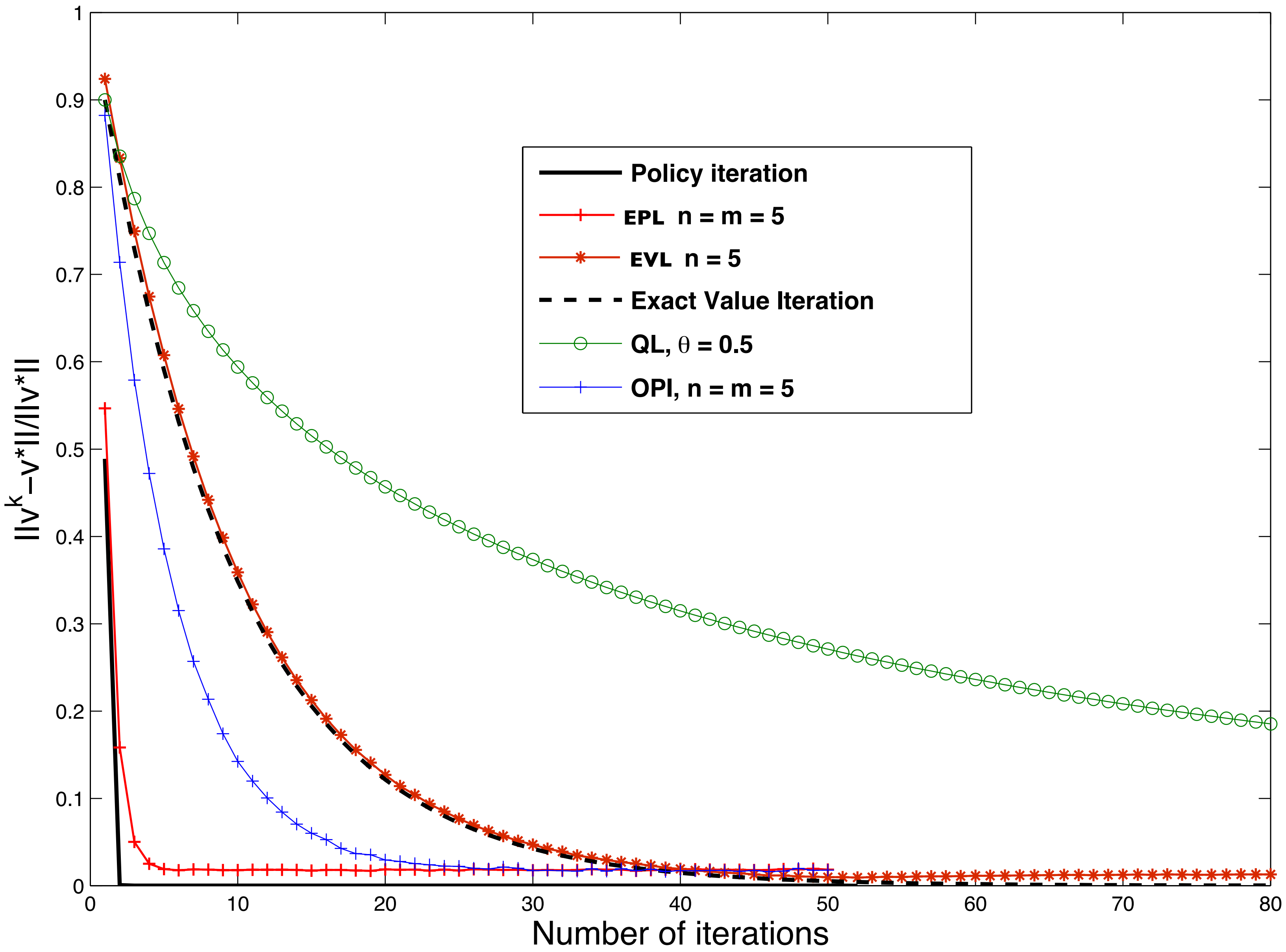
Numerical Evidence

100 States, 5 actions, Random MDP



Approx. Opt. in finite-time (w.h.p.)

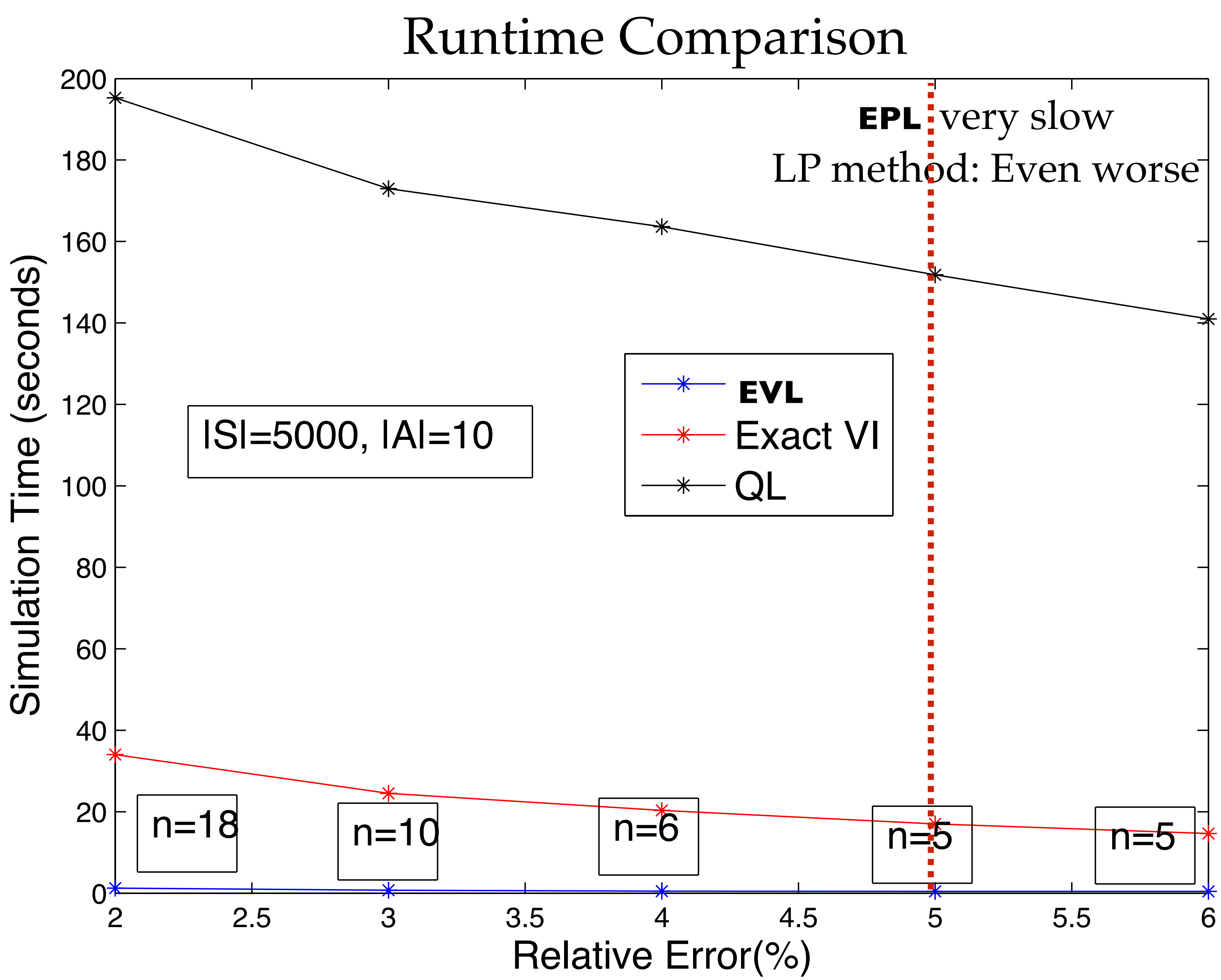
How do they compare?



- ★ States=100, Actions=5, random MDP
- ★ Offline QL with n=5 samples/iteration:

$$Q_{k+1} = (1-\alpha_k)Q_k + \alpha_k G Q_k, \qquad \sum_k \alpha_k = \infty, \sum_k \alpha_k^2 < \infty$$

Actual Runtime



- States=5000, Actions=10, random MDP.
  - All simulations run on a *Macbook Pro* under near-identical conditions

The Empirical Bellman Operator and *its Iterations*

Q. Can we prove convergence?

$$\hat{V}_k = \hat{T}(\omega_k)\hat{V}_{k-1} = \hat{T}(\omega_k)\cdots\hat{T}(\omega_1)V_0$$

- This is like product of random matrices
- *(V_k)* is a Markov chain. [Diaconis & Freedman'99]
  - Converges weakly
- Another way to look at it... whether $\hat{T}$ is *probabilistically contracting*, and has a *(probabilistic) fixed point*?

$$\boxed{\hat{V} = \hat{T}\hat{V}}$$

[5] A. Gupta, R. Jain and P. Glynn, "Probabilistic contraction analysis for iterated random operators", *submitted: Ann. Appl. Prob.*, Feb 2019.

Sample Complexity of EVL

n samples, k iterations

Theorem [1]:

Given $\epsilon \in (0, 1)$, $\delta \in (0, 1)$, select

$$n \geq \frac{C_1}{\epsilon^2} \log \frac{2|\mathbb{X}||\mathbb{A}|}{\delta}, \qquad k \geq \log \frac{1}{\delta \mu_{n,min}}$$

Then,

$$\mathbb{P}(||\hat{V}_k - V^*|| \leq \epsilon) \geq 1 - \delta.$$

- ★ `Sample Complexity' of EVL: $O(\frac{1}{\epsilon^2}, \ \log \frac{1}{\delta}, \log |\mathbb{X}||\mathbb{A}|)$
- ★ No assumptions on MDPs needed!
- ★ `Online' EVL converges under suitable recurrence conditions

[1] W. Haskell, R. Jain and D. Kalathil, "Empirical Dynamic Programming", *Mathematics of Operations Research*, 2016.

Outline

1. A `Quasi-Model-free' RL Algorithm for finite MDPs
2. **Continuous state MDPs**
3. Continuous state-action MDPs
4. `Online' RL for Continuous state MDPs

The Probabilistic Contraction Analysis Framework

MDPs with Continuous States

$x_{k+1}=f(x_k,u_k,w_k)$

`Universal’

Computationally *simple*

Arbitrarily good approximation

Non-asymptotic (*Probabilistic*) Guarantees

Continuous State Space MDPs

- State space Aggregation methods often don't work
- Function approximation *via* ϕ: $X \times \Theta \to \mathbb{R}$

$$V^*(x) \approx \sum_{j=1}^{J} \alpha_j \phi(x, \vartheta_j)$$

  - Approximation error depends on $d(\mathbf{\Phi}(\Theta), V^*)$, J, basis functions picked

$$\inf_{\alpha,\vartheta} \frac{1}{N} \sum_{n=1}^{N} |\tilde{V}(x_n) - \sum_{j=1}^{J} \alpha_j \phi(x_n, \vartheta_j)|^2$$

- (Deep) Neural Nets
  - Universal function approximators [Cybenko'89, Hornik, et al'89, Barron'93]
  - No guarantees: *How much data? How many layers/arch.? When to stop?* Lot of Computation.

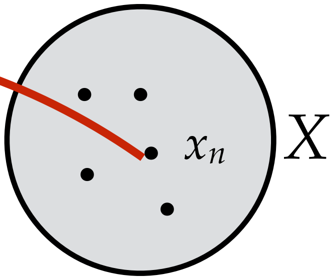
Use ‘Universal’ Function Approx. Spaces

Randomized Function Approximation
in a Universal Function Approximation Space

$$V_{k+1}(x) = \widehat{\Pi}_{\mathcal{H}_K}[\tilde{V}_k(x_1), \cdots, \tilde{V}_k(x_n)]$$

EVL+RKHS: A simple random basis function fitting algorithm

1. Sample $x_n \sim \mu$, basis functions $\boxed{\phi_n(x) = K(x_n, x)}$



2. EVL update:

$$\tilde{V}_{k+1}(x_n) = [\widehat{T}_M V_k](x_n) = \max_u \{r(x_n, u) + \frac{\gamma}{M} \sum_{m=1}^{M} V_k(X'_m)$$

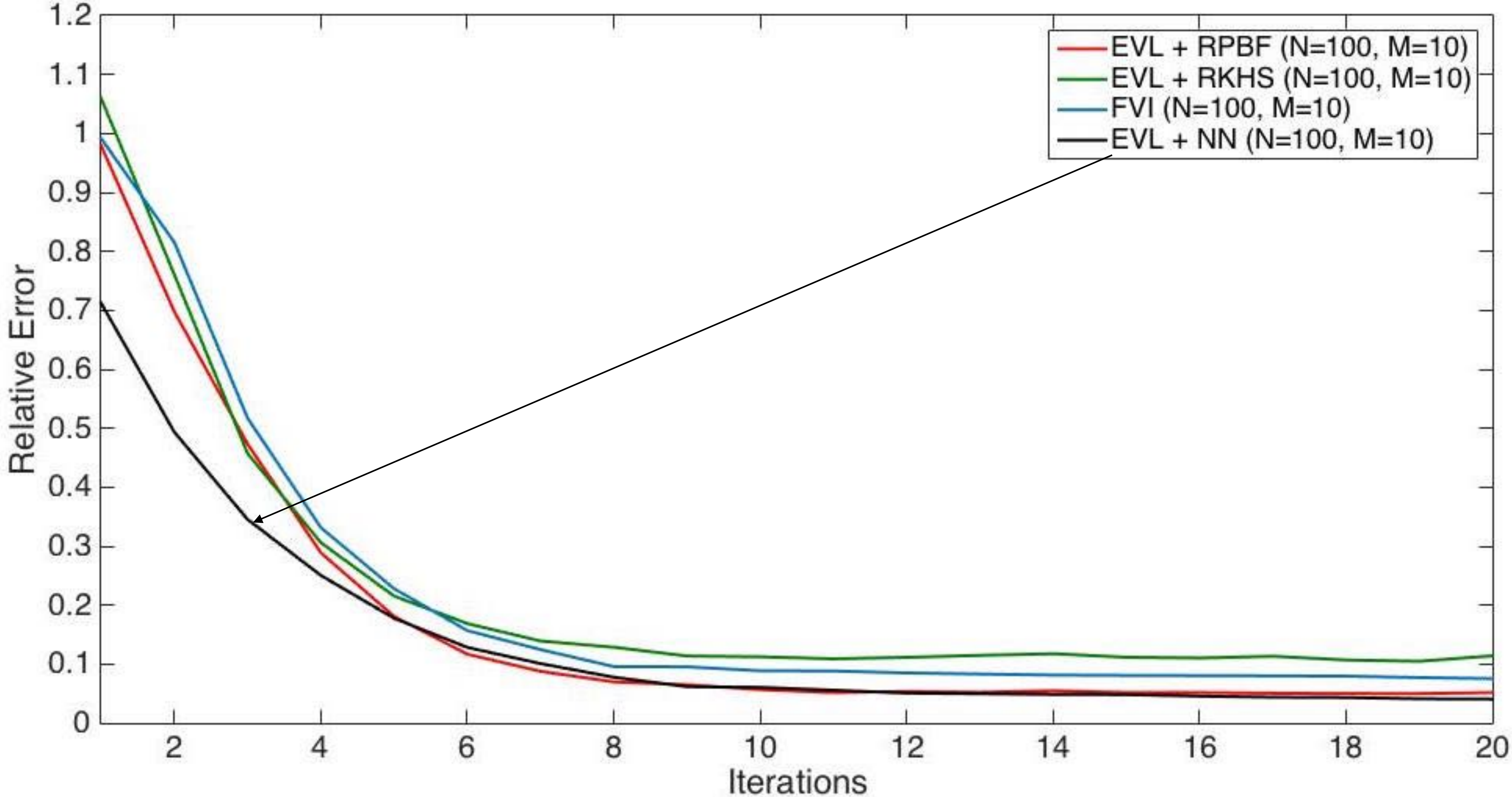
Next state from x_n

3. Randomized Function Approximation

$$V_{k+1}(x) = \sum_{n=1}^{N} \alpha_n K(x_n, x) = \widehat{\Pi}_{\mathcal{H}_K}[\tilde{V}_k(x_1), \cdots, \tilde{V}_k(x_n)]$$

Numerical Evidence

Optimal replacement problem



Sample Complexity of EVL+RPBF

N sampled points, J(=N) basis functions, M next states, K iterations

Theorem [2]:

Given $\epsilon \in (0, 1)$, $\delta \in (0, 1)$, select

$$N \geq N_\infty(\frac{1}{\epsilon^2}, \log\frac{1}{\delta}), \quad M \geq M_\infty(\frac{1}{\epsilon^2}), \quad K \geq K_\infty(\log\frac{1}{\delta})$$

Then,

$$||\hat{V}_k - V^*||_1 \leq \epsilon$$

with probability $> 1-\delta$.

- Dependence of N on ϵ is bad! but we get sup-error
- *Assumptions:* Absolute continuity of θ wrt μ and boundedness of Radon-Nikodym derivative $d\theta/d\mu$ needed!
- *Proof:* Randomized function fitting error concentration + Probabilistic Contraction Analysis of Iterated Random Operators

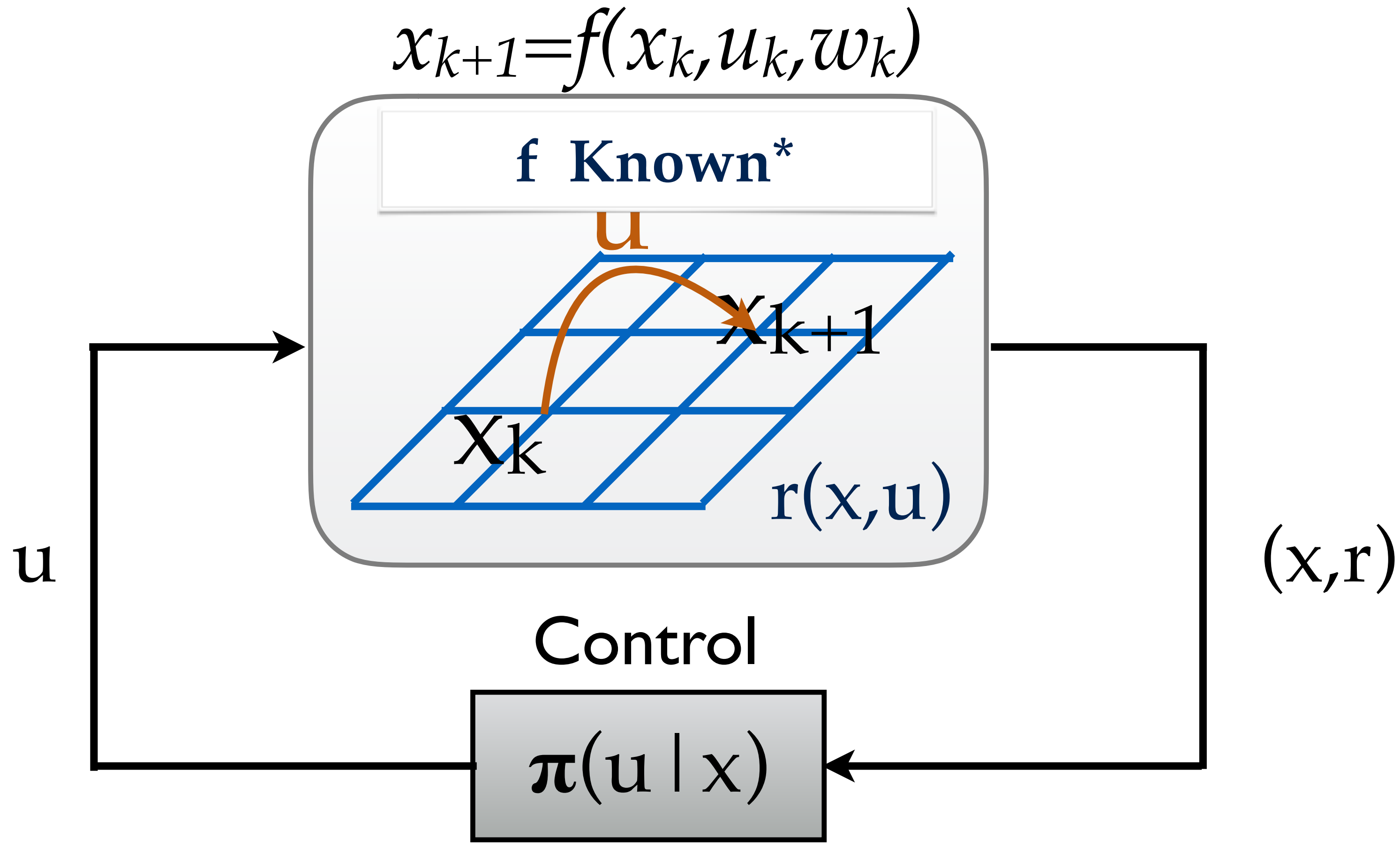
[2] W. Haskell, R. Jain, H. Sharma and P. Yu,,``An empirical dynamic programming algorithm for continuous state space MDPs'', *to appear, IEEE TAC*

Outline

1. A `Quasi-Model-free' RL Algorithm for finite MDPs
2. Continuous state MDPs
3. **Continuous state-action MDPs**
4. `Online' RL for Continuous state MDPs

The Probabilistic Contraction Analysis Framework

Continuous MDPs



MDP *Continuous* State space **X** *Continuous* Action space e.g., **U=[-1,1]**

*Samples from a generative model available

$$\tilde{V}_{k+1}(x_n) = [\widehat{T}_M V_k](x_n) = \max_u \{r(x_n, u) + \frac{\gamma}{M} \sum_{m=1}^{M} V_k(X'_m)$$

A simple RL Algorithm for Cont. state-action MDPs

RAEVL: Random Actions for Empirical Value Learning

1. Sample $x_n \sim \mu$, basis functions $\phi_n(x) = K(x_n,x)$

2. Sample $u_1,\ldots,u_n \sim$ Unif[U] (Can also do Adaptive Sampling, e.g., MCTS)

3. EVL update:

$$\tilde{V}_{k+1}(x_n) = [\widehat{T}_M V_k](x_n) = \max_{u_1^n}\{r(x_n,u_i) + \frac{\gamma}{M}\sum_{m=1}^{M} V_k(X'_m)$$

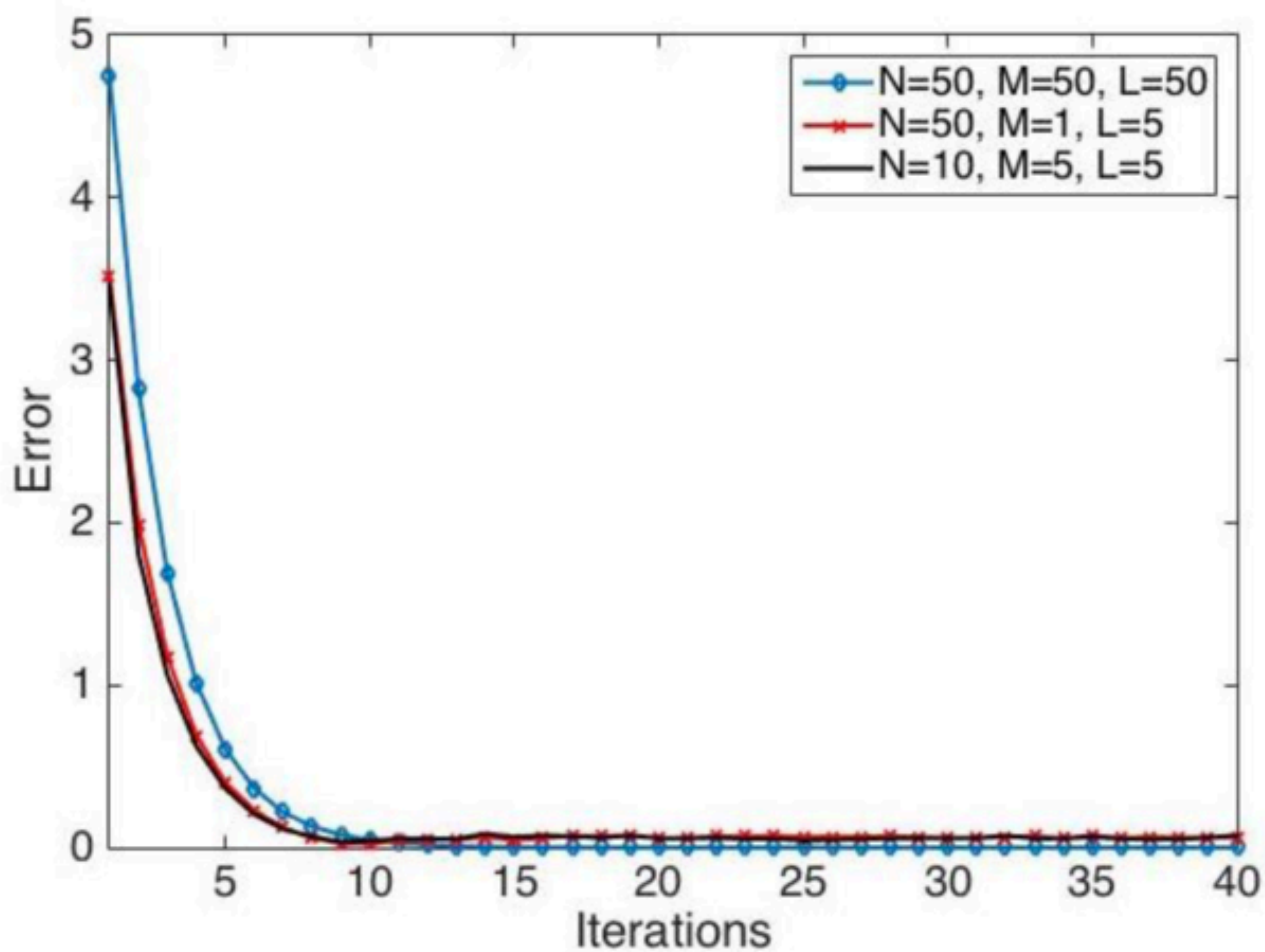
4. Randomized Function Approximation

$$V_{k+1}(x) = \sum_{n=1}^{N} \alpha_n K(x_n,x) = \widehat{\Pi}_{\mathcal{H}_K}[\tilde{V}_k(x_1),\cdots,\tilde{V}_k(x_n)]$$

Numerical Evidence

An MDP with X=[0,1], U=[0,1], $r(x,u) = -(x-u)^2$

$V^*(x)=0$



N sampled points, M next states, L actions

Sample Complexity of RAEVL

N sampled points, J basis functions, M next states, L actions, K iterations

Theorem [3]:

Given $\epsilon \in (0, 1)$, $\delta \in (0, 1)$, select $J \geq J_2(\frac{1}{\epsilon^2}, \log \frac{1}{\delta})$

$N \geq N_2(\frac{1}{\epsilon^4}, \log \frac{1}{\delta})$, $M \geq M_2(\frac{1}{\epsilon^2}), L \geq L_2(\frac{1}{\epsilon}, \log \frac{1}{\delta})$ $K \geq K_2(\log \frac{1}{\delta})$

Then,

$$||\hat{V}_k - V^*||_2 \leq C_1\epsilon + C_2\gamma^K$$

with probability > 1-δ.

- *Assumptions:* Lipschitz continuity of $r(x,.)$ and $\boldsymbol{\theta}(B\,|\,x,.)$
- *Assumptions:* Absolute continuity of $\boldsymbol{\theta}$ wrt μ and boundedness of Radon-Nikodym derivative $d\boldsymbol{\theta}/d\mu$ needed!
- *Proof: V^* is Lipschitz-cont., and bound sample complexity for approx optimal of a Lipschitz continuous function maximization by sampling*

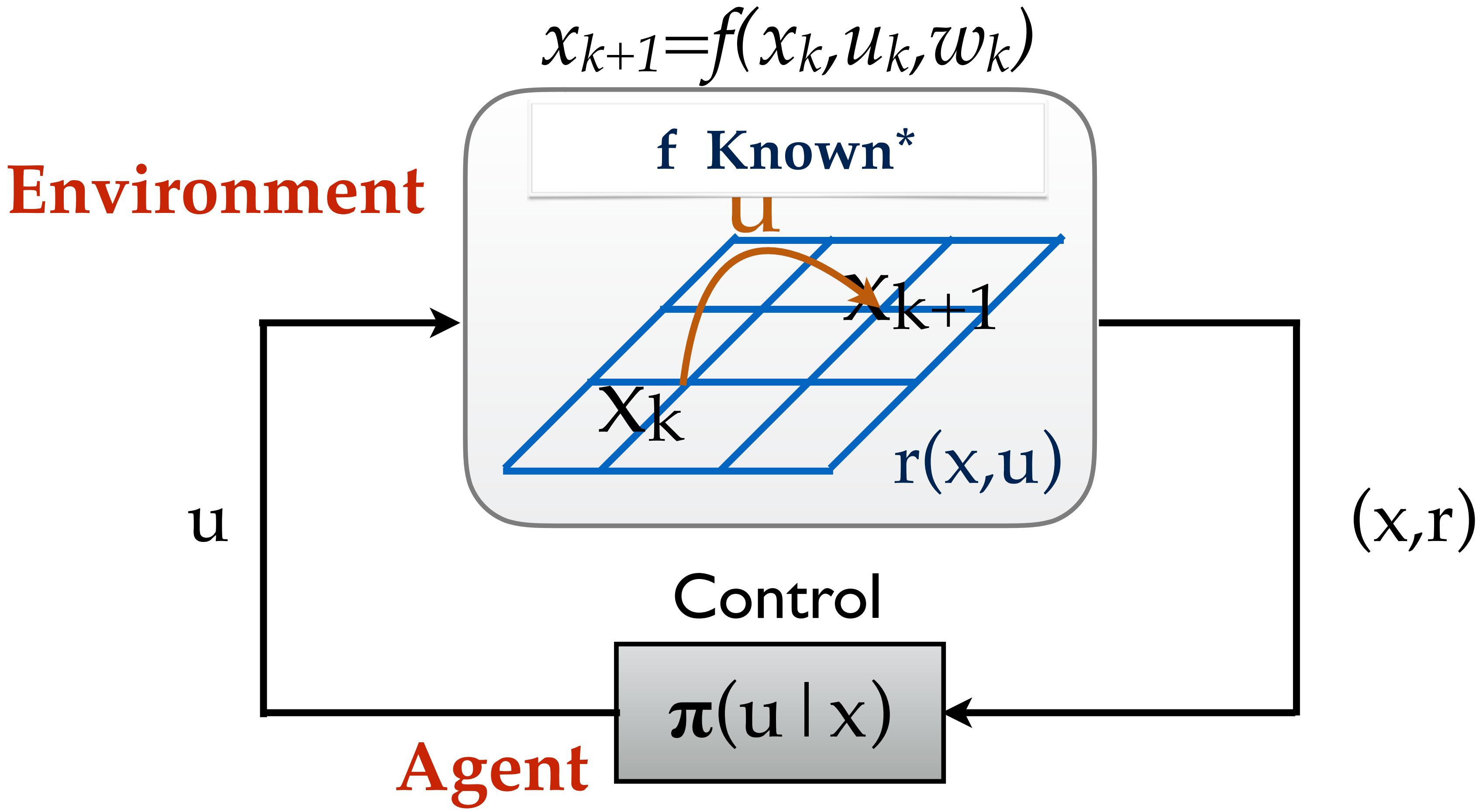
[3] H. Sharma and R. Jain,``Empirical algorithms for stochastic systems with continuous states and actions'', *submitted, March 2019*

Outline

1. A `Quasi-Model-free' RL Algorithm for finite MDPs
2. Continuous state MDPs
3. Continuous state-action MDPs
4. **`Online' RL for Continuous state MDPs**

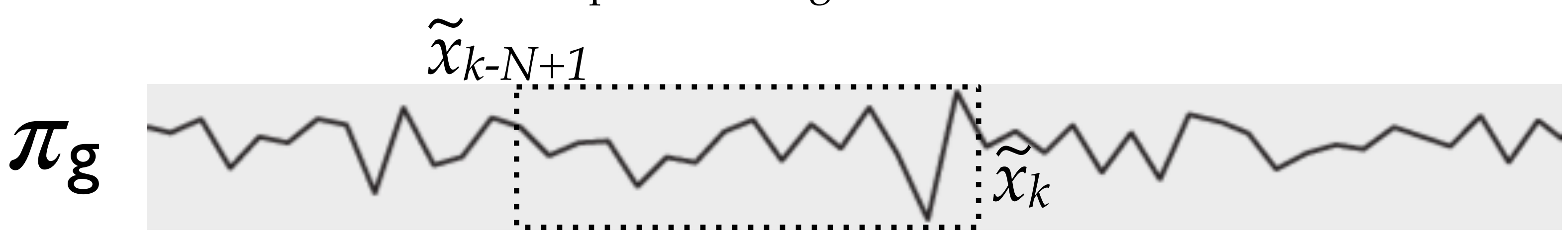
The Probabilistic Contraction Analysis Framework

An ‘Online’ RL Algorithm



MDP *Continuous* State space **X** *Finite* Action space **U**

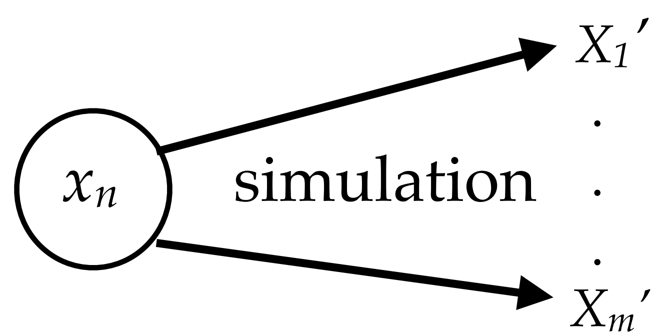
*Samples from a generative model available



- ⋆ *Fully randomized* policy, π_g: β-mixing with geometric rate [Nummelin-Tuominen’82]
- ⋆ Use N previous states, or from those visited so far

A ‘Online’ RL Algorithm for Cont. state MDPs

The Online-EVL algorithm



★ Pick basis functions randomly, optimize over weights

1. $x_n \sim [\tilde{x}_{k-N+1},\ldots,\tilde{x}_k]$, basis functions $\phi_n(x) = K(x_n,x)$

2. EVL update:

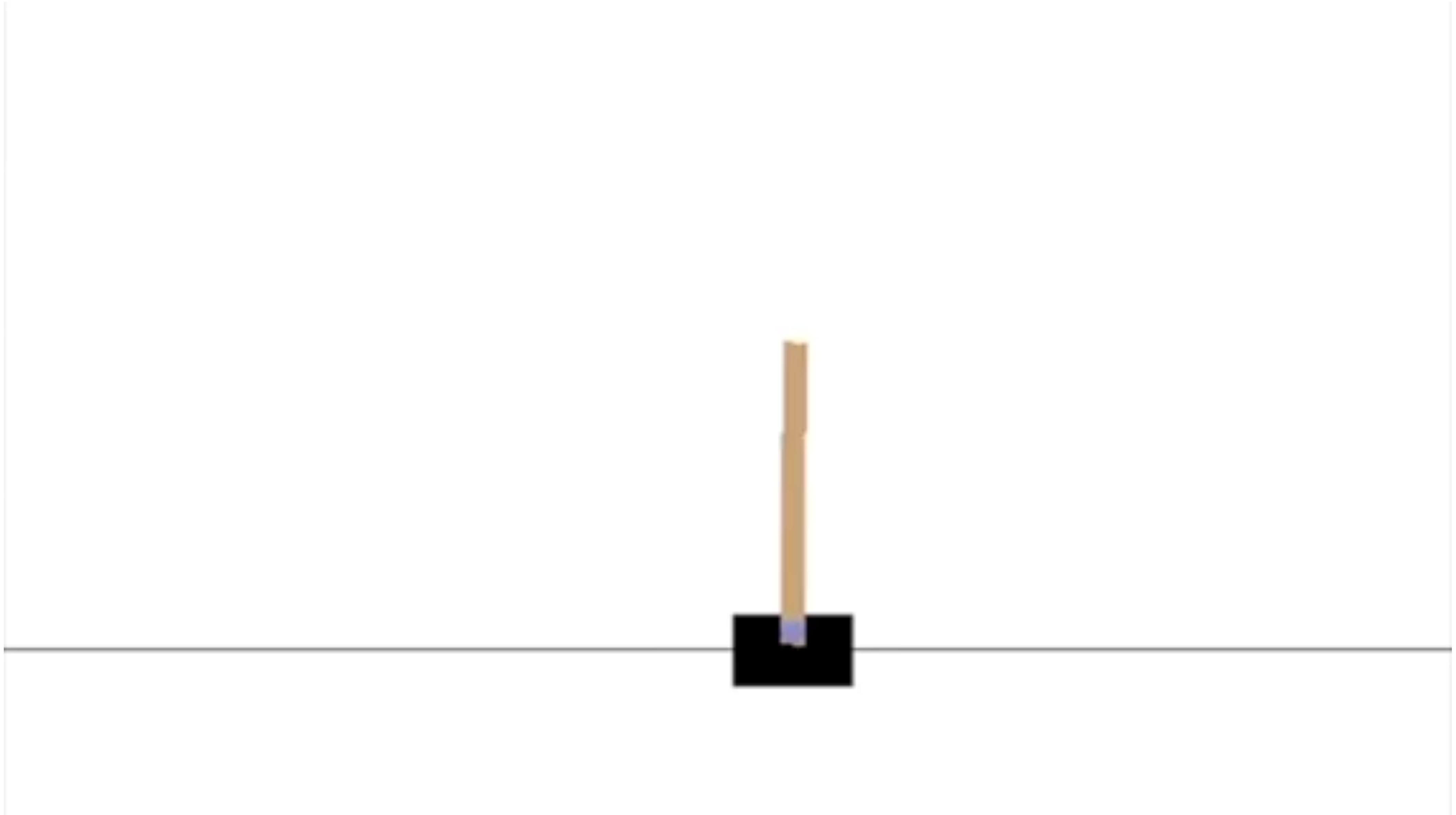
$$\tilde{V}_{k+1}(x_n) = [\widehat{T}_M V_k](x_n) = \max_u\{r(x_n,u) + \frac{\gamma}{M}\sum_{m=1}^{M} V_k(X_m')$$

3. Randomized Function Approximation

$$V_{k+1}(x) = \sum_{n=1}^{N} \alpha_n K(x_n,x) = \widehat{\Pi}_{\mathcal{H}_K}[\tilde{V}_k(x_1),\cdots,\tilde{V}_k(x_n)]$$

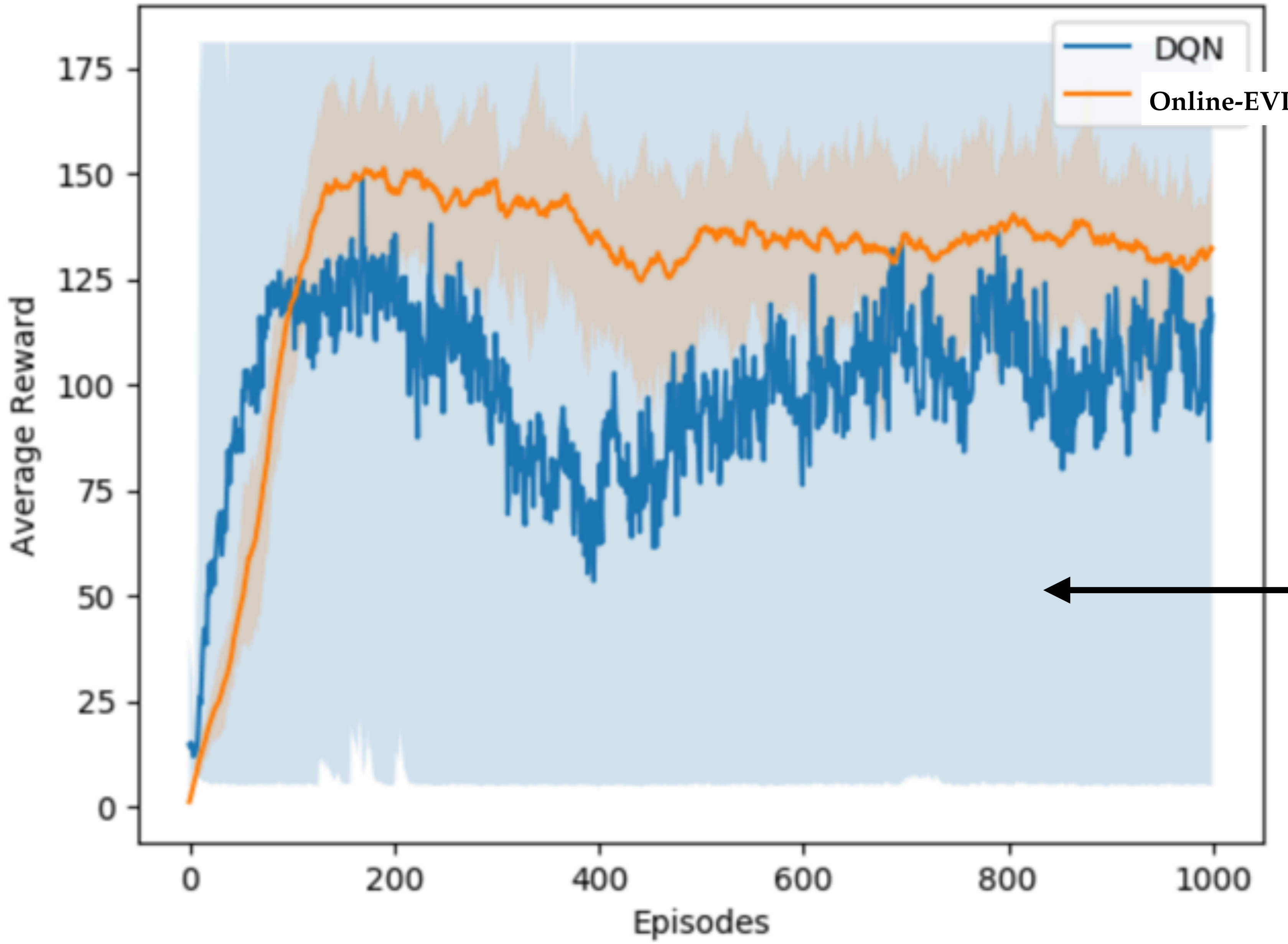
Does Online EVL work?

The Cartpole problem



Numerical Evidence

The Cartpole problem



DQN has very high variance

Of various other algorithms (ridge regression, Nystrom, Nearest-neighbor), DQN performs best. Runtime better than all except ridge regression which has poor performance

Sample Complexity of Online EVL

N sampled points, J basis functions, M next states, K iterations

Theorem [4]:

Given $\epsilon \in (0, 1)$, $\delta \in (0, 1)$, select $J \geq J_2(\frac{1}{\epsilon^2}, \log \frac{1}{\delta})$

$$N \geq N_2(\frac{1}{\epsilon^4}, \log \frac{1}{\delta}), \quad M \geq M_2(\frac{1}{\epsilon^4}), L \geq L_2(\frac{1}{\epsilon}, \log \frac{1}{\delta}) \quad K \geq K_2(\log \frac{1}{\delta})$$

Then,

$$||\hat{V}_k - V^*||_2 \leq C_1 \epsilon$$

with probability > $1-\delta$.

- *Assumptions:* Lipschitz continuity of $r(.,u)$ and $\boldsymbol{\theta}(B \mid .,u)$
- *Assumptions:* Absolute continuity of $\boldsymbol{\theta}$ wrt μ and boundedness of Radon-Nikodym derivative $d\boldsymbol{\theta}/d\mu$ needed!
- *Proof: Use beta-mixing to treat Markov chain samples as independent*

[4] H. Sharma and R. Jain,``A reinforcement learning algorithm for continuous state MDPs with finite time guarantees'', *submitted, May 2019*

Outline

1. A `Quasi-Model-free' RL Algorithm for finite MDPs
2. Continuous state MDPs
3. Continuous state-action MDPs
4. `Online' RL for Continuous state MDPs

The Probabilistic Contraction Analysis Framework

Key Analysis Idea:
View Stochastic Recursive Algorithms as Iteration of a Random Operator

Contraction Operator:

(for example)

$$V^* = TV^*, \quad \text{where} \quad [TV](x) = \sup_a \{r(x,a) + \gamma \mathbb{E}_\omega [V(\psi(x,a,\omega))]\}$$

$$||TV_1 - TV_2|| < \beta ||V_1 - V_2||, \text{ with } \beta < 1$$

Random Operators:

(for example)

$$\hat{V}_{k+1} = \hat{T}_n \hat{V}_k, \quad \text{where} \quad [\hat{T}_n V](x) = \sup_a \{r(x,a) + \gamma \frac{1}{n} \sum_{i=1}^{n} V(\psi(x,a,\omega_i))\}$$

$$||\hat{T}_n V_1 - \hat{T}_n V_2|| < \beta ||V_1 - V_2|| \quad \text{w.h.p.}$$

Probabilistic Contraction Property:

$$\text{PCP}_1: \qquad \mathbb{P}\left(||TV - \hat{T}_n V|| < \epsilon\right) > p_n(\epsilon),$$

where $p_n(\epsilon) \uparrow 1$ as $n \to \infty$ for all $\epsilon > 0$.

[4] W. Haskell, R. Jain and D. Kalathil, "Empirical Dynamic Programming", *Mathematics of Operations Research*, 2016.

Convergence to Probabilistic Fixed Points

★ $\hat{V}$ is a *Strong Probabilistic Fixed Point (SPFP)* of $\{\hat{T}_n\}$ if

$$\lim_{n \to \infty} \mathbb{P}\left(||\hat{T}_n \hat{V} - \hat{V}|| > \epsilon \right) = 0, \qquad \forall \epsilon > 0.$$

Theorem. [4] *We can obtain sample complexity bounds such that if* $n \geq n_0(\epsilon,\delta)$ *and* $k > k_0(\epsilon,\delta)$*, then*

$$\mathbb{P}(||\hat{V}_k^n - V^*|| > \epsilon) < \delta.$$

(where $n_0 = O(\frac{1}{\epsilon^2}, \log\frac{1}{\delta})$ *and* $k_0 = O(\log\frac{1}{\delta})$ *can be given explicitly)*

$$\text{PCP}_2: \quad ||\hat{S}_n(\omega)V_1 - \hat{S}_n(\omega)V_2|| < \boxed{\beta_n(\omega)}||V_1 - V_2||,$$

$$\text{and} \quad \mathbb{P}\left(\beta_n(\omega) \in (1-\epsilon, 1)\right) < \delta_n(\epsilon),$$

where $\epsilon < \epsilon_0$ for some ϵ_0, $\delta_n(\epsilon) \downarrow 0$ as $n \to \infty$ and $\beta_n(\omega) < 1$ a.s.

Probabilistic Contraction Analysis of Iterated Random Operators

- Algorithm converges to `*Weak probabilistic fixed points*' of random operators [1,5]
  - Stochastic dominance via a Markov chain
  - Stochastic optimization algorithms such as mini-batch versions of SGD, and SVRG can be shown to satisfy PCP_1 *and* PCP_2, and converge to WPFPs [5]

Problem↓/Methods→	Direct/Alt	Lyapunov	Contraction
Deterministic	Many	Well-established	Well-established
Stochastic	Martingale/Markov	Difficult	*None!*

[5] A. Gupta, R. Jain and P. Glynn, "Probabilistic contraction analysis for iterated random operators", *submitted: Ann. Appl. Prob.*, Feb 2019.

Conclusions

- *`Empirical' (RL) Algorithms are simple, `universal', have good numerical performance, average-case also [6]*
  - `Quasi-model free': Need a generative model
  - Weaker performance guarantees, but good numerical performance
- *A new analytical tool for Stochastic Iterative Algorithms:*
  - "Probabilistic Contraction analysis" *v.* Stochastic Lyapunov techniques *v.* Direct methods
  - Also useful for stochastic optimization algorithms: minibatch-SGD, SVRG, streaming variants
- *Future:*
  - Solving the robotic problem
  - Incorporating (safety) constraints

[6] H. Sharma, A. Gupta, R. Jain, "An Empirical Relative Value Learning Algorithm for Non-parametric MDPs with Continuous State Space", ECC 2019.

RL: Challenges

- RL Literature has focused on discrete (finite) state and action spaces
  - Continuous state and action space problems are way harder
- Online R. Learning for continuous state (and action) spaces needs ideas beyond Posterior Sampling
  - Search over Value function space
- RL with constraints?
- Formal RL for safety-critical applications
- Multi-Agent RL