Generalization and Exploration via Randomized Value Functions

Ian Osband, Benjamin Van Roy, Zheng Wen
iosband@google.com, bvr@stanford.edu, zwen@adobe.com

Stanford University

Contribution

Propose randomized least-squares value iteration (RLSVI), a new reinforcement learning (RL) algorithm designed to explore and generalize efficiently via linearly parameterized value functions. RLSVI is:

- BOTH provably efficient in the tabular learning case
- AND empirically efficient in several representative RL problems with value function generalization

Problem formulation

Learn to optimize a random finite horizon MDP $M = (\mathcal{S}, \mathcal{A}, R, P, H)$ in repeated episodes of interaction.

Figure 1: the reinforcement learning problem.

- State space $\mathcal{S}$, action space $\mathcal{A}$
- Rewards $r_t \sim R^M(s_t, a_t)$
- Transitions $s_{t+1} \sim P^M(s_t, a_t)$
- Finite episode length H

For MDP M policy μ, define value function:

$$Q^M_{\mu,h}(s,a) := \mathbb{E}_{M,\mu}\left[\sum_{j=h}^{H} \overline{r}^M(s_j,a_j)\Big| s_h=s, a_h=a\right],$$

We define the value $V^M_{\mu,h}(s) := Q^M_{\mu,h}(s,\mu(s,h))$ and the regret in episode k using μ_k on M^*

$$\Delta_k := \underbrace{V^{M^*}_{\mu^*,1}(s)}_{\text{optimal value}} - \underbrace{V^{M^*}_{\mu_k,1}(s)}_{\text{actual value}},$$

and $\text{Regret}(T,\pi,M^*) := \sum_{k=1}^{\lceil T/H\rceil} \Delta_k$.

Our goal is to design algorithms which can guarantee low regret (statistical efficiency) while remaining computationally tractable, even in large problems.

Linear Value Functions

The agent models that,

$$Q^*_h \in \text{span}[\Phi_h] \;\text{ for some } \Phi_h \in \mathbb{R}^{SA\times K}.$$

- We call Φ_h the generalization matrix.
- Φ_h is given a priori and is *not* learned.
- $Q^*_h \in \text{span}[\Phi_h] \implies$ coherent learning.
- $Q^*_h \notin \text{span}[\Phi_h] \implies$ agnostic learning.

Inefficient Exploration Schemes

There is a large literature on efficient exploration in RL. Most of these are motived by some combination of:

- Bayes-optimal tree search.
- Optimism in the face of uncertainty.
- Thompson sampling.

However, most of these algorithms become computationally intractable for large problems with generalization.

For this reason, most practical approaches to large-scale RL resort to simple **dithering exploration**.

- Dithering selectively takes random actions.
- e.g. ϵ-greedy and Boltzmann exploration
- can lead to regret that grows exponentially in H and/or $\mathcal{S}$ (see Kearns & Singh, 2002; Kakade, 2003)
- Efficient RL requires exploration which is directed over multiple timesteps = "deep exploration".

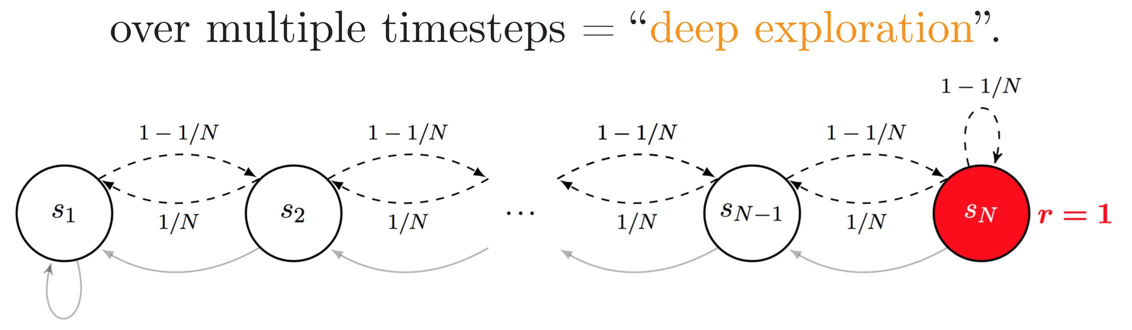

Figure 2: An MDP where dithering is highly inefficient.

- Consider a long chain with $\mathcal{S} = H = N$.
- Two actions "left" and "right" as shown in Figure 2.
- Optimal policy is to go right $V_h^*(s_1) = (1 - \frac{1}{N})^{N-1}$.
- Any other of the $2^{N\times N}$ policies will have 0 reward.
- Before reward dithering strategies explore at random.
- Thus, dithering has $\liminf_{T\to\infty}\text{Regret}(T) \geq 2^{\mathcal{S}-1}-1$.

High-Level Motivation

- Inspired by Thompson sampling for RL = PSRL.
- PSRL demonstrates efficient exploration with generalization (Osband and Van Roy 2014a;b) BUT
  - Requires model-based MDP planning.
  - Does not allow value function generalization.
- RLSVI uses an approximate posterior for PSRL.
- Bayesian linear regression for the value function.
- Posterior is wrong... but it might still be useful.

RLSVI Algorithm

Input: $\Phi_0(s_{i0},a_{i0}), r_{i0},..,\Phi_{H-1}(s_{iH-1},a_{iH-1}), r_{iH}: i<L$, Parameters $\lambda>0$, $\sigma>0$
Output: $\tilde\theta_{l0},...,\tilde\theta_{l,H-1}$
for $h=H-1,...,1,0$ **do**
  Generate regression problem $A\in\Re^{l\times K}$, $b\in\Re^l$:

$$A \leftarrow \begin{bmatrix}\Phi_h(s_{0h},a_{0h})\\ \vdots \\ \Phi_h(s_{l-1,h},a_{l-1,h})\end{bmatrix}$$

$$b_i \leftarrow \begin{cases} r_{ih} + \max_\alpha\left(\Phi_{h+1}\tilde\theta_{l,h+1}\right)(s_{i,h+1},\alpha) & \text{if } h<H-1\\ r_{ih}+r_{i,h+1} & \text{if } h=H-1\end{cases}$$

  Bayesian linear regression for the value function

$$\overline\theta_{lh} \leftarrow \frac{1}{\sigma^2}\left(\frac{1}{\sigma^2}A^\top A + \lambda I\right)^{-1} A^\top b$$

$$\Sigma_{lh} \leftarrow \left(\frac{1}{\sigma^2}A^\top A + \lambda I\right)^{-1}$$

  Sample $\tilde\theta_{lh} \sim N(\overline\theta_{lh}, \Sigma_{lh})$ from Gaussian posterior
end for

RLSVI with Greedy Action

Input: Features $\Phi_0,..,\Phi_{H-1}$; $\sigma>0$, $\lambda>0$
for $l=0,1,..$ **do**
  Compute $\tilde\theta_{l0},..,\tilde\theta_{l,H-1}$ using RLSVI algorithm
  Observe s_{l0}
  for $h=0,..,H-1$ **do**
    Sample $a_{lh} \in \arg\max_{\alpha\in\mathcal{A}}\left(\Phi_h\tilde\theta_{lh}\right)(s_{lh},\alpha)$
    Observe r_{lh} and $s_{l,h+1}$
  end for
  Observe r_{lH}
end for

Regret Bound for Tabula Rasa

We study a simple tabular setting without prior knowledge, $\Phi_h = I$ for all period h (i.e. without generalization).

Non-essential simplifying assumptions:

- $\mathcal{S}$, $\mathcal{A}$, H, and π, are deterministic
- rewards $R(s,a,h)$ are drawn from independent Dirichlet priors $\alpha^R(s,a,h) \in \Re^2_+$ on $\{-1,0\}$.
- transition probabilities $P(s,a,h,\cdot)$ are drawn from independent Dirichlet priors $\alpha^P(s,a,h) \in \Re^{\mathcal{S}}_+$.

Theorem: For RLSVI with $\Phi_h = I\;\forall h$, $\lambda \geq \max_{(s,a,h)}(\mathbb{1}^T\alpha^R(s,a,h) + \mathbb{1}^T\alpha^P(s,a,h))$ and $\sigma \geq \sqrt{H^2+1}$:

$$\mathbb{E}\left[\text{Regret}(T,\pi^{\text{RLSVI}},M^*)\right] \leq \tilde O\left(\sqrt{H^3\mathcal{S}AT}\right)$$

Remark: better than state-of-the-art $\tilde O(\sqrt{H^3\mathcal{S}^2AT})$ regret for tabular RL (see Jaksch et al., 2010)

Key Idea for Proof: the notion of stochastic optimism. It is not crucial that PSRL samples from the *exact* posterior distribution. RLSVI will succeed whenever the samples are sufficiently spread out but still concentrate with the data.

Experiment 1 - a chain MDP

Consider the MDP of Figure 2 with $\mathcal{S}=H=N=50$, where dithering strategies are provably inefficient.

Coherent learning: 10 basis functions are generated randomly to span a space which *does* include Q^*_h.

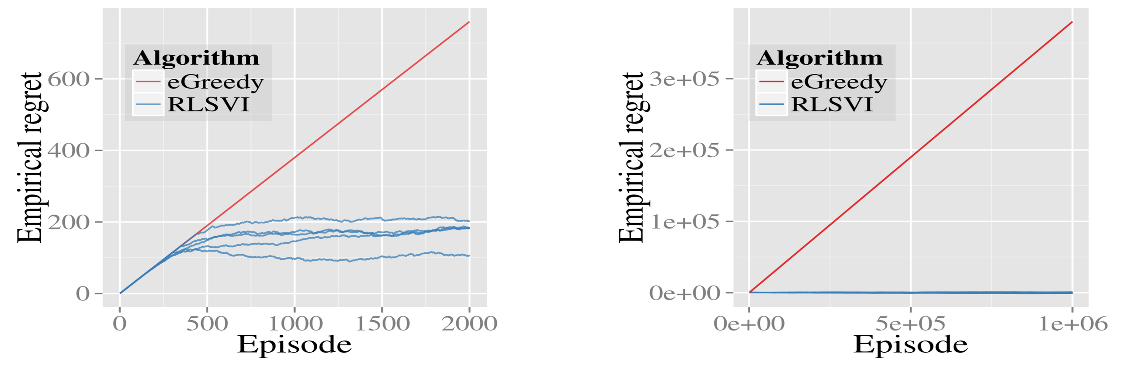

Figure 3: RLSVI demonstrates efficient exploration.

Experiment 1 - a chain MDP

- Dashed line: dithering lower bound 2^{N-1}.
- Solid line: $\frac{1}{10}H^2\mathcal{S}\mathcal{A}$ lower bound for any tabular learning algorithm (Dann & Brunskill, 2015)

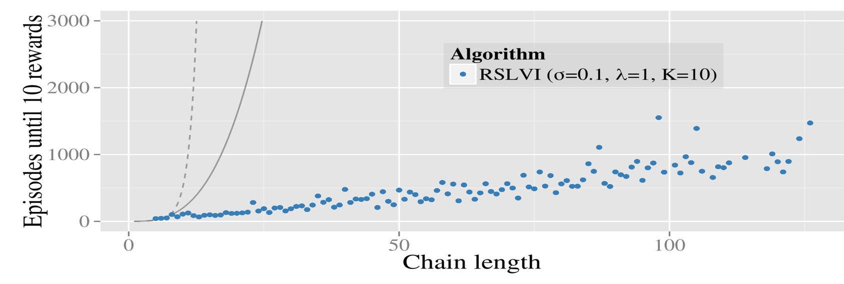

Figure 4: Examine RLSVI as we vary chain length N

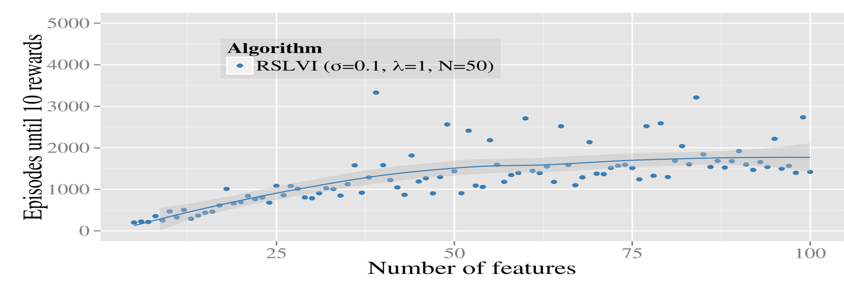

Figure 5: Examine RLSVI as we vary basis functions K

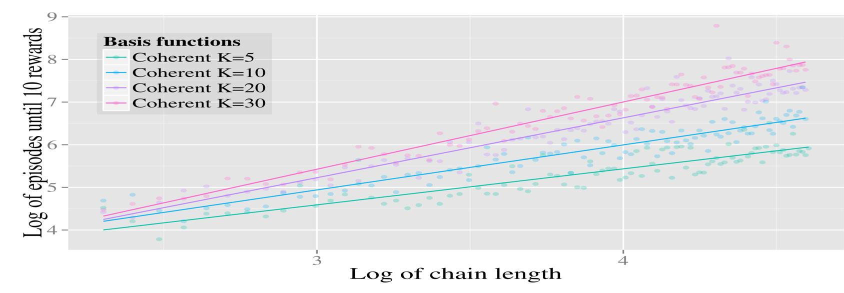

Figure 6: Empirical support for polynomial learning in RLSVI.

- Generate agnostic basis functions $\phi_{hk} \sim N(Q^*_h, \rho I)$

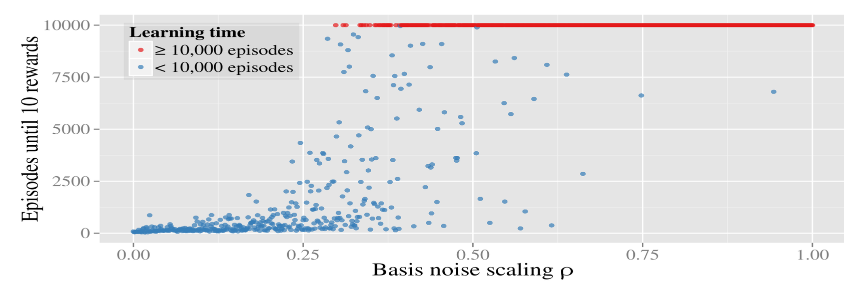

Figure 7: RLSVI is somewhat robust to model mis-specification.

Experiment 2 - Tetris

Apply RLSVI and LSVI (with tuned ϵ) to Tetris:

- 2D grid with 20 rows and 10 columns
- objective: maximize the total number of rows removed before the game ends
- 22 benchmark features (Bertsekas & Ioffe, 1996)
- no fixed episode length: adapt RLSVI/LSVI by approximating a time-homogenous Q^*

RLSVI/LSVI vs. LSPI with same features:

- higher final performance: RLSVI $\simeq$ 4500, LSVI $\simeq$ 3500, best score of LSPI: 3183
- RLSVI and LSVI learn from scratch while LSPI requires an initial policy

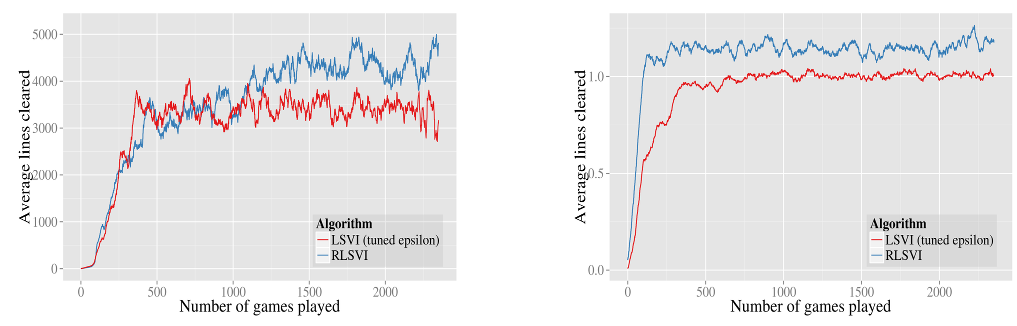

Figure 8: Learning curves for LSVI + RLSVI (left). Improvement magnified on difficult 4-row tetris with SZ pieces (right).

Experiment 3 - Recommendations

Recommend J out of N products sequentially. State $x \in \{\pm1, 0\}^N$ indicates what products the customer has observed, and whether she likes or dislikes each one. The probability the customer will like a new product a is

$$\mathbb{P}(a|x) = 1/\left(1+\exp\left(-[\beta_a + \textstyle\sum_n \gamma_{an}x_n]\right)\right)$$

RL setting: (1) $\mathbb{P}(a|x)$ is unknown; (2) each customer is modeled as an episode with horizon $H=J$; (3) $\beta=0$ and γ is randomly sampled; (4) $K = N^2+N$ basis functions: $\phi_m(x,a) = \mathbf{1}\{a=m\}$ and $\phi_{mn}(x,a) = x_n\mathbf{1}\{a=m\}$

- RLSVI outperforms LSVI with Boltzmann exploration (with a wide range of temperatures)
- RLSVI outperforms bandit algorithms (both contextual and non-contextual) and optimal myopic policy

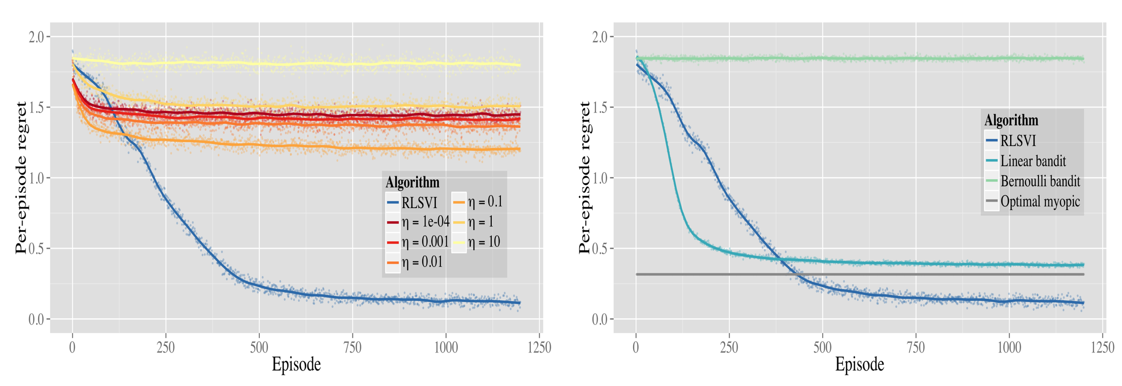

Figure 9: RLSVI drives an efficient recommendation system.