Finite-sample analysis of Least Squares Temporal Differences

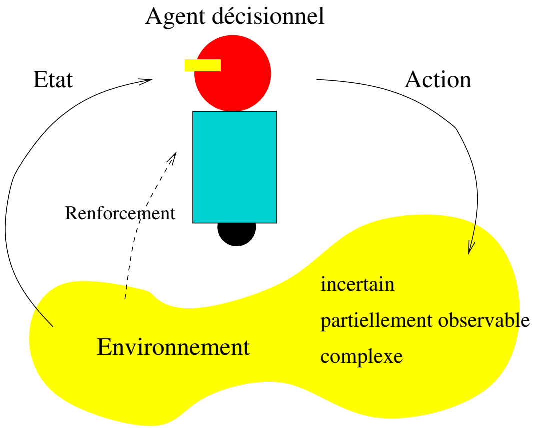
Rémi Munos, joint work with A. Lazaric, M. Ghavamzadeh

Sequential Learning project
INRIA Lille - Nord Europe, France

Journées MAS 2010, Bordeaux

Reinforcement Learning, the big picture

Learning to make decisions from interactions with an unknown environment.



Markov decision Process

A MDP is defined by

- State space $\mathcal{X}$,
- Action space A,
- Transition probabilities $P(\cdot|x, a)$,
- Reward function $r : \mathcal{X} \times A \mapsto \mathbb{R}$.

Goal: Find policy $\pi : \mathcal{X} \mapsto A$ that maximizes the (expected) sum of discounted rewards

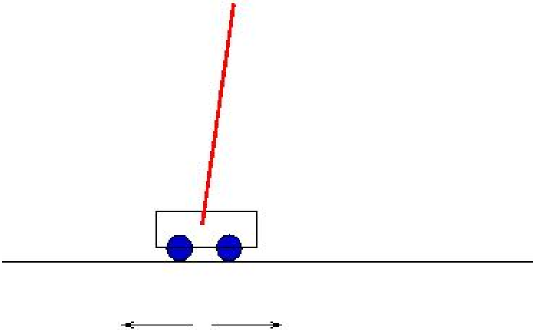
$$V^\pi(x) = \mathbb{E}\Big[\sum_{t\geq 0} \gamma^t r(X_t, \pi(X_t))|X_0 = x; \pi\Big],$$

where the discount factor $\gamma < 1$.
Definitions:

- V^π is called the value funcion for policy π,
- $V^*(x) = \sup_\pi V^\pi(x) = V^{\pi^*}(x)$ is the optimal value function and π^* an optimal policy.

Policy iteration: illustration

Inverted pendulum:



(click to start movie. Thanks to Martin Riedmiller)

Policy iteration: setting

Start with a policy π_0, then iterate: for $k \geq 0$,

- **Policy evaluation step:** For policy π_k, compute an approximation V_k of the value function V^{π_k}
- **Policy improvement step:** Build a new policy

$$\pi_{k+1}(x) \stackrel{\text{def}}{=} \arg\max_{a \in A} \Big[r(x,a) + \gamma \int_{\mathcal{X}} P(dy|x,a) V_k(y) \Big].$$

How good is π_k compared to π^*?

Policy iteration: results

Known results:

- **Exact policy evaluation:**
  If $V_k = V^{\pi_k}$, then $V^{\pi_{k+1}} \geq V^{\pi_k}$ and $\lim_{k\to\infty} V^{\pi_k} = V^*$.
- **Approximate policy evaluation in L_∞-norm** [Bertsekas and Tsitsiklis, 1996]:

$$\limsup_{k\to\infty} ||V^* - V^{\pi_k}||_\infty \leq \frac{2}{(1-\gamma)^2} \limsup_{k\to\infty} ||V_k - V^{\pi_k}||_\infty.$$

- **Approximate policy evaluation in L_p-norm** [Munos, 2003]:

$$\limsup_{k\to\infty} ||V^* - V^{\pi_k}||_{p,\mu} \leq \frac{2}{(1-\gamma)^2} C(\mu,\rho)^{1/p} \limsup_{k\to\infty} ||V_k - V^{\pi_k}||_{p,\rho}.$$

Performance of PI results from performance of the policy evaluation steps.

What this talk is about...

For a given policy, the MDP reduces to a Markov chain. Our goal is to approximate the corresponding value function V

$$V(x) \stackrel{\text{def}}{=} \mathbb{E}\Big[\sum_{t\geq 0} \gamma^t r(X_t) | X_0 = x\Big]$$

Methodology:

- Choose a function space $\mathcal{F}$
- Observe a trajectory $X_1, \dots, X_n$ following the policy
- Build an estimate $\widehat{V} \in \mathcal{F}$ of V
- Derive bounds on the approximation error $||\widehat{V} - V||$ in terms of
  - How well the function space $\mathcal{F}$ can approximate V
  - Capacity of $\mathcal{F}$
  - Number of samples n (sample complexity)

Some properties of the value function

- V is unique solution to the Bellman equation:

$$V(x) = r(x) + \gamma \int P(dy|x)V(y) \tag{1}$$

- Define the Bellman operator T:

$$TW(x) \stackrel{\text{def}}{=} r(x) + \gamma \int P(dy|x)W(y).$$

  Then (1) writes

$$V = TV.$$

- Property: T is a contration in $||\cdot||_\infty$.
- Thus from Banach fixed point theorem, T has a unique fixed point, which is V.

Linear approximation

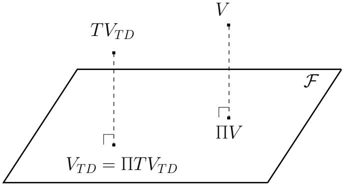
Let $\varphi_1, \dots, \varphi_d$ be a set of functions $\mathcal{X} \to \mathbb{R}$, and the linear space

$$\mathcal{F} \stackrel{\text{def}}{=} \left\{ f_\alpha(x) \stackrel{\text{def}}{=} \sum_{i=1}^{d} \alpha_i \varphi_i(x), \alpha \in \mathbb{R}^d \right\}$$

Best approximation of V in $\mathcal{F}$ is

$$\Pi V = \arg\min_{f \in \mathcal{F}} ||V - f||$$

(i.e. Π is the projection onto $\mathcal{F}$)



LSTD solution: fixed point of ΠT, i.e. $V_{TD} = \Pi T V_{TD}$.

Question: what norm should we use in the projection?

Known result [Tsitsiklis and Van Roy, 1997]

Assuming that the Markov chain has a stationary distribution μ (i.e. $\mu P = \mu$), then T is a contraction mapping in $L_{2,\mu}$-norm (i.e. such that $||f||_\mu^2 = \int f(x)^2 \mu(dx)$).
Thus ΠT is a contraction mapping and there exits a TD solution V_{TD}, fixed-point of ΠT. We have

$$||V - V_{TD}||_\mu \leq \frac{1}{\sqrt{1-\gamma^2}} ||V - \Pi V||_\mu.$$

Now we wish to address those questions:

- Is it possible to approximate V_{TD} using a finite number of samples?
- What is the quality of that approximation?
- What if the chain does not possess a stationary distribution?

Pathwise LSTD

Observe a sample path $(X_1, \dots, X_n)$ of the Markov chain.

- Consider $\mathcal{F}_n = \{(f_\alpha(X_1), \dots, f_\alpha(X_n))^T, \alpha \in \mathbb{R}^d\} \subset \mathbb{R}^n$.
- Define the empirical projection: $\widehat{\Pi} u = \inf_{w \in \mathcal{F}_n} ||u - w||$,
- Define the empirical Bellman operator:

$$(\widehat{T} u)_t = \begin{cases} r(X_t) + \gamma u_{t+1} & \text{for } t < n, \\ r(X_n) & \text{otherwise} \end{cases}$$

Property: $\widehat{T}$ is a contraction mapping. Thus $\widehat{\Pi}\widehat{T}$ has a unique fixed-point, $\widehat{v} \in \mathcal{F}_n$, whose corresponding $\widehat{\alpha}$ solves the linear system $\widehat{A}\alpha = \widehat{b}$ with

$$\widehat{A}_{i,j} \stackrel{\text{def}}{=} \frac{1}{n}\Big(\sum_{t=1}^{n-1} \varphi_i(X_t)[\varphi_j(X_t) - \gamma\varphi_j(X_{t+1})] + \varphi_i(X_n)\varphi_j(X_n)\Big)$$

$$\widehat{b}_i \stackrel{\text{def}}{=} \frac{1}{n}\sum_{t=1}^{n} r(X_t)\varphi_i(X_t).$$

$\widehat{V} = f_{\widehat{\alpha}}$ is called the **pathwise LSTD** solution.

Finite-time analysis of pathwise LSTD

Define the empirical norm $||f||_n \stackrel{\text{def}}{=} \left[\frac{1}{n}\sum_{t=1}^{n} f(X_t)^2\right]^{1/2}$.

Theorem

With probability $1-\delta$ *(w.r.t. the sample path),*

$$||V-\widehat{V}||_n \leq \frac{1}{\sqrt{1-\gamma^2}} \inf_{f\in\mathcal{F}} ||V-f||_n + \frac{2\gamma V_{\max} L}{1-\gamma}\sqrt{\frac{2d\log(2d/\delta)}{n\,\nu}} + O\Big(\frac{1}{n}\Big),$$

where $L = \max_{1\leq i\leq d} ||\varphi_i||_\infty$ *and* $\nu > 0$ *is the smallest strictly positive eigenvalue of the Gram matrix:*

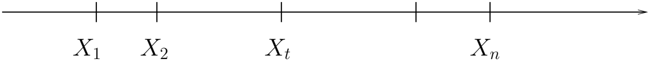
$$M \stackrel{\text{def}}{=} \frac{1}{n}\sum_{t=1}^{n} \varphi(X_t)\varphi(X_t)^T.$$

Properties

- This is a no-assumption theorem...
- $\widehat{V}$ is well-defined for any n and any Markov chain.
- No assumption about stationarity!

Example:

- Markov chain on the real line where transitions always move to the right $\rightarrow$ no stationary distribution



- A good estimate of the value function at a state X_t is learned from noisy pieces of information at states that may be far away from X_t.

Learning the value function at a given state does not require making an average over many samples close to that state.

Sketch of proof

Let $v, \widehat{v} \in \mathbb{R}^n$, $v_t = V(X_t)$, $\widehat{v}_t = \widehat{V}(X_t)$,

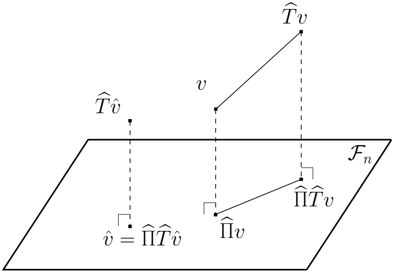
$\mathcal{F}_n = \{(f(X_1), \ldots, f(X_n)), f \in \mathcal{F}\} \subset \mathbb{R}^n$

Empirical projection operator: $\widehat{\Pi}$

Empirical Bellman operator: $\widehat{T}$

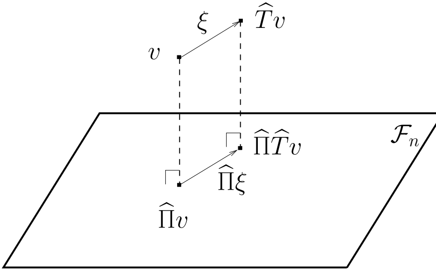
$$(\widehat{T}u)_t = \begin{cases} r_t + \gamma v_{t+1} & \text{for } t < n, \\ r_n & \text{otherwise} \end{cases}$$

Property: $\widehat{T}$ is a contraction



$$\begin{aligned} ||\widehat{v} - v||_n^2 &\leq \big(\underbrace{||\widehat{v} - \widehat{\Pi}\widehat{T}v||_n} + ||\widehat{\Pi}\widehat{T}v - \widehat{\Pi}v||_n\big)^2 + ||\widehat{\Pi}v - v||_n^2 \\ &\qquad = ||\widehat{\Pi}\widehat{T}\widehat{v} - \widehat{\Pi}\widehat{T}v||_n \leq ||\widehat{T}\widehat{v} - \widehat{T}v||_n \leq \gamma||\widehat{v} - v||_n \\ &\leq \big(\gamma||\widehat{v} - v||_n + \underbrace{||\widehat{\Pi}\widehat{T}v - \widehat{\Pi}v||_n}_{\text{estimation error}}\big)^2 + \underbrace{||\widehat{\Pi}v - v||_n^2}_{\text{approx. error}}. \end{aligned}$$

Estimation error term



Estimation error:
$||\widehat{\Pi}v - \widehat{\Pi}\widehat{T}v||_n^2 = ||\widehat{\Pi}\xi||_n^2$, where

$$\xi_t = V(X_t) - \big[r(X_t) + \gamma V(X_{t+1})\big]$$

We have $\mathbb{E}[\xi_t | X_t] = 0$,
thus $\mathbb{E}[\xi] = 0$.

... but the ξ_t are NOT independent! and $\widehat{\Pi}$ is itself random...
Thus $\mathbb{E}[\widehat{\Pi}\xi] \neq 0$ and

$$\mathbb{E}[||\widehat{\Pi}\xi||_n^2] = \frac{1}{n}\mathbb{E}[\xi^T\widehat{\Pi}\xi] \neq \mathbb{E}[||\xi||_n^2]\mathrm{tr}(\widehat{\Pi}) \leq C\frac{d}{n}.$$

(which would be the case with a deterministic design).

Regression with Markov design

Let $X_1, \ldots, X_n$ be a sample path of the Markov chain. Let

$$Y_t = f(X_t) + \xi_t, \quad \text{with} \quad \mathbb{E}[\xi_t | X_1, \ldots, X_t] = 0,$$

and ξ_t adapted to the filtration generated by $X_1, \ldots, X_{t+1}$.
Write $\widehat{\Pi}\xi$ the projection of the noise ξ onto $\mathcal{F}_n$.

Lemma

For any $\delta > 0$, with probability $1 - \delta$,

$$||\widehat{\Pi}\xi||_n \leq CL\sqrt{\frac{2d \log(2d/\delta)}{n\nu}},$$

where C is a bound on $||\xi_t||_\infty$, L is a bound on $||f||_\infty$, and ν is the smallest strictly-positive eigenvalue of the Gram matrix $\frac{1}{n}\sum_{t=1}^{n} \varphi(X_t)\varphi(X_t)^T$.

Corollary: This concludes the proof of the Theorem since the estimation error $||\widehat{\Pi}v - \widehat{\Pi}\widehat{T}v||_n = ||\widehat{\Pi}\xi||_n$.

Proof of the Lemma

Since $\widehat{\Pi}\xi \in \mathcal{F}_n$, there exists $\widehat{\alpha} \in \mathbb{R}^d$ such that $\widehat{\Pi}\xi = \sum_{i=1}^d \varphi_i \alpha_i$ (choose the one of minimal norm if there are several). Thus

$$\begin{aligned} ||\widehat{\Pi}\xi||_n^2 &= \langle \xi, \widehat{\Pi}\xi \rangle_n = \frac{1}{n}\sum_{t=1}^n \xi_t \sum_{i=1}^d \varphi_i(X_t)\widehat{\alpha}_i = \frac{1}{n}\sum_{i=1}^d \widehat{\alpha}_i \sum_{t=1}^n \xi_t \varphi_i(X_t) \\ &\leq \frac{1}{n}||\widehat{\alpha}||_2 \Big[\sum_{i=1}^d \Big(\underbrace{\sum_{t=1}^n \xi_t \varphi_i(X_t)}_{\text{martingale}} \Big)^2\Big]^{1/2}. \end{aligned}$$

Concentration for martingale: $O(\sqrt{n \log 1/\delta})$, w.p. $1-\delta$.
Now, $\widehat{\alpha}$ is orthogonal to the null-space of the Gram matrice:

$$||\widehat{\alpha}||_2^2 = \widehat{\alpha}^\top \widehat{\alpha} \leq \frac{1}{n\nu}\widehat{\alpha}^\top \Phi^\top \Phi \widehat{\alpha} = \frac{1}{\nu}||\widehat{\Pi}\xi||_n^2.$$

from which we deduce that $||\widehat{\Pi}\xi||_n = O(\sqrt{\frac{d \log d/\delta}{n\nu}})$.

Generalization bound

Recall the result in empirical norm:

$$||V - \widehat{V}||_n \leq \frac{1}{\sqrt{1-\gamma^2}} \inf_{f \in \mathcal{F}} ||V - f||_n + O\Big(\sqrt{\frac{d \log(d/\delta)}{n\nu}}\Big),$$

Now, in the case the Markov chain possesses a stationary distribution μ and is β-mixing, then we have the generalization bound: with probability $1-\delta$,

$$||\widehat{V} - V||_\mu \leq \frac{c}{\sqrt{1-\gamma^2}} \inf_{f \in \mathcal{F}} ||V - f||_\mu + O\Big(\sqrt{\frac{d \log(d/\delta)}{n\nu}}\Big),$$

expressed in terms of

- the best possible approximation of V in $\mathcal{F}$ measured with μ
- the smallest eigenvalue ν of the Gram matrix $\big(\int \varphi_i \varphi_j d\mu\big)_{i,j}$
- β-mixing coefficients of the chain (hidden in O).

Conclusions

We derived finite-sample high probability bounds for LSTD:

- Empirical bound at the states of the Markov chain, without any assumption about the chain
- Generalization bound in the case the Markov chain has a stationary distribution and is β-mixing.

Those approximation error bounds can be used to derive performance bounds for Policy Iteration (i.e. bounds on $||V^* - V^\pi||$).

Open questions:

- can we get rid of ν?
- Similar analysis for Bellman residual minimization?
- Similar analysis for off-policy LSTD?