Introduction to Reinforcement Learning and multi-armed bandits

Rémi Munos

INRIA Lille - Nord Europe
Currently on leave at MSR-NE
http://researchers.lille.inria.fr/~munos/

NETADIS Summer School 2013, Hillerod, Denmark
Part 2: Reinforcement Learning and dynamic programming with function approximation

- Approximate policy iteration
- Approximate value iteration
- Analysis of sample-based algorithms
Example: Tetris

- **State:** wall configuration + new piece
- **Action:** possible positions of the new piece on the wall,
- **Reward:** number of lines removed
- **Next state:** Resulting configuration of the wall + random new piece.

Size state space: $\approx 10^{61}$ states!
Approximate methods

When the state space is finite and small, use DP or RL techniques. However in most interesting problems, the state-space $X$ is huge, possibly infinite:

- Tetris, Backgammon, ...
- Control problems often consider continuous spaces

We need to use function approximation:

- Linear approximation $\mathcal{F} = \{ f_\alpha = \sum_{i=1}^{d} \alpha_i \phi_i, \alpha \in \mathbb{R}^d \}$
- Neural networks: $\mathcal{F} = \{ f_\alpha \}$, where $\alpha$ is the weight vector
- Non-parametric: $k$-nearest neighbors, Kernel methods, SVM, ...

Write $\mathcal{F}$ the set of representable functions.
Approximate dynamic programming

**General approach**: build an approximation $V \in \mathcal{F}$ of the optimal value function $V^*$ (which may not belong to $\mathcal{F}$), and then consider the policy $\pi$ greedy policy w.r.t. $V$, i.e.,

$$\pi(x) \in \arg\max_{a \in A} \left[ r(x, a) + \gamma \sum_y p(y|x, a) V(y) \right].$$

(for the case of *infinite horizon with discounted rewards.*)

We expect that if $V \in \mathcal{F}$ is close to $V^*$ then the policy $\pi$ will be close-to-optimal.
Bound on the performance loss

Proposition 1.

Let $V$ be an approximation of $V^*$, and write $\pi$ the policy greedy w.r.t. $V$. Then

$$\|V^* - V^\pi\|_\infty \leq \frac{2\gamma}{1 - \gamma} \|V^* - V\|_\infty.$$ 

Proof.

From the contraction properties of the operators $T$ and $T^\pi$ and that by definition of $\pi$ we have $T V = T^\pi V$, we deduce

$$\|V^* - V^\pi\|_\infty \leq \|V^* - T^\pi V\|_\infty + \|T^\pi V - T^\pi V^\pi\|_\infty$$
$$\leq \|T V^* - T V\|_\infty + \gamma \|V - V^\pi\|_\infty$$
$$\leq \gamma \|V^* - V\|_\infty + \gamma (\|V - V^*\|_\infty + \|V^* - V^\pi\|_\infty)$$
$$\leq \frac{2\gamma}{1 - \gamma} \|V^* - V\|_\infty.$$
Approximate Value Iteration

Approximate Value Iteration: builds a sequence of $V_k \in \mathcal{F}$:

$$V_{k+1} = \Pi \mathcal{T} V_k,$$

where $\Pi$ is a projection operator onto $\mathcal{F}$ (under some norm $\| \cdot \|$).

Property: the algorithm may not converge.
Apply AVI for $K$ iterations.

**Proposition 2 (Bertsekas & Tsitsiklis, 1996).**

The performance loss $\|V^* - V^{\pi_K}\|_\infty$ resulting from using the policy $\pi_K$ greedy w.r.t. $V_K$ is bounded as:

$$\|V^* - V^{\pi_K}\|_\infty \leq \frac{2\gamma}{(1 - \gamma)^2} \max_{0 \leq k < K} \|TV_k - V_{k+1}\|_\infty + \frac{2\gamma^{K+1}}{1 - \gamma} \|V^* - V_0\|_\infty.$$
Proof of Proposition 2

Write $\varepsilon = \max_{0 \leq k < K} \| T V_k - V_{k+1} \|_{\infty}$. For all $0 \leq k < K$, we have

\[
\| V^* - V_{k+1} \|_{\infty} \leq \| T V^* - T V_k \|_{\infty} + \| T V_k - V_{k+1} \|_{\infty} \\
\leq \gamma \| V^* - V_k \|_{\infty} + \varepsilon,
\]

thus, $\| V^* - V_K \|_{\infty} \leq (1 + \gamma + \cdots + \gamma^{K-1})\varepsilon + \gamma^K \| V^* - V_0 \|_{\infty}$

\[
\leq \frac{1}{1 - \gamma} \varepsilon + \gamma^K \| V^* - V_0 \|_{\infty}
\]

and we conclude by using Proposition 1.
A possible numerical implementation

Makes use of a generative model. At each round $k$,

1. Sample $n$ states $(x_i)_{1 \leq i \leq n}$
2. From each state $x_i$, for each action $a \in A$, use the model to generate a reward $r(x_i, a)$ and $m$ next-state samples $(y_{i,a}^j)_{1 \leq j \leq m} \sim p(\cdot|x_i, a)$
3. Define

$$V_{k+1} = \arg \min_{V \in \mathcal{F}} \max_{1 \leq i \leq n} \left| V(x_i) - \max_{a \in A} \left[ r(x_i, a) + \gamma \frac{1}{m} \sum_{j=1}^m V_k(y_{i,a}^j) \right] \right|$$

sample estimate of $\mathcal{T}V_k(x_i)$

This is still a numerically hard problem.
Approximate Policy Iteration

Choose an initial policy $\pi_0$ and iterate:

1. **Approximate policy evaluation** of $\pi_k$: compute an approximation $V_k$ of $V^{\pi_k}$.

2. **Policy improvement**: $\pi_{k+1}$ is greedy w.r.t. $V_k$:

$$
\pi_{k+1}(x) \in \arg \max_{a \in A} \left[ r(x, a) + \gamma \sum_{y \in X} p(y|x, a) V_k(y) \right].
$$

Property: the algorithm may not converge.
Proposition 3 (Bertsekas & Tsitsiklis, 1996).

We have

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

Thus if we are able to compute a good approximation of the value function $V^{\pi_k}$ at each iteration then the performance of the resulting policies will be good.
Proof of Proposition 3 [part 1]

Write $e_k = V_k - V^\pi_k$ the approximation error, $g_k = V^\pi_{k+1} - V^\pi_k$ the performance gain between iterations $k$ and $k+1$, and $l_k = V^* - V^\pi_k$ the loss of using policy $\pi_k$ instead of $\pi^*$. The next policy cannot be much worst that the current one:

$$g_k \geq -\gamma(I - \gamma P^{\pi_{k+1}})^{-1}(P^{\pi_{k+1}} - P^{\pi_k}) e_k \tag{1}$$

Indeed, since $T^{\pi_{k+1}} V_k \geq T^{\pi_k} V_k$ (as $\pi_{k+1}$ is greedy w.r.t. $V_k$), we have:

$$g_k = T^{\pi_{k+1}} V^{\pi_{k+1}} - T^{\pi_{k+1}} V^{\pi_k} + T^{\pi_{k+1}} V^{\pi_k} - T^{\pi_{k+1}} V_k$$
$$+ T^{\pi_{k+1}} V_k - T^{\pi_k} V_k + T^{\pi_k} V_k - T^{\pi_k} V^{\pi_k}$$
$$\geq \gamma P^{\pi_{k+1}} g_k - \gamma (P^{\pi_{k+1}} - P^{\pi_k}) e_k$$
$$\geq -\gamma(I - \gamma P^{\pi_{k+1}})^{-1}(P^{\pi_{k+1}} - P^{\pi_k}) e_k$$
Proof of Proposition 3 [part 2]

The loss at the next iteration is bounded by the current loss as:

\[ l_{k+1} \leq \gamma P^{\pi^*} l_k + \gamma \left[ P^{\pi_{k+1}} (I - \gamma P^{\pi_{k+1}})^{-1} (I - \gamma P^{\pi_k}) - P^{\pi^*} \right] e_k \]

Indeed, since \( T^{\pi^*} V_k \leq T^{\pi_{k+1}} V_k \),

\[
l_{k+1} = T^{\pi^*} V^* - T^{\pi^*} V^{\pi_k} + T^{\pi^*} V^{\pi_k} - T^{\pi^*} V_k \\
+ T^{\pi^*} V_k - T^{\pi_{k+1}} V_k + T^{\pi_{k+1}} V_k - T^{\pi_{k+1}} V^{\pi_k} \\
+ T^{\pi_{k+1}} V^{\pi_k} - T^{\pi_{k+1}} V^{\pi_{k+1}} \leq \gamma \left[ P^{\pi^*} l_k - P^{\pi_{k+1}} g_k + (P^{\pi_{k+1}} - P^{\pi^*}) e_k \right]
\]

and by using (1),

\[
l_{k+1} \leq \gamma P^{\pi^*} l_k + \gamma \left[ P^{\pi_{k+1}} (I - \gamma P^{\pi_{k+1}})^{-1} (I - \gamma P^{\pi_k}) + P^{\pi_{k+1}} - P^{\pi^*} \right] e_k \\
\leq \gamma P^{\pi^*} l_k + \gamma \left[ P^{\pi_{k+1}} (I - \gamma P^{\pi_{k+1}})^{-1} (I - \gamma P^{\pi_k}) - P^{\pi^*} \right] e_k.
\]
Proof of Proposition 3 [part 3]

Writing \( f_k = \gamma[P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1}(I - \gamma P^{\pi_k}) - P^{\pi^*}]e_k \), we have:

\[
l_{k+1} \leq \gamma P^{\pi^*} l_k + f_k.
\]

Thus, by taking the limit sup.,

\[
(I - \gamma P^{\pi^*}) \limsup_{k \to \infty} l_k \leq \limsup_{k \to \infty} f_k
\]

\[
\limsup_{k \to \infty} l_k \leq (I - \gamma P^{\pi^*})^{-1} \limsup_{k \to \infty} f_k,
\]

since \( I - \gamma P^{\pi^*} \) is invertible. In \( L_\infty \)-norm, we have

\[
\limsup_{k \to \infty} \|l_k\| \leq \frac{\gamma}{1 - \gamma} \limsup_{k \to \infty} \|P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1}(I + \gamma P^{\pi_k}) + P^{\pi^*}\| \|e_k\|
\]

\[
\leq \frac{\gamma}{1 - \gamma} \left(\frac{1 + \gamma}{1 - \gamma} + 1\right) \limsup_{k \to \infty} \|e_k\| = \frac{2\gamma}{(1 - \gamma)^2} \limsup_{k \to \infty} \|e_k\|.
\]
Case study: TD-Gammon [Tesauro, 1994]

State = game configuration $x + \text{player } j \rightarrow N \approx 10^{20}$.

Reward 1 or 0 at the end of the game.

The neural network returns an approximation of $V^*(x, j)$: probability that player $j$ wins from position $x$, assuming that both players play optimally.
Approximate Value Iteration

Approximate Policy Iteration

Analysis of sample-based algo

TD-Gammon algorithm

- At time $t$, the current game configuration is $x_t$
- Roll dices and select the action that maximizes the value $V_\alpha$ of the resulting state $x_{t+1}$
- Set the temporal difference $d_t = V_\alpha(x_{t+1},j_{t+1}) - V_\alpha(x_t,j_t)$ (if this is a final position, replace $V_\alpha(x_{t+1},j_{t+1})$ by $+1$ or $0$)
- Update $\alpha_t$ according to a gradient descent

$$
\alpha_{t+1} = \alpha_t + \eta_t d_t \sum_{0 \leq s \leq t} \lambda^{t-s} \nabla_\alpha V_\alpha(x_s).
$$

After several weeks of self playing → world best player.
According to human experts it developed new strategies, specially in openings.
Least Squares Temporal Difference (LSTD)

[Bradtke & Barto, 1996] Consider a linear space $\mathcal{F}$. Let $\Pi_\mu$ be the projection onto $\mathcal{F}$ defined by a weighted norm $L_2(\mu)$. The **Least Squares Temporal Difference** solution $V_{TD}$ is the fixed-point of $\Pi_\mu T^\pi$. 

\[
V_{TD} = \Pi_\mu T^\pi V_{TD}
\]
Performance bound for LSTD

In general, no guarantee that there exists a fixed-point to $\Pi_\mu T^\pi$ (since $T^\pi$ is not a contraction in $L_2(\mu)$-norm).
However, when $\mu$ is the stationary distribution associated to $\pi$ (i.e., such that $\mu P^\pi = \mu$), then there exists a unique LSTD solution.

**Proposition 4.**
Consider $\mu$ to be the stationary distribution associated to $\pi$. Then $T^\pi$ is a contraction mapping in $L_2(\mu)$-norm, thus $\Pi_\mu T^\pi$ is also a contraction, and there exists a unique LSTD solution $V_{TD}$. In addition, we have the approximation error:

$$
\| V^\pi - V_{TD} \|_\mu \leq \frac{1}{\sqrt{1 - \gamma^2}} \inf_{V \in \mathcal{F}} \| V^\pi - V \|_\mu. \quad (2)
$$
Proof of Proposition 4 [part 1]

First let us prove that \( \| P_\pi \|_\mu = 1 \). We have:

\[
\| P_\pi V \|_\mu^2 = \sum_x \mu(x) \left( \sum_y p(y|x, \pi(x)) V(y) \right)^2 \\
\leq \sum_x \sum_y \mu(x) p(y|x, \pi(x)) V(y)^2 \\
= \sum_y \mu(y) V(y)^2 = \| V \|_\mu^2.
\]

We deduce that \( T_\pi \) is a contraction mapping in \( L_2(\mu) \):

\[
\| T_\pi V_1 - T_\pi V_2 \|_\mu = \gamma \| P_\pi (V_1 - V_2) \|_\mu \leq \gamma \| V_1 - V_2 \|_\mu,
\]

and since \( \Pi_\mu \) is a non-expansion in \( L_2(\mu) \), then \( \Pi_\mu T_\pi \) is a contraction in \( L_2(\mu) \). Write \( V_{TD} \) its (unique) fixed-point.
Proof of Proposition 4 [part 2]

We have \( \| V^\pi - V_{TD} \|_\mu^2 = \| V^\pi - \Pi_\mu V^\pi \|_\mu^2 + \| \Pi_\mu V^\pi - V_{TD} \|_\mu^2 \),

but \( \| \Pi_\mu V^\pi - V_{TD} \|_\mu^2 = \| \Pi_\mu V^\pi - \Pi_\mu \mathcal{T}^\pi V_{TD} \|_\mu^2 \)
\[ \leq \| \mathcal{T}^\pi V^\pi - \mathcal{T} V_{TD} \|_\mu^2 \leq \gamma^2 \| V^\pi - V_{TD} \|_\mu^2. \]

Thus \( \| V^\pi - V_{TD} \|_\mu^2 \leq \| V^\pi - \Pi_\mu V^\pi \|_\mu^2 + \gamma^2 \| V^\pi - V_{TD} \|_\mu^2 \),
from which the result follows.
Characterization of the LSTD solution

The Bellman residual $\mathcal{T}^\pi V_{TD} - V_{TD}$ is orthogonal to the space $\mathcal{F}$, thus for all $1 \leq i \leq d$,

$$\langle r^\pi + \gamma P^\pi V_{TD} - V_{TD}, \phi_i \rangle_{\mu} = 0$$

$$\langle r^\pi, \phi_i \rangle_{\mu} + \sum_{j=1}^{d} \langle \gamma P^\pi \phi_j - \phi_j, \phi_i \rangle_{\mu} \alpha_{TD,j} = 0,$$

where $\alpha_{TD}$ is the parameter of $V_{TD}$. We deduce that $\alpha_{TD}$ is solution to the linear system (of size $d$):

$$A\alpha = b, \text{ with } \begin{cases} A_{i,j} & = \langle \phi_i, \phi_j - \gamma P^\pi \phi_j \rangle_{\mu} \\ b_i & = \langle \phi_i, r^\pi \rangle_{\mu} \end{cases}$$
Empirical LSTD

Consider a trajectory \((x_1, x_2, \ldots, x_n)\) generated by following \(\pi\)
Build the matrix \(\hat{A}\) and the vector \(\hat{b}\) as

\[
\hat{A}_{ij} = \frac{1}{n} \sum_{t=1}^{n} \phi_i(x_t) [\phi_j(x_t) - \gamma \phi_j(x_{t+1})],
\]

\[
\hat{b}_i = \frac{1}{n} \sum_{t=1}^{n} \phi_i(x_t) r_{x_t}.
\]

and compute the empirical LSTD solution \(\hat{V}_{TD}\) whose parameter is
the solution to \(\hat{A}\alpha = \hat{b}\).

We have \(\hat{V}_{TD} \xrightarrow{a.s.} V_{TD}\) when \(n \to \infty\), since \(\hat{A}\xrightarrow{a.s.} A\) and \(\hat{b}\xrightarrow{a.s.} b\).
Finite-time analysis of LSTD

Define the empirical norm $\|f\|_n = \sqrt{\frac{1}{n}\sum_{t=1}^{n} f(x_t)^2}$.

**Theorem 1 (Lazaric et al., 2010).**

With probability $1 - \delta$ (w.r.t. the trajectory),

$$\|V^\pi - \hat{V}_{TD}\|_n \leq \frac{1}{\sqrt{1 - \gamma^2}} \inf_{V \in \mathcal{V}} \|V^\pi - V\|_n + \frac{c}{1 - \gamma} \sqrt{\frac{d \log(1/\delta)}{n}}$$

- Approximation error
- Estimation error

This type of bounds is similar to results in Statistical Learning.
Least-Squares Policy Iteration

[Lagoudakis & Parr, 2003] Consider $Q(x, a) = \sum_{i=1}^{d} \alpha_i \phi_i(x, a)$

- **Policy evaluation:** At round $k$, run a trajectory $(x_t)_{1 \leq t \leq n}$ by following policy $\pi_k$. Build $\hat{A}$ and $\hat{b}$ as

\[
\hat{A}_{ij} = \frac{1}{n} \sum_{t=1}^{n} \phi_i(x_t, a_t)[\phi_j(x_t, a_t) - \gamma \phi_j(x_{t+1}, a_{t+1})],
\]

\[
\hat{b}_i = \frac{1}{n} \sum_{t=1}^{n} \phi_i(x_t, a_t) r(x_t, a_t).
\]

and $\hat{Q}_k$ is the Q-function defined by the solution to $\hat{A} \alpha = \hat{b}$.

- **Policy improvement:** $\pi_{k+1}(x) \in \arg \max_{a \in A} \hat{Q}_k(x, a)$.

We would like guarantees on $\|Q^* - Q^{\pi_k}\|$.
Theoretical guarantees so far

Approximate Value Iteration:

\[ \| V^* - V^{\pi_K} \|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \max_{0 \leq k < K} \| TV_k - V_{k+1} \|_\infty + O(\gamma^K). \]

\( \{z\} \) projection error

Approximate Policy Iteration:

\[ \| V^* - V^{\pi_K} \|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \max_{0 \leq k < K} \| V^{\pi_k} - V_k \|_\infty + O(\gamma^K). \]

\( \{z\} \) approximation error

Problem: hard to control \( L_\infty \)-norm using samples. We could minimize an empirical \( L_\infty \)-norm, but

- Numerically intractable
- Hard to relate \( L_\infty \)-norm to empirical \( L_\infty \)-norm.
Instead use empirical $L^2$-norm

- For AVI this is just a linear regression problem:

$$V_{k+1} = \arg \min_{V \in \mathcal{F}} \sum_{i=1}^{n} |\mathcal{T} V_k(x_i) - V(x_i)|^2,$$

- For API this is just LSTD: fixed-point of an empirical Bellman operator projected onto $\mathcal{F}$ using an empirical norm.

In both cases, $V_k$ is solution to a linear problem, which is

- Numerically tractable
- For which generalization bounds exits (using VC theory):

$$\|\mathcal{T} V_k - V_{k+1}\|_2^2 \leq \frac{1}{n} \sum_{i=1}^{n} |\mathcal{T} V_k(x_i) - V(x_i)|^2 + c \sqrt{\frac{VC(\mathcal{F})}{n}}$$
Approximate Value Iteration

Approximate Policy Iteration

Analysis of sample-based algo

$L_p$-norm analysis of ADP

Under smoothness assumptions on the MDP, the propagation error of all usual ADP algorithms can be analyzed in $L_p$-norm ($p \geq 1$).

**Proposition 5 (Munos, 2003, 2007).**

- **Approximate Value Iteration:** Assume there is a constant $C \geq 1$ and a distribution $\mu$ such that $\forall x \in X, \forall a \in A,$
  
  $$p(\cdot|x, a) \leq C\mu(\cdot).$$

  $$\|V^* - V^{\pi_K}\|_\infty \leq \frac{2\gamma}{(1 - \gamma)^2} C^{1/p} \max_{0 \leq k < K} \|TV_k - V_{k+1}\|_{p,\mu} + O(\gamma^K).$$

- **Approximate Policy Iteration:** Assume $p(\cdot|x, a) \leq C\mu_{\pi}(\cdot)$, for any policy $\pi$

  $$\|V^* - V^{\pi_K}\|_\infty \leq \frac{2\gamma}{(1 - \gamma)^2} C^{1/p} \max_{0 \leq k < K} \|V_k - V^{\pi_k}\|_{p,\mu_{\pi}} + O(\gamma^K).$$

We have all ingredients for a finite-sample analysis of RL/ADP.
Finite-sample analysis of LSPI

Perform $K$ policy iterations steps. At stage $k$, run one trajectory of length $n$ following $\pi_k$ and compute the LSTD solution $\hat{V}_k$ (by solving a linear system).

**Proposition 6 (Lazaric et al., 2010).**

For any $\delta > 0$, with probability at least $1 - \delta$, we have:

$$\|V^* - V^{\pi_K}\|_\infty \leq \frac{2\gamma}{(1 - \gamma)^3} C^{1/2} \sup_{k} \inf_{\mathcal{F}} \|V^{\pi_k} - V\|_{2,\mu_k} + O\left(\frac{d \log(1/\delta)}{n}\right)^{1/2} + O(\gamma^K)$$
Finite-sample analysis of AVI

$K$ iterations of AVI with $n$ samples $x_i \sim \mu$. From each state $x_i$, each $a \in A$, generate $m$ next state samples $y_{i,a}^j \sim p(\cdot|x_i, a)$.

**Proposition 7 (Munos and Szepesvári, 2007).**

For any $\delta > 0$, with probability at least $1 - \delta$, we have:

$$\|V^* - V^{\pi,K}\|_\infty \leq \frac{2\gamma}{(1 - \gamma)^2} C^{1/p} d(\mathcal{TF}, \mathcal{F}) + O(\gamma^K)$$

$$+ O\left(\frac{V(\mathcal{F}) \log(1/\delta)}{n}\right)^{1/4} + O\left(\frac{\log(1/\delta)}{m}\right)^{1/2},$$

where $d(\mathcal{TF}, \mathcal{F}) \overset{\text{def}}{=} \sup_{g \in \mathcal{F}} \inf_{f \in \mathcal{F}} \|Tg - f\|_{2,\mu}$ is the Bellman residual of the space $\mathcal{F}$, and $V(\mathcal{F})$ the pseudo-dimension of $\mathcal{F}$. 
More works on finite-sample analysis of ADP/RL

This is important to know how many samples $n$ are required to build an $\epsilon$-approximation of the optimal policy.

- Policy iteration using a single trajectory [Antos et al., 2008]
- BRM [Maillard et al., 2010]
- LSTD with random projections [Ghavamzadeh et al., 2010]
- Lasso-TD [Ghavamzadeh et al., 2011]

Active research topic which links RL and statistical learning theory.