Introduction to Reinforcement Learning
Part 2: Approximate Dynamic Programming

Rémi Munos

SequeL project: Sequential Learning
http://researchers.lille.inria.fr/∼munos/

INRIA Lille - Nord Europe

Machine Learning Summer School, September 2011, Bordeaux

Outline of Part 2: Approximate dynamic programming

- Function approximation
- Bellman residual minimization
- Approximate value iteration: fitted VI
- Approximate policy iteration, LSTD, BRM
- Analysis of sample-based algorithms

References

General references on Approximate Dynamic Programming:

- *Neuro Dynamic Programming*, Bertsekas et Tsitsiklis, 1996.
- *Markov Decision Processes in Artificial Intelligence*, Sigaud and Buffet ed., 2008.
- *Algorithms for Reinforcement Learning*, Szepesvári, 2009.

BRM, TD, LSTD/LSPI:

- BRM [Williams and Baird, 1993]
- TD learning [Tsitsiklis and Van Roy, 1996]
- LSTD [Bradtke and Barto, 1993], [Boyan, 1999], LSPI [Lagoudakis and Parr, 2003], [Munos, 2003]

Finite-sample analysis:

- AVI [Munos and Szepesvári, 2008]
- API [Antos et al., 2009]
- LSTD [Lazaric et al., 2010]

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 $\mathcal{F} = \{f_\alpha = \sum_{i=1}^{d} \alpha_i \phi_i, \alpha \in \boldsymbol{R}^d\}$
- Neural networks: $\mathcal{F} = \{f_\alpha\}$, where α is the weight vector
- Non-parametric: k-nearest neighboors, 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 π 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 π will be close-to-optimal.

Bound on the performance loss

Proposition 1.

Let V be an approximation of V^*, and write π 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 $\mathcal{T}$ and $\mathcal{T}^\pi$ and that by definition of π we have $\mathcal{T}V = \mathcal{T}^\pi V$, we deduce

$$\begin{aligned}
\|V^* - V^\pi\|_\infty &\leq \|V^* - \mathcal{T}^\pi V\|_\infty + \|\mathcal{T}^\pi V - \mathcal{T}^\pi V^\pi\|_\infty \\
&\leq \|\mathcal{T}V^* - \mathcal{T}V\|_\infty + \gamma\|V - V^\pi\|_\infty \\
&\leq \gamma\|V^* - V\|_\infty + \gamma(\|V - V^*\|_\infty + \|V^* - V^\pi\|_\infty) \\
&\leq \frac{2\gamma}{1-\gamma}\|V^* - V\|_\infty.
\end{aligned}$$

Bellman residual

- Let us define the **Bellman residual** of a function V as the function $\mathcal{T}V - V$.
- Note that the Bellman residual of V^* is 0 (Bellman equation).
- If a function V has a low $||\mathcal{T}V - V||_\infty$, then is V close to V^*?

Proposition 2 (Williams and Baird, 1993).

We have

$$\begin{aligned} \|V^* - V\|_\infty &\leq \frac{1}{1-\gamma}\|\mathcal{T}V - V\|_\infty \\ \|V^* - V^\pi\|_\infty &\leq \frac{2}{1-\gamma}\|\mathcal{T}V - V\|_\infty \end{aligned}$$

Proof of Proposition 2

Point 1: we have

$$\begin{aligned} \|V^* - V\|_\infty &\leq \|V^* - \mathcal{T}V\|_\infty + \|\mathcal{T}V - V\|_\infty \\ &\leq \gamma\|V^* - V\|_\infty + \|\mathcal{T}V - V\|_\infty \\ &\leq \frac{1}{1-\gamma}\|\mathcal{T}V - V\|_\infty \end{aligned}$$

Point 2: We have $\|V^* - V^\pi\|_\infty \leq \|V^* - V\|_\infty + \|V - V^\pi\|_\infty$.
Since $\mathcal{T}V = \mathcal{T}^\pi V$, we deduce

$$\begin{aligned} \|V - V^\pi\|_\infty &\leq \|V - \mathcal{T}V\|_\infty + \|\mathcal{T}V - V^\pi\|_\infty \\ &\leq \|\mathcal{T}V - V\|_\infty + \gamma\|V - V^\pi\|_\infty \\ &\leq \frac{1}{1-\gamma}\|\mathcal{T}V - V\|_\infty, \end{aligned}$$

thus, by using Point 1, it comes

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

Bellman residual minimizer

Given a function space $\mathcal{F}$ we can search for the function with minimum Bellman residual:

$$V_{BR} = \arg\min_{V \in \mathcal{F}} \|\mathcal{T}V - V\|_\infty.$$

What is the performance of the policy π_{BR} greedy w.r.t. V_{BR}?

Proposition 3.

We have:

$$\|V^* - V^{\pi_{BR}}\|_\infty \leq \frac{2(1+\gamma)}{1-\gamma} \inf_{V \in \mathcal{F}} \|V^* - V\|_\infty. \qquad (1)$$

Thus minimizing the Bellman residual in $\mathcal{F}$ is a sound approach whenever $\mathcal{F}$ is rich enough.

Proof of Proposition 3

We have

$$\begin{aligned} \|\mathcal{T}V - V\|_\infty &\leq \|\mathcal{T}V - \mathcal{T}V^*\|_\infty + \|V^* - V\|_\infty \\ &\leq (1+\gamma)\|V^* - V\|_\infty. \end{aligned}$$

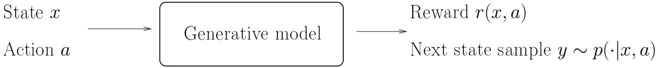
Thus V_{BR} satisfies:

$$\begin{aligned} \|\mathcal{T}V_{BR} - V_{BR}\|_\infty &= \inf_{V \in \mathcal{F}} \|\mathcal{T}V - V\|_\infty \\ &\leq (1+\gamma) \inf_{V \in \mathcal{F}} \|V^* - V\|_\infty. \end{aligned}$$

Combining with the result of Proposition 2, we deduce (1).

Possible numerical implementation

Assume that we possess a generative model:



- Sample n states $(x_i)_{1\le i\le n}$ uniformly over the state space X,
- For each action $a \in A$, generate a reward sample $r(x,a)$ and m next state samples $(y^j_{i,a})_{1\le j\le m}$.
- Return the empirical Bellman residual minimizer:

$$\widehat{V}_{BR} = \arg\min_{V\in\mathcal{F}} \max_{1\le i\le n} \Big| \underbrace{\max_{a\in A} \Big[r(x_i,a) + \gamma \frac{1}{m}\sum_{j=1}^{m} V(y^j_{i,a}) \Big]}_{\text{sample estimate of } \mathcal{T}V(x_i)} - V(x_i) \Big|.$$

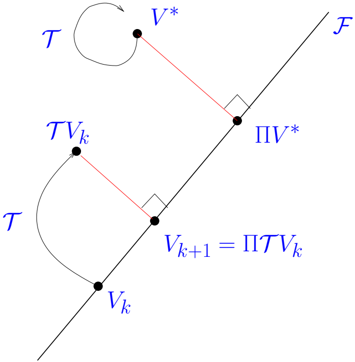
This problem is numerically hard to solve...

Approximate Value Iteration

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

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

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



Remark: Π is a non-expansion under $\|\cdot\|$, and $\mathcal{T}$ is a contraction under $\|\cdot\|_\infty$. Thus if we use $\|\cdot\|_\infty$ for Π, then AVI converges. If we use another norm for Π (e.g., L_2), then AVI may not converge.

Performance bound for AVI

Apply AVI for K iterations.

Proposition 4 (Bertsekas & Tsitsiklis, 1996).

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

$$\|V^* - V^{\pi_K}\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \max_{0 \leq k < K} \underbrace{\|\mathcal{T} V_k - V_{k+1}\|_\infty}_{\text{projection error}} + \frac{2\gamma^{K+1}}{1-\gamma} \|V^* - V_0\|_\infty.$$

Now if we use $\|\cdot\|_\infty$-norm for Π, then AVI converges, say to $\widetilde{V}$ which is such that $\widetilde{V} = \Pi \mathcal{T} \widetilde{V}$. Write $\tilde{\pi}$ the policy greedy w.r.t. $\widetilde{V}$. Then

$$\|V^* - V^{\tilde{\pi}}\|_\infty \leq \frac{2}{(1-\gamma)^2} \inf_{V \in \mathcal{F}} \|V^* - V\|_\infty$$

Proof of Proposition 4

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

$$
\begin{aligned}
\|V^* - V_{k+1}\|_\infty &\le \|\mathcal{T} V^* - \mathcal{T} V_k\|_\infty + \|\mathcal{T} V_k - V_{k+1}\|_\infty \\
&\le \gamma \|V^* - V_k\|_\infty + \varepsilon,
\end{aligned}
$$

$$
\begin{aligned}
\text{thus, } \|V^* - V_K\|_\infty &\le (1 + \gamma + \cdots + \gamma^{K-1})\varepsilon + \gamma^K \|V^* - V_0\|_\infty \\
&\le \frac{1}{1-\gamma}\varepsilon + \gamma^K \|V^* - V_0\|_\infty
\end{aligned}
$$

and we conclude by using Proposition 1.

Point 2: If Π uses $\|\cdot\|_\infty$ then $\Pi\mathcal{T}$ is a γ-contraction mapping, thus AVI converges, say to $\widetilde{V}$ satisfying $\widetilde{V} = \Pi\mathcal{T}\widetilde{V}$. And

$$
\|V^* - \widetilde{V}\|_\infty \le \|V^* - \Pi V^*\|_\infty + \|\Pi V^* - \widetilde{V}\|_\infty
$$

$$
\text{with } \|\Pi V^* - \widetilde{V}\|_\infty = \|\Pi\mathcal{T} V^* - \Pi\mathcal{T}\widetilde{V}\|_\infty \le \gamma \|V^* - \widetilde{V}\|_\infty,
$$

and the result follows from Proposition 1.

A possible numerical implementation

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 generative model to obtain a reward $r(x_i,a)$ and m next state samples $(y^j_{i,a})_{1\leq j\leq m}\sim p(\cdot|x_i,a)$
3. Define the next approximation (say using L_∞-norm)

$$V_{k+1}=\arg\min_{V\in\mathcal{F}}\max_{1\leq i\leq n}\Big|V(x_i)-\underbrace{\max_{a\in A}\Big[r(x_i,a)+\gamma\frac{1}{m}\sum_{j=1}^m V_k(y^j_{i,a})\Big]}_{\text{sample estimate of } \mathcal{T}V_k(x_i)}\Big|$$

This is still a numerically hard problem. However, using L_2 norm:

$$V_{k+1}=\arg\min_{V\in\mathcal{F}}\sum_{i=1}^n\Big|V(x_i)-\max_{a\in A}\Big[r(x_i,a)+\gamma\frac{1}{m}\sum_{j=1}^m V_k(y^j_{i,a})\Big]\Big|^2$$

is much easier!

Example: optimal replacement problem

1d-state: accumulated utilization of a product (ex. car).
Decisions: each year,

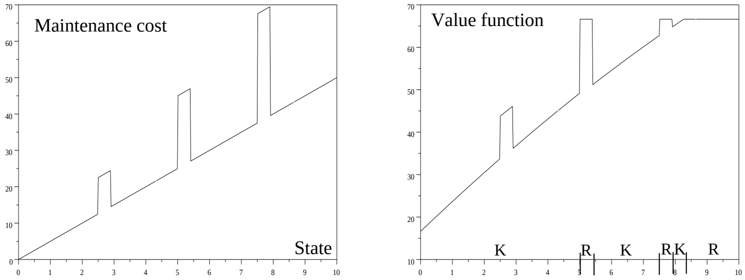
- **Replace**: replacement cost C, next state $y \sim d(\cdot)$,
- **Keep**: maintenance cost $c(x)$, next state $y \sim d(\cdot - x)$.

Goal: Minimize the expected sum of discounted costs.
The optimal value function solves the Bellman equation:

$$V^*(x) = \min\left\{c(x) + \gamma \int_0^\infty d(y-x)V^*(y)dy,\ C + \gamma \int_0^\infty d(y)V^*(y)dy\right\}$$

and the optimal policy is the argument of the min.

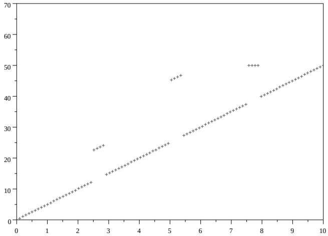
Maintenance cost and value function



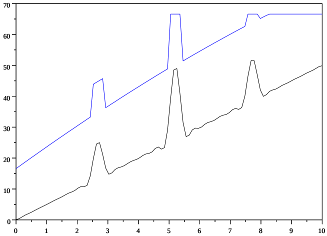
Here, $\gamma = 0.6$, $C = 50$, $d(y) = \beta e^{-\beta y} \mathbf{1}_{y \geq 0}$, with $\beta = 0.6$.
Maintenance costs = increasing function + punctual costs.

Linear approximation

Function space $\mathcal{F} = \left\{ f_\alpha(x) = \sum_{i=1}^{20} \alpha_i \cos(i\pi \frac{x}{x_{\max}}), \alpha \in \boldsymbol{R}^{20} \right\}$.
Consider a uniform discretization grid with $n = 100$ states,
$m = 100$ next-states.
First iteration: $V_0 = 0$,

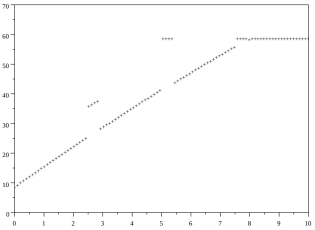


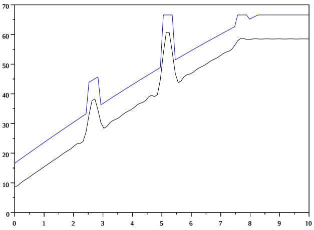
Bellman values $\{\widehat{\mathcal{T}V}_0(x_i)\}_{1\le i\le n}$

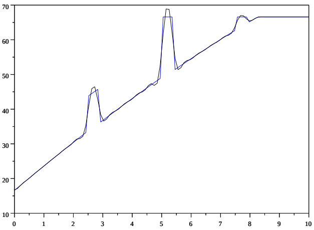


Approximation $V_1 \in \mathcal{F}$ of $\widehat{\mathcal{T}V}_0$

Next iterations





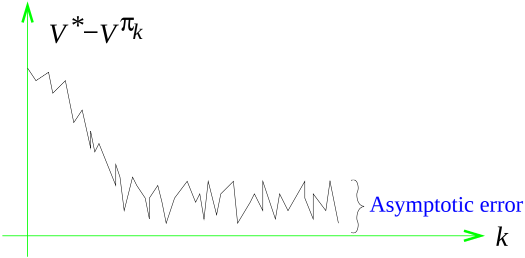


Approximate Policy Iteration

Choose an initial policy π_0 and iterate:

1. **Approximate policy evaluation** of π_k:
   compute an approximation V_k of V^{π_k}.
2. **Policy improvement**: π_{k+1} is greedy w.r.t. V_k:

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



The algorithm may not converge but we can analyze the asymptotic performance.

Performance bound for API

We relate the asymptotic performance $||V^* - V^{\pi_k}||_\infty$ of the policies π_k greedy w.r.t. the iterates V_k, in terms of the approximation errors $||V_k - V^{\pi_k}||_\infty$.

Proposition 5 (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 well approximate the value functions V^{π_k} at each iteration then the performance of the resulting policies will be close to the optimum.

Proof of Proposition 5 [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 π_k instead of π^*.
The next policy cannot be much worst that the current one:

$$g_k \geq -\gamma(I - \gamma P^{\pi_{k+1}})^{-1}(P^{\pi_{k+1}} - P^{\pi_k})\, e_k \qquad (2)$$

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

$$\begin{aligned} g_k &= T^{\pi_{k+1}} V^{\pi_{k+1}} - T^{\pi_{k+1}} V^{\pi_k} + T^{\pi_{k+1}} V^{\pi_k} - T^{\pi_{k+1}} V_k \\ &\quad + T^{\pi_{k+1}} V_k - T^{\pi_k} V_k + T^{\pi_k} V_k - T^{\pi_k} V^{\pi_k} \\ &\geq \gamma P^{\pi_{k+1}} g_k - \gamma (P^{\pi_{k+1}} - P^{\pi_k})\, e_k \\ &\geq -\gamma(I - \gamma P^{\pi_{k+1}})^{-1}(P^{\pi_{k+1}} - P^{\pi_k})\, e_k \end{aligned}$$

Proof of Proposition 5 [part 2]

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

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

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

$$\begin{aligned} l_{k+1} &= T^{\pi^*} V^* - T^{\pi^*} V^{\pi_k} + T^{\pi^*} V^{\pi_k} - T^{\pi^*} V_k \\ &\quad + T^{\pi^*} V_k - T^{\pi_{k+1}} V_k + T^{\pi_{k+1}} V_k - T^{\pi_{k+1}} V^{\pi_k} \\ &\quad + T^{\pi_{k+1}} V^{\pi_k} - T^{\pi_{k+1}} V^{\pi_{k+1}} \\ &\leq \gamma[P^{\pi^*} l_k - P^{\pi_{k+1}} g_k + (P^{\pi_{k+1}} - P^{\pi^*})e_k] \end{aligned}$$

and by using (2),

$$\begin{aligned} l_{k+1} &\leq \gamma P^{\pi^*} l_k + \gamma[P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1}(P^{\pi_{k+1}} - P^{\pi_k}) + P^{\pi_{k+1}} - P^{\pi^*}]e_k \\ &\leq \gamma P^{\pi^*} l_k + \gamma[P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1}(I - \gamma P^{\pi_k}) - P^{\pi^*}]e_k. \end{aligned}$$

Proof of Proposition 5 [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.,

$$\begin{aligned}(I - \gamma P^{\pi^*}) \limsup_{k\to\infty} l_k &\leq \limsup_{k\to\infty} f_k \\ \limsup_{k\to\infty} l_k &\leq (I - \gamma P^{\pi^*})^{-1} \limsup_{k\to\infty} f_k,\end{aligned}$$

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

$$\begin{aligned}\limsup_{k\to\infty} ||l_k|| &\leq \frac{\gamma}{1-\gamma} \limsup_{k\to\infty} ||P^{\pi_{k+1}}(I - \gamma P^{\pi_{k+1}})^{-1}(I + \gamma P^{\pi_k}) + P^{\pi^*}|| \, ||e_k|| \\ &\leq \frac{\gamma}{1-\gamma}(\frac{1+\gamma}{1-\gamma} + 1) \limsup_{k\to\infty} ||e_k|| = \frac{2\gamma}{(1-\gamma)^2} \limsup_{k\to\infty} ||e_k||.\end{aligned}$$

Approximate policy evaluation

For a given policy π we search for an approximation $V_\alpha \in \mathcal{F}$ of V^π. For example, by minimizing the approximation error

$$\inf_{V_\alpha \in \mathcal{F}} ||V_\alpha - V^\pi||_2^2.$$

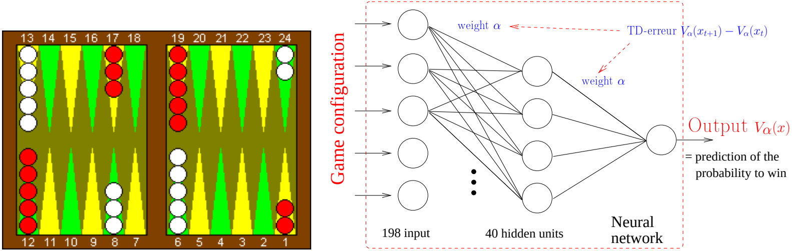
Writing $g(\alpha) = \frac{1}{2}\|V_\alpha - V^\pi\|_2^2$, we may consider a stochastic gradient algorithm:

$$\alpha \leftarrow \alpha - \eta \widehat{\nabla g}(\alpha)$$

where an estimate $\widehat{\nabla g}(\alpha) = \langle \nabla V_\alpha, V_\alpha - \sum_{t\geq 0} \gamma^t r_t \rangle$ of the gradient $\nabla g(\alpha) = \langle \nabla V_\alpha, V_\alpha - V^\pi \rangle$ may be obtained by using MC sampling of trajectories (x_t) following π.
Extension to **TD(λ)** algorithms have been introduced:

$$\alpha \leftarrow \alpha + \eta \sum_{s\geq 0} \nabla_\alpha V_\alpha(x_s) \sum_{t\geq s} (\gamma\lambda)^{t-s} d_t.$$

TD-Gammon [Tesauro, 1994]



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

TD-Gammon algorithm

- At time t, the current game configuration is x_t
- Roll dices and select the action that maximizes the value V_α of the resulting state x_{t+1}
- Compute 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 α_t according to

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

This is a variant of API using TD(λ) where there is a policy improvement step after each update of the parameter.
After several weeks of self playing $\rightarrow$ **world best player.**
According to human experts it developed new strategies, specially in openings.

TD(λ) with linear space

Consider a set of features $(\phi_i : X \to \boldsymbol{R})_{1\leq i\leq d}$ and the linear space

$$\mathcal{F} = \{V_\alpha(x) = \sum_{i=1}^{d} \alpha_i \phi_i(x), \alpha \in \boldsymbol{R}^d\}.$$

Run a trajectory (x_t) by following policy π.
After the transition $x_t \xrightarrow{r_t} x_{t+1}$, compute the temporal difference $d_t = r_t + \gamma V_\alpha(x_{t+1}) - V_\alpha(x_t)$, and update

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

Proposition 6 (Tsitsiklis & Van Roy, 1996).

Assume that $\sum \eta_t = \infty$ and $\sum \eta_t^2 < \infty$, and there exists $\mu \in \boldsymbol{R}^N$ such that $\forall x, y \in X$, $\lim_{t\to\infty} \mathbb{P}(x_t = y | x_0 = x) = \mu(y)$. Then α_t converges, say to α^*. And we have

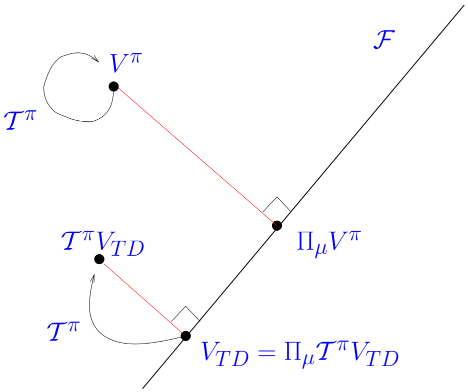
$$||V_{\alpha^*} - V^\pi||_\mu \leq \frac{1-\lambda\gamma}{1-\gamma} \inf_\alpha ||V_\alpha - V^\pi||_\mu.$$

Least Squares Temporal Difference

[Bradtke & Barto, 1996, Lagoudakis & Parr, 2003]

Consider a linear space $\mathcal{F}$ and Π_μ the projection with norm $L_2(\mu)$, where μ is a distribution over X.

When the fixed-point of $\Pi_\mu T^\pi$ exists, we call it **Least Squares Temporal Difference** solution V_{TD}.



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 $1 \leq i \leq d$,

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

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

$$A\alpha = b, \text{ with} \left\{ \begin{array}{rcl} A_{i,j} &=& \langle \phi_i, \phi_j - \gamma P^\pi \phi_j \rangle_\mu \\ b_i &=& \langle \phi_i, r^\pi \rangle_\mu \end{array} \right.$$

Performance bound for LSTD

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

Proposition 7.

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

$$\|V^\pi - V_{TD}\|_\mu \le \frac{1}{\sqrt{1-\gamma^2}} \inf_{V \in \mathcal{F}} \|V^\pi - V\|_\mu. \quad (3)$$

Proof of Proposition 7 [part 1]

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

$$\begin{aligned}
\|P^\pi V\|_\mu^2 &= \sum_x \mu(x)\Big(\sum_y p(y|x,\pi(x))V(y)\Big)^2 \\
&\leq \sum_x \sum_y \mu(x) p(y|x,\pi(x)) V(y)^2 \\
&= \sum_y \mu(y) V(y)^2 = \|V\|_\mu^2.
\end{aligned}$$

We deduce that $\mathcal{T}^\pi$ is a contraction mapping in $L_2(\mu)$:

$$\|\mathcal{T}^\pi V_1 - \mathcal{T}^\pi V_2\|_\mu = \gamma \|P^\pi(V_1 - V_2)\|_\mu \leq \gamma \|V_1 - V_2\|_\mu,$$

and since Π_μ is a non-expansion in $L_2(\mu)$, then $\Pi_\mu \mathcal{T}^\pi$ is a contraction in $L_2(\mu)$. Write V_{TD} its (unique) fixed-point.

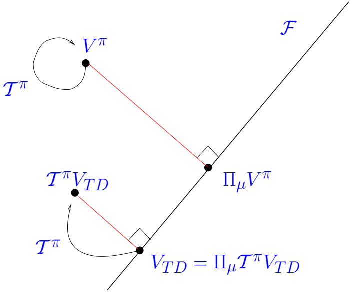
Proof of Proposition 7 [part 2]

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

$$\begin{aligned}\text{but } \|\Pi_\mu V^\pi - V_{TD}\|_\mu^2 &= \|\Pi_\mu V^\pi - \Pi_\mu \mathcal{T}^\pi V_{TD}\|_\mu^2 \\ &\leq \|\mathcal{T}^\pi V^\pi - \mathcal{T} V_{TD}\|_\mu^2 \leq \gamma^2 \|V^\pi - V_{TD}\|_\mu^2.\end{aligned}$$

$$\text{Thus } \|V^\pi - V_{TD}\|_\mu^2 \leq \|V^\pi - \Pi_\mu V^\pi\|_\mu^2 + \gamma^2 \|V^\pi - V_{TD}\|_\mu^2,$$

from which the result follows.

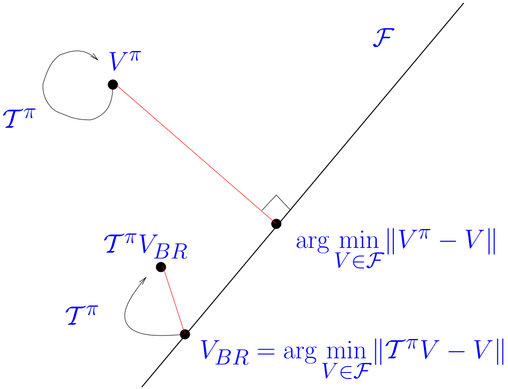


Bellman Residual Minimization (BRM)

Another approach consists in searching for the function $\mathcal{F}$ that minimizes the Bellman residual for the policy π:

$$V_{BR} = \arg\min_{V \in \mathcal{F}} \| T^\pi V - V \|, \tag{4}$$

for some norm $\| \cdot \|$.



Characterization of the BRM solution

Let μ be a distribution and V_{BR} be the BRM using $L_2(\mu)$-norm. The mapping $\alpha \to \|\mathcal{T}^\pi V_\alpha - V_\alpha\|_\mu^2$ is quadratic and its minimum is characterized by its gradient $= 0$: for all $1 \le i \le d$,

$$\begin{aligned}
\langle r^\pi + \gamma P^\pi V_\alpha - V_\alpha, \gamma P^\pi \phi_i - \phi_i \rangle_\mu &= 0 \\
\langle r^\pi + (\gamma P^\pi - I) \sum_{j=1}^d \phi_j \alpha_j, (\gamma P^\pi - I)\phi_i \rangle_\mu &= 0
\end{aligned}$$

We deduce that α_{BR} is solution to the linear system (of size d):

$$A\alpha = b, \text{ with } \left\{ \begin{array}{lcl} A_{i,j} &=& \langle \phi_i - \gamma P^\pi \phi_i, \phi_j - \gamma P^\pi \phi_j \rangle_\mu \\ b_i &=& \langle \phi_i - \gamma P^\pi \phi_i, r^\pi \rangle_\mu \end{array} \right.$$

Performance of BRM

Proposition 8.

We have

$$\|V^\pi - V_{BR}\| \leq \|(I - \gamma P^\pi)^{-1}\|(1 + \gamma\|P^\pi\|) \inf_{V \in \mathcal{F}} \|V^\pi - V\|. \quad (5)$$

Now, if μ *is the stationary distribution for* π*, then* $\|P^\pi\|_\mu = 1$ *and* $\|(I - \gamma P^\pi)^{-1}\|_\mu = \frac{1}{1-\gamma}$*, thus*

$$\|V^\pi - V_{BR}\|_\mu \leq \frac{1+\gamma}{1-\gamma} \inf_{V \in \mathcal{F}} \|V^\pi - V\|_\mu.$$

Note that the BRM solution has performance guarantees even when μ is not the stationary distribution (contrary to LSTD). See discussion in [Lagoudakis & Parr, 2003] and [Munos, 2003].

Proof of Proposition 8

Point 1: For any fonction V, we have

$$\begin{aligned} V^\pi - V &= V^\pi - T^\pi V + T^\pi V - V \\ &= \gamma P^\pi (V^\pi - V) + T^\pi V - V \\ (I - \gamma P^\pi)(V^\pi - V) &= T^\pi V - V, \end{aligned}$$

thus

$$\|V^\pi - V_{BR}\| \leq \|(I - \gamma P^\pi)^{-1}\| \|\mathcal{T}^\pi V_{BR} - V_{BR}\|$$

and $\|\mathcal{T}^\pi V_{BR} - V_{BR}\| = \inf_{V \in \mathcal{F}} \|\mathcal{T}^\pi V - V\| \leq (1 + \gamma \|P^\pi\|) \inf_{V \in \mathcal{F}} \|V^\pi - V\|,$

and (5) follows.

Point 2: Now when we consider the stationary distribution, we have already seen that $\|P^\pi\|_\mu = 1$, which implies that $\|(I - \gamma P^\pi)^{-1}\|_\mu \leq \sum_{t \geq 0} \gamma^t \|P^\pi\|_\mu^t \leq \frac{1}{1-\gamma}$.

Back to RL

Approximate Policy Iteration algorithm: We studied how to compute an approximation V_k of the value function V^{π_k} for any policy π_k. Now the policy improvement step is:

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

In RL, the transition probabilities and rewards are unknown. How to adapt this methodology? Again, two same ideas:

1. Use sampling methods
2. Use Q-value functions

API with Q-value functions

We now wish to approximate the Q-value function $Q^\pi : X \times A \to \boldsymbol{R}$ for any policy π, where

$$Q^\pi(x,a) = \mathbb{E}\big[\sum_{t\geq 0} \gamma^t r(x_t,a_t) | x_0 = x, a_0 = a, a_t = \pi(x_t), t \geq 1\big].$$

Consider a set of features $\phi_i : X \times A \to \boldsymbol{R}$ and the linear space $\mathcal{F}$

$$\mathcal{F} = \{Q_\alpha(x,a) = \sum_{i=1}^{d} \alpha_i \phi_i(x,a), \alpha \in \boldsymbol{R}^d\}.$$

Least-Squares Policy Iteration

[Lagoudakis & Parr, 2003]

- **Policy evaluation**: At round k, run a trajectory $(x_t)_{1\le t\le n}$ by following policy π_k. Write $a_t = \pi_k(x_t)$ and $r_t = r(x_t, a_t)$. Build the matrix $\hat{A}$ and the vector $\hat{b}$ as

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

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

  and we compute the solution $\hat{\alpha}_{TD}$ of $\hat{A}\alpha = \hat{b}$.
  (Note that $\hat{\alpha}_{TD} \stackrel{a.s.}{\to} \alpha_{TD}$ when $n \to \infty$, since $\hat{A} \stackrel{a.s.}{\to} A$ and $\hat{b} \stackrel{a.s.}{\to} b$).

- **Policy improvement**:

$$\pi_{k+1}(x) \in \arg\max_{a\in A} Q_{\hat{\alpha}_{TD}}(x, a).$$

BRM alternative

We require a *generative model*. At each iteration k, we generate n i.i.d. samples $x_t \sim \mu$, and for each sample, we make a call to the generative model to obtain 2 independent samples y_t and $y'_t \sim p(\cdot|x_t, a_t)$. Write $b_t = \pi_k(y_t)$ and $b'_t = \pi_k(y'_t)$.

We build the matrix $\hat{A}$ and the vector $\hat{b}$ as

$$\begin{aligned} \widehat{A}_{i,j} &= \frac{1}{n}\sum_{t=1}^{n}\left[\phi_i(x_t, a_t) - \gamma\phi_i(y_t, b_t)\right]\left[\phi_j(x_t, a_t) - \gamma\phi_j(y'_t, b'_t)\right], \\ \widehat{b}_i &= \frac{1}{n}\sum_{t=1}^{n}\left[\phi_i(X_t, a_t) - \gamma\frac{\phi_i(y_t, b_t) + \phi_i(y'_t, b'_t)}{2}\right] r_t. \end{aligned}$$

We also have the property that $\hat{A} \stackrel{a.s.}{\to} A$ and $\hat{b} \stackrel{a.s.}{\to} b$ of the BRM system, thus $\hat{\alpha}_{BR} \stackrel{a.s.}{\to} \alpha_{BR}$.

Theoretical guarantees so far

For example, Approximate Value Iteration:

$$\|V^* - V^{\pi_K}\|_\infty \le \frac{2\gamma}{(1-\gamma)^2} \max_{0\le k<K} \underbrace{\|\mathcal{T}V_k - V_{k+1}\|_\infty}_{\text{projection error}} + O(\gamma^K).$$

Sample-based algorithms minimizing an empirical L_∞-norm

$$V_{k+1} = \arg\min_{V\in\mathcal{F}} \max_{1\le i\le n} \left|\widehat{\mathcal{T}V}_k(x_i) - V(x_i)\right|$$

suffer from 2 problems:

- Numerically intractable
- Cannot relate $\|\mathcal{T}V_k - V_{k+1}\|_\infty$ to $\max_i |\widehat{\mathcal{T}V}_k(x_i) - V_{k+1}(x_i)|$

L_2-based algorithms

We would like to use sample-based algorithms minimizing an empirical L_2-norm:

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

which is just a **regression problem!**

- Numerically tractable
- Generalization bounds exits: with high probability,

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

But we need $\|\mathcal{T} V_k - V_{k+1}\|_\infty$, not $\|\mathcal{T} V_k - V_{k+1}\|_2$!

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 9 (Munos, 2003, 2007).

Assume there is a constant $C \geq 1$ and a distribution μ such that $\forall x \in X$, $\forall a \in A$,

$$p(\cdot|x,a) \leq C\mu(\cdot).$$

- *Approximate Value Iteration:*

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

- *Approximate Policy Iteration:*

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

We now have all ingredients for a finite-sample analysis of ADP.

Finite-sample analysis of AVI

Sample n states i.i.d. $x_i \sim \mu$. From each state x_i, each $a \in A$, generate m next state samples $y^j_{i,a} \sim p(\cdot|x_i, a)$. Iterate K times:

$$V_{k+1} = \arg\min_{V \in \mathcal{F}} \sum_{i=1}^{n} \Big| V(x_i) - \max_{a \in A} \Big[r(x_i, a) + \gamma \frac{1}{m} \sum_{j=1}^{m} V_k(y^j_{i,a}) \Big] \Big|^2$$

Proposition 10 (Munos and Szepesvári, 2007).

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

$$\begin{aligned} ||V^* - V^{\pi_K}||_\infty \quad \leq \quad & \frac{2\gamma}{(1-\gamma)^2} C^{1/p} \, d(\mathcal{T}\mathcal{F}, \mathcal{F}) + O(\gamma^K) \\ & + O\Big(\frac{V(\mathcal{F}) \log(1/\delta)}{n} \Big)^{1/4} + O\Big(\frac{\log(1/\delta)}{m} \Big)^{1/2}, \end{aligned}$$

where $d(\mathcal{T}\mathcal{F}, \mathcal{F}) \stackrel{\text{def}}{=} \sup_{g \in \mathcal{F}} \inf_{f \in \mathcal{F}} ||\mathcal{T}g - f||_{2,\mu}$ 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 ϵ-approximation of the optimal policy.

- Policy iteration using a single trajectory [Antos et al., 2008]
- LSTD/LSPI [Lazaric et al., 2010]
- 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.