On the convergence of the Rolling Horizon procedure

A. Jean-Marie
joint work with E. Della Vecchia and S. Di Marco,
Conicet-Univ. Nacional de Rosario, Argentina.

2nd GdR RO/COS Meeting 2015, Paris, 3 December 2015

Outline

Progress

Problem elements

Very generally, decision in a dynamic random environment involves:

- A dynamic system evolving according to a random process
- Transitions which depend on a sequence of actions $a_1, a_2, \ldots$
- A performance evaluated trough a criterion

$$\mathbb{E}G^{(\infty)}(s_1, a_1, s_2, a_2, \ldots, s_n, a_n, \ldots)$$

The objective is to find the optimal sequence $\{a_t; t \in \mathbb{N}\}$, usually in the form of a decision rule that maps histories to actions. For each n:

$$(s_1, a_1, s_2, a_2, \ldots, s_n) \quad \mapsto \quad a_n \ .$$

The Rolling Horizon procedure

Computing the exact optimal decision rule is usually very difficult
$\rightarrow$ approximations and heuristics.

One such heuristic is the Rolling Horizon method:

Step 1: At time t, and for state s_t, solve the finite horizon control problem with a given horizon H, taking s_t as initial state:

$$\mathbb{E}G^{(H)}(s_t, a_t, s_{t+1}, a_{t+1}, \ldots, a_{t+H-1}, s_{t+H})$$

Step 2: Apply just the first policy obtained in the state s_t.

Step 3: Observe the achieved state at time $t+1$

Step 4: $t \leftarrow t+1$; go to Step 1.

Why RH might work

In principle, solving exactly a finite-horizon problem from a known state requires only exploring the event tree

$$\Longrightarrow \mathcal{O}(T^H),$$

(T transition per state, horizon H).
Solving exactly the infinite-horizon problem requires exploring the whole state space and solving linear systems. In practice, let us say:

$$\Longrightarrow \quad \sim C \times N^3$$

(N states), but C may be as large as 2^T.
Solving approximately the infinite-horizon problem can be done with Value Iteration. In practice, about:

$$\Longrightarrow \quad \sim C' \times N \times T$$

and C' depends on the precision required and many other factors.

If H relatively small and N relatively large, this might be worth it.

Research Objectives for RH

The **RH** procedure is a heuristic method. We have studied its precision in different models

- **MDPs** with finite state space
- **MDPs** with general state space, bounded or unbounded reward

Semi-Markov Games the most general case

with different points of view

Convergence Find sufficient conditions for convergence (asymptotic zero error when the horizon $\to \infty$);

Approximation Find error bounds

About the notion of "convergence"

In Rolling Horizon, the horizon length H is fixed and a design parameter: compromise computational complexity/precision.

The notion of convergence refers to the fact that errors should vanish when H is large, and to the quantification of these errors: not to the fact that H changes over time.

Progress

Relation with the Value Iteration procedure

The Markov case:

- transitions $s_t \to s_{t+1} \equiv T_{a_t} s_t$ do not depend on the past, only on the current state s_t and current action a_t.
- the reward function is recursive:

$$G^{(\infty)}(s_1, a_1, s_2, a_2, \ldots) \;=\; f(r(s_1, a_1), h \circ G^{(\infty)}(s_2, a_2, \ldots))$$

(typically, $f(x, y) = x + y$, $h(x) = \beta x$).

Bellman Equation

Result: the optimal decision rule does not depend on the past either and can be chosen deterministic; the value function solves the equation

$$V(s) \;=\; \max_a \mathbb{E}\{f\,(r(s,a), h \circ V(T_a s))\}\;.$$

and the optimal decision $s \mapsto a = d(s)$ realizes the arg max.

Relation with the Value Iteration procedure (ctd)

The Bellman equation is a fixed-point equation of operator T: for a real-valued function w on the state space,

$$(Tw)(s) \;:=\; \max_a \mathbb{E}\{f(r(s,a), h \circ w(T_a s))\} \;.$$

Idea to solve the Bellman equation: iterate until convergence.

Value iteration algorithm

Step 1: $n = 0,\ v^0 = 0,\ S,\ A_s,\ \forall s \in S$

Step 2: Compute $v^{n+1} = Tv^n$ and $d_{n+1}(s) = \arg\max\{\dots v^n\}$

Step 3: If an *adequate stopping rule*, stop. Otherwise, do $n \leftarrow n+1$ and go to Step 2.

Observation: **VI** is computed *offline*, whereas **RH** is supposed to be computed *online*.

Relation with the Value Iteration procedure (ctd)

Both **VI** and **RH** are approximations:

- The gain obtained by **RH** with horizon n is produced by the stationary policy

$$(d_n)^\infty \equiv (d_n, d_n, d_n, \ldots) \quad \Longrightarrow \quad g[((d_n)^\infty)]$$

- The gain obtained by **VI** is produced by the non-stationary, periodic policy

$$(d_n, d_{n-1}, \ldots, d_1)^\infty \equiv (d_n, d_{n-1}, \ldots, d_1, d_n, d_{n-1}, \ldots)$$
$$\Longrightarrow g[(d_n, d_{n-1}, \ldots, d_1)^\infty]$$

The literature concentrates on the convergence:

$$g[(d_n, d_{n-1}, \ldots, d_1)^\infty] \quad \to_{n\to+\infty} \quad g[(d^*)^\infty]$$

but almost ignores the convergence of

$$g[(d_n)^\infty] \quad \to_{n\to+\infty} \quad g[(d^*)^\infty] \qquad !!$$

Missing the point...

In practice, what do people do?

They use **VI** to obtain an approximation of the value function, stopping when they feel satisfied.

They obtain a Markov policy d_N.

They use it as a stationary policy: $(d_N)^\infty$. This is **RH**!
They do not use the non-stationary policy $(d_N, d_{N-1}, \ldots, d_1)$ of which they had computed the value!

Progress

Markov Decision Processes, finite case, average cost

State space and action space are finite.
Performance function is the average gain:

$$\mathbb{E}G^{(\infty)}(s_1, a_1, \ldots) \;=\; \lim_{n\to+\infty} \sup \frac{1}{n}\mathbb{E}\sum_{m=1}^{n} r(s_m, a_m)$$

Convergence concepts

There always exist an optimal average gain for each starting state s: $g^*(s)$. **VI** is said to converge if:

$$\lim_{n\to\infty} v_n \;-\; ng^* \;=\; h^* \;.$$

RH is said to converge if:

$$\lim_{n\to+\infty} g^{(d_n)^\infty} \;=\; g^* \;.$$

Convergence Results

Theorem: Hernández-Lerma (1990)

A sufficient condition for the convergence of RH is:

> *There exists a positive number $\delta < 1$ such that $sp(p(.|s,a) - p(.|s',a')) \leq 2\delta$ for every (s,a) and (s',a') with $s, s' \in S$, $a \in A_s$, $a' \in A_{s'}$,*

where, for $B \subset S$,

$$\mathrm{sp}(\lambda) := \sup_B \lambda(B) - \inf_B \lambda(B) \ .$$

In addition, convergence is geometric with rate δ.

Principal result

Geometric convergence for **RH**: $\exists C, \alpha, \forall n$,

$$0 \leq g^*(s) - g^{(d_n)^\infty}(s) < C\alpha^n,$$

and for **VI**:

$$||v_n - ng^* - h^*||_\infty \; < \; C\alpha^n \; .$$

Theorem (Della Vecchia *et al.* (2011))

If the **VI** algorithm converges geometrically then also does the **RH** procedure.

The converse does not hold: there are cases where **RH** converges, but not **VI**.

Convergence conditions

Condition 1: Schweitzer and Federgruen (1977)

There exists a randomized maximal gain policy whose transition probability matrix is aperiodic (but not necessarily unichain) and has $R^* = \{i \in S :$ i is recurrent for some pure maximal gain policy $\}$ as its set of recurrent states.

Condition 2: Schweitzer and Federgruen (1977)

Every optimal (pure) stationary policy gives rise to an aperiodic (but not necessarily unichain) transition matrix.

Condition 3: *weak unichain condition*, Tijms (1986), p. 199, Assumption 3.3.1.

Every optimal stationary policy has a transition probability matrix unichain and aperiodic

Convergence conditions (ctd.)

Condition 4: Puterman (1994), p. 370

Every stationary policy is unichain and gives rise to an aperiodic transition matrix.

Condition 5: Hernández Lerma and Lasserre (1990), Assumption 5.1.

Cf. supra.

Relative strength of conditions

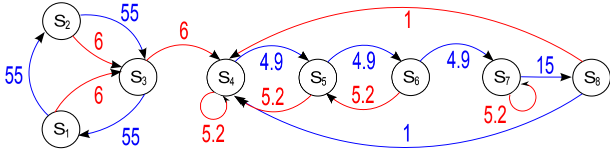
Condition 5 $\Longrightarrow$ Condition 4 $\Longrightarrow$ Condition 3 $\Longrightarrow$ Condition 3 $\Longrightarrow$ Condition 1 $\Longleftrightarrow$ **VI** converges geometrically $\Longrightarrow$ **RH** converges geometrically

All conditions need aperiodicity.

An Example

And in effect, neither procedure converges on the following case:

- $\mathcal{S} = \{s_1, s_2, \ldots, s_8\}$; for each state s_i, $A_{s_i} = \left\{a_1^i, a_2^i\right\}$ for $i = 1, 2, \ldots, 8$.



- The **RH** procedure applied directly on this problem gives us infinitely many times four policies, three of them have gains $(55, 55, 55, 5.2, 5.2, 5.2, 5.2, 5.2)$ and the fourth one, which is the optimal one, has gain $(55, 55, 55, 6.14, 6.14, 6.14, 6.14, 6.14)$.

Approximations

We are generally looking for conditions which provide bounds such as

$$V^* - V_N \;\leq\; C_1 \delta^N \;+\; C_2 \varepsilon$$

where:

- $V_N = g^{(d_N)^\infty}$ is the value obtained with the N-horizon **RH** procedure
- V^* the optimal value
- ε represents some error bound on transition model parameters, such as transition probabilities or gains.

Cases studied

Cases we have studied:

- infinite state space, continuous-state spaces
- semi-Markov transitions
- stochastic games
- discounted, undiscounted rewards, time-consistent risk measures (Ruszczynski *et al.*)
- approximate Rolling Horizon, *à la* Chang & Markus.

The semi-Markov case with unbounded state space

Setting: Semi-Markov decision process.

- discrete state space (possibly infinite) $\mathcal{S}$
- compact/finite action space in state s: $\mathcal{A}_s$
- joint state/action space:

$$\mathbb{K} \ := \ \{(s,a) : s \in \mathcal{S}, a \in \mathcal{A}_s\}$$

- reward function $\ell(s,a)$
- distribution of inter-decision times $F(\cdot|s,a)$
- total expected discounted reward under policy π

$$V^\pi(s) := \mathbb{E}^\pi_s \left[\int_0^\infty e^{-\alpha u} \ell(S_u, A_u) \mathrm{d}u \right] \quad \to \quad \max$$

Assumptions

$$\beta(s,a) := \int_0^\infty e^{-\alpha t} F(\mathrm{d}t|s,a) \ .$$

Assumption 1

$\rho := \sup_{(s,a)\in\mathbb{K}} \beta(s,a) < 1.$

Proposition: Luque-Vásquez (2002)

If there exists a pair of positive numbers θ and ϵ such that

$$F(\theta|s,a) \leqq 1-\epsilon$$

for all $(s,a) \in \mathbb{K}$, then **Assumption 1** holds with $\rho = 1-\epsilon+\epsilon e^{\alpha\theta}$.

Controlled growth conditions

Assumption 2

There exist a function $\mu : S \to [1, \infty)$ and a constant m such that for all $(s, a) \in \mathbb{K}$,

(a)

$$|r(s,a)| \leqq m\,\mu(s),$$

(b)

$$\int_S \mu(z) Q(\mathrm{d}z|s,a) \;\leqq\; \mu(s)\;.$$

$\mathcal{M}_\mu(\mathcal{S})$: the linear space functions v such that $\exists C$,

$$|v(s)| \;\leq\; C\mu(s)$$

for all s (finite μ-weighted norm). This is a Banach space.

Approximate Rolling Horizon

Idea: computing exactly value iterations may be difficult/expensive $\Longrightarrow$ settle for an approximation.

For instance:

- replace V^*_{N-1} with an approximation V,
- compute TV:

$$(Tv)(s) := \sup_{a \in \mathcal{A}_s} \left\{ r(s,a) + \beta(s,a) \int_{\mathcal{S}} v(z) Q(\mathrm{d}z | s, a) \right\} .$$

- return TV instead of $V^*_N = TV^*_{N-1}$.

Typical approximation result

Theorem (Della Vecchia et al. (2012))

Suppose that **Assumptions 1** and **2** hold, and that, for some N, $||V^*_N - V^*_{N-1}||_\mu \leqq \varepsilon_1$.
Let $V \in \mathcal{M}_\mu(\mathcal{S})$ be a function such that $||V^*_{N-1} - V||_\mu \leqq \varepsilon_2$.
Consider a stationary strategy f such that $T^f V = TV$, and let $\widetilde{U}_N = V^f$. Then, $\forall s$,

$$|V^*(s) - \widetilde{U}_N(s)| \;\leq\; \frac{2\rho(\varepsilon_1 + \varepsilon_2)}{1-\rho}\,\mu(s)\;.$$

The goodness of V^*_{N-1} is measured with ε_1: $\varepsilon_1 = 0$ if it is the optimal V^*.

The goodness of V is measured with ε_2: $\varepsilon_2 = 0$ if it is exact.

Approximation result, ctd.

It can be shown that $V_N^* \to V^*$ μ-geometrically. Then:

Corollary

Suppose that **Assumptions 1** and **2** hold. Let $V \in \mathcal{M}_\mu(\mathcal{S})$ be a function such that for some $N \geqq 1$, $||V_{N-1}^* - V||_\mu \leqq \varepsilon$. Consider a policy f_N such that $T^{f_N} V = TV$, and let $U_N = V^{f_N}$. Then,

$$|V^*(s) - U_N(s)| \leqq \left(\frac{2m\rho^N}{1-\rho} + \frac{2\rho\varepsilon}{1-\rho} \right) \mu(s).$$

References

- Alden J.M. and Smith R.L. *Rolling horizon procedure in nonhomogeneus Markov decision processes*, Tech.Report IOE Dept., Univ. Michigan , 1988.
- Hernández-Lerma O., Lasserre J.B., *Error Bounds for Rolling Horizon Policies in Discrete-Time Markov Control Processes.* IEEE Transactions on Automatic Control, 35, 10, 1990, pp. 1118 - 1124
- Kallenberg L., *Finite state and action MDPS*, in Handbook of Markov Decision Processes. Methods and applications, E. Feinberg and A. Shwartz (Eds), Kluwer's international Series, 2002.
- Puterman L., *Markov Decision Processes.* Wiley and Sons (2005).
- Schweitzer P. J.,*Iterative solution of the functional equation of undiscounted Markov renewal programming*, Journal of Mathematical Analysis and Applications 34, (1971), 495-501
- Schweitzer P. J.and Federgruen A, *The asymptotic behavior of undiscounted value iteration in Markov decision problems*, Mathematics of Operations Research, 2(4), 1977, 360-381.
- Schweitzer P. J.and Federgruen A, *Geometric convergence of the value itteration in multichain Markov decision problems*, Adv.Appl.Prob. , 11, 1979, 188-217.

Self references

- Della Vecchia, E., Di Marco S. and Jean-Marie, A., *On the Convergence of Rolling Horizon Procedure and the Average Criterion*, ALIO-INFORMS Joint International Meeting, Buenos Aires, June 2010.
- Della Vecchia, E., Di Marco S. and Jean-Marie, A., *Illustrated review of convergence conditions of the value iteration algorithm and the rolling horizon procedure for average-cost MDPs*, INRIA RR 7710, Aug. 2011. `http://hal.inria.fr/docs/00/61/72/71/PDF/RR-7710.pdf`
- Della Vecchia, E., Di Marco S. and Jean-Marie, A., *Rolling Horizon and State Space Truncation Approximations for Zero-Sum Semi-Markov Games with Discounted Payoff*, INFORMS Applied Probability Society Conf., Stockholm, July 2011.
- Della Vecchia E., Di Marco S., Jean-Marie A., RR INRIA N^{o}8019, *Rolling horizon procedures in Semi-Markov Games: The Discounted Case*. `http://hal.inria.fr/docs/00/72/03/51/PDF/RR-8019.pdf`
- Della Vecchia E., Di Marco S., Jean-Marie A., RR INRIA N^{o}8162, *Structural approximations in discounted semi-Markov games*. `http://hal.inria.fr/docs/00/76/42/17/PDF/RR-8162.pdf`