On the Locality of Action Domination in Sequential Decision Making

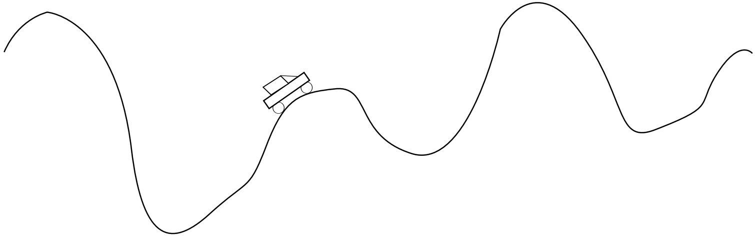
E. Rachelson M. G. Lagoudakis

Technical University of Crete

ISAIM, January 6th, 2010

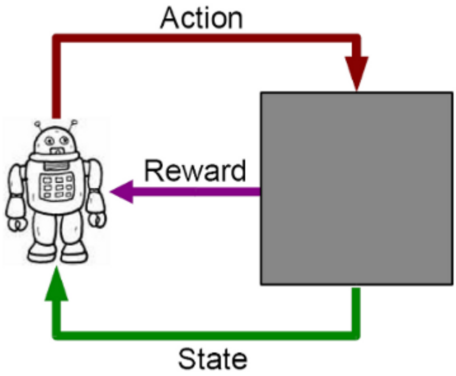
Background

Sequential decision making



Find the best sequence of L/R actions
or the best control policy
to reach the summit.

Background



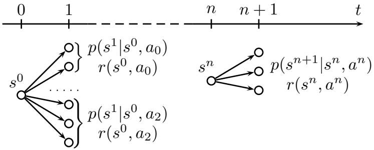
Background

Sequential decision making in Markov Decision Processes.

Markov Decision Process

Tuple $\langle S, A, p, r, T \rangle$
Markovian transition model $p(s'|s, a)$
Reward model $r(s, a)$
T is a set of timed decision epochs $\{0, 1, \ldots, H\}$

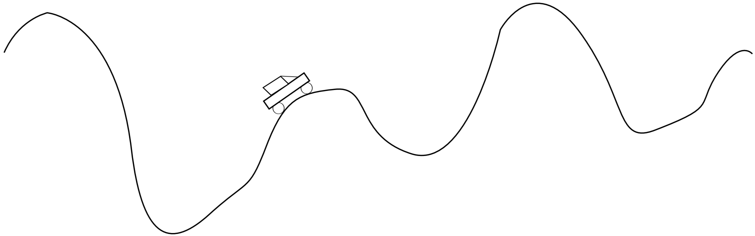
Infinite (unbounded) horizon: $H \longrightarrow \infty$



Goal: optimize a cumulative reward.

How local is the knowledge gained from experience?

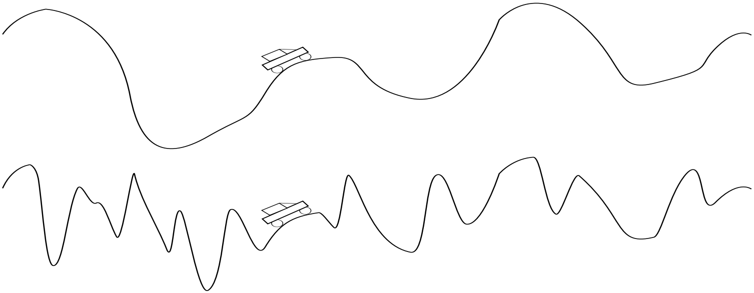
Learning an improved policy



Intuition indicates that a “good” action in a given position remains “good” **around** this position.

Environment smoothness

Ability to generalize $\leftrightarrow$ regularity of the environment



But the environment's model is unknown:

it is still possible to make an hypothesis on learn its **smoothness**.

Focus of this contribution

- Formalize the notion of smoothness for the underlying model,
- Derive properties for the optimal policy and value function,
- Exploit these properties in an algorithm for RL problems.

1. General intuition
2. Key results
3. Localized Policy Iteration

Characterizing the environment's regularity?

Model smoothness $\leftrightarrow$ Lipschitz continuity

Lipschitz continuity

$$f : X \rightarrow Y, \forall (x_1, x_2) \in X^2, \quad d_Y\big(f(x_1) - f(x_2)\big) \leq L \cdot d_X(x_1 - x_2)$$

Characterizing the environment's regularity?

Model smoothness $\leftrightarrow$ Lipschitz continuity

Transition model's smoothness

- The results of two "similar" actions, in two "similar" states, are "similar".
- LC on probability distributions.
- Kantorovich distance:

$$K(p,q) = \sup_f \left\{ \left| \int_X f dp - \int_X f dq \right| : \|f\|_L \leq 1 \right\}$$

- L_p-LC transition model:

$$K\big(p(\cdot|s,a), p(\cdot|\hat{s},\hat{a})\big) \leq L_p\big(\|s-\hat{s}\| + \|a-\hat{a}\|\big)$$

Characterizing the environment's regularity?

Model smoothness $\leftrightarrow$ Lipschitz continuity

Reward model's smoothness

- The rewards of two "similar" transitions are "similar".
- L_r-LC reward model:

$$\left|r(s,a) - r(\hat{s},\hat{a})\right| \leq L_r\left(\|s-\hat{s}\| + \|a-\hat{a}\|\right)$$

Characterizing the environment's regularity?

Model smoothness $\leftrightarrow$ Lipschitz continuity

Policy's smoothness

LC on actions or action distributions. L_π-LC policies:

$$d_\Pi\left(\pi(s) - \pi(\hat{s})\right) \leq L_\pi \|s - \hat{s}\|$$

Characterizing the environment's regularity?

Model smoothness $\leftrightarrow$ Lipschitz continuity

Model smoothness hypothesis

- (L_p, L_r)-LC MDP.

$$K\big(p(\cdot|s,a), p(\cdot|\hat{s},\hat{a})\big) \leq L_p\big(\|s-\hat{s}\| + \|a-\hat{a}\|\big)$$
$$\big|r(s,a) - r(\hat{s},\hat{a})\big| \leq L_r\big(\|s-\hat{s}\| + \|a-\hat{a}\|\big)$$

- L_π-LC policies.

$$d_\Pi\big(\pi(s) - \pi(\hat{s})\big) \leq L_\pi \|s-\hat{s}\|$$

Intermediate results on LC of value functions

Lemma (Lipschitz continuity of the value function)

$$\left.\begin{array}{l} L_Q\text{-}LC\ Q\text{-}function\ Q \\ L_\pi\text{-}LC\ policy\ \pi \end{array}\right\} \Rightarrow [L_Q(1+L_\pi)]\text{-}LC\ value\ function\ V^\pi\ w.r.t.\ Q$$

Intermediate results on LC of value functions

Lemma (Lipschitz continuity of the value function)

$$\left.\begin{array}{l} L_Q\text{-}\mathit{LC}\ Q\text{-}\mathit{function}\ Q \\ L_\pi\text{-}\mathit{LC}\ \mathit{policy}\ \pi \end{array}\right\} \Rightarrow [L_Q(1+L_\pi)]\text{-}\mathit{LC}\ \mathit{value\ function}\ V^\pi\ \mathit{w.r.t.}\ Q$$

Lemma (Lipschitz continuity of the n-step Q-value)

$$\left.\begin{array}{l} (L_p, L_r)\text{-}\mathit{LC\ MDP} \\ L_\pi\text{-}\mathit{LC}\ \mathit{policy}\ \pi \end{array}\right\} \Rightarrow \left\{\begin{array}{l} \mathit{the}\ n\text{-}\mathit{step,\ finite\ horizon,}\ \gamma\text{-}\mathit{discounted} \\ \mathit{value\ function}\ Q^\pi_n\ \mathit{is}\ L_{Q_n}\text{-}\mathit{LC,\ with:} \end{array}\right.$$

$$L_{Q_{n+1}} = L_r + \gamma(1+L_\pi)L_pL_{Q_n}\ .$$

LC of value functions

Theorem (Lipschitz-continuity of the Q-values)

$$\left.\begin{array}{l} (L_p, L_r)\text{-LC MDP} \\ L_\pi\text{-LC policy } \pi \\ \gamma L_p (1 + L_\pi) < 1 \end{array}\right\} \Rightarrow \left\{ \begin{array}{l} \text{the infinite horizon, } \gamma\text{-discounted} \\ \text{value function } Q^\pi \text{ is } L_Q\text{-LC, with:} \end{array}\right.$$

$$L_Q = \frac{L_r}{1 - \gamma L_p (1 + L_\pi)}$$

Short discussion

- Value of L_π.
  For most common discrete policies, almost everywhere in the state space, one can prove the previous result with $L_\pi = 0$.
- $\gamma L_p (1 + L_\pi) < 1$?
  With $L_\pi = 0$, $\gamma L_p < 1$.
  $\Rightarrow$ The environment's spatial variations (L_p)
  need to be compensated by
  the discount on temporal variations (γ)
  to obtain smoothness guarantees on the Q-function.
- No guarantees $\not\Rightarrow$ no smoothness.

Short discussion

- Value of L_π.
  For most common discrete policies, almost everywhere in the state space, one can prove the previous result with $L_\pi = 0$.
- $\gamma L_p(1 + L_\pi) < 1$?
  With $L_\pi = 0$, $\gamma L_p < 1$.
  $\Rightarrow$ The environment's spatial variations (L_p)
  need to be compensated by
  the discount on temporal variations (γ)
  to obtain smoothness guarantees on the Q-function.
- No guarantees $\not\Rightarrow$ no smoothness.

Short discussion

- Value of L_π.
  For most common discrete policies, almost everywhere in the state space, one can prove the previous result with $L_\pi = 0$.
- $\gamma L_p(1 + L_\pi) < 1$?
  With $L_\pi = 0$, $\gamma L_p < 1$.
  $\Rightarrow$ The environment's spatial variations (L_p)
  need to be compensated by
  the discount on temporal variations (γ)
  to obtain smoothness guarantees on the Q-function.
- No guarantees $\not\Rightarrow$ no smoothness.

Local validity of dominating actions

Definition (Sample)

$\left(s, \Delta^{\pi}(s), a^{*}(s)\right)$ with:

- $a^{*}(s)$ the one step lookahead dominating action in s
- $\Delta^{\pi}(s)$ the domination gap

Local validity of dominating actions

Definition (Sample)

$\left(s, \Delta^\pi(s), a^*(s)\right)$ with:

- $a^*(s)$ the one step lookahead dominating action in s
- $\Delta^\pi(s)$ the domination gap

Theorem (Influence radius of a sample)

Given a policy π, with:

$$\left.\begin{array}{l} L_Q\text{-}LC \textit{ value function } Q^\pi \\ \left(s, \Delta^\pi(s), a^*(s)\right) \end{array}\right\} \Rightarrow a^*(s) \textit{ dominates in all } s' \in B\left(s, \rho(s)\right)$$

$$\rho(s) = \frac{\Delta^\pi(s)}{2L_Q}.$$

Exploiting influence radii

"Sampling"

Acquiring information concerning the dominating action in a given state.

Two parallel processes:

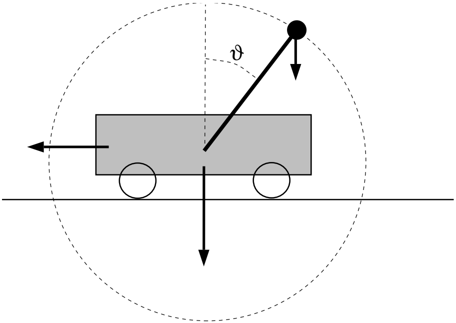
- Focus sampling on states providing high domination values (large ρ).
- Removing chunks of the state space where local validity is guaranteed.

LPI — The algorithm

Init: threshold ε_c, π_0, $n=0$, $W = \textsc{Draw}(m, d(), S)$
while $\pi_n \neq \pi_{n-1}$ **do**
  $n \leftarrow n+1$, $c=1$, $\mathscr{B} = \emptyset$
  while $c > \varepsilon_c$ **do**
    $\left(s, a^*(s), \Delta^{\pi_{n-1}}(s)\right) \leftarrow \textsc{GetSample}(\pi_{n-1}, W)$
    $\mathscr{B} \leftarrow \mathscr{B} \cup \left\{\left(B\left(s, \rho(s)\right), a^*(s)\right)\right\}$
    for all $s' \in W \cap B\left(s, \rho(s)\right)$, remove s' and repopulate W
    $c = 1 - \textsc{Volume}(\mathscr{B}) / \textsc{Volume}(S)$
  $\pi_n = \textsc{Policy}(\pi_{n-1}, T)$

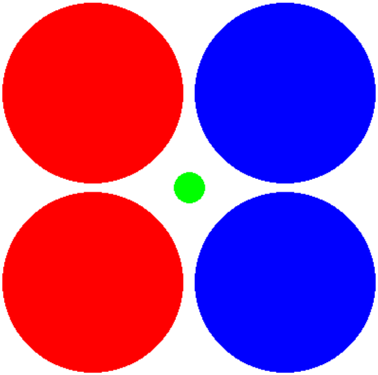
$\textsc{GetSample}(\pi, W)$
loop
  select state s in W with highest utility $U(s)$
  for all $a \in A$, update $Q^\pi(s,a), \Delta^\pi(s), U(s)$, statistics
  if there are sufficient statistics for s, **return** $\left(s, a^*(s), \Delta^\pi(s)\right)$

Results on the inverted pendulum



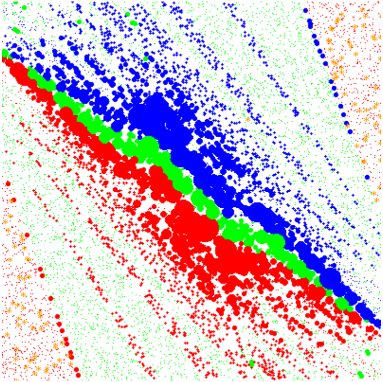
Goal: move left/right to balance the pendulum.
State space: $(\theta, \dot{\theta})$

Successive policies



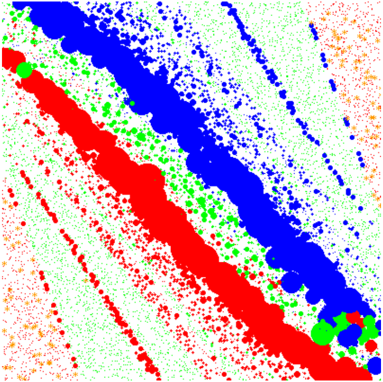
π_0

Successive policies



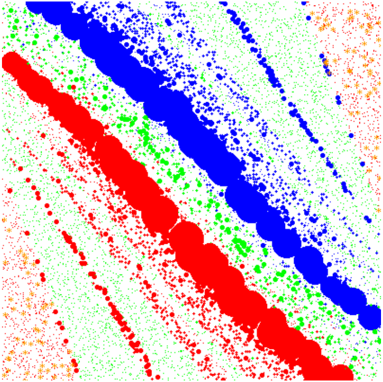
π_1

Successive policies



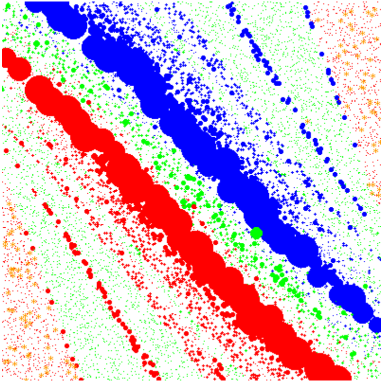
π_2

Successive policies



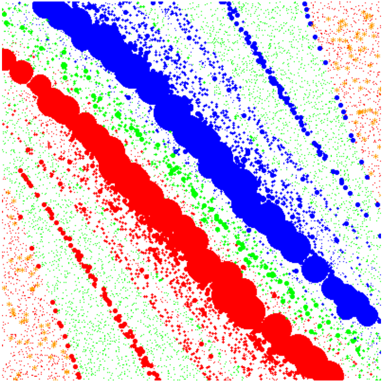
π_3

Successive policies



π_4

Successive policies



π_5

Remarks

- A balancing policy is found very early.
- No *a priori* discretization of the state space, nor parameterization of the value function.
- Focus on the global shape of π_n before refinement.
- Reduced computational effort.

Conclusion

- Original question:
  How **local** is the info gathered in one state about the dominating action?

Conclusion

- Original question:
  How **local** is the info gathered in one state about the dominating action?
- Formalize the notion of **smoothness** for the environment's underlying model:
  Kantorovich distance, Lipschitz continuity $\longrightarrow$ MDP smoothness.
  Other metrics? Other continuity criterion?
  Other similarity measure?

Conclusion

- Original question:
  How **local** is the info gathered in one state about the dominating action?
- Formalize the notion of **smoothness** for the environment's underlying model:
  Kantorovich distance, Lipschitz continuity $\longrightarrow$ MDP smoothness.
  ? Other metrics? Other continuity criterion?
  Other similarity measure?
- Derive properties for the **policies** and **value functions**:
  LC of the (optimal) value functions, influence radius of a sample.

Conclusion

- Original question:
  How **local** is the info gathered in one state about the dominating action?

- Formalize the notion of **smoothness** for the environment's underlying model:
  Kantorovich distance, Lipschitz continuity $\rightarrow$ MDP smoothness.

  Other metrics? Other continuity criterion?
  Other similarity measure?

- Derive properties for the **policies** and **value functions**:
  LC of the (optimal) value functions, influence radius of a sample.

- Exploit these properties in an **algorithm** for RL problems:
  Localized Policy Iteration combines UCB-like methods with influence radii into an active learning method.

  Deeper study of incremental/asynchronous PI methods.

Thank you for your attention!