Probabilistic Aspects of Computer Science: MDP

Serge Haddad

LMF, ENS Paris-Saclay & CNRS & INRIA

MPRI M1

Plan

Mixing non determinism and probability

Numerous systems present both non deterministic and probabilistic features.

Acting in an uncertain world

- **non determinism:** decisions of an agent;
- **probability:** effects of the decisions;
- **goal:** maximizing some utility function.

Randomness against the environment

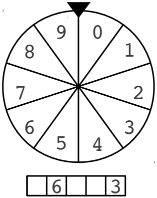
- **probability:** distributed randomized algorithm;
- **non determinism:** network behaviour;
- **goal:** evaluating the worst case behaviour.

Optimization problems

The spinner game

The player has to compose a five-digit number.

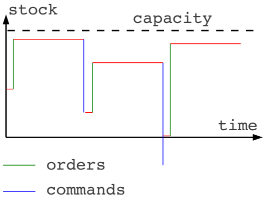
- The digits are randomly chosen by a spinner during five rounds.
- After every round (except the last one),
  the player chooses in which position he inserts the current digit.
- The goal of the player is to obtain the largest number as possible.



Management of a stock

The stock is in a warehouse with fixed capacity.

- The manager decides at the beginning of every month, which additional stock he will order.
- The monthly commands randomly arrive following some distribution. If the commands exceed the inventory the commands are lost.
- Every unit of a stock has a monthly cost while selling it provides a benefit.
- The aim of the manager is to maximize the expected profit.



Introduction to Markov decision process

A Markov decision process MDP is a (finite) transition system.

The dynamic of the system is defined as follows.

- Non deterministically, one chooses an *action* enabled in the current *state*.
- Then one randomly selects the next state.
  The *distribution* depends on the current state and on the selected action.
- There is a numerical *reward* per pair of (current) state and (selected) action.
- For *finite horizon* problems, there is a *terminal reward* per state.

For some problems, rewards are not required.

Syntax of MDP

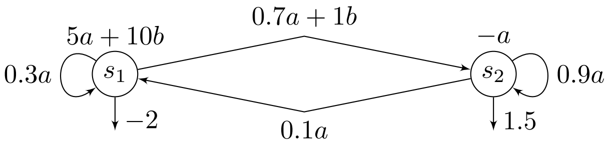
An MDP $\mathcal{M} \stackrel{\text{def}}{=} (S, \{A_s\}_{s\in S}, p, r, rend)$ is defined by:

- S, the finite set of states;
- For every state s, A_s, the finite set of actions enabled in s.
  $A \stackrel{\text{def}}{=} \bigcup_{s\in S} A_s$ is the whole set of actions.
- p, a mapping from $\{(s,a) \mid s \in S, a \in A_s\}$ to the set of distributions over S.
  $p(s'|s,a)$ denotes the probability to go from s to s' if a is selected.
- r, a mapping from $\{(s,a) \mid s \in S, a \in A_s\}$ to $\mathbb{R}$.
  $r(s,a)$ is the reward associated with the selection of a in state s.
- $rend$, a mapping from S to $\mathbb{R}$. $rend(s)$ is the reward when ending in state s.

An example of MDP

An MDP with two states (s_1 and s_2)

- In s_1 actions a and b are enabled while in s_2 only action a is possible.
- A vertex s is labelled by $\sum_{a \in A_s} r(s,a)a$;
- An edge from s to s' is labelled by $\sum_{a \in A_s} p(s'|s,a)a$
- The ending edge of s is labelled by $rend(s)$.



When a is chosen in state s_1,
the probability that the next state is s_2, is 0.7 and the reward is 5.

The terminal reward of s_2 is 1.5.

The rewards could depend on the destination state letting unchanged the theory.

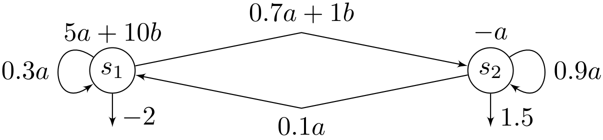
Rewards for histories

A *history* $\sigma \stackrel{\text{def}}{=} (s_0, a_0, \ldots, s_i, a_i, \ldots)$ is a sequence alternating states and actions. $\lg(\sigma) \in \mathbb{N} \cup \{\infty\}$ denotes the number of actions of σ.

Let σ be an history and $0 < \lambda < 1$. Then:

- When $\lg(\sigma) < \infty$, the *total reward* of σ is: $u(\sigma) \stackrel{\text{def}}{=} \sum_{0 \le i < \lg(\sigma)} r(s_i, a_i) + rend(s_{\lg(\sigma)})$. and $v(\sigma) \stackrel{\text{def}}{=} \sum_{0 \le i < \lg(\sigma)} r(s_i, a_i)$ is the *pure total reward*.
- When $\lg(\sigma) = \infty$, the *discounted reward* of σ w.r.t. λ is: $v_\lambda(\sigma) \stackrel{\text{def}}{=} \sum_{0 \le i} r(s_i, a_i)\lambda^i$.
- When $\lg(\sigma) = \infty$, the *lim sup average reward* of σ is: $g_+(\sigma) \stackrel{\text{def}}{=} \limsup_{n\to\infty} \frac{1}{n} \sum_{0 \le i < n} r(s_i, a_i)$.
- When $\lg(\sigma) = \infty$, the *lim inf average reward* of σ is: $g_-(\sigma) \stackrel{\text{def}}{=} \liminf_{n\to\infty} \frac{1}{n} \sum_{0 \le i < n} r(s_i, a_i)$.

Examples of rewards



$$\sigma \stackrel{\text{def}}{=} (s_1, a, s_2, a, s_1, b, s_2)$$
$$u(\sigma) = 5 - 1 + 10 + 1.5 = 15.5$$

$$\sigma \stackrel{\text{def}}{=} (s_1, a)^\omega$$
$$v_{\frac{2}{3}}(\sigma) = 5(1 + \tfrac{2}{3} + (\tfrac{2}{3})^2 + \cdots) = 15$$

$$\sigma \stackrel{\text{def}}{=} (s_1, a, s_2, a)(s_1, b, s_2, a)\ldots(s_1, a, s_2, a)^{2^i}(s_1, b, s_2, a)^{2^i}\ldots$$
$$g_+(\sigma) = \lim_{i\to\infty} \tfrac{13(2^{i+1}-1)+5}{4(2^{i+1}-1)+1} = \tfrac{13}{4}$$
$$g_-(\sigma) = \lim_{i\to\infty} \tfrac{13(2^i-1)+4(2^i)}{4(2^i-1)+2^{i+1}} = \tfrac{17}{6}$$

From MDP to DTMC: principles

In order to obtain a stochastic process,
one needs to fix the non deterministic features of the MDP.

- *Decision rules* select at some time instant the next action depending on the history of the execution.
- *Policies* specify which decision rules should be used at any time instant.

Classes of decision rules and policies are defined depending on two criteria.

- the information used in the history;
- the way the selection is performed (deterministically or randomly).

From MDP to DTMC: decision rules

A *decision rule* d_t maps every history σ of length $t < \infty$
to a distribution $d_t(\sigma)$ over A_{s_t}.

- D_t^{HR} is the set of decision rules at time t.
  It is also called *history-dependent randomized* decision rules.

- D_t^{HD} is the subset of *history-dependent deterministic* decision rules at time t.
  It consists in selecting a single action. In this case $d_t(\sigma) \in A_{s_t}$.

- D_t^{MR} is the subset of *Markovian randomized* decision rules at time t.
  D_t^{MR}, also denoted D^{MR}, only depends on the final state of the history.
  So one denotes $d_t(s)$ the distribution that depends on s.

- D^{MD} is the subset of *Markovian deterministic* decision rules at time t.
  D^{MD} only depends on the final state of the history and selects a single action.
  So $d_t(s) \in A_s$.

From MDP to DTMC: policies

A *policy* (also called a *strategy*) $\boldsymbol{\pi} \stackrel{\text{def}}{=} (d_0, \ldots, d_t, \ldots)$ is a finite or infinite sequence of decision rules such that d_t is a decision rule at time t.

The set of policies such that for all t, $d_t \in D_t^K$ is denoted Π^K.

When decisions d_t are Markovian and all equal to some d,
$\boldsymbol{\pi}$ is said *stationary* and denoted d^∞.

Π^{SR} (resp. Π^{SD}) is the set of stationary randomized (resp. deterministic) policies.

Once a policy is chosen, an MDP becomes a DTMC
whose states are information used in histories.

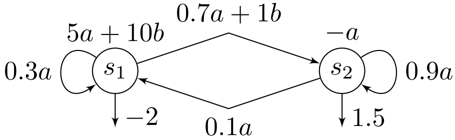
Given d^∞, the states of the DTMC are those of the MDP and the matrix $\mathbf{P}_d$ is:

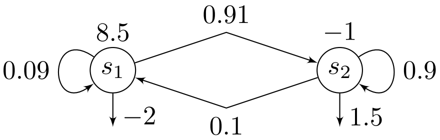
$$\mathbf{P}_d[s, s'] \stackrel{\text{def}}{=} \sum_{a \in A_s} d(s)(a) p(s'|s, a)$$

The (expected) reward in state s is: $\mathbf{r}_d[s] \stackrel{\text{def}}{=} \sum_{a \in A_s} d(s)(a) r(s, a)$

A randomized stationary policy

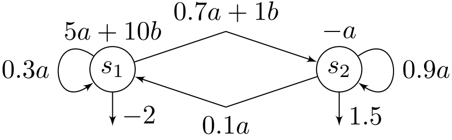
In state s_1, choose a with probability 0.3 and b with probability 0.7.

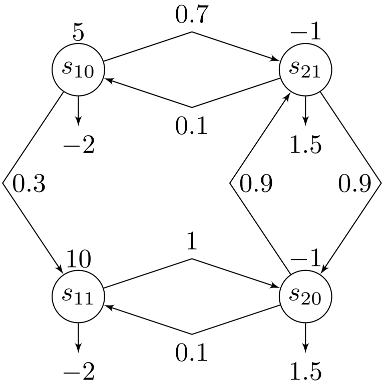




A Markovian non stationary policy

In state s_1, choose a on even instants and b on odd instants.





Rewards for policies

X_n denotes the random state at time n and Y_n denotes the action at time n.

Let $\boldsymbol{\pi}$ be a policy with $\mathbf{E}^{\boldsymbol{\pi}}$ the corresponding expectations, $t \in \mathbb{N}$ and $0 < \lambda < 1$. Then:

- The *total (expected) reward* at time t of $\boldsymbol{\pi}$ is:
  $$u_t^{\boldsymbol{\pi}} \stackrel{\text{def}}{=} \sum_{0 \leq i < t} \mathbf{E}^{\boldsymbol{\pi}}(r(X_i, Y_i)) + \mathbf{E}^{\boldsymbol{\pi}}(rend(X_t))$$
- The *pure total (expected) reward* at time t of $\boldsymbol{\pi}$ is:
  $$v_t^{\boldsymbol{\pi}} \stackrel{\text{def}}{=} \sum_{0 \leq i < t} \mathbf{E}^{\boldsymbol{\pi}}(r(X_i, Y_i))$$
- The *discounted (expected) reward* of $\boldsymbol{\pi}$ w.r.t. λ is:
  $$v_\lambda^{\boldsymbol{\pi}} \stackrel{\text{def}}{=} \sum_{0 \leq i} \lambda^i \mathbf{E}^{\boldsymbol{\pi}}(r(X_i, Y_i))$$
- The *lim sup average (expected) reward* of $\boldsymbol{\pi}$ is:
  $$g_+^{\boldsymbol{\pi}} \stackrel{\text{def}}{=} \limsup_{n \to \infty} \frac{1}{n} \sum_{0 \leq i < n} \mathbf{E}^{\boldsymbol{\pi}}(r(X_i, Y_i))$$
- The *lim inf average (expected) reward* of $\boldsymbol{\pi}$ is:
  $$g_-^{\boldsymbol{\pi}} \stackrel{\text{def}}{=} \liminf_{n \to \infty} \frac{1}{n} \sum_{0 \leq i < n} \mathbf{E}^{\boldsymbol{\pi}}(r(X_i, Y_i))$$

Optimization problems

Let $u_t^* \stackrel{\text{def}}{=} \sup(u_t^{\boldsymbol{\pi}} \mid \boldsymbol{\pi} \in \Pi^{HR})$

- Compute u_t^*;
- When there is some policy $\boldsymbol{\pi}$ such that $u_t^* = u_t^{\boldsymbol{\pi}}$ compute such a policy;
- In general given $\varepsilon > 0$, compute some policy $\boldsymbol{\pi}_\varepsilon$ such that $u_t^* \leq u_t^{\boldsymbol{\pi}_\varepsilon} + \varepsilon$.

Solve similar problems for:

- the discounted reward: $v_\lambda^* \stackrel{\text{def}}{=} \sup(v_\lambda^{\boldsymbol{\pi}} \mid \boldsymbol{\pi} \in \Pi^{HR})$;
- the $\limsup$ and $\liminf$ average rewards: $g_+^* \stackrel{\text{def}}{=} \sup(g_+^{\boldsymbol{\pi}} \mid \boldsymbol{\pi} \in \Pi^{HR})$ and $g_-^* \stackrel{\text{def}}{=} \sup(g_-^{\boldsymbol{\pi}} \mid \boldsymbol{\pi} \in \Pi^{HR})$.

From policies to Markovian policies (1)

Let $\boldsymbol{\pi} \in \Pi^{HR}$ be a policy.
Then there exists $\boldsymbol{\pi}' \in \Pi^{MR}$ such that for all $n \in \mathbb{N}$, $s_0, s \in S$ and $a \in A_s$:

$$\mathbf{Pr}^{\boldsymbol{\pi}'}(X_n = s, Y_n = a \mid X_0 = s_0) = \mathbf{Pr}^{\boldsymbol{\pi}}(X_n = s, Y_n = a \mid X_0 = s_0)$$

Proof. Let us define a Markovian policy $\boldsymbol{\pi}' = (d'_0, d'_1, \ldots)$ by:

$$d'_n(s)(a) \stackrel{\text{def}}{=} \mathbf{Pr}^{\boldsymbol{\pi}}(Y_n = a \mid X_n = s, X_0 = s_0)$$

For $n = 0$, the equality

$$\mathbf{Pr}^{\boldsymbol{\pi}'}(X_n = s, Y_n = a \mid X_0 = s_0) = \mathbf{Pr}^{\boldsymbol{\pi}}(X_n = s, Y_n = a \mid X_0 = s_0)$$

is only relevant for $s = s_0$ and holds by definition of $\boldsymbol{\pi}'$.

Assume that the equality holds up to n. Then:

$$\mathbf{Pr}^{\boldsymbol{\pi}'}(X_{n+1} = s \mid X_0 = s_0) = \sum_{s' \in S, a \in A_{s'}} \mathbf{Pr}^{\boldsymbol{\pi}'}(X_n = s', Y_n = a \mid X_0 = s_0) p(s|s', a)$$
$$= \sum_{s' \in S, a \in A_{s'}} \mathbf{Pr}^{\boldsymbol{\pi}}(X_n = s', Y_n = a \mid X_0 = s_0) p(s|s', a) = \mathbf{Pr}^{\boldsymbol{\pi}}(X_{n+1} = s \mid X_0 = s_0)$$

Now:

$$\mathbf{Pr}^{\boldsymbol{\pi}'}(X_{n+1} = s, Y_{n+1} = a \mid X_0 = s_0) = d'_{n+1}(s)(a)\mathbf{Pr}^{\boldsymbol{\pi}'}(X_{n+1} = s \mid X_0 = s_0)$$
$$= \mathbf{Pr}^{\boldsymbol{\pi}}(Y_{n+1} = a \mid X_{n+1} = s, X_0 = s_0)\mathbf{Pr}^{\boldsymbol{\pi}'}(X_{n+1} = s \mid X_0 = s_0)$$
$$= \mathbf{Pr}^{\boldsymbol{\pi}}(X_{n+1} = s, Y_{n+1} = a \mid X_0 = s_0)$$

From policies to Markovian policies (2)

$$\mathbf{E}^{\boldsymbol{\pi}}(r(X_i, Y_i)) = \sum_{s \in S, a \in A_s} r(s,a)\mathbf{Pr}^{\boldsymbol{\pi}}(X_i = s, Y_i = a)$$

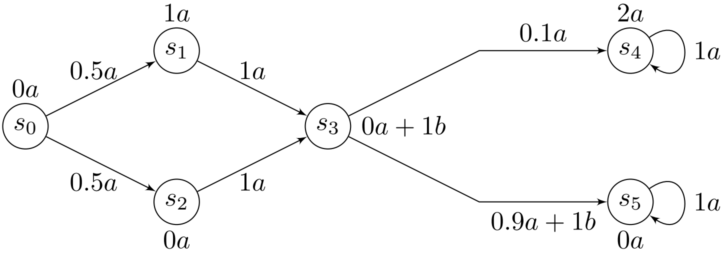
Thus $\boldsymbol{\pi}'$ achieves the same rewards that those of $\boldsymbol{\pi}$.

Warning: the result is only valid for these kinds of rewards.

Can you find a kind of rewards for which it does not hold?

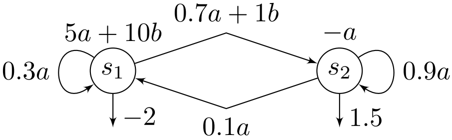
A counter-example

The *maximal (expected) reward*: $\mathbf{E}^{\boldsymbol{\pi}}(\max_{i\in\mathbb{N}}(r(X_i, Y_i)))$



Plan

An introductive example (1)



$u_0^{\boldsymbol{\pi}}$ is independent from $\boldsymbol{\pi}$ and so here: $u_0^*[s_1] = -2$ and $u_0^*[s_2] = 1.5$

Consider horizon $t = 1$. Then in state s_1:

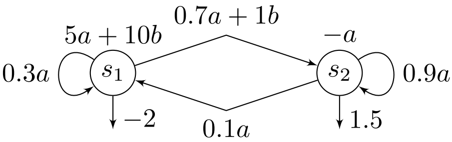
- either one selects a and gets $5 + 0.3u_0^*[s_1] + 0.7u_0^*[s_2] = 5.45$;
- either one selects b and gets $10 + u_0^*[s_2] = 11.5$;
- or one performs a random choice getting $5.45\alpha + 11.5(1 - \alpha)$ with $0 < \alpha < 1$.

Thus $u_1^*[s_1] = 11.5$.

In state s_2, one selects a and gets $-1 + 0.1u_0^*[s_1] + 0.9u_0^*[s_2] = 0.15$

The optimal decision rule d_1 is: $d_1(s_1) = b$ and $d_1(s_2) = a$

An introductive example (2)



Consider horizon $t = 2$. Then in state s_1:

- either one selects a and gets $5 + 0.3u_1^*[s_1] + 0.7u_1^*[s_2] = 8.555$;
- either one selects b and gets $10 + u_1^*[s_1] = 10.15$;
- or one performs a random choice getting $8.555\alpha + 10.15(1 - \alpha)$ with $0 < \alpha < 1$.

Thus $u_2^*[s_1] = 10.15$.

In state s_2, one selects a and gets $-1 + 0.1u_1^*[s_1] + 0.9u_1^*[s_2] = 0.285$

The optimal decision policy is (d_1, d_1).

The algorithm

This algorithm is based on dynamic programming.

It computes the optimal values $optval$ and decisions $optdec$ by increasing horizons.

For $s \in S$ **do** $optval[s, 0] \leftarrow rend(s)$
For i **from** 1 **to** n **do**
  For $s \in S$ **do**
    $best \leftarrow -\infty$
    For $a \in A_s$ **do**
      $temp \leftarrow r(s, a)$
      For $s' \in S$ **do** $temp \leftarrow temp + p(s'|s, a)optval[s', i-1]$
      If $best < temp$ **then** $best \leftarrow temp$; $optdec[s, i] \leftarrow a$
    $optval[s, i] \leftarrow best$

It performs in $O(n|S|^2|A|)$.

Correctness of the algorithm

The proof is done by induction on the time horizon.

Assume optimality of $\boldsymbol{\pi}_{n-1} \stackrel{\text{def}}{=} (d_{n-1}, \ldots, d_1)$ (indexed in a backward way), the policy computed by the algorithm for time horizon $n-1$.

Let d_n be the decision rule computed at the n^{th} iteration.
Pick an arbitrary policy $\boldsymbol{\pi}'_n \stackrel{\text{def}}{=} d'_n, \ldots, d'_1$ and denote $\boldsymbol{\pi}'_{n-1} \stackrel{\text{def}}{=} d'_{n-1}, \ldots, d'_1$.

Let $s \in S$,

$$\mathbf{u}_n^{\boldsymbol{\pi}_n}[s] = r(s, d_n(s)) + \sum_{s' \in S} p(s'|s, d_n(s)) \mathbf{u}_{n-1}^{\boldsymbol{\pi}_{n-1}}[s']$$

$$\geq r(s, d'_n(s)) + \sum_{a \in A_s} d'_n(s)(a) \sum_{s' \in S} p(s'|s, a) \mathbf{u}_{n-1}^{\boldsymbol{\pi}_{n-1}}[s']$$

(due to the iterative step of the algorithm)

$$\geq r(s, d'_n(s)) + \sum_{a \in A_s} d'_n(s)(a) \sum_{s' \in S} p(s'|s, a) \mathbf{u}_{n-1}^{\boldsymbol{\pi}'_{n-1}}[s'] = \mathbf{u}_n^{\boldsymbol{\pi}'_n}[s]$$

(due to the inductive hypothesis)

Plan

Presentation

Finite Horizon Analysis

3. Discounted Reward Analysis

Average Reward Analysis

Preliminary observations and notations

Let $\boldsymbol{\pi} \stackrel{\text{def}}{=} (d_0, \dots, d_n, \dots)$ be some Markovian policy. Then:

$$\mathbf{v}_\lambda^{\boldsymbol{\pi}}(s) = \mathbf{r}_{d_0}(s) + \lambda \sum_{s' \in S} \mathbf{P}_{d_0}[s, s'] \mathbf{r}_{d_1}(s') + \lambda^2 \sum_{s' \in S} \sum_{s'' \in S} \mathbf{P}_{d_0}[s, s'] \mathbf{P}_{d_1}[s', s''] \mathbf{r}_{d_2}(s'') + \cdots$$

$$\mathbf{v}_\lambda^{\boldsymbol{\pi}} = \sum_{i \in \mathbb{N}} \lambda^i \left(\prod_{0 \leq j < i} \mathbf{P}_{d_j} \right) \mathbf{r}_{d_i}$$

Let $\boldsymbol{\pi} \stackrel{\text{def}}{=} d^\infty$, this reward can be rewritten as: $\mathbf{v}_\lambda^{\boldsymbol{\pi}} = \sum_{i \in \mathbb{N}} \left(\lambda \mathbf{P}_d\right)^i \mathbf{r}_d$

$\mathbf{Id} - \lambda \mathbf{P}_d$ is invertible and its inverse is $\sum_{i \in \mathbb{N}} \left(\lambda \mathbf{P}_d\right)^i$. So:

$$\mathbf{v}_\lambda^{\boldsymbol{\pi}} = \left(\mathbf{Id} - \lambda \mathbf{P}_d\right)^{-1} \mathbf{r}_d \text{ and consequently } \mathbf{v}_\lambda^{\boldsymbol{\pi}} = \mathbf{r}_d + \lambda \mathbf{P}_d \mathbf{v}_\lambda^{\boldsymbol{\pi}}$$

Let L be the mapping from $\mathbb{R}^S$ to $\mathbb{R}^S$ defined by:

$$L(\mathbf{v})[s] \stackrel{\text{def}}{=} \max \left(r(s, a) + \lambda \sum_{s' \in S} p(s' | s, a) \mathbf{v}[s'] \mid a \in A_s \right)$$

L "selects" the best decision rule for time horizon 1 and terminal reward $\lambda \mathbf{v}$.

Characterization of optimality (1)

Theorem Let $\mathbf{v} \in \mathbb{R}^S$. Then:

- If $\mathbf{v} \leq L(\mathbf{v})$ then $\mathbf{v} \leq \mathbf{v}^*_\lambda$
- If $\mathbf{v} \geq L(\mathbf{v})$ then $\mathbf{v} \geq \mathbf{v}^*_\lambda$
- If $\mathbf{v} = L(\mathbf{v})$ then $\mathbf{v} = \mathbf{v}^*_\lambda$ *(as a consequence of the previous assertions)*

Proof

Let $\mathbf{v} \leq L(\mathbf{v})$.

By definition, there is a decision rule d such that: $L(\mathbf{v}) = \mathbf{r}_d + \lambda \mathbf{P}_d \mathbf{v}$.

Thus:

$$\mathbf{v} - \lambda \mathbf{P}_d \mathbf{v} \leq \mathbf{r}_d$$

Applying the *non negative* matrix $(\mathbf{Id} - \lambda \mathbf{P}_d)^{-1}$ to the inequality yields:

$$\mathbf{v} \leq (\mathbf{Id} - \lambda \mathbf{P}_d)^{-1} \mathbf{r}_d = \mathbf{v}^{d^\infty} \leq \mathbf{v}^*_\lambda$$

Characterization of optimality (2)

Let $\mathbf{v} \geq L(\mathbf{v})$. Let $\boldsymbol{\pi} \stackrel{\text{def}}{=} (d_0, \ldots, d_n, \ldots)$ be a Markovian policy.
$\mathbf{v} \geq L(\mathbf{v}) \geq \mathbf{r}_{d_0} + \lambda \mathbf{P}_{d_0} \mathbf{v}$. By induction for $n \geq 1$,

$$\mathbf{v} \geq \sum_{0 \leq i < n} \lambda^i \left(\prod_{0 \leq j < i} \mathbf{P}_{d_j} \right) \mathbf{r}_{d_i} + \lambda^n \left(\prod_{0 \leq j < n} \mathbf{P}_{d_j} \right) \mathbf{v}$$

On the other hand,

$$\mathbf{v}_\lambda^{\boldsymbol{\pi}} = \sum_{i \in \mathbb{N}} \lambda^i \left(\prod_{0 \leq j < i} \mathbf{P}_{d_j} \right) \mathbf{r}_{d_i}$$

Let us define $B \stackrel{\text{def}}{=} \max(\max_s(|\mathbf{v}[s]|), \max_{s,a}(|r(s,a)|)$.
Then for all $s \in S$ and $n \in \mathbb{N}$:

$$\mathbf{v}[s] - \mathbf{v}_\lambda^{\boldsymbol{\pi}}[s] \geq -\lambda^n B (1 + \sum_{i \in \mathbb{N}} \lambda^i)$$

Letting n go to ∞, one gets: $\mathbf{v} \geq \mathbf{v}_\lambda^{\boldsymbol{\pi}}$. Since $\boldsymbol{\pi}$ is arbitrary, one obtains: $\mathbf{v} \geq \mathbf{v}_\lambda^*$.

Existence of a fixed-point

Let $\mathbf{v}$ and $\mathbf{v}'$ be two vectors.
Let d be a decision rule such that $L(\mathbf{v}) = \mathbf{r}_d + \lambda \mathbf{P}_d \mathbf{v}$.
Since $L(\mathbf{v}') \geq \mathbf{r}_d + \lambda \mathbf{P}_d \mathbf{v}'$:

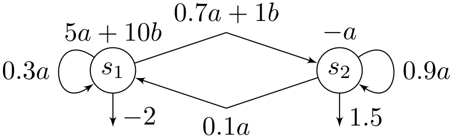
$$L(\mathbf{v})[s] - L(\mathbf{v}')[s] \leq \lambda \left(\mathbf{P}_d(\mathbf{v} - \mathbf{v}')\right)[s] \leq \lambda \|\mathbf{v} - \mathbf{v}'\|_\infty$$

Thus: $\|L(\mathbf{v}) - L(\mathbf{v}')\|_\infty \leq \lambda \|\mathbf{v} - \mathbf{v}'\|_\infty$
So L is Lipschitz-continuous with Lipschitz constant equal to $\lambda < 1$.

Using the Banach fixed-point theorem *(easy to prove)*,
given an arbitrary $\mathbf{v}_0$ and inductively defining $\mathbf{v}_{n+1} \stackrel{\text{def}}{=} L(\mathbf{v}_n)$.

- L admits a (unique) fixed-point equals to $\mathbf{v}^*_\lambda$
- $\lim_{n\to\infty} \mathbf{v}_n = \mathbf{v}^*_\lambda$
- For all n, $\|\mathbf{v}^*_\lambda - \mathbf{v}_n\|_\infty \leq \frac{\lambda^n}{1-\lambda} \|\mathbf{v}_1 - \mathbf{v}_0\|_\infty$

An example of convergence



Let $\lambda \stackrel{\text{def}}{=} \frac{1}{2}$ and $\mathbf{v}_0 \stackrel{\text{def}}{=} (0, 0)$.

Then:

$\mathbf{v}_1 = (10, -1)$

$\mathbf{v}_2 = (9.5, -0.95)$

$\mathbf{v}_3 = (9.525, -0.9525)$

$\ldots$

$\mathbf{v}^*_\lambda = (9.5238095238, -0.9523809524)$

Optimal policies (1)

Let d be a decision rule in D^{MD} that fulfills: $\mathbf{v}^*_\lambda = \mathbf{r}_d + \lambda \mathbf{P}_d \mathbf{v}^*_\lambda$.
Then d^∞ is an optimal policy since $\mathbf{v}^*_\lambda = (\mathbf{Id} - \lambda \mathbf{P}_d)^{-1} \mathbf{r}_d$.

Theorem. There exist $k \in \mathbb{N}$, $0 = \lambda_0 < \lambda_1 < \cdots < \lambda_k < \lambda_{k+1} = 1$ and $d_0, \ldots, d_k$ deterministic rules such that:

$$\forall 0 \leq i \leq k \ \forall \lambda \in [0,1[\qquad \lambda \in [\lambda_i, \lambda_{i+1}] \Rightarrow d_i^\infty \text{ is an optimal policy for } \lambda$$

Proof

Let d be an arbitrary deterministic decision rule.
Since $\mathbf{v}_\lambda^{d^\infty} = (\mathbf{Id} - \lambda \mathbf{P}_d)^{-1} \mathbf{r}_d$, every item of $\mathbf{v}_\lambda^{d^\infty}$ is a rational fraction of λ with poles outside $[0,1[$.
Let us consider $\mathbf{v}_x^{d^\infty}[s]$ as a function of x.
Define $Zero \stackrel{\text{def}}{=} \{\lambda \mid \exists d, d' \in D^{MD} \ \exists s \in S \ \mathbf{v}_x^{d^\infty}[s] \neq \mathbf{v}_x^{d'^\infty}[s] \wedge \mathbf{v}_\lambda^{d^\infty}[s] = \mathbf{v}_\lambda^{d'^\infty}[s]\}$
Then $Zero$ is finite.

Optimal policies (2)

Proof (continued)

Let $I \stackrel{\text{def}}{=}]a, b[$ be an interval such that $Zero \cap I = \emptyset$.

Pick an arbitrary $c \in I$ and let d be an optimal decision rule w.r.t. to c.

We claim that d is optimal for the whole interval I.

Otherwise, due to the continuity of $\mathbf{v}_x^{d^\infty}[s]$,
there should exist $\lambda \in I$, d' and s with $\mathbf{v}_x^{d^\infty}[s] \neq \mathbf{v}_x^{d'^\infty}[s] \wedge \mathbf{v}_\lambda^{d^\infty}[s] = \mathbf{v}_\lambda^{d'^\infty}[s]$.

Furthermore again by continuity d is also optimal at a and b (when $b \neq 1$).

So the appropriate decomposition of $[0, 1[$ is the one of $[0, 1[\setminus Zero$.

□

A policy $\boldsymbol{\pi}$ is *Blackwell optimal*
if there exists $0 \leq \lambda_0 < 1$ such that $\boldsymbol{\pi}$ is optimal for every $\lambda \in [\lambda_0, 1[$.

The theorem implies that
there exist deterministic stationary Blackwell optimal policies.

The value iteration algorithm

The value iteration algorithm implements the fixed-point approach while maintaining the current decision rule.

For $s \in S$ **do** $optval[s] \leftarrow 0$
Repeat
  $oldval \leftarrow optval$
  For $s \in S$ **do**
    $best \leftarrow -\infty$
    For $a \in A_s$
      $temp \leftarrow r(s, a)$
      For $s' \in S$ **do** $temp \leftarrow temp + \lambda p(s'|s, a) oldval[s']$
      If $best < temp$ **then** $best \leftarrow temp$; $optdec[s] \leftarrow a$
  $optval[s] \leftarrow best$
  $stop \leftarrow$ **true**
  For $s \in S$ **do If** $|optval[s] - oldval[s]| > \frac{\varepsilon(1-\lambda)}{2\lambda}$ **then** $stop \leftarrow$ **false**
Until stop

Why $\frac{\varepsilon(1-\lambda)}{2\lambda}$?

Criterium of convergence

Proposition. Let d be the decision rule computed by the algorithm. Then:

$$\|\mathbf{v}_\lambda^{d^\infty} - \mathbf{v}_\lambda^*\|_\infty \leq \varepsilon$$

Proof

Using Banach theorem, $\|\mathbf{v}_{n+1} - \mathbf{v}_\lambda^*\|_\infty \leq \frac{\lambda}{1-\lambda}\|\mathbf{v}_{n+1} - \mathbf{v}_n\|_\infty \leq \frac{\lambda}{1-\lambda}\frac{\varepsilon(1-\lambda)}{2\lambda} = \frac{\varepsilon}{2}$

$$\|\mathbf{v}_\lambda^{d^\infty} - \mathbf{v}_{n+1}\|_\infty \leq \|\mathbf{v}_\lambda^{d^\infty} - (\mathbf{r}_d + \lambda \mathbf{P}_d \mathbf{v}_{n+1})\|_\infty + \|(\mathbf{r}_d + \lambda \mathbf{P}_d \mathbf{v}_{n+1}) - \mathbf{v}_{n+1}\|_\infty$$

$$= \lambda\|\mathbf{P}_d\mathbf{v}_\lambda^{d^\infty} - \mathbf{P}_d\mathbf{v}_{n+1}\|_\infty + \lambda\|\mathbf{P}_d\mathbf{v}_{n+1} - \mathbf{P}_d\mathbf{v}_n\|_\infty \leq \lambda\|\mathbf{v}_\lambda^{d^\infty} - \mathbf{v}_{n+1}\|_\infty + \lambda\|\mathbf{v}_{n+1} - \mathbf{v}_n\|_\infty$$

So

$$\|\mathbf{v}_\lambda^{d^\infty} - \mathbf{v}_{n+1}\|_\infty \leq \frac{\lambda}{1-\lambda}\|\mathbf{v}_{n+1} - \mathbf{v}_n\|_\infty \leq \frac{\varepsilon}{2}$$

Thus:

$$\|\mathbf{v}_\lambda^{d^\infty} - \mathbf{v}_\lambda^*\|_\infty \leq \|\mathbf{v}_\lambda^{d^\infty} - \mathbf{v}_{n+1}\|_\infty + \|\mathbf{v}_{n+1} - \mathbf{v}_\lambda^*\|_\infty \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon$$

Principles of policy iteration

In the value iteration approach,
the current value is an approximation of the reward of the current policy.

Unlike value iteration approach,
the *policy iteration* approach maintains the exact reward of the current policy.

It tries to improve this reward using another decision rule.

More precisely, let d be the current decision rule.
Then a deterministic decision rule d' is chosen such that:

$$L(\mathbf{v}_\lambda^{d^\infty}) = \mathbf{r}_{d'} + \lambda \mathbf{P}_{d'} \mathbf{v}_\lambda^{d^\infty}$$

with d' equal to d if possible.

Properties of policy iteration

$\boxed{\text{If } d' = d \text{ then } d^\infty \text{ is an optimal policy.}}$

$$L(\mathbf{v}_\lambda^{d^\infty}) = \mathbf{r}_d + \lambda \mathbf{P}_d \mathbf{v}_\lambda^{d^\infty} = \mathbf{v}_\lambda^{d^\infty}$$

So $\mathbf{v}_\lambda^{d^\infty}$ is the optimal value and d is an optimal decision rule.

$\boxed{\text{If } d' \neq d \text{ then } \mathbf{v}_\lambda^{d'^\infty} > \mathbf{v}_\lambda^{d^\infty}}$

One has:

$$\mathbf{r}_{d'} + \lambda \mathbf{P}_{d'} \mathbf{v}_\lambda^{d^\infty} \geq \mathbf{r}_d + \lambda \mathbf{P}_d \mathbf{v}_\lambda^{d^\infty} = \mathbf{v}_\lambda^{d^\infty}$$

with at least one strict inequality.
Thus:

$$\mathbf{r}_{d'} \geq (\mathbf{Id} - \lambda \mathbf{P}_{d'}) \mathbf{v}_\lambda^{d^\infty}$$

Applying $(\mathbf{Id} - \lambda \mathbf{P}_{d'})^{-1}$ $(= \sum_{i \in N} (\lambda \mathbf{P}_{d'})^i)$

$$\mathbf{v}_\lambda^{d'^\infty} \geq \mathbf{v}_\lambda^{d^\infty}$$

Moreover since $(\mathbf{Id} - \lambda \mathbf{P}_{d'})^{-1} \geq \mathbf{Id}$, the strict inequality is preserved.

The policy iteration algorithm

For $s \in S$ **do** $optdec[s] \leftarrow$ some $a \in A_s$
Repeat
  $stop \leftarrow \mathbf{true}$
  For $s \in S$ **do**
    $\mathbf{rd}[s] \leftarrow r(s, optdec[s])$
    For $s' \in S$ **do**
      If $s = s'$ **then** $\mathbf{Md}[s, s'] \leftarrow 1 - \lambda p(s'|s, optdec[s])$
      Else $\mathbf{Md}[s, s'] \leftarrow -\lambda p(s'|s, optdec[s])$
  $optval \leftarrow \mathbf{LinearSolve}(\mathbf{Md}, \mathbf{rd})$
  For $s \in S$ **do**
    $best \leftarrow optval[s]$
    For $a \in A_s$ **do**
      $temp \leftarrow r(s, a)$
      For $s' \in S$ **do** $temp \leftarrow temp + \lambda p(s'|s, a) optval[s']$
      If $best < temp$ **then** $best \leftarrow temp$; $optdec[s] \leftarrow a$; $stop \leftarrow$ **false**
Until stop

Convergence of policy iteration

Termination.

Since there is a finite number of deterministic policies
and such a policy is never visited twice the algorithm terminates.

However this number is $\Omega(|A|^{|S|})$.

Comparison with value iteration.

Denote $\mathbf{v}_n$ (resp. $\mathbf{u}_n$) the reward
computed by policy (resp. value) iteration at the n^{th} iteration.

Denote dv_n (resp. du_n) the decision rule
corresponding to the n^{th} iteration of the policy (resp. value) iteration.

Assume that $\mathbf{v}_0 = \mathbf{u}_0$.

We claim that for all n, $\mathbf{v}_n \geq \mathbf{u}_n$.

$\mathbf{v}_{n+1} = \mathbf{r}_{dv_{n+1}} + \lambda \mathbf{P}_{dv_{n+1}} \mathbf{v}_{n+1} \geq \mathbf{r}_{dv_{n+1}} + \lambda \mathbf{P}_{dv_{n+1}} \mathbf{v}_n$
(since $\mathbf{v}_{n+1} \geq \mathbf{v}_n$)

$\mathbf{r}_{dv_{n+1}} + \lambda \mathbf{P}_{dv_{n+1}} \mathbf{v}_n \geq \mathbf{r}_{du_{n+1}} + \lambda \mathbf{P}_{du_{n+1}} \mathbf{v}_n$
(since $\mathbf{r}_{dv_{n+1}} + \lambda \mathbf{P}_{dv_{n+1}} \mathbf{v}_n = L(\mathbf{v}_n)$)

$\mathbf{r}_{du_{n+1}} + \lambda \mathbf{P}_{du_{n+1}} \mathbf{v}_n \geq \mathbf{r}_{du_{n+1}} + \lambda \mathbf{P}_{du_{n+1}} \mathbf{u}_n = \mathbf{u}_{n+1}$
(since $\mathbf{v}_n \geq \mathbf{u}_n$)

Principles of linear programming

A linear program is:

- the specification of an optimization problem;
- where both constraints and objective are expressed by linear expressions related to the variables of the problem.
- different equivalent formulations are possible: general, canonic or standard ones.

$$\texttt{Maximize } \mathbf{c}\cdot\mathbf{x} \texttt{ such that } \mathbf{A}\mathbf{x}=\mathbf{b}\wedge\mathbf{x}\geq 0$$

There are *a priori* three possible outputs:

- The set of feasible solutions is empty.
- The problem is unbounded, i.e. there exists a sequence of feasible solutions $\{\mathbf{x}_n\}$ such that $\lim_{n\to\infty}\mathbf{c}\cdot\mathbf{x}_n=\infty$.
- The problem admits an optimal value v, i.e. for all feasible solution $\mathbf{x}$, $\mathbf{c}\cdot\mathbf{x}\leq v$ and for all $\varepsilon>0$ there exists a feasible solution $\mathbf{x}$ with $\mathbf{c}\cdot\mathbf{x}\geq v-\varepsilon$. In this case, there exists an optimal solution.

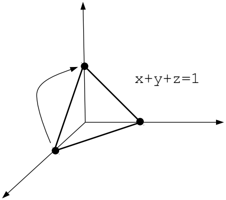
Solving a linear program

The simplex algorithm first decides whether the problem is empty or exhibits a *basic* solution:
a vertex of the polyhedron defined by the constraints.

The algorithm tries to improve a basic solution by selecting a neighbour of the current vertex.

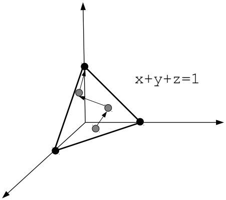
It stops when the solution is (locally) optimal or the problem is unbounded.

It performs well in practice but its worst case complexity is *exponential*.



The interior point approaches follow a path inside the polyhedron of solutions toward an optimal solution.

They are mathematically involved but perform in *polynomial time*.



In practice, whatever the algorithm
the number of constraints is the main factor of complexity.

The dual problem

Assume that we have a linear combination $\mathbf{y}$ of the row vectors of $\mathbf{A}$,

$$\mathbf{d} \stackrel{\text{def}}{=} \mathbf{yA} \left(= \sum_{i \in I} \mathbf{y}[i]\mathbf{A}[i, -] \right) \text{ such that } \mathbf{d} \geq \mathbf{c}$$

Then for all feasible solution $\mathbf{x}$,

$$\mathbf{c} \cdot \mathbf{x} \leq \mathbf{d} \cdot \mathbf{x} = \sum_{i \in I} \mathbf{y}[i](\mathbf{A}[i, -] \cdot \mathbf{x}) = \sum_{i \in I} \mathbf{y}[i]\mathbf{b}[i]$$

Otherwise stated, $\sum_{i \in I} \mathbf{y}[i]\mathbf{b}[i]$ is an upper bound of the optimal value.

$$\texttt{The dual problem} : \texttt{Minimize } \mathbf{y} \cdot \mathbf{b} \texttt{ such that } \mathbf{yA} \geq \mathbf{c} \wedge \mathbf{y} \in \mathbb{R}^I$$

Duality Theorem. Let **P** be a linear problem and **D** be its dual. Then:

- If **P** is unbounded then **D** does not admit a feasible solution.
- If **D** is unbounded then **P** does not admit a feasible solution.
- **P** admits an optimal solution if and only if **D** admits an optimal solution. In that case, the optimal values are equal.

A linear programming characterization

The previous characterization

- Any $\mathbf{v}$ that fulfills $\mathbf{v} \geq L(\mathbf{v})$ is an upper bound of $\mathbf{v}^*_\lambda$.
- $\mathbf{v}^*_\lambda$ also fulfills this inequation.

A linear programming reformulation

$$\text{Minimize } \sum_{s \in S} \alpha_s \mathbf{v}[s]$$

$$\text{subject to } \forall s \in S\ \forall a \in A_s\ \mathbf{v}[s] - \sum_{s' \in S} \lambda p(s'|s,a)\mathbf{v}[s'] \geq r(s,a)$$

- the variables are the components of vector $\mathbf{v}$.
- the α_s's are arbitrary constants that fulfill: $\forall s\ 0 < \alpha_s$ and $\sum_{s \in S} \alpha_s = 1$ *(this equality introduced only for probabilistic reasoning)*

The problem has $\sum_{s \in S} |A_s|$ constraints.

The dual characterization

Dual linear program

$$\text{Maximize} \sum_{s\in S}\sum_{a\in A_s} r(s,a)x(s,a)$$

$$\text{subject to } \forall s \in S \ \sum_{a\in A_s} x(s,a) - \sum_{s'\in S}\sum_{a\in A_{s'}} \lambda p(s|s',a)x(s',a) = \alpha_s$$

$$\forall s \in S \ \forall a \in A_s \ x(s,a) \geq 0$$

- The variables are the $x(s,a)$'s.
- Observation: a feasible solution fulfills for all s, $\sum_{a\in A_s} x(s,a) \geq \alpha_s > 0$.

The dual problem has $|S|$ constraints.

Decision rules and feasible solutions

- Let d be a Markovian decision rule. Then x_d is defined by:

$$x_d(s,a) \stackrel{\text{def}}{=} d(s)(a) \sum_{s' \in S} \alpha_{s'} \sum_{n \in \mathbb{N}} \lambda^n (\mathbf{P}_d)^n [s', s]$$

Probabilistic interpretation

- For all s, a, $x_d(s,a)$ is the average discounted number of times that action a is selected in state s knowing that the initial distribution is given by $\{\alpha_s\}$;
- $\sum_{s \in S} \sum_{a \in A_s} r(s,a) x_d(s,a)$ is the expected discounted reward of policy d^∞ knowing that the initial distribution is given by $\{\alpha_s\}$;
- For all s, $\sum_{a \in A_s} x_d(s,a) \geq \alpha_s > 0$.

x_d is a feasible solution of the dual linear program

- Let x be a feasible solution of the dual linear program.

Then the decision rule d_x is defined by by: $d_x(s)(a) \stackrel{\text{def}}{=} \frac{x(s,a)}{\sum_{a \in A_s} x(s,a)}$.

$$d_{x_d} = d \qquad \text{and} \qquad x_{d_x} = x$$

Plan

Different kinds of limits

Let $\{u_n\}_{n\in\mathbb{N}}$ be a sequence of reals (real vectors, real matrices, etc.). Then:

- $\{u_n\}_{n\in\mathbb{N}}$ is *Cesaro convergent* to a limit l if $\lim_{n\to\infty} \frac{1}{n+1}\sum_{i\leq n} u_i = l$. One denotes it by $u_n \to_c l$.
- $\{u_n\}_{n\in\mathbb{N}}$ is *Abel convergent* to a limit l if for all $0 \leq \lambda < 1$, $u(\lambda) \stackrel{\text{def}}{=} \sum_{n\in\mathbb{N}} u_n\lambda^n$ exists and $\lim_{\lambda\uparrow 1}(1-\lambda)u(\lambda) = l$. One denotes it by $u_n \to_a l$.

Observe the analogy with the discounted and average rewards.

Let $\{u_n\}_{n\in\mathbb{N}}$ be a sequence of reals.

- If $u_n \to l$ then $u_n \to_c l$.
- If $u_n \to_c l$ then $u_n \to_a l$.

Asymptotic behaviour of a finite DTMC

Let $\mathbf{P}$ be a stochastic matrix. Then $\{\mathbf{P}^n\}$ is Cesaro convergent to a stochastic matrix, denoted $\mathbf{P}^*$ and one has:

$$\mathbf{P}^*\mathbf{P} = \mathbf{P}\mathbf{P}^* = \mathbf{P}^*\mathbf{P}^* = \mathbf{P}^*$$

Proof. Let $\tilde{\mathbf{P}}_n \overset{\text{def}}{=} \frac{1}{n}\sum_{0\leq i<n} \mathbf{P}^i$ for $n > 0$.

$\tilde{\mathbf{P}}_n$ is a stochastic matrix thus the sequence $\{\tilde{\mathbf{P}}_n\}$ is bounded.

Pick a sequence of indices $n_0 < n_1 < \cdots$ such that $\mathbf{L} \overset{\text{def}}{=} \lim_{k\to\infty} \tilde{\mathbf{P}}_{n_k}$ exists.

$$\tilde{\mathbf{P}}_n\mathbf{P} = \mathbf{P}\tilde{\mathbf{P}}_n = \tilde{\mathbf{P}}_n + \frac{1}{n}(\mathbf{P}^n - \mathbf{Id})$$

Applying these equalities to n_k letting k go to ∞ yields: $\mathbf{LP} = \mathbf{PL} = \mathbf{L}$

Let $\mathbf{L}'$ be another limit of a subsequence of $\{\tilde{\mathbf{P}}_n\}$. Then: $\mathbf{PL}' = \mathbf{L}'\mathbf{P} = \mathbf{L}'$.
By iteration, $\mathbf{P}^n\mathbf{L}' = \mathbf{L}'\mathbf{P}^n = \mathbf{L}'$ for all n.
By linear combination, $\tilde{\mathbf{P}}_n\mathbf{L}' = \mathbf{L}'\tilde{\mathbf{P}}_n = \mathbf{L}'$ for all n.
Applying this equality for n_k and letting k go to ∞ yields $\mathbf{L}'\mathbf{L} = \mathbf{L}\mathbf{L}' = \mathbf{L}'$.
Swapping $\mathbf{L}$ and $\mathbf{L}'$ yields $\mathbf{L}\mathbf{L}' = \mathbf{L}'\mathbf{L} = \mathbf{L}$. Thus $\mathbf{L}' = \mathbf{L}$.

So $\tilde{\mathbf{P}}_n$ is convergent and the limit is stochastic. (why?)

Fundamental and deviation matrices

Let $\mathbf{P}$ be a stochastic matrix. Then $\mathbf{Id} - \mathbf{P} + \mathbf{P}^*$ is invertible and its inverse called the *fundamental matrix* and denoted $\mathbf{Z}$ fulfills:

$$\sum_{i=0}^{n} (\mathbf{P} - \mathbf{P}^*)^i \rightarrow_c \mathbf{Z}$$

The *deviation matrix* $\mathbf{D}$ is defined by $\mathbf{D} \overset{\text{def}}{=} \mathbf{Z} - \mathbf{P}^*$.

$\boxed{\text{Probabilistic interpretation in the aperiodic case}}$

- $\mathbf{P}^n \rightarrow \mathbf{P}^*$
- $\mathbf{P}^n - \mathbf{P}^* = (\mathbf{P} - \mathbf{P}^*)^n$ implying that the greatest module of eigenvalues of $\mathbf{P} - \mathbf{P}^*$ is smaller than 1.
- So $\mathbf{Z} = \mathbf{Id} + \sum_{n \geq 1} (\mathbf{P}^n - \mathbf{P}^*)$ and $\mathbf{D} = \sum_{n \in \mathbb{N}} (\mathbf{P}^n - \mathbf{P}^*)$

$\mathbf{D}[s, s']$ is the limit when n goes to ∞ of the difference between:

1. the mean number of visits of s' until time n starting from s;
2. the mean number of visits of s' until time n starting from the steady-state distribution reached when the initial state is s.

Properties of the deviation matrix

Let P be a stochastic matrix. Its deviation matrix D fulfills:

- $\mathbf{P}^*\mathbf{D} = \mathbf{D}\mathbf{P}^* = 0$
  (no deviation starting from a stationary distribution)
- $(\mathbf{Id} - \mathbf{P})\mathbf{D} = \mathbf{Id} - \mathbf{P}^*$
  (decomposing deviation between the initial and the remaining instants)

Application to the average reward.

Let d be a decision rule. Then the average reward of d^∞ is:

$$\mathbf{g}^{d^\infty} = \lim_{n \to \infty} \frac{1}{n} \sum_{i=0}^{n-1} \mathbf{P}_{\mathbf{d}}^{\mathbf{i}} \mathbf{r}_d = \mathbf{P}_{\mathbf{d}}^* \mathbf{r}_d$$

Define $\mathbf{h}^{d^\infty} \stackrel{\text{def}}{=} \mathbf{D}_{\mathbf{d}} \mathbf{r}_d$. Then:

$$\mathbf{g}^{d^\infty} = \mathbf{P}_{\mathbf{d}} \mathbf{g}^{d^\infty} \text{ and } \mathbf{g}^{d^\infty} + \mathbf{h}^{d^\infty} = \mathbf{P}_{\mathbf{d}} \mathbf{h}^{d^\infty} + \mathbf{r}_d$$

Characterization of optimality

1. Establish a condition for upper bounds and a conditional characterization

2. Relate average and discounted values

3. Prove that a Blackwell policy meets the characterization (using 1 and 2)

A first bound of the optimal value

Idea: Transforming $\mathbf{g}^{d^\infty} = \mathbf{P_d}\mathbf{g}^{d^\infty}$, $\mathbf{g}^{d^\infty} + \mathbf{h}^{d^\infty} = \mathbf{P_d}\mathbf{h}^{d^\infty} + \mathbf{r}_d$ into inequations.

Assume there exist two vectors $\mathbf{g}, \mathbf{h}$ over states such that for all $d \in D^{MD}$:

$$\mathbf{g} \geq \mathbf{P}_d\mathbf{g} \text{ and } \mathbf{g} + \mathbf{h} \geq \mathbf{P}_d\mathbf{h} + \mathbf{r}_d$$

Then: $\mathbf{g} \geq \mathbf{g}_+^*$.

Proof. Let $\boldsymbol{\pi} = (d_1, d_2, \ldots)$ be a Markovian policy. Then:

$$\mathbf{g} \geq \mathbf{r}_{d_k} + (\mathbf{P}_{d_k} - \mathbf{Id})\mathbf{h}$$

Then one applies the first inequation with d_{k-1} getting:

$$\mathbf{g} \geq \mathbf{P}_{d_{k-1}}\mathbf{g} \geq \mathbf{P}_{d_{k-1}}\mathbf{r}_{d_k} + \mathbf{P}_{d_{k-1}}(\mathbf{P}_{d_k} - \mathbf{Id})\mathbf{h}$$

Applying iteratively the first inequation with $\mathbf{P}_{d_{k-2}}, \ldots, \mathbf{P}_{d_1}$ one obtains:

$$\mathbf{g} \geq \mathbf{P}_{d_1} \ldots \mathbf{P}_{d_{k-1}}\mathbf{r}_{d_k} + \mathbf{P}_{d_1} \ldots \mathbf{P}_{d_{k-1}}(\mathbf{P}_{d_k} - \mathbf{Id})\mathbf{h}$$

Summing this inequation for k from 1 to n, one gets:

$$n\mathbf{g} \geq \mathbf{v}_n^{\boldsymbol{\pi}} + (\mathbf{P}_{d_1} \ldots \mathbf{P}_{d_{n-1}}\mathbf{P}_{d_n} - \mathbf{Id})\mathbf{h}$$

Since the last term is bounded by $\|h\|$, dividing by n and letting n go to ∞ yields:

$$\mathbf{g} \geq \limsup_{n\to\infty} \frac{1}{n}\mathbf{v}_n^{\boldsymbol{\pi}} = \mathbf{g}_+^{\boldsymbol{\pi}}$$

Refining the bound

Assume there exists two vectors $\mathbf{g}, \mathbf{h}$ such that for all $d \in D^{MD}$, for all $s \in S$:

- either $\mathbf{g}[s] > \sum_{s' \in S} \mathbf{P}_d[s, s']\mathbf{g}[s']$
- or $\mathbf{g}[s] = \sum_{s' \in S} \mathbf{P}_d[s, s']\mathbf{g}[s'] \wedge \mathbf{g}[s] + \mathbf{h}[s] \geq \sum_{s' \in S} \mathbf{P}_d[s, s']\mathbf{h}[s'] + \mathbf{r}_d[s]$

Then $\mathbf{g} \geq \mathbf{g}^*_+$.

Proof. Let $\mathbf{g}, \mathbf{h}$ be a solution of this system.

We claim that $\mathbf{g}, \mathbf{h} + M\mathbf{g}$ for M large enough fulfil the previous hypotheses.

Consider the possibly unsatisfied equation:

$$\mathbf{g}(s) + (\mathbf{h}[s] + M\mathbf{g}[s]) \overset{?}{\geq} \sum_{s' \in S} \mathbf{P}_d[s, s'](\mathbf{h}[s'] + M\mathbf{g}[s']) + \mathbf{r}_d[s]$$

for which $\mathbf{g}[s] > \sum_{s' \in S} \mathbf{P}_d[s, s']\mathbf{g}[s']$

- $M\mathbf{g}[s]$ occurs on the left side.
- $\sum_{s' \in S} \mathbf{P}_d[s, s']M\mathbf{g}[s']$ occurs on the right side.
- So there exists M large enough that satisfies such an equation.

A conditional characterization

Assume that $\mathbf{g}$ and $\mathbf{h}$ fulfill:

$$\forall s \in S \ \mathbf{g}[s] = \max_{a \in A_s} \left(\sum_{s' \in S} p(s'|s,a)\mathbf{g}[s'] \right)$$

$$\forall s \in S \ \mathbf{g}[s] + \mathbf{h}[s] = \max_{a \in B_s} \left(\sum_{s' \in S} p(s'|s,a)\mathbf{h}[s'] + r(s,a) \right)$$

$$\text{where } B_s \stackrel{\text{def}}{=} \arg\max_{a \in A_s} \left(\sum_{s' \in S} p(s'|s,a)\mathbf{g}[s'] \right)$$

Then $\mathbf{g} = \mathbf{g}^*_+ = \mathbf{g}^*_-$ and it is obtained by a stationary policy.

Proof of the conditional characterization

$(\mathbf{g}, \mathbf{h})$ fulfills the requirements to be a bound. So: $\mathbf{g} \geq \mathbf{g}_+^*$.

Define d by choosing some optimal $d(s) \in B_s$.
The equation system can be rewritten:

$$\mathbf{g} = \mathbf{P}_d \mathbf{g} \text{ and } \mathbf{g} + \mathbf{h} = \mathbf{P}_d \mathbf{h} + \mathbf{r}_d$$

Using the second equation, one gets: $\mathbf{g} = \mathbf{r}_d + (\mathbf{P}_d - \mathbf{Id})\mathbf{h}$

Applying the first equation: $\mathbf{g} = \mathbf{P}_d \mathbf{g} = \mathbf{P}_d \mathbf{r}_d + \mathbf{P}_d(\mathbf{P}_d - \mathbf{Id})\mathbf{h}$

By iteration: $\mathbf{g} = \mathbf{P}_d^k \mathbf{r}_d + \mathbf{P}_d^{k-1}(\mathbf{P}_d - \mathbf{Id})\mathbf{h}$

Summing, one gets: $n\mathbf{g} = \mathbf{u}_n^{d^\infty} + (\mathbf{P}_d^n - \mathbf{Id})\mathbf{h}$

Since the last term is bounded by $\|\mathbf{h}\|$, dividing by n and letting n go to ∞ yields:

$$\mathbf{g} = \lim_{n \to \infty} \frac{1}{n} \mathbf{u}_n^{d^\infty} = \mathbf{g}_+^{d^\infty} = \mathbf{g}_-^{d^\infty}$$

Relating average and discounted values

A limit relation

Let $d \in D^{MD}$, then:

$$\mathbf{g}_-^{d^\infty} = \mathbf{g}_+^{d^\infty} = \mathbf{P}_d^* \mathbf{r}_d = \lim_{\lambda \uparrow 1} (1-\lambda) \mathbf{v}_\lambda^{d^\infty} \stackrel{\text{def}}{=} \mathbf{g}^{d^\infty}$$

due to Cesaro (and so Abel) convergence towards $\mathbf{P}_d^*$

The exact relation

Let us define $\rho \stackrel{\text{def}}{=} \frac{1-\lambda}{\lambda}$ and assume that $\frac{\|\mathbf{D}_d\|}{1+\|\mathbf{D}_d\|} < \lambda < 1$ (so $\rho \|\mathbf{D}_d\| < 1$) then:

$$\mathbf{v}_\lambda^{d^\infty} = \frac{1}{1-\lambda} \left(\mathbf{P}_d^* \mathbf{r}_d - \sum_{n=1}^{\infty} (-\rho \mathbf{D}_d)^n \mathbf{r}_d \right)$$

since the right hand term fulfills equation $(\mathbf{Id} - \lambda \mathbf{P}_d) \mathbf{X} = \mathbf{r}_d$
whose single solution is $\mathbf{v}_\lambda^{d^\infty}$ *(using properties of* $\mathbf{P}_d^*$ *and* $\mathbf{D}_d$*).*

The first-order relation

$$\mathbf{v}_\lambda^{d^\infty} = \frac{1}{1-\lambda} \mathbf{P}_d^* \mathbf{r}_d + \mathbf{D}_d \mathbf{r}_d + O(1-\lambda)$$

Existence of optimal policies

Let d^∞ be (Blackwell) optimal for $\lambda \in [\lambda_0, 1[$. Then $(\mathbf{P}_d^*\mathbf{r}_d, \mathbf{D}_d\mathbf{r}_d)$ fulfills the characterization and $\mathbf{g}_{d^\infty} = \mathbf{P}_d^*\mathbf{r}_d$ is the optimal value.

Proof.

By optimality: $\forall s \in S \; \forall a \in A_s \; \mathbf{v}_\lambda^{d^\infty}[s] \geq r(s,a) + \lambda \sum_{s' \in S} p(s'|s,a)\mathbf{v}_\lambda^{d^\infty}[s']$

- Using first-order development one gets:

$$\frac{1}{1-\lambda}\left((\mathbf{P}_d^*\mathbf{r}_d)[s] - \sum_{s' \in S} p(s'|s,a)(\mathbf{P}_d^*\mathbf{r}_d)[s']\right) + (\mathbf{D}_d\mathbf{r}_d)[s] - r(s,a) - \sum_{s' \in S} p(s'|s,a)(\mathbf{D}_d\mathbf{r}_d - \mathbf{P}_d^*\mathbf{r}_d)[s'] + O(1-\lambda) \geq 0$$

- So: $(\mathbf{P}_d^*\mathbf{r}_d)[s] - \sum_{s' \in S} p(s'|s,a)(\mathbf{P}_d^*\mathbf{r}_d)[s'] \geq 0$

- When equality holds:

$$(\mathbf{D}_d\mathbf{r}_d)[s] - r(s,a) - \sum_{s' \in S} p(s'|s,a)(\mathbf{D}_d\mathbf{r}_d - \mathbf{P}_d^*\mathbf{r}_d)[s'] \geq 0$$

Implying: $(\mathbf{D}_d\mathbf{r}_d)[s] - r(s,a) - \sum_{s' \in S} p(s'|s,a)(\mathbf{D}_d\mathbf{r}_d)[s'] + (\mathbf{P}_d^*\mathbf{r}_d)[s] \geq 0$

Policy iteration: principles

As seen for the discounted reward, the policy approach is based on two key items.

- Computing values provided by a stationary policy d^∞.
  Here we are going to compute:
  1. the reward $\mathbf{P}_d^*\mathbf{r}_d$;
  2. the second term of the above Taylor development $\mathbf{D}_d\mathbf{r}_d$.
- Designing a rule that:
  1. either identifies an optimal stationary policy;
  2. or provides a way to improve it.

Values associated with a policy

Let d be a decision rule and consider the following equation system where the variables are vectors $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$.

$$\begin{aligned} (\mathbf{Id} - \mathbf{P}_d)\mathbf{x} &= \mathbf{0} && (1) \\ \mathbf{x} + (\mathbf{Id} - \mathbf{P}_d)\mathbf{y} &= \mathbf{r}_d && (2) \\ \mathbf{y} + (\mathbf{Id} - \mathbf{P}_d)\mathbf{z} &= \mathbf{0} && (3) \end{aligned}$$

Then:

- Vectors $\mathbf{P}_d^*\mathbf{r}_d$, $\mathbf{D}_d\mathbf{r}_d$ and $-\mathbf{D}_d^2\mathbf{r}_d$ are solutions of this system.
- Any $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ solution of this system fulfills $\mathbf{x} = \mathbf{P}_d^*\mathbf{r}_d$ and $\mathbf{y} = \mathbf{D}_d\mathbf{r}_d$.

Thus one computes $\mathbf{P}_d^*\mathbf{r}_d$ and $\mathbf{D}_d\mathbf{r}_d$ in polynomial time.

Correctness of the equation system

Let us check that $\mathbf{P}_d^*\mathbf{r}_d$, $\mathbf{D}_d\mathbf{r}_d$ and $-\mathbf{D}_d^2\mathbf{r}_d$ are solutions of this system.

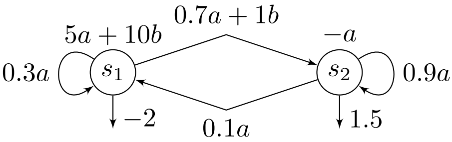
- $(\mathbf{Id} - \mathbf{P}_d)\mathbf{P}_d^*\mathbf{r}_d = (\mathbf{P}_d^* - \mathbf{P}_d^*)\mathbf{r}_d = \mathbf{0}$
- $\mathbf{P}_d^*\mathbf{r}_d + (\mathbf{Id} - \mathbf{P}_d)\mathbf{D}_d\mathbf{r}_d = \left(\mathbf{P}_d^* + (\mathbf{Id} - \mathbf{P}_d)\mathbf{D}_d\right)\mathbf{r}_d = \mathbf{r}_d$
- $\mathbf{D}_d\mathbf{r}_d - (\mathbf{Id} - \mathbf{P}_d)\mathbf{D}_d^2\mathbf{r}_d = \left(\mathbf{Id} - (\mathbf{Id} - \mathbf{P}_d)\mathbf{D}_d\right)\mathbf{D}_d\mathbf{r}_d = \mathbf{P}_d^*\mathbf{D}_d\mathbf{r}_d = \mathbf{0}$

Let $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ be a solution of this system.
From (1), $\mathbf{P}_d\mathbf{x} = \mathbf{x}$ which entails $\mathbf{P}_d^*\mathbf{x} = \mathbf{x}$.
So: $\mathbf{x} = \mathbf{P}_d^*\mathbf{x} = \mathbf{P}_d^*\mathbf{r}_d - \mathbf{P}_d^*(\mathbf{Id} - \mathbf{P}_d)\mathbf{y} = \mathbf{P}_d^*\mathbf{r}_d$ using (2)

$\mathbf{0} = \mathbf{P}_d^*\left(\mathbf{y} + (\mathbf{Id} - \mathbf{P}_d)\mathbf{z}\right) = \mathbf{P}_d^*\mathbf{y}$ using (3)

Thus using second equation of the system:
$\mathbf{r}_d - \mathbf{P}_d^*\mathbf{r}_d = (\mathbf{Id} - \mathbf{P}_d)\mathbf{y} = (\mathbf{Id} - \mathbf{P}_d + \mathbf{P}_d^*)\mathbf{y}$ which can be rewritten as:
$\mathbf{y} = (\mathbf{Id} - \mathbf{P}_d + \mathbf{P}_d^*)^{-1}(\mathbf{Id} - \mathbf{P}_d^*)\mathbf{r}_d = (\mathbf{D}_d + \mathbf{P}_d^*)(\mathbf{Id} - \mathbf{P}_d^*)\mathbf{r}_d = \mathbf{D}_d\mathbf{r}_d$

Illustration



Let us study the (already described) policies d and d'.

$$\mathbf{Id} - \mathbf{P}_d = \begin{pmatrix} 1 & -1 \\ -0.1 & 0.1 \end{pmatrix} \text{ and } \mathbf{Id} - \mathbf{P}_{d'} = \begin{pmatrix} 0.7 & -0.7 \\ -0.1 & 0.1 \end{pmatrix}$$

The range of $\mathbf{Id} - \mathbf{P}_d$ is $\alpha(1, -0.1)$. So $\mathbf{x} = \alpha(1, -0.1) + (10, -1)$ for some α.
Furthermore $\mathbf{x}$ is in the kernel of $\mathbf{Id} - \mathbf{P}_d$.
So we get $\alpha + 10 = -0.1\alpha - 1$ yielding $\alpha = -10$ and $\mathbf{x} = (0, 0)$.

The range of $\mathbf{Id} - \mathbf{P}_{d'}$ is $\alpha(0.7, -0.1)$. So $\mathbf{x} = \alpha(0.7, -0.1) + (5, -1)$ for some α.
Furthermore $\mathbf{x}$ is in the kernel of $\mathbf{Id} - \mathbf{P}_{d'}$.
So we get $0.7\alpha + 5 = -0.1\alpha - 1$ yielding $\alpha = -\frac{15}{2}$ and $\mathbf{x} = (-\frac{1}{4}, -\frac{1}{4})$.

Improving a policy

Let d be a decision rule and s be a state. Define:

$$\text{Improve}(d,s) \stackrel{\text{def}}{=} \{a \in A_s \mid (\mathbf{P}_d^*\mathbf{r}_d)[s] < \sum_{s' \in S} p(s'|s,a)(\mathbf{P}_d^*\mathbf{r}_d)[s']\}$$

$$\cup \{a \in A_s \mid (\mathbf{P}_d^*\mathbf{r}_d)[s] = \sum_{s' \in S} p(s'|s,a)(\mathbf{P}_d^*\mathbf{r}_d)[s']$$

$$\wedge ((\mathbf{P}_d^* + \mathbf{D}_d)\mathbf{r}_d)[s] < r(s,a) + \sum_{s' \in S} p(s'|s,a)(\mathbf{D}_d\mathbf{r}_d)[s']\}$$

Then if for all s, Improve$(d,s) = \emptyset$ then d^∞ is average optimal.

Otherwise let d' be any policy such that for all s,

1. Improve$(d,s) = \emptyset$ implies $d'(s) = d(s)$;
2. Improve$(d,s) \neq \emptyset$ implies $d'(s) \in$ Improve(d,s).

Then $\mathbf{P}_d^*\mathbf{r}_d \leq \mathbf{P}_{d'}^*\mathbf{r}_{d'}$ and there exists λ_0 such that for all $\lambda_0 < \lambda$, $\mathbf{v}_\lambda^{d^\infty} < \mathbf{v}_\lambda^{d'^\infty}$.

The proof of improvement is based on the first-order development and the analysis of policy $\boldsymbol{\pi} \stackrel{\text{def}}{=} (d', d, d, \ldots)$.

Linear programming

Using bounding results, for every pair of vectors $(\mathbf{g}, \mathbf{h})$ such that for all $d \in D^{MD}$, $\mathbf{g} \geq \mathbf{P}_d\mathbf{g}$ and $\mathbf{g} + \mathbf{h} \geq \mathbf{P}_d\mathbf{h} + \mathbf{r}_d$ one gets: $\mathbf{g} \geq \mathbf{g}^*$.

For any Blackwell optimal policy d^∞, $(\mathbf{P}_d^*\mathbf{r}_d, \mathbf{D}_d\mathbf{r}_d + M\mathbf{P}_d^*\mathbf{r}_d)$ is a solution of such a system as soon as M is large enough.

Thus the following linear program has its g component equal to the optimal expected average reward.

Primal Linear Program

$$\text{Minimize } \sum_{s \in S} \alpha_s \mathbf{g}[s] \text{ subject to } \forall s \in S\ \forall a \in A_s,$$

$$\mathbf{g}[s] - \sum_{s' \in S} p(s'|s,a)\mathbf{g}[s'] \geq 0 \text{ and } \mathbf{g}[s] + \mathbf{h}[s] - \sum_{s' \in S} p(s'|s,a)\mathbf{h}[s'] \geq r(s,a)$$

The variables are vectors $\mathbf{g}$ and $\mathbf{h}$ while the α_s's are positive constants.

As for the discounted case, solving the dual program is preferred.