

An Inverse Method for Policy-Iteration Based Algorithms*

Étienne André

LSV, ENS Cachan & CNRS
Cachan, France

andre@lsv.ens-cachan.fr

Laurent Fribourg

LSV, ENS Cachan & CNRS
Cachan, France

fribourg@lsv.ens-cachan.fr

We present an extension of two policy-iteration based algorithms on weighted graphs (viz., Markov Decision Problems and Max-Plus Algebras). This extension allows us to solve the following *inverse problem*: considering the weights of the graph to be unknown constants or *parameters*, we suppose that a reference instantiation of those weights is given, and we aim at computing a constraint on the parameters under which an optimal policy for the reference instantiation is still optimal. The original algorithm is thus guaranteed to behave well around the reference instantiation, which provides us with some criteria of robustness. We present an application of both methods to simple examples. A prototype implementation has been done.

1 Introduction

We consider the *inverse problem* initially defined in the context of timed models (more precisely, Timing Constraint Graphs [5] and Timed Automata [1]). We present here this problem in the context of systems modeled by directed graphs with (parametric) weights associated to their edges, and more specifically in the cases of Markov Decision Processes (MDPs) [2, 6] and Max-Plus Algebras [3].

Let us first present the *direct* problem in this context. The model is given under the form of a directed graph G , with weights that are unknown constants or *parameters*. We also assume that a *reference instantiation* π_0 is given for these parametric weights. Roughly speaking, a *policy* is a function which associates with each state of the graph an action which goes from the state to (a set of) successor state(s). Each action has a specific weight. The weight of a *path* (or sequence of actions) is the sum of the weights of its constitutive actions. The *value* (or cost) of a given policy μ for a given state s corresponds to the mean weight of the paths induced by μ , which go from s to a final state of the graph. Given a specific instantiation π_0 of the parameters, the direct problem consists in computing an *optimal* policy, that is a policy which gives the *minimal value* (or *maximal value*) when the parameters are instantiated with π_0 .

The optimal policy is classically found using the method of *policy iteration* (PI) (see [6]). The corresponding value is then computed by the *value determination* procedure (VD) (see, e.g., [3]). We show in this paper that the inverse problem can be simply posed, and solved via a natural generalization of the procedures of policy iteration and value determination. We focus here on two classes of models: Markov Decision Processes and Max-Plus Algebras.

Given a reference valuation π_0 , the inverse algorithm generalizes the direct algorithm “around” π_0 , and infers a constraint on the parameters guaranteeing a similar behavior as under π_0 . This ensures that the original algorithm continues to behave well around π_0 , thus giving some criteria of *robustness*.

We first give the general framework of our method (Sect. 2). We then present the adaptation of the inverse method to Markov Decision Processes (Sect. 3) and Max-Plus Algebras (Sect. 4). We conclude by giving some final remarks (Sect. 5).

*This work is partially supported by the Agence Nationale de la Recherche, grant ANR-06-ARFU-005, and by Institut Farman (ENS Cachan).

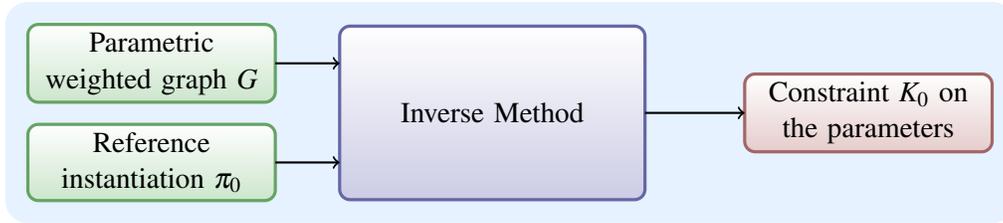


Figure 1: Our generic framework

2 General Framework

2.1 Preliminaries

Throughout this paper, we assume a fixed set $P = \{p_1, \dots, p_N\}$ of *parameters*. A *parameter instantiation* π is a function $\pi : P \rightarrow \mathbb{R}$ assigning a real constant to each parameter. There is a one-to-one correspondence between instantiations and points in \mathbb{R}^N . We will often identify an instantiation π with the point $(\pi(p_1), \dots, \pi(p_N))$.

Definition 1 A linear inequality on the parameters P is an inequality $e \prec e'$, where $\prec \in \{<, \leq\}$, and e, e' are two linear terms of the form

$$\sum_i \alpha_i p_i + d$$

where $1 \leq i \leq N$, $\alpha_i \in \mathbb{R}$ and $d \in \mathbb{R}$.

A (convex) constraint on the parameters P is a conjunction of inequalities on P .

We say that a parameter instantiation π *satisfies* a constraint K on the parameters, denoted by $\pi \models K$, if the expression obtained by replacing each parameter p in K with $\pi(p)$ evaluates to true. We will consider *True* as a constraint on the parameters, corresponding to the set of all possible instances for P .

2.2 Overview of the Inverse Method

We assume given a weighted graph, and an algorithm *PI* of policy iteration. We define a *parametric* version of the weighted graph, i.e., a weighted graph whose weights are unknown constants, or *parameters*. Given a parametric weighted graph G and an instantiation π of the parameters, we denote by $G[\pi]$ the (standard) weighted graph, where the parameters p_i have been replaced by their instance $\pi(p_i)$. For a given graph G , a given reference instance π_0 of the parameters, and an optimal policy μ_0 found by *PI* for $G[\pi_0]$, our goal is to generate a constraint K_0 on the parameters such that:

1. $\pi_0 \models K_0$, and
2. μ_0 is optimal for $G[\pi]$, for any instantiation π satisfying K_0 (i.e., $\pi \models K_0$).

A trivial solution is $K_0 = \{\pi_0\}$. However, our method will always generate something more general than $K_0 = \{\pi_0\}$, under the form of a conjunction of inequalities on the parameters (without any constant, apart from 0). Given *PI*, the framework of our inverse method is given in Fig. 1. Given an algorithm *PI* of policy iteration from the literature, calling itself an algorithm *VD* of value determination, our approach can be summarized as follows:

1. Compute an optimal policy μ_0 for the (standard) weighted graph $G[\pi_0]$, using *PI*;

2. Compute a *generic value* (or generic cost) corresponding to G for the policy μ_0 , using a parameterized version of VD ;
3. From the generic value computed above, infer a constraint K_0 such that μ_0 is optimal for $G[\pi]$, for any instantiation π satisfying K_0 .

We now present such an inverse method in the case of two policy-based iteration algorithms.

3 Markov Decision Processes

3.1 Preliminaries

We consider in this section *Markov Decision Processes* [2] as an extension of weighted labeled directed graphs. We associate to every edge of the graph a *probability* such that, for a given state and a given action (or label), the sum of the probabilities of the edges leaving this state through this action is equal to 1. Markov Decision Processes are widely used to model, e.g., the power consumption of devices (see, e.g., [8]). Formally:

Definition 2 A Markov Decision Process (MDP) is a tuple $M = (S, A, Prob, w)$, where

- $S = \{s_1, \dots, s_n\}$ is a set of states;
- A is a set of actions (or labels);
- $Prob : S \times A \times S \rightarrow [0, 1]$ is a probability function such that $Prob(s_1, a, s_2)$ (which will also be denoted by $Prob_a(s_1, s_2)$) is the probability that action a in state s_1 will lead to state s_2 , and $\forall s \in S, \forall a \in A : \sum_{s' \in S} Prob_a(s, s') = 1$;
- $w : S \times A \rightarrow \mathbb{R}$ is a weight function such that $w(s, a)$ (also denoted by $w_a(s)$) is the weight associated to the action a when leaving s .

In the following, we consider the MDP $M = (S, A, Prob, w)$. Given a state $s \in S$, we denote by $e(s)$ (for *enabled*) the set of possible actions for s , i.e., $\{a \in A \mid \exists s' \in S : Prob_a(s, s') > 0\}$. We suppose that, for any state $s \in S$, $e(s) \neq \emptyset$. We also suppose that M has a *unique* “absorbing state”, i.e., a state which is reachable (with positive probability) from any other state for any policy, and which has a self-loop outgoing transition with weight 0 and probability 1. We suppose in the following that the absorbing state is s_n . For the sake of simplicity, we will not depict, in the graphs describing MDPs in this paper, the self-loop outgoing transition of the absorbing state.

In every state s of $S \setminus \{s_n\}$, we can choose *non-deterministically* an action a in $e(s)$. Then, for this action, the system will evolve to a state s' such that $Prob_a(s, s') > 0$. A way of removing non-determinism from an MDP is to introduce a *policy* μ , i.e., a function from states to actions. A policy is of the form $\mu = \{s_1 \rightarrow a_{s_1}, s_2 \rightarrow a_{s_2}, \dots, s_{n-1} \rightarrow a_{s_{n-1}}\}$. We denote by $\mu[s]$ the action associated to state s . The MDP, associated to a policy, behaves as a *Markov chain* [7].

Given a policy μ , the *associated value* is a function mapping each state s to the mean sum of weights attached to the paths induced by μ , which go from s to s_n . (By convention, the value associated to s_n is null.) A classical problem for MDPs is to find an *optimal policy*, i.e., a policy under which the value function is maximum (or minimum), for every $s \in S$. Note that, under the assumption of the existence of an absorbing state, such an optimal policy always exists, but is not necessarily unique (see, e.g., [6]). We focus here on finding an optimal policy for which the value function is *minimal*.

We give in Fig. 12 in Appendix A the classical algorithm *mdpPI* for policy iteration on MDPs. This algorithm computes the optimal policy for an MDP, and it makes use of the algorithm *mdpVD* for value

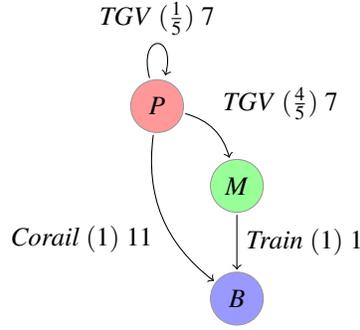


Figure 2: An example of Markov Decision Process

determination in MDPs (see Fig. 11 in Appendix A). Given an MDP and a policy, this second algorithm computes the mean sum of weights attached to the paths reaching s_n , for every starting state in $S \setminus \{s_n\}$. We denote by $v[s]$ the value associated to state s . The value v computed by Algorithm *mdpVD* is obtained by solving a system of linear equations, and is computed by applying the inverse of a real-valued matrix to a parametric vector. The fact that there is a single solution to this system is due to the invertibility of the matrix, which comes itself from the existence of an absorbing state.

3.2 An Illustrating Example

Consider the case of a researcher getting by train from Paris to Bologna. He can either take a night train *Corail*, or use the French high-speed train TGV. When there is no strike impacting the TGV service, the TGV usually needs 7 hours to go from Paris to Milan (with probability $4/5$). It is then possible to take an Italian train, reaching Bologna from Milan in 1 hour with probability 1. However, in case of strike (with probability $1/5$), the TGV does not leave Paris, and the researcher should wait 7 more hours until the next TGV. Note that this next TGV may also be on strike (with the same probability $1/5$), and so on. The night train can not be impacted by any strike, and it goes directly from Paris to Bologna in 11 hours with probability 1.

The MDP depicted in Fig. 2 summarizes those different possibilities, where P stands for Paris, M for Milan and B for Bologna. We denote by “*TGV* ($4/5$) 7” a transition using label *TGV* with probability $4/5$ and weight 7 (i.e., 7 hours). Note that the only source of non-determinism is in state Paris (P), where it is possible to choose between the *TGV* and the *Corail* actions.

We are first interested in the following question: considering the probability of strike, what is the best option, i.e., should we use the TGV or the night train? This problem corresponds to finding an *optimal policy* for this MDP, i.e., a policy minimizing the global weight of the system w.r.t. the probabilities. An application of the (standard) algorithm *mdpPI* [6] (see Fig. 12 in Appendix A) to the MDP modeling the train journey from Paris to Bologna gives the following optimal policy: $\mu = \{P \rightarrow \textit{TGV}, M \rightarrow \textit{Train}\}$ ¹. For this policy, the value for state P (given by the last call to Algorithm *mdpVD*), i.e., the expected time to reach Bologna, is 9.75.

We now suppose that the train between Milan and Bologna can be subject to delays due, e.g., to some works on the track. Our problem is the following: until which delay of the train between Milan and Bologna the option “TGV” in Paris remains the best option? In other words, until which delay of the train between Milan and Bologna the optimal policy remains optimal? We are thus interested in computing a

¹ As B is the absorbing state, recall that we do not define a policy for it.

constraint on all the delays of the system, viewed as parameters, such that, for any instantiation of this constraint, the policy μ remains the optimal policy for this MDP.

3.3 The Algorithm *P-mdpPI*

We first adapt the notion of MDP to the parametric case. We now consider that the weights of the MDP are *parameters*.

Definition 3 Given a set P of parameters, a Parametric Markov Decision Process (PMDP) is a tuple $M = (S, A, Prob, W)$, where

- $S = \{s_1, \dots, s_n\}$ is a set of states;
- A is a set of actions;
- $Prob : S \times A \times S \rightarrow [0, 1]$ is a probability function such that $Prob(s_1, a, s_2)$ (which will also be denoted by $Prob_a(s_1, s_2)$) is the probability that action a in state s_1 will lead to state s_2 , and $\forall s \in S, \forall a \in A : \sum_{s' \in S} Prob_a(s, s') = 1$;
- $W : S \times A \rightarrow P$ is a parametric weight function such that $W(s, a)$ (also denoted by $W_a(s)$) is a parameter associated to the action a when leaving s .

We consider in the following the PMDP $M = (S, A, Prob, W)$. Given an instantiation π of the parameters, we denote by $W[\pi]$ the function from $S \times A$ to \mathbb{R} obtained by replacing each occurrence of a parameter p_i in W with the value $\pi(p_i)$, for $1 \leq i \leq N$. By extension, we denote by $M[\pi]$ the (standard) MDP $(S, A, Prob, W[\pi])$.

We first introduce the algorithm *P-mdpVD*, given in Fig. 3, which computes, given a policy μ , the parametric value associated to every state s (i.e., the mean sum of the parametric weights of paths induced by μ going from s to s_n). This algorithm is a straightforward adaptation to the parametric case of the classical algorithm *mdpVD* of value determination for MDPs (see Fig. 11 in Appendix A). We denote by $V[s]$ the parametric value associated to state s .

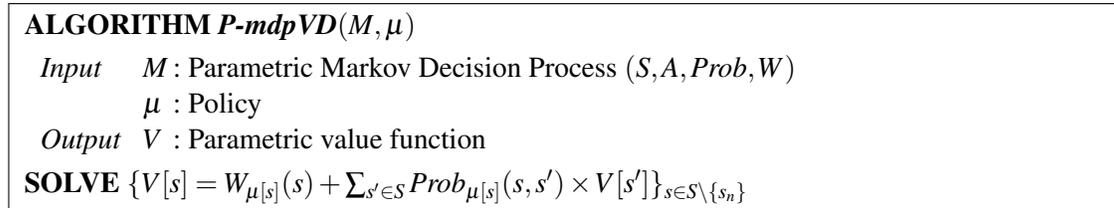


Figure 3: Algorithm for parametric value determination for MDPs

The value V computed by this algorithm *P-mdpVD* is obtained by solving a system of linear equations. Since this system is of the form $V = A \times V + B$, it is equivalent to $V = (1 - A)^{-1} \times B$, and can be implemented using the inversion of matrix $(1 - A)$. Note that this matrix A is computed from matrix $Prob$ and vector μ , and is therefore a constant real-valued matrix (i.e., containing no parameters). As for the algorithm *mdpVD*, the fact that there is a single solution to this system comes from the existence of an absorbing state. Note also that the parametric value associated to a state is a *linear term*, as defined in Def. 1.

We state in the following Lemma that, given M and μ , the instantiation with π of the parametric value associated to M w.r.t. μ is equal to the value associated to $M[\pi]$ w.r.t. μ . We use $V[\pi]$ to denote the parametric value V instantiated with π .

<p>ALGORITHM P-mdpPI(M, μ_0)</p> <p><i>Input</i> M : Parametric Markov Decision Process ($S, A, Prob, W$) μ_0 : Optimal policy for the reference instantiation of the parameters</p> <p><i>Output</i> K_0 : Constraint on the set of parameters</p> <p>$V := P$-mdpVD(M, μ_0) $K_0 := True$</p> <p>FOR EACH $s \in S \setminus \{s_n\}$ DO FOR EACH $a \in e(s)$ s.t. $a \neq \mu_0[s]$ DO $K_0 := K_0 \wedge \{W_a(s) + \sum_{s' \in S} Prob_a(s, s')V[s'] \geq V[s]\}$ OD OD</p>

Figure 4: Algorithm solving the inverse problem for MDPs

Lemma 1 Let $M = (S, A, Prob, W)$ be a PMDP, π an instantiation of the parameter, and μ a policy for M . Let $V = P$ -mdpVD(M, μ). Then $V[\pi] = mdpVD(M[\pi], \mu)$.

Proof. The algorithm P -mdpVD(M, μ) consists in solving a system of the form $V = A \times V + W_{\mu[s]}$. Hence, $V = (1 - A)^{-1} \times W_{\mu[s]}$. Moreover, the algorithm $mdpVD(M[\pi], \mu)$ consists in solving a system of the form $v = A' \times v + W[\pi]_{\mu[s]}$, i.e., $v = (1 - A')^{-1} \times W[\pi]_{\mu[s]}$. It is easy to see on the two algorithms that $A = A'$. We trivially have: for all s , $W[\pi]_{\mu[s]}(s) = (W_{\mu[s]}(s))[\pi]$, where $(W_{\mu[s]}(s))[\pi]$ denotes the linear term $W_{\mu[s]}(s)$ where every occurrence of a parameter p_i was replaced by its instantiation π_i . Hence, $V[\pi] = mdpVD(M[\pi], \mu)$. \square

We now introduce the algorithm P -mdpPI, which fits in our general framework of Fig. 1. Given a reference instantiation π_0 of the parameters, this algorithm takes as input a PMDP M , and an optimal policy μ_0 associated to $M[\pi_0]$ (which can be computed using $mdpPI(M[\pi_0])$). Recall that, by ‘‘optimal’’, we mean here a policy under which the value of states is *minimal*. The algorithm outputs a constraint K_0 on the parameters such that:

1. $\pi_0 \models K_0$, and
2. for any $\pi \models K_0$, μ_0 is an optimal policy of $M[\pi]$.

The algorithm P -mdpPI is given in Fig. 4. We can summarize this algorithm as follows:

1. Compute the parametric value function, which associates to any state a parametric value w.r.t. μ_0 , using Algorithm P -mdpVD;
2. For every state $s \neq s_n$, for every action a different from the action $\mu_0[s]$ given by the the optimal policy, generate the following inequality stating that a is not a better action (i.e., an action which would lead to a better policy) than $\mu_0[s]$:

$$W_a(s) + \sum_{s' \in S} Prob_a(s, s')V[s'] \geq V[s]$$

The above set of inequalities implies that, for any s and a , the policy obtained from μ_0 by changing $\mu_0[s]$ with a , does not improve policy μ_0 (i.e., does not lead to any smaller value of state).

3.4 Properties

We first show that π_0 models the constraint K_0 output by our algorithm.

Proposition 1 *Let $\mu_0 = \text{mdpPI}(M[\pi_0])$, and $K_0 = P\text{-mdpPI}(M, \mu_0)$. Then $\pi_0 \models K_0$.*

Proof. (By *reductio ad absurdum*) Suppose $\pi_0 \not\models K_0$. Then, there exists an inequality J in K_0 such that $\pi_0 \not\models J$. By construction, this inequality J is of the form $W_a(s) + \sum_{s' \in S} \text{Prob}_a(s, s')V[s'] \geq V[s]$, for some s and some a . If this inequality J is not satisfied by π_0 , this means that a is a strictly better policy for s than the policy $\mu_0[s]$ in $M[\pi_0]$, which is not possible since μ_0 is an optimal policy for $M[\pi_0]$. \square

Proposition 2 *Algorithm $P\text{-mdpPI}$ terminates.*

Proof. Since M contains exactly one absorbing state, the computation of the parametric value in $P\text{-mdpVD}$ is guaranteed to terminate with a single solution. Since the number of generated inequalities is finite, it is easy to see that Algorithm $P\text{-mdpPI}$ terminates. \square

Note that the size (in term of number of inequalities) of the constraint K_0 output by our algorithm is in $O(|S| \times |A|)$, where $|S|$ (resp. $|A|$) denotes the number of states (resp. actions) of M .

We now state that our algorithm $P\text{-mdpPI}$ solves the inverse problem as described in Sect. 2.2.

Theorem 1 *Let $\mu_0 = \text{mdpPI}(M[\pi_0])$, and $K_0 = P\text{-mdpPI}(M, \mu_0)$. Then:*

1. $\pi_0 \models K_0$, and
2. for all $\pi \models K_0$, policy μ_0 is optimal for $M[\pi]$.

Proof. Let us prove item (2) by *reductio ad absurdum*. Recall that $M = (S, A, \text{Prob}, W)$. Let $\pi \models K_0$. We have $M[\pi] = (S, A, \text{Prob}, W[\pi])$.

Suppose that μ_0 is not an optimal policy for $M[\pi]$. Let μ be an optimal policy for $M[\pi]$. Then there exists some state s such that $\mu[s]$ is a strictly better policy than $\mu_0[s]$ for $M[\pi]$. Let $a = \mu[s]$ and $a_0 = \mu_0[s]$. Let $v = \text{mdpVD}(M[\pi], \mu)$. Since a is a strictly better policy than a_0 for state s in $M[\pi]$, then, from the last iteration of Algorithm $\text{mdpPI}(M[\pi])$, we have: $W[\pi]_{a_0}(s) + \sum_{s' \in S} \text{Prob}_{a_0}(s, s')v[s'] > W[\pi]_a(s) + \sum_{s' \in S} \text{Prob}_a(s, s')v[s']$.

Moreover, since $a \neq \mu_0[s]$, Algorithm $P\text{-mdpPI}(M, \mu_0)$ generates the following inequality in K_0 : $W_a(s) + \sum_{s' \in S} \text{Prob}_a(s, s')V[s'] \geq V[s]$. Since $V[s] = W_{a_0}(s) + \sum_{s' \in S} \text{Prob}_{a_0}(s, s') \times V[s']$ (from the call to Algorithm $P\text{-mdpVD}(M, \mu_0)$), this inequality is equal to $W_a(s) + \sum_{s' \in S} \text{Prob}_a(s, s')V[s'] \geq W_{a_0}(s) + \sum_{s' \in S} \text{Prob}_{a_0}(s, s') \times V[s']$. Since $\pi \models K_0$, the instantiation of K_0 , and in particular of this inequality, with π should evaluate to true. By Lemma 1, we have $V[\pi] = v$. Hence, by instantiating the inequality with π , we get: $W[\pi]_a(s) + \sum_{s' \in S} \text{Prob}_a(s, s')v[s'] \geq W[\pi]_{a_0}(s) + \sum_{s' \in S} \text{Prob}_{a_0}(s, s') \times v[s']$, which is exactly the contrary of what was stated before. \square

3.5 Application to the Example

Consider again the journey from Paris to Bologna described in Sect. 3.2. We give in Fig. 5 the PMDP M adapted from Fig. 2 to the parametric case. The set of parameters is $P = \{p_1, p_2, p_3\}$. The reference instantiation π_0 of the parameters is the following one²:

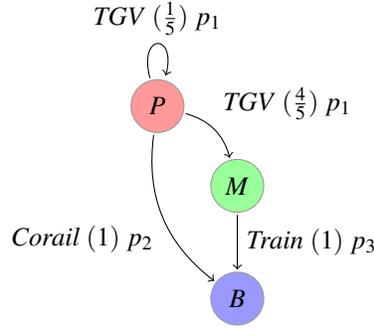


Figure 5: An example of Parametric Markov Decision Process

$$p_1 = 7 \quad p_2 = 11 \quad p_3 = 1$$

Note that $M[\pi_0]$ corresponds to the (standard) MDP depicted in Fig. 2.

Let us briefly explain the application of P - $mdpPI$ to this example. We first compute the optimal policy μ_0 for $M[\pi_0]$. As said in Sect. 3.2, $\mu_0 = \{P \rightarrow TGV, M \rightarrow Train\}$. Applying Algorithm P - $mdpVD(M, \mu_0)$, we then compute the parametric value of each state w.r.t. the optimal policy μ_0 . As B is an absorbing state, we have $V[B] = 0$. Thus, we trivially have $V[M] = p_3$. We then have $V[P] = W_{\mu_0[P]}(P) + 1/5 \times V[P] + 4/5 \times V[M]$, which gives $V[P] = 5/4 \times p_1 + p_3$. Note that, by replacing the parameters p_i by $\pi_0(p_i)$ in $V[P]$ for $i = 1, 2, 3$, we get $5/4 \times 7 + 1 = 9.75$, which is equal to the value computed by the classical algorithm $mdpPI$ (from Lemma 1).

We now compute the constraint K_0 . The only non-determinism being in state P , we generate the following inequality: $1 \times (p_2 + V[B]) \geq V[P]$, which gives:

$$p_2 \geq \frac{5}{4}p_1 + p_3$$

By instantiating all the parameters except the one corresponding to the duration of the train between Milan and Bologna (i.e., p_3), we get the following inequality:

$$p_3 \leq \frac{9}{4}$$

Thus, if the train between Milan and Bologna takes more than 2h15 (i.e., is impacted by a delay of more than 1h15), then the optimal policy of the TGV will not be optimal anymore, and we should consider another option.

Remark. This example being simple, it was rather easy to predict this result from the direct application of the classical algorithm $mdpPI$ to the MDP described in Fig. 2. Indeed, the expected value $v[P]$ in state P is equal to 9.75 so, if a delay of more than $11 - 9.75 = 1.25$ (i.e., 1h15) occurs somewhere between Paris and Bologna using the TGV option (in particular between Milan and Bologna), the TGV policy will not be optimal anymore. Our algorithm P - $mdpPI$ is of course interesting for more complex systems.

² From the definition of the MDP and PMDP, the weight corresponding to leaving state P through action TGV must be the same for any destination state. This is the reason why, in state P , the duration corresponding to waiting the next train (7 hours) is the same as the time needed to reach Milan. In the case where we would need different weights, it is possible to set an average value for the weight by taking into account the respective probabilities.

3.6 Implementation

The algorithm *P-mdpPI* has been implemented under the form of a program named *IMPERATOR* (standing for *Inverse Method for Policy with Reward AbstracT behaviOR*). This program, containing about 4300 lines of code, is written in Caml, and uses matrix inversion to compute the parametric value V in Algorithm *P-mdpVD*. We applied our program to various examples of MDPs modeling devices. For a system containing 11 states, 4 actions and 132 transitions, corresponding to the model of a robot evolving in a bounded physical space [9], our program *IMPERATOR* generates a constraint in 0.17 s.

4 Max-Plus Algebra

We consider in this section the algorithm 4.4 “Max-Plus Policy Iteration” of [3] (which will be here denoted by *maxPI*), used to compute the maximal circuit mean of a weighted directed graph in the framework of max-plus algebra. We are interested in computing a constraint on the weights attached to a directed graph, such that the circuit of maximal mean remains the same, under any instantiation satisfying this constraint.

We use in this section a formalism similar to the one in [3]³.

4.1 Preliminaries

The max-plus semiring \mathbb{R}_{max} is the set $\mathbb{R} \cup \{-\infty\}$, equipped with *max* and $+$. The zero element will be denoted by $\mathbb{0}$ ($\mathbb{0} = -\infty$). The unit element will not be used in this paper.

Definition 4 A directed weighted graph (or DWG) G is a triple (S, E, w) , where:

- S is a finite set of states,
- E is a set of oriented edges $E \subset S \times S$,
- $w : E \rightarrow \mathbb{R}$ is a function associating to every edge a real-valued weight.

We denote by $w(e)$, or alternatively by $w_{i,j}$, the weight associated to the edge $e = (i, j)$. We associate to G a matrix $M \in (\mathbb{R}_{max})^{S \times S}$, such that

$$M_{ij} = \begin{cases} w_{i,j} & \text{if } (i, j) \in E, \\ \mathbb{0} & \text{otherwise} \end{cases}$$

Conversely, we associate to any matrix $M \in \mathbb{R}^{n \times n}$ the graph $G_M = (S, E, w)$, where $S = \{1, \dots, n\}$, $E = \{(i, j) \in S \times S \mid M_{ij} \neq \mathbb{0}\}$, and $w_{i,j} = M_{ij}$ for any $(i, j) \in S \times S$.

In the following, we will mainly consider the formalism of matrices rather than the graphs. We consider in the following the matrix M , whose associated graph G_M is strongly connected.

Definition 5 Given a DWG $G = (S, E, w)$, the maximal circuit mean is

$$\rho = \max_c \frac{\sum_{e \in c} w(e)}{\sum_{e \in c} 1},$$

where the max is taken over all the circuits c of G , and the sums are taken over all the edges e of c .

³However, we will denote by μ the policy, in order to avoid confusion with π , standing in our framework for an instantiation of the parameters.

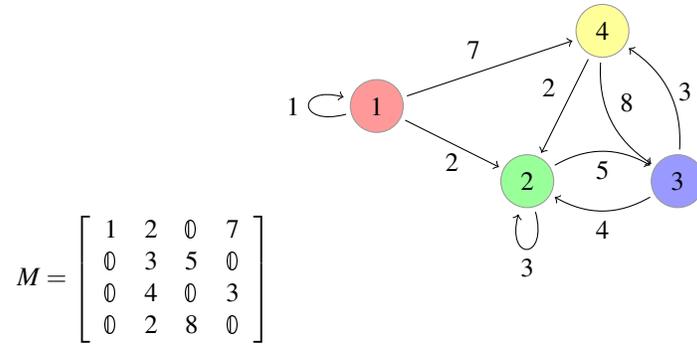


Figure 6: A matrix and its graph

Note that the denominator in the definition of ρ is the length of circuit c , whereas the numerator is the weight of c .

In the context of DWGs, given a matrix M , a *policy* is a function μ from S to E , such that for all $i \in S$, $\mu[i]$ is an edge starting from i . In the following, without loss of understanding, we will sometimes abbreviate the edge $\mu[i] = (i, j)$ as its target state j . Given a policy μ for M , we denote by $\mu[i]$ the policy associated to state i . Moreover, we denote by M^μ the matrix such that, for any i, j , $M_{ij}^\mu = M_{ij}$ if $j = \mu[i]$, and $M_{ij}^\mu = 0$ otherwise.

Given a matrix M and a policy μ , the *value function*, denoted by (η, x) , associates to each state i of S a couple $(\eta_i, x_i) \in \mathbb{R} \times \mathbb{R}$ (called “(generalized) eigenmode” in [3]).

An *optimal policy* μ for M induces a *circuit* c of maximal mean in graph G . More precisely, $\mu[i]$ is an edge of c if i belongs to c , and there is a path from i to a state of c otherwise. Moreover, the associated value (η, x) is such that all the η_i s are identical, and equal to the *maximal circuit mean* ρ of G .⁴

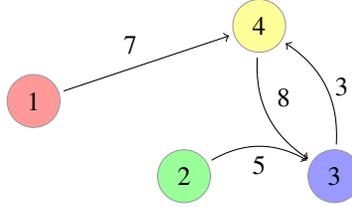
The algorithm *maxPI* (see Fig. 15 in Appendix B) computes an optimal policy for a given DWG. Starting from an arbitrary policy, it iteratively improves the current policy using Algorithm *maxPImpr* (see Fig. 14 in Appendix B) and Algorithm *maxVD* (see Fig. 13 in Appendix B), which computes the associated value function (η, x) .

4.2 An Illustrating Example

We give in Fig. 6 an example of DWG (coming from [3]) with its corresponding matrix. We are interested in finding the maximal circuit mean of a DWG. Let us briefly apply Algorithm *maxPI* to the matrix M of Fig. 6. As in [3], we choose the initial policy $\pi_1 : 1 \rightarrow 1, i \rightarrow 2$, for $i = 2, 3, 4$. Applying Algorithm *maxVD*, we find a first circuit $c_1 : 1 \rightarrow 1$, with $\bar{\eta} = w(c_1) = 1$. We set $\eta_1^1 = 1$, and $x_1^1 = 0$. Since 1 is the only state which has access to 1, we apply algorithm *maxVD* to the subgraph of G_M with states 2, 3, 4. We find the circuit $c_2 : 2 \rightarrow 2$ and set $\bar{\eta} = w(c_2) = 3$, $\eta_2^1 = 3$, and $x_2^1 = 0$. Since 3, 4 have access to 2, we set $\eta_i^1 = 3$ for $i = 3, 4$. Moreover, an application of (7) yields $x_3^1 = 4 - 3 + x_2^1$, and $x_4^1 = 2 - 3 + x_2^1$. To summarize:

$$\eta^1 = \begin{pmatrix} 1 \\ 3 \\ 3 \\ 3 \end{pmatrix}, \quad x^1 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ -1 \end{pmatrix}$$

⁴Note that the η_i s are also equal to the (unique) eigenvalue of M , and x is an eigenvector of M (see Theorem 3.1 in [3]).

Figure 7: The graph corresponding to the matrix M^μ

We improve the policy using Algorithm *maxPImpr*. Since $J = \{1\} \neq \emptyset$, we have a type 3a improvement. This yields $\pi_2 : i \rightarrow 2$ for $i = 2, 3, 4$. Only the entry 1 of x^1 and η^1 has to be modified, which yields

$$\eta^2 = \begin{pmatrix} 3 \\ 3 \\ 3 \\ 3 \end{pmatrix}, \quad x^2 = \begin{pmatrix} -1 \\ 0 \\ 1 \\ -1 \end{pmatrix}$$

We tabulate with less details the end of the run of the algorithm. Algorithm *maxPImpr*, type 3b, policy improvement. $\pi_3 : 1 \rightarrow 4, 2 \rightarrow 3, 3 \rightarrow 2, 4 \rightarrow 3$. Algorithm *maxVD*. Circuit found $c : 3 \rightarrow 2 \rightarrow 3$, $\bar{\eta} = (w_{2,3} + w_{3,2})/2 = 9/2$.

$$\eta^3 = \begin{pmatrix} \frac{9}{2} \\ \frac{9}{2} \\ \frac{9}{2} \\ \frac{9}{2} \end{pmatrix}, \quad x^3 = \begin{pmatrix} \frac{11}{2} \\ 0 \\ -\frac{1}{2} \\ 3 \end{pmatrix}$$

Algorithm *maxPImpr*, type 3b, policy improvement. The only change is $\pi_4(3) = 4$. Algorithm *maxVD*. Circuit found $c : 3 \rightarrow 4 \rightarrow 3$, $\bar{\eta} = (w_{3,4} + w_{4,3})/2 = 11/2$.

$$\eta^4 = \begin{pmatrix} \frac{11}{2} \\ \frac{11}{2} \\ \frac{11}{2} \\ \frac{11}{2} \end{pmatrix}, \quad x^4 = \begin{pmatrix} 4 \\ -\frac{1}{2} \\ 0 \\ \frac{5}{2} \end{pmatrix}$$

Algorithm *maxPImpr*. Stop. Hence, we get the following result:

$$\eta = \begin{pmatrix} \frac{11}{2} \\ \frac{11}{2} \\ \frac{11}{2} \\ \frac{11}{2} \end{pmatrix}, \quad x = \begin{pmatrix} 4 \\ -\frac{1}{2} \\ 0 \\ \frac{5}{2} \end{pmatrix}, \quad \mu = \begin{pmatrix} 4 \\ 3 \\ 4 \\ 3 \end{pmatrix} \quad (1)$$

Thus, $11/2$ is an eigenvalue of M , and x is an eigenvector. The subgraph M^μ of M restricted to the policy μ is given in Fig. 7. We note that the mean of circuit $4 \rightarrow 3 \rightarrow 4$ is $(8 + 3)/2 = 11/2$, and it is easy to check that this circuit has the maximal circuit mean of the graph associated to M .

We are now interested in the following problem. Suppose that one wants to minimize the weight associated to the edge $4 \rightarrow 3$ (of weight $w_{4,3} = 8$). What is the minimal value for $w_{4,3}$ so that circuit $4 \rightarrow 3 \rightarrow 4$ remains the circuit of maximal mean in the graph M ? In other words, we are interested in computing a constraint on the weights of the system, viewed as *parameters*, so that the circuit of maximal mean remains the circuit of maximal mean.

4.3 The Algorithm P -maxPI

We first adapt the notion of DWG to the parametric case. We now consider that the weights of the graph are *parameters*.

Definition 6 Given a set P of parameters, a parametric directed weighted graph (or PDWG) G is a triple (S, E, w) , where:

- S is a finite set of states,
- E is a set of oriented edges $E \subset S \times S$,
- $W : E \rightarrow P$ is a parametric function associating to every edge a parametric weight.

We denote by $W(e)$, or alternatively by $W_{i,j}$, the parametric weight associated to the edge $e = (i, j)$. We associate to G a parametric matrix $M \in (P \cup \mathbb{0})^{S \times S}$, such that

$$M_{ij} = \begin{cases} W_{i,j} & \text{if } (i, j) \in E, \\ \mathbb{0} & \text{otherwise} \end{cases}$$

Conversely, we associate to any parametric matrix $M \in (P \cup \mathbb{0})^{n \times n}$ the graph $G_M = (S, E, W)$, where $S = \{1, \dots, n\}$, $E = \{(i, j) \in S \times S \mid M_{ij} \neq \mathbb{0}\}$, and $W_{i,j} = M_{ij}$ for any $(i, j) \in S \times S$.

We consider in the following the PDWG (S, E, W) , and its associated matrix M . Given an instantiation π of the parameters, we denote by $W[\pi]$ the weight function from E to \mathbb{R} obtained by replacing each occurrence of a parameter p_i in W with the value $\pi(p_i)$, for $1 \leq i \leq N$. Similarly, we denote by $M[\pi]$ the matrix (in $(\mathbb{R}_{max})^{n \times n}$) obtained by replacing in M each occurrence of a parameter p_i with the value $\pi(p_i)$, for $1 \leq i \leq N$. The notion of policy can be extended to the parametric framework in a natural way.

Following the idea of our framework of Sect. 2, we first give in Fig. 8 the algorithm P -maxVD. This algorithm is an adaptation to the parametric case of the algorithm for value determination $maxVD$ from [3] (see Fig. 13 in Appendix B). Given a policy μ , it computes a *parametric* eigenmode (H, X) of M^μ . In other words, it associates to every state i of M^μ two parametric values H_i and X_i , which are two linear terms (as defined in Def. 1).

We now introduce the algorithm P -maxPI, which fits in our general framework of Fig. 1. We give the algorithm P -maxPI in Fig. 9. As in the MDP Section, we first apply the standard algorithm for policy iteration from the literature, i.e., we first call Algorithm $maxPI$, given in Fig. 15 in Appendix B (which makes itself use of Algorithms $maxVD$ and $maxPImpr$, available in Fig. 13 and Fig. 14 respectively). This algorithm computes the eigenmode (η, x) of the maximal circuit mean of M , and the corresponding policy μ_0 . Then we compute the parametric eigenmode of M associated to μ_0 , using Algorithm P -maxVD. Finally, we compute a set of inequalities ensuring that the policy μ_0 is the optimal policy w.r.t. maximal circuit mean. This generation of inequalities is the adaptation to the parametric case of the test of optimality performed in the classical algorithm $maxPImpr$ (given in Fig. 14 in Appendix B).

We now state that our algorithm P -maxPI solves the inverse problem as described in Sect. 2.2.

Theorem 2 Let $((\eta, x), \mu_0) = maxPI(M[\pi_0])$ and $K_0 = P$ -maxPI($M, ((\eta, x), \mu_0)$). Then:

1. $\pi_0 \models K_0$, and
2. for all $\pi \models K_0$, policy μ_0 corresponds to a maximal mean circuit of $M[\pi]$.

Note that, although we guarantee that the circuit of maximal mean in $M[\pi]$ is always the same, for any $\pi \models K$, the mean value itself varies with π .

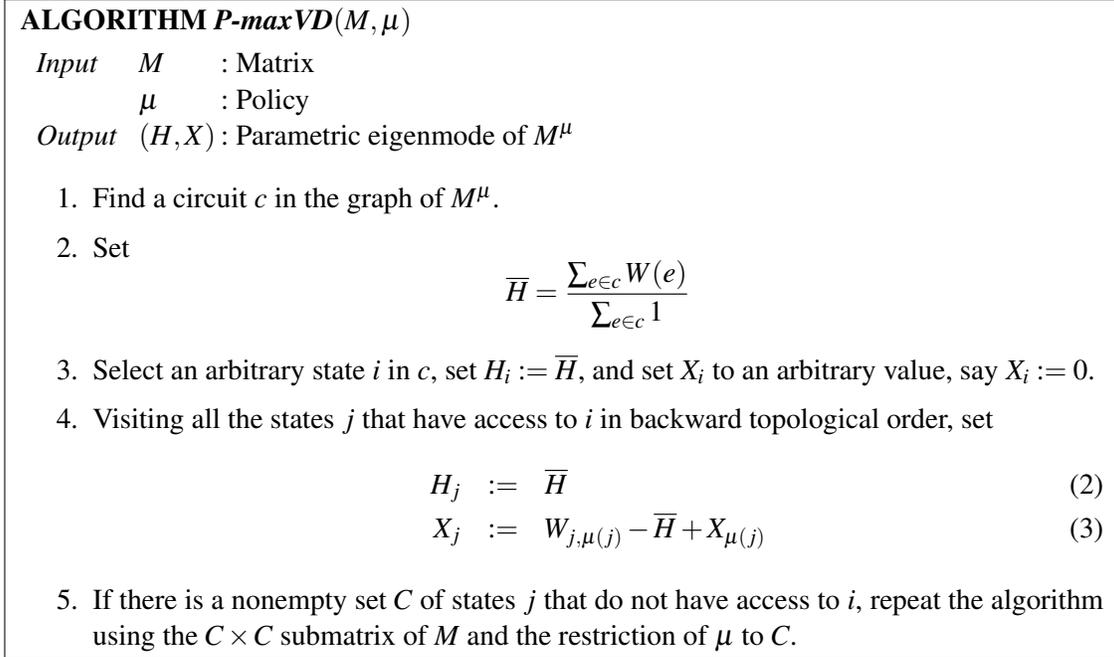


Figure 8: Algorithm for parametric value determination for maximal circuit mean

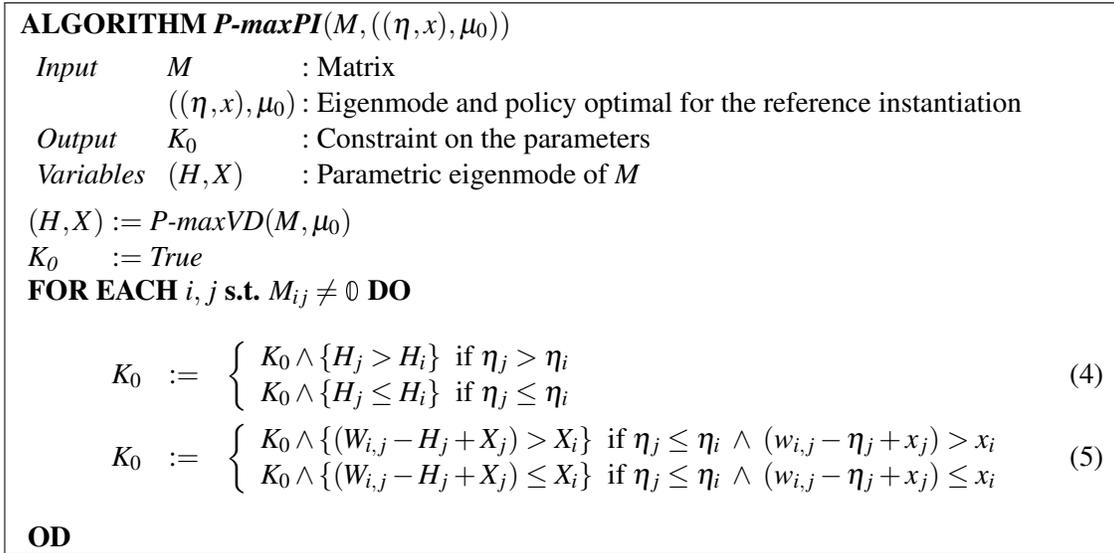


Figure 9: Algorithm solving the maximal circuit mean inverse problem

1 → 1	$W_{1,1} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} + W_{1,4} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} + W_{4,3} - \frac{1}{2}W_{3,4} + \frac{1}{2}W_{4,3} \leq \frac{1}{2}W_{3,4} + \frac{1}{2}W_{4,3}$ $W_{1,1} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} + W_{1,4} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} + W_{4,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} \leq W_{1,4} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} + W_{4,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3}$
1 → 2	$W_{1,2} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} + W_{2,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} \leq \frac{1}{2}W_{3,4} + \frac{1}{2}W_{4,3}$ $W_{1,2} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} + W_{2,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} \leq W_{1,4} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} + W_{4,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3}$
1 → 4	$W_{1,4} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} + W_{4,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} \leq \frac{1}{2}W_{3,4} + \frac{1}{2}W_{4,3}$ $W_{1,4} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} + W_{4,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} \leq W_{1,4} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} + W_{4,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3}$
2 → 2	$W_{2,2} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} + W_{2,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} \leq \frac{1}{2}W_{3,4} + \frac{1}{2}W_{4,3}$ $W_{2,2} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} + W_{2,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} \leq W_{2,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3}$
2 → 3	$W_{2,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} + 0 \leq \frac{1}{2}W_{3,4} + \frac{1}{2}W_{4,3}$ $W_{2,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} + 0 \leq W_{2,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3}$
3 → 2	$W_{3,2} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} + W_{2,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} \leq \frac{1}{2}W_{3,4} + \frac{1}{2}W_{4,3}$ $W_{3,2} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} + W_{2,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} \leq 0$
3 → 4	$W_{3,4} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} + W_{4,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} \leq \frac{1}{2}W_{3,4} + \frac{1}{2}W_{4,3}$ $W_{3,4} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} + W_{4,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} \leq 0$
4 → 2	$W_{4,2} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} + W_{2,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} \leq \frac{1}{2}W_{3,4} + \frac{1}{2}W_{4,3}$ $W_{4,2} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} + W_{2,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} \leq W_{4,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3}$
4 → 3	$W_{4,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} + 0 \leq \frac{1}{2}W_{3,4} + \frac{1}{2}W_{4,3}$ $W_{4,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} + 0 \leq W_{4,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3}$

Figure 10: The generation of K_0 for our example of graph

4.4 Application to the Example

Let us apply the algorithm P -maxPI given in Fig. 9 to the graph from [3] depicted in Fig. 6 in Sect. 4.2.

We first apply Algorithm $maxPI$, which gives the result (1) of Sect. 4.2.

Then, we call Algorithm P -maxVD. The circuit c found is $3 \rightarrow 4 \rightarrow 3$. We set $\bar{H} = (W_{3,4} + W_{4,3})/2$. We then pick up, say, state 3 in c , set $\bar{H}_3 := \bar{H}$ and $X_3 := 0$. Then, visiting all the states j that have access to i in backward topological order, we have:

- For state 1: $H_1 = \bar{H}$, and $X_1 = W_{1,4} - \bar{H} + X_4$
- For state 2: $H_2 = \bar{H}$, and $X_2 = W_{2,3} - \bar{H} + X_3$
- For state 4: $H_4 = \bar{H}$, and $X_4 = W_{4,3} - \bar{H} + X_3$

Since the set C of states j that do not have access to i is empty, the algorithm P -maxVD terminates. After resolution of the system above, we get

$$H = \begin{pmatrix} \frac{W_{3,4} + W_{4,3}}{2} \\ \frac{W_{3,4} + W_{4,3}}{2} \\ \frac{W_{3,4} + W_{4,3}}{2} \\ \frac{W_{3,4} + W_{4,3}}{2} \end{pmatrix}, \quad X = \begin{pmatrix} W_{1,4} - W_{3,4} \\ W_{2,3} - \frac{1}{2}W_{3,4} - \frac{1}{2}W_{4,3} \\ 0 \\ \frac{1}{2}W_{4,3} - \frac{1}{2}W_{3,4} \end{pmatrix}$$

We now generate the inequalities. For every edge (i, j) of the graph, we generate two inequalities, i.e., inequalities (4) and (5) of the algorithm P -maxPI. All generated inequalities, including the trivial ones (i.e., of the form $a \leq a$, for some linear term a), are depicted in Fig. 10. For every edge (i, j) of the graph, we first give the inequality corresponding to (4), and then the inequality corresponding to (5). The conjunction of those inequalities gives the constraint K_0 output by the algorithm.

After simplification (trivially done by hand) of the constraint K_0 , we get the following constraint:

$$\begin{aligned}
& 2W_{1,1} \leq W_{3,4} + W_{4,3} \\
\wedge & \quad W_{1,2} + W_{2,3} \leq W_{1,4} + W_{4,3} \\
\wedge & \quad 2W_{2,2} \leq W_{3,4} + W_{4,3} \\
\wedge & \quad W_{2,3} + W_{3,2} \leq W_{3,4} + W_{4,3} \\
\wedge & \quad 2W_{2,3} + 2W_{4,2} \leq W_{3,4} + 3W_{4,3}
\end{aligned}$$

Recall that we were interested in Sect. 4.2 in knowing until which value it was possible to minimize $W_{4,3}$ so that the circuit $4 \rightarrow 3 \rightarrow 4$ remained the circuit of maximal mean in the graph M of Fig. 6. Let us instantiate all parameters except $W_{4,3}$ in the constraint output by $P\text{-maxPI}(M, ((\eta, x), \mu_0))$. We then get the following inequality:

$$W_{4,3} \geq 6$$

Thus, provided this inequality is verified, the circuit $4 \rightarrow 3 \rightarrow 4$ remains the maximal mean circuit in the graph M of Fig. 6. Note that it is actually easy to see on the graph in Fig. 6 that, if $W_{4,3} < 6$, the maximal mean circuit then becomes $2 \rightarrow 3 \rightarrow 2$, with $\eta = 9/2$.

5 Final Remarks

We have presented an extension of two algorithms based on policy-iteration for two models: Markov Decision Problems and Max-Plus Algebras. For these models, we introduced a natural generalization of the policy-iteration method that solves the inverse problem, i.e: considering the weights of the models to be unknown constants or *parameters*, and given a reference instantiation π_0 of those weight parameters, we compute a constraint under which an optimal policy for π_0 is still optimal. This increases our confidence in the robustness of policy-iteration based methods.

Such an extension seems to work on several other policy-iteration algorithms. In particular, we are studying the adaptation of the method to Markov decision processes with *two* weights, as used in the problem of *dynamic power management* [8] for real-time systems where one wants to minimize the power consumption while keeping a certain level of efficiency. We also plan to adapt the method to an extension of Algorithm *maxPI* allowing to treat deterministic games with mean payoff [4].

Acknowledgments. We thank an anonymous referee for his/her helpful comments.

References

- [1] Étienne André, Thomas Chatain, Emmanuelle Encrenaz & Laurent Fribourg (2009): *An Inverse Method for Parametric Timed Automata*. *International Journal of Foundations of Computer Science* Available at <http://www.lsv.ens-cachan.fr/Publis/PAPERS/PDF/ACEF-ijfcs09.pdf>. To appear.
- [2] R. Bellman (1957): *A Markov decision process*. *Journal of Mathematical Mechanics* 6, pp. 679–684.
- [3] J. Cochet-terrasson, G. Cohen, S. Gaubert, M. Mc Gettrick & J-P. Quadrat (1998): *Numerical Computation of Spectral Elements in Max-Plus Algebra*. In: *IFAC Conf. on Syst. Structure and Control*.
- [4] V. Dhingra & S. Gaubert (2006): *How to solve large scale deterministic games with mean payoff by policy iteration*. In: *Valuetools '06*. ACM, New York, NY, USA, p. 12.
- [5] E. Encrenaz & L. Fribourg (2008): *Time Separation of Events: An Inverse Method*. In: *Proceedings of the LIX Colloquium '06, ENTCS 209*. Elsevier Science Publishers, Palaiseau, France.
- [6] R. A. Howard (1960): *Dynamic Programming and Markov Processes*. John Wiley and Sons, Inc.
- [7] J. Kemeny, H. Mirkil, J. Snell & G. Thompson (1959): *Finite mathematical structures*. Prentice-Hall, Englewood Cliffs, N.J.

- [8] G. A. Paleologo, L. Benini, A. Bogliolo & G. De Micheli (1998): *Policy optimization for dynamic power management*. In: *DAC '98*. ACM, New York, NY, USA, pp. 182–187.
- [9] L. Stachniss & W. Burgard (2003): *The Markov Decision Problem - Autonomous Mobile Systems*. Available at <http://ais.informatik.uni-freiburg.de/teaching/ss03/ams/DecisionProblems.pdf>. Course notes, University Freiburg, Germany.

A Markov Decision Processes Algorithms

```

ALGORITHM mdpVD( $M, \mu$ )
  Input    $M$  : Markov Decision Process ( $S, A, Prob, w$ )
            $\mu$  : Policy
  Output   $v$  : Value function
  SOLVE  $\{v[s] = w_{\mu[s]}(s) + \sum_{s' \in S} Prob_{\mu[s]}(s, s') \times v[s']\}_{s \in S \setminus s_n}$ 

```

Figure 11: Algorithm for value determination for MDPs

```

ALGORITHM mdpPI( $M$ )
  Input    $M$  : Markov Decision Process ( $S, A, Prob, w$ )
  Output   $\mu$  : Policy optimal w.r.t.  $w$  (initially random)
            $v$  : Value function
  REPEAT
     $v$       := mdpVD( $M, \mu$ )
    fixpoint := True
    for each  $s \in S \setminus s_n$  DO
      optimum :=  $v[s]$ 
      for each  $a \in e(s)$  DO
        IF  $w_a(s) + \sum_{s' \in S} Prob_a(s, s')v[s'] < \textit{optimum}$  THEN
          optimum :=  $w_a(s) + \sum_{s' \in S} Prob_a(s, s')v[s']$ 
           $\mu[s]$     :=  $a$ 
          fixpoint := False
        FI
      OD
    OD
  UNTIL fixpoint

```

Figure 12: Algorithm of policy iteration for MDPs

B Maximal Circuit Mean Algorithms

ALGORITHM $\text{maxVD}(M, \mu)$

Input M : Matrix

μ : Policy

Output (η, x) : Eigenmode of M^μ

1. Find a circuit c in the graph of M^μ .
2. Set

$$\bar{\eta} = \frac{\sum_{e \in c} w(e)}{\sum_{e \in c} 1}$$

3. Select an arbitrary state i in c , set $\eta_i := \bar{\eta}$, and set x_i to an arbitrary value, say $x_i := 0$.
4. Visiting all the states j that have access to i in backward topological order, set

$$\eta_j := \bar{\eta} \tag{6}$$

$$x_j := w_{j, \mu(j)} - \bar{\eta} + x_{\mu(j)} \tag{7}$$

5. If there is a nonempty set C of states j that do not have access to i , repeat the algorithm using the $C \times C$ submatrix of M and the restriction of μ to C .

Figure 13: Algorithm of value determination for maximal circuit mean in max-plus algebras

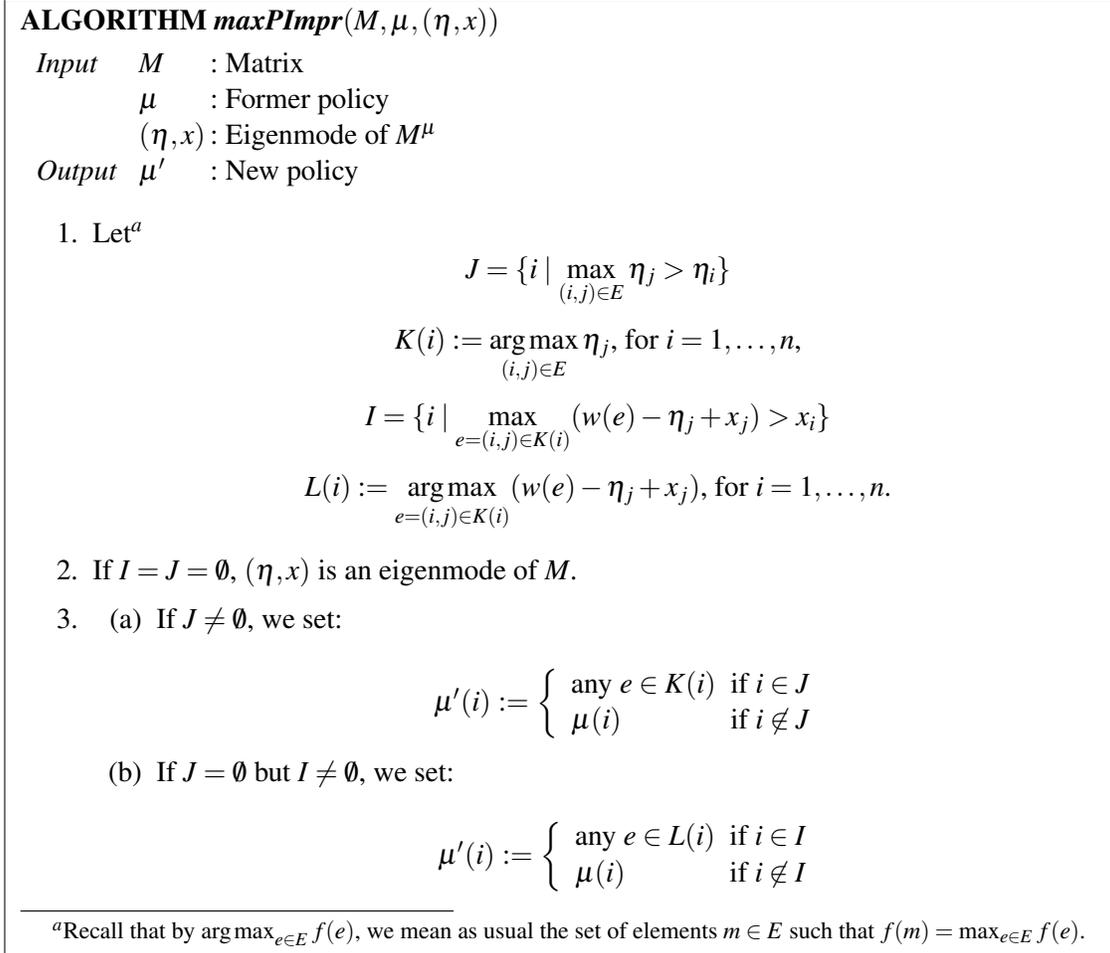


Figure 14: Algorithm of policy improvement for maximal circuit mean in max-plus algebras

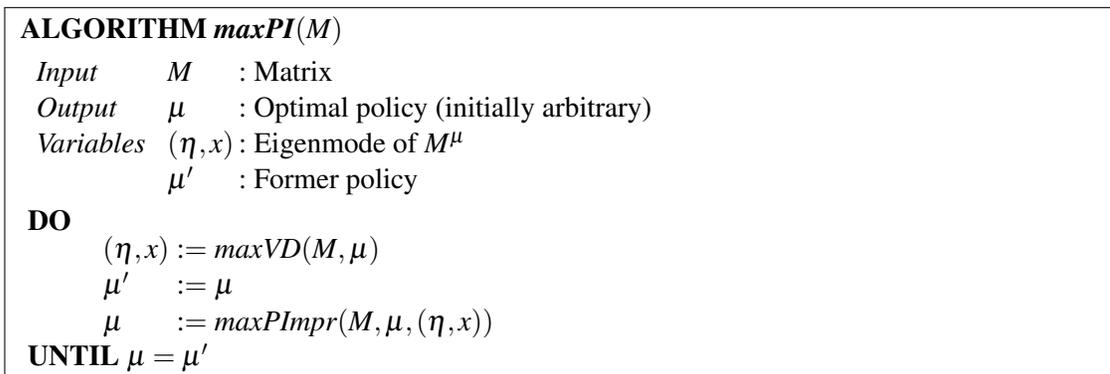


Figure 15: Algorithm of policy iteration for maximal circuit mean in max-plus algebras