

# A Constrained Randomized Shortest-Paths Framework for Optimal Exploration

(draft manuscript submitted for publication and subject to changes)

Bertrand Lebichot<sup>1</sup>, Guillaume Guex<sup>1</sup>,  
Ilkka Kivimäki<sup>1,2</sup> & Marco Saerens<sup>1,3</sup>

<sup>1</sup>ICTEAM and Machine Learning Group (MLG)  
Université catholique de Louvain (UCLouvain), Belgium

<sup>2</sup>Department of Computer Science  
Aalto University, Helsinki, Finland

<sup>3</sup>IRIDIA Laboratory  
Université Libre de Bruxelles (ULB), Belgium

July 13, 2018

## Abstract

The present work extends the randomized shortest-paths framework (RSP), interpolating between shortest-path and random-walk routing in a network, in three directions. First, it shows how to deal with equality constraints on a subset of transition probabilities and develops a generic algorithm for solving this constrained RSP problem using Lagrangian duality. Second, it derives a surprisingly simple iterative procedure to compute the optimal, randomized, routing policy generalizing the previously developed “soft” Bellman-Ford algorithm. The resulting algorithm allows balancing exploitation and exploration in an optimal way by interpolating between a pure random behavior and the deterministic, optimal, policy (least-cost paths) while satisfying the constraints. Finally, the two algorithms are applied to Markov decision problems by considering the process as a constrained RSP on a bipartite state-action graph. In this context, the derived “soft” value iteration algorithm appears to be closely related to dynamic policy programming [9, 10] as well as “Kullback-Leibler” and “path integral” control [76, 66, 27, 43, 74, 73], and similar to the reinforcement learning exploration strategy recently introduced in [7, 8]. This shows that this strategy is optimal in the RSP sense – it minimizes expected path cost subject to relative entropy constraint. Simulation results on illustrative examples show that the model behaves as expected.

## 1 Introduction

### 1.1 General introduction

The present work aims to study **randomized shortest-paths** (RSP) problems with *equality constraints* on the transition probabilities issued from a subset of

nodes, in the context of a single source and a single destination. This extension allows to fix some transition probabilities and then finding the optimal policy which is compatible with these probabilities. It therefore extends previous work dedicated to the RSP [68, 80, 48], initially inspired by stochastic traffic assignment models developed in transportation science [4].

The studied problem can be described informally as follows. Our aim is to find the optimal policy for reaching a goal node from a source node in a network by minimizing the expected cost of paths connecting these two nodes, where costs are associated to local decisions/actions. Usually, deterministic and stochastic shortest-path algorithms provide pure deterministic policies: when standing in a given state, we just choose the best path leading to minimal expected cost. In this work, we investigate the possibility of *optimally randomizing* the policy (exploration) while fixing a subset of transition probabilities. More precisely, the agent chooses a path to the goal node within a bag of paths according to an optimal probability distribution minimizing expected cost of paths subject to a relative entropy constraint, while satisfying transition probabilities constraints on a subset of nodes. In other words, the policy is expressed in terms of paths to the goal node. Interestingly, it can be shown that this method actually defines an optimal, biased, Markov chain in which the agent is “attracted” by the goal node (see later for details).

The degree of randomness is controlled by a *temperature parameter* allowing interpolating between the least-cost solution given by the (constrained) shortest-path algorithm and a random behavior provided by a predefined, reference, random policy (a reference random walk). Randomizing the policy thus introduces a *continual exploration* of the network. Standard Markov decision problems are a special case of this framework.

The originality of the work, in comparison with other models, lies in the fact that we adopt a *paths-based formalism* with entropy regularization. That is, the quantities of interest are defined on the set of full paths (or trajectories) connecting the source node to the goal node in the network. By using this paths-based formalism, as in the standard RSP [68, 48] and some models in transportation science [4], it is shown that the optimal randomized policy (both at the path level and at the edge level) can be computed by either (i) iteratively solving a system of linear equations or (ii) using a soft Bellman-Ford-like iteration algorithm.

## 1.2 Why consider randomized policies?

In practice, randomization corresponds to the association of a probability distribution on the set of admissible decisions in each node ([68], choice randomization or mixed strategy in game theory). If no randomization is present, only the best policy is exploited. Randomization thus appears when this probability distribution is no more peaked on the best choice: the agent is willing to sacrifice efficiency for exploration. Note that randomized choices are common in a variety of fields [68]; for instance game theory (mixed strategies; see for instance [58]), computer sciences [54], Markov games [51], decision sciences [63], reinforcement learning [70], etc. A comprehensive related work and a detailed discussion of the reasons for randomizing the policy can be found in [68, 2, 1], which are quickly summarized here:

- ▶ It is sometimes necessary to explore the environment, for instance when performing exploration in reinforcement learning [70].
- ▶ If the environment is changing over time (non-stationary), the system could benefit from randomization by performing continual exploration.
- ▶ A deterministic policy would lead to a totally predictable behavior; on the contrary, randomness introduces unpredictability and therefore renders interception more difficult. Randomization (randomized, or mixed, strategies) has proved to be useful for this reason in game theory [58].
- ▶ A randomized policy spreads the traffic over multiple paths, therefore reducing the danger of congestion.
- ▶ In some applications, like social networks analysis, computing a distance accounting for all paths – and thus integrating the concept of high connectivity – could provide better results than relying on the optimal, shortest, paths only [26].
- ▶ In computer gaming, it is often desirable to be able to adapt the strength of the digital opponent [31]. This allows modeling the behavior of incompletely rational players.

Within the context of the RSP framework, the randomness associated to paths connecting the source node and the goal node is quantified by the relative entropy, or Kullback-Leibler divergence (see, e.g., [22]), between the probability distribution defined on the paths and their likelihood according to a reference random walk on the graph – usually following a uniform distribution on the set of available decisions. This relative entropy captures the degree of randomness of the system. The optimal randomized policy is then obtained by minimizing the free energy – the expected cost plus the relative entropy weighted by temperature. As already mentioned, in this work, constraints are added to the optimisation problem by considering equality constraints on some transition probabilities, which are assumed provided by the environment and which have to be verified exactly.

### 1.3 Integrating constraints to the RSP framework

Being able to deal with constraints on the transition probabilities is important in a number of applications. Indeed, we do not always have a complete control on the behavior of the system: some state transitions are intrinsically stochastic and the model has to integrate this fact. For instance, in **Markov decision processes** (MDP), part of the environment is stochastic and is modeled by a Markov chain. By the way, it will be shown that our introduced constrained randomized shortest-paths formalism subsumes simple Markov decision processes in Section 6.

Based on this constrained RSP formalism, a first, generic, algorithm for solving the constrained problem is developed by exploiting Lagrange duality. Then, a simple, easy-to-implement, iterative algorithm, related to the “soft” Bellman-Ford algorithm [28, 29], is derived and its convergence to a fixed point is proved.

As an illustrative example, the framework is then used in order to solve randomized MDP problems, therefore providing a randomized policy. Markov decision processes [61, 62, 70, 75], also called stochastic shortest-path problems [12, 14], are currently used in a wide range of application areas including transportation networks, medical imaging, wide-area network routing, artificial intelligence, to name a few (see, e.g., [62, 70, 77, 78, 79]).

Interestingly, when applied to MDPs, the derived Bellman-Ford-like iterative algorithm – called here the **soft value iteration** – is closely related to dynamic policy programming [9, 10] as well as Kullback-Leibler and path integral control [76, 66, 27, 43, 74, 73]. It is also similar to the exploration strategy recently introduced in [7, 8]. This shows that this proposed exploration strategy is *globally optimal* in the following sense: it minimizes expected cost subject to constant relative entropy of paths probabilities when the goal state is absorbing and reachable from any other state. Interestingly, as in [28, 29] for the standard RSP without constraints, the soft value iteration algorithm extends the Bellman-Ford value iteration algorithm by simply replacing the minimum operator by a soft minimum operator. Note that still another way of solving the problem was developed in [16], but this algorithm is not included here because it is less generic.

## 1.4 Contributions and organization of the paper

In brief, this work contains the following contributions:

- ▶ It extends randomized shortest paths to problems with constrained transition probabilities on a subset of nodes.
- ▶ A generic algorithm solving the problem is introduced.
- ▶ An alternative, simple and easy-to-implement, iterative algorithm for computing the optimal randomized policy is derived.
- ▶ The constrained randomized shortest-paths framework is applied to solve standard Markov decision problems by introducing a soft value iteration algorithm.
- ▶ Simulations on concrete problems show that the algorithms behave as expected.

As far as the organization of the paper is concerned, Section 2 introduces the standard randomized shortest-paths framework. Section 3 considers randomized shortest-path problems with constraints on transition probabilities, which are then solved in Section 4 by using Lagrange duality. Section 5 then develops an alternative iterative algorithm, reminiscent from the Bellman-Ford recurrence, for computing the free energy and the optimal randomized policy. In Section 6, the standard Markov decision problem is recast as a constrained randomized shortest-path problem on a bipartite graph and a soft value iteration algorithm is developed for solving it. Section 7 shows some simulation examples and Section 8 is the conclusion.

## 2 The standard randomized shortest-path framework

As already stated, our formulation of the problem is based on the randomized shortest-path (RSP) framework defining, among others, a dissimilarity measure interpolating between the shortest-path distance and the commute-cost distance<sup>1</sup> in a graph [80, 68, 48]. The RSP framework relies on full paths instead of standard “local” flows [3].

In this section, we start by providing the necessary background and notation. Then, we proceed with a short summary of the randomized shortest-path formalism before introducing, in the next section, randomized shortest paths with constraints on the transition probabilities.

### 2.1 Some background and notation

Let us consider a weighted directed graph or network,  $G$ , with a set of  $n$  nodes  $\mathcal{V}$  (or vertices) and a set of arcs  $\mathcal{E}$  (or edges). The graph is assumed to be *strongly connected* and is represented by its  $n \times n$  adjacency matrix  $\mathbf{A}$ , containing binary values if the graph is unweighted or non-negative, local, affinities between nodes in the case of a weighted graph. To each edge linking node  $i$  to node  $j$ , we also associate a non-negative number  $c_{ij}$  representing the immediate cost of following this edge. The costs should be non-negative and are gathered in matrix  $\mathbf{C}$ . Note that self-loops are forbidden; in other words, the diagonal elements of the adjacency matrix are equal to 0. Similarly, diagonal elements of the cost are equal to  $\infty$ .

Moreover, a **reference random walk** (Markov chain) on  $G$  is defined in the usual manner. The choice to follow an edge from node  $i$  will be made according to a probability distribution (transition probabilities) defined on the set  $\text{Succ}(i)$  of successor nodes of  $i$ . These transition probabilities, defined on each node  $i$ , will be denoted as

$$p_{ij}^{\text{ref}} = \text{P}_{\text{ref}}(s(t+1) = j | s(t) = i) = \frac{a_{ij}}{\sum_{k \in \text{Succ}(i)} a_{ik}} \quad (1)$$

where  $a_{ij}$  is element  $i, j$  of the adjacency matrix and  $s(t)$  is a random variable representing the node visited by the random walker at time  $t$ . Furthermore,  $\mathbf{P}_{\text{ref}}$  will be the matrix containing the transition probabilities  $p_{ij}^{\text{ref}}$  as elements. For consistency, if there is no edge between  $i$  and  $j$  ( $a_{ij} = 0$ ), we consider that  $c_{ij}$  takes a large value, denoted by  $\infty$ ; in this case, the corresponding transition probability must also be equal to zero,  $p_{ij}^{\text{ref}} = 0$ .

Finally, in this work, we will assume that there is a unique *goal node*, which will be the last node  $n$ . This goal node is turned into an absorbing, killing, state in the corresponding Markov chain. Thus, any transition from this node is forbidden, that is,  $p_{nj}^{\text{ref}} = 0$  for all  $j$  – the random walker is killed after reaching goal state  $n$ .

---

<sup>1</sup>On an undirected graph, the commute-cost distance appears to be proportional to the commute-time distance [25, 48] and to the effective resistance [18] (also called resistance distance [49]) for a given graph – see [26] for a discussion.

## 2.2 The standard randomized shortest-path formalism

The main idea behind the RSP is as follows [68, 80, 48, 28, 29]. We consider the set of all **hitting paths**, or walks,  $\varphi \in \mathcal{P}$  from node 1 to the (unique) absorbing and killing node  $n$  on  $G$  (a bag of paths). Since the original graph is strongly connected, state  $n$  can be reached from any other node of the graph. Each path  $\varphi$  consists in a sequence of connected nodes starting in node 1 and ending in  $n$ . Then, we assign a probability distribution  $P(\cdot)$  (denoted as  $\mathbb{P}$  for convenience) on the set of paths  $\mathcal{P}$  by minimizing the relative **free energy**<sup>2</sup> of statistical physics [40, 59, 65],

$$\left\{ \begin{array}{l} \text{minimize}_{\{P(\varphi)\}_{\varphi \in \mathcal{P}}} \quad \phi(\mathbb{P}) = \underbrace{\sum_{\varphi \in \mathcal{P}} P(\varphi) \tilde{c}(\varphi)}_{\text{expected cost}} + T \underbrace{\sum_{\varphi \in \mathcal{P}} P(\varphi) \log \left( \frac{P(\varphi)}{\tilde{\pi}(\varphi)} \right)}_{\text{relative entropy}} \\ \text{subject to} \quad \sum_{\varphi \in \mathcal{P}} P(\varphi) = 1 \end{array} \right. \quad (2)$$

where  $\tilde{c}(\varphi) = \sum_{\tau=1}^t c_{s(\tau-1)s(\tau)}$  is the total cumulated cost along path  $\varphi$  when visiting the sequence of nodes, or states,  $(s(\tau))_{\tau=0}^t$  and  $t$  is the length of path  $\varphi$ . Furthermore,  $\tilde{\pi}(\varphi) = \prod_{\tau=1}^t p_{s(\tau-1)s(\tau)}^{\text{ref}}$  is the product of the reference transition probabilities (see Equation (1)) along path  $\varphi$  connecting node 1 to hitting node  $n$  – the likelihood of path  $\varphi$ . It defines a **reference** probability distribution over paths as  $\sum_{\varphi \in \mathcal{P}} \tilde{\pi}(\varphi) = 1$  [28, 29]. Note that, instead of a pure random walk, the reference probabilities  $p_{ij}^{\text{ref}}$  can also be set according to some prior knowledge.

The objective function in Equation (2) is a mixture of two dissimilarity terms with the temperature  $T$  balancing the trade-off between their relative contributions. The first term is the expected cost for reaching goal node from source node (favoring shorter paths – *exploitation*). The second term corresponds to the relative entropy [22, 45], or Kullback-Leibler divergence, between the path probability distribution and the path likelihood distribution (introducing randomness – *exploration*). When the temperature  $T$  is low, shorter paths are favored while when  $T$  is large, paths are chosen according to their likelihood in the reference random walk on the graph  $G$ . Note that we should add non-negativity constraints on the path probabilities, but this is not necessary as the resulting quantities will automatically be non-negative [22, 45]. Note that, instead of minimizing free energy, it is equivalent to minimize expected cost subject to a fixed relative entropy constraint [28, 26, 29].

This argument, akin to maximum entropy [40, 22, 44, 45], leads to a **Gibbs-Boltzmann distribution** on the set of paths (see, e.g., [28, 29] for a detailed derivation),

$$P^*(\varphi) = \frac{\tilde{\pi}(\varphi) \exp[-\theta \tilde{c}(\varphi)]}{\sum_{\varphi' \in \mathcal{P}} \tilde{\pi}(\varphi') \exp[-\theta \tilde{c}(\varphi')]} = \frac{\tilde{\pi}(\varphi) \exp[-\theta \tilde{c}(\varphi)]}{\mathcal{Z}} \quad (3)$$

<sup>2</sup>Alternatively, we can adopt a maximum entropy point of view, which is equivalent when the reference probability distribution is uniform [39, 41]. Moreover, the free energy could also be defined as  $\phi(\mathbb{P}) = \sum_{\varphi \in \mathcal{P}} P(\varphi)(\tilde{c}(\varphi) - c^*) + T \sum_{\varphi \in \mathcal{P}} P(\varphi) \log \left( \frac{P(\varphi)}{\tilde{\pi}(\varphi)} \right)$  where  $c^*$  is the least cost from starting node 1 to goal node  $n$ . In this case, costs are computed relatively to the shortest-path cost. This choice leads to the same probability distribution over paths (Equation (3)).

where  $\theta = 1/T$  is the inverse temperature and the denominator  $\mathcal{Z} = \sum_{\varphi \in \mathcal{P}} \tilde{\pi}(\varphi) \exp[-\theta \tilde{c}(\varphi)]$  is the **partition function** of the system.

This equation defines the **optimal randomized policy** at the *paths level*, in terms of *probabilities of choosing a particular path or trajectory*,  $P^*(\varphi)$ . It has been shown that this set of path probabilities is exactly equivalent to the one generated by a Markov chain with biased transition probabilities  $p_{ij}^*$ , favouring shorter paths, depending on the temperature  $T$  (see Equations (A.8), (A.17) and [68] for details). Contrary to (3) defined at the path level, these transition probabilities define the optimal policy at the *local, edge, level* in terms of probabilities of choosing an edge in each node. Note that a method for computing the RSP on large sparse graphs by restricting the set to paths with a finite predefined length was developed in [52, Section 4].

Several important quantities can easily be computed from this framework by, e.g., taking the partial derivative of the minimum free energy (see Equation (A.1) and [80, 68, 48, 28, 29, 26]). The quantities of interest that will be needed in this paper are introduced in the Appendix A. Readers who are not familiar with the RSP framework are invited to go through this appendix before continuing the reading.

### 3 Randomized shortest paths with constrained transition probabilities

Interestingly, the randomized shortest-path formulation just introduced in previous Section 2.2 can easily be extended to account for some types of constraints. The goal here will thus be to determine the best randomized policy – the optimal transition probabilities  $p_{ij}^*$  transporting the agent to the goal state  $n$  with minimum expected cost for a given level of relative entropy, and subject to *equality constraints* on some transition probabilities. We therefore have to derive the equivalent of the optimal biased transition probabilities provided by Equations (A.8), (A.17) in the standard RSP, but dealing now with equality constraints. This new model will be called the **constrained RSP**. As for the standard RSP, the goal node  $n$  is made absorbing and killing so that all the other nodes are *transient*.

As already discussed, constraints on transition probabilities are common in real-life applications where, in some (unconstrained) nodes, the agent has the control on the probability of choosing the next node while, in some other (constrained) nodes, the transition probabilities are provided by the environment and cannot be changed. An obvious example is Markov decision processes, which will be studied in the light of constrained RSP in Section 6. The constrained RSP therefore extends the range of applications of the standard RSP framework.

More concretely, we proceed as in previous section with the standard RSP, but we now constrain the transition probabilities associated to some nodes to be equal to predefined values provided by the user. In other words, we fix the relative flow passing through the edges incident to the nodes belonging to the subset of nodes  $\mathcal{C} \subset \mathcal{V} \setminus \{n\}$  (the absorbing goal node is excluded). These nodes will be called the **constrained**, transient, nodes. The optimal transition probabilities on the remaining, unconstrained and transient, nodes (the equivalent

of Equation (A.8) to be adapted for the constrained RSP) define the optimal policy that has to be adopted by the agents at the edge level. The subset of transient, **unconstrained**, nodes will be denoted as  $\mathcal{U} = \mathcal{V} \setminus (\mathcal{C} \cup \{n\})$ .

### 3.1 The Lagrange function

More precisely, from Equation (A.8), the considered constraints on the nodes  $i \in \mathcal{C}$  state that, on these nodes, the optimal randomized policy (transition probabilities) followed by an agent (i.e., the relative flow passing through an edge  $(i, j)$ ) should be equal to some given values  $q_{ij}$ ,

$$p_{ij}^*(T) = \frac{\bar{n}_{ij}(T)}{\bar{n}_i(T)} = q_{ij} \text{ for the edges starting in nodes } i \in \mathcal{C} \quad (4)$$

which should be independent of the temperature  $T$ . Here,  $\bar{n}_i(T)$  is the expected number of visits through node  $i$  and  $\bar{n}_{ij}(T)$  is the expected number of passages through edge  $(i, j)$ , when choosing trajectories thanks to the Gibbs-Boltzmann distribution in Equation (3) (see Equations (A.2) and (A.7)). The fixed transition probabilities  $q_{ij}$  must be specified by the user for all the nodes in  $\mathcal{C}$ . Of course, we have to assume that these constraints are feasible. In particular, we must have  $\sum_{j \in \text{Succ}(i)} q_{ij} = 1$  for all  $i \in \mathcal{C}$  with  $\text{Succ}(i)$  being the set of successor nodes of  $i$ .

Moreover, the RSP model (see Equation (2)) implies that, when  $T \rightarrow \infty$ , we should recover a pure random walk behavior with reference probabilities provided by Equation (1). Therefore, to be consistent, these reference probabilities and the  $q_{ij}$  must verify  $p_{ij}^{\text{ref}} = p_{ij}^*(T = \infty) = q_{ij}$  for nodes  $i \in \mathcal{C}$ . Therefore the constrained transition probabilities  $q_{ij}$  must be equal to the reference transition probabilities  $p_{ij}^{\text{ref}}$  on these constrained nodes. It will be assumed that this is the case in the sequel.

Consequently, let us consider the following Lagrange function integrating equality constraints (4)

$$\begin{aligned} \mathcal{L}(\mathbb{P}, \boldsymbol{\lambda}) = & \underbrace{\sum_{\varphi \in \mathcal{P}} \text{P}(\varphi) \tilde{c}(\varphi) + T \sum_{\varphi \in \mathcal{P}} \text{P}(\varphi) \log \left( \frac{\text{P}(\varphi)}{\tilde{\pi}(\varphi)} \right)}_{\text{relative free energy, } \phi(\mathbb{P})} + \mu \left( \sum_{\varphi \in \mathcal{P}} \text{P}(\varphi) - 1 \right) \\ & + \sum_{i \in \mathcal{C}} \sum_{j \in \text{Succ}(i)} \lambda_{ij} \left[ \underbrace{\sum_{\varphi \in \mathcal{P}} \text{P}(\varphi) \eta((i, j) \in \varphi)}_{\bar{n}_{ij}(T)} - q_{ij} \underbrace{\sum_{\varphi \in \mathcal{P}} \text{P}(\varphi) \eta(i \in \varphi)}_{\bar{n}_i(T)} \right] \quad (5) \end{aligned}$$

where, as before,  $\mathcal{P}$  is the set of paths connecting node 1 to node  $n$ , and with  $\eta((i, j) \in \varphi)$  being the number of times edge  $(i, j)$  appears on path  $\varphi$ . In a similar way,  $\eta(i \in \varphi)$  is the number of visits to node  $i$  when following path  $\varphi$ . Therefore, the last term in the previous equation states that the constraints (4) must be verified on each node  $i \in \mathcal{C}$ . Note that in our paths-based formalism, the expected number of visits to node  $i$  is expressed by  $\bar{n}_i(T) = \sum_{\varphi \in \mathcal{P}} \text{P}(\varphi) \eta(i \in \varphi)$  and the number of passages through edge  $(i, j)$  by  $\bar{n}_{ij}(T) = \sum_{\varphi \in \mathcal{P}} \text{P}(\varphi) \eta((i, j) \in \varphi)$  (see Equation (A.2)).



Now, the Lagrange function can be rearranged as

$$\begin{aligned}
\mathcal{L}(\mathbb{P}, \boldsymbol{\lambda}) &= \sum_{\wp \in \mathcal{P}} P(\wp) \left[ \underbrace{\tilde{c}(\wp) + \sum_{i \in \mathcal{C}} \sum_{j \in \text{Succ}(i)} \lambda_{ij} \eta((i, j) \in \wp) - \sum_{i \in \mathcal{C}} \eta(i \in \wp) \sum_{j' \in \text{Succ}(i)} q_{ij'} \lambda_{ij'}}_{\tilde{c}'(\wp)} \right] \\
&\quad + T \sum_{\wp \in \mathcal{P}} P(\wp) \log \left( \frac{P(\wp)}{\tilde{\pi}(\wp)} \right) + \mu \left( \sum_{\wp \in \mathcal{P}} P(\wp) - 1 \right) \\
&= \sum_{\wp \in \mathcal{P}} P(\wp) \sum_{i \in \mathcal{V}} \sum_{j \in \text{Succ}(i)} \eta((i, j) \in \wp) \left[ \underbrace{c_{ij} + \delta(i \in \mathcal{C}) \lambda_{ij} - \delta(i \in \mathcal{C}) \sum_{j' \in \text{Succ}(i)} q_{ij'} \lambda_{ij'}}_{\text{augmented costs } c'_{ij}} \right] \\
&\quad + T \sum_{\wp \in \mathcal{P}} P(\wp) \log \left( \frac{P(\wp)}{\tilde{\pi}(\wp)} \right) + \mu \left( \sum_{\wp \in \mathcal{P}} P(\wp) - 1 \right) \\
&= \underbrace{\sum_{\wp \in \mathcal{P}} P(\wp) \tilde{c}'(\wp)}_{\text{free energy } \phi'(\mathbb{P}) \text{ based on augmented costs, } \tilde{c}'(\wp)} + T \sum_{\wp \in \mathcal{P}} P(\wp) \log \left( \frac{P(\wp)}{\tilde{\pi}(\wp)} \right) + \mu \left( \sum_{\wp \in \mathcal{P}} P(\wp) - 1 \right) \quad (6)
\end{aligned}$$

where we used the Kronecker delta  $\delta(i \in \mathcal{C})$  which is equal to 1 when  $i \in \mathcal{C}$  and 0 otherwise, as well as  $\eta(i \in \wp) = \sum_{j \in \text{Succ}(i)} \eta((i, j) \in \wp)$  and  $\tilde{c}(\wp) = \sum_{i \in \mathcal{V}} \sum_{j \in \text{Succ}(i)} \eta((i, j) \in \wp) c_{ij}$  the total cost along path  $\wp$ . Thus, in (6) the local costs  $c_{ij}$  are redefined as

$$c'_{ij} = \begin{cases} c_{ij} + \lambda_{ij} - \underbrace{\sum_{j' \in \text{Succ}(i)} q_{ij'} \lambda_{ij'}}_{\text{extra cost } \Delta_{ij}} & \text{when node } i \in \mathcal{C} \\ c_{ij} & \text{when node } i \in \mathcal{U} \end{cases} \quad (7)$$

and  $\mathbf{C}'$  will be the matrix containing these new costs  $c'_{ij}$  where the **extra costs** are defined as  $\Delta_{ij} \triangleq \lambda_{ij} - \sum_{j' \in \text{Succ}(i)} q_{ij'} \lambda_{ij'}$ .

These new costs  $c'_{ij}$ , augmented by the extra costs coming from the Lagrange multipliers, will be called the **augmented costs**. We observe that Equation (6) is exactly a randomized shortest-paths problem (see Equation (2)) containing augmented costs instead of the initial costs, which can be solved by a standard RSP algorithm.

We further observe that the weighted (by transition probabilities) means of the extra costs must be equal to zero on each node  $i \in \mathcal{C}$ :

$$\sum_{j \in \text{Succ}(i)} q_{ij} \Delta_{ij} = 0 \quad \text{for each } i \in \mathcal{C} \quad (8)$$

In other words, the extra costs are centered with respect to the weights  $q_{ij}$  on each constrained node. Interestingly, this implies that the weighted average of the augmented costs is equal to the weighted average of the original costs on each constrained node  $i$ ,  $\sum_{j \in \text{Succ}(i)} q_{ij} c'_{ij} = \sum_{j \in \text{Succ}(i)} q_{ij} c_{ij}$ . In this case, the perceived cost (cost plus extra cost) when visiting any node using the augmented costs is exactly the same in average as the perceived real cost (cost only) as in the case where no constraint is introduced.

Thus, in Equation (6), everything happens as if the costs have been redefined by taking into account the Lagrange parameters. The extra costs, depending on these Lagrange parameters, can therefore be interpreted as extra virtual costs necessary to exactly satisfy the equality constraints, in the same way as when considering the dual problem in linear programming [35].

Let  $\phi'(\mathbb{P}) = \sum_{\wp \in \mathcal{P}} \mathbb{P}(\wp) \check{c}'(\wp) + T \sum_{\wp \in \mathcal{P}} \mathbb{P}(\wp) \log\left(\frac{\mathbb{P}(\wp)}{\bar{\pi}(\wp)}\right)$  be the relative free energy obtained from these augmented costs (see Equation (6)). We now address the problem of computing the Lagrange parameters  $\lambda_{ij}$  and the extra costs  $\Delta_{ij}$  by Lagrange duality.

## 4 Solving constrained RSP problems by Lagrange duality

In this section, we will take advantage of the fact that, in our formulation of the problem, the Lagrange dual function and its gradient with respect to a set of Lagrange parameters associated to a node are easy to compute. Indeed, the situation is equivalent to maximum entropy problems under constraints (see, e.g., [39, 41]), so that the same methodology can be used for optimising the objective function. This will provide a *generic algorithm* for solving constrained RSP problems based on Lagrange duality.

As the objective function is convex and all the equality constraints are linear, there is only one global minimum and the duality gap is zero [13, 23, 35]. The optimum is a saddle point of the Lagrange function and a common optimization procedure ([13, 23, 35], related to the Arrow-Hurwicz-Uzawa method [6]) consists in sequentially (i) solving the primal while considering the Lagrange parameters as fixed, which provides the dual Lagrange function  $\mathcal{L}^*(\boldsymbol{\lambda})$ , and then (ii) optimizing the obtained dual Lagrange function (which is concave) with respect to a subset of Lagrange parameters (a block  $\mathcal{B}$ ) until convergence.

In our context, this provides the two following steps [35], which are computed iteratively on blocks of Lagrange parameters  $\mathcal{B}$ ,

$$\begin{cases} \mathcal{L}^*(\boldsymbol{\lambda}^{(t)}) = \min_{\mathbb{P} \equiv \{\mathbb{P}(\wp)\}_{\wp \in \mathcal{P}}} \mathcal{L}(\mathbb{P}, \boldsymbol{\lambda}^{(t)}) & \text{(compute the dual Lagrange function)} \\ \lambda_{ij}^{(t+1)} = \arg \max_{\lambda_{ij}^{(t)} \in \mathcal{B}^{(t)}} \mathcal{L}^*(\boldsymbol{\lambda}^{(t)}) \text{ for } \lambda_{ij}^{(t)} \in \mathcal{B}^{(t)} & \text{(maximize the dual Lagrange function)} \\ \lambda_{ij}^{(t+1)} = \lambda_{ij}^{(t)} \text{ for } \lambda_{ij}^{(t)} \notin \mathcal{B}^{(t)} & \text{(keep the other Lagrange parameters)} \end{cases} \quad (9)$$

and the first maximization is performed subject to non-negativity and sum-to-one constraints. This is the procedure that will be followed, where the dual function will be maximized through a simple block coordinate ascend on Lagrange parameters. Each block at a given step  $t$  of the iteration will contain the Lagrange parameters associated to the node  $i$  processed at that time step (the edges incident to node  $i$ ,  $\mathcal{B}^{(t)} = \text{Succ}(i)$ ) and the procedure is iterated on the set of constrained nodes ( $i \in \mathcal{C}$ ).

## 4.1 Computing the dual Lagrange function

We already know from (3) that in the first step in Equation (9) the optimal probability distribution is obtained with

$$P^*(\varphi) = \frac{\tilde{\pi}(\varphi) \exp[-\theta \tilde{c}'(\varphi)]}{\sum_{\varphi' \in \mathcal{P}} \tilde{\pi}(\varphi') \exp[-\theta \tilde{c}'(\varphi')]} = \frac{\tilde{\pi}(\varphi) \exp[-\theta \tilde{c}'(\varphi)]}{\mathcal{Z}'} \quad (10)$$

where  $\tilde{c}'(\varphi)$  is the augmented cost of path  $\varphi$ .

Then, from Equations (A.1) and (6), the dual Lagrange function can easily be computed in function of the partition function defined from the augmented costs [41],

$$\mathcal{L}^*(\boldsymbol{\lambda}) = -T \log \mathcal{Z}' \quad (11)$$

and will be maximized at each time step with respect to the  $\{\lambda_{ij}\}$  with  $i \in \mathcal{C}$  and  $j \in \text{Succ}(i)$ . In addition, by extension of Equation (A.1) to any transient nodes (see Equation (A.14)), the minimum free energy from any node  $i$  (see [48, 28, 29] for details) is given by

$$\phi_i^* = -T \log z_{in} = -\frac{1}{\theta} \log z_{in} \quad (12)$$

where the backward variable  $z_{in}$  (element  $i, n$ , of the fundamental matrix  $\mathbf{Z}$ , see Equation (A.5)) is now computed from the *augmented costs*.

## 4.2 Maximizing the dual Lagrange function

Let us now maximize the dual function by using a simple block coordinate ascend. Because  $\bar{n}_i(T) = \sum_{j \in \text{Succ}(i)} \bar{n}_{ij}(T)$ , by following the reasoning of previous subsection (see Equation (A.2)), we obtain for constrained nodes  $i \in \mathcal{C}$

$$\begin{aligned} \frac{\partial \mathcal{L}^*(\boldsymbol{\lambda})}{\partial \lambda_{ij}} &= \frac{\partial(-T \log \mathcal{Z}')}{\partial \lambda_{ij}} = \sum_{j' \in \text{Succ}(i)} \frac{\partial(-T \log \mathcal{Z}')}{\partial c'_{ij'}} \frac{\partial c'_{ij'}}{\partial \lambda_{ij}} \\ &= \sum_{j' \in \text{Succ}(i)} \bar{n}_{ij'}(T) (\delta_{jj'} - q_{ij}) = \bar{n}_{ij}(T) - q_{ij} \bar{n}_i(T) \end{aligned} \quad (13)$$

Quite naturally, and similarly to maximum entropy problems [45], setting the result to zero provides the constraints on nodes  $i \in \mathcal{C}$ ,

$$\frac{\bar{n}_{ij}(T)}{\bar{n}_i(T)} = q_{ij} \quad (14)$$

and we now have to solve these equations in terms of the Lagrange parameter  $\lambda_{ij}$ .

---

**Algorithm 1** Computing the optimal randomized policy of a constrained RSP problem.

---

**Input:**

- The  $n \times n$  adjacency matrix  $\mathbf{A}$  of a strongly connected directed graph, containing non-negative edge affinities. Node 1 is the starting node and node  $n$  the goal node.
- The  $n \times n$  cost matrix  $\mathbf{C}$  of the graph, containing non-negative edge costs.
- The set of unconstrained nodes  $\mathcal{U}$  and constrained nodes  $\mathcal{C}$ .
- The positive inverse temperature parameter  $\theta$ .

**Output:**

- The  $(n-1) \times n$  matrix  $\mathbf{P}^*$  containing optimal transition probabilities (the policy).

1.  $\mathbf{D} \leftarrow \text{Diag}(\mathbf{A}\mathbf{e})$   $\triangleright$  the diagonal out-degree matrix;  $\mathbf{e}$  is a vector full of 1's
  2.  $\mathbf{P}_{\text{ref}} \leftarrow \mathbf{D}^{-1}\mathbf{A}$   $\triangleright$  the  $n \times n$  reference transition probabilities matrix
  3.  $\mathbf{C}' \leftarrow \mathbf{C}$   $\triangleright$  initialise the augmented costs matrix
  4. Set row  $n$  of  $\mathbf{P}_{\text{ref}}$  to  $\mathbf{0}^T$   $\triangleright$  row  $n$  set to zero: node  $n$  is made absorbing and killing
  5. **repeat**  $\triangleright$  main iteration loop
  6.     **for each**  $i \in \mathcal{C}$  **do**  $\triangleright$  loop on constrained nodes in  $\mathcal{C}$
  7.          $\mathbf{W} \leftarrow \mathbf{P}_{\text{ref}} \circ \exp[-\theta\mathbf{C}']$   $\triangleright$  compute the auxiliary matrix  $\mathbf{W}$  in terms of current augmented costs;  $\circ$  is the elementwise matrix product
  8.         Solve  $(\mathbf{I} - \mathbf{W})\mathbf{z}_b = \mathbf{e}_n$  with respect to  $\mathbf{z}_b$   $\triangleright$  compute the backward variable  $\mathbf{z}_b = \mathbf{Z}\mathbf{e}_n$  (column  $n$  of the fundamental matrix  $\mathbf{Z} = (\mathbf{I} - \mathbf{W})^{-1}$ ) where  $\mathbf{e}_n$  is a vector full of 0's except element  $n$  which is equal to 1
  9.          $\phi^* \leftarrow -\frac{1}{\theta} \log \mathbf{z}_b$  and then  $\phi_n^* \leftarrow 0$   $\triangleright$  elementwise natural logarithm: compute the vector of free energies, and force 0 on the goal node  $n$
  10.        **for each**  $j \in \text{Succ}(i)$  **do**  $\triangleright$  update the augmented costs on edges incident to constrained node  $i$
  11.            $c'_{ij} \leftarrow -\phi_j^* + \sum_{k \in \text{Succ}(i)} p_{ik}^{\text{ref}}(c_{ik} + \phi_k^*)$   $\triangleright$  augmented cost update for edge  $(i, j)$
  12.        **end for**
  13.     **end for**
  14. **until** convergence of the free energy vector
  15.  $\mathbf{Q} \leftarrow \mathbf{P}_{\text{ref}} \circ \exp[-\theta(\mathbf{C}' + \mathbf{e}(\phi^*)^T)]$   $\triangleright$  compute the numerator of the optimal transition probabilities matrix
  16. Remove row  $n$  of matrix  $\mathbf{Q}$   $\triangleright$  delete the zero row corresponding to the absorbing, goal, node  $n$
  17.  $\mathbf{s} \leftarrow \mathbf{Q}\mathbf{e}$   $\triangleright$  the row sums vector for normalization
  18.  $\mathbf{P}^* \leftarrow \mathbf{Q} \div (\mathbf{s}\mathbf{e}^T)$   $\triangleright$  the  $(n-1) \times n$  optimal transition probabilities matrix (the policy);  $\div$  is the elementwise division. We divide each row of  $\mathbf{Q}$  by its sum.
  19. **return**  $\mathbf{P}^*$
- 

### 4.3 Computing the Lagrange parameters and the augmented costs

Recalling that  $\bar{n}_i(T) = \sum_{j \in \text{Succ}(i)} \bar{n}_{ij}(T)$  and Equations (A.6)-(A.7), we obtain by imposing the constraint (14) for a node  $i \in \mathcal{C}$  and  $j \in \text{Succ}(i)$ ,

$$\begin{aligned}
& \frac{p_{ij}^{\text{ref}} \exp[-\theta c'_{ij}] z_{jn}}{\sum_{j' \in \text{Succ}(i)} p_{ij'}^{\text{ref}} \exp[-\theta c'_{ij'}] z_{j'n}} \\
&= \frac{p_{ij}^{\text{ref}} z_{jn} \exp[-\theta c_{ij}] \exp[-\theta \Delta_{ij}]}{\sum_{j' \in \text{Succ}(i)} p_{ij'}^{\text{ref}} z_{j'n} \exp[-\theta c_{ij'}] \exp[-\theta \Delta_{ij'}]} = q_{ij} \quad (15)
\end{aligned}$$

The goal now is to compute the new augmented cost (and thus the new extra costs  $\Delta_{ij}$  and the Lagrange parameters  $\lambda_{ij}$ , see Equation (7)) corresponding to node  $i \in \mathcal{C}$  by isolating the  $\Delta_{ij}$  with  $j \in \text{Succ}(i)$  in the previous Equation (15). In Appendix B, it is shown that we obtain (see Equation (B.4))

$$\Delta_{ij} = -(c_{ij} + \phi_j^*) + \sum_{k \in \text{Succ}(i)} p_{ik}^{\text{ref}}(c_{ik} + \phi_k^*), \text{ for each } j \in \text{Succ}(i) \quad (16)$$

which allows to directly compute the new augmented costs

$$c'_{ij} = c_{ij} + \Delta_{ij} = \sum_{k \in \text{Succ}(i)} p_{ik}^{\text{ref}}(c_{ik} + \phi_k^*) - \phi_j^*, \text{ for each } j \in \text{Succ}(i) \quad (17)$$

and, after convergence, this expression must be exactly verified by the augmented costs on all the constrained nodes.

Equation (17) suggests the following updating rule (bloc coordinate ascend) to be applied on all the edges incident to  $i$  at each iteration

$$c'_{ij} \leftarrow \sum_{k \in \text{Succ}(i)} p_{ik}^{\text{ref}}(c_{ik} + \phi_k^*) - \phi_j^*, \text{ for each } j \in \text{Succ}(i) \quad (18)$$

to be repeated on all constrained nodes (one constrained node  $i$  processed at each iteration step) until convergence.

Moreover, it can easily be shown from the previous results that the Lagrange multipliers are given<sup>3</sup> by

$$\lambda_{ij} = -(c_{ij} + \phi_j^*) \quad (19)$$

Let us now summarize the whole procedure.

#### 4.4 The complete procedure

Therefore, after specifying a parameter  $\theta$  and initializing the augmented costs  $c'_{ij}$  to the real costs  $c_{ij}$ , the final block coordinate ascend procedure iterates the following steps for updating the augmented costs associated to a constrained node  $i$ :

1. The elements of the fundamental matrix are computed thanks to Equation (A.4) from the *current augmented costs*  $c'_{ij}$  (defined in Equation (7)) and from the transition matrix of the natural random walk on  $G$  (Equation (1)), where goal node  $n$  is made absorbing and killing,  $\mathbf{Z} = (\mathbf{I} - \mathbf{W})^{-1}$  with  $\mathbf{W} = \mathbf{P}_{\text{ref}} \circ \exp[-\theta \mathbf{C}']$ .
2. Compute the minimum free energies on node  $i$  and its adjacent nodes ( $j \in \text{Succ}(i)$ ) thanks to Equation (12),  $\phi_i^* = -\frac{1}{\theta} \log z_{in}$ .
3. The augmented costs are updated on all edges incident to node  $i$  ( $j \in \text{Succ}(i)$ ) thanks to Equation (18),  $c'_{ij} \leftarrow \sum_{k \in \text{Succ}(i)} p_{ik}^{\text{ref}}(c_{ik} + \phi_k^*) - \phi_j^*$ . Then, go back to step 1 and proceed with another constrained node  $i$ .

---

<sup>3</sup>Up to the addition of a constant, as they must be centered.

The previous steps are thus performed repeatedly on the constrained nodes  $i \in \mathcal{C}$  and the whole procedure is iterated until convergence. Then, the optimal policy is obtained from Equation (A.17) by using the augmented costs  $c'_{ij}$  instead of  $c_{ij}$  (also for computing the backward variables  $z_{in}$ ). This provides the optimal transition probabilities  $p_{ij}^*(T)$  on the unconstrained nodes – for the constrained nodes, the optimal transition probabilities are of course equal to the reference transition probabilities. The resulting algorithm is presented in Algorithm 1. Note that in this algorithm (line 8), instead of computing the fundamental matrix  $\mathbf{Z}$ , we prefer to simply calculate the backward variables vector  $\mathbf{z}_b = \mathbf{Z}\mathbf{e}_n$  containing the elements  $z_{in}$ .

Let us now present an alternative, iterative, procedure, reminiscent of the Bellman-Ford formula for finding the shortest-path distance in a graph and the value iteration in Markov decision problems, solving the constrained RSP problem.

## 5 Solving constrained RSP problems by a simple iterative algorithm

This section introduces an alternative way of solving constrained randomized shortest-paths problems, based on an extension of Equation (A.15) computing the free energy from each transient node to the goal node [28, 29]. Once the free energy has been computed for all nodes, the optimal policy is easily obtained by the closed-form expression (A.17).

### 5.1 An optimality condition in terms of free energy

Recall that the quantity  $\phi_i^*(T) = -\frac{1}{\theta} \log z_{in}$  with  $\theta = 1/T$  (see Equation (12)), where  $z_{in}$  is the backward variable introduced in Equation (A.5), is called the (minimum) relative, directed, free energy potential<sup>4</sup> of the constrained RSP system associated to the different nodes  $i \in \mathcal{V}$ . As before, the dependence of the free energy on  $T$  will be omitted.

Inspired by the standard bag-of-paths framework [28, 29], it is shown in Appendix C that, at optimality, the recurrence relations computing the minimal free energy of the constrained RSP system are of the following form

$$\phi_i^* = \begin{cases} -\frac{1}{\theta} \log \left[ \sum_{j \in \text{Succ}(i)} p_{ij}^{\text{ref}} \exp \left[ -\theta(c_{ij} + \phi_j^*) \right] \right] & \text{if } i \in \mathcal{U} \\ \sum_{j \in \text{Succ}(i)} p_{ij}^{\text{ref}} (c_{ij} + \phi_j^*) & \text{if } i \in \mathcal{C} \\ 0 & \text{if } i = n \end{cases} \quad (20)$$

where, as usual,  $\text{Succ}(i)$  is the set of successor nodes of node  $i$  in the network and  $\mathcal{U}, \mathcal{C}$  are respectively the sets of unconstrained and constrained nodes. This equation states the necessary optimality conditions for the constrained RSP in terms of the free energy. The first line of this equation is simply the optimality condition previously obtained for the standard RSP (see Equation (A.15) or [28,

<sup>4</sup>Often simply called the free energy.

29)), which should apply on unconstrained nodes. The second line also makes sense as it corresponds to the recurrence expression for computing expected cost for reaching goal node  $n$  from constrained node  $i$  (transition probabilities are fixed on these nodes) [46, 57, 72].

## 5.2 Computing the randomized policy

The previous Equation (20) suggests a simple fixed-point iteration algorithm for computing the solution of the constrained RSP by replacing the equality “=” by an update “ $\leftarrow$ ”. The update is iterated until convergence to a fixed point, in the same way as the value iteration algorithm in Markov decision processes, eventually providing the values of the free energy on each node. Then, the optimal, local, randomized policy can be obtained by Equation (A.17) for unconstrained nodes  $i \in \mathcal{U}$ . For constrained nodes, the transition probabilities are of course fixed to  $p_{ij}^{\text{ref}} = q_{ij}$ .

In [71], it was shown that the iterative update of an expression similar (but somewhat simpler) to Equation (A.15) converges and its limit is independent of the initial values. We prove the same property for the iteration of Equation (20) in Appendix D by using a fixed-point theorem point of view, showing that the update of (20) is a contraction mapping. Besides theoretical convergence, we observed empirically in all our experiments that both techniques (the iterative and the generic constrained RSP procedures) converge and provide exactly the same policies.

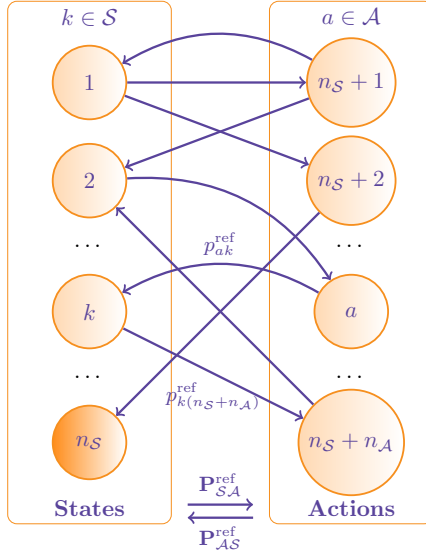
## 6 Markov decision processes as a constrained RSP on a bipartite graph

The previous sections developed all the needed tools for computing an optimal randomized policy on a Markov decision process (MDP), which is done in this section.

Recall that, as in [14], we assume that there is a special cost-free goal state  $n_S$ ; once the system has reached that state, it simply disappears (killing state – state  $n_S$  has no outgoing link). As in [68], we will also consider a problem structure such that termination is inevitable. Thus, the horizon is in effect finite, but its length is random and it depends on the policy being used. The conditions for which this is true are, basically, related to the fact that the goal state can be reached in a finite number of steps from any potential initial state; for a rigorous treatment, see e.g. [14, 15].

The main objective is thus, as before, to design a randomized policy minimizing the expected cost-to-go subject to an (relative) entropy constraint controlling the total randomness spread in the Markov process, and therefore the exploration effort. In other words, we are looking for an optimal policy or, in our case, the optimal transition probabilities matrix of a finite, first-order, Markov chain minimizing the expected cost needed to reach the goal state from the initial state, while fixing the entropy spread in the chain as well as the transition probabilities provided by the environment.

Therefore, the solution is obtained by the algorithms described in Sections 4 and 5, solving the constrained RSP, applied to a bipartite graph, as described now.



**Figure 1:** A simple Markov decision process modeled as a bipartite graph  $G_b$  with states on the left side ( $\mathcal{S}$ ) and control actions on the right ( $\mathcal{A}$ ). Node 1 is the initial state while node  $n_S$  is the absorbing, goal, state of the process. The reference transition probabilities from states to actions  $p_{ka}^{\text{ref}}$  (the reference policy) are gathered in matrix  $\mathbf{P}_{\mathcal{S}\mathcal{A}}^{\text{ref}}$  while the transition probabilities from actions to states  $p_{ak}^{\text{ref}}$ , provided by the environment, are gathered in matrix  $\mathbf{P}_{\mathcal{A}\mathcal{S}}^{\text{ref}}$ .

## 6.1 The basic model

The Markov decision process is now viewed as a constrained randomized shortest paths problem on a bipartite graph (see Figure 1). Let us first describe the structure of this bipartite graph. Then, we examine how the reference transition probabilities, corresponding to the natural random walk on this graph, are defined. Finally, the way to compute the optimal randomized policy is detailed.

### 6.1.1 Definition of the bipartite graph

The process can be modeled as a directed **bipartite graph**  $G_b$  (Figure 1) in which the *left nodes* are the original states  $\mathcal{S}$  and the *right nodes* correspond to the possible actions associated to the states,  $\mathcal{A} = \mathcal{A}(1) \cup \mathcal{A}(2) \cup \dots \cup \mathcal{A}(n_S - 1)$  where  $\mathcal{A}(k)$  is the set of actions available in state  $k$ . Note that the last, goal, state  $n_S$  is absorbing and has no associated action. We thus have  $n_S = |\mathcal{S}|$  left nodes (called *states* or *state nodes*) and  $n_A = |\mathcal{A}|$  right nodes (called *actions* or *action nodes*).

Note that each action associated to a state is a node of  $G_b$ , even if the same action is also available in some other states. In other words, action nodes are duplicated for each state in which they appear. Therefore, the number of such right states is  $|\mathcal{A}| = |\mathcal{A}(1)| + |\mathcal{A}(2)| + \dots + |\mathcal{A}(n_S - 1)| = n_A$ .

Moreover, it is assumed that, in this bipartite graph  $G_b$ , the nodes corresponding to states  $\mathcal{S}$  are numbered first (from 1 to  $n_S$ ) and actions  $\mathcal{A}$  are



following (from  $n_{\mathcal{S}} + 1$  to  $n_{\mathcal{S}} + n_{\mathcal{A}}$ ). Moreover, the set of available actions in any state  $k$  is nothing else than the successor nodes of  $k$  in  $G_b$ ,  $\mathcal{A}(k) = \text{Succ}(k)$ .

### 6.1.2 Defining reference probabilities on the bipartite graph

We will now describe how the reference transition probabilities (see Equation (1)) as well as the constrained nodes are assigned on our graph  $G_b$ . In the case of a *pure, natural, random walk* on  $G_b$ , corresponding to  $T \rightarrow \infty$  in Equation (2), we consider that agents are sent from the initial state 1 and that, at each state  $s = k$  ( $n_{\mathcal{S}}$  states in total), they choose an action  $a$  with probability mass  $p_{ka}^{\text{ref}}$ ,  $k \in \mathcal{S}$  and  $a \in \mathcal{A}(k)$ . When no prior information on the system is available, these are usually set to  $p_{ka}^{\text{ref}} = 1/|\mathcal{A}(k)|$ , a uniform distribution. Agents in state  $k$  then jump to some action node  $a$  with probability  $p_{ka}^{\text{ref}}$ , meaning that they perform the action  $a$  and incur a finite cost  $c_{ka}$  associated to the execution of action  $a$  in state  $k$ .

Furthermore, the agent then moves from action node  $a$  to the next state  $s = l$  with a reference transition probability  $p_{al}^{\text{ref}}$  provided by the environment as in standard Markov decision processes, where  $l \in \mathcal{S}$ , depending on the chosen action. These transition probabilities from action nodes to state nodes cannot be controlled or changed, and correspond therefore to the *constrained* transition probabilities,  $q_{al}$ , as discussed in the previous section describing the constrained RSP.

Thus, in our bipartite graph  $G_b$ , *the set of state nodes  $\mathcal{S}$  is nothing else than the set of unconstrained nodes  $\mathcal{U}$* , together with the absorbing, goal, node  $n_{\mathcal{S}}$ , in the constrained RSP framework. Conversely, *the set of action nodes  $\mathcal{A}$  corresponds exactly to the constrained nodes  $\mathcal{C}$*  because the transition probabilities are fixed by the environment. Consequently, the transition and the cost matrices defined on the bipartite graph  $G_b$  are

$$\mathbf{P}_{\text{ref}} = \begin{matrix} & \mathcal{S} & \mathcal{A} \\ \begin{matrix} \mathcal{S} \\ \mathcal{A} \end{matrix} & \begin{bmatrix} \mathbf{O} & \mathbf{P}_{\mathcal{S}\mathcal{A}}^{\text{ref}} \\ \mathbf{P}_{\mathcal{A}\mathcal{S}}^{\text{ref}} & \mathbf{O} \end{bmatrix} \end{matrix}, \quad \mathbf{C}_b = \begin{matrix} & \mathcal{S} & \mathcal{A} \\ \begin{matrix} \mathcal{S} \\ \mathcal{A} \end{matrix} & \begin{bmatrix} \mathbf{O} & \mathbf{C} \\ \mathbf{O} & \mathbf{O} \end{bmatrix} \end{matrix}, \quad (21)$$

where  $\mathbf{O}$  is a 0 matrix of the appropriate size.

Note that, as for standard Markov decision processes, it is assumed that there is a non-negative cost associated to the transitions between state nodes and action nodes (the cost of choosing the action in the state), while no cost is associated to the transitions between action nodes and state nodes<sup>5</sup>.

### 6.1.3 Computing the optimal randomized policy

Now that the bipartite graph  $G_b$  is defined, solving the MDP problem simply aims at applying the constrained RSP procedure defined in the Section 4 on  $G_b$  (see Algorithm 1). This procedure returns matrix  $\mathbf{P}^*$ , containing the optimal randomized policy  $p_{ka}^*(T)$  for each state node  $k$ . More precisely, the elements  $\{p_{ka}^*(T) : (k \in \mathcal{S}) \wedge (a \in \mathcal{A}(k))\}$  contain, for each state  $k$ , an optimal probability

<sup>5</sup>Note that an additional cost could also be assigned to the transition to state node, after action  $a$  is performed, as, e.g., in [70], but in this work we adopt the simpler setting where the cost is a function of the action  $a$  only. However, our algorithm can straightforwardly be adapted to costs on actions-to-states [61, 70, 75].

distribution on the set  $\mathcal{A}(k)$  of actions available in this state, provided by Equation (A.17), and gradually biasing the walk towards the optimal, deterministic, policy when temperature is low. Indeed, when the temperature  $T$  decreases, the agents are more and more exploiting good policies while still exploring the environment – they interpolate between a purely random behavior (guided by the reference probabilities) and the best, deterministic, policy solving the Markov decision process, provided, e.g., by the well-known value iteration algorithm [62, 70, 12, 14]. This policy is optimal in the sense that it minimizes expected cost for a given degree of relative entropy (see Equation (2)).

In summary, the MDP problem tackled in this section simply corresponds to a constrained randomized shortest-path problem (RSP) on  $G_b$ . We now describe a more direct way for obtaining the optimal randomized policy avoiding the construction of  $G_b$ , and inspired by the value iteration algorithm. It is derived as a special case of the iterative procedure for solving constrained RSP problems developed in Section 5.

## 6.2 A soft value iteration algorithm

Interestingly and surprisingly, we will show in this section that, as for the standard RSP (see Equation (A.15) and its discussion below), replacing the minimum operator by a softmin operator (A.16) in the standard value iteration algorithm recovers exactly the iterative procedure solving the constrained RSP of Section 5 – and providing an optimal randomized policy in the RSP sense to our Markov decision problem. This was already observed in the context of the standard RSP where we obtained a randomized Bellman-Ford recurrence expression where the min operator is replaced by a softmin operator [28, 29].

This implies that the recent propositions of using the softmin function for exploration in reinforcement learning [9, 10, 7, 8, 66, 27, 43, 74, 73] are globally optimal in that they minimize expected path cost subject to a fixed total relative entropy of paths constraint (see Equation (2)), at least in our setting of a absorbing, goal, node  $n_S$  reachable from any other node of the graph.

Interestingly, from Equations (A.10) and (A.13), the cost function (2) can be rewritten at the local, edge, level as  $\sum_{i,j \in \mathcal{V} \setminus n} \bar{n}_{ij} c_{ij} + T \sum_{i \in \mathcal{V} \setminus n} \bar{n}_i \sum_{j \in \text{Succ}(i)} p_{ij} \log(p_{ij}/p_{ij}^{\text{ref}})$  where  $\bar{n}_{ij}$  is the expected flow through edge  $(i, j)$  and  $\bar{n}_i$  the expected number of visits to  $i$  (see [5, 11, 36] and [68], section 6.2). In this expression, the entropy term defined on each node is weighted by the expected number of visits to the node. The policy can thus also be obtained by minimizing this “local” cost function in function of the transition probabilities defined on unconstrained nodes.

### 6.2.1 The standard value iteration algorithm

Let us first recall the standard value iteration procedure, computing the expected cost until absorption by the goal state  $n_S$  [62, 70, 12, 14] when starting from a state  $k \in \mathcal{S}$ , denoted by  $v_k$ , based on the following recurrence formula verified at optimality

$$v_k = \begin{cases} \min_{a \in \mathcal{A}(k)} \left\{ c_{ka} + \sum_{l \in \text{Succ}(a)} p_{al}^{\text{ref}} v_l \right\} & \text{if } k \in \mathcal{S} \setminus \{n_S\} \\ 0 & \text{if } k = n_S \end{cases} \quad (22)$$

where  $v_k$  is the value (expected cost) from state  $k$  and  $p_{al}^{\text{ref}}$  is element  $a, l$  (with  $a \in \mathcal{A}$  and  $l \in \mathcal{S}$ ) of the transition matrix of the reference random walk on the bipartite graph. This expression is iterated until convergence, which is guaranteed under some mild conditions, for any set of nonnegative initial values (see, e.g., [61, 62, 70, 12, 14] for details).

### 6.2.2 The soft value iteration algorithm

Let us start from the standard softmin-based expression computing the free energy directed distance in a regular graph (Equation (A.15); see also [28, 29, 26]). We observe that it corresponds to the Bellman-Ford expression providing the shortest-path distance in which the min operator has been replaced by the softmin operator defined in Equation (A.16).

Substituting in the same way the min operator (A.16) for the softmin, with the  $p_{ka}^{\text{ref}}$  playing the role of the weighting factors  $q_i$ , in the value iteration update formula (22) provides a “soft” equivalent of the Bellman-Ford optimality conditions on the set of state nodes  $\mathcal{S}$ ,

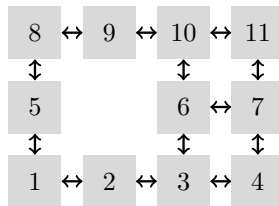
$$\phi_k^{\mathcal{S}} = \begin{cases} -\frac{1}{\theta} \log \left[ \sum_{a \in \mathcal{A}(k)} p_{ka}^{\text{ref}} \exp \left[ -\theta \left( c_{ka} + \sum_{l \in \text{Succ}(a)} p_{al}^{\text{ref}} \phi_l^{\mathcal{S}} \right) \right] \right] & \text{if } k \in \mathcal{S} \setminus \{n_{\mathcal{S}}\} \\ 0 & \text{if } k = n_{\mathcal{S}} \end{cases} \quad (23)$$

In the case of our bipartite graph of Figure 1, this equation can exactly be obtained by applying the recurrence expression computing the free energy in the constrained RSP (Equation (20)), after recalling that the cost of the transition between an action node and a state node is equal to zero. More precisely, we simply substitute  $\phi_j^*$  in the first line of Equation (20) by the expression in the second line,  $\phi_j^* = \sum_{l \in \text{Succ}(j)} p_{jl}^{\text{ref}} (0 + \phi_l^*)$ , which directly provides Equation (23). Recall that the  $p_{ka}^{\text{ref}}$ ,  $k \in \mathcal{S}$  and  $a \in \mathcal{A}(k)$ , correspond to the reference, prior, policy commonly set to a uniform distribution on the possible actions in state  $k$ ,  $p_{ka}^{\text{ref}} = 1/|\mathcal{A}(k)|$ . Conversely, the  $p_{ak}^{\text{ref}}$  with  $a \in \mathcal{A}$  and  $k \in \mathcal{S}$  are provided by the environment.

Note that it can easily be shown by following the same reasoning as in the appendix of [28, 29] that this recurrence formula reduces to the standard optimality conditions for Markov decision processes (Equation (22)) when  $\theta \rightarrow \infty$ . Conversely, when  $\theta \rightarrow 0^+$ , it reduces to the expression allowing to compute the expected cost until absorption by the goal state  $n_{\mathcal{S}}$ , also called the average first-passage cost [46, 57, 72],  $\phi_k = \sum_{a \in \mathcal{A}(k)} p_{ka}^{\text{ref}} (c_{ka} + \sum_{l \in \text{Succ}(a)} p_{al}^{\text{ref}} \phi_l)$ .

The idea is to iterate (23) until convergence of the free energies to a fixed point where the optimality conditions (23) are verified (no change occurs any more). The procedure converges to a unique solution as it corresponds to a particular case of the iterative procedure for solving the constrained RSP (Equation (20)); see Appendix D for the proof. Then, the optimal policy for each state  $k \in \mathcal{S}$ ,  $k \neq n_{\mathcal{S}}$ , is computed thanks to Equation (A.17), which provides the probability of choosing action  $a$  within state  $k$ .

This procedure, involving the iteration of Equation (23) and the computation of the optimal policy from Equation (A.17), will be called the **soft value iteration** algorithm. As already stated, such soft variants of value iteration already appeared in control [76] and exploration strategies for which



**Figure 2:** The maze problem. The goal of the agent is to reach node 11 from node 1. Notice that some transitions with no resulting displacement are possible (example in node 1: going west or south). The costs related to the actions are detailed in the text.

an additional Kullback-Leibler cost term is incorporated in the immediate cost [66, 27, 43, 74, 73, 9, 10]. It was also recently proposed as an operator guiding exploration in reinforcement learning, and more specifically for the SARSA algorithm in [7, 8]. The present work therefore provides a new interpretation to this exploration strategy. We apply this algorithm in the next Section 7 in order to solve simple Markov decision problems, for illustration.

### 6.3 Markov decision processes with discounting

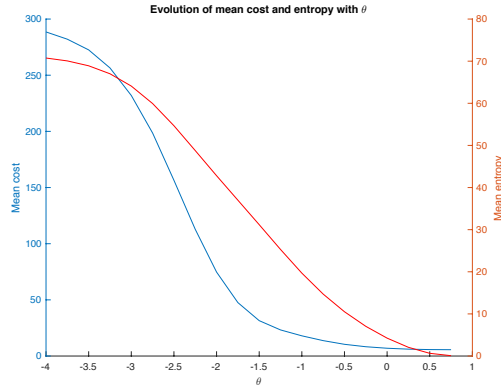
Finally, let us briefly discuss the concept of MDP with discounting. In this setting, we still consider the random walk on the graph  $G_b$  with reference transition probabilities  $\mathbf{P}_{\text{ref}}$ . However, in contrast with our previous setting, here, no goal node is defined – the Markov chain defining the random walk is regular. In addition, a **discounting factor**  $\gamma \in ]0, 1[$  is introduced to decrease the impact of future costs with time [70]: immediate costs are more important than postponed ones.

In standard MDPs, the introduction of the discounting factor can be interpreted from two different points of view:

- ▶ each future cost, for instance appearing at time step  $t$ , is reduced by a factor  $\gamma^t$ .
- ▶ at each time step, the random walker has a small chance  $(1 - \gamma)$  of quitting the process (the contract is cancelled, the agent is killed, etc).

In the case of standard MDPs, these two interpretations lead to the same model; however, in the RSP framework, they take distinct forms. They are left for further work, but we quickly introduce the intuition behind them.

The first interpretation leads to a new soft value iteration expression that has to be iterated for paths with increasing length. This can be done by unfolding the network in time and then apply the RSP on this new directed acyclic graph, as described in [52]. For the second interpretation, the problem can be tackled by introducing a cemetery node (a killing, absorbing state). The agent then has a  $(1 - \gamma)$  probability of being teleported to this cemetery state with a zero cost after choosing any action. The soft value iteration expression (23) can be adapted to this new setting. These two RSP with discounting models will be investigated in further work.



**Figure 3:** Results, averaged over  $10^6$  runs, obtained by simulating the policy provided by the constrained RSP when increasing  $\theta$  (in log scale). The blue curve depicts the evolution of the average cost (mean number of turns to reach square 11 – the smaller the best) in function of  $\theta$ . The red curve indicates the corresponding entropy of the state nodes (entropy of the randomized policies). Naturally, the largest entropy and average cost are achieved when  $\theta$  is small and are minimum when  $\theta$  is large.

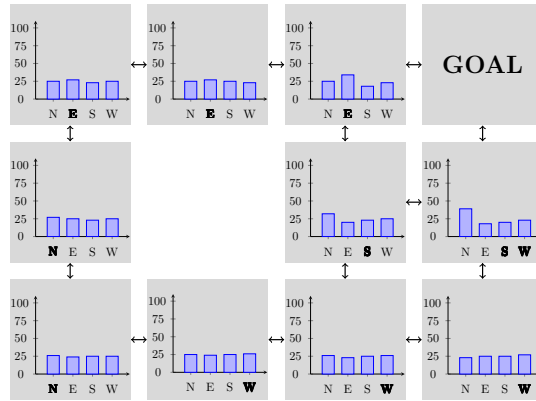
## 7 Some simulations: a simple illustration on the maze problem

This section illustrates the application of constrained randomized shortest paths to Markov decision problems. Several simulations were run on four different problems [50] but, in order to save space and because the conclusions are similar, we decided to report only one simple application: the probabilistic maze game inspired by [67], as illustrated in Figure 2.

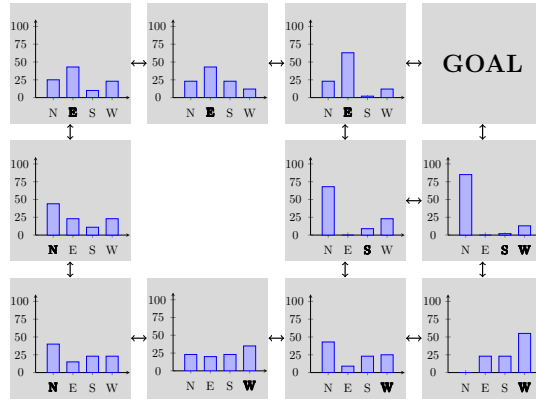
An agent, initially starting on square 1, is asked to reach square 11 in a minimum number of time steps (see Figure 2) and incurs some additional costs described below. To do so, the agent can choose between four actions in each square:

- ▶ Go north. However, a bug (for instance due to adverse wind conditions) can occur when the agent decides to go north so that it has only a 0.8 probability to actually go north (no bug). When this bug occurs (0.2 probability), it then has a 0.5 probability to go east (globally, a 10% chance) and a 0.5 probability to go west (also a 10% chance globally).
- ▶ Go east, west or south with probability one.

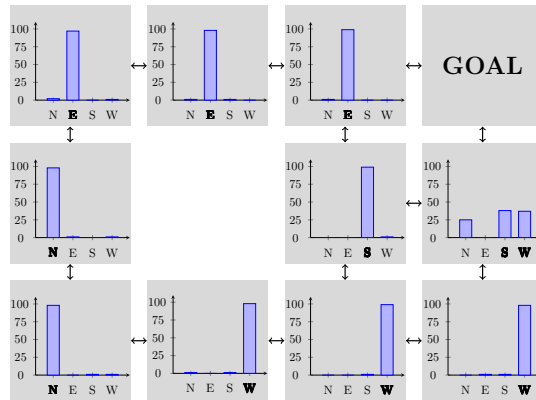
A unit cost is associated to each time step and, in addition, the cost for visiting square 7 is +100. This implies that if the above mentioned bug occurs on square 7 and the agent is redirected to east, the cost is increased again by +100 because the agent re-enters square 7. Indeed, if the agent selects a direction which leads nowhere (a wall), for example selecting “go east” or “go west” in square 5, it stays in its current position, but incurs (again) the cost associated to the current square.



(a)  $\uparrow$  Inverse temperature  $\theta = 10^{-2.5}$ .



(b)  $\uparrow$  Inverse temperature  $\theta = 10^{-1}$ .



(c)  $\uparrow$  Inverse temperature  $\theta = 10^{+0.5}$ .

**Figure 4:** Optimal randomized policy obtained after convergence of the soft value iteration algorithm for three different, increasing, values of the inverse temperature parameter  $\theta$ . In each square, the agent has to choose between going north (N), east (E), south (S) and west (W). A larger  $\theta$  corresponds to a more deterministic policy. Note that the optimal deterministic policy is indicated in bold in each case.

Concerning the reference probabilities  $p_{ij}^{\text{ref}}$ , these are defined by the environment on action nodes and are set to  $1/4$  on state nodes (a purely random policy).

Note that the optimal policy from square 1 is  $1 \rightarrow 5 \rightarrow 8 \rightarrow 9 \rightarrow 10 \rightarrow 11$ . We ran a large number of simulations of the process until the agent reaches the goal node (each simulation from initial state 1 to goal state 11 is called a run), for a range of randomized policies (depending on the parameter  $\theta$ ) obtained after running the soft value iteration algorithm (23).

Figure 3 represents the evolution of the mean cost to reach goal node 11 and the entropy (computed only on state nodes, thus on the randomized policies) in function of  $\theta$ . The results are averaged over  $10^6$  runs for each value of  $\theta$ , with the policy obtained after convergence of the soft value iteration algorithm for this  $\theta$  (see Equation (23)). We observe that the largest average cost and entropy are achieved when  $\theta$  is small, and are smallest when  $\theta$  is large. The resulting functions are both logistic-shaped between two bounds:

- ▶ When  $\theta$  is small, entropy is maximum as each action has approximately a  $1/4$  probability to be chosen and therefore the expected scores are the same as for a pure random walk.
- ▶ When  $\theta$  becomes large, entropy and expected cost (as well as the policy) are almost the same as for the standard value iteration algorithm providing the optimal deterministic policy.

Moreover, the optimal randomized policies obtained after convergence of the soft value iteration algorithm for three different, increasing, values of the inverse temperature parameter  $\theta$  are illustrated in Figure 4. This example clearly shows that using a randomized strategy allows to balance the strength of the player.

## 8 Conclusion

This work presented two procedures for solving constrained randomized shortest-paths problems, together with an application to randomized Markov decision processes where the problem is viewed as a bipartite graph. The main objective is to reach a goal node from an initial node in a graph while minimizing expected cost subject to a relative entropy equality constraint and transition probabilities constraints on some edges. The model provides a randomized policy encouraging exploration, balancing exploitation and exploration. The amount of exploration is monitored by the inverse temperature parameter.

The problem is expressed in terms of full paths connecting the initial node to the goal node and can easily be solved. The solution is a Gibbs-Boltzmann probability distribution on the set of paths with virtual extra costs associated to the constrained edges.

Two algorithms for computing the local policy at the edge level are developed. The first algorithm is based on Lagrange duality and requires solving iteratively the standard randomized shortest-paths problem until convergence. The second algorithm is reminiscent of Bellman-Ford’s algorithm for solving the shortest-path distance problem. It simply aims to replace the min operator by a softmin operator in Bellman-Ford’s recurrence relation to update the expected cost on unconstrained nodes. For the constrained nodes, because the transition

probabilities are fixed, we simply use the expression for computing the expected cost until absorption in a Markov chain. The convergence of the procedure is guaranteed for the two algorithms.

The usefulness of these algorithms is then illustrated on standard Markov decision problems. Indeed, a standard Markov decision process can be reinterpreted as a randomized shortest-paths problem on a bipartite graph. Standard Markov decision problems can thus easily be solved by the two introduced algorithms: they provide a randomized policy minimizing expected cost under entropy and transition probabilities constraints.

This shows that the exploration strategy using the softmin instead of the min in the value iteration algorithm is optimal in the predefined sense. Therefore it justifies the previous work [7, 8, 9, 10, 27, 43, 66, 74, 73, 76] from a randomized shortest-paths point of view.

Future work will focus on extending the randomized shortest-paths model in order to deal with other types of constraints. In particular we will work on inequality constraints on transition probabilities, as well as flow equality and inequality constraints, on both node flows and edge flows. Another interesting extension of the randomized shortest-paths model is the multi-sources multi-destinations randomized optimal transport on a graph generalizing the deterministic optimal transport on a graph problem. We also plan to investigate constrained randomized shortest paths with a discounting factor as well as average-reward Markov decision processes which were recently studied in the light of entropy regularization [56].

## Acknowledgements

This work was partially supported by the Immediate and the Brufence projects funded by InnovIris (Brussels region), as well as former projects funded by the Walloon region, Belgium. We thank these institutions for giving us the opportunity to conduct both fundamental and applied research.

We also thank Benjamin Blaise, a former Master student, who helped us to investigate the randomized Markov decision processes during his masters thesis at UCLouvain [16], as well as Prof. Fabrice Rossi for his useful remarks.

---

## Appendix: Additional material and proofs of the main results

### A Computing quantities of interest

In this appendix, several important quantities derived from the standard randomized shortest-paths (RSP) framework, and which will be needed in the paper, are detailed. The material is mainly taken from the previous work [80, 68, 48, 28, 29, 26].



**The minimum free energy.** Interestingly, if we replace the probability distribution  $\mathbb{P}$  by the optimal distribution  $\mathbb{P}^*$  provided by Equation (3) in the objective function (2), we obtain for the minimum **free energy** between node 1 and node  $n$

$$\begin{aligned}\phi_1^*(T) &= \phi(\mathbb{P}^*) = \sum_{\varphi \in \mathcal{P}} P^*(\varphi) \tilde{c}(\varphi) + T \sum_{\varphi \in \mathcal{P}} P^*(\varphi) \log \left( \frac{P^*(\varphi)}{\tilde{\pi}(\varphi)} \right) \\ &= \sum_{\varphi \in \mathcal{P}} P^*(\varphi) \tilde{c}(\varphi) + T \sum_{\varphi \in \mathcal{P}} P^*(\varphi) \left( -\frac{1}{T} \tilde{c}(\varphi) - \log \mathcal{Z} \right) \\ &= -T \log \mathcal{Z} = -\frac{1}{\theta} \log \mathcal{Z}\end{aligned}\tag{A.1}$$

**The expected number of passages through edges.** For the *expected number of passages through an edge*  $(i, j)$  at temperature  $T = 1/\theta$ , that is, the flow in  $(i, j)$ , we obtain from last result (A.1) and the definition of the partition function  $\mathcal{Z}$  (Equation (3)),

$$\begin{aligned}\frac{\partial \phi_1^*}{\partial c_{ij}} &= \frac{\partial(-\frac{1}{\theta} \log \mathcal{Z})}{\partial c_{ij}} = -\frac{1}{\theta \mathcal{Z}} \frac{\partial \mathcal{Z}}{\partial c_{ij}} = -\frac{1}{\theta \mathcal{Z}} \sum_{\varphi \in \mathcal{P}} \tilde{\pi}(\varphi) \exp[-\theta \tilde{c}(\varphi)] (-\theta) \frac{\partial \tilde{c}(\varphi)}{\partial c_{ij}} \\ &= \sum_{\varphi \in \mathcal{P}} \frac{\tilde{\pi}(\varphi) \exp[-\theta \tilde{c}(\varphi)]}{\mathcal{Z}} \frac{\partial \tilde{c}(\varphi)}{\partial c_{ij}} = \sum_{\varphi \in \mathcal{P}} P^*(\varphi) \eta((i, j) \in \varphi) \triangleq \bar{n}_{ij}(T)\end{aligned}\tag{A.2}$$

where we used  $\partial \tilde{c}(\varphi) / \partial c_{ij} = \eta((i, j) \in \varphi)$ , with  $\eta((i, j) \in \varphi)$  being the number of times edge  $(i, j)$  appears on path  $\varphi$  at temperature  $T$ . Therefore, we have for the flow in  $(i, j)$

$$\bar{n}_{ij}(T) = -T \frac{\partial \log \mathcal{Z}}{\partial c_{ij}}\tag{A.3}$$

**Computation of the partition function.** Now, it turns out that the partition function can easily be computed in closed form (see, e.g., [68, 47, 26] for details). Let us first introduce the **fundamental matrix** of the RSP system,

$$\mathbf{Z} = \mathbf{I} + \mathbf{W} + \mathbf{W}^2 + \dots = (\mathbf{I} - \mathbf{W})^{-1}, \quad \text{with } \mathbf{W} = \mathbf{P}_{\text{ref}} \circ \exp[-\theta \mathbf{C}]\tag{A.4}$$

where  $\mathbf{C}$ ,  $\mathbf{P}_{\text{ref}}$  are respectively the cost and the reference transition probabilities matrices (see Equation (1)) while  $\circ$  is the elementwise (Hadamard) product. Elementwise, the entries of the  $\mathbf{W}$  matrix are  $w_{ij} = [\mathbf{W}]_{ij} = p_{ij}^{\text{ref}} \exp[-\theta c_{ij}]$ , except for the goal node where  $w_{nj} = 0$  for all  $j$  (killing, absorbing, node). Note that this matrix is sub-stochastic because the costs are non-negative and node  $n$  is absorbing and killing (row  $n$  contains only 0 values).

Then, the partition function is simply  $\mathcal{Z} = [\mathbf{Z}]_{1n} = z_{1n}$  (see [80, 68, 48, 28, 29]). More generally [30, 28], it can be shown that the elements  $z_{in}$  of the fundamental matrix correspond to

$$z_{in} = \sum_{\varphi \in \mathcal{P}_{in}} \tilde{\pi}(\varphi) \exp[-\theta \tilde{c}(\varphi)]\tag{A.5}$$

with  $z_{nn} = 1$ , and where  $\mathcal{P}_{in}$  is the set of hitting paths starting in node  $i$  and ending in killing absorbing node  $n$ . The  $z_{in}$  quantities are usually called the backward variables. They can be interpreted as probabilities of surviving during a killed random walk with transition matrix  $\mathbf{W}$ , that is, reaching hitting node  $n$  without being killed during the walk (see, e.g., [28, 29] for details).

**Computation of the expected number of passages and visits.** Moreover, the flow in  $(i, j)$  can be obtained from (A.4) and the expression  $\partial \mathbf{M}^{-1} / \partial x = -\mathbf{M}^{-1} (\partial \mathbf{M} / \partial x) \mathbf{M}^{-1}$  (see, e.g., [37]),

$$\bar{n}_{ij}(T) = -\frac{1}{\theta} \frac{\partial \log \mathcal{Z}}{\partial c_{ij}} = \frac{z_{1i} p_{ij}^{\text{ref}} \exp[-\theta c_{ij}] z_{jn}}{z_{1n}} = \frac{z_{1i} w_{ij} z_{jn}}{z_{1n}} \quad (\text{A.6})$$

and because only the first row and the last column of  $\mathbf{Z}$  are needed, two systems of linear equations can be solved instead of matrix inversion in Equation (A.4).

From the last equation and  $z_{in} = \sum_{j=1}^n w_{ij} z_{jn} + \delta_{in}$  (the elementwise form of  $(\mathbf{I} - \mathbf{W})\mathbf{Z} = \mathbf{I}$ ), the *expected number of visits to a node  $j$*  can be computed from

$$\bar{n}_i(T) \triangleq \sum_{j=1}^n \bar{n}_{ij}(T) + \delta_{in} = \frac{z_{1i} z_{in}}{z_{1n}} \quad \text{for } i \neq n \quad (\text{A.7})$$

where we assume  $i \neq n$  for the last equality because we already know that  $\bar{n}_n(T) = 1$  at the goal node, which is absorbing and killing.

**The optimal randomized policy.** Furthermore, from (A.6)-(A.7), the optimal transition probabilities of following an edge  $(i, j)$  with  $i \neq n$  are

$$p_{ij}^*(T) = \frac{\bar{n}_{ij}(T)}{\bar{n}_i(T)} = p_{ij}^{\text{ref}} \exp[-\theta c_{ij}] \frac{z_{jn}}{z_{in}} = \frac{w_{ij} z_{jn}}{z_{in}} = \frac{w_{ij} z_{jn}}{\sum_{j'=Succ(i)} w_{ij'} z_{j'n}} \quad (\text{A.8})$$

because  $p_{ij}^{\text{ref}} \exp[-\theta c_{ij}] = w_{ij}$  and  $z_{in} = \sum_{j=Succ(i)} w_{ij} z_{jn}$  for all  $i \neq n$  (the elementwise form of  $(\mathbf{I} - \mathbf{W})\mathbf{Z} = \mathbf{I}$ , coming from Equation (A.4)). This expression defines a **biased random walk** on  $G$  – the random walker is “attracted” by the goal node  $n$ . These transition probabilities define a first-order Markov chain and do not depend on the source node. They correspond to the optimal, randomized, “routing” strategy, or policy, minimizing free energy from the current node. This policy will therefore be called the **randomized policy** in the sequel (a mixed policy or strategy in game theory [58]).

**The expected cost until destination.** In addition, the expected cost until reaching goal node  $n$  from node 1 is [68, 47, 26]

$$\langle \tilde{c} \rangle = \sum_{\varphi \in \mathcal{P}} \mathbf{P}^*(\varphi) \tilde{c}(\varphi) = \sum_{\varphi \in \mathcal{P}} \frac{\tilde{\pi}(\varphi) \exp[-\theta \tilde{c}(\varphi)]}{\mathcal{Z}} \tilde{c}(\varphi) \quad (\text{A.9})$$

After defining the matrix containing the expected number of passages through the edges by  $\mathbf{N}$  with  $[\mathbf{N}]_{ij} = \bar{n}_{ij}(T)$ , it can be shown by proceeding in the same way as for Equation (A.2) (see [68] for details) that the expected cost spent in the network is

$$\langle \tilde{c} \rangle = -\frac{\partial \log \mathcal{Z}}{\partial \theta} = \mathbf{e}^T (\mathbf{N} \circ \mathbf{C}) \mathbf{e} \quad (\text{A.10})$$

where  $\mathbf{e}$  is a column vector of 1s and  $\circ$  is the elementwise (Hadamard) matrix product. This quantity is just the cumulative sum of the expected number of passages through each edge times the cost of following the edge,  $\sum_{i=1}^{n-1} \sum_{j \in Succ(i)} \bar{n}_{ij}(T) c_{ij}$  [30].

**The entropy of the paths.** In Equation (2), the relative entropy of the set of paths, for the optimal probability distribution, was defined as

$$J(\mathbb{P}^*|\tilde{\pi}) = \sum_{\varphi \in \mathcal{P}} \mathbb{P}^*(\varphi) \log \left( \frac{\mathbb{P}^*(\varphi)}{\tilde{\pi}(\varphi)} \right) \quad (\text{A.11})$$

and, from Equations (2), (A.1) and (A.10), can be computed thanks to

$$J(\mathbb{P}^*|\tilde{\pi}) = -(\log \mathcal{Z} + \frac{1}{T} \langle \tilde{c} \rangle) \quad (\text{A.12})$$

where the partition function  $\mathcal{Z} = [\mathbf{Z}]_{1n} = z_{1n}$ .

In addition, it can be shown that the total entropy of the set of paths is [5, 68]

$$J(\mathbb{P}^*) = - \sum_{\varphi \in \mathcal{P}} \mathbb{P}^*(\varphi) \log \mathbb{P}^*(\varphi) = - \sum_{i=1}^{n-1} \bar{n}_i \sum_{j \in \text{Succ}(i)} p_{ij}^*(T) \log(p_{ij}^*(T)) \quad (\text{A.13})$$

which sums the local entropies over the transient (non-absorbing) nodes weighted by the expected number of visits to each node.

**The free energy distance.** It was already shown in Equation (A.1) that the minimal free energy (A.1) at temperature  $T$  is provided by

$$\phi_1^*(T) = \phi(\mathbb{P}^*) = -T \log \mathcal{Z} = -\frac{1}{\theta} \log z_{1n} \quad (\text{A.14})$$

In [28, 29], it was proved that the free energy from any starting node  $i$  to absorbing, goal, node  $n$ ,  $\phi_i^*(T) = -\frac{1}{\theta} \log z_{in}$ , can be computed thanks to the following recurrence formula to be iterated until convergence

$$\phi_i^*(T) = \begin{cases} -\frac{1}{\theta} \log \left[ \sum_{j \in \text{Succ}(i)} p_{ij}^{\text{ref}} \exp[-\theta(c_{ij} + \phi_j^*(T))] \right] & \text{if } i \neq n \\ 0 & \text{if } i = n \end{cases} \quad (\text{A.15})$$

This equation is an extension of Bellman-Ford's formula for computing the shortest-path distance in a graph (see, e.g., [14, 19, 21, 33, 42, 64, 69]). Moreover, the recurrence expression (A.15) is also a generalization of the distributed consensus algorithm developed in [71], considering binary costs only.

It was also shown [28, 29] that this minimal free energy interpolates between the least cost ( $T = \theta^{-1} \rightarrow \infty$ ;  $\phi_i^*(\infty) = \min_{j \in \text{Succ}(i)} \{c_{ij} + \phi_j^*(\infty)\}$  and  $\phi_n^*(\infty) = 0$ ) and the expected cost before absorption ( $T = \theta^{-1} \rightarrow 0^+$ ;  $\phi_i^*(0) = \sum_{j \in \text{Succ}(i)} p_{ij}^{\text{ref}} (c_{ij} + \phi_j^*(0))$  and  $\phi_n^*(0) = 0$ ) [48, 28, 29]. In addition, this quantity defines a **directed distance** between any node and absorbing node  $n$  [48, 28, 29]. This directed free energy distance has a nice interpretation: it corresponds (up to a scaling factor) to minus the logarithm of the probability of reaching node  $n$  without being killed during a killed random walk defined by the sub-stochastic transition probabilities  $w_{ij} = p_{ij}^{\text{ref}} \exp[-\theta c_{ij}]$  [28, 29]. In other words, it is minus the logarithm of the probability of *surviving* during the walk. Still another interesting result is that, when computing the continuous time – continuous state equivalent to the RSP model by densifying the graph, the free energy becomes a *potential* attracting the agents to the goal state [32].

**The softmin operator.** In fact, as discussed in [28, 29], this last expression (A.15) is obtained by simply substituting the min operator by a weighted version of the **softmin operator** ([20]; also called the **log-sum-exp** function [17, 55, 71]) in the Bellman-Ford recurrence formula,

$$\text{softmin}_{\mathbf{q},\theta}(\mathbf{x}) = -\frac{1}{\theta} \log \left( \sum_{j=1}^n q_j \exp[-\theta x_j] \right), \text{ with all } q_j \geq 0 \text{ and } \sum_{j=1}^n q_j = 1 \quad (\text{A.16})$$

which is a smooth approximation of the min operator and interpolates between weighted average and minimum operators, depending on the parameter  $\theta$  [20, 71]. This expression also appeared in control [76] and exploration strategies for which an additional Kullback-Leibler cost term is incorporated in the immediate cost [66, 27, 43, 74, 73, 9, 10]. Moreover, this function<sup>6</sup> was recently proposed as an operator guiding exploration in reinforcement learning, and more specifically for the SARSA algorithm [7, 8] – see these references for a discussion of its properties.

**The randomized policy in terms of free energy.** Note that the optimal randomized policy derived in Equation (A.8) can be rewritten in function of the free energy as

$$p_{ij}^*(T) = \frac{p_{ij}^{\text{ref}} \exp[-\theta c_{ij}] z_{jn}}{\sum_{j'=1}^n p_{ij'}^{\text{ref}} \exp[-\theta c_{ij'}] z_{j'n}} = \frac{p_{ij}^{\text{ref}} \exp[-\theta(c_{ij} + \phi_j^*(T))]}{\sum_{j'=1}^n p_{ij'}^{\text{ref}} \exp[-\theta(c_{ij'} + \phi_{j'}^*(T))]} \quad (\text{A.17})$$

because  $z_{in} = \exp[-\theta \phi_i^*(T)]$  and  $z_{in} = \sum_{j=1}^n w_{ij} z_{jn} = \sum_{j=1}^n p_{ij}^{\text{ref}} \exp[-\theta c_{ij}] z_{jn}$  for all  $i \neq n$ . This corresponds to a multinomial logistic function.

## B Solving the system of logistic equations

In this appendix, we are mainly interested in deriving the solution of a simple system of multinomial logistic equations. Assume we have to solve the following equations

$$\frac{\gamma_i \exp[-\theta x_i]}{\sum_{j=1}^n \gamma_j \exp[-\theta x_j]} = q_i \quad \text{with each } q_i, \gamma_i \geq 0 \quad (\text{B.1})$$

with respect to the  $x_i$ , together with the following equality constraints

$$\begin{cases} \sum_{i=1}^n q_i = 1 \\ \sum_{i=1}^n q_i x_i = 0 \end{cases} \quad (\text{B.2})$$

The multinomial logistic function in (B.1) is often encountered in applied statistics, for instance it forms the main functional form of the multivariate logistic model [38]. In this appendix, we derive the solution  $\mathbf{x}^*$  of this equation satisfying the given constraints and then use it in order to solve Equation (15). The second equality constraint in (B.2) is introduced because any shift of a solution vector,  $\mathbf{x}^* - \mathbf{c}$ , is also a solution. Adding this second constraint solves the problem of degeneracy.

---

<sup>6</sup>They actually study the softmax counterpart.

Taking the ratio between the two equations (B.1) involving  $q_i$  and  $q_j$  and taking  $-\frac{1}{\theta}$  log of both sides gives  $x_i - x_j = -\frac{1}{\theta}[\log(q_i/\gamma_i) - \log(q_j/\gamma_j)]$ . This provides  $n - 1$  independent equations and a common practice is to set one value to 0, for instance  $x_n = 0$  [38]. Here, we will instead force the second equality constraint (B.2). Multiplying both sides by  $q_j$  and summing over  $j$  provides  $x_i - \sum_{j=1}^n q_j x_j = -\frac{1}{\theta}[\log(q_i/\gamma_i) - \sum_{j=1}^n q_j \log(q_j/\gamma_j)]$  (recall that the  $q_i$  sum to 1 and  $\sum_{i=1}^n q_i x_i = 0$ ) gives

$$x_i = -\frac{1}{\theta} \left( \log(q_i/\gamma_i) - \sum_{j=1}^n q_j \log(q_j/\gamma_j) \right) \quad (\text{B.3})$$

We now apply this result in order to solve Equation (15) with  $x_j = \Delta_{ij}$  (we condition the computation on an arbitrary node  $i$ ). By comparing (15) with (B.1) as well as recalling that  $p_{ij}^{\text{ref}} = q_{ij}$  and  $\phi_i^* = -\frac{1}{\theta} \log z_{in}$  (Equation (12)), we observe that  $\gamma_j = p_{ij}^{\text{ref}} \exp[-\theta c_{ij}] z_{jn}$  and therefore  $-\frac{1}{\theta} \log(q_j/\gamma_j) = \frac{1}{\theta} \log(\exp[-\theta c_{ij}] z_{jn}) = -(c_{ij} + \phi_j^*)$ . Injecting this result in (B.3) finally provides for constrained nodes

$$\Delta_{ij} = -(c_{ij} + \phi_j^*) + \sum_{k \in \text{Succ}(i)} p_{ik}^{\text{ref}} (c_{ik} + \phi_k^*) \quad (\text{B.4})$$

which is the required result.

## C Derivation of the iterative algorithm

In order to compute the optimal policy  $p_{ij}^*$ , we observe from Equation (A.17) that we need to find the free energy  $\phi_j^* = -\log z_{jn}$ , and thus the backward variable  $z_{jn}$  starting from a node  $j$ ,

$$p_{ij}^* \propto p_{ij}^{\text{ref}} \exp[-\theta(c_{ij} + \phi_j^*)]$$

where  $\propto$  means ‘‘proportional to’’. The quantity  $p_{ij}^*$  then needs to be normalized so that  $\sum_{j \in \text{Succ}(i)} p_{ij}^* = 1$ . We will therefore have to compute the backward variable  $z_{jn}$  for the two sets of nodes of interest, the constrained nodes  $\mathcal{C}$  and the unconstrained nodes  $\mathcal{U}$ .

From the definition of the backward variable (Equation (A.5), but now including the augmented costs on constrained nodes), we obtain by decomposing the paths  $i \rightsquigarrow n$  into the first step  $i \rightarrow j$ , and then the remaining steps  $j \rightsquigarrow n$  (see [30] for a related derivation),

$$\begin{aligned} z_{in} &= \sum_{\wp_{in} \in \mathcal{P}_{in}} \tilde{\pi}(\wp_{in}) \exp[-\theta \tilde{c}'(\wp_{in})] \\ &= \sum_{j \in \text{Succ}(i)} \sum_{\wp_{jn} \in \mathcal{P}_{jn}} p_{ij}^{\text{ref}} \tilde{\pi}(\wp_{jn}) \exp[-\theta(c'_{ij} + \tilde{c}'(\wp_{jn}))] \\ &= \sum_{j \in \text{Succ}(i)} p_{ij}^{\text{ref}} \exp[-\theta c'_{ij}] \underbrace{\sum_{\wp_{jn} \in \mathcal{P}_{jn}} \tilde{\pi}(\wp_{jn}) \exp[-\theta \tilde{c}'(\wp_{jn})]}_{z_{jn}} \\ &= \sum_{j \in \text{Succ}(i)} p_{ij}^{\text{ref}} \exp[-\theta c'_{ij}] z_{jn} \end{aligned} \quad (\text{C.1})$$

where  $\wp_{in}$  is a path starting in a node  $i$  and ending in the killing, absorbing, node  $n$ . We will now express this recurrence formula in terms of the free energy, which will lead to an interesting extension of the Bellman-Ford formula.

Taking  $-\frac{1}{\theta}$  log of this last expression and recalling that  $\phi_i^* = -\frac{1}{\theta} \log z_{in}$  yields, for any node  $i \neq n$ ,

$$\phi_i^* = -\frac{1}{\theta} \log \sum_{j \in \text{Succ}(i)} p_{ij}^{\text{ref}} \exp[-\theta(c'_{ij} + \phi_j^*)] \quad (\text{C.2})$$

The remainder of the development depends on the type of node  $i$ ; we therefore continue with the unconstrained nodes, followed by the constrained ones.

### C.1 Computation of the free energy on unconstrained nodes

For unconstrained nodes,  $c'_{ij} = c_{ij}$  and we simply have

$$\phi_i^* = -\frac{1}{\theta} \log \sum_{j \in \text{Succ}(i)} p_{ij}^{\text{ref}} \exp[-\theta(c_{ij} + \phi_j^*)] \quad \text{for each } i \in \mathcal{U} \quad (\text{C.3})$$

because there is no augmented cost associated to the transitions from an unconstrained node – they are not part of the set of constrained transitions (see Subsection 6.1). This corresponds to the standard recurrence formula for computing the free energy in the RSP framework (see Equation (A.15) or [28, 29]). Let us now compute this quantity on constrained nodes.

### C.2 Computation of the free energy on constrained nodes

In the case of constrained nodes, we have to use the augmented costs  $c'_{ij}$  in order to ensure that the relative flow in the edge  $(i, j)$  is equal to the predefined transition probability  $p_{ij}^{\text{ref}}$  provided by the environment. Remember that the value of these augmented costs can be expressed in function of the free energy,  $c'_{ij} = \sum_{l \in \text{Succ}(i)} p_{il}^{\text{ref}} (c_{il} + \phi_l^*) - \phi_j^*$  (Equation (17)). Injecting this result in Equation (C.2) provides

$$\begin{aligned} \phi_i^* &= -\frac{1}{\theta} \log \sum_{j \in \text{Succ}(i)} p_{ij}^{\text{ref}} \exp[-\theta(c'_{ij} + \phi_j^*)] \\ &= -\frac{1}{\theta} \log \sum_{j \in \text{Succ}(i)} p_{ij}^{\text{ref}} \exp \left[ -\theta \left( \sum_{l \in \text{Succ}(i)} p_{il}^{\text{ref}} (c_{il} + \phi_l^*) \right) \right] \\ &= -\frac{1}{\theta} \log \left[ \exp \left[ -\theta \left( \sum_{l \in \text{Succ}(i)} p_{il}^{\text{ref}} (c_{il} + \phi_l^*) \right) \right] \left( \sum_{j \in \text{Succ}(i)} p_{ij}^{\text{ref}} \right) \right] \\ &= \sum_{l \in \text{Succ}(i)} p_{il}^{\text{ref}} (c_{il} + \phi_l^*), \quad \text{for each } i \in \mathcal{C} \end{aligned} \quad (\text{C.4})$$

Moreover, for the goal node  $n$ ,  $z_{nn} = 1$  so that  $\phi_n^* = 0$ . This last result as well as Equations (C.3)-(C.4) therefore justify the recurrence formula (20).

## D Convergence of the iterative algorithm

In this appendix, the convergence of the iteration algorithm based on Equation (20) is shown based on the fixed point theorem.

First, let us observe that the solution to the recurrence relation (20) (two first lines of the equation) is invariant up to a translation of the origin of the free energy. Indeed, it can easily be shown that if  $\phi^*$  is a solution of (20), a shift of the free energy by a quantity  $\alpha$ , that is  $\phi_i^{*'} = \phi_i^* + \alpha$  for each  $i$ , is also a solution to (20). To overcome this underdetermination, the free energy is set to zero on the absorbing, goal, node  $n$ ,  $\phi_n^* = 0$ .

We will now study the following fixed point iteration after permuting the index of the nodes so that the unconstrained nodes appear before the constrained nodes,

$$\phi_i^* \leftarrow \begin{cases} -\frac{1}{\theta} \log \left[ \sum_{j=1}^n p_{ij}^{\text{ref}} \exp \left[ -\theta (c_{ij} + \phi_j^*) \right] \right] & \text{if } 1 \leq i \leq |\mathcal{U}| \\ \sum_{j=1}^n p_{ij}^{\text{ref}} (c_{ij} + \phi_j^*) & \text{if } |\mathcal{U}| + 1 \leq i \leq |\mathcal{U}| + |\mathcal{C}| \\ 0 & \text{if } i = n \end{cases} \quad (\text{D.1})$$

Then, it is well-known that this kind of fixed-point iteration converges to a unique solution in a convex domain (here, the positive quadrant) if the Jacobian matrix,  $\mathbf{J}$ , of the transformation has a matrix norm (for instance its spectral radius) strictly smaller than 1 everywhere in this domain [24, 60]. In that case, the fixed-point transformation is what is called a contraction mapping. We will thus compute the spectral radius of the Jacobian matrix and verify that it is smaller than one for all non-negative values of  $\phi^*$ .

The element  $i, j$  of this Jacobian matrix can easily be computed from Equation (D.1). For unconstrained nodes,

$$[\mathbf{J}]_{ij} = \frac{\partial \phi_i^*}{\partial \phi_j^*} = \frac{p_{ij}^{\text{ref}} \exp \left[ -\theta (c_{ij} + \phi_j^*) \right]}{\sum_{k=1}^n p_{ik}^{\text{ref}} \exp \left[ -\theta (c_{ik} + \phi_k^*) \right]} \quad \text{for } 1 \leq i \leq |\mathcal{U}| \quad (\text{D.2})$$

For constrained nodes,

$$\frac{\partial \phi_i^*}{\partial \phi_j^*} = p_{ij}^{\text{ref}} \quad \text{for } |\mathcal{U}| + 1 \leq i \leq |\mathcal{U}| + |\mathcal{C}| \quad (\text{D.3})$$

and of course  $\partial \phi_n^* / \partial \phi_j^* = 0$  for all  $j$ .

Then, we can verify that this Jacobian matrix  $\mathbf{J}$  is sub-stochastic. Indeed, row sums are equal to 1 for rows 1 to  $(n-1)$ , and the last row sum (for node  $n$ ) is strictly less than 1 (it is equal to 0). Consequently, because, in addition, all the elements of the matrix are non-negative,  $\mathbf{J}$  is sub-stochastic [53]. Thus,  $\mathbf{J}$  defines a transition probability matrix of a killing, absorbing, Markov chain with a killing absorbing node  $n$  [26].

Now, from the definition of the Jacobian matrix (D.2)-(D.3), the graph induced by  $\mathbf{J}$  connects the  $n$  nodes in exactly the same way as the original graph

$G$ : node  $i$  and node  $j$  are connected if and only if they are connected in the original graph (the connectivity pattern is induced by  $p_{ij}^{\text{ref}}$ ).

Moreover, as it is assumed that the original graph  $G$  is strongly connected, the absorbing, killing, node  $n$  can be reached from any initial node of the graph and this property is kept for  $\mathbf{J}$ . This means that, exactly as in the case of a standard absorbing Markov chain, the total probability mass in the transient states of the network (nodes 1 to  $n - 1$ ) will gradually decrease and  $\lim_{t \rightarrow \infty} \mathbf{J}^t \rightarrow 0$  [34]. This implies that the spectral radius of the Jacobian matrix  $\mathbf{J}$  is strictly less than 1 [53]. Therefore, as the spectral radius is a matrix norm, the iteration (D.1) converges to a unique solution independently of the (positive) initial conditions [24, 60].

---

## References

- [1] Y. Achbany, F. Fouss, L. Yen, A. Pirotte, and M. Saerens. Optimal tuning of continual exploration in reinforcement learning. *Proceedings of the 16th International Conference on Artificial Neural Networks (ICANN 06). Lecture notes in Computer Science*, LNCS 4131:734–749, 2006.
- [2] Y. Achbany, F. Fouss, L. Yen, A. Pirotte, and M. Saerens. Tuning continual exploration in reinforcement learning: an optimality property of the Boltzmann strategy. *Neurocomputing*, 71:2507–2520, 2008.
- [3] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin. *Network flows: Theory, algorithms, and applications*. Prentice Hall, 1993.
- [4] T. Akamatsu. Cyclic flows, markov process and stochastic traffic assignment. *Transportation Research B*, 30(5):369–386, 1996.
- [5] T. Akamatsu. Decomposition of path choice entropy in general transport networks. *Transportation Science*, 31(4):349–362, 1997.
- [6] K. Arrow, L. Hurwicz, and H. Uzawa. *Studies in linear and non-linear programming*. Stanford University Press, 1958.
- [7] K. Asadi and M. L. Littman. An alternative softmax operator for reinforcement learning. *ArXiv preprint arXiv:1612.05628*, 2016.
- [8] K. Asadi and M. L. Littman. An alternative softmax operator for reinforcement learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 243–252, 2017.
- [9] M. G. Azar, V. Gómez, and B. Kappen. Dynamic policy programming with function approximation. In *International Conference on Artificial Intelligence and Statistics (AISTAT)*, pages 119–127, 2011.
- [10] M. G. Azar, V. Gómez, and H. J. Kappen. Dynamic policy programming. *Journal of Machine Learning Research*, 13(Nov):3207–3245, 2012.
- [11] F. Bavaud and G. Guex. Interpolating between random walks and shortest paths: a path functional approach. In *Proceedings of the International Conference on Social Informatics (SocInfo 2012)*, pages 68–81. Springer, 2012.
- [12] D. P. Bertsekas. *Network optimization: continuous and discrete models*. Athena Scientific, 1998.
- [13] D. P. Bertsekas. *Nonlinear Programming, 2nd ed.* Athena Scientific, 1999.



- [14] D. P. Bertsekas. *Dynamic programming and optimal control*. Athena Scientific, 2000.
- [15] D. P. Bertsekas and J. Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.
- [16] B. Blaise. Randomized markov decision processes: a study of two new algorithms. Master’s thesis, Universite de Louvain, 2013. Supervisor: Prof. Marco Saerens.
- [17] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.
- [18] A. K. Chandra, P. Raghavan, W. L. Ruzzo, R. Smolensky, and P. Tiwari. The electrical resistance of a graph captures its commute and cover times. *Annual ACM Symposium on Theory of Computing*, pages 574–586, 1989.
- [19] N. Christofides. *Graph theory: An algorithmic approach*. Academic Press, 1975.
- [20] J. Cook. Basic properties of the soft maximum. Unpublished manuscript available from [www.johndcook.com/blog/2010/01/13/soft-maximum](http://www.johndcook.com/blog/2010/01/13/soft-maximum), 2011.
- [21] T. Cormen, C. Leiserson, R. Rivest, and C. Stein. *Introduction to algorithms, 3th ed.* The MIT Press, 2009.
- [22] T. M. Cover and J. A. Thomas. *Elements of information theory, 2nd ed.* John Wiley and Sons, 2006.
- [23] J. Culioli. *Introduction a l’optimisation*. Ellipses, 2012.
- [24] G. Dahlquist and A. Bjorck. *Numerical methods*. Prentice-Hall, 1974.
- [25] F. Fouss, A. Pirotte, J.-M. Renders, and M. Saerens. Random-walk computation of similarities between nodes of a graph, with application to collaborative recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 19(3):355–369, 2007.
- [26] F. Fouss, M. Saerens, and M. Shimbo. *Algorithms and models for network data and link analysis*. Cambridge University Press, 2016.
- [27] R. Fox, A. Pakman, and N. Tishby. G-learning: taming the noise in reinforcement learning via soft updates. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence (UAI 2016)*, pages 202–211, 2001.
- [28] K. Françoisse, I. Kivimäki, A. Mantrach, F. Rossi, and M. Saerens. A bag-of-paths framework for network data analysis. *ArXiv preprint arXiv:1302.6766*, 2013.
- [29] K. Françoisse, I. Kivimäki, A. Mantrach, F. Rossi, and M. Saerens. A bag-of-paths framework for network data analysis. *Neural Networks*, 90:90–111, 2017.
- [30] S. García-Díez, F. Fouss, M. Shimbo, and M. Saerens. A sum-over-paths extension of edit distances accounting for all sequence alignments. *Pattern Recognition*, 44(6):1172–1182, 2011.
- [31] S. García-Díez, J. Laforge, and M. Saerens. Rminimax: an optimally randomized minimax algorithm. *IEEE Transactions on Systems, Man and Cybernetics, part B: Cybernetics*, 43(1):385–393, 2013.
- [32] S. García-Díez, E. Vandebussche, and M. Saerens. A continuous-state version of discrete randomized shortest-paths. *Proceedings of the 50th IEEE International Conference on Decision and Control (IEEE CDC 2011)*, pages 6570–6577, 2011.
- [33] M. Gondran and M. Minoux. *Graphs and algorithms*. John Wiley & Sons, 1984.
- [34] C. Grinstead and J. L. Snell. *Introduction to probability*. The Mathematical Association of America, 2nd edition, 1997.
- [35] I. Griva, S. Nash, and A. Sofer. *Linear and nonlinear optimization*. SIAM, 2nd edition, 2008.

- [36] G. Guex and F. Bavaud. Flow-based dissimilarities: shortest path, commute time, max-flow and free energy. In B. Lausen, S. Krolak-Schwerdt, and M. Bohmer, editors, *Data science, learning by latent structures, and knowledge discovery*, volume 1564 of *Studies in Classification, Data Analysis, and Knowledge Organization*, pages 101–111. Springer, 2015.
- [37] D. A. Harville. *Matrix algebra from a statistician’s perspective*. Springer-Verlag, 1997.
- [38] D. Hosmer and S. Lemeshow. *Applied logistic regression, 2nd ed.* Wiley, 2000.
- [39] T. Jaakkola, M. Meila, and T. Jebara. Maximum entropy discrimination. In *Advances in Neural Information Processing Systems 16 (NIPS 2000)*, pages 470–476. MIT Press, 2000.
- [40] E. T. Jaynes. Information theory and statistical mechanics. *Physical review*, 106:620–630, 1957.
- [41] T. Jebara. *Machine learning, discriminative and generative*. Kluwer Academic Publishers, 2004.
- [42] D. Jungnickel. *Graphs, networks, and algorithms, 3th ed.* Springer, 2008.
- [43] H. J. Kappen, V. Gómez, and M. Opper. Optimal control as a graphical model inference problem. *Machine learning*, 87(2):159–182, 2012.
- [44] J. N. Kapur. *Maximum-entropy models in science and engineering*. Wiley, 1989.
- [45] J. N. Kapur and H. K. Kesavan. *Entropy optimization principles with applications*. Academic Press, 1992.
- [46] J. G. Kemeny and J. L. Snell. *Finite Markov chains*. Springer-Verlag, 1976.
- [47] I. Kivimäki, B. Lebichot, J. Saramäki, and M. Saerens. Two betweenness centrality measures based on randomized shortest paths. *Scientific Reports*, 6:srep19668, 2016.
- [48] I. Kivimäki, M. Shimbo, and M. Saerens. Developments in the theory of randomized shortest paths with a comparison of graph node distances. *Physica A: Statistical Mechanics and its Applications*, 393:600–616, 2014.
- [49] D. J. Klein and M. Randic. Resistance distance. *Journal of Mathematical Chemistry*, 12:81–95, 1993.
- [50] B. Lebichot. *Network analysis based on bag-of-paths: classification, node criticality and randomized policies*. PhD thesis, Ecole Polytechnique, Université catholique de Louvain, Belgium, 2018. Supervisor: Prof. Marco Saerens.
- [51] M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the 11th International Conference on Machine Learning (ICML-94)*, pages 157–163, 1994.
- [52] A. Mantrach, N. V. Zeebroeck, P. Francq, M. Shimbo, H. Bersini, and M. Saerens. Semi-supervised classification and betweenness computation on large, sparse, directed graphs. *Pattern recognition*, 44(6):1212–1224, 2011.
- [53] C. D. Meyer. *Matrix analysis and applied linear algebra*. SIAM, 2000.
- [54] R. Motwani and P. Raghavan. *Randomized algorithms*. Cambridge University Press, 1995.
- [55] K. Murphy. *Machine learning: a probabilistic perspective*. MIT Press, 2012.
- [56] G. Neu, A. Jonsson, and V. Gómez. A unified view of entropy-regularized markov decision processes. *ArXiv preprint arXiv:1705.07798*, 2017.
- [57] J. Norris. *Markov chains*. Cambridge University Press, 1997.
- [58] M. J. Osborne. *An introduction to game theory*. Oxford University Press, 2004.

- [59] L. Peliti. *Statistical mechanics in a nutshell*. Princeton University Press, 2011.
- [60] G. Phillips and P. Taylor. *Theory and applications of numerical analysis, 2nd ed.* Academic Press, 1996.
- [61] W. Powell. *Approximate dynamic programming, 2nd ed.* John Wiley and Sons, 2011.
- [62] M. Puterman. *Markov decision processes: discrete stochastic programming*. John Wiley and Sons, 1994.
- [63] H. Raiffa. *Decision analysis*. Addison-Wesley, 1970.
- [64] R. Rardin. *Optimization in operations research*. Prentice Hall, 1998.
- [65] L. Reichl. *A modern course in statistical physics, 2nd ed.* Wiley, 1998.
- [66] J. Rubin, O. Shamir, and N. Tishby. *Trading value and information in MDPs*, pages 57–74. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
- [67] S. Russell and P. Norvig. *Artificial intelligence: A modern approach, 3d ed.* Prentice-Hall, 2010.
- [68] M. Saerens, Y. Achbany, F. Fouss, and L. Yen. Randomized shortest-path problems: Two related models. *Neural Computation*, 21(8):2363–2404, 2009.
- [69] R. Sedgewick. *Algorithms, 4th ed.* Addison-Wesley, 2011.
- [70] R. S. Sutton and A. G. Barto. *Reinforcement learning: an introduction, 2nd ed. Draft manuscript in progress*. The MIT Press, 2017.
- [71] A. Tahbaz and A. Jadbabaie. A one-parameter family of distributed consensus algorithms with boundary: from shortest paths to mean hitting times. In *Proceedings of IEEE Conference on Decision and Control*, pages 4664–4669, 2006.
- [72] H. M. Taylor and S. Karlin. *An introduction to stochastic modeling, 3th Ed.* Academic Press, 1998.
- [73] E. Theodorou, D. Krishnamurthy, and E. Todorov. From information theoretic dualities to path integral and kullback-leibler control: Continuous and discrete time formulations. In *The Sixteenth Yale Workshop on Adaptive and Learning Systems*, 2013.
- [74] E. A. Theodorou and E. Todorov. Relative entropy and free energy dualities: Connections to path integral and kl control. In *Proceedings of the 51st IEEE Conference on Decision and Control (CDC 2012)*, pages 1466–1473. IEEE, 2012.
- [75] H. C. Tijms. *A first course in stochastic models*. John Wiley and Sons, 2003.
- [76] E. Todorov. Linearly-solvable markov decision problems. In *Advances in Neural Information Processing Systems 19 (NIPS 2006)*, pages 1369–1375. MIT Press, 2006.
- [77] D. White. Real applications of Markov decision processes. *Interfaces*, 15(6):73–83, 1985.
- [78] D. White. Further real applications of Markov decision processes. *Interfaces*, 18(5):55–61, 1988.
- [79] D. J. White. A survey of applications of Markov decision processes. *Journal of the Operational Research Society*, 44(11):1073–1096, 1993.
- [80] L. Yen, A. Mantrach, M. Shimbo, and M. Saerens. A family of dissimilarity measures between nodes generalizing both the shortest-path and the commute-time distances. In *Proceedings of the 14th SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2008)*, pages 785–793, 2008.