On the Metric-based Approximate Minimization of Markov Chains[*]

Giovanni Bacci, Giorgio Bacci, Kim G. Larsen, and Radu Mardare

Dept. of Computer Science, Aalborg University, Denmark.
{giovbacci,grbacci,klg,mardare}@cs.aau.dk

Abstract

We address the behavioral metric-based approximate minimization problem of Markov Chains (MCs), i.e., given a finite MC and a positive integer k, we are interested in finding a k-state MC of *minimal* distance to the original. By considering as metric the bisimilarity distance of Desharnais at al., we show that optimal approximations always exist; show that the problem can be solved as a bilinear program; and prove that its threshold problem is in PSPACE and NP-hard. Finally, we present an approach inspired by expectation maximization techniques that provides suboptimal solutions. Experiments suggest that our method gives a practical approach that outperforms the bilinear program implementation run on state-of-the-art bilinear solvers.

1998 ACM Subject Classification F.1.1 Models of Computation.

Keywords and phrases Behavioral distances, Probabilistic Models, Automata Minimization.

Digital Object Identifier 10.4230/LIPIcs.ICALP.2017.??

1 Introduction

Minimization of finite automata, i.e., the process of transforming a given finite automaton into an equivalent one with minimum number of states, has been a major subject since the 1950s due to its fundamental importance for any implementation of finite automata tools.

The first algorithm for the minimization of deterministic finite automata (DFAs) is due to Moore [27], with time complexity $O(n^2 s)$, later improved by the now classical Hopcroft's algorithm [17] to $O(ns \log n)$, where n is the number of states and s the size of the alphabet. Their algorithms are based on a partition refinement of the states into equivalence classes of the *Myhill-Nerode equivalence relation.* Partition refinement has been employed in the definition of efficient minimization procedures for a wide variety of automata: by Kanellakis and Smolka [19, 20] for the minimization of labelled transition systems (LTSs) w.r.t. Milner's strong bisimulation [26]; by Baier [4] for the reduction of Markov Chains (MCs) w.r.t. Larsen and Skou's probabilistic bisimulation [23]; by Alur et al. [2] and by Yannakakis and Lee [30], respectively, for the minimization of timed transition systems and timed-automata. This technique was used also in parallel and distributed implementations of the above algorithms [31, 8], and in the online reachability analysis of transition systems [24].

In [18], Jou and Smolka observed that for reasoning about the behavior of probabilistic systems (and more in general, all type of quantitative systems), rather than equivalences, a notion of distance is more reasonable in practice, since it permits "*a shift in attention from*

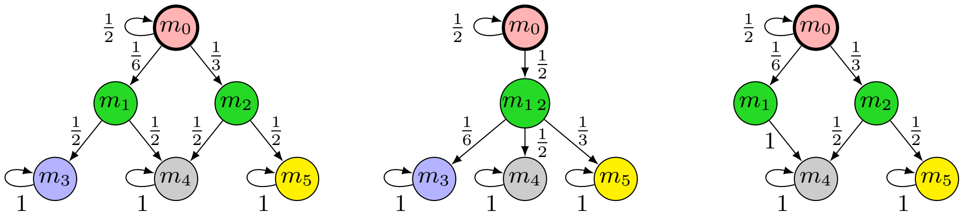
[*] Work supported by the EU 7th Framework Programme (FP7/2007-13) under Grants Agreement nr.318490 (SENSATION), nr.601148 (CASSTING), the Sino-Danish Basic Research Center IDEA4CPS funded by Danish National Research Foundation and National Science Foundation China, the ASAP Project (4181-00360) funded by the Danish Council for Independent Research, the ERC Advanced Grant LASSO, and the Innovation Fund Denmark center DiCyPS.

44th International Colloquium on Automata, Languages, and Programming (ICALP 2017).
Editors: Ioannis Chatzigiannakis, Piotr Indyk, Fabian Kuhn, and Anca Muscholl;
Article No. ??; pp. ??:1–??:19
Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

equivalent processes to probabilistically similar processes". This observation motivated the development of metric-based semantics for quantitative systems, that consists in proposing 1-bounded pseudometrics capturing the similarities of the behaviors in the presence of small variations of the quantitative data. These pseudometrics generalize behavioral equivalences in the sense that, two processes are at distance 0 iff they are equivalent, and at distance 1 if no significant similarities can be observed between them.

The first proposal of a behavioral pseudometric is due to Desharnais et al. [12] on labelled MCs, a.k.a. *probabilistic bisimilarity distance*, with the property that two MCs are at distance 0 iff they are probabilistic bisimilar. Its definition is parametric on a discount factor $\lambda \in (0, 1]$ that controls the significance of the future steps in the measurement. This pseudometric has been greatly studied by van Breugel and Worrell [28, 29, 10] who noticed, among other notable results, its relation with the Kantorovich distance on probability distributions and provided a polynomial-time algorithm for its computation.

The introduction of metric-based semantics motivated the interest in the approximate minimization of quantitative systems. The goal of approximate minimization is to start from a minimal automaton and produce a smaller automaton that is close to the given one in a certain sense. The desired size of the approximating automaton is given as input. Inspired by the aggregation of equivalent states typical of partition refinement techniques, in [15], the approximate minimization problem has been approached by aggregating states having relative smaller distance. An example of this approach on MCs using the λ-bisimilarity distance of Desharnais et al. is shown below.



Let $\mathcal{M}$ be the MC on the left and assume we want to approximate it by an MC with at most 5 states. Since m_1, m_2 are the only two states at distance less than 1, the most natural choice for an aggregation shall collapse (via convex combination) m_1 and m_2, obtaining the MC in the middle, which has distance $\frac{4}{9}(\frac{\lambda^2}{2-\lambda})$ from $\mathcal{M}$. However, the approximate aggregation of states does not necessarily yield the closest optimal solution. Indeed, the MC on the right is a closer approximant of $\mathcal{M}$, at distance $\frac{1}{6}(\frac{\lambda^2}{2-\lambda})$ from it.

In this paper we address the issue of finding *optimal* solutions to the approximate minimization problem. Specifically we aim to answer to the following problem, left open in [15]: "*given a finite MC and a positive integer k, what is its 'best' k-state approximant? Here by 'best' we mean a k-state MC at minimal distance to the original*". We refer to this problem as *Closest Bounded Approximant* (CBA) and we present the following results related to it.

1. We characterize CBA as a bilinear optimization problem, proving the existence of *optimal* solutions. As a consequence of this result, approximations of optimal solutions can be obtained by checking the feasibility of bilinear matrix inequalities (BMIs) [22, 21].

2. We provide upper- and lower-bound complexity results for the threshold problem of CBA, called *Bounded Approximant* problem (BA), that asks whether there exists a k-state approximant with distance from the original MC bounded by a given rational threshold. We show that BA is in PSPACE and NP-hard. As a corollary we obtain NP-hardness for CBA.

3. We introduce the *Minimum Significant Approximant Bound* (MSAB) problem, that asks what is the minimum size k for an approximant to have some significant similarity to the original MC (i.e., at distance strictly less than 1). We show that this problem is NP-complete when one considers the undiscounted bisimilarity distance.

4. Finally, we present an algorithm for finding suboptimal solutions of CBA that is inspired by Expectation Maximization (EM) techniques [25, 7]. Experiments suggest that our method gives a practical approach that outperforms the bilinear program implementation —state-of-the-art bilinear solvers [21] fails to handle MCs with more than 5 states!

Related Work In [16], the approximate minimization of MCs is addressed via the notion of *quasi-lumpability*. An MC is quasi-lumpable if the given aggregations of the states can be turned into actual bisimulation-classes by a small perturbation of the transition probabilities. This approach differs from ours since there is no relation to a proper notion of behavioral distance (the approximation is w.r.t. the supremum norm of the difference of the stochastic matrices) and we do not consider any approximate aggregation of states. In [6], Balle et al. consider the approximate minimization of weighted finite automata (WFAs). Their method is via a truncation of a canonical normal form for WFAs that they introduced for the SVD decomposition of infinite Hankel matrices. Both [16] and [6] do not consider the issue of finding the *closest* approximant, which is the main focus of this paper, instead they give upper bounds on the distance from the given model.

2 Markov Chains and Bisimilarity Pseudometric

In this section we introduce the notation and recall the definitions of (discrete-time) *Markov chains* (MCs), *probabilistic bisimilarity* of Larsen and Skou [23], and the *probabilistic bisimilarity pseudometric* of Desharnais et al. [13].

For $R \subseteq X \times X$ an equivalence relation, $X/_R$ denotes its quotient set and $[x]_R$ denotes the R-equivalence class of $x \in X$. $\mathcal{D}(X)$ denotes the set of discrete probability distributions on X, i.e., functions $\mu \colon X \to [0,1]$, s.t. $\mu(X) = 1$, where $\mu(E) = \sum_{x \in E} \mu(x)$ for $E \subseteq X$.

In what follows we fix a countable set L of labels.

▶ **Definition 1** (Markov Chain)**.** A *Markov chain* is a tuple $\mathcal{M} = (M, \tau, \ell)$ consisting of a finite nonempty *set of states* M, a *transition distribution function* $\tau \colon M \to \mathcal{D}(M)$, and a *labelling function* $\ell \colon M \to L$.

Intuitively, if $\mathcal{M}$ is in state m it moves to state m' with probability $\tau(m)(m')$. Labels represent atomic properties that hold in certain states. The set of labels of $\mathcal{M}$ is denoted by $L(\mathcal{M}) = \{\ell(m) \mid m \in M\}$. Hereafter, we use $\mathcal{M} = (M, \tau, \ell)$ and $\mathcal{N} = (N, \theta, \alpha)$ to range over MCs and we refer to their constituents implicitly.

▶ **Definition 2** (Probabilistic Bisimulation [23])**.** An equivalence relation $R \subseteq M \times M$ is a *probabilistic bisimulation* on $\mathcal{M}$ if whenever $m \mathrel{R} n$, then

1. $\ell(m) = \ell(n)$, and
2. for all $C \in M/_R$, $\tau(m)(C) = \tau(n)(C)$.

Two states $m, n \in M$ are *probabilistic bisimilar w.r.t.* $\mathcal{M}$, written $m \sim_{\mathcal{M}} n$ if they are related by some probabilistic bisimulation on $\mathcal{M}$. In fact, probabilistic bisimilarity is the greatest probabilistic bisimulation.

Any bisimulation R on $\mathcal{M}$ induces a quotient construction, the *R-quotient of $\mathcal{M}$*, denoted $\mathcal{M}/_R = (M/_R, \tau/_R, \ell/_R)$, having R-equivalence classes as states, transition function $\tau/_R([m]_R)([n]_R) = \sum_{u \in [n]_R} \tau(m)(u)$, and labelling function $\ell/_R([m]_R) = \ell(m)$. An MC $\mathcal{M}$ is said *minimal* if it is isomorphic to its quotient w.r.t. probabilistic bisimilarity.

A 1-bounded *pseudometric* on X is a function $d \colon X \times X \to [0,1]$ such that, for any $x, y, z \in X$, $d(x,x) = 0$, $d(x,y) = d(y,x)$, and $d(x,y) + d(y,z) \geq d(x,z)$. 1-bounded pseudometrics on X forms a complete lattice under the point-wise partial order $d \sqsubseteq d'$ iff, for all $x, y \in X$, $d(x,y) \leq d'(x,y)$.

A pseudometric is said to lift an equivalence relation if it enjoys the property that two points are at distance zero iff they are related by the equivalence. A lifting for the probabilistic bisimilarity is provided by the *bisimilarity distance* of Desharnais et al. [13]. Its definition is based on the *Kantorovich (pseudo)metric* on probability distributions over a finite set X, defined as $\mathcal{K}(d)(\mu,\nu) = \min\left\{\int d \,\mathrm{d}\omega \mid \omega \in \Omega(\mu,\nu)\right\}$, where d is a (pseudo)metric on X and $\Omega(\mu,\nu)$ denotes the set of *couplings* for (μ,ν), i.e., distributions $\omega \in \mathcal{D}(X \times X)$ such that, for all $E \subseteq X$, $\omega(E \times X) = \mu(E)$ and $\omega(X \times E) = \nu(E)$.

▶ **Definition 3** (Bisimilarity Distance)**.** Let $\lambda \in (0,1]$. The λ-discounted *bisimilarity pseudometric* on $\mathcal{M}$, denoted by δ_λ, is the least fixed-point of the following functional operator on 1-bounded pseudometrics over M (ordered point-wise)

$$\Psi_\lambda(d)(m,n) = \begin{cases} 1 & \text{if } \ell(m) \neq \ell(n) \\ \lambda \cdot \mathcal{K}(d)(\tau(m),\tau(n)) & \text{otherwise}\,. \end{cases}$$

The operator Ψ_λ is monotonic, hence, by Tarski fixed-point theorem, δ_λ is well defined.

Intuitively, if two states have different labels δ_λ considers them as “incomparable” (i.e., at distance 1), otherwise their distance is given by the Kantorovich distance w.r.t. δ_λ between their transition distributions. The *discount factor* $\lambda \in (0,1]$ controls the significance of the future steps in the measurement of the distance; if $\lambda = 1$, the distance is said *undiscounted.*

The distance δ_λ has also a characterization based on the notion of coupling structure.

▶ **Definition 4** (Coupling Structure)**.** A function $\mathcal{C} \colon M \times M \to \mathcal{D}(M \times M)$ is a *coupling structure* for $\mathcal{M}$ if for all $m, n \in M$, $\mathcal{C}(m,n) \in \Omega(\tau(m),\tau(n))$.

Intuitively, a coupling structure can be thought of as an MC on the cartesian product $M \times M$, obtained as the probabilistic combination of two copies of $\mathcal{M}$.

Given a coupling structure $\mathcal{C}$ for $\mathcal{M}$ and $\lambda \in (0,1]$, let $\gamma_\lambda^{\mathcal{C}}$ be the least fixed-point of the following operator on $[0,1]$-valued functions $d \colon M \times M \to [0,1]$ (ordered point-wise)

$$\Gamma_\lambda^{\mathcal{C}}(d)(m,n) = \begin{cases} 1 & \text{if } \ell(m) \neq \ell(n) \\ \lambda \int d \,\mathrm{d}\mathcal{C}(m,n) & \text{otherwise}\,. \end{cases}$$

The function $\gamma_\lambda^{\mathcal{C}}$ is called λ*-discounted discrepancy* of $\mathcal{C}$, and the value $\gamma_\lambda^{\mathcal{C}}(m,n)$ is the λ-discounted probability of hitting from (m,n) a pair of states with different labels in $\mathcal{C}$.

▶ **Theorem 5** (Minimal coupling criterion [10])**.** *For arbitrary MCs $\mathcal{M}$ and discount factors $\lambda \in (0,1]$, $\delta_\lambda = \min\left\{\gamma_\lambda^{\mathcal{C}} \mid \mathcal{C} \text{ coupling structure for } \mathcal{M}\right\}$.*

Usually, MCs are associated with an initial state to be thought of as their initial configurations. In the rest of the paper when we talk about the distance between two MCs, written $\delta_\lambda(\mathcal{M},\mathcal{N})$, we implicitly refer to the distance between their initial states computed over the disjoint union of their MCs.

3 The Closest Bounded Approximant Problem

In this section we introduce the *Closest Bounded Approximant* problem w.r.t. δ_λ (CBA-λ), and give a characterization of it as a bilinear optimization problem.

$$
\begin{array}{lllr}
\text{mimimize} & d_{m_0,n_0} & & \\
\text{such that} & \lambda \sum_{(u,v) \in M \times N} c^{m,n}_{u,v} \cdot d_{u,v} \leq d_{m,n} & m \in M,\, n \in N & (5) \\
 & 1 - \alpha_{n,l} \leq d_{m,n} \leq 1 & n \in N,\, l \in L(\mathcal{M}),\, \ell(m) \neq l & (6) \\
 & \alpha_{n,l} \cdot \alpha_{n,l'} = 0 & n \in N,\, l, l' \in L(\mathcal{M}),\, l \neq l' & (7) \\
 & \sum_{l \in L(\mathcal{M})} \alpha_{n,l} = 1 & n \in N & (8) \\
 & \sum_{v \in N} c^{m,n}_{u,v} = \tau(m)(u) & m, u \in M,\, n \in N & (9) \\
 & \sum_{u \in M} c^{m,n}_{u,v} = \theta_{n,v} & m \in M,\, n, v \in N & (10) \\
 & c^{m,n}_{u,v} \geq 0 & m, u \in M,\, n, v \in N & (11)
\end{array}
$$

Figure 1 Characterization of CBA-λ as a bilinear optimization problem.

▶ **Definition 6** (Closest Bounded Approximant). Let $k \in \mathbb{N}$ and $\lambda \in (0, 1]$. The *closest bounded approximant problem* w.r.t. δ_λ for an MC $\mathcal{M}$ is the problem of finding an MC $\mathcal{N}$ with at most k states minimizing $\delta_\lambda(\mathcal{M}, \mathcal{N})$.

Clearly, when k is greater than or equal to the number of bisimilarity classes of $\mathcal{M}$, an optimal solution of CBA-λ is the bisimilarity quotient. Therefore, without loss of generality, we will assume $1 \leq k < |M|$ and $\mathcal{M}$ to be minimal. Note that, under these assumptions $\mathcal{M}$ must have at least two nodes with different labels.

Let $\mathrm{MC}(k)$ denote the set of MCs with at most k states and $\mathrm{MC}_A(k)$ its restriction to those using only labels in $A \subseteq L$. Using this notation, the optimization problem CBA-λ on the instance $\langle \mathcal{M}, k \rangle$ can be reformulated as finding an MC $\mathcal{N}^*$ such that

$$\delta_\lambda(\mathcal{M}, \mathcal{N}^*) = \min \{\delta_\lambda(\mathcal{M}, \mathcal{N}) \mid \mathcal{N} \in \mathrm{MC}(k)\} \ , \tag{1}$$

In general, it is not obvious that for arbitrary instances $\langle \mathcal{M}, k \rangle$ a minimum in (1) exists. At the end of the section, we will show that such a minimum always exists (Corollary 9).

A useful property of CBA-λ is that an optimal solution can be found among the MCs using labels from the given MC.

▶ **Lemma 7** (Meaningful labels). *Let $\mathcal{M}$ be an MC. Then, for any $\mathcal{N}' \in \mathrm{MC}(k)$ there exists $\mathcal{N} \in \mathrm{MC}_{L(\mathcal{M})}(k)$ such that $\delta_\lambda(\mathcal{M}, \mathcal{N}) \leq \delta_\lambda(\mathcal{M}, \mathcal{N}')$.*

In the following, fix $\langle \mathcal{M}, k \rangle$ as instance of CBA-λ, let $m_0 \in M$ be the initial state of $\mathcal{M}$. By Lemma 7, Theorem 5 and Tarski fixed-point theorem

$$
\begin{array}{lr}
\inf \{\delta_\lambda(\mathcal{M}, \mathcal{N}) \mid \mathcal{N} \in \mathrm{MC}(k)\} = & (2) \\
\quad = \inf \left\{\gamma^{\mathcal{C}}_\lambda(\mathcal{M}, \mathcal{N}) \mid \mathcal{N} \in \mathrm{MC}_{L(\mathcal{M})}(k) \text{ and } \mathcal{C} \in \Omega(\mathcal{M}, \mathcal{N})\right\} & (3) \\
\quad = \inf \left\{d(\mathcal{M}, \mathcal{N}) \mid \mathcal{N} \in \mathrm{MC}_{L(\mathcal{M})}(k),\, \mathcal{C} \in \Omega(\mathcal{M}, \mathcal{N}), \text{ and } \Gamma^{\mathcal{C}}_\lambda(d) \sqsubseteq d\right\} \ , & (4)
\end{array}
$$

where $\Omega(\mathcal{M}, \mathcal{N})$ denotes the set of all coupling structures for the disjoint union of $\mathcal{M}$ and $\mathcal{N}$. This simple change in perspective yields a translation of the problem of computing the optimal value of CBA-λ to the bilinear program in Figure 1.

In our encoding, $N = \{n_0, \dots, n_{k-1}\}$ are the states of an arbitrary $\mathcal{N} = (N, \theta, \alpha) \in \mathrm{MC}(k)$ and n_0 is the initial one. The variable $\theta_{n,v}$ is used to encode the transition probability $\theta(n)(v)$. Hence, a feasible solution satisfying (11–13) will have the variable $c^{m,n}_{u,v}$ representing the value $\mathcal{C}(m, n)(u, v)$ for a coupling structure $\mathcal{C} \in \Omega(\mathcal{M}, \mathcal{N})$. An assignment for the variables $\alpha_{n,l}$ satisfying (7–10) encodes (uniquely) a labeling function $\alpha\colon N \to L(\mathcal{M})$ satisfying the following property:

$$\text{for all } n \in N, l \in L(\mathcal{M}) \qquad \alpha_{n,l} = 1 \quad \text{iff} \quad \alpha(n) = l\,. \tag{12}$$

The constraint (7) models the fact that each node $n \in N$ is assigned to at most one label $l \in L(\mathcal{M})$, and the constraint (10) ensures that each node is assigned to at least one label. Conversely, any labeling $\alpha\colon N \to L(\mathcal{M})$ admits an assignment of the variables $\alpha_{n,l}$ that satisfy (7–10) and (14). Finally, an assignment for the variables $d_{m,n}$ satisfying the constraints (6–5) represents a prefix point of $\Gamma^{\mathcal{C}}_{\lambda}$. Note that (5) guarantees that $d_{m,n} = 1$ whenever $\alpha(n) \neq \ell(m)$ —indeed, by (7), $\alpha_{n,l} = 0$ iff $\alpha(n) \neq \ell(m)$.

Let $F_\lambda\langle \mathcal{M}, k\rangle$ denote the bilinear optimization problem in Figure 1. Directly from the arguments stated above we obtain the following result.

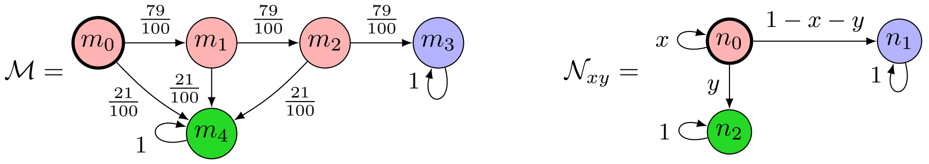
▶ **Theorem 8.** $\inf\{\delta_\lambda(\mathcal{M},\mathcal{N}) \mid \mathcal{N} \in \mathrm{MC}(k)\}$ *is the optimal value of* $F_\lambda\langle \mathcal{M}, k\rangle$.

▶ **Corollary 9.** *Any instance of CBA-λ admits an optimal solution.*

Proof. Let h be the number of variables in $F_\lambda\langle \mathcal{M}, k\rangle$. The constraints (5–13) describe a compact subset of $\mathbb{R}^h$ —it is an intersection of closed sets bounded by $[0,1]^h$. The objective function of $F_\lambda\langle \mathcal{M}, k\rangle$ is linear, hence the infimum is attained by a feasible solution. The thesis follows by Theorem 8. ◀

The following example shows that even by starting with a MC with rational transition probabilities, optimal solutions for CBA-λ may have irrational transition probabilities.

▶ **Example 10.** Consider the MC $\mathcal{M}$ depicted below, with initial state m_0 and labeling represented by colors. An optimal solution of CBA-1 on $\langle \mathcal{M}, 3\rangle$ is the MC $\mathcal{N}_{xy}$ depicted below, with initial state n_0 and parameters $x = \frac{1}{30}\left(10 + \sqrt{163}\right)$, $y = \frac{21}{100}$.



Since the distance $\delta_1(\mathcal{M},\mathcal{N}_{xy}) = \frac{436}{675} - \frac{163\sqrt{163}}{13500} \approx 0.49$ is irrational, by [10, Proposition 13], any optimal solution must have some irrational transition probability.

Next we show that the above is indeed an optimal solution. Assume by contradiction that $\mathcal{N}^* \not\sim \mathcal{N}_{xy}$ is an optimal solution. By Lemma 7, we can assume $L(\mathcal{N}^*) \subseteq L(\mathcal{M})$. If $L(\mathcal{N}^*) = L(\mathcal{M})$, then $\delta_1(\mathcal{M},\mathcal{N}^*) = \min\{\delta_1(\mathcal{M},\mathcal{N}_{zy}) \mid z \in [0, 1-y]\}$ since one can show that for any $y' \neq y$ and z', there exists $z \in [0, 1-y]$, such that $\delta_1(\mathcal{M},\mathcal{N}_{zy}) \leq \delta_1(\mathcal{M},\mathcal{N}_{z'y'})$. $\delta_1(\mathcal{M},\mathcal{N}_{zy})$ is analytically solved by $z^3 - z^2 - \frac{21}{100}z - \frac{79}{100}$ and its minimum value is achieved at $z = \frac{1}{30}\left(10 + \sqrt{163}\right)$. This contradicts $\mathcal{N}^* \not\sim \mathcal{N}_{xy}$. Assume $L(\mathcal{N}^*) \subsetneq L(\mathcal{M})$. By [10, Corollary 11], for any measurable set $A \subseteq L^\omega$, $\delta_1(\mathcal{M},\mathcal{N}^*) \geq |\mathbb{P}_{\mathcal{M}}(A) - \mathbb{P}_{\mathcal{N}^*}(A)|$, where $\mathbb{P}_{\mathcal{N}}(A)$ denotes the probability that a run of $\mathcal{N}$ is in A. If $\ell(m_0) \notin L(\mathcal{N}^*)$, we have that $\delta_1(\mathcal{M},\mathcal{N}^*) \geq |\mathbb{P}_{\mathcal{M}}(\ell(m_0)L^\omega) - \mathbb{P}_{\mathcal{N}^*}(\ell(m_0)L^\omega)| = \mathbb{P}_{\mathcal{M}}(\ell(m_0)L^\omega) = 1 > \delta_1(\mathcal{M},\mathcal{N}_{xy})$. Analogously, if $\ell(m_3) \notin L(\mathcal{N}^*)$ we have $\delta_1(\mathcal{M},\mathcal{N}^*) \geq \mathbb{P}_{\mathcal{M}}(L^*\ell(m_3)L^\omega) = \left(\frac{79}{100}\right)^3 > \delta_1(\mathcal{M},\mathcal{N}_{xy})$. Finally, if $\ell(m_4) \notin L(\mathcal{N}^*)$, $\delta_1(\mathcal{M},\mathcal{N}^*) \geq \mathbb{P}_{\mathcal{M}}(L^*\ell(m_4)L^\omega) = \frac{21}{100}\sum_{i=0}^{2}\left(\frac{79}{100}\right)^i > \delta_1(\mathcal{M},\mathcal{N}_{xy})$. ◀

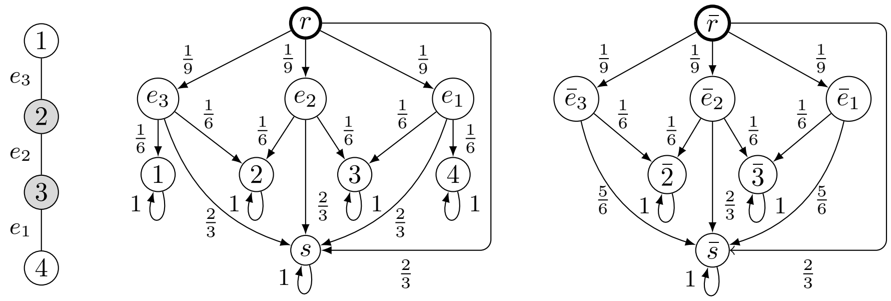


Figure 2 (Left) An undirected graph G; (Center) The MC $\mathcal{M}_G$ associated to the graph G; (Right) The MC $\mathcal{M}_C$ associated to the vertex cover $C = \{2, 3\}$ of G. (see Thm. 14).

4 The Bounded Approximant Threshold Problem

The *Bounded Approximant problem* w.r.t. δ_λ (BA-λ) is the threshold decision problem of CBA-λ, that, given MC $\mathcal{M}$, integer $k \geq 1$, and rational $\epsilon \geq 0$, asks whether there exists $\mathcal{N} \in \mathrm{MC}(k)$ such that $\delta_\lambda(\mathcal{M}, \mathcal{N}) \leq \epsilon$.

From the characterization of CBA-λ as a bilinear optimization problem (Theorem 8) we immediately get the following complexity upper-bound for BA-λ.

▶ **Theorem 11.** *For any $\lambda \in (0, 1]$, BA-λ is in PSPACE.*

Proof. By Theorem 8, deciding an instance $\langle \mathcal{M}, k, \epsilon \rangle$ of BA-λ can be encoded as a decision problem for the existential theory of the reals, namely, checking the feasibility of the constraints (5–13) in conjunction with $d_{m_0,n_0} \leq \epsilon$. The encoding is polynomial in the size of $\langle \mathcal{M}, k, \epsilon \rangle$, thus it can be solved in PSPACE (*cf.* Canny [9]). ◀

In the rest of the section we provide a complexity lower-bound for BA-λ, by showing that BA-λ is NP-hard via a reduction from Vertex Cover. Recall that, a vertex cover of an undirected graph G is a subset C of vertices such that every edge in G has at least one endpoint in C. Given a graph G and a positive integer h, the Vertex Cover problem asks if G has a cover of size at most h.

Before presenting the reduction we establish structural properties for an optimal solution of CBA-λ in the case the given MC has injective labeling (i.e., no two distinct states with the same label). Specifically, we show that an optimal solution for an instance $\langle \mathcal{M}, k \rangle$ of CBA-λ can be found among MCs with injective labeling into $L(\mathcal{M})$.

▶ **Lemma 12.** *If $\mathcal{M}$ has injective labeling, there exists $\mathcal{N} \in \mathrm{MC}_{L(\mathcal{M})}(k)$ with injective labeling that minimizes the distance $\delta_\lambda(\mathcal{M}, \mathcal{N})$.*

▶ **Lemma 13.** *For all $m \in M$ and $n \in N$, $\delta_\lambda(m, n) \geq \lambda \cdot \tau(m)(\{u \in M \mid \ell(u) \notin L(\mathcal{N})\})$.*

Note that Lemma 13 provides a lower-bound on the optimal distance between $\mathcal{M}$ and any $\mathcal{N} \in \mathrm{MC}(k)$. This lower-bound will be useful in the proof of the following result.

▶ **Theorem 14.** *For any $\lambda \in (0, 1]$, BA-λ is NP-hard.*

Proof. We provide a polynomial-time many-one reduction from Vertex Cover.

Let $\langle G = (V, E), h \rangle$ be an instance of Vertex Cover and let $e = |E|$. Without loss of generality we assume $e \geq 2$ and $k < n$. From G we construct the MC $\mathcal{M}_G = (M, \tau, \ell)$ as follows. The set of states M is given as the union of V and E to which we add two extra states: a root r (thought of as the initial state) and a sink s. Each node of $\mathcal{M}_G$ is associated with a unique label (i.e., ℓ is injective). The sink state s and all $v \in V$ loop to themselves with probability 1. All the other states go with probability $1-\frac{1}{e}$ to the sink state s. The rest of their transition probability mass is assigned as follows. The root r goes with probability $\frac{1}{e^2}$ to each $a \in E$, and all $(u,v) \in E$ go with probability $\frac{1}{2e}$ to their endpoints u, v. An example of construction for $\mathcal{M}_G$ is given in Figure 2. Next we show that

$$\langle G, h\rangle \in \textsc{Vertex Cover} \qquad \text{iff} \qquad \langle \mathcal{M}_G, e+h+2, \tfrac{\lambda^2}{2e^2}\rangle \in \text{BA-}\lambda\,.$$

($\Rightarrow$) Let C be a h-vertex cover of G. Construct $\mathcal{M}_C \in \text{MC}(e+h+2)$ by taking a copy of $\mathcal{M}_G$, removing all states in $V \setminus C$, and redirecting the exceeding transition probability to the sink state s (an example is shown in Figure 2). Next we show that $\delta_\lambda(\mathcal{M}_G, \mathcal{M}_C) \leq \frac{\lambda^2}{2e^2}$. For convenience, the states in $\mathcal{M}_C$ will be marked with a bar. By construction of $\mathcal{M}_G, \mathcal{M}_C$, for each $a \in E$, $\delta_\lambda(a, \overline{a}) \leq \frac{\lambda}{2e}$. Thus, $\delta_\lambda(\mathcal{M}_G, \mathcal{M}_C) = \delta_\lambda(r, \bar{r}) = \frac{\lambda}{e^2}\sum_{a\in E} \delta_\lambda(a, \overline{a}) \leq \frac{\lambda^2}{2e^2}$.

($\Leftarrow$) By contradiction, assume there exists $\mathcal{N} = (N, \theta, \alpha) \in \text{MC}(e+h+2)$ such that $\delta_\lambda(\mathcal{M}_G, \mathcal{N}) \leq \frac{\lambda^2}{2e^2}$ but no vertex cover of G of size h. Since ℓ is injective, by Lemma 12 we can assume α to be injective and $L(\mathcal{N}) \subseteq L(\mathcal{M}_G)$. We consider three cases separately:

Case: $\ell(s) \notin L(\mathcal{N})$. By Lemma 13 and the fact that $e > 1$ and $\lambda \in (0,1]$, we get the following contradiction: $\delta_\lambda(\mathcal{M}_G, \mathcal{N}) = \delta_\lambda(r, n_0) \geq \lambda \cdot \tau(r)(s) = \frac{\lambda(e-1)}{e} > \frac{\lambda^2}{2e^2}$.

Case: $\ell((u,v)) \notin L(\mathcal{N})$, for some $(u,v) \in E$. By Lemma 13 and the fact that $\lambda \in (0,1]$ and $e > 1$, leading to the contradiction $\delta_\lambda(\mathcal{M}_G, \mathcal{N}) = \delta_\lambda(r, n_0) \geq \lambda \cdot \tau(r)((u,v)) = \frac{\lambda}{e^2} > \frac{\lambda^2}{2e^2}$.

Case: $\ell(s) \in L(\mathcal{N})$ and $\{\ell((u,v)) \mid (u,v) \in E\} \subseteq L(\mathcal{N})$. Let $N' \subseteq N$ be the states with labels in $\{\ell(u) \mid u \in V\}$. By the structural hypothesis assumed on $\mathcal{N}$, we have $|N'| \leq h$. For each $(u,v) \in E$, two possible cases apply: if $\alpha(n) \in \{\ell(u), \ell(v)\}$, for some $n \in N'$, then $\delta_\lambda((u,v), \overline{(u,v)}) \geq \frac{\lambda}{2e}$; otherwise $\delta_\lambda((u,v), \overline{(u,v)}) \geq \frac{\lambda}{e} > \frac{\lambda}{2e}$. By hypothesis, there is no vertex cover of size h, hence there is at least one edge $(u,v) \in E$ for which the second case applies. Therefore, $\delta_\lambda(\mathcal{M}_G, \mathcal{N}) = \delta_\lambda(r, n_0) = \frac{\lambda}{e^2}\sum_{(u,v)\in E} \delta_\lambda((u,v), \overline{(u,v)}) > \frac{\lambda}{e^2} \cdot e \cdot \frac{\lambda}{2e} = \frac{\lambda^2}{2e^2}$.

The instance $\langle \mathcal{M}_G, e+h+2, \frac{\lambda^2}{2e^2}\rangle$ of BA-λ can be constructed in polynomial time in the size of $\langle G, h\rangle$. Thus, since Vertex Cover is NP-hard, so is BA-λ. ◀

5 Minimum Significant Approximant Bound

Recall that, two MCs are at distance 1 from each other when there is no significant similarity between their behaviors. Thus an MC $\mathcal{N}$ is said to be a *significant approximant* for the MC $\mathcal{M}$ w.r.t. δ_λ if $\delta_\lambda(\mathcal{M}, \mathcal{N}) < 1$.

Given an MC $\mathcal{M}$, the *Minimum Significant Approximant Bound* problem w.r.t. δ_λ (MSAB-λ) looks for the smallest k such that $\delta_\lambda(\mathcal{M}, \mathcal{N}) < 1$, for some $\mathcal{N} \in \text{MC}(k)$. The decision version of this problem is called *Significant Bounded Approximant problem* w.r.t. δ_λ (SBA-λ), and asks whether, for a given positive integer k, there exists $\mathcal{N} \in \text{MC}(k)$ such that $\delta_\lambda(\mathcal{M}, \mathcal{N}) < 1$.

When the discount factor $\lambda < 1$, the two problems above turn out to be trivial. Indeed, $\delta_\lambda(\mathcal{M}, \mathcal{N}) \leq \lambda$ when the initial states of $\mathcal{M}$ and $\mathcal{N}$ have the same label. On the contrary, in the case the distance is undiscounted ($\lambda = 1$), these problems are NP-complete. Before presenting the result, we provide the following technical lemma.

▶ **Lemma 15.** *Let $\mathcal{M}$ be a MC (assumed to be minimal) with initial state m_0 and $G(\mathcal{M})$ its underlying directed graph. Then, $\langle \mathcal{M}, k\rangle \in$ SBA-1 iff there exists a bottom strongly connected component (SCC) $G' = (V, E)$ in $G(\mathcal{M})$ and a path $m_0 \dots m_h$ in $G(\mathcal{M})$ such that $m_h \in V$ and $|\{\ell(m_i) \mid i < h, \nexists$ a path $v_i \dots v_{h-1} m_h$ in G' s.t. $\forall i \leq j < h.\, \ell(m_j) = \ell(v_j)\}| + |V| \leq k$.*

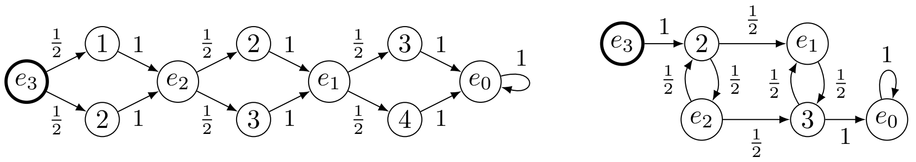


Figure 3 (Left) The MC $\mathcal{M}_G$ associated to the graph G in Figure 2 and (right) an MC $\mathcal{N}$ associated to the vertex cover $C = \{1, 2\}$ of G such that $\delta_1(\mathcal{M}_G, \mathcal{N}) < 1$ (cf. Theorem 16).

▶ **Theorem 16.** *SBA-1 is NP-complete.*

Proof. The membership in NP is easily proved by using the characterization in Lemma 15 and exploiting Tarjan's algorithm for generating bottom SCCs. As for the NP-hardness, we provide a polynomial-time many-one reduction from VERTEX COVER. Let $G = (V, E)$ be a graph with $E = \{e_1, \dots, e_n\}$. We construct the MC $\mathcal{M}_G$ as follows. The set of states is given by the set of edges E along with two states e_i^1 and e_i^2, for each edge $e_i \in E$, representing the two endpoints of e_i and an extra sink state e_0. The initial state is e_n. The transition probabilities are given as follows. The sink state e_0 loops with probability 1 to itself. Each edge $e_i \in E$ goes with probability $\frac{1}{2}$ to e_i^1 and e_i^2, respectively. For $1 \leq i \leq n$, the states e_i^1 and e_i^2 go with probability 1 to the state e_{i-1}. The edge states and the sink state are labelled by pairwise distinct labels, while the endpoints states e_i^1 and e_i^2 are labelled by the node in V they represent. An example of construction for $\mathcal{M}_G$ is shown in Figure 3.

Next we show the following equivalence:

$$\langle G, h \rangle \in \text{VERTEX COVER} \qquad \text{iff} \qquad \langle \mathcal{M}_G, h + n + 1 \rangle \in \text{SBA-1} \tag{13}$$

By construction, $\mathcal{M}_G$ is minimal and its underlying graph H has a unique bottom strongly connected component, namely the self-loop in e_0. Each path $p = e_n \rightsquigarrow e_0$ in H passes through all edge states, and the set of labels of the endpoint states in p is a vertex cover of G. Since $e_0, \dots, e_n$ have pairwise distinct labels, we have that G has a vertex cover of size at most h iff there exists a path in H from e_n to e_0 that has at most $n + 1 + h$ different labels. Thus, (15) follows by Lemma 15. ◀

6 An Expectation Maximization-like Heuristic

In this section we describe an approximation algorithm for determining suboptimal solutions of CBA-λ for an arbitrary instance $\langle \mathcal{M}, k \rangle$.

Given an initial approximant $\mathcal{N}_0 \in \text{MC}(k)$, the algorithm produces a sequence of MCs $\mathcal{N}_0, \mathcal{N}_1, \dots$ in $\text{MC}(k)$ having successively decreased distance from $\mathcal{M}$. We defer until later a discussion of how the initial MC $\mathcal{N}_0$ is chosen. The procedure is described in Algorithm 1.

The intuitive idea of the algorithm is to iteratively update the initial MC by assigning relatively greater probability to transitions that are most representative of the behavior of the MC $\mathcal{M}$ w.r.t. δ_λ. The procedure stops when the last iteration has not yield an improved approximant w.r.t. the preceding one. The input also includes a parameter $h \in \mathbb{N}$ that bounds the number of iterations.

The rest of the section explains two heuristics for implementing the UPDATETRANSITION function invoked at line 5. This function shall return the transition probabilities for the successive approximant (see line 6).

Algorithm 1 Approximate Minimization – Expectation Maximization-like heuristic

```
Input: $\mathcal{M} = (M, \tau, \ell)$, $\mathcal{N}_0 = (N, \theta_0, \alpha)$, and $h \in \mathbb{N}$.
1. $i \leftarrow 0$
2. repeat
3.     $i \leftarrow i + 1$
4.     compute $\mathcal{C} \in \Omega(\mathcal{M}, \mathcal{N}_{i-1})$ such that $\delta_\lambda(\mathcal{M}, \mathcal{N}_{i-1}) = \gamma_\lambda^{\mathcal{C}}(\mathcal{M}, \mathcal{N}_{i-1})$
5.     $\theta_i \leftarrow \textsc{UpdateTransition}(\theta_{i-1}, \mathcal{C})$
6.     $\mathcal{N}_i \leftarrow (N, \theta_i, \alpha)$
7. until $\delta_\lambda(\mathcal{M}, \mathcal{N}_i) > \delta_\lambda(\mathcal{M}, \mathcal{N}_{i-1})$ or $i \geq h$
8. return $\mathcal{N}_{i-1}$
```

Define $\beta_\lambda^{\mathcal{C}}$ to be the least fixed-point of the following functional operator on 1-bounded real-valued functions $d \colon M \times N \to [0,1]$ (ordered point-wise):

$$B_\lambda^{\mathcal{C}}(d)(m,n) = \begin{cases} 1 & \text{if } \gamma_\lambda^{\mathcal{C}}(m,n) = 0 \\ 0 & \text{if } \ell(m) \neq \alpha(n) \\ (1-\lambda) + \lambda \int_{M \times N} d \, \mathrm{d}\mathcal{C}(m,n) & \text{otherwise}\,. \end{cases}$$

By Theorem 5, the relation $R_{\mathcal{C}} = \big\{(m,n) \mid \gamma_\lambda^{\mathcal{C}}(m,n) = 0\big\}$ is easily shown to be a bisimulation, specifically, the greatest bisimulation induced by $\mathcal{C}$.

Define $\mathcal{C}_\lambda$ as the MC obtained by augmenting $\mathcal{C}$ with an ‘sink’ state $\perp$ to which any other state moves with probability $(1-\lambda)$. Intuitively, the value $\beta_\lambda^{\mathcal{C}}(m,n)$ can be interpreted as the reachability probability in $\mathcal{C}_\lambda$ of either hitting the sink state or a pair of bisimilar states in $R_{\mathcal{C}}$ along a path formed only by pairs of states with identical labels starting from (m,n).

▶ **Lemma 17.** *For all $m \in M$ and $n \in N$, $\beta_\lambda^{\mathcal{C}}(m,n) = 1 - \gamma_\lambda^{\mathcal{C}}(m,n)$.*

From equation (3) and Lemma 18, we can turn the problem CBA-λ as

$$\operatorname{argmax} \big\{\beta_\lambda^{\mathcal{C}}(\mathcal{M}, \mathcal{N}) \mid \mathcal{N} \in \mathrm{MC}_{L(\mathcal{M})}(k),\, \mathcal{C} \in \Omega(\mathcal{M}, \mathcal{N})\big\}\ . \tag{14}$$

Equation (16) says that a solution of CBA-λ is the right marginal of a coupling structure $\mathcal{C}$ such that $\mathcal{C}_\lambda$ maximizes the probability of generating paths with prefix in $\cong^*(R_{\mathcal{C}} \cup \perp)$ starting from the pair (m_0, n_0) of initial states[1], where $\cong = \{(m,n) \notin R_{\mathcal{C}} \mid \ell(m) = \alpha(n)\}$.

In the rest of the section we assume $\mathcal{N}_{i-1} \in \mathrm{MC}(k)$ to be the current approximant with associated coupling structure $\mathcal{C} \in \Omega(\mathcal{M}, \mathcal{N}_{i-1})$ as in line 4 in Algorithm 1.

The “Averaged Marginal” Heuristic The first heuristic is inspired by the Expectation Maximization (EM) algorithm described in [7]. The idea is to count the expected number of occurrences of the transitions in $\mathcal{C}$ in the set of paths $\cong^* R_{\mathcal{C}}$ and, in accordance with (16), updating $\mathcal{C}$ by increasing the probability of the transitions that were contributing the most.

For each $m, u \in M$ and $n, v \in N$ let $Z_{u,v}^{m,n} \colon (M \times N)^\omega \to \mathbb{N}$ be the random variable that counts the number of occurrences of the edge $((m,n)(u,v))$ in a prefix in $\cong^*(R_{\mathcal{C}} \cup \perp)$ of the given path. We denote by $\mathbf{E}[Z_{u,v}^{m,n} | \mathcal{C}]$ the expected value of $Z_{u,v}^{m,n}$ w.r.t. the probability distribution induced by $\mathcal{C}_\lambda$. Using these values we define the optimization problem

[1] We borrowed notation from regular expressions, such as language union, concatenation, and Kleene star, to express the set of finite paths $\cong^* R_{\mathcal{C}}$ as a language over the alphabet $M \times N$.

EM$\langle\mathcal{N},\mathcal{C}\rangle$:

$$\begin{array}{lll r}
\text{maximize} & \sum_{m,u\in M}\sum_{n,v\in N}\mathbf{E}[Z^{m,n}_{u,v}|\,\mathcal{C}]\cdot\ln(c^{m,n}_{u,v}) & & \\
\text{such that} & \sum_{v\in N} c^{m,n}_{u,v} = \tau(m)(u) & m,u\in M,\ n\in N & (15)\\
 & \sum_{u\in M} c^{m,n}_{u,v} = \theta_{n,v} & m\in M,\ n,v\in N & (16)\\
 & c^{m,n}_{u,v}\geq 0 & m,u\in M,\ n,v\in N &
\end{array}$$

A solution of EM$\langle\mathcal{N},\mathcal{C}\rangle$ can be used to improve a pair $\langle\mathcal{N},\mathcal{C}\rangle$ in the sense of (16).

▶ **Theorem 18.** *If $\beta^{\mathcal{C}}_{\lambda}(\mathcal{M},\mathcal{N})>0$, then an optimal solution for EM$\langle\mathcal{N},\mathcal{C}\rangle$ describes an MC $\mathcal{N}'\in\mathrm{MC}(k)$ and a coupling structure $\mathcal{C}'\in\Omega(\mathcal{M},\mathcal{N}')$ such that $\beta^{\mathcal{C}'}_{\lambda}(\mathcal{M},\mathcal{N}')\geq\beta^{\mathcal{C}}_{\lambda}(\mathcal{M},\mathcal{N})$.*

Unfortunately, EM$\langle\mathcal{N},\mathcal{C}\rangle$ does not have an easy analytic solution and turns out to be inefficiently solved by nonlinear optimization methods. On the contrary, the relaxed optimization problem obtained by dropping the constraints (18) has a simple analytic solution, and the first heuristic at line 5, updates θ_i as follows[2]

$$c^{m,n}_{u,v}=\frac{\tau(m)(n)\cdot\mathbf{E}[Z^{m,n}_{u,v}|\,\mathcal{C}]}{\sum_{x\in N}\mathbf{E}[Z^{m,n}_{u,x}|\,\mathcal{C}]}\,,\quad \theta_i(n)(v)=\begin{cases}\theta_{i-1}(n)(v) & \text{if } \exists m\in M.\, n\ R_{\mathcal{C}}\ m\\ \dfrac{\sum_{m,u\in M}c^{m,n}_{u,v}}{\sum_{x\in N}\sum_{m,u\in M}c^{m,n}_{u,x}} & \text{otherwise}\end{cases}$$

Note that, the $c^{m,n}_{u,v}$ above may not describe a coupling structure. Nevertheless we recover the transition probability θ_i, from it by averaging the right marginals.

The "Averaged Expectations" Heuristic In contrast to the previous case, the second heuristic will update θ_i by directly averaging the expected values of $Z^{m,n}_{u,v}$ as follows

$$\theta_i(n)(v)=\begin{cases}\theta_{i-1}(n)(v) & \text{if } \exists m\in M.\, n\ R_{\mathcal{C}}\ m\\ \dfrac{\sum_{m,u\in M}\mathbf{E}[Z^{m,n}_{u,v}|\,\mathcal{C}]}{\sum_{x\in N}\sum_{m,u\in M}\mathbf{E}[Z^{m,n}_{u,x}|\,\mathcal{C}]} & \text{otherwise}\,.\end{cases}$$

Computing the Expected Values We compute $\mathbf{E}[Z^{m,n}_{u,v}|\,\mathcal{C}]$ using a variant of the *forward-backward* algorithm for hidden Markov models. Let $Z^{m,n}\colon(M\times N)^{\omega}\to\mathbb{N}$ be the random variable that counts the number of occurrences of the pair (m,n) in a prefix in $\cong^*(R_{\mathcal{C}}\cup\bot)$ of the path. We compute the expected value of $Z^{m,n}$ w.r.t. the probability induced by $\mathcal{C}_\lambda$ as the solution $z_{m,n}$ of the following system of equations

$$z_{m,n}=\begin{cases}0 & \text{if } m\ncong n\\ \iota(m,n)+\lambda\sum_{u,v}(z_{u,v}+1)\cdot\mathcal{C}(u,v)(m,n) & \text{otherwise}\,,\end{cases}$$

where ι denotes the characteristic function for $\{(m_0,n_0)\}$. Then, the expected value of $Z^{m,n}_{u,v}$ w.r.t. the probability induced by $\mathcal{C}_\lambda$ is given by $\mathbf{E}[Z^{m,n}_{u,v}|\,\mathcal{C}]=\lambda\cdot z_{m,n}\cdot\mathcal{C}(m,n)(u,v)\cdot\beta^{\mathcal{C}}_{\lambda}(u,v)$.

Choosing the initial approximant Similarly to EM algorithms, the choice of the initial approximant $\mathcal{N}_0$ may have a significant effect on the quality of the solution. For the labeling of the states, one should follow Lemma 7. As for the choice of the underlying structure one shall be guided by Lemma 15. However, due to Theorem 14, it seems unlikely to have generic good strategies for a starting approximant candidate. Nevertheless, good selections for the transition probabilities may be suggested by looking at the problem instance.

[2] By abusing the notation, whenever the nominator is 0, we consider entire expression equal to 0, regardless of any division by 0. The same convention is used implicitly in the rest of the section.

Case	$\|M\|$	k	$\lambda = 1$				$\lambda = 0.8$			
			δ_λ-init	δ_λ-final	h	time	δ_λ-init	δ_λ-final	h	time
IPv4 (AM)	53	5	0.856	0.062	3	25.7	0.667	0.029	3	25.9
	103	5	0.923	0.067	3	116.3	0.734	0.035	3	116.5
	53	6	0.757	0.030	3	39.4	0.544	0.011	3	39.4
	103	6	0.837	0.032	3	183.7	0.624	0.017	3	182.7
IPv4 (AE)	53	5	0.856	0.110	2	14.2	0.667	0.049	3	21.8
	103	5	0.923	0.110	2	67.1	0.734	0.049	3	100.4
	53	6	0.757	0.072	2	21.8	0.544	0.019	3	33.0
	103	6	0.837	0.072	2	105.9	0.624	0.019	3	159.5
DrkW (AM)	39	7	0.565	0.466	14	259.3	0.432	0.323	14	252.8
	49	7	0.568	0.460	14	453.7	0.433	0.322	14	420.5
	59	8	0.646	–	–	TO	0.423	–	–	TO
DrkW (AE)	39	7	0.565	0.435	11	156.6	0.432	0.321	2	28.6
	49	7	0.568	0.434	10	247.7	0.433	0.316	2	46.2
	59	8	0.646	0.435	10	588.9	0.423	0.309	2	115.7

Table 1 Comparison of the performance of Algorithm 1 on the IPv4 zeroconf protocol and the classic Drunkard's Walk w.r.t. the heuristics AM and AE.

Experimental Results Table 1 shows the results of some tests[3] on Algorithm 1. run on a number of instances $\langle \mathcal{M}, k \rangle$ of increasing size, where $\mathcal{M}$ is the bisimilarity quotient of either the IPv4 protocol [5, Ex.10.5] or the drunkard's walk, parametric on the number of states $|M|$. As initial approximant we use a suitably small instance of the same model. Each row reports the distance to the original model respectively from $\mathcal{N}_0$ and $\mathcal{N}_h$, where h is the total number of iterations; and execution time (in seconds). We compare the two heuristics, averaged marginals (AM) and averaged expectation (AE), on the same initial approximant.

The results obtained on the IPv4 protocol show significant improvements between the initial and the returned approximant. Notably, these are obtained in very few iterations. On this model, AM gives approximants of better quality compared with those obtained using AE; however AE seems to be slightly faster than AM. On the drunkard's walk model, the two heuristics exhibit opposite results w.r.t. the previous experiment: AE provides the best solutions with fewer iterations and significantly lower execution times.

7 Conclusions and Future Work

To the best of our knowledge, this is the first paper addressing the complexity of the *optimal* approximate minimization of MCs w.r.t. a behavioral metric semantics. Even though for a good evaluation of our heuristics more tests are needed, the current results seem promising. Moreover, in the light of [10, 3], relating the probabilistic bisimilarity distance to the LTL-model checking problem as $\delta_1(\mathcal{M},\mathcal{N}) \geq |\mathbb{P}_\mathcal{M}(\varphi) - \mathbb{P}_\mathcal{N}(\varphi)|$, for all $\varphi \in$ LTL, our results might be used to lead saving in the overall model checking time. A deeper study of this topic will be the focus of future work. We close with an interesting open problem. Membership of BA-λ in NP is left open. However, by arguments analogous to [11, 14] and leveraging on the ideas that made us produce the MC in Example 10, we suspect that BA-λ is hard for the square-root-sum problem. The latter is known to be NP-hard and in PSPACE, but membership in NP has been open since 1976. Allender et al. [1] showed that it can be decided in the 4th level of the counting hierarchy, thus it is unlikely its PSPACE-completeness.

[3] The tests are done on a prototype implementation coded in Mathematica® (`people.cs.aau.dk/giovbacci/tools.html`) running on an Intel Core-i5 2.5GHz with 8GB of DDR3 RAM 1600MHz.

References

1 Eric Allender, Peter B urgisser, Johan Kjeldgaard-Pedersen, and Peter Bro Miltersen. On the complexity of numerical analysis. *SIAM Journal on Computing*, 38(5):1987–2006, 2009. doi:10.1137/070697926.

2 Rajeev Alur, Costas Courcoubetis, Nicolas Halbwachs, David L. Dill, and Howard Wong-Toi. Minimization of timed transition systems. In *CONCUR*, volume 630 of *Lecture Notes in Computer Science*, pages 340–354. Springer, 1992. doi:10.1007/BFb0084802.

3 Giorgio Bacci, Giovanni Bacci, Kim G. Larsen, and Radu Mardare. Converging from Branching to Linear Metrics on Markov Chains. In *ICTAC*, volume 9399 of *LNCS*, pages 349–367. Springer, 2015. doi:10.1007/978-3-319-25150-9_21.

4 Christel Baier. Polynomial time algorithms for testing probabilistic bisimulation and simulation. In *CAV*, volume 1102 of *Lecture Notes in Computer Science*, pages 50–61. Springer, 1996. doi:10.1007/3-540-61474-5_57.

5 Christel Baier and Joost-Pieter Katoen. *Principles of Model Checking*. MIT Press, 2008.

6 Borja Balle, Prakash Panangaden, and Doina Precup. A canonical form for weighted automata and applications to approximate minimization. In *LICS*, pages 701–712. IEEE Computer Society, 2015. doi:10.1109/LICS.2015.70.

7 Michael Benedikt, Rastislav Lenhardt, and James Worrell. LTL Model Checking of Interval Markov Chains. In *TACAS*, volume 7795 of *Lecture Notes in Computer Science*, pages 32–46. Springer, 2013. doi:10.1007/978-3-642-36742-7_3.

8 Stefan Blom and Simona Orzan. A distributed algorithm for strong bisimulation reduction of state spaces. *International Journal on Software Tools for Technology Transfer*, 7(1):74–86, 2005. doi:10.1007/s10009-004-0159-4.

9 John F. Canny. Some Algebraic and Geometric Computations in PSPACE. In *Proceedings of the 20th Annual ACM Symposium on Theory of Computing (STOC'88)*, pages 460–467. ACM, 1988. doi:10.1145/62212.62257.

10 Di Chen, Franck van Breugel, and James Worrell. On the Complexity of Computing Probabilistic Bisimilarity. In *FoSSaCS*, volume 7213 of *LNCS*, pages 437–451. Springer, 2012.

11 Taolue Chen and Stefan Kiefer. On the Total Variation Distance of Labelled Markov Chains. In *CSL-LICS'14*, pages 33:1–33:10. ACM, 2014. doi:10.1145/2603088.2603099.

12 Josee Desharnais, Vineet Gupta, Radha Jagadeesan, and Prakash Panangaden. Metrics for Labeled Markov Systems. In *CONCUR*, volume 1664 of *LNCS*, pages 258–273. Springer, 1999. doi:10.1007/3-540-48320-9_19.

13 Josee Desharnais, Vineet Gupta, Radha Jagadeesan, and Prakash Panangaden. Metrics for labelled Markov processes. *Theoretical Compututer Science*, 318(3):323–354, 2004.

14 Kousha Etessami and Mihalis Yannakakis. Recursive Markov chains, stochastic grammars, and monotone systems of nonlinear equations. *J. ACM*, 56(1):1:1–1:66, 2009. doi:10.1145/1462153.1462154.

15 Norm Ferns, Prakash Panangaden, and Doina Precup. Metrics for finite Markov Decision Processes. In *UAI*, pages 162–169. AUAI Press, 2004.

16 Giuliana Franceschinis and Richard R. Muntz. Bounds for quasi-lumpable markov chains. *Perform. Eval.*, 20(1-3):223–243, 1994. doi:10.1016/0166-5316(94)90015-9.

17 John Hopcroft. An $n \log n$ algorithm for minimizing states in a finite automaton. In Zvi Kohavi and Azaria Paz, editors, *Theory of Machines and Computations*, pages 189–196. Academic Press, 1971. doi:10.1016/B978-0-12-417750-5.50022-1.

18 Chi-Chang Jou and Scott A.Smolka. Equivalences, congruences, and complete axiomatizations for probabilistic processes. In *CONCUR'90 Theories of Concurrency: Unification and Extension*, volume 458 of *LNCS*, pages 367–383, 1990.

19 Paris C. Kanellakis and Scott A. Smolka. CCS expressions, finite state processes, and three problems of equivalence. In *Proceedings of the 2nd Annual ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing*, pages 228–240. ACM, 1983. doi:10.1145/800221.806724.

20 Paris C. Kanellakis and Scott A. Smolka. CCS expressions, finite state processes, and three problems of equivalence. *Information and Computation*, 86(1):43–68, 1990. doi:http://dx.doi.org/10.1016/0890-5401(90)90025-D.

21 Michal Ko1cmcvara and Michael Stingl. PENBMI 2.0. http://www.penopt.com/penbmi.html. Accessed: 2016-08-28.

22 Michal Ko1cmcvara and Michael Stingl. PENNON: A code for convex nonlinear and semidefinite programming. *Optimization Methods and Software*, 18(3):317–333, 2003. doi:10.1080/1055678031000098773.

23 Kim Guldstrand Larsen and Arne Skou. Bisimulation through probabilistic testing. *Information and Computation*, 94(1):1–28, 1991.

24 David Lee and Mihalis Yannakakis. Online minimization of transition systems (extended abstract). In *Annual ACM Symposium on Theory of Computing*, pages 264–274. ACM, 1992. doi:10.1145/129712.129738.

25 Geoffrey J. McLachlan and Thriyambakam Krishnan. *The EM Algorithm and Extensions.* Wiley-Interscience, 2 edition, 2008.

26 Robin Milner. *A Calculus of Communicating Systems*, volume 92 of *Lecture Notes in Computer Science*. Springer, 1980. doi:10.1007/3-540-10235-3.

27 Edward F. Moore. Gedanken Experiments on Sequential Machines. In *Automata Studies*, pages 129–153. Princeton University, 1956.

28 Franck van Breugel and James Worrell. Towards Quantitative Verification of Probabilistic Transition Systems. In *ICALP*, volume 2076 of *LNCS*, pages 421–432, 2001.

29 Franck van Breugel and James Worrell. Approximating and computing behavioural distances in probabilistic transition systems. *Theoretical Computer Science*, 360(3):373–385, 2006.

30 Mihali Yannakakis and David Lee. An efficient algorithm for minimizing real-time transition systems. *Formal Methods in System Design*, 11(2):113–136, 1997. doi:10.1023/A:1008621829508.

31 Shipei Zhang and Scott A. Smolka. Towards efficient parallelization of equivalence checking algorithms. In *FORTE*, volume C-10 of *IFIP Transactions*, pages 121–135. North-Holland, 1992.

A Missing Proofs of the Technical Results

Proof of Lemma 7. Let $\mathcal{N}' = (N', \theta', \alpha')$. If $L(\mathcal{N}') \subseteq L(\mathcal{M})$, take $\mathcal{N} = \mathcal{N}'$. Otherwise, define $\mathcal{N} = (N, \theta, \alpha)$ as follows: $N = N'$, $\theta = \theta'$, and $\alpha(n) = \alpha'(n)$ if $\alpha'(n) \in L(\mathcal{M})$, otherwise $\alpha(n) = \ell(m_0)$, where m_0 is the initial state of $\mathcal{M}$. The initial state of $\mathcal{N}$ is the one of $\mathcal{N}'$. Clearly, $\mathcal{N} \in \mathrm{MC}(k)$ and $L(\mathcal{N}) \subseteq L(\mathcal{M})$.

Let $\mathcal{A} = \mathcal{M} \oplus \mathcal{N}$ and $\mathcal{B} = \mathcal{M} \oplus \mathcal{N}'$ be the disjoint union of $\mathcal{M}$ with $\mathcal{N}$ and $\mathcal{N}'$ respectively. We prove $\mathcal{N} \preceq_\lambda \mathcal{N}'$ by showing $\delta^{\mathcal{A}}_\lambda \sqsubseteq \delta^{\mathcal{B}}_\lambda$. By Tarski fixed-point theorem, it suffices to show $\Psi^{\mathcal{A}}_\lambda(\delta^{\mathcal{B}}_\lambda) \sqsubseteq \delta^{\mathcal{B}}_\lambda$. Let $u, v \in M \cup N$. When u and v have different labels in $\mathcal{B}$, then, $\Psi^{\mathcal{A}}_\lambda(\delta^{\mathcal{B}}_\lambda)(u, v) \leq 1 = \delta^{\mathcal{B}}_\lambda(u, v)$ follows by definition of Ψ_λ and the fact that $\delta^{\mathcal{B}}_\lambda = \Psi^{\mathcal{B}}_\lambda(\delta^{\mathcal{B}}_\lambda)$. Assume u and v have the same label in $\mathcal{B}$. Then, by construction of $\mathcal{N}$ (i.e, by definition of α'), u and v have the same label in $\mathcal{A}$. By the fact that $\mathcal{N}$ and $\mathcal{N}'$ have the same transition distribution function, one readily checks that $\Psi^{\mathcal{A}}_\lambda(\delta^{\mathcal{B}}_\lambda)(u, v) = \delta^{\mathcal{B}}_\lambda(u, v)$. ◀

Proof of Lemma 12. By Lemma 7 it suffices to show that for any $\mathcal{N}' \in \mathrm{MC}_{L(\mathcal{M})}(k)$ there exists $\mathcal{N} \in \mathrm{MC}_{L(\mathcal{M})}(k)$ with injective labeling such that $\delta_\lambda(\mathcal{M}, \mathcal{N}) \leq \delta_\lambda(\mathcal{M}, \mathcal{N}')$.

Assume $\mathcal{N}'$ does not have injective labeling. By Theorem 5, there exists $\mathcal{C} \in \Omega(\mathcal{M}, \mathcal{N}')$ such that $\delta_\lambda(\mathcal{M}, \mathcal{N}') = \gamma^{\mathcal{C}}_\lambda(\mathcal{M}, \mathcal{N}')$. Consider $m \in M$ and $n \in N$ with same label (i.e., $\ell(m) = \alpha(n)$). We can construct $\omega_n \in \mathcal{D}(M \times N)$ satisfying the following

$$\omega_n(u, v) \geq \mathcal{C}(m, n)(u, v) \qquad \text{for all } u \in M \text{ and } v \in N \text{ s.t. } \ell(u) = \ell(v) \tag{17}$$

$$\textstyle\sum_{v \in N} \omega_n(u, v) = \tau(m)(u) \qquad \text{for all } u \in M \tag{18}$$

$$\textstyle\sum_{v \in \alpha^{-1}(\ell(u))} \omega_n(u, v) = \tau(m)(u) \qquad \text{for all } u \in M \text{ s.t. } \ell(u) \in L(\mathcal{N}) \tag{19}$$

$$\omega_n(u, n) = \tau(m)(u) \qquad \text{for all } u \in M \text{ s.t. } \ell(u) \notin L(\mathcal{N}) \tag{20}$$

Let $\mathcal{N}'' = (N, \theta'', \alpha)$ where, for $n, v \in N$, $\theta''(n)(v) = \sum_{u \in M} \omega_n(u, v)$. Note that θ'' is well defined because $\mathcal{M}$ has injective labeling. Next we show that $\delta_\lambda(\mathcal{M}, \mathcal{N}'') \leq \delta_\lambda(\mathcal{M}, \mathcal{N}')$.

By Theorem 5 and Tarski's fixed point theorem it suffices to find $\mathcal{C}'' \in \Omega(\mathcal{M}, \mathcal{N}'')$ such that $\Gamma^{\mathcal{C}''}_\lambda(\gamma^{\mathcal{C}}_\lambda) \sqsubseteq \gamma^{\mathcal{C}}_\lambda$. Take any $\mathcal{C}''$ such that $\mathcal{C}''(m, n) = \omega_n$ whenever $\ell(m) = \alpha(n)$. Note that by (20), $\omega_n \in \Omega(\tau(m), \theta''(n))$. Next we show that, for all $m \in M$ and $n \in N$, $\Gamma^{\mathcal{C}''}_\lambda(\gamma^{\mathcal{C}}_\lambda)(m, n) \leq \gamma^{\mathcal{C}}_\lambda(m, n)$. If $\ell(m) \neq \alpha(n)$, the inequality holds because $\gamma^{\mathcal{C}}_\lambda(m, n) = 1$. If $\ell(m) \neq \alpha(n)$, the following hold

$$\begin{aligned}
\textstyle\sum_{u,v} \mathcal{C}(m, n)(u, v) - \omega_n(u, v) &= 0 && (\text{by } \mathcal{C}(m, n), \omega_n \in \mathcal{D}(M \times N)) \\
\textstyle\sum_{\ell(u) \neq \alpha(v)} \mathcal{C}(m, n)(u, v) - \omega_n(u, v) &= \textstyle\sum_{\ell(u) = \alpha(v)} \omega_n(u, v) - \mathcal{C}(m, n)(u, v) \\
\textstyle\sum_{\ell(u) \neq \alpha(v)} \mathcal{C}(m, n)(u, v) - \omega_n(u, v) &\geq \textstyle\sum_{\ell(u) = \alpha(v)} \gamma^{\mathcal{C}}_\lambda(u, v)\big(\omega_n(u, v) - \mathcal{C}(m, n)(u, v)\big) \\
&&& (\text{by (19) and } \gamma^{\mathcal{C}}_\lambda \sqsubseteq \mathbf{1}) \\
\textstyle\sum_{u,v} \gamma^{\mathcal{C}}_\lambda(u, v) \cdot \mathcal{C}(m, n)(u, v) &\geq \textstyle\sum_{u,v} \gamma^{\mathcal{C}}_\lambda(u, v) \cdot \omega_n(u, v) \\
&&& (\text{for } \ell(u) \neq \alpha(v)\ \gamma^{\mathcal{C}}(u, v) = 1) \\
\gamma^{\mathcal{C}}_\lambda(m, n) &\geq \lambda \textstyle\sum_{u,v} \gamma^{\mathcal{C}}_\lambda(u, v) \cdot \omega_n(u, v) && (\text{by def. } \gamma^{\mathcal{C}}_\lambda \text{ and } \lambda > 0) \\
\gamma^{\mathcal{C}}_\lambda(m, n) &\geq \Gamma^{\mathcal{C}''}_\lambda(\gamma^{\mathcal{C}}_\lambda)(m, n) && (\text{by def. } \mathcal{C}'' \text{ and } \Gamma^{\mathcal{C}''}_\lambda)
\end{aligned}$$

So far we proved that $\delta_\lambda(\mathcal{M}, \mathcal{N}'') \leq \delta_\lambda(\mathcal{M}, \mathcal{N}')$. $\mathcal{N}''$ may not have injective labeling, however we will show that its bisimilarity quotient has injective labeling function. To prove that we will show that the relation $R = \{(n, v) \mid \alpha(n) = \alpha(v)\} \subseteq N \times N$ is a probabilistic bisimulation. R is readily seen to be equivalence relation that preserves labeling. It only remains to prove that if $n \mathrel{R} v$, then for any $C \in N/_R$, $\theta''(n)(C) = \theta''(v)(C)$. Let $m \in M$

be the unique state of $\mathcal{M}$ such that $\ell(m) = \alpha(n) = \alpha(v)$ and $m' \in M$ the one that has the same label as any element of C. We consider two cases.

1. If $n, v \in C$. Let $M' = \{m \in M \mid \ell(u) \notin L(\mathcal{N})\}$, then the following equalities hold

$$\begin{aligned}
\theta''(n)(C) &= \textstyle\sum_{c \in C} \sum_{u \in M} \omega_n(u, c) && \text{(by def. } \theta''\text{)} \\
&= \textstyle\sum_{c \in C} \omega_n(m', c) + \sum_{u \in M'} \omega_n(u, n) \\
& \qquad \text{(by (20), (21) and (22), } \omega_n(u, c) > 0 \text{ implies } \ell(u) = \alpha(c) \text{ or } c = n\text{)} \\
&= \tau(m)(m') + \tau(m)(M') && \text{(by (21) and (22))} \\
&= \textstyle\sum_{c \in C} \omega_v(m', c) + \sum_{u \in M'} \omega_v(u, v) \\
&= \textstyle\sum_{c \in C} \sum_{u \in M} \omega_v(u, c) = \theta''(v)(C)
\end{aligned}$$

2. If $n, v \notin C$, then the following equalities hold

$$\begin{aligned}
\theta''(n)(C) &= \textstyle\sum_{c \in C} \sum_{u \in M} \omega_n(u, c) && \text{(by def. } \theta''\text{)} \\
&= \textstyle\sum_{c \in C} \omega_n(m', c) && \text{(by (20) and (21) } \omega_n(u, c) > 0 \text{ implies } \ell(u) = \alpha(c)\text{)} \\
&= \tau(m)(m') + \tau(m)(M') && \text{(by (21))} \\
&= \textstyle\sum_{c \in C} \omega_v(m', c) \\
&= \textstyle\sum_{c \in C} \sum_{u \in M} \omega_v(u, c) = \theta''(v)(C)
\end{aligned}$$

This proves the thesis. ◀

Proof of Lemma 13. The thesis holds trivially when $\ell(m) \neq \alpha(n)$, since $\delta_\lambda(m, n) = 1$.

Let $\ell(m) = \alpha(n)$, and $M' = \{u \in M \mid \ell(u) \notin L(\mathcal{N})\}$, then the following hold

$$\begin{aligned}
\delta_\lambda(m, n) &= \lambda \textstyle\sum_{u \in M} \sum_{v \in N} \delta_\lambda(u, v) \cdot \omega(u, v) && \text{(for some } \omega \in \Omega(\tau(m), \theta(n))\text{)} \\
&\geq \lambda \textstyle\sum_{u \in M'} \sum_{v \in N} \delta_\lambda(u, v) \cdot \omega(u, v) && \text{(by } M' \subseteq M\text{)} \\
&= \lambda \textstyle\sum_{u \in M'} \sum_{v \in N} \omega(u, v) && (\delta_\lambda(u, v) = 1 \text{ for all } u \in M' \text{ and } n \in N) \\
&= \lambda \cdot \tau(m)(M') && \text{(by } \omega \in \Omega(\tau(m), \theta(n))\text{)}
\end{aligned}$$

◀

Proof of Lemma 15. ($\Rightarrow$) By hypothesis there exists $\mathcal{N} \in \mathrm{MC}(k)$ such that $\delta_1(\mathcal{M}, \mathcal{N}) < 1$. We can assume without loss of generality that $\mathcal{N}$ is minimal (otherwise one can replace it with its bisimilarity quotient). By Lemma 18 and Theorem 5, there exists $\mathcal{C} \in \Omega(\mathcal{M}, \mathcal{N})$ such that $\beta_1^{\mathcal{C}}(\mathcal{M}, \mathcal{N}) > 0$. Therefore there exists a path $(m_0, n_0) \dots (m_p, n_p)$ in $G(\mathcal{C}_1)$, such that $\ell(m_i) = \alpha(n_i)$ (for $i = 0..h$) and $\gamma_1^{\mathcal{C}}(m_p, n_p) = 0$.

Note that for arbitrary $m \in M$ and $n \in N$ such that $\gamma_1^{\mathcal{C}}(m, n) = 0$, the following hold

$$0 = \gamma_1^{\mathcal{C}}(m, n) = \textstyle\sum_{u \in M} \sum_{v \in N} \gamma_1^{\mathcal{C}}(u, v) \cdot \mathcal{C}(m, n)(u, v) \,.$$

Therefore, for any $u \in M$ and $v \in N$ we have that $\mathcal{C}(m, n)(u, v) > 0$ implies $\gamma_1^{\mathcal{C}}(u, v) = 0$.

Let $R \subseteq M \times N$ be the set of states reachable from (m_p, n_p) in $G(\mathcal{C}_1)$. By $\gamma_1^{\mathcal{C}}(m_p, n_p) = 0$ and what have been said before we have that $(m, n) \in R$ implies that $m \sim n$. Let $G = (V, E)$ be a bottom strongly connected component of $G(\mathcal{C}_1)$ such that $V \subseteq R$.

Consider now the graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$ where

$$V_1 = \{m \mid (m, n) \in V\} \qquad E_1 = \{(m, u) \mid \tau(m)(u) > 0 \text{ and } m, u \in V_1\}, \tag{21}$$

$$V_2 = \{n \mid (m, n) \in V\} \qquad E_2 = \{(n, v) \mid \theta(n)(v) > 0 \text{ and } n, v \in V_1\}. \tag{22}$$

Since $\mathcal{M}$ and $\mathcal{N}$ are minimal and $m \sim n$ for all $(m,n) \in V$ we have that for all $(m,n),(u,v) \in V$, $C(m,n)(u,v) = \tau(m)(u) = \theta(n)(v)$. Therefore G_1 and G_2 are bottom strongly connected components of $G(\mathcal{M})$ and $G(\mathcal{N})$ respectively, and are isomorphic with each other.

Let now take the path $(m_0,n_0)\dots(m_h,n_h)$ in $G(\mathcal{C}_1)$ such that $(m_h,n_h) \in V$ obtained by appending extending the path $(m_0,n_0)\dots(m_p,n_p)$ by following a path $(m_p,n_p)\dots(m_h,n_h)$. Note that such a path exists since $(m_h,n_h) \in R$ and, $\ell(m_i) = \alpha(n_i)$ for all $0 \leq i \leq h$.

There are two possible cases:

If $n_i \notin V_2$ for all $0 \leq i < h$ we have that

$$|\{\ell(m_i) \mid i < h\}| + |V_1| = |\{\ell(n_i) \mid i < h\}| + |V_2| \leq |\{n_i \mid i < h\}| + |V_2| \leq |N| \leq k\,.$$

If $n_i \in V_2$ for some $0 \leq i < h$. Let $q < h$ be the smallest index such that $n_q \in V_2$. Since $C(m_q,n_q)(n_{q+1},n_{q+1}) > 0$ implies $\theta(n_q)(n_{q+1}) > 0$ and G_2 is a bottom strongly connected component, we have that also $n_{q+1} \in V$. This shows that $n_q \dots n_h$ is a path in G_2. Since the isomorphism between G_2 and G_1 preserves the labels (indeed, any $n \in V_2$ is mapped with the unique state $m \in V_1$ such that $m \sim n$) we can see that there exists a path $v_p \dots v_{h-1}m_h$ in G_1 such that $\ell(m_i) = \ell(v_i)$ for all $p \leq i < h$. For this we have that

$$\begin{aligned}
&|\{\ell(m_i) \mid i < h, \nexists \text{ a path } v_i \dots v_{h-1}m_h \text{ in } G' \text{ s.t } \forall j.\, \ell(m_j) = \ell(v_j)\}| + |V_1| \\
&\leq |\{\ell(n_i) \mid i < p\}| + |V_2| \leq |\{n_i \mid i < p\}| + |V_2| \leq |N| \leq k\,.
\end{aligned}$$

($\Leftarrow$) Let $M' = \{m_0,\dots,m_h\}$. Assume w.l.o.g. that $M' \cap V = \{m_h\}$ (otherwise one can consider a prefix of the path that enjoys the assumption). Let $p \in \{0,\dots,h\}$ be the smallest index such that there exists a path $v_p \dots v_{h-1}m_h$ in G' with $\ell(m_j) = \ell(v_j)$ for all $p \leq j < h$. To simplify the notation later on we will also use v_h to refer to m_h.

Let $\{l_0,\dots,l_q\} = \{\ell(m_i) \mid i < p\}$ and $N' = \{n_0,\dots,n_q\}$. Consider the chain $\mathcal{N} = (N,\theta,\alpha)$ where $N = N' \cup V$, and

$$\theta(n) = \begin{cases} \tau(n) & \text{if } n \in V \\ \mu & \text{if } n \in N' \end{cases} \qquad\qquad \alpha(n) = \begin{cases} \ell(n) & \text{if } n \in V \\ l_i & \text{if } n = n_i \text{ for some } 0 \leq i \leq q \end{cases}$$

where μ denotes the uniform distribution with support N, i.e., $\mu(n) = 1/|N|$ for all $n \in N$. Note that θ is well defined because the support of $\tau(v)$ is included in V for all $v \in V$.

By construction states in N' have pairwise distinct labels, therefore we can define the function $f\colon M' \to N$ as $f(m_i) = n$ if $0 \leq i < p$ and $n \in N'$ such that $\alpha(n) = \ell(m_i)$; and $f(m_i) = v_i$ if $p \leq i \leq h$. In the following we will prove that for all $m_i \in M'$, $\delta_1(m_i, f(m_i)) < 1$. We proceed by induction on $r = h - i$.

Base Case ($i = h$): One can readily show that $\delta_1(m_h,m_h) = 0$.
Inductive Step ($i < h$): Let $n = f(m_i)$ and $n' = f(m_{i+1})$ then the following hold

$$\begin{aligned}
&\delta_1(m_i, f(m_i)) = \mathcal{K}(d)(\tau(m_i),\theta(n)) && (\ell(m_i) = \ell) \\
&\leq \textstyle\sum_{u \in M}\sum_{v \in N} \tau(m_i)(u)\cdot\theta(n)(v)\cdot\delta_1(u,v) && (\tau(m_i)\odot\theta(n) \in \Omega(\tau(m_i),\theta(n))) \\
&\leq \tau(m_i)(m_{i+1})\cdot\theta(n)(n')\cdot\delta_1(m_{i+1},n') + (1 - \tau(m_i)(m_{i+1})\cdot\theta(n)(n')) && (\delta_1 \sqsubseteq \mathbf{1}) \\
&< 1 && (\delta_1(m_{i+1}, f(m_{i+1})) < 1 \text{ and } \tau(m_i)(m_{i+1})\cdot\theta(n)(n') > 0)
\end{aligned}$$

This proves that $\delta_1(\mathcal{M},\mathcal{N}) = \delta_1(m_0, f(m_0)) < 1$. By construction $|N| \leq k$, therefore $\langle\mathcal{M},k\rangle \in$ SBA-1. ◀

Proof of Lemma 18. We prove the equivalent statement $\gamma_\lambda^{\mathcal{C}} = \mathbf{1} - \beta_\lambda^{\mathcal{C}}$. Consider the following operator

$$G_\lambda^{\mathcal{C}}(d)(m,n) = \begin{cases} 0 & \text{if } \gamma_\lambda^{\mathcal{C}}(m,n) = 0 \\ 1 & \text{if } \ell(m) \neq \ell(n) \\ \lambda \int d \, \mathrm{d}\mathcal{C}(m,n) & \text{otherwise}\,. \end{cases}$$

One can prove that $G_\lambda^{\mathcal{C}}$ has a unique fixed point and that $\gamma_\lambda^{\mathcal{C}} = G_\lambda^{\mathcal{C}}(\gamma_\lambda^{\mathcal{C}})$. We will complete the proof by showing that also $\mathbf{1} - \beta_\lambda^{\mathcal{C}}$ is a fixed point for $G_\lambda^{\mathcal{C}}$. Let $m \in M$ and $n \in N$, if $\gamma_\lambda^{\mathcal{C}}(m,n) = 0$ or $\ell(m) \neq \alpha(n)$ it trivially holds that $1 - \beta_\lambda^{\mathcal{C}}(m,n) = G_\lambda^{\mathcal{C}}(\mathbf{1} - \beta_\lambda^{\mathcal{C}})(m,n)$. If $\gamma_\lambda^{\mathcal{C}}(m,n) > 0$ and $\ell(m) = \alpha(n)$, then the following equalities hold

$$\begin{aligned} G_\lambda^{\mathcal{C}}(\mathbf{1} - \beta_\lambda^{\mathcal{C}})(m,n) &= \lambda \textstyle\int (\mathbf{1} - \beta_\lambda^{\mathcal{C}}) \, \mathrm{d}\mathcal{C}(m,n) && (\text{by def. } G_\lambda^{\mathcal{C}}) \\ &= \lambda - \lambda \textstyle\int \beta_\lambda^{\mathcal{C}} \, \mathrm{d}\mathcal{C}(m,n) && (\textstyle\int \mathbf{1} \, \mathrm{d}\mathcal{C}(m,n) = 1) \\ &= 1 - \beta_\lambda^{\mathcal{C}}(m,n) && (\beta_\lambda^{\mathcal{C}}(m,n) = (1-\lambda) + \lambda \textstyle\int \beta_\lambda^{\mathcal{C}} \, \mathrm{d}\mathcal{C}(m,n)) \end{aligned}$$

◀

Proof of Theorem 5. For any measurable set $A \subseteq (M \times N)^\omega$, and $\mathcal{C} \in \Omega(\mathcal{M}, \mathcal{N})$, we denote by $\mathbb{P}_{\mathcal{C}_\lambda}(A)$ the probability that a run of the chain $\mathcal{C}_\lambda$ belongs to A. To shorten the notation, for $B \subseteq (M \times N)^*$ (resp. $\pi \in (M \times N)^*$) we write $\mathbb{P}_{\mathcal{C}}(B)$ (resp. $\mathbb{P}_{\mathcal{C}}(\pi)$) to indicate $\mathbb{P}_{\mathcal{C}_\lambda}(B(M \times N)^\omega)$ (resp. $\mathbb{P}_{\mathcal{C}_\lambda}(\pi(M \times N)^\omega)$).

Racall that, $\beta_\lambda^{\mathcal{C}}(m_0, n_0) = \mathbb{P}_{\mathcal{C}}(G)$ where $G = (\cong^*(R_{\mathcal{C}} \cup \bot))$ that is the probability that $\mathcal{C}_\lambda$ generates a path with prefix in $\cong^* R_{\mathcal{C}}$ or $\cong^* \bot$ starting from (m_0, n_0).

Consider the following inequalities

$$\begin{aligned} &\ln \beta_\lambda^{\mathcal{C}'}(m_0,n_0) - \ln \beta_\lambda^{\mathcal{C}}(m_0,n_0) = \ln \frac{\mathbb{P}_{\mathcal{C}'}(G)}{\mathbb{P}_{\mathcal{C}}(G)} \\ &= \ln \frac{\sum_{\pi \in G} \mathbb{P}_{\mathcal{C}'}(G \mid \pi) \cdot \mathbb{P}_{\mathcal{C}'}(\pi)}{\mathbb{P}_{\mathcal{C}}(G)} = \ln \sum_{\pi \in G} \frac{\mathbb{P}_{\mathcal{C}'}(G \mid \pi) \cdot \mathbb{P}_{\mathcal{C}'}(\pi)}{\mathbb{P}_{\mathcal{C}}(G)} \cdot \frac{\mathbb{P}_{\mathcal{C}}(\pi \mid G)}{\mathbb{P}_{\mathcal{C}}(\pi \mid G)} \\ &\geq \sum_{\pi \in G} \mathbb{P}_{\mathcal{C}}(\pi \mid G) \cdot \ln \frac{\mathbb{P}_{\mathcal{C}'}(G \mid \pi) \cdot \mathbb{P}_{\mathcal{C}'}(\pi)}{\mathbb{P}_{\mathcal{C}}(G) \cdot \mathbb{P}_{\mathcal{C}}(\pi \mid G)} \qquad (\text{by Jensen's inequality}) \\ &= \frac{1}{\beta_\lambda^{\mathcal{C}}(m_0,n_0)} \sum_{\pi \in G} \mathbb{P}_{\mathcal{C}}(\pi) \cdot \ln \frac{\mathbb{P}_{\mathcal{C}'}(\pi)}{\mathbb{P}_{\mathcal{C}}(\pi)} = \frac{1}{\beta_\lambda^{\mathcal{C}}(m_0,n_0)} (Q' - Q) \end{aligned}$$

where $Q' = \sum_{\pi \in G} \mathbb{P}_{\mathcal{C}}(\pi) \cdot \ln \mathbb{P}_{\mathcal{C}'}(\pi)$ and $Q = \sum_{\pi \in G} \mathbb{P}_{\mathcal{C}}(\pi) \cdot \ln \mathbb{P}_{\mathcal{C}}(\pi)$. Rearranging we have

$$\beta_\lambda^{\mathcal{C}}(m_0,n_0) \cdot \big(\ln \beta_\lambda^{\mathcal{C}'}(m_0,n_0) - \ln \beta_\lambda^{\mathcal{C}}(m_0,n_0)\big) \geq Q' - Q\,. \tag{23}$$

We have that $\ln \beta_\lambda^{\mathcal{C}'}(m_0,n_0) - \ln \beta_\lambda^{\mathcal{C}}(m_0,n_0) \geq 0$ iff $\beta_\lambda^{\mathcal{C}'}(m_0,n_0) \geq \beta_\lambda^{\mathcal{C}}(m_0,n_0)$, therefore by (25) we conclude that $Q' \geq Q$ implies $\beta_\lambda^{\mathcal{C}'}(m_0,n_0) \geq \beta_\lambda^{\mathcal{C}}(m_0,n_0)$.

Thus, inequality 25 suggests that the best choice of $\mathcal{C}'$ is that which maximizes Q' as a function of $\mathcal{C}'$. Expanding the definition of Q' we obtain

$$Q' = \sum_{\pi \in G} \mathbb{P}_{\mathcal{C}}(\pi) \cdot \ln \mathbb{P}_{\mathcal{C}'}(\pi) = \sum_{\pi \in G} \mathbb{P}_{\mathcal{C}}(\pi) \Big(\ln \iota(\pi_0) + \sum_{i=0}^{|\pi|-1} \ln \mathcal{C}'_\lambda(\pi_i)(\pi_{i+1}) \Big)\,,$$

where ι denotes the characteristic function for $\{(m_0, n_0)\}$.

For each $m, u \in M$ and $n, v \in N$ let $Z^{m,n}_{u,v} \colon (M \times N)^\omega \to \mathbb{N}$ be the random variable that counts the number of occurrences of the edge $((m,n)(u,v))$ in a prefix in G of the given path. Then Q' can be rewritten as

$$\sum_{\pi \in G} \mathbb{P}_{\mathcal{C}}(\pi) \Big(\ln(\iota(\pi_0)) + \sum_{m,u \in M} \sum_{n,v \in N} Z^{m,n}_{u,v}(\pi) \ln \mathcal{C}'_\lambda(m,n)(u,v) \Big) .$$

Therefore the coupling structure $\mathcal{C}'$ that maximize the above is obtained as

$$\begin{aligned} &\operatorname*{argmax}_{c} \sum_{\pi \in G} \mathbb{P}_{\mathcal{C}}(\pi) \sum_{m,u \in M} \sum_{n,v \in N} Z^{m,n}_{u,v}(\pi) \ln c^{m,n}_{u,v} && \text{(eliminating constants)} \\ &\operatorname*{argmax}_{c} \sum_{m,u \in M} \sum_{n,v \in N} \mathbf{E}[Z^{m,n}_{u,v} \mid \mathcal{C}] \cdot \ln c^{m,n}_{u,v} \end{aligned}$$

Since c has to range among coupling structures of the form $\mathcal{C}' \in \Omega(\mathcal{M}, \mathcal{N}')$ for some chain $\mathcal{N}'$ with the same states as $\mathcal{N}$ we conclude that an optimal solution of the following optimization problem describes a coupling $\mathcal{C}'$ such that $Q' \geq Q$.

$$\begin{aligned} \text{maximize} \quad & \textstyle\sum_{m,u \in M} \sum_{n,v \in N} \mathbf{E}[Z^{m,n}_{u,v} | \mathcal{C}] \cdot \ln(c^{m,n}_{u,v}) \\ \text{such that} \quad & \textstyle\sum_{v \in N} c^{m,n}_{u,v} = \tau(m)(u) && m, u \in M,\ n \in N \\ & \textstyle\sum_{u \in M} c^{m,n}_{u,v} = \theta_{n,v} && m \in M,\ n, v \in N \\ & c^{m,n}_{u,v} \geq 0 && m, u \in M,\ n, v \in N \end{aligned}$$

As above said, this implies $\beta^{\mathcal{C}'}_\lambda(\mathcal{M}, \mathcal{N}') \geq \beta^{\mathcal{C}}_\lambda(\mathcal{M}, \mathcal{N})$. ◀