Approximating Fixpoints of Approximated Functions

Paolo Baldan[1], Sebastian Gurke[2], Barbara König[2], Tommaso Padoan[3], and Florian Wittbold[2]

[1] Dipartimento di Matematica "Tullio Levi-Civita", Università di Padova
baldan@math.unipd.it
[2] Universität Duisburg-Essen
{sebastian.gurke,barbara_koenig,florian.wittbold}@uni-due.de
[3] Università di Trieste
tommaso.padoan@units.it

Abstract. Fixpoints are ubiquitous in computer science and when dealing with quantitative semantics and verification one often considers least fixpoints of (higher-dimensional) functions over the non-negative reals. We show how to approximate the least fixpoint of such functions, focusing on the case in which they are not known precisely, but represented by a sequence of approximating functions that converge to them. We concentrate on monotone and non-expansive functions, for which uniqueness of fixpoints is not guaranteed and standard fixpoint iteration schemes might get stuck at a fixpoint that is not the least. Our main contribution is the identification of an iteration scheme, a variation of Mann iteration with a dampening factor, which, under suitable conditions, is shown to guarantee convergence to the least fixpoint of the function of interest. We then argue that these results are relevant in the context of model-based reinforcement learning for Markov decision processes, showing how the proposed iteration scheme instantiates and allows us to derive convergence to the optimal expected return. More generally, we show that our results can be used to iterate to the least fixpoint almost surely for systems where the function of interest can be approximated with given probabilistic error bounds, as it happens for probabilistic systems which can be explored via sampling.

Keywords: Fixpoints, Approximation, Mann iteration, Markov Decision Processes

1 Introduction

Fixpoints are fundamental in computer science as they arise as a natural way of providing a meaning to inductive and recursive definitions. When dealing with systems or programming languages embodying quantitative aspects, such as probability, time, or cost, we are often led to consider fixpoints of functions associating states with real values. Fixpoints also occur in optimization methods

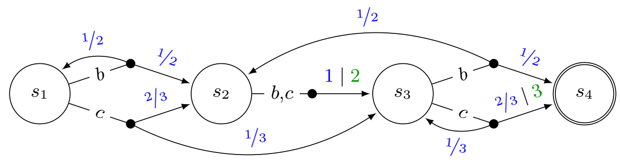


Fig. 1: A Markov decision process

where one typically determines solutions that are stable under suitable transformations. For instance, in probabilistic systems such as finite Markov decision processes (MDPs) [5], the expected payoff or the likelihood of satisfying a property can be characterized as a (least) fixpoint, from which one can then derive optimal policies.

The aim of this paper is to develop a theory of fixpoints for functions that are not completely known but can be approximated. Estimating fixpoints of unknown functions is a common task in computer science. In the aforementioned optimization setting, a standard framework is the one in which the parameters of a dynamic system of interest (e.g. probabilities, rewards, costs) are partially unknown and can only be estimated by experimenting with the system. This is exactly what happens in reinforcement learning, a branch of machine learning that is intended to provide methods for learning optimal policies for an agent in order to maximize rewards in an unknown dynamic environment.

We exemplify our motivations by focusing on MDPs and (a variant of) the Dyna-Q algorithm [41,42,27]. Consider the MDP M in Figure 1 with state set $S = \{s_1, s_2, s_3, s_4\}$. In each state $s \in S$, the agent can choose an action a from an action set $A = \{b, c\}$. Performing an action $a \in A$ in state s yields a successor state s' with a probability $T(s, a, s')$ and results in reward $R(s, a, s')$ (in the picture, arrows are labelled $T(s, a, s') \mid R(s, a, s')$; if a reward is 0, the vertical line and the number 0 are omitted). Rewards can be discounted by a factor $\gamma < 1$ so that rewards obtained in the future are valued less than immediate rewards.

A typical aim is to determine a policy for the agent which maximises the expected return. The expected return (also called payoff or total reward) $q(s, a)$ at state s taking action a can be naturally expressed by means of a recursive equation, the so-called Bellman equation

$$q(s,a) = \sum_{s' \in S} T(s,a,s')(R(s,a,s') + \gamma \cdot \max_{a' \in A} q(s',a')) \tag{1}$$

stating that $q(s, a)$ coincides with the (discounted) expected returns based on the successor states. If we denote by $g_M : \mathbb{R}^{S\times A} \to \mathbb{R}^{S\times A}$ the operator associated with the MDP, defined by $g_M(q)(s,a) = \sum_{s' \in S} T(s,a,s')(R(s,a,s') + \gamma \cdot \max_{a' \in A} q(s',a'))$, the equation reduces to $q = g_M(q)$ and the maximal expected reward is the least fixpoint of g_M (unique when $\gamma < 1$ and thus g_M contractive).

Various techniques have been devised for determining such fixpoints. However, in reinforcement learning, the MDP describes the interaction with an environment that is unknown or only partially known. In particular, the probabilities

$T(s,a,s')$ of arriving in s' after choosing action a in state s can only be estimated, e.g., by sampling. In fact, a natural approach consists in interacting with the model and record how many times one arrives in each state s', after choosing action a in s. If we denote by $N(s,a,s')$ this number, then $T(s,a,s')$ can be estimated as $N(s,a,s')/\sum_{s''} N(s,a,s'')$. Clearly, as we proceed and the number of interactions increases, we can expect to obtain better and better approximations of the MDP. For instance, the Dyna-Q algorithm starts from arbitrary values. At each step, the model is sampled and updated, and then also the values $q(s,a)$ are updated by taking a weighted sum of the previously computed value and the (γ-discounted) expectation of the return according to equation (1) (either for selected pairs or for all pairs, the variant that we consider here) according to the following schema:

$$q(s,a) := (1-\alpha)\cdot q(s,a) + \alpha \cdot \sum_{s'\in S} T(s,a,s')(R(s,a,s') + \gamma \cdot \max_{a'\in A} q(s',a')) \qquad (2)$$

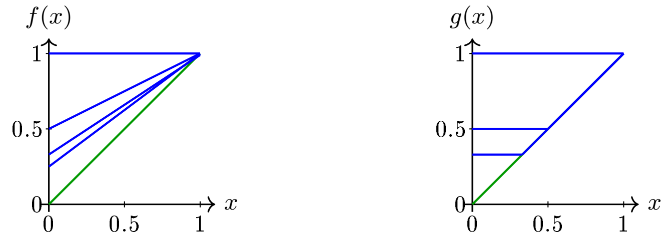
which, using function g_M, becomes $q := (1-\alpha)\cdot q + \alpha \cdot g_M(q)$. The weighted sum makes the update more conservative, with a value of α closer to 0 giving more relevance to the past knowledge. As the number of samplings increases, better approximations of the MDP are obtained, leading to better q-values.

Several other model-based algorithms use variations of this scheme, where the updates (determining the value vector and the optimal policy) and the exploration of the model are interleaved [33,27]. Model-free versions, such as Q-learning [45] or SARSA [40], do not explicitly build a model of the MDP.

The schema (2) above resembles iteration algorithms for determining the fixpoint of the function $g_M \colon \mathbb{R}^{S\times A} \to \mathbb{R}^{S\times A}$ associated to the MDP. When $\alpha = 1$, we obtain Kleene iteration which converges to the least fixpoint for monotone functions over a complete lattice, starting the iteration from the least element. In general, it corresponds to a Mann iteration [6] which, under suitable hypotheses, converges to a fixpoint of a continuous function starting from any initial state. The twist here is that the function f is not fixed, but can only be approximated, since the probabilities $T(s,a,s')$ (and possibly also the rewards $R(s,a,s')$) change and will be updated during the iteration.

Aim of the paper. Our aim is to develop a fixpoint theory of approximated functions of which the scenario above is a special case. There is a large body of work guaranteeing the existence of fixpoints for certain classes of functions and providing methods for computing them. This includes, for instance, Banach's fixpoint theorem for contractive functions over complete metric spaces [4], Knaster-Tarski theorem for monotone functions over complete lattices [43] that is frequently employed in computer science, and Kleene iteration [14].

There is much less work on how to compute (least) fixpoints for a function which is not known exactly, but can only be obtained by a sequence of subsequently better approximations. As we will see, developing such a theory is relatively simple when the functions of interest are contractions (or power contractions) whose (repeated) application decreases the distance of two values by a factor $\gamma < 1$. This is, for instance, true for reinforcement learning in the



(a) Sequence (f_n) converging to f (b) Sequence (g_n) converging to g

discounted case (cf. the proof of correctness of Q-learning [45] based on stochastic approximation theory [7]). However, in this paper, we are also and, actually, mostly interested in the non-discounted case ($\gamma = 1$) where the functions are just non-expansive (the distance of two vectors after function application is bounded by their original distance) and the aim is to determine their least fixpoints. We remark that working on the non-discounted case is sometimes the only appropriate choice, e.g., when the reward represents the likelihood of eventually satisfying a given property.

We will concentrate on functions that are monotone and non-expansive with respect to the supremum norm. Besides policy computation for MDPs, a number of other applications consider value vectors which can be characterized as (least) fixpoints of functions of this kind, e.g., computing bisimilarity metrics [1], solving (simple) stochastic games [12,28], or model-checking quantitative logics on probabilistic systems [26,32].

Least fixpoints of simple functions are not easy to compute. As a warm up, we consider some simple examples showing that computing the least fixpoint of an approximated function is non-trivial. Let $f\colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be a non-expansive and monotone function over the non-negative reals. A non-expansive function is also continuous. Moreover, if it has a fixpoint, it has a least fixpoint μf which can be computed by Kleene iteration, that is, the sequence $0, f(0), f^2(0), \ldots$ converges to μf. Now, assume that f is not known exactly, but can only be approximated via a sequence of functions f_n converging to f. One might think that Kleene iteration can be easily adapted to deal with this situation. We provide some examples showing that, instead, non-trivial problems arise.

Example 1.1 (Non-continuity of the least fixpoint operator). Take $f, f_n\colon [0,1] \to [0,1]$, $n \in \mathbb{N}$, with $f_n(x) = \frac{1}{n} + (1 - \frac{1}{n}) \cdot x$, $f(x) = x$, as depicted in Figure 2a.

The least (and only) fixpoint of each approximation f_n is 1, while the least fixpoint of f is 0. Hence, $\lim_{n\to\infty} \mu f_n = 1 \neq 0 = \mu f$, i.e., the least fixpoint operator is non-continuous. In particular, trying to obtain an approximation of the least fixpoint of f by naively performing a Kleene iteration on any fixed approximation f_n is doomed to fail, as we will always converge to 1 instead of 0.

Example 1.2 (Kleene over improving approximations fails). Take $g, g_n\colon [0,1] \to [0,1]$ with $g_n(x) = \frac{1}{n}$ if $x \leq \frac{1}{n}$, $g_n(x) = x$ otherwise, and $g(x) = x$, as depicted in Figure 2b. In this case, $\lim_{n\to\infty} \mu g_n = 0 = \mu g$, hence the sequence of least fixpoints of the approximations will converge to the least fixpoint of g. But imagine that we want to reuse intermediate results once we have obtained a better approximating function. This is what is done in reinforcement learning where one alternates sampling and iteration of the estimated value function. In this scenario, this would mean that we have already obtained an approximation of μg, based on a Kleene iteration with some g_n. This will, however, over-estimate the least fixpoint of g by $1/n$ and further iterating on the previously obtained approximation of μg with "better" functions g_m $(m > n)$ will never decrease it.

Example 1.3 (Approximations might not have fixpoints). Let $h, h_n \colon \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ where $h_n(x) = x + 1/n$ and $h(x) = x$. Here, functions h_n have no fixpoint while the least fixpoint of h is 0. We want to be able to compute μh even in such cases.

A fixpoint theory for approximated functions. As a first option, one could think of confining the attention to contractive or power contractive functions, where the problem mentioned above do not occur. This, however, would restrict applicability to probabilistic systems that have some form of stopping condition which typically produce functions of this kind (e.g. MDPs without end components) or force the use of a discount factor.

Another option would be to restart the (Kleene) fixpoint iteration from scratch for each newly obtained approximation f_n, a technique, called *resetting* in the paper, which resembles certainty equivalent methods in reinforcement learning [30]. Even when this approach is viable, it precludes the reuse of approximations obtained in a previous step, making it very inefficient and inadequate in scenarios where live value estimates are required.

We propose a solution that does neither: we allow non-expansive functions and interleave sampling and fixpoint iteration, which allows us to continue from previously computed approximations. The central idea is to combine the fixpoint iteration with a dampening factor that can be seen as a form of discount factor that "vanishes" over time. It is inspired by work in [29] which, however, does not apply to our setting since it does not deal with approximations. Furthermore, it puts stronger restrictions on the iteration parameters and the space where the functions are defined, preventing its use for the applications we have in mind.

To be more precise, given a function f over d-tuples of positive reals we propose what we call a *dampened Mann iteration*:

$$x_{n+1} = (1 - \beta_n) \cdot (\alpha_n \cdot x_n + (1 - \alpha_n) \cdot f(x_n)) \tag{3}$$

The genesis of this recursion scheme can be explained as follows: As a first step the right-hand side of a simple Kleene-iteration is replaced, as in the case of Dyna-Q, with a weighted sum $\alpha \cdot x_n + (1 - \alpha) \cdot f(x_n)$ for some parameter α. Next, we allow this parameter to be potentially different in every step, i.e. for a fixed sequence (α_n) in the unit interval we consider the sequence generated by the recursion $x_{n+1} := \alpha_n \cdot x_n + (1 - \alpha_n) \cdot f(x_n)$. Recursion schemes of this form are called Mann iterations and can in some cases be used to improve convergence properties of the generated sequence to a larger class of functions

f than only contractive functions. Indeed, one of the more elementary results on Mann iterations states that if $f\colon [a,b] \to [a,b]$ is any continuous function on a bounded interval, then choosing a parameter sequence (α_n) with $\alpha_n \to 1$ and $\sum_n (1-\alpha_n) = \infty$ guarantees that the corresponding Mann iteration always converges to some fixed point of f for any choice of initial value $x_0 \in [a,b]$ [9].

There are many results that guarantee the (weak) convergence of a Mann iteration in more general spaces under more restrictive assumptions on the function f and/or other assumptions on the parameter sequences [6,17]. The existing results in the literature are, however, not directly applicable to the problem we want to solve since the assumptions on the surrounding space are often too restrictive and, more importantly, since we specifically want to find the *least fixed point* of a function. Note that a Mann iteration that starts on some fixed point x_0 of f which is not the least will get stuck there forever. For the notion of a least fixed point to be well-defined we, on the other hand, will – unlike most literature on Mann iterations we are aware of – always assume that the map f is monotone with respect to the pointwise order.

To be able to always guarantee the convergence to the least fixed point and also to deal with approximations later on, we add – as a last step – the dampening factor $(1-\beta_n)$ to the iteration scheme.

Depending on the setting, we will require conditions on the coefficients α_n, they must either converge to 0 (Mann-Kleene scheme) or to a value strictly less than 1 (relaxed Mann-Kleene scheme). This is combined with a sequence $(1-\beta_n)$ of dampening factors such that $\lim_{n\to\infty} \beta_n = 0$ and $\sum_n \beta_n = \infty$. The first condition ensures that we dampen less and less during the iteration so that we converge to a Mann iteration, while the second condition guarantees that the dampening factors always have enough "power" left to decrease an over-approximation to the least fixpoint.

We will show that, for a fixed non-expansive and monotone function f, this form of iteration converges to the least fixpoint of f starting from every point in the domain. We will also identify sufficient conditions so that the scheme works when in (3) function f is replaced by f_n, a sequence of approximations converging to f. The resulting theory will allow to treat the case of MDPs, possibly non-discounted and with end-components.

For more general probabilistic systems which can be explored by sampling, we will show how to iterate to the least fixpoint by making sure that the functions are sampled "fast enough": using the law of large numbers and Bernstein's inequality, one can estimate the probability that the functions f_n are close enough to f after some sampling. Then, based on the Borel-Cantelli lemma, we can give an algorithm which guarantees convergence to the least fixpoint almost surely.

In summary, the main contributions of our work are:

- The definition of an iteration scheme, referred to as dampened Mann iteration, which converges to the least fixpoint of a given non-expansive and monotone function f from any starting point (Section 3).

- The identification of various sufficient conditions for applying such iteration scheme in a setting wherethe function f is unknown and can only be approximated via a sequence of functions f_n converging to it (Section 4).
- The instantiation of dampened Mann iteration to the computation of the maximal expected return for MDPs where the model is unknown and explored via sampling (Section 5).
- The instantiation of dampened Mann iteration to systems where the function of interest can be approximated with given probabilistic error bounds (this includes various probabilistic systems which can be explored via sampling) (Section 5.4).

We view this work as an important step towards extending fixpoint theory to deal with scenarios arising from machine learning and data-driven applications.

An extended version of the paper with full proofs of the results and some additional material is available as [3].

2 Preliminaries and Notation

We introduce some basic notions and notation on finite-dimensional real spaces and Markov decision processes [5], that will be our main application.

2.1 Finite-Dimensional Real Spaces

We denote by $\mathbb{R}_{\geq 0}$ the set of non-negative reals, by $\mathbb{N}_0$ and $\mathbb{N}$ the sets of naturals with and without 0, respectively. For sets X, Y, we will denote by Y^X the set of functions $f\colon X \to Y$. If X is finite, we will identify Y^X with the set of vectors $Y^{|X|}$. For $x \in \mathbb{R}^d$, we will write $(x)^i \in \mathbb{R}$ for its i-th component and, by abuse of notation, we will sometimes identify $x \in \mathbb{R}$ with the vector $(x, \dots, x) \in \mathbb{R}^d$.

We equip $\mathbb{R}^d$ with the *supremum norm* defined as $\|x\| = \sup\{|(x)^i| \mid i \in \{1, \dots, d\}\}$ for $x \in \mathbb{R}^d$ and extended to functions $f\colon A \to B$, where $A, B \subseteq \mathbb{R}^d$ by $\|f\| = \sup_{x \in A} \|f(x)\| \in [0, \infty]$. Given $x \in \mathbb{R}^d$, we denote by $|x|$ the vector in $\mathbb{R}^d_{\geq 0}$ defined by $(|x|)^i = |(x)^i|$. Note that $\|x\| = \||x|\|$.

For $x, y \in \mathbb{R}^d$, we write $x \leq y$ for the pointwise (partial) order, i.e., $(x)^i \leq (y)^i$ for all $i \in \{1, \dots, d\}$. A function $f \in Y^X$ with $X, Y \subseteq \mathbb{R}^d$ is *monotone* if, for all $x, x' \in X$, we have that $x \leq x'$ implies $f(x) \leq f(x')$. A *fixpoint* of $f\colon X \to X$ is $x \in X$ such that $f(x) = x$. The *set of fixpoints* of a function f is denoted by $\mathrm{Fix}(f)$ and, in case it exists, the *least fixpoint* of f is denoted by μf.

A function $f \in Y^X$ with $X, Y \subseteq \mathbb{R}^d$ is L-Lipschitz for a constant $L \in \mathbb{R}_{\geq 0}$ if, for all $x, x' \in X$, $\|f(x) - f(x')\| \leq L\|x - x'\|$. The function is *non-expansive* if it is 1-Lipschitz. It is *contractive* if it is q-Lipschitz for some $q < 1$, and it is a *power contraction* if there is some $n \in \mathbb{N}$ such that f^n, the n-fold composition of f, is a contraction. If $f\colon X \to X$ is a contraction (and, more generally, power contraction) with $X \subseteq \mathbb{R}^d$ closed (hence complete), by Banach's theorem, it has a unique fixpoint given by $\lim_{n \to \infty} f^n(x_0)$ where $x_0 \in X$ is arbitrary.

2.2 Markov Decision Processes

We will focus on Markov decision processes [5] in a non-discounted non-negative reward setting.

Definition 2.1 (Markov decision process). *A* Markov decision process (MDP) *is a tuple $M = (S, A, T, R)$ where S and A are the finite sets of states and actions. Moreover $T\colon S \times A \times S \to [0, 1]$ provides the probability $T(s, a, s')$ of transitioning from state $s \in S$ to state $s' \in S$ when action $a \in A$ is chosen, in a way that $\sum_{s' \in S} T(s, a, s') \in \{0, 1\}$. When $\sum_{s' \in S} T(s, a, s') = 1$ we say that action a is* enabled *in s and we write $A(s)$ for the set of actions enabled in s. Finally, $R\colon S \times A \times S \to \mathbb{R}_{\geq 0}$ is a non-negative step-wise reward function. We let $F = F(M) \coloneqq \{s \in S \mid A(s) = \emptyset\}$ and call it the set of* final states. *The Markov decision process is a Markov chain if, for all $s \in S$, we have $|A(s)| \leq 1$.*

The idea is that M describes an interactive system that, when in state s, transitions to another state s' with probability $T(s, a, s')$ based on an action $a \in A(s)$ chosen by an external agent. Possible strategies of the agent can be described by (positional) policies $\pi\colon S \backslash F \to A$ with $\pi(s) \in A(s)$ for any non-final state $s \in S \setminus F$. We denote the set of all such policies by $\mathit{\Pi}(M)$.

For MDPs, one is interested in finding a policy that optimizes the expected return. It should be noted that positional policies (as defined above) are sufficient for optimal behaviour in *finite* MDPs where the goal is to maximize the non-discounted total reward [36]. The essential step often lies in finding the (least) fixpoint of a so-called Bellman operator. In reinforcement learning, one often considers the state-action-value operator, discussed in the introduction, whose fixpoint gives the expected optimal return $q^*(s, a)$ at state s when taking action a. Alternatively, one can consider the state-value operator f_M whose fixpoint gives the expected optimal return $v^*(s)$ at state s. The two options are tightly related, e.g., $q^*(s, a) = \sum_{s' \in S} T(s, a, s') \cdot (R(s, a, s') + v^*(s'))$. Here we will focus on the state-value operator since it allows for a simpler presentation. As we are assuming a model-based scenario, approximations to the state-action values $q^*(s, a)$ can be retrieved using the above equation (with an appropriately close approximation of T).

Definition 2.2 (Bellman operator). *Given an MDP $M = (S, A, T, R)$, the (*state-value*) Bellman operator $f_M\colon \mathbb{R}^S_{\geq 0} \to \mathbb{R}^S_{\geq 0}$ is defined, for $v \in \mathbb{R}^S_{\geq 0}$, by*

$$f_M(v)(s) = \max_{a \in A(s)} \sum_{s' \in S} T(s, a, s')(R(s, a, s') + v(s')) \tag{4}$$

In contrast to the usually treated *discounted* case, the operator f_M is not necessarily contractive. Hence, it might have several fixpoints or might not have any fixpoint in $\mathbb{R}^S_{\geq 0}$, intuitively because the expected return might sum up to infinity. Hereafter, we will work with so-called *MDPs with finite value*, for which a fixpoint exists. In Section 5.3, they will be characterized as those MDPs where positive rewards are given only outside end-components. By a later result (Lemma 3.1), existence of a fixpoint for the Bellman operator implies the existence of a least fixpoint. For an MDP with finite value $M = (S, A, T, R)$, we let $v^*_M = \mu f_M$.

3 Dampened Mann Iteration for Known Functions

We define an iteration scheme for approximating, with arbitrary precision, the least fixpoint of monotone and non-expansive functions over the non-negative reals. It is a variation of Mann iteration [6] suitably modified in order to be "robust" with respect to perturbations in the computation with the introduction of a dampening factor [29,24]. In this section, we will focus on the case in which the function of interest is known and identify conditions which ensure convergence to the least fixpoint. The case in which the function can only be approximated will be discussed in the next section.

3.1 Dampened Mann Iteration Scheme

We start by clarifying the class of functions of interest in the paper.

Assumption 1. *Given a* closed *and* convex *set $X \subseteq \mathbb{R}^d_{\geq 0}$ with $0 \in X$, where $d \in \mathbb{N}_0$ is a fixed arity, let $f\colon X \to X$ be* monotone *and* non-expansive *with* $\mathrm{Fix}(f) \neq \emptyset$.

Note that, under the above assumption, if f is a (power) contraction, by Banach's theorem, it has a unique fixpoint, i.e., $\mathrm{Fix}(f) = \{\mu f\}$. Otherwise, if f is just non-expansive, the condition $\mathrm{Fix}(f) \neq \emptyset$ is non-trivial (e.g., the function $x \mapsto x + 1$ over $\mathbb{R}_{\geq 0}$ is monotone and non-expansive, but it has no fixpoints). However, when $\mathrm{Fix}(f) \neq \emptyset$, we can show that f also admits a least fixpoint μf.

Lemma 3.1 (Existence of least fixpoints). *Let $X \subseteq \mathbb{R}^d_{\geq 0}$ be a closed set with $0 \in X$ and let $f\colon X \to X$ be a monotone function with $\mathrm{Fix}(f) \neq \emptyset$. Then f has a least fixpoint μf. Moreover, if f is continuous then $\mu f = \sup_{n\in\mathbb{N}} f^n(0)$.*

Since our functions are non-expansive and thus continuous, by the lemma above, the least fixpoint can be obtained by iterating the function on 0, according to what is often called *Kleene iteration* (see, e.g. [14]).

However, as noted in the introduction, Kleene iteration is not "robust" with respect to perturbations in the computation. For this reason, we introduce a variation of a Mann iteration scheme.

Definition 3.2 (Mann scheme/iteration). *A* (dampened) Mann scheme *$\mathcal{S} = ((\alpha_n)_{n\in\mathbb{N}_0}, (\beta_n)_{n\in\mathbb{N}_0})$ is a pair of parameter sequences in $[0, 1)$ such that*

$$\lim_{n\to\infty} \beta_n = 0, \tag{5}$$

$$\sum_{n\in\mathbb{N}_0} \beta_n = \infty \quad \textit{(or equivalently, } \prod_{n\in\mathbb{N}_0} (1 - \beta_n) = 0\textit{).} \tag{6}$$

A Mann-Scheme is called a Mann-Kleene scheme *whenever $\lim_{n\to\infty} \alpha_n = 0$ and a* relaxed Mann-Kleene scheme *if $\lim_{n\to\infty} \alpha_n < 1$.*

Given a convex $X \subseteq \mathbb{R}^d_{\geq 0}$ with $0 \in X$, a Mann scheme $\mathcal{S}$ defines a sequence $(T^{\mathcal{S}}_n)_{n\in\mathbb{N}_0}$ of operators $T^{\mathcal{S}}_n\colon X^X \times X \to X$ given by

$$T^{\mathcal{S}}_n(f, x) = (1 - \beta_n) \cdot (\alpha_n x + (1 - \alpha_n) f(x))\,.$$

Together with a sequence $(f_n)_{n\in\mathbb{N}_0}$ *of functions* $f_n\colon X \to X$ *and an initial point* $x_0 \in X$*, it gives rise to a* (dampened) Mann iteration $\mathcal{F} = (\mathcal{S}, (f_n), x_0)$*, determining a sequence* $(x_n^{\mathcal{F}})_{n\in\mathbb{N}_0}$ *defined as*

$$x_0^{\mathcal{F}} = x_0, \qquad x_{n+1}^{\mathcal{F}} = T_n^{\mathcal{S}}(f_n, x_n^{\mathcal{F}}).$$

Note that the iteration can start at any point (not just 0). This might look irrelevant in cases where the function of interest is known exactly, but it will be extremely useful when it can only be approximated.

Intuitively, when trying to approximate the least fixpoint, Condition (5) ensures that dampening eventually reduces – meaning that, in the long run, when we are close to the least fixpoint of f, we stay close to it. On the other hand, Condition (6) guarantees that at any stage there is still enough "dampening power" left to correct possible over-approximations. In the exact case, when $f_n = f$ for all n, this is needed for the iteration to be convergent for *all* initial points. In the approximated case, when (f_n) converges to f, dampening will be indispensable since over-approximations at some components can be introduced along the way.

3.2 Approximating the Fixpoint of Known Functions

For an approximation scheme to be reasonable, it should at least converge to the correct solution μf when we use the exact function f at every step. Hence, we first focus on the case in which $f_n = f$ for all $n \in \mathbb{N}_0$ and identify conditions which ensure convergence. By abuse of notation, we will sometimes identify f with the sequence (f_n) with $f_n = f$ for all $n \in \mathbb{N}_0$.

Definition 3.3 (Exact Mann scheme). *Let* $\mathcal{S} = ((\alpha_n), (\beta_n))$ *be a Mann scheme. We call it* exact *if for all* $f\colon X \to X$ *as in Assumption 1 and* $x_0 \in X$*, the sequence* $(x_n^{\mathcal{F}})$ *generated by the iteration* $\mathcal{F} = (\mathcal{S}, f, x_0)$ *converges to* μf*.*

One can prove that, for functions satisfying Assumption 1, even if the domain X might not be bounded, the sequence generated by a Mann scheme is bounded. However, in general, the sequence does not converge to the least fixpoint of f (and it might not converge at all) without additional restrictions on the parameters.

We next prove that convergence of the iteration to the least fixed point is ensured by working with Mann-Kleene schemes, i.e., Mann schemes satisfying

$$\lim_{n\to\infty} \alpha_n = 0. \tag{7}$$

Intuitively, Condition (7) implies that the iteration gets closer and closer to a Kleene iteration (in fact, when $\alpha_n \to 0$, the operator $T_n^{\mathcal{S}}(f, \cdot)$ tends to f).

Canonical choices of the parameters for obtaining a Mann-Kleene scheme are $\beta_n = 1/n$ and either $\alpha_n = 1/n$ as well or $\alpha_n = 0$ constantly. In the sequel we sometimes start at x_n with $n > 0$ rather than x_0 to ensure that all parameters are well-defined, particular when they are based on fractions such as $1/n$.

Theorem 3.4 (Approximating the fixpoint of known functions). *Every Mann-Kleene scheme* $\mathcal{S}$ *is exact.*

Although so far we have only considered the case in which $f_n = f$ for every $n \in \mathbb{N}_0$, Theorem 3.4 already yields a significant improvement to the usual Kleene approximation, where one sets $x_{n+1} = f(x_n)$: The Mann-Kleene iteration with a non-expansive function converges to the least fixpoint of f for *every* starting point x_0 and not just for $x_0 = 0$.

With some additional effort one can prove convergence also for relaxed Mann-Kleene schemes where Condition (7) $\alpha_n \to 0$ is relaxed to $\alpha_n \to \alpha < 1$. This for instance allows to set $\alpha_n = \alpha \in [0, 1)$.

Corollary 3.5 (Convergence for relaxed Mann-Kleene). *Every relaxed Mann-Kleene scheme is exact.*

Theorem 3.4 and Corollary 3.5 above instantiate to our main application scenario, i.e., finding the optimal value function of an MDP M and imply that, given a (relaxed) Mann-Kleene scheme $\mathcal{S}$, the iterations $\mathcal{F} = (\mathcal{S}, f_M, v_0)$ yields a converging sequence $x_n^{\mathcal{F}} \to v_M^*$ for all initial values v_0, q_0. In fact, the domain $X = \mathbb{R}_{\geq 0}$ is closed and the Bellman operator f_M is monotone, non-expansive and admit fixpoints (since we work with MDPs with finite value).

4 Approximated Fixpoints of Approximated Functions

In this section, we are, more generally, interested in the least fixpoint of a function f under the assumption that we can get access only to a sequence (f_n) of approximations converging to f.

Assumption 2. *Given a closed and convex set $X \subseteq \mathbb{R}^d_{\geq 0}$ with $0 \in X$ where $d \in \mathbb{N}_0$ is a fixed arity, let (f_n) be a sequence of monotone and non-expansive functions $f_n \colon X \to X$, pointwise converging to $f \colon X \to X$ with $\mathrm{Fix}(f) \neq \emptyset$.*

Note that, under Assumption 2, also f is guaranteed to be monotone and non-expansive. Furthermore, a standard result from analysis ensures that, given a sequence (f_n) of L-Lipschitz functions, $f_n : X \to X$, with $X \subseteq \mathbb{R}^d$ compact, if the sequence converges to a function f, then the convergence is uniform. Therefore, if we consider the function sequence (f_n) only on a compact subset of X, we can assume w.l.o.g. that it converges uniformly. Indeed, in the sequel we will often show that the sequences generated by Mann iterations are bounded and use this fact to restrict to a bounded and thus compact domain $X \subseteq \mathbb{R}^d_{\geq 0}$.

Now, given an iteration scheme $\mathcal{S}$ that is exact, i.e., which works when iterating the exact function, a naive idea to construct a sequence of approximations (x_n) of the least fixpoint of the target function f might be to perform *resetting*, i.e., to restart the iteration in each step and calculate x_k as $T^{\mathcal{S}}_{n_k}(f_k, T^{\mathcal{S}}_{n_k-1}(f_k, \ldots T^{\mathcal{S}}_0(f_k, x_0) \ldots))$, namely approximate μf by iterating appropriately many times with an approximation f_k sufficiently close to f.

However, as already noted, the least fixpoint operator μ is not continuous for non-expansive functions, i.e., the sequence μf_n might not converge to μf (see Example 1.1). Even worse, the approximations might not even have a fixpoint (see Example 1.3). Hence, simply choosing an approximation f_n sufficiently close to f, and iterating with it, will not guarantee to get close to μf.

In the rest of this section we show that, under suitable assumptions, dampened Mann schemes, which instead avoid restarting and properly reuse previous iterates, work in the approximated setting given by Assumption 2. Here, Condition (6) on the dampening parameters (β_n) is essential – even when the iteration starts from 0 (bottom value) – as the approximations f_n constantly introduce errors, possibly causing over-approximations which need to be "dampened".

One can show that the least fixpoint μf of the target function serves as an asymptotic lower bound for all relaxed Mann-Kleene schemes.

Lemma 4.1. *Under Assumption 2, given a relaxed Mann-Kleene scheme $\mathcal{S}$ and arbitrary $x_0 \in X$, consider the iteration $\mathcal{F} = (\mathcal{S}, (f_n), x_0)$. If $(x_n^{\mathcal{F}})$ is bounded, we have $\liminf_{n\to\infty} x_n^{\mathcal{F}} \geq \mu f$.*

Moreover, restricting to the one dimensional case, i.e., for functions $f_n : X \to X$ with $X \subseteq \mathbb{R}_{\geq 0}$, the condition $\mu f = \lim_{n\to\infty} \mu f_n$, ensures convergence for every (relaxed) Mann-Kleene scheme.

Theorem 4.2 (Approximated Mann-Kleene - one dimension). *Let $f, f_n : X \to X$ be as in Assumption 2 with $d = 1$, let $\mathcal{S}$ be a (relaxed) Mann-Kleene scheme, $x_0 \in X$, and consider the iteration $\mathcal{F} = (\mathcal{S}, (f_n), x_0)$. If $\mu f = \lim_{n\to\infty} \mu f_n$ then $(x_n^{\mathcal{F}})$ converges to μf.*

In the above result the hypothesis that $\lim_{n\to\infty} \mu f_n = \mu f$ is essential, witnessed by the example below.

Example 4.3 (Mann-Kleene iteration converging to the wrong value). Consider again Example 1.1 from the introduction, with $f, f_n \colon [0,1] \to [0,1]$ where $f(x) = x$ and $f_n(x) = (1 - 1/n) \cdot x + 1/n$ for $n \in \mathbb{N}$. We choose the Mann-Kleene scheme $\alpha_n = 0, \beta_n = 1/n$, and iterate starting with $x_1 = 0$. This generates a sequence $(x_n^{\mathcal{F}})$ such that $x_{n+1}^{\mathcal{F}} = \frac{n-1}{2n}$ and thus $x_n^{\mathcal{F}} \to 1/2 \neq 0 = \mu f$.

Theorem 4.2 does not generalise to the multidimensional case. As shown by the example below, there are Mann-Kleene schemes satisfying $\lim_{n\to\infty} \mu f_n = \mu f$ where iterating with the approximations does not even converge.

Example 4.4 (Mann-Kleene iteration diverging). Let $f : [0,1]^2 \to [0,1]^2$ be the flip map given by $f(x,y) = (y,x)$ and consider a sequence of maps $f_n \colon [0,1]^2 \to [0,1]^2$ as follows:

1. If n is even let $f_n = f$ and
2. if n is odd define f_n by $f_n(x,y) = \big(y \ominus \varepsilon_n, x \oplus \varepsilon_n\big)$ where $\varepsilon_n := 2/n$ and $\ominus, \oplus$ stand for truncated subtraction/addition in the interval $[0,1]$.

Then, the maps f and f_n for $n \in \mathbb{N}$ are monotone and non-expansive. Moreover, f_n converges (uniformly) to f and since $f_n(0, \varepsilon_n) = (0, \varepsilon_n)$ we also have $\lim_{n\to\infty} \mu f_n = \mu f = (0,0)$.

However, if we choose the Mann-Kleene scheme $\mathcal{S} = ((\alpha_n), ({}^1\!/\!{}_{n+1}))$, with $\alpha_n = 0$ for all n, it is easy to check that the sequence $(x_n^{\mathcal{F}})$ generated by the iteration $\mathcal{F} = (\mathcal{S}, (f_n), (0, 1))$ (starting at index 1) is given by

$$x_n^{\mathcal{F}} = \begin{cases} (\frac{n-2}{n}, 0) & \text{if } n \text{ is odd} \\ (0, \frac{n-1}{n}) & \text{if } n \text{ is even} \end{cases}$$

for $n > 1$. Thus, both $(1, 0)$ and $(0, 1)$ are cluster points of $(x_n^{\mathcal{F}})$. In particular, the sequence $(x_n^{\mathcal{F}})$ is not even asymptotically regular, i.e. $\|x_n^{\mathcal{F}} - x_{n+1}^{\mathcal{F}}\|$ does not converge to 0.

We provide a first positive result for the multidimensional case when the limit function f is a power contraction. Power contractions naturally arise in the study of MDPs. Hence this result will play an important role in Section 5.

Theorem 4.5 (Approximated Mann-Kleene - power contractions). *Let $f, f_n : X \to X$ be as in Assumption 2 with f a power contraction. Given a Mann-Kleene scheme $\mathcal{S}$ and $x_0 \in X$, consider the iteration $\mathcal{F} = (\mathcal{S}, (f_n), x_0)$.*

Then, if $(x_n^{\mathcal{F}})$ is bounded, it converges to the unique fixpoint of f.

Next we come to one of our main results for general non-expansive functions. It provides sufficient conditions on the approximation sequence (f_n), ensuring that an exact Mann scheme converges to the least fixpoint when iterating with approximations. Note that it applies, in particular, to (relaxed) Mann-Kleene schemes which are known to be exact by Theorem 3.4 and Corollary 3.5.

Main Theorem 4.6 (Approximated Mann iteration). *Let $f, f_n : X \to X$ be as in Assumption 2, let $\mathcal{S}$ be an exact Mann scheme, $x_0 \in X$, and consider the iteration $\mathcal{F} = (\mathcal{S}, (f_n), x_0)$. Then $(x_n^{\mathcal{F}})$ converges to μf if either*

1. *the sequence (f_n) is monotone and $\mu f = \lim_{n\to\infty} \mu f_n$ or*
2. *the sequence (f_n) converges normally to f, i.e. $\sum_{n\in\mathbb{N}_0} \|f - f_n\| < \infty$.*

It is remarkable that under the assumption of normal convergence of (f_n) to f, the sequence $(x_n^{\mathcal{F}})$ converges to the correct solution even if $\lim_{n\to\infty} \mu f_n \neq \mu f$. Combining with estimates obtained in Theorem 3.4 one can even calculate error bounds ε_n (converging to 0) such that $\|x_n^{\mathcal{F}} - \mu f\| \leq \varepsilon_n$ for all $n \in \mathbb{N}_0$, provided that we can calculate such estimates for $\|f^n(0) - \mu f\|$.

Example 4.7. Consider again the sequence of functions (f_n) from Example 4.3. We observed that the Mann-Kleene iteration does not converge to the least fixpoint in this case. Indeed, (f_n) does not converge normally as $\|f_n - f\| = {}^1\!/\!{}_n$.

However, let us consider a sped up sequence of approximations as to guarantee normal convergence, e.g., put $g_n = f_{n^2}$ and thus $\|g_n - f\| = {}^1\!/\!{}_{n^2}$. If $(x_n^{\mathcal{G}})$ is the sequence generated from (g_n) by Theorem 4.6 we get $\lim_{n\to\infty} x_n^{\mathcal{G}} = \mu f = 0$.

Normal convergence intuitively expresses the fact that (f_n) converges sufficiently fast to f. In the context of a sampling-based approximation such as model-based reinforcement learning, this can be obtained by taking a sufficiently large number of samples. As we will discuss in Section 5.4, this number can be bounded using quantitative versions of the law of large numbers.

5 Value Functions of MDPs and Probabilistic Systems

5.1 MDP Sampling

We now come to our main application: approximating the optimal value functions of MDPs in a model-based reinforcement learning setting. We show that dampened Mann iteration on the Bellman operators of sampled MDPs always almost surely converges to the optimal value function for any Mann-Kleene scheme.

For an MDP $M = (S, A, T, R)$ we assume that the states, actions, and reward function are known to the agent, and that the transition function is learned through sampling in the following sense: For every state $s \in S$ and every action $a \in A(s)$, there is a family of independent discrete random variables $(X_n^{s,a})_{n\in\mathbb{N}}$ with range S defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ whose distributions are given by $\mathbb{P}[X_n^{s,a} = s'] = T(s, a, s')$.

We let T_n be the maximum likelihood approximation to T after performing n random experiments on each state-action pair, that means

$$T_n(s, a, s') := \frac{|\{i \in \{1, \dots, n\} \mid X_i^{s,a} = s'\}|}{n}.$$

Note that, in particular, we never over-estimate a probability 0, i.e.

$$T(s, a, s') = 0 \Rightarrow T_n(s, a, s') = 0 \qquad \text{for all } s, s' \in S, a \in A(s). \tag{8}$$

The n-th approximation of M is given by $M_n = (S, A, T_n, R)$ with f_n being the value iteration functions of M_n. By the law of large numbers, we then infer that $\mathbb{P}\big[\lim_{n\to\infty} T_n(s, a, s') = T(s, a, s')\big] = 1$ for all $s, s' \in S, a \in A(s)$. Hence the sequence (f_n) converges to the value iteration function f of M almost surely.

Assumption 3. *Let $M = (S, A, T, R)$ be an MDP with finite value, let (M_n) be given by a sampling process as described above with $T_n \to T$. We denote by $f_n \colon \mathbb{R}^S_{\geq 0} \to \mathbb{R}^S_{\geq 0}$ and $f \colon \mathbb{R}^S_{\geq 0} \to \mathbb{R}^S_{\geq 0}$ the Bellman operators of M_n and M, respectively (see* (1)*).*

Note that Assumption 3 implies Assumption 2. We can show that iterations based on a Mann scheme using MDP functions stay bounded. Still, Theorem 4.6 does not apply directly since approximations are obtained by a generic sampling process and thus convergence is neither monotone nor normal, in general. Still, we can conclude using a combination of the results about convergence in the "exact case" (Theorem 3.4) and for power contractions (Theorem 4.5). In order to state the result more abstractly and concisely, we give a name to Mann schemes satisfying the two mentioned properties.

Definition 5.1 (Regular Mann scheme). *A Mann scheme $\mathcal{S}$ is* regular *if*

1. *it is exact and*
2. *in the approximated setting (Assumption 2), the sequence $(x_n^{\mathcal{F}})$ generated by the iteration $\mathcal{F} = (\mathcal{S}, (f_n), x_0)$ converges to μf for all $x_0 \in X$, provided that it is bounded and f is a power contraction.*

By Theorem 3.4 and Theorem 4.5, all Mann-Kleene schemes are regular, while it is currently unknown whether all relaxed Mann-Kleene schemes are regular.

5.2 Simple MDPs

In general, the Bellman operators of MDPs are not contractions. We show that if we restrict to a suitable subclass MDPs, they are power contractions. To single out such class of MDPs, we recall the notion of (maximal) end-components [2,16].

Definition 5.2 (Maximal end-components, simple MDPs). *Let an MDP $M = (S, A, T, R)$ be given. A tuple (S', A') identifying a subprocess $M' = (S', A', T', R')$ with $T'(s, a, s') = T(s, a, s')$ for all $s, s' \in S', a \in A'(s)$ and $R' = R|_{S' \times A' \times S'}$ is called an* end-component *if the following conditions hold:*

- *$F(M') = \emptyset$,*
- *for all $s \in S', s' \in S \backslash S', a \in A'(s)$, we have $T(s, a, s') = 0$*
- *for all $s, s' \in S'$, there exists a path in M' from s to s'.*[4]

An end-component (S', A') is called maximal end-component (MEC) iff there exists no larger end-component (with respect to (pointwise) inclusion). We will write $\mathrm{MEC}(M)$ for the set of all maximal end-components of the MDP M.

We call an MDP M simple *if $\mathrm{MEC}(M) = \emptyset$.*

Intuitively, an end-component is a strongly connected subset S' of non-final states such that each state has an action giving probability zero of leaving S'.

We now show convergence of regular Mann schemes for simple MDPs. The proof proceeds by introducing a class of functions which we refer to as witness-bounded functions, and proving convergence of a regular Mann scheme for functions in this class. The result for simple MDPs follows by observing that the corresponding Bellman operators are witness-bounded. A property playing a prominent role is that the Bellman operator for simple MDPs is a power contraction, a result which is probably folklore (see e.g. [19]).

Theorem 5.3 (Convergence for simple MDPs). *Let M, M_n be MDPs as in Assumption 3. Given a regular Mann scheme $\mathcal{S}$ and an initial point $v \in \mathbb{R}^S_{\geq 0}$, consider the iteration $\mathcal{F} = (\mathcal{S}, (f_n), v)$. If M is simple, then $x_n^{\mathcal{F}} \to v_M^*$.*

5.3 General MDPs

When MDPs have end-components [2,11,22], the iteration functions are no longer power contractions, which implies that they might not admit fixpoints. As anticipated in Section 2.2, we will work under the assumption that all MDPs have finite value, i.e., the Bellman operator has a (finite) fixpoint and thus a least fixpoint. We observe here that MDPs with finite value can be naturally characterized as MDPs where no reward is given inside end-components. As observed later in Remark 5.9 this setting generalises MDPs with discounted return.

[4] Given two states $s, t \in S$, a path from s to t is a sequence $s_0 = s, s_1, \ldots, s_k = t \in S$ and $a_1, \ldots, a_k \in A$ with $a_j \in A(s_{j-1})$ and $T(s_{j-1}, a_j, s_j) > 0$ for all $1 \leq j \leq k$.

Lemma 5.4 (Characterising MDPs with finite value). *An MDP M has a finite value, i.e., its Bellman operator f_M has a (least) fixpoint, if and only if for all maximal end-components (S', A') and all $s \in S', a \in A'(s)$, and $s' \in S'$ with $T(s, a, s') > 0$, we have $R(s, a, s') = 0$.*

Still, the existence of multiple fixpoints, with the possibility of getting stuck at a fixpoint that is not the least, brings additional complications. An idea explored in the literature [11,22] is to work on an MDP obtained from the original one by quotienting each MEC into a single node. In order to deal with this quotient we need some notation. Given a function $r : S \to Q$, we define:

- *Reindexing*: $r^\bullet : \mathbb{R}^Q_{\geq 0} \to \mathbb{R}^S_{\geq 0}$ defined, for all $w \in \mathbb{R}^Q_{\geq 0}$, by $r^\bullet(w) = w \circ r$
- *Maximum*: $\max_r : \mathbb{R}^S_{\geq 0} \to \mathbb{R}^Q_{\geq 0}$ defined, for all $v \in \mathbb{R}^Q_{\geq 0}$ and $e \in Q$ by $\max_r(v)(e) = \max\{v(s) \mid s \in S \wedge r(s) = e\}$.

When dealing with quotients, it is convenient to work with MDPs $M = (S, A, T, R)$ where the action sets for different states are disjoint, i.e., for $s, s' \in S$ if $s \neq s'$ then $A(s) \cap A(s') = \emptyset$. It is clear that any MDP can be transformed into an equivalent MDP with disjoint action sets simply by relabelling the actions, without changing the value or Bellman function, whence we will, from now on, w.l.o.g., work under the disjointness assumption.

Definition 5.5 (Quotient MDP). *Let $M = (S, A, T, R)$ be an MDP with finite value. Let $\equiv$ denote the equivalence identifying states in the same MEC. Consider $r : S \to S/_\equiv$ defined by mapping each state to its equivalence class: $r(s) = [s]$. The quotient MDP is $M_r = (S_r, A_r, T_r, R_r)$ where $S_r = S/\equiv$, $A_r([s]) = \bigcup_{s' \in [s]} A(s')$. Moreover, for $s, s' \in S$, $a \in A(s)$:*

$$T_r([s], a, [s'])) = \sum_{s'' \in [s']} T(s, a, s'') \qquad R_r([s], a, [s']) = \frac{\sum_{s'' \in [s']} T(s, a, s')R(s, a, s')}{T_r([s], a, [s'])}.$$

The reduced quotient *$\hat{M}_r$ is defined as the MDP derived from M_r by removing self-loops, i.e., modifying the definition of the set of enabled actions as follows:*

$$\hat{A}_r([s]) = A_r([s]) \setminus \{a \in A_r([s]) \mid T_r([s], a, [s]) = 1\}.$$

and adapting $\hat{T}_r$ accordingly.

The quotient is an MDP with only singleton end-components, i.e., states with actions producing self-loops. Since the reduced quotient is obtained from the quotient by removing self-loops, it clearly has no end-components, allowing to reuse results from Section 5.2.

The following proposition restates results from [11,22] on collapsing MECs.

Proposition 5.6 (Fixpoints of non-discounted MDPs). *Let M be an MDP with finite value. Let $\hat{M}_r$ be the reduced quotient. Then $v^*_M = r^\bullet(v^*_{\hat{M}_r})$.*

We show that for the successive approximation of (the value iteration function of) an MDP by sampling as described above, the least fixpoints of the approximations always converge to the least fixpoint of the underlying MDP.

Theorem 5.7. *Under Assumption 3, we have* $\lim_{n\to\infty} \mu f_n = \mu f$.

Now, we can show that, also for general MDPs, a dampened Mann iteration based on a regular Mann scheme converges to the optimal value function.

Main Theorem 5.8 (Mann iteration over MDPs). *Let* M, M_n *be MDPs as in Assumption 3. Given a regular Mann scheme* $\mathcal{S}$ *and any* $v \in \mathbb{R}^S_{\geq 0}$, *consider the iteration* $\mathcal{F}^{\mathrm{a}} = (\mathcal{S}, (f_n), v)$, *generating a sequence* $(v_n^{\mathcal{F}^{\mathrm{a}}})$. *Then* $\lim_{n\to\infty} v_n^{\mathcal{F}^{\mathrm{a}}} = v_M^* = \mu f_M$.

The proof first shows that μf_M serves as a lower bound to the limit of $(v_n^{\mathcal{F}^{\mathrm{a}}})$, making use of the fact that the expected return for an MDP is the maximum of the expected return of the Markov chains over all possible policies. As, furthermore, any state in an end-component of a Markov chain has expected return 0, the iteration generated by their reduced quotients can serve as a lower bound to the iteration of interest, which together with the above observation yields μf as a lower bound. The crux of the proof is to show that μf also serves as an upper bound. Roughly the idea is to show that iterating on the quotient MDPs generates a sequence (v_n) over-approximating the actual sequence $(v_n^{\mathcal{F}^{\mathrm{a}}})$. Using the fact that reduced quotients are simple MDPs one can build on the results of Section 5.2 to deduce that $\lim_{n\to\infty} v_n^{\mathcal{F}^{\mathrm{a}}}$ is below μf_M.

Remark 5.9. In the context of (model-based) reinforcement learning, usually the discounted total reward criterion is used, that is the aim of the agent is to optimize the discounted total reward for some discount factor $\gamma < 1$, as described in Section 1. This setting yields some desirable properties, mainly that the resulting discounted Bellman operator is contractive, making the optimal value its unique fixpoint and always well-defined. It can also be motivated by interpreting it as a scenario in which the agent prefers an immediate reward over long-term rewards or where the system terminates at any step with probability $1-\gamma$.

However, we remark that the discounted setting can be interpreted as a special case of the non-discounted one by adding a sink state which is reached by every state with probability $1-\gamma$ (akin to its interpretation as a system that stops with probability $1-\gamma$ in each step). Then the resulting MDP is not only simple in the sense of Definition 5.2 but its Bellman operators are, with slight modifications, also contractions whence we can apply Theorem 5.8 to derive convergence results. In particular, this gives a convergence result of simple versions of model-based algorithms such as the Dyna-Q algorithm [41,42] where we consider only full-sweep updates and no direct update step.

5.4 Reaching the Least Fixpoint by Fast Iteration

In the previous section we proved a convergence result for MDPs, approximated by sampling. It would be desirable to extend such results to more general settings involving probabilistic systems which are explored by sampling. Indeed, our results allow the formulation of a generic algorithm that can compute a sequence of values converging to the least fixpoint of f almost surely. This can be

Algorithm 1: Dampened Mann Iteration with Probabilistic Guarantees

1 **Input:** A map $n \mapsto f_n$ such that (f_n) uniformly converges to f almost surely. Parameters $\gamma_i, \delta_i \in \mathbb{R}_{\geq 0}$ such that $\gamma_i > 0$ and $\sum_{i \in \mathbb{N}} \gamma_i < \infty$ (analogously for δ_i)
2 **Output:** A sequence (x_i) in $\mathbb{R}^d_{\geq 0}$
3 $x_0 := 0$;
4 $i := 0$;
5 **while** *true* **do**
6   Choose n_i such that $\mathbb{P}\big[\|f_{n_i} - f\| \geq \gamma_i\big] \leq \delta_i$;
7     $x_{i+1} := (1 - \beta_i) \cdot \big(\alpha_i \cdot x_i + (1 - \alpha_i) \cdot f_{n_i}(x_i)\big)$;
8     $i := i + 1$;
9 **end**

done in all cases where we can estimate the error $\|f_n - f\|$ at least with a certain probability. The procedure, reported in Algorithm 1, relies on a very simple idea: make sure algorithmically that the sequence (f_n) converges normally to f almost surely. The correctness of the procedure follows by the Borel-Cantelli Lemma [8].

Theorem 5.10. *For exact Mann schemes the sequence (x_i) produced by Algorithm 1 converges to the least fixpoint of f almost surely.*

The only precondition for the algorithm is that the indices n_i in line 6 can be chosen effectively. When the sequence (f_n) is generated by a sampling process as described above, this can often be done by using Bernstein's inequality [25] – a quantitative version of the law of large numbers. If we sample the transition probabilities $T_n(s, a, s')$ as detailed in Section 5.1, we can estimate the error as below, which allows us to computationally determine the indices needed in line 4 of the algorithm:

$$\mathbb{P}\big[|T_n(s, a, s') - T(s, a, s')| \geq \varepsilon\big] \leq 2e^{-2\varepsilon^2 n}$$

As an example, the described technique can be used to deal with (simple) stochastic games (This is fully detailed in the extended version of the paper [3]).

6 Conclusion

Related Work. There has been a significant amount of research on convergence properties of Mann iterations of non-expansive maps. However, most of the existing literature deals with the undampened case and only a single function f, i.e., with a scheme defined by $x_{n+1} = \alpha_n x_n + (1 - \alpha_n) f(x_n)$.

It is known [10] that in every Banach space and for every non-expansive f with at least one fixpoint this scheme produces a sequence which is f-asymptotically regular (i.e. $\|x_n - f(x_n)\| \to 0$) for all parameter sequences (α_n) with $\sum(1 - \alpha_n) = \infty$. In general the sequence does not need to converge. (Weak) convergence is shown under additional assumptions on the map f and/or the surrounding space in various places (see for example [38,6,17]).

Dampened versions similar to ours have been considered in [29] where the authors show convergence to a fixpoint under strong additional assumptions on the parameter sequences in a uniformly smooth Banach space and in [46] where convergence is shown for more general parameter sequences in Hilbert spaces.

Here, in addition, we deal with the case where the function f can only be approximated. Motivated by the applicability to MDPs, we consider the case of monotone non-expansive functions in finite-dimensional Banach spaces with the supremum norm, which are neither uniformly smooth nor Hilbert spaces. They are not even uniformly convex as is assumed e.g. in [38]. We also introduced a probabilistic setting and devised a generic algorithm, which applies in many situations and guarantees almost sure convergence to the least fixpoint of f.

As mentioned earlier, our work is inspired by the integration of reinforcement learning algorithms with fixpoint theory [45,33,40,30], and in particular by the Dyna-Q algorithm [41,42,27]. Reinforcement learning techniques are by now a central ingredient in machine learning and we plan to strengthen the foundations behind this method.

There is some similarity of our results to the theory of stochastic approximation [7,18,31], going back to [39], which has also been employed in the correctness proof of Q-learning [45]. Stochastic approximation deals with a root finding problem where the function contains an error term with expected value zero. In that line of work, the function itself is stochastic with known expected value, while we separate the (approximated) fixpoint theory from the probabilistic setting by checking that we converge almost surely via sampling. More importantly, the focus in stochastic approximation is to consider either contractions (with unique fixpoints) or convergence to *some* fixpoint, while the case of convergence to the *least* fixpoint is not treated. Hence there is no machinery for leaving a fixpoint once it is reached, such as we have by employing a dampening factor.

In order to obtain our results on MDPs, the notion of end-components is quite fundamental. For an introduction to end-components see [2], while the connection of non-uniqueness of fixpoints of MDP functions to end-components was exploited in [11,22] in order to compute the correct expected return by collapsing maximal end components. In [11] the authors assume – as we do – that MDPs are not known exactly but are sampled. They present a different solution by collapsing maximal end-components, which is unnecessary in our approach.

Numerical Experiments. We made several numerical experiments that are reported in the extended version [3]. For unknown functions the outcome can be summarized as follows: in terms of precision, the results are comparable to "resetting", where – for some n – we perform the iteration from 0 with the current approximation f_n. Compared to our suggested form of iteration, resetting has the disadvantage that the results vary highly with the quality of the current approximation and that intermediate results are not available.

Future Work. For future work we are in particular interested in the case of parameter sequences where α_n converges to one and the question is whether we can still guarantee convergence to μf. This would be interesting, since such parameters are for instance used in the context of Q-learning [45]. Our long-term aim is to devise a fixpoint theory of approximated functions for which a large number of reinforcement learning algorithms are special cases. Here we concentrated on model-based reinforcement learning, for model-free reinforcement learning we need a theory that allows updates with a sampled value rather than with the current best approximation and we plan to work with the weaker assumption that the limit-average of the functions f_n converges to f (rather than the functions themselves). This has some overlap to stochastic approximation theory as discussed above, even though we do not consider stochastic functions. Hence we plan to partially build on the results obtained in that area.

To properly match reinforcement learning algorithms we will look into chaotic and asynchronous iteration [13,21] where one can iterate at different speeds at various states. This is useful when traversing an MDP or stochastic system, making updates on-the-fly and locally when better estimates are obtained.

We will also investigate how to iterate to the greatest fixpoint rather than the least, either by dualizing or adapting ideas from [29,24] on iterating to the fixpoint closest to a given value. Including rewards from different domains, such as negative rewards, is also a direction of future work.

We plan to identify more cases where approximated fixpoint iteration works out of the box. For instance, we mentioned that simple stochastic games are covered by the results in Section 5.4, but there we needed a way to enforce fast convergence. We will investigate whether the case of simple stochastic games (and related applications) works with the standard iteration scheme. Due to Example 4.4 it is clear that this will not be true for all approximated functions, but we plan to find requirements weaker than the current ones.

Moreover we will further look into error estimates, making guarantees of the form "with probability p the error is below ε". In the case of MDPs, even when they are known exactly, determining upper/lower bounds and thus error estimates already requires some non-trivial techniques, such as the results in [22,37,23]. It is our aim to generalize these results to the case where the probabilities associated to the MDP are not known exactly and, in this case, we expect to obtain probabilistic error estimates.

Disclosure of interests. The authors have no competing interests to declare that are relevant to the content of this article.

Acknoledgements. This work is partially supported by the European Union – NextGenerationEU under the National Recovery and Resilience Plan (NRRP) - Call PRIN 2022 PNRR - Project P2022HXNSC "Resource Awareness in Programming: Algebra, Rewriting, and Analysis"

References

1. Giorgio Bacci, Giovanni Bacci, Kim G. Larsen, and Radu Mardare. On-the-fly exact computation of bisimilarity distances. *Logical Methods in Computer Science*, 13(2:13):1–25, 2017.
2. Christel Baier and Joost-Pieter Katoen. *Principles of Model Checking*. MIT Press, 2008.
3. Paolo Baldan, Sebastian Gurke, Barbara König, Tommaso Padoan, and Florian Wittbold. Approximating fixpoints of approximated functions. *CoRR*, abs/2501.08950, 2025.
4. Stefan Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fundamenta Mathematicae*, 3(1):133–181, 1922.
5. Richard Bellman. A Markovian decision process. *Journal of Mathematics and Mechanics*, 6(5):679–684, 1957.
6. Vasile Berinde. *Iterative Approximation of Fixed Points*, volume 1912 of *Lecture Notes in Mathematics*. Springer, 2007.
7. Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, Massachusetts, 1996.
8. Patrick Billingsley. *Probability and measure*. Wiley series in probability and mathematical statistics. Wiley, New York [u.a, 3. ed. edition, 1995.
9. David Borwein and Jonathan Borwein. Fixed point iterations for real functions. *Journal of Mathematical Analysis and Applications*, 157(1):112–126, 1991.
10. Jonathan Borwein, Simeon Reich, and Itai Shafrir. Krasnoselski-Mann Iterations in normed spaces. *Canadian Mathematical Bulletin*, 35(1):21–28, 1992.
11. Tomás Brázdil, Krishnendu Chatterjee, Martin Chmelik, Vojtech Forejt, Jan Kretínský, Marta Z. Kwiatkowska, David Parker, and Mateusz Ujma. Verification of Markov decision processes using learning algorithms. In *Proc. of ATVA '14*, pages 98–114. Springer, 2014. LNCS 8837.
12. Anne Condon. The complexity of stochastic games. *Information and Computation*, 96(2):203–224, 1992.
13. Patrick Cousot. Asynchronous iterative methods for solving a fixed point system of monotone equations in a complete lattice. Research report R.R. 88, Laboratoire IMAG, Université scientifique et médicale de Grenoble, Grenoble, France, September 1977.
14. Radhia Cousot and Patrick Cousot. Constructive versions of Tarski's fixed point theorems. *Pacific Journal of Mathematics*, 82(1):43–57, 1979.
15. Brian A. Davey and Hilary A. Priestley. *Introduction to Lattices and Order, Second Edition*. Cambridge University Press, 2002.
16. Luca de Alfaro. *Formal Verification of Probabilistic Systems*. PhD thesis, Stanford University, 1997.
17. Qiao-Li Dong, Yeol Je Cho, Songnian He, Panos M. Pardalos, and Themistocles M. Rassias. *The Krasnosel'skiĭ-Mann Iterative Method Recent Progress and Applications*. SpringerBriefs in Optimization. Springer International Publishing, Cham, 1st ed. 2022 edition, 2022.
18. Marie Duflo. *Random Iterative Models*. Number 34 in Applications of Mathematics – Stochastic Modelling and Applied Probability. Springer, 1991. Translated by Stephen S. Wilson.
19. J.H. Eaton and L.A. Zadeh. Optimal pursuit strategies in discrete-state probabilistic systems. *Journal of Basic Engineering*, 84(1):23–29, 1962.
20. David A. Freedman. *Markov Chains*. Springer NY, 1983.

21. Andreas Frommer and Daniel B. Szyld. On asynchronous iterations. *Journal of Computational and Applied Mathematics*, 123(1):201–216, 2000. Numerical Analysis 2000. Vol. III: Linear Algebra.
22. Serge Haddad and Benjamin Monmege. Interval iteration algorithm for MDPs and IMDPs. *Theoretical Computer Science*, 735:111–131, 2018.
23. Arnd Hartmanns and Benjamin Lucien Kaminski. Optimistic value iteration. In *Proc. of CAV '20 (2)*, LNCS 12225, pages 488–511. Springer, 2020.
24. Songnian He and Wenlong Zhu. A modified Mann iteration by boundary point method for finding minimum-norm fixed point of nonexpansive mappings. *Abstract and Applied Analysis*, 2013, 2013. Article ID 768595.
25. Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.
26. Michael Huth and Marta Kwiatkowska. Quantitative analysis and model checking. In *Proc. of LICS '97*. IEEE, 1997.
27. Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.
28. Edon Kelmendi, Julia Krämer, Jan Křetínský, and Maximilian Weininger. Value iteration for simple stochastic games: Stopping criterion and learning algorithm. In *Proc. of CAV '18*, pages 623–642. Springer, 2018. LNCS 10981.
29. Tae-Hwa Kim and Hong-Kun Xu. Strong convergence of modified Mann iterations. *Nonlinear Analysis: Theory, Methods & Applications*, 61(1):51–60, 2005.
30. P. R. Kumar and Pravin Varaiya. *Stochastic Systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2015.
31. Harold J. Kushner and Dean S. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Number 26 in Applied Mathematical Sciences. Springer, 1978.
32. Matteo Mio and Alex Simpson. Łukasiewicz μ-calculus. *Fundamenta Informaticae*, 150(3–4):317–346, 2017.
33. Andrew W. Moore and Christopher G. Atkeson. Prioritized sweeping: Reinforcement learning with less data and less time. *Machine Learning*, 13:103–130, 1990.
34. Sam B. Nadler, Jr. Sequences of contractions and fixed points. *Pacific Journal of Mathematics*, 27:579–585, 1968.
35. D. Pataraia. A constructive proof of Tarski's fixed-point theorem for dcpo's. Presented at the 65th Peripatetic Seminar on Sheaves and Logic, Aarhus, Denmark, 1997.
36. Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Statistics. John Wiley and Sons, 1994.
37. Tim Quatmann and Joost-Pieter Katoen. Sound value iteration. In *Proc. of CAV '18 (1)*, LNCS 10981, pages 643–661. Springer, 2018.
38. Simeon Reich. Weak convergence theorems for nonexpansive mappings in banach spaces. *Journal of Mathematical Analysis and Applications*, 67(2):274–276, 1979.
39. H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.
40. Gavin A Rummery and Mahesan Niranjan. *On-line Q-learning using connectionist systems*, volume 37. University of Cambridge, Department of Engineering Cambridge, UK, 1994.
41. Richard S. Sutton. Dyna, an integrated architecture for learning, planning, and reacting. *SIGART Bulletin*, 2(4):160–163, 1991.

42. Richard S. Sutton. Planning by incremental dynamic programming. In *Proc. of ML '91 (Conference on Machine Learning)*, pages 353–357. Morgan Kaufmann, 1991.
43. Alfred Tarski. A lattice-theoretical fixpoint theorem and its applications. *Pacific Journal of Mathematics*, 5:285–309, 1955.
44. Wolfgang Walter. A note on contraction. *SIAM Review*, 18(1):107–111, 1976.
45. Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8:279–292, 1992.
46. Yonghong Yao, Haiyun Zhou, and Yeong-Cheng Liou. Strong convergence of a modified Krasnoselski-Mann iterative algorithm for non-expansive mappings. *Journal of Applied Mathematics and Computing*, 29:383–389, 2008.

A Proofs and Additional Material for §2 (Preliminaries and Notation)

For a fixed policy $\pi \in \Pi(M)$, the MDP can be interpreted as a Markov chain by removing all actions not prescribed by the policy. We denote this Markov chain by M^π, the corresponding state-value iteration functions $f_M^\pi = f_{M^\pi}$, and its least fixpoints by $v_M^\pi = \mu f_M^\pi$. Then $f_M(v)(s) = \max_{\pi \in \Pi(M)} f_M^\pi(v)(s)$ and, by existence of memoryless optimal policies, $v_M^*(s) = \max_{\pi \in \Pi(M)} v_M^\pi(s)$.

B Proofs and Additional Material for §3 (Dampened Mann Iteration for Known Functions)

Lemma 3.1 (Existence of least fixpoints). *Let $X \subseteq \mathbb{R}^d_{\geq 0}$ be a closed set with $0 \in X$ and let $f\colon X \to X$ be a monotone function with $\mathrm{Fix}(f) \neq \emptyset$. Then f has a least fixpoint μf. Moreover, if f is continuous then $\mu f = \sup_{n\in\mathbb{N}} f^n(0)$.*

Proof. Let $Y = \{x \in X \mid \forall z \in \mathrm{Fix}(f) : x \leq z\}$, i.e., Y is the set of elements of X below all fixpoints of f.

Observe that Y is a pointed dcpo. In fact, it has a bottom element $0 \in Y$. Moreover, given any directed set $D \subseteq Y$, its supremum $d = \sup D$, obtained as the pointwise supremum, is in Y. In fact, for all components $i = \{1, \ldots, d\}$, we can consider a sequence $(x_{i,n})$ such that $\sup_n (x_{i,n})^i = (d)^i$. Using the fact that D is directed we can obtain a sequence (y_n) in D such that $x_{n,i} \leq y_n$ for all $i \in \{1, \ldots, d\}$. Then $\sup_n y_n = d$ and thus $d \in X$ by closedness. Moreover, since each $y_n \in Y$, for all $z \in \mathrm{Fix}(f)$, we have $y_n \leq z$ and thus $d = \sup_n y_n \leq z$. Therefore $d \in Y$, as desired.

Additionally observe that for all $x \in Y$ we have $f(x) \in Y$. In fact, take $x \in Y$. For all $z \in \mathrm{Fix}(f)$ it holds $x \leq z$ and thus, by monotonicity $f(x) \leq f(z) \leq z$. Hence $f(x) \in Y$.

By the above observation, we can consider the restriction of f to Y, denoted $f\big|_Y\colon Y \to Y$. We have that $f\big|_Y$ is a monotone function over a dcpo. By Pataraia's theorem [35,15], $f\big|_Y$ has a least fixpoint, which is clearly a fixpoint of f and it is the least by definition of Y.

The last part follows immediately by the fact that whenever f and thus $f\big|_Y$ is continuous $\mu f = \mu f\big|_Y = \sup_{n\in\mathbb{N}} f\big|_Y{}^n(0) = \sup_{n\in\mathbb{N}} f^n(0)$. $\square$

Conditions (5) and (6) are indeed necessary as shown by the examples below:

– for $f\colon [0,1] \to [0,1]$, $x \mapsto 1$ and $x_0 \in X$ arbitrary, we have

$$x_{n+1}^{\mathcal{F}} = (1-\beta_n)(\alpha_n x_n^{\mathcal{F}} + (1-\alpha_n) f(x_n^{\mathcal{F}})) = (1-\beta_n)(\alpha_n x_n^{\mathcal{F}} + 1 - \alpha_n) \leq (1-\beta_n).$$

Hence, if $x_n^{\mathcal{F}} \to 1 = \mu f$, then also $\beta_n \to 0$.

– for $f = \mathrm{id}\colon [0,1] \to [0,1]$, $x \mapsto x$ and $x_0 > 0$ arbitrary, we have

$$x_{n+1}^{\mathcal{F}} = (1-\beta_n)(\alpha_n x_n^{\mathcal{F}} + (1-\alpha_n) x_n^{\mathcal{F}}) = (1-\beta_n) x_n^{\mathcal{F}} = x_0 \prod_{k=0}^{n} (1-\beta_k).$$

Hence, if $x_n^{\mathcal{F}} \to 0 = \mu f$, then necessarily also $\prod_{n \in \mathbb{N}_0}(1-\beta_n) = 0$.

Lemma B.1. *Under Assumption 1, for any Mann scheme $\mathcal{S} = ((\alpha_n),(\beta_n))$ and $x_0 \in X$, the sequence $(x_n^{\mathcal{F}})$ generated by the iteration $\mathcal{F} = (\mathcal{S}, f, x_0)$ is bounded. In fact, for all $\bar{x} \in \mathrm{Fix}(f)$ and $n \in \mathbb{N}$*

$$\|x_n^{\mathcal{F}} - \bar{x}\| \leq \max(\|\bar{x}\|, \|x_0 - \bar{x}\|)$$

Proof. Let $\bar{x} \in \mathrm{Fix}(f)$. Then, for all $n \in \mathbb{N}_0$, we have

$$\begin{aligned}\|x_{n+1}^{\mathcal{F}} - \bar{x}\| &\leq \beta_n\|\bar{x}\| + (1-\beta_n)\|\alpha_n(x_n^{\mathcal{F}} - \bar{x}) + (1-\alpha_n)(f(x_n^{\mathcal{F}}) - \bar{x})\| \\ &\leq \beta_n\|\bar{x}\| + (1-\beta_n)\|x_n^{\mathcal{F}} - \bar{x}\| \\ &\leq \max(\|\bar{x}\|, \|x_n^{\mathcal{F}} - \bar{x}\|).\end{aligned}$$

Inductively, we immediately conclude. □

Theorem 3.4 (Approximating the fixpoint of known functions). *Every Mann-Kleene scheme $\mathcal{S}$ is exact.*

Proof. let $\mathcal{S}$ be a Mann-Kleene scheme. Let f as in Assumption 1 and let $x_0 \in X$. Consider the iteration $\mathcal{F} = (\mathcal{S}, f, x_0)$ and let $(x_n^{\mathcal{F}})$ the sequence generated.

Assume without loss of generality that $\alpha_n = 0$ for all n. Define a map $s\colon \mathbb{N} \to \mathbb{R}$ by

$$s(n) := \|f^n(0) - \mu f\|$$

By Kleene's Theorem the map s is non-increasing and $\lim_{n\to\infty} s(n) = 0$. We already know by Lemma B.1 that the sequence $(x_n^{\mathcal{F}})$ is norm bounded, that means we can choose M such that $\|\mu f\| + \|x_n^{\mathcal{F}} - \mu f\| \leq M$ for all $n \in \mathbb{N}_0$. Define for all $n \in \mathbb{N}$ a number δ_n as

$$\delta_n := \min\left\{ s(k) + M \sum_{i=m}^{m+k-1} \beta_i \mid n = m+k \right\}$$

It is easy to see that $\delta_n \to 0$. We claim that $x_n^{\mathcal{F}} \geq \mu f - \delta_n$. Fix a partition $n = m+k$ and note that $f^k(x_m^{\mathcal{F}}) \geq f^k(0)$ and therefore

$$x_{m+k}^{\mathcal{F}} \geq f^k(0) - M \sum_{i=m}^{m+k-1} \beta_i \geq \mu f - \left(s(k) + M \sum_{i=m}^{m+k-1} \beta_i \right)$$

For an upper bound define $\eta_n := M \cdot \prod_{i=0}^{n-1}(1-\beta_i)$. By assumption on the sequence (β_n) we know that $\lim_{n\to\infty} \eta_n = 0$. Let (y_n) be the dampened Mann Iteration with the same parameters but starting from 0, then $y_n \leq \mu f$ for all $n \in \mathbb{N}$ and by induction we see that

$$\|x_n^{\mathcal{F}} - y_n\| \leq \|x_0 - y_0\| \cdot \prod_{i=0}^{n-1}(1-\beta_i) \leq M \cdot \prod_{i=0}^{n-1}(1-\beta_i)$$

In particular $x_n^{\mathcal{F}} \leq \mu f + \eta_n$. Putting $\varepsilon_n := \max\{\delta_n, \eta_n\}$ we obtain

$$\|x_n^{\mathcal{F}} - \mu f\| \leq \varepsilon_n$$

with $\lim_{n\to\infty} \varepsilon_n = 0$. The case where (α_n) only converges to 0 is analogous (see the proof of Lemma 4.1). □

Corollary 3.5 (Convergence for relaxed Mann-Kleene). *Every relaxed Mann-Kleene scheme is exact.*

Proof. Let $\mathcal{S}$ be a relaxed Mann-Kleene scheme and let $x_0 \in X$. Consider the sequence $(x_n^{\mathcal{F}})$ determined by the iteration $\mathcal{F} = (\mathcal{S}, f, x_0)$. For showing that it converges to μf, we can use the upper bound given in the proof of Theorem 3.4 and, for the lower bound, we use the more general Lemma 4.1. □

C Proofs and Additional Material for §4 (Approximated Fixpoints of Approximated Functions)

Lemma 4.1. *Under Assumption 2, given a relaxed Mann-Kleene scheme $\mathcal{S}$ and arbitrary $x_0 \in X$, consider the iteration $\mathcal{F} = (\mathcal{S}, (f_n), x_0)$. If $(x_n^{\mathcal{F}})$ is bounded, we have $\liminf_{n\to\infty} x_n^{\mathcal{F}} \geq \mu f$.*

Proof. For the sake of readability, we write $x_n \coloneqq x_n^{\mathcal{F}}$. Since (x_n) is bounded, we can w.l.o.g. assume that (f_n) is uniformly converging.

We define $M \coloneqq \sup_{n\in\mathbb{N}_0} \|x_n - \mu f\| + \|\mu f\|$ ($< \infty$ by assumption). Furthermore, we define $\alpha'_n \coloneqq \frac{\alpha_n - \alpha}{1-\alpha}$ (with $\alpha = \lim_{n\to\infty} \alpha_n < 1$) and $f'_n = (1-\alpha) f_n + \alpha$ so that

$$x_{n+1} = (1-\beta_n)(\alpha_n x_n + (1-\alpha_n) f_n(x_n)) = (1-\beta_n)(\alpha'_n x_n + (1-\alpha'_n) f'_n(x_n)).$$

Note that $\alpha'_n \in (-\infty, 1], \alpha'_n \to 0$, and all f'_n are monotone, non-expansive, and converge to $f' = (1-\alpha) f + \alpha$.

We write $s_n = \|(f')^n(0) - \mu f\|$ and, for $k, m \in \mathbb{N}_0$,

$$t_m^k \coloneqq \sum_{i=m}^{m+k-1} \Big(\prod_{\ell=i+1}^{m+k-1} |1-\alpha'_\ell| \Big)(|1-\alpha'_i| \|f'_i - f'\| + M(2|\alpha'_i| + \beta_i)).$$

For $n \in \mathbb{N}_0$, we then define

$$\delta_n \coloneqq \min\{s_k + t_m^k \mid n = m + k\}.$$

We note that, by Kleene's theorem and $\mu f = \mu f'$, we have $s_k \to 0$ and, for each $k \in \mathbb{N}_0$, also $t_m^k \to 0$ since $\alpha'_i \to 0, \|f'_i - f'\| \to 0$, and $\beta_i \to 0$. Thus, it is clear that also $\delta_n \to 0$.

Now, for $n + 1 = m + k$, we have

$$\|x_{n+1} - (f')^k(x_m)\|$$

$$
\begin{aligned}
&\leq (1-\beta_n)\|\alpha'_n(x_n-(f')^k(x_m))+(1-\alpha'_n)(f'_n(x_n)-(f')^k(x_m))\| \\
&\qquad +\beta_n\|(f')^k(x_m)\| \\
&\leq |1-\alpha'_n|\|f'_n(x_n)-(f')^k(x_m)\|+|\alpha'_n|\|x_n-(f')^k(x_m)\|+\beta_n M \\
&\leq |1-\alpha'_n|(\|x_n-f'^{k-1}(x_m)\|+\|f'_n-f'\|)+M(2|\alpha'_n|+\beta_n) \\
&\leq |1-\alpha'_n|\|x_n-f'^{k-1}(x_m)\|+|1-\alpha'_n|\|f'_n-f'\|+M(2|\alpha'_n|+\beta_n) \\
&\leq \sum_{i=m}^{n}(\prod_{\ell=i+1}^{n}|1-\alpha_\ell|)(|1-\alpha'_i|\|f'_i-f'\|+M(2|\alpha'_i|+\beta_i)) \\
&= t^k_m
\end{aligned}
$$

In particular, since $f'^k(x_m)\geq f'^k(0)$ (f' is monotone), we have

$$x_{m+k}\geq f'^k(0)-t^k_m\geq \mu f-\delta_{m+k}.$$

□

From this, the following corollary immediately follows.

Corollary C.1. *Under Assumption 2, given a relaxed Mann-Kleene scheme $\mathcal{S}$ and arbitrary $x_0\in X$, we consider the iteration $\mathcal{F}=(\mathcal{S},(f_n),x_0)$. Then, we have $\lim_{n\to\infty}x^{\mathcal{F}}_n=\mu f$ if and only if $\limsup_{n\to\infty}x^{\mathcal{F}}_n\leq\mu f$.*

Theorem 4.2 (Approximated Mann-Kleene - one dimension). *Let $f,f_n: X\to X$ be as in Assumption 2 with $d=1$, let $\mathcal{S}$ be a (relaxed) Mann-Kleene scheme, $x_0\in X$, and consider the iteration $\mathcal{F}=(\mathcal{S},(f_n),x_0)$. If $\mu f=\lim_{n\to\infty}\mu f_n$ then $(x^{\mathcal{F}}_n)$ converges to μf.*

Proof. Assume that (f_n) converges pointwise to f with $\lim\mu f_n=\mu f$ in a one-dimensional problem. In particular, the sequence (μf_n) is bounded by some constant C. Consider the monotone and non-expansive function $f_C\colon\mathbb{R}_{\geq 0}\to\mathbb{R}_{\geq 0}$ given by

$$f_C(x)=\begin{cases}C & \text{if } x\leq C\\ x & \text{if } x\geq C\end{cases}$$

Note that by monotonicity and non-expansiveness we have $f_n\leq f_C$ for all n. Denote by (y_n) the sequence obtained from the iteration $(\mathcal{S},f_C,x)$. Then $x^{\mathcal{F}}_n\leq y_n$ for all n and since (y_n) converges by exactness of $\mathcal{S}$, we can conclude that $(x^{\mathcal{F}}_n)$ is bounded. In fact, for any $C>\mu f$ the inequality $f_n\leq f_C$ does hold for all sufficiently large n and therefore $\limsup_{n\to\infty}x^{\mathcal{F}}_n\leq\mu f$.

Since $(x^{\mathcal{F}}_n)$ is bounded we may assume that (f_n) converges to f uniformly. This implies that $\liminf_{n\to\infty}x^{\mathcal{F}}_n\geq\mu f$ by Lemma 4.1. □

Theorem 4.5 (Approximated Mann-Kleene - power contractions). *Let $f,f_n: X\to X$ be as in Assumption 2 with f a power contraction. Given a Mann-Kleene scheme $\mathcal{S}$ and $x_0\in X$, consider the iteration $\mathcal{F}=(\mathcal{S},(f_n),x_0)$.*

Then, if $(x^{\mathcal{F}}_n)$ is bounded, it converges to the unique fixpoint of f.

An inspection of the proof shows that the result above actually holds even if the scheme $\mathcal{S}$ does not satisfy Condition (6).

For the proof we use the following result of Walter:

Theorem C.2 ([44]). *Let (X, d) be a complete metric space, $f\colon X \to X$ be a Lipschitz continuous power contraction and x^* be the unique fixpoint of f. Let (y_n) be any sequence in X and define $\varepsilon_n := d(y_{n+1}, f(y_n))$, then*

$$\lim_{n\to\infty} y_n = x^* \iff \lim_{n\to\infty} \varepsilon_n = 0$$

Proof (Proof of Theorem 4.5). Since $\alpha_n \to 0$ and $\beta_n \to 0$, the sequence $(T_n^{\mathcal{F}}(f_n, \cdot))$ converges pointwise to f. Since by assumption $(x_n^{\mathcal{F}})$ lives on a bounded domain, we can assume without loss of generality that the convergence is uniform. Hence (instantiating Theorem C.2 with $y_n = x_n^{\mathcal{F}}$) we get

$$\varepsilon_n = \|x_{n+1}^{\mathcal{F}} - f(x_n^{\mathcal{F}})\| = \|T_n(f_n, x_n^{\mathcal{F}}) - f(x_n^{\mathcal{F}})\| \leq \|T_n(f_n, \cdot) - f\| \to 0$$

Therefore $(x_n^{\mathcal{F}})$ converges to the fixpoint of f. □

At the cost of restricting to proper contractions, we can prove a version of this result for other choices of parameter sequences (in particular, ones with $\alpha_n \to 1$).

Theorem C.3. *Under Assumption 2 with a* contraction *f, given a Mann scheme $\mathcal{S}$ fulfilling the additional conditions $\lim_{n\to\infty} \alpha_n = 1$, $\sum_{n\in\mathbb{N}}(1-\alpha_n) = \infty$ and $\lim_{n\to\infty} \beta_n/(1-\alpha_n) = 0$ (but not necessarily Condition (6)), and $x_0 \in X$, consider the iteration $\mathcal{F} = (\mathcal{S}, (f_n), x_0)$.*

Then, if the sequence $(x_n^{\mathcal{F}})$ is bounded, it converges to the fixpoint of f.

Proof. Let $q < 1$ be a contraction factor for f. Since the sequence $(x_n^{\mathcal{F}})$ is bounded by hypothesis, we can assume that $f_n \to f$ uniformly. By Theorem C.2, it suffices to show that $\|x_{n+1}^{\mathcal{F}} - f(x_n^{\mathcal{F}})\| \to 0$. We will show the stronger statement that the sequence (s_n) with

$$s_n := \|x_{n+1}^{\mathcal{F}} - f(x_n^{\mathcal{F}})\| + q\|x_n^{\mathcal{F}} - x_{n+1}^{\mathcal{F}}\|$$

converges to zero.

To this end, note that

$$\begin{aligned}
&\|x_{n+1}^{\mathcal{F}} - f(x_n^{\mathcal{F}})\| \\
&= \|(1-\beta_n)(\alpha_n x_n^{\mathcal{F}} + (1-\alpha_n) f_n(x_n^{\mathcal{F}})) - f(x_n^{\mathcal{F}})\| \\
&\leq (1-\beta_n)\alpha_n \|f(x_n^{\mathcal{F}}) - x_n^{\mathcal{F}}\| + \beta_n \|f(x_n^{\mathcal{F}})\| + (1-\alpha_n)\|f - f_n\|
\end{aligned}$$

and

$$\begin{aligned}
&\|x_n^{\mathcal{F}} - x_{n+1}^{\mathcal{F}}\| \\
&= \|(1-\beta_n)(\alpha_n x_n^{\mathcal{F}} + (1-\alpha_n) f_n(x_n^{\mathcal{F}})) - x_n^{\mathcal{F}}\|
\end{aligned}$$

$$\leq (1-\beta_n)(1-\alpha_n)\|f(x_n^{\mathcal{F}}) - x_n^{\mathcal{F}}\| + \beta_n\|x_n^{\mathcal{F}}\| + (1-\alpha_n)\|f - f_n\|$$

Therefore, we can choose a constant C such that

$$\begin{aligned}
s_n &= \|x_{n+1}^{\mathcal{F}} - f(x_n^{\mathcal{F}})\| + q\|x_n^{\mathcal{F}} - x_{n+1}^{\mathcal{F}}\| \\
&\leq (1-(1-q)(1-\alpha_n))\|f(x_n^{\mathcal{F}}) - x_n^{\mathcal{F}}\| + \beta_n C \\
&\qquad + (1-\alpha_n)\|f - f_n\| \\
&= (1-(1-q)(1-\alpha_n))\|f(x_n^{\mathcal{F}}) - x_n^{\mathcal{F}}\| \\
&\qquad + (1-\alpha_n)(\beta_n/(1-\alpha_n))C + (1-\alpha_n)\|f - f_n\| \\
&\leq (1-(1-q)(1-\alpha_n))\big(\|f(x_{n-1}^{\mathcal{F}}) - x_n^{\mathcal{F}}\| \\
&\qquad + \|f(x_n^{\mathcal{F}}) - f(x_{n-1}^{\mathcal{F}})\|\big) + (1-\alpha_n)(\beta_n/(1-\alpha_n))C + (1-\alpha_n)\|f - f_n\| \\
&\leq (1-(1-q)(1-\alpha_n))\big(\|f(x_{n-1}^{\mathcal{F}}) - x_n^{\mathcal{F}}\| + q\|x_n^{\mathcal{F}} - x_{n-1}^{\mathcal{F}}\|\big) \\
&\qquad + (1-\alpha_n)(\beta_n/(1-\alpha_n))C + (1-\alpha_n)\|f - f_n\| \\
&= (1-(1-q)(1-\alpha_n))s_{n-1} + (1-\alpha_n)((\beta_n/(1-\alpha_n))C + \|f - f_n\|)
\end{aligned}$$

It now follows from Lemma 4 in [29] that $s_n \to 0$. □

Main Theorem 4.6 (Approximated Mann iteration). *Let $f, f_n : X \to X$ be as in Assumption 2, let $\mathcal{S}$ be an exact Mann scheme, $x_0 \in X$, and consider the iteration $\mathcal{F} = (\mathcal{S}, (f_n), x_0)$. Then $(x_n^{\mathcal{F}})$ converges to μf if either*

1. *the sequence (f_n) is monotone and $\mu f = \lim_{n\to\infty} \mu f_n$ or*
2. *the sequence (f_n) converges normally to f, i.e. $\sum_{n\in\mathbb{N}_0} \|f - f_n\| < \infty$.*

Proof. Throughout the proof we denote by (x_n) the sequence obtained from the iteration $(\mathcal{S}, f, x_0)$. By exactness of $\mathcal{S}$ we know that (x_n) converges to μf.

1. Note that in the case where (f_n) is increasing, that means $f_1 \leq f_2 \cdots \leq f$, the condition $\mu f = \lim \mu f_n$ is satisfied automatically, by the Scott-continuity of the fixpoint operator μ.
   Assume without loss of generality that $f_1 \geq f_2 \geq \cdots \geq f$, then also $\mu f_1 \geq \mu f_2 \geq \cdots \geq \mu f$. For a fixed m consider the sequence (y_n) that is inductively defined by $y_0 = x_m^{\mathcal{F}}$ and

   $$y_{n+1} = \big(\alpha_n y_n + (1-\alpha_n) f_m(y_n)\big) \cdot (1-\beta_n)$$

   It is easy to see by induction that $y_n \geq x_{m+n}^{\mathrm{a}}$ for all $n \in \mathbb{N}$. Also, by exactness of $\mathcal{S}$, the sequence (y_n) converges to μf_m and therefore every cluster point p of $(x_n^{\mathcal{F}})$ satisfies $p \leq \mu f_m$. Since $m \in \mathbb{N}$ is arbitrary we obtain that $p \leq \inf_m \mu f_m = \mu f$. Similarly, since $f \leq f_n$ for all $n \in \mathbb{N}$, we have $x_n \leq x_n^{\mathcal{F}}$ for all $n \in \mathbb{N}$. Since the sequence (x_n) converges to μf every cluster point p of $(x_n^{\mathcal{F}})$ is above μf. It follows that μf is the unique cluster point of the sequence $(x_n^{\mathcal{F}})$. The case where $f_1 \leq f_2 \leq \cdots \leq f$ is analogous.

2. Assume that (f_n) converges normally to f. In order to conclude that $(x_n^{\mathcal{F}})$ converges to the same limit as (x_n) we show that $\lim_{n\to\infty} \|x_n - x_n^{\mathcal{F}}\| = 0$. To this aim, by induction we can show that

$$\|x_n - x_n^{\mathcal{F}}\| \leq \sum_{i=0}^{n-1} \prod_{j=i}^{n-1} (1-\beta_j) \cdot (1-\alpha_i) \cdot \|f_i - f\| \tag{9}$$

In fact, first note

$$x_{n+1} - x_{n+1}^{\mathcal{F}} = \big(\alpha_n(x_n - x_n^{\mathcal{F}}) + (1-\alpha_n)(f(x_n) - f_n(x_n^{\mathcal{F}})\big) \cdot (1-\beta_n)$$

and thus

$$\begin{aligned}
&\|x_{n+1} - x_{n+1}^{\mathcal{F}}\| \\
&\quad \leq (\alpha_n \|x_n - x_n^{\mathcal{F}}\| + (1-\alpha_n)\|f(x_n) - f_n(x_n^{\mathcal{F}})\|) \cdot (1-\beta_n) \\
&\quad \leq (\alpha_n \|x_n - x_n^{\mathcal{F}}\| + (1-\alpha_n)(\|f(x_n) - f(x_n^{\mathcal{F}})\| \\
&\qquad + \|f(x_n^{\mathcal{F}}) - f_n(x_n^{\mathcal{F}})\|)) \cdot (1-\beta_n) \\
&\quad \leq (\alpha_n \|x_n - x_n^{\mathcal{F}}\| + (1-\alpha_n)(\|x_n - x_n^{\mathcal{F}}\| + \|f - f_n\|)) \cdot (1-\beta_n) \\
&\quad = (\|x_n - x_n^{\mathcal{F}}\| + (1-\alpha_n)\|f - f_n\|)) \cdot (1-\beta_n) \\
&\quad \leq (\sum_{i=0}^{n-1} \prod_{j=i}^{n-1} (1-\beta_j) \cdot (1-\alpha_i)\|f_i - f\| + (1-\alpha_n)\|f - f_n\|)) \cdot (1-\beta_n) \\
&\quad = (\sum_{i=0}^{n-1} \prod_{j=i}^{n} (1-\beta_j) \cdot (1-\alpha_i)\|f_i - f\|) + (1-\alpha_n)(1-\beta_n)\|f - f_n\| \\
&\quad = \sum_{i=0}^{n} \prod_{j=i}^{n} (1-\beta_j) \cdot (1-\alpha_i)\|f_i - f\|
\end{aligned}$$

as desired.

Now observe that the right hand side of (9) converges to 0. Let $\varepsilon > 0$ and choose $N \in \mathbb{N}$ such that $\sum_{i=N}^{\infty} \|f_i - f\| \leq \varepsilon/2$. Now choose $M \in \mathbb{N}$ such that $\prod_{j=N}^{M} (1-\beta_j) \leq \varepsilon/(2N \max_{n\in\mathbb{N}_0} \|f_n - f\|)$. Then for all $n \geq M$ we have

$$\begin{aligned}
&\sum_{i=0}^{n} \prod_{j=i}^{n} (1-\beta_j) \cdot \|f_i - f\| \\
&\quad \leq \sum_{i=0}^{N} \prod_{j=i}^{n} (1-\beta_j) \cdot \|f_i - f\| + \sum_{i=N+1}^{\infty} \prod_{j=i}^{n} (1-\beta_j) \cdot \|f_i - f\| \\
&\quad \leq N \max_{n\in\mathbb{N}_0} \|f_n - f\| \cdot \prod_{j=N}^{n} (1-\beta_j) + \sum_{i=N+1}^{\infty} \|f_i - f\| \\
&\quad \leq \varepsilon/2 + \varepsilon/2 = \varepsilon
\end{aligned}$$

$\square$

D Proofs and Additional Material for §5 (Value Functions of MDPs and Probabilistic Systems)

Since we are we are working in the *non-discounted* setting, in order to ensure well-definedness of the least fixpoint of the Bellman iteration functions f and g of a MDP, which correspond to the optimal state- or state-action-values of M, respectively, we need to restrict the reward functions by asking that when staying inside a MEC the reward is zero.

Definition D.1 (Proper reward). *Let $M = (S, A, T)$ be an MDP and let R be a corresponding reward function. We say that R is a* proper reward *for M if for all $(S', A') \in \mathrm{MEC}(M)$ and $s, s' \in S'$, $a \in A'(s)$ it holds $R(s, a, s') = 0$.*

Lemma D.2. *Under Assumption 3, given a Mann scheme $\mathcal{S}$ and an initial point $v \in \mathbb{R}^S_{\geq 0}$, consider the iteration $\mathcal{F} = (\mathcal{S}, (f_n), v)$. Then, the sequence $(x^{\mathcal{F}}_n)$ is bounded.*

Proof. Let us consider the iteration $\mathcal{F} \sqcup \mathrm{id} = ((f_n \sqcup \mathrm{id}), (\alpha_n), (\beta_n), v)$ instead.

For all states s, t and actions a with $0 < T(s, a, t) < 1$ fix numbers

$$0 < \underline{T}^{\mathrm{a}}(s, t) < T(s, a, t) < \overline{T}^{\mathrm{a}}(s, t) < 1$$

Note that eventually

$$\underline{T}^{a}(s, t) < T_n(s, a, t) < \overline{T}^{a}(s, t) \text{ for all such } s, t, a \tag{10}$$

Now for a state s of M and each action $a \in A(s)$ let $\mathbb{R}_a(s) := \{t_1, \ldots, t_k\}$ be all the states that can be reached from s via a with positive probability. For all $t_i \in \mathbb{R}_a(s)$ we add a new action $b = b(s, a, t_i)$ to M with $\mathbb{R}_b(s) = \mathbb{R}_a(s)$ and

$$T(s, b, t_j) = \begin{cases} \underline{T}^{a}(s, t_j) & \text{if } j \neq i \\ 1 - \sum_{\gamma \neq i} \underline{T}^{a}(s, t_\gamma) & \text{if } j = i \end{cases}$$

Let N be the MDP which arises from M by adding all these new actions and setting all positive rewards to some large R. Note that N still has the same MECs as M, therefore its value function g has a fixpoint and so does $g \sqcup \mathrm{id}$. In particular the Mann iteration with $g \sqcup \mathrm{id}$ in every step stays bounded from every starting point. Thus, is remains to prove that almost surely we eventually have $f_n \sqcup \mathrm{id} \leq g \sqcup \mathrm{id}$.

It suffices to show that this is implied by (10). Fix $v \in \mathbb{R}^d_{\geq 0}$ and $s \in S$, clearly $\mathrm{id}(v)(s) \leq (g \sqcup \mathrm{id})(v)(s)$. It remains to bound $f_n(v)(s)$, let $a \in A(s)$ be the action where $f_n(v)(s)$ is maximized, then

$$f_n(v)(s) \leq R + \sum_{t \in S} T_n(s, a, t) \cdot v(t) = R + \sum_{t \in \mathbb{R}_a(s)} T_n(s, a, t) \cdot v(t)$$

Let $\mathbb{R}_a(s) = \{t_1, \ldots, t_k\}$ with $v(t_1) \leq \cdots \leq v(t_k)$, then successively shifting probability mass to the maximal value, we see that

$$f_n(v)(s) \leq \sum_{i=1}^{k-1} R + \underline{T}^{\mathrm{a}}(s, t_i) \cdot v(t_i) + \Big(1 - \sum_{i=1}^{k-1} \underline{T}^{\mathrm{a}}(s, t_i)\Big) \cdot v(t_k)$$

$$= R + \sum_{t \in S} T(s,b,t) \cdot v(t) \le g(v)(s) \le (g \sqcup \mathrm{id})(v)(s)$$

with $b = b(s,a,t_k)$. □

Definition D.3 (Witness-bounded functions). *A monotone function u over $\mathbb{R}^S_{\geq 0}$ is called a* witness *if*

1. *for all $v, v' \in \mathbb{R}^S_{\geq 0}$ it holds $|u(v) - u(v')| \le u(|v - v'|)$;*
2. *there are $k \in \mathbb{N}$, $0 \le c < 1$ such that for all $v \in \mathbb{R}^S_{\geq 0}$ we have $\|u^k(v)\| \le c\|v\|$.*

The function $h \colon \mathbb{R}^S_{\geq 0} \to \mathbb{R}^S_{\geq 0}$ is called witness-bounded *if there is a witness $u \colon \mathbb{R}^S_{\geq 0} \to \mathbb{R}^S_{\geq 0}$ such that $|h(v) - h(v')| \le u(|v - v'|)$ for all $v, v' \in \mathbb{R}^S_{\geq 0}$.*

Note that a witness u is witness-bounded, as one can take u itself as its witness. Also, we can show that witness-bounded functions are power contractions and they admit a least fixpoint.

Proposition D.4 (Fixpoints of witness-bounded functions). *We assume a witness-bounded function $h \colon \mathbb{R}^S_{\geq 0} \to \mathbb{R}^S_{\geq 0}$ with witness $u \colon \mathbb{R}^S_{\geq 0} \to \mathbb{R}^S_{\geq 0}$, $k \in \mathbb{N}$, and $0 \le c < 1$ such that $\|u^k(v)\| \le c\|v\|$. Then*

1. *h is a power contraction;*
2. *h admits a least fixpoint μh such that $\|\mu h\| \le \frac{1}{1-c}\|h^k(0)\|$.*

Proof. (1) We show that h^k is a contraction with factor c. For all $v, v' \in \mathbb{R}^S_{\geq 0}$, repeatedly applying property (1) of Definition D.3 and using monotonicity of u we obtain

$$|h^k(v) - h^k(v')| \le u(|h^{k-1}(v) - h^{k-1}(v')|) \le \ldots \le u^k(|v - v'|)$$

hence

$$\|h^k(v) - h^k(v')\| \le \|u^k(|v - v'|)\| \le c\| \; |v - v'| \; \| = c\|v - v'\|.$$

where the second inequality uses (2) of Definition D.3.

(2) First observe that, by the previous point, for all $i \in \mathbb{N}$ it holds

$$\|h^{k(i+1)}(0) - h^{ki}(0)\| \le c^i \|h^k(0) - h^0(0)\| = c^i \|h^k(0)\| \tag{11}$$

where we write 0 to denote the constant function 0.

This means that if we iterate h over 0, we obtain

$$\begin{aligned} \|h^{k(i+1)}(0)\| &= \| \sum_{j=0}^{i} (h^{k(j+1)}(0) - h^{kj}(0)) \| \\ &\le \sum_{j=0}^{i} \|h^{k(j+1)}(0) - h^{kj}(0)\| && \text{[by (11)]} \end{aligned}$$

$$\leq \sum_{j=0}^{i} c^j \|h^k(0)\| \qquad\qquad [\text{since } c<1 \text{ and } \textstyle\sum_{j=0}^{\infty} c^j = \frac{1}{1-c}]$$
$$\leq \frac{1}{1-c}\|h^k(0)\|$$

Since all iterates are below a fixed bound $b = \frac{1}{1-c}\|h^k(0)\|$, we can restrict h to the lattice $[0,b]^S$. By Knaster-Tarski the restriction has a least fixpoint μh which is also the least fixpoint of the original function. Hence h has a fixpoint μh and $\|\mu h\| \leq \frac{1}{1-c}\|h^k(0)\|$. $\square$

We need a preliminary technical lemma.

Lemma D.5 (Ranking MDPs). *Let $M = (S, A, T, R)$ be a simple MDP. Then there is a function $r\colon S \to \mathbb{N}_0$ such that for all $s \in S$, it holds $r(s) < |S|$, $r(s) = 0$ implies $s \in F(M)$ and for all $a \in A(s)$ there exists $s' \in S$ with $T(s, a, s') > 0$ and $r(s') < r(s)$.*

Proof. The function r can be defined inductively as follows.

- $r(s) = 0$ for $s \in F(M)$;
- Assume that the states with $r(s) \leq k$ have been already identified and let S_k be the set of such states. Consider the set of states

  $$S' = \{s \in S \setminus S_k \mid \forall a \in A(s). \exists s' \in S_k.\ T(s, a, s') > 0\}.$$

  We have $S' \neq \emptyset$, otherwise, if for all $s \in S \setminus S_k$ there is $a \in A(s)$ such that for all s' with $T(s', a, s'') > 0$ it holds $s'' \in S \setminus S_k$, then $S \setminus S_k$ would contain an end-component, violating the hypotheses.
  Then let $r(s) = k + 1$ for all all $s \in S'$.

Function r satisfies the desired properties by construction. In particular, since at each step $S' \neq \emptyset$, there will be at most $|S|$ iterations, hence $r(s) \leq |S|$. $\square$

The Bellman operators of a simple MDP are witness-bounded, which will shed more intuition on the concept. First, such functions are power contractions (cf. Definition D.3(2)) and second, observe that expectation and maximum are the central operators of the Bellman equations and they satisfy Definition D.3(1)) (unlike, for instance, the minimum). In fact the witness u for an MDP function is based on an MDP with the same parameters, but setting all rewards to 0.

The fact that the Bellman operators of simple MDPs are power contractions is probably folklore, but we could not find it in the literature. Here we will prove a more general statement.

Proposition D.6 (Simple MDP functions are witness-bounded). *Let $M = (S, A, T, R)$ be a simple MDP. Then, its Bellman operator f_M is witness-bounded.*

Proof. Let us prove that f_M is witness-boundedwith witness $u\colon \mathbb{R}^S_{\geq 0} \to \mathbb{R}^S_{\geq 0}$ defined by

$$u(v)(s) = \max_{a \in A(s)} \sum_{s' \in S} T(s, a, s')v(s').$$

In fact, for all $v, v' \in \mathbb{R}^S_{\geq 0}$ we have, for all $s \in S$:

$$\begin{aligned}
&|f(v)(s) - f(v')(s)| = \\
&\quad = |\max_{a \in A(s)} \sum_{s' \in S} T(s, a, s')(R(s, a, s') + v(s')) - \\
&\qquad \max_{a \in A(s)} \sum_{s' \in S} T(s, a, s')(R(s, a, s') + v'(s'))| \\
&\quad \leq \max_{a \in A(s)} |\sum_{s' \in S} T(s, a, s')(R(s, a, s') + v(s')) - \\
&\qquad\qquad \sum_{s' \in S} T(s, a, s')(R(s, a, s') + v'(s'))| \\
&\quad \leq |\max_{a \in A(s)} \sum_{s' \in S} T(s, a, s')(v(s') - v'(s')) \\
&\quad \leq \max_{a \in A(s)} \sum_{s' \in S} T(s, a, s')|v(s') - v'(s')| \\
&\quad = u(|v - v'|)(s)
\end{aligned}$$

Thus, $|f(v) - f(v')| \leq u(|v - v'|)$ as desired.

The function u is a witness, in fact:

1. for all $v, v' \in \mathbb{R}^S_{\geq 0}$ we have $|f(v) - f(v')| \leq u(|v - v'|)$. The proof is analogous to the one above.
2. In order to show that u is a power contraction, we consider the function $r(s)$ from Lemma D.5 and prove, by induction that for all $s \in S$ it holds that $u^{r(s)+1}(v)(s) \leq c_s \|v\|$ for a suitable constant c_s, with $0 \leq c_s < 1$.
   - ($r(s) = 0$) In this case $s \in F(M)$ and thus $u^{0+1}(v)(s) = u(v)(s) = 0 \leq 0 \cdot \|v\|$, as desired.
   - ($r(s) > 0$) First observe that, given any $a \in A(s)$, by Lemma D.5, there is is a state $s'_a \in S$ such that $T(s, a, s'_a) > 0$ and $r(s') < r(s)$. Hence

$$\begin{aligned}
&u^{r(s)}(v)(s) = \\
&= \max_{a \in A(s)} \sum_{s'' \in S} T(s, s'')u^{r(s)-1}(v)(s'') \\
&= \max_{a \in A(s)} T(s, s'_a)u^{r(s)-1}(v)(s'_a) + \sum_{s'' \neq s'_a} T(s, s'')u^{r(s)-1}(v)(s'') \\
&\leq \max_{a \in A(s)} T(s, s'_a)c_{s'_a}\|v\| + \sum_{s'' \neq s'_a} T(s, s'')\|v\| \\
&= \max_{a \in A(s)} (T(s, s'_a)c_{s'_a} + \sum_{s'' \neq s'_a} T(s, s''))\|v\| \\
&= c_s \cdot \|v\|
\end{aligned}$$

where $c_s = \max_{a \in A(s)}(T(s, s'_a)c_{s'_a} + \sum_{s'' \neq s'_a} T(s, s'')) < 1$.
Given the above, we conclude that $u^{|S|}(v)(s) \leq c_s \cdot \|v\|$ for all $s \in S$ and thus $\|u^{|S|}(v)\| \leq c \cdot \|v\|$ where $c = \max_{s \in S} c_s < 1$, as desired.

□

In particular, we can use Theorem 4.5 to get convergence of Mann iteration.

Lemma 5.4 (Characterising MDPs with finite value). *An MDP M has a finite value, i.e., its Bellman operator f_M has a (least) fixpoint, if and only if for all maximal end-components (S', A') and all $s \in S', a \in A'(s)$, and $s' \in S'$ with $T(s, a, s') > 0$, we have $R(s, a, s') = 0$.*

Proof. By existence of an optimal stationary policy, the value v^*_M is finite (i.e., f_M has a fixpoint) if and only if $v^\pi = v^\pi_M$ is finite for all policies $\pi \in \Pi(M)$ (i.e., f^π_M has a fixpoint).

Now, fix a policy $\pi \in \Pi(M)$. Note that end-components of M^π are also end-components of M.

Assume that $R^\pi(s, s') = 0$ for all $s, s' \in S'$ with $T^\pi(s, s') > 0$ in every end-component given by S' of M^π.[5]

Then, we have for all $v \in \mathbb{R}^S_{\geq 0}$ with $v\big|_{S'} = 0$ that

$$\begin{aligned} f^\pi_M(v)(s) &= \sum_{s' \in S} T^\pi(s, s')(R^\pi(s, s') + v(s')) \\ &= \sum_{s' \in S'} T^\pi(s, s')R^\pi(s, s') \\ &= 0 \end{aligned}$$

for all $s \in S'$ whence

$$v^\pi\big|_{S'} = \lim_{n \to \infty} (f^\pi_M)^n(0)\big|_{S'} = 0.$$

On the other hand, let $s \in S$ be such that $v^\pi(s) > 0$. By the above, we have that $s \notin S'$ for all end-components S'.[6] Then, there exists a path of length $|S|$ to a state $s' \in S'$ for some end-component S' of M^π as otherwise, s together with the set of states reachable from s would form an end component. Let us denote the probability of this path by $p(s)$. But then

$$\begin{aligned} \max_{s \in S} v^\pi(s) &\leq \max_{s \in S, v^\pi(s) > 0} |S| \cdot R + \sum_{s' \in S} \mathbb{P}^\pi(\mathbf{s}_n = s' \mid \mathbf{s}_0 = s) v^\pi(s') \\ &\leq |S| \cdot R + (1 - \min_{s \in S, v^\pi(s) > 0} p(s)) \max_{s' \in S} v^\pi(s') \\ &\leq |S| \cdot R / \min_{s \in S, v^\pi(s) > 0} p(s) < \infty. \end{aligned}$$

[5] Here, we abbreviate $R^\pi(s, s') = R(s, \pi(s), s')$ and $T^\pi(s, s') = T(s, \pi(s), s')$.
[6] For the sake of readability, we assume that $F = \emptyset$ (the argument is the same with final states treated analogously to end-component states).

Thus, if $R(s,a,s') = 0$ for all $s,s' \in S', a \in A'(s)$ with $T(s,a,s') > 0$ in end-components (S',A') of M, whence $R^\pi(s,s') = 0$ for all $s,s' \in S'$ with $T^\pi(s,s') > 0$ in end components S' of M^π for all $\pi \in \Pi(M)$, we have v^π is finite for all $\pi \in \Pi(M)$, whence also v^*_M is finite.

Now assume that $r = R^\pi(s,s') > 0$ for some $s,s' \in S'$ with $T^\pi(s,s') > 0$ in some end component of M^π given by S'. Then, we have

$$\begin{aligned}
v^\pi(s) &= \mathbb{E}^\pi[\sum_{n\in\mathbb{N}_0} \mathbf{r}_n \mid \mathbf{s}_0 = s] \\
&\geq r \sum_{n\in\mathbb{N}_0} \mathbb{P}^\pi(\mathbf{s}_n = s, \mathbf{s}_{n+1} = s' \mid \mathbf{s}_0 = s) \\
&= rT^\pi(s,s') \sum_{n\in\mathbb{N}_0} \mathbb{P}^\pi(\mathbf{s}_n = s \mid \mathbf{s}_0 = s) \\
&= \infty
\end{aligned}$$

since the state s is essential (see, e.g., [20]).

Thus, if v^*_M is finite, whence all v^π are finite, also $R^\pi(s,s') = 0$ for all $s,s' \in S'$ with $T^\pi(s,s') > 0$ in end-components S' of M^π and $\pi \in \Pi(M)$. By definition of end-components, this gives $R(s,a,s') = 0$ for all $s,s' \in S', a \in A'(s)$ with $T(s,a,s') > 0$ in end-components (S',A') of M. □

We need another technical lemma.

Lemma D.7 (Relating MDPs and quotients). *Let $M = (S,A,T,R)$ be an MDP with finite value, let $\equiv$ by the equivalence identifying states in the same MEC, with $r\colon S \to S/_\equiv$ sending each state to its equivalence class, and let $M_r = (S_r, A_r, T_r, R_r)$ be its quotient as in Definition 5.5. Let $f_r = f_{M_r}$ and $\hat{f}_r = f_{\hat{M}_r}$, then $f_r = \max_r \circ f \circ r^\bullet$. Moreover, $\hat{f}_r \leq f_r \leq \hat{f}_r \sqcup id$.*

Proof. Let $w \in \mathbb{R}^Q_{\geq 0}$. Then for $e \in S/_\equiv = Q$ we have

$$\begin{aligned}
&\max_r \circ f \circ r^\bullet(w)(e) = \\
&= \max_{s\in e} f(w \circ r)(s) \\
&= \max_{s\in e} \max_{a\in A(s)} \sum_{s'\in S} T(s,a,s')(R(s,a,s') + w(r(s'))) \\
&= \max_{s\in e} \max_{a\in A(s)} \sum_{e'\in S} (\sum_{s'\in e'} T(s,a,s')(R(s,a,s') + w(e'))) \\
&= \max_{s\in e} \max_{a\in A(s)} \sum_{e'\in S} (T_r(e,a,e')w(e') + \sum_{s'\in e'} T(s,a,s')R(s,a,s')) \\
&= \max_{s\in e} \max_{a\in A(s)} \sum_{e'\in S} (T_r(e,a,e')w(e') + T_r(e,a,e')R_r(e,a,e')) \\
&= \max_{a\in A_r(e)} \sum_{e'\in S} (T_r(e,a,e')(w(e') + R_r(e,a,e'))
\end{aligned}$$

$$= f_r(w)(e)$$

The second part, $\hat{f}_r \leq f_r \leq f_r \sqcup id$ is immediate from the fact that M_r and $\hat{M}_r$ only differ in the enabled actions. The first inequality $\hat{f}_r \leq f_r$ is due to the fact that $\hat{M}_r$ has less enabled actions. The second $f_r \leq f_r \sqcup id$ derives from the fact that the only additional actions in M_r are self-loops with probability 1. □

Proposition 5.6 (Fixpoints of non-discounted MDPs). *Let M be an MDP with finite value. Let $\hat{M}_r$ be the reduced quotient. Then $v^*_M = r^\bullet(v^*_{\hat{M}_r})$.*

Proof. First of all observe that, since $\hat{M}_r = (\hat{S}_r, \hat{A}_r, \hat{T}_r, \hat{R}_r)$ is a simple MDP, by Proposition D.6, $\hat{f}_r$ is a witness-bounded function and thus, by Proposition D.4(2), it admits a least fixpoint $\mu\hat{f}_r$.

By Lemma D.7, $\hat{f}_r \leq f_r \leq \hat{f}_r \sqcup id$ and thus $\mu\hat{f}_r \leq \mu f_r \leq \mu(\hat{f}_r \sqcup id)$. It is easy to see that $\mu(\hat{f}_r \sqcup id) = \mu\hat{f}_r$, since they produce the same sequence of Kleene iterates from 0, and thus $\mu f_r = \mu\hat{f}_r$

In order to conclude we show $\mu f = r^\bullet(\mu f_r)$. First note that $r^\bullet(\mu f_r)$ is a pre-fixpoint of f and thus $\mu f \leq r^\bullet(\mu f_r)$. In fact

$$\begin{aligned}
&f(r^\bullet(\mu f_r)) \leq \\
&\leq r^\bullet(\max_r(f(r^\bullet(\mu f_r)))) \\
&= r^\bullet(f_r(\mu f_r)) && \text{[by Lemma D.7, } f_r = \max_r \circ f \circ r^\bullet \text{]} \\
&= r^\bullet(\mu f_r)
\end{aligned}$$

For the converse inequality, first note that if $v \in \mathbb{R}^S_{\geq 0}$ is a fixpoint of f then

$$v = r^\bullet(\max_r(v)),$$

i.e., for all MECs E, for any two states $s, s' \in E$, $v(s) = v(s')$. In fact, assume that $v \neq r^\bullet(\max_r(v))$ and let E be a MEC where v has not the same value on all states. Let π be a policy which "cycles" in E, i.e., such that for all $s \in E$ it holds $T(s, \pi(s), s') > 0$ implies $s' \in E$. Let $s \in E$ be such that $v(s) = \min\{v(s') \mid s' \in E\}$ and $T(s, \pi(s), s') > 0$ for some s' such that $v(s) < v(s')$. A simple calculation shows that $f(v)(s) \geq f^\pi(v)(s) > v(s)$, hence v cannot be a fixpoint for f.

Using the above, we show that $\max_r(\mu f)$ is a pre-fixpoint of f_r:

$$\begin{aligned}
&f_r(\max_r(\mu f)) = \\
&= \max_r(f(r^\bullet(\max_r(\mu f)))) && \text{[by Lemma D.7, } f_r = \max_r \circ f \circ r^\bullet \text{]} \\
&= \max_r(f(\mu f)) && \text{[since } r^\bullet(\max_r(\mu f)) = \mu f\text{]} \\
&= \max_r(\mu f)
\end{aligned}$$

Therefore $\mu f_r \leq \max_r(\mu f)$ and thus

$$r^\bullet(\mu f_r) \leq \mu r^\bullet(\max_r(\mu f))) = \mu f,$$

as desired. □

Lemma D.8 ([34]). *Let (f_n) be a sequence of contractions $f_n\colon [0,1]^S \to [0,1]^S$ with respective fixpoints a_n and $f\colon [0,1]^S \to [0,1]^S$ be a contraction with fixpoint a. If (f_n) converges pointwise to f, then (a_n) converges to a.*

Theorem 5.7. *Under Assumption 3, we have $\lim_{n\to\infty} \mu f_n = \mu f$.*

Proof. Fix a policy $\pi\colon S \to A$ and consider the "Markov chain" M^π together with its value iteration function $f^\pi\colon \mathbb{R}^S_{\geq 0} \to \mathbb{R}^S_{\geq 0}$.

Let S' be the set of states $s \in S$ with $\mu f^\pi(s) = 0$ and consider

$$\hat{f}^\pi(v)(s) = \begin{cases} f^\pi(v)(s) & \text{if } s \notin S' \\ 0 & \text{if } s \in S' \end{cases}$$

Note that, if $\equiv$ is the function identifying states in the same MEC of M^π and $r\colon S \to S/_\equiv$ is the corresponding quotient function then $\hat{f}^\pi$ is $r^\bullet \circ \hat{f}^\pi$. Thus, by Proposition D.6, it is a *power contraction* and, by Proposition 5.6, its unique fixpoint is equal to μf^π.

We can analogously consider for $n \in \mathbb{N}$ the Markov Chain M^π_n, the set S'_n and the maps f^π_n and $\hat{f}^\pi_n$. Now note that (f_n) converges to f almost surely and moreover if $T(s,a,t) = 0$ for some $a \in A$ and $s,t \in S$ then also $T_n(s,a,t) = 0$ for all $n \in \mathbb{N}$. It follows that almost surely $S'_n = S'$ for all n above a fixed threshold N. But this implies that $\hat{f}^\pi_n$ converges to $\hat{f}^\pi$ and the same is true for the k-fold iterations. Thus, in this case we have

$$\lim_{n\to\infty} \mu f^\pi_n = \lim_{n\to\infty} \mu \hat{f}^\pi_n = \lim_{n\to\infty} \mu(\hat{f}^\pi_n)^k = \mu(\hat{f}^\pi)^k = \mu \hat{f}^\pi = \mu f^\pi$$

Since π is arbitrary and there are only finitely many policies, we obtain

$$\lim_{n\to\infty} \mu f_n = \lim_{n\to\infty} \max_\pi \mu f^\pi_n = \max_\pi \lim_{n\to\infty} \mu f^\pi_n = \max_\pi \mu f^\pi = \mu f$$

almost surely. □

We first prove a preliminary result concerning witness-bounded functions.

Proposition D.9 (Mann iteration for witness-bounded functions). *Let $h\colon \mathbb{R}^S_{\geq 0} \to \mathbb{R}^S_{\geq 0}$ be witness-bounded. Let (h_n) be a sequence of functions converging to h. Given a regular Mann scheme $\mathcal{S} = ((\alpha_n),(\beta_n))$ and arbitrary $v \in \mathbb{R}^S_{\geq 0}$, consider the iteration $\mathcal{F}^{\mathrm{a}} = (\mathcal{S}, (h_n \sqcup \mathrm{id}), v)$.*

If $(x^{\mathcal{F}^{\mathrm{a}}}_n)$ is bounded, then $x^{\mathcal{F}^{\mathrm{a}}}_n \to \mu h$.

Proof. Let us write v^{a}_n instead of $v^{\mathcal{F}^{\mathrm{a}}}_n$. Consider the iteration $\mathcal{F} = (\mathcal{S}, h \sqcup id, v_0)$ and let us write v_n for $v^{\mathcal{F}}_n$.

Since $h \sqcup id$ is non-expansive, monotone and has at least one fixpoint, we know that $v_n \to \mu(h \sqcup id) = \mu h$ by regularity.

In order to conclude we show that v^{a}_n converges to the same limit of v_n. Consider

$$|v^{\mathrm{a}}_{n+1} - v_{n+1}| =$$

$$= |(1-\beta_n)(\alpha_n(v_n^{\mathrm{a}} - v_n) + (1-\alpha_n)((h_n \sqcup id)(v_n^{\mathrm{a}}) - (h \sqcup id)(v_n)))|$$
$$\leq (1-\beta_n)(\alpha_n|v_n^{\mathrm{a}} - v_n| + (1-\alpha_n)|(h_n \sqcup id)(v_n^{\mathrm{a}}) - (h \sqcup id)(v_n)|$$
$$\leq (1-\beta_n)(\alpha_n|v_n^{\mathrm{a}} - v_n| + (1-\alpha_n)(|h_n(v_n^{\mathrm{a}}) - h(v_n)| \sqcup |id(v_n^{\mathrm{a}}) - id(v_n)|)$$

Now, if u is the function witnessing that h is witness-bounded we have

$$\begin{aligned}
&|h_n(v_n^{\mathrm{a}}) - h(v_n)| \leq \\
&\leq |h_n(v_n^{\mathrm{a}}) - h(v_n^{\mathrm{a}})| + |h(v_n^{\mathrm{a}}) - h(v_n)| \\
&\leq \varepsilon_n + u(|v_n^{\mathrm{a}} - v_n|)
\end{aligned}$$

where $\varepsilon_n = \sup_{i \geq n} \|h_n(v_n^{\mathrm{a}}) - h(v_n^{\mathrm{a}})\|$ ($\to 0$ since $h_n \to h$ and (v_n^{a}) bounded). We can use this in the analysis of $|v_{n+1}^{\mathrm{a}} - v_{n+1}|$ and deduce that

$$\begin{aligned}
&|v_{n+1}^{\mathrm{a}} - v_{n+1}| \leq \\
&\leq (1-\beta_n)(\alpha_n|v_n^{\mathrm{a}} - v_n| + (1-\alpha_n)\max\{u(|v_n^{\mathrm{a}} - v_n|) + \varepsilon_n, |v_n^{\mathrm{a}} - v_n|\} \\
&= (1-\beta_n)(\alpha_n|v_n^{\mathrm{a}} - v_n| + (1-\alpha_n)(h'_n \sqcup id)(|v_n^{\mathrm{a}} - v_n|))
\end{aligned}$$

where $h'_n(v) = u(v) + \varepsilon_n$.

If we denote by (v'_n) the sequence

$$\begin{aligned}
&v'_0 = |v_0^{\mathrm{a}} - v_0| \\
&v'_{n+1} = (1-\beta_n)\,(\alpha_n v'_n + (1-\alpha_n)((h'_n \sqcup id)(v'_n)))
\end{aligned}$$

then $|v_n^{\mathrm{a}} - v_n| \leq v'_n$. Observe that for v'_n we again obtain a dampened Mann iteration.

Now, observe that ε_n is a decreasing sequence, converging to 0 since $h_n \to h$ (and (v_n^{a}) bounded). Therefore, we have that $h'_n \sqcup id \to u \sqcup id$ monotonically from above. Moreover h'_n is witness-bounded, in fact $|h'_n(v) - h'_n(v')| = |u(v) - u(v')| \leq u(|v - v'|)$.

Hence by Proposition D.4 there exists c such that:

$$|\mu h'_n| \leq \tfrac{1}{1-c}\|(h'_n)^k(0)\| \leq \tfrac{k}{1-c}\varepsilon_n$$

since $(h'_n)^k(0) \leq k\varepsilon_n$. Hence $\mu h'_n \to 0 = \mu h'$ where $h' = u \sqcup \mathrm{id}$.

Thus, by Theorem 4.6(1), we have that $v_n \to \mu(h' \sqcup id) = \mu h' = 0$. Since v'_n is an upper bound for $|v_n^{\mathrm{a}} - v_n|$ we conclude that v_n^{a} converges to the same limit of v_n, i.e., $\mu(h \sqcup id) = \mu h$, as desired. □

Lemma D.10. *Let (h_n) and (h'_n) be sequences of functions with $h_n \colon \mathbb{R}_{\geq 0}^S \to \mathbb{R}_{\geq 0}^S$, $h'_n \colon \mathbb{R}_{\geq 0}^{S'} \to \mathbb{R}_{\geq 0}^{S'}$ and $r \colon S \to S'$ be a function such that $h_n \circ r^\bullet \leq r^\bullet \circ h'_n$.*

Let $x_n^{\mathrm{a}}, (x')_n^{\mathrm{a}}$ be the corresponding dampened Mann iterations, for which we assume that $x_0^{\mathrm{a}} \leq r^\bullet((x')_0^{\mathrm{a}})$. Then $x_n^{\mathrm{a}} \leq r^\bullet((x')_n^{\mathrm{a}})$ holds for all n.

Proof. Let $w \in [0,1]^{S'}$ and $s \in S$. We first observe that $(c \cdot r^\bullet(w))(s) = c \cdot w(r(s)) = (c \cdot w)(r(s)) = r^\bullet(c \cdot w)(s)$, hence $c \cdot r^\bullet(w) = r^\bullet(c \cdot w)$. Additionally for $w_1, w_2 \in [0,1]^{S'}$ we obtain $(r^\bullet(w_1) + r^\bullet(w_2))(s) = r^\bullet(w_1)(s) + r^\bullet(w_2)(s) = w_1(r(s)) + w_2(r(s)) = (w_1 + w_2)(r(s)) = r^\bullet(w_1 + w_2)(s)$, hence $r^\bullet(w_1) + r^\bullet(w_2) = r^\bullet(w_1 + w_2)$.

Using these facts we can show by induction that

$$\begin{aligned}
x^{\mathrm{a}}_{n+1} &= (1-\beta_n)\cdot(\alpha_n\cdot x_n + (1-\alpha_n)\cdot h_n(x^{\mathrm{a}}_n))\\
&\leq (1-\beta_n)\cdot(\alpha_n\cdot r^\bullet((x')^{\mathrm{a}}_n) + (1-\alpha_n)\cdot h_n(r^\bullet((x')^{\mathrm{a}}_n)))\\
&\leq (1-\beta_n)\cdot(\alpha_n\cdot r^\bullet((x')^{\mathrm{a}}_n) + (1-\alpha_n)\cdot r^\bullet(h'_n((x')^{\mathrm{a}}_n)))\\
&= (1-\beta_n))\cdot r^\bullet((\alpha_n\cdot(x')^{\mathrm{a}}_n + (1-\alpha_n)\cdot h'_n((x')^{\mathrm{a}}_n))\\
&= r^\bullet((x')^{\mathrm{a}}_{n+1}),
\end{aligned}$$

□

With this, we can now observe that the dampened Mann iteration for an arbitrary sampled MDP indeed converges to the optimal value function.

Main Theorem 5.8 (Mann iteration over MDPs). *Let M, M_n be MDPs as in Assumption 3. Given a regular Mann scheme $\mathcal{S}$ and any $v \in \mathbb{R}^S_{\geq 0}$, consider the iteration $\mathcal{F}^{\mathrm{a}} = (\mathcal{S}, (f_n), v)$, generating a sequence $(v^{\mathcal{F}^{\mathrm{a}}}_n)$. Then $\lim_{n\to\infty} v^{\mathcal{F}^{\mathrm{a}}}_n = v^*_M = \mu f_M$.*

Proof. Let us write v^{a}_n instead of $v^{\mathcal{F}}_n$. We show that the cluster points of the sequence can be bounded from above and from below by μf_M, and then we conclude.

First of all, observe that we can assume that all approximations M_n have the same MECs which are in turn the MECs of M. In fact, by Assumption 3, the approximations M_n are obtained by sampling and $T_n \to T$ (almost surely). Hence we can find an index n_0 such that for all $n \geq n_0$ $T(s,a,s') > 0$ if and only if $T_n(s,a,s') > 0$ and consider the iteration starting from $v^{\mathrm{a}}_{n_0}$.

Hereafter we work under this assumption, indicate by $\equiv$ is the function identifying states in the same MEC of M^π and by $r\colon S \to S/_\equiv$ the corresponding quotient function.

Upper bound Let (f_r) be the function associated to the quotient $(M_n)_r$ defined as in Definition 5.5. Let $\mathcal{S} = ((\alpha_n),(\beta_n))$. By Lemma D.10, the sequence (w^{a}_n) generated by the iteration $\mathcal{F} = (\mathcal{S}, (f_r), w^{\mathrm{a}}_0)$ with w^{a}_0 such that $v^{\mathrm{a}}_0 \leq r^\bullet(w^{\mathrm{a}}_0)$ is such that that

$$v^{\mathrm{a}}_n \leq r^\bullet(w^{\mathrm{a}}_n)$$

If we denote $(\hat{f}_r)$ be the function associated with the reduced quotient $(\hat{M}_n)_r$ then, by Lemma D.7, $(f_r) \leq (\hat{f}_r) \sqcup id$. Hence, if we consider the iteration $\mathcal{F} = (\mathcal{S}, (\hat{f}_r) \sqcup id, z^{\mathrm{a}}_0)$, with $z^{\mathrm{a}}_0 = w^{\mathrm{a}}_0$, and call (z^{a}_n) the induced sequence then

$$w^{\mathrm{a}}_n \leq z^{\mathrm{a}}_n$$

Moreover, it is clear that $(\hat{f}_r)$ converges to $\hat{f}_r$, the function associated to the reduced quotient $\hat{M}_r$ of M. Since $\hat{f}_r$ is witness-bounded and (z^{a}_n) is bounded by Lemma D.2, by Proposition D.9, z^{a}_n converges to $\mu\hat{f}_r$.

Putting things together and using monotonicity and continuity of $r^\bullet$, we have that $v^{\mathrm{a}}_n \leq r^\bullet(w^{\mathrm{a}}_n) \leq r^\bullet(z^{\mathrm{a}}_n)$. Moreover $r^\bullet(z^{\mathrm{a}}_n)$ converges to $r^\bullet(\mu\hat{f}_r) = \mu f_M$, with the last equality given by Proposition 5.6.

We thus conclude that μf_M is an upper bound for the cluster points of x^{a}_n.

Lower bound: Let $f = f_M$. Consider some fixed policy $\pi\colon S \to A$ and let M_n^π be the Markov chain obtained by the n-th approximation of M. Clearly the corresponding iteration function satisfies $f_n^\pi \le f_n$. Moreover, if we indicate by $E = \bigcup_n MEC(M_n)$, the set of states in some end-component of M_n^π and, given $h\colon \mathbb{R}^S_{\ge 0} \to \mathbb{R}^S_{\ge 0}$ we write h^E for the function

$$h^E(v)(s) = \begin{cases} 0 & \text{if } s \in E \\ h(v)(s) & \text{otherwise.} \end{cases}$$

Then, if $\hat{f}_r^\pi$ is the function associated to the reduced quotient (see Definition 5.5), it holds $(f^\pi)^E = r^\bullet \circ \hat{f}_r^\pi \circ \max_r$. Since the reduced quotient is simple, we infer that $(f^\pi)^E$ is a power contraction (note that $\max_r \circ r^\bullet = id$ whence $((f^\pi)^E)^n = r^\bullet \circ (\hat{f}_r^\pi)^n \circ \max_r$) and, by Proposition 5.6, $\mu (f^\pi)^E = r^\bullet(\mu \hat{f}_r^\pi) = \mu f^\pi$.

Consider the sequence y_n^π generated by the iteration $\mathcal{F} = (\mathcal{S}, (f_n^\pi)^E, y_0^\pi)$ with $y_0^\pi \le x_0^{\mathrm{a}}$. Remembering that $(f_n^E)^\pi \le f_n^\pi \le f_n$ we have that

$$y_n^\pi \le x_n^{\mathrm{a}}$$

Since $(f_n^\pi)^E$ converges to $(f^\pi)^E$ and $(f^\pi)^E$ is a power contraction, by boundedness of y_n^π and regularity of the scheme, y_n^π converges to $\mu (f^\pi)^E = \mu f^\pi$, which is therefore a lower bound for the cluster points of x_n^{a}.

The above applies to all possible policies $\pi\colon S \to A$, hence also $\max_\pi \mu f^\pi = \mu f$ is a lower bound.

Given that the upper and lower bound coincide with μf this is necessarily the limit of x_n^{a}. □

Theorem 5.10. *For exact Mann schemes the sequence (x_i) produced by Algorithm 1 converges to the least fixpoint of f almost surely.*

Proof. By construction we have

$$\sum_{i=1}^{\infty} \mathbb{P}\big[\|f_{n_i} - f\| > \gamma_i\big] \le \sum_{i=1}^{\infty} \delta_i < \infty$$

The Borel-Cantelli Lemma [8] states:

Let $(A_n)_{n\in\mathbb{N}}$ be a sequence of events in a probability space and assume that $\sum_{n=1}^\infty \mathbb{P}[A_n] < \infty$, i.e., the sum of probabilities is finite. Then

$$\mathbb{P}[\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k] = \mathbb{P}[\{\omega \in \Omega \mid \text{there are infinitely many } n \in \mathbb{N} \text{ with } \omega \in A_n\}] = 0$$

So by applying this lemma we get that

$$\mathbb{P}\big[\|f_{n_i} - f\| > \gamma_i \text{ for infinitely many } i\big] = 0$$

Hence, almost surely we have $\|f_{n_i} - f\| \le \gamma_i$ eventually and by Theorem 4.6(2) the sequence produced by the algorithm converges to the solution vector of its input in that case. □

E Application: Simple Stochastic Games

As an example application of the technique in Section 5.4 we consider (simple) stochastic games or SSGs. An SSG is given by a finite set of nodes partitioned into $V_{\min}$, $V_{\max}$, V_{av} and V_{sink}, and the following data: $\eta\colon V_{\min} \cup V_{\max} \to \mathcal{P}(V)$ (successor function for Min and Max nodes), $\eta_{\mathrm{av}}\colon V_{\mathrm{av}} \to \mathcal{D}(V)$ (where $\mathcal{D}(V)$ is the set of probability distributions over V) and a map $w\colon V_{\mathrm{sink}} \to [0,1]$ that assigns to each sink state a payoff. The value vector of an SSG is the least fixpoint of the monotone and non-expansive map $f\colon [0,1]^V \to [0,1]^V$ given by

$$f(p)(v) = \begin{cases} \min_{u\in\eta(v)} p(u) & \text{if } v \in V_{\min} \\ \max_{u\in\eta(v)} p(u) & \text{if } v \in V_{\max} \\ \sum_{u\in V} \eta_{\mathrm{av}}(v)(u)\cdot p(u) & \text{if } v \in V_{\mathrm{av}} \\ w(v) & \text{if } v \in V_{\mathrm{sink}} \end{cases}$$

As above we assume that the transition probabilities of the average nodes are not known precisely but can only be approximated by sampling. Formally, for each average node $v \in V_{\mathrm{av}}$ we have a collection of independent and identically distributed random variables $(X_n^v)_{n\in\mathbb{N}}$ with values in V and distribution $\mathbb{P}[X_n^v = u] = \eta_{\mathrm{av}}(v)(u)$ for all $u \in V$, $n \in \mathbb{N}$. For all n we can estimate the true transition probabilities by the relative frequency

$$(\eta_{\mathrm{av}})_n(v)(u) = \frac{|\{1 \le k \le n \mid X_k^v = u\}|}{n}$$

If we denote by f_n the fixpoint function of the SSG based on $(\eta_{\mathrm{av}})_n$, we have

$$\|f - f_n\| = \max_{v\in V_{\mathrm{av}}} \max_{u\in V} |\eta_{\mathrm{av}}(v)(u) - (\eta_{\mathrm{av}})_n(v)(u)|$$

In particular, by the law of large numbers the sequence of maps (f_n) converges to the correct map f with probability 1. But more importantly – using Bernstein's inequality or similar estimates – we can, for every $\delta, \varepsilon > 0$ effectively produce an index $n \in \mathbb{N}$ such that $\mathbb{P}[\|f - f_n\| \ge \delta] \le \varepsilon$. Using Algorithm 1 we can therefore iterate to the value of an SSG without knowing the true transition probabilities by simply using approximations obtained from sampling for each value update provided that sampling and value updates are interleaved in a suitable way.

In Section 5 we have seen by a careful analysis that for the special case of Bellman objective functions we do not have to pay attention to the way in which sampling and value updates are interleaved. Here, on the other hand, we showed that our methods are flexible enough to be applicable to many other problems which can be phrased as computing least fixpoints of monotone and non-expansive functions. Along the lines of the example presented here, it is not strictly necessary to redo an analysis of the new objective function as we did in Section 5. Rather, it is sufficient to be able to effectively compute probabilistic error bounds for Algorithm 1 to work.

F Numerical Experiments

While our main goal was to merge the existing theory behind the dampened Mann iteration and approximative settings such as model-based reinforcement learning, with a focus on formal proofs of convergence, in this section, we want to gain some practice-based intuition about how a dampened Mann iteration approximates the least fixpoints and how it compares to classical Kleene or (undampened) Mann iteration.

Here, an iteration scheme $\mathcal{S} = ((\alpha_n), (\beta_n))$ with $\alpha_n = 0$ will be called a *Kleene* scheme, otherwise (mostly, choosing $\alpha_n = 1/n$) a *Mann* scheme. It is called *undampened* if $\beta_n \equiv 0$ and *dampened* otherwise (again, choosing $\beta_n = 1/n$ in general). Furthermore, we will consider schemes *with resetting*, which iterate a single approximation from an initial point, i.e., $x_n^{\mathcal{F}} = T_n(f_n, T_{n-1}(f_n, \ldots T_0(f_n, x) \ldots))$ without reusing previous results. We will typically start with $x = 0$ in a resetting iteration.

The experiments were performed on an Intel i7-11850H processor with 16GB of RAM.

Known function. First of all, the dampened Mann iteration does not guarantee, in any way, faster convergence when compared to classical Kleene iteration. Its benefits lie in a more general applicability not in a better efficiency. In fact, as to be expected when the exact function is known and the iteration starts at 0, we can see both in a simple example using the function

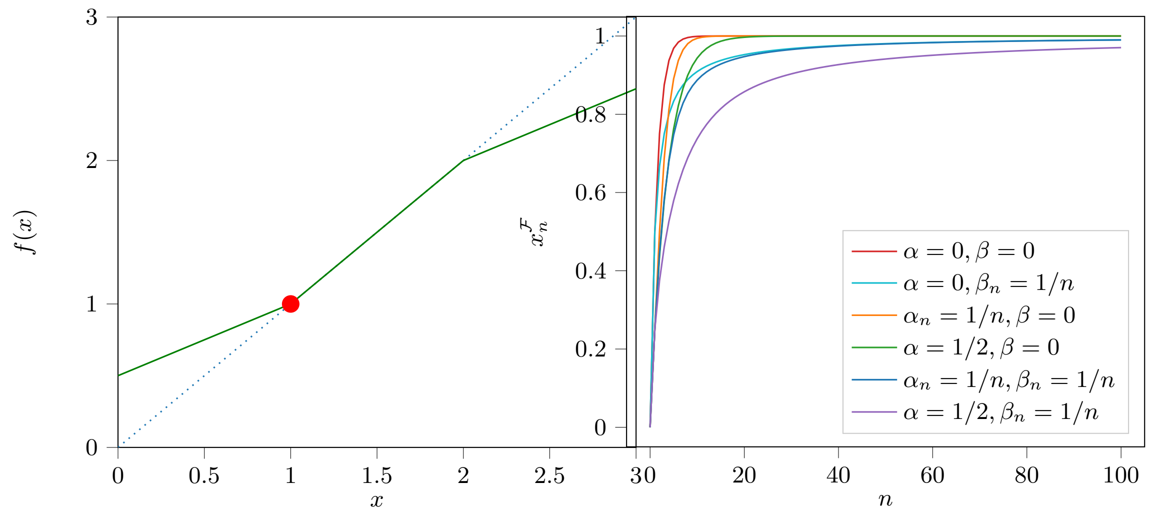
$$f(x) = \min(\max({}^1\!/\!_2 x + {}^1\!/\!_2, x), {}^1\!/\!_2 x + 1)$$

depicted in Figure 3a, as well as a more complex example using the (state-value) Bellman operator of the Markov decision process depicted in Figure 4, that Kleene iteration converges faster than iteration schemes using positive parameters α_n, β_n (Figures 3b and 5a).

On the other hand, as we already discussed in Section 3, positive dampening factors allow the iteration to converge to the least fixpoint for *any* initial point whereas it is easy to check that both (undampened) Kleene and undampened Mann iterations, in general, do not converge or get stuck at another fixpoint (e.g., take the flip map $f(x, y) = (y, x)$ with initial point $x_0 = (1, 0)$).

Unsurprisingly, this happens also when considering more complex functions. Note that the Markov decision process in Figure 4 is not simple whence its Bellman operator is not a power contraction. Here, starting from an initial point that overestimates the state-values in end-components, no undampened iteration converges to the least fixpoint. Indeed, the Kleene iteration gets stuck in a loop while the undampened Mann iterations converge to an overapproximation of the true optimal value function. The dampened versions, however, do converge to the least fixpoint (see Figure 5b).

Unknown function. Now, the main focus of this paper was a setting in which we do not have access to the exact function. In this approximative setting, we



(a) The simple piecewise linear function f (green) with its least fixpoint $\mu f = 1$ (red)

(b) Generated sequences $(x_n^{\mathcal{F}})$ of the different iteration schemes (starting from 0)

Fig. 3: Results of different iteration schemes for an exact function f.

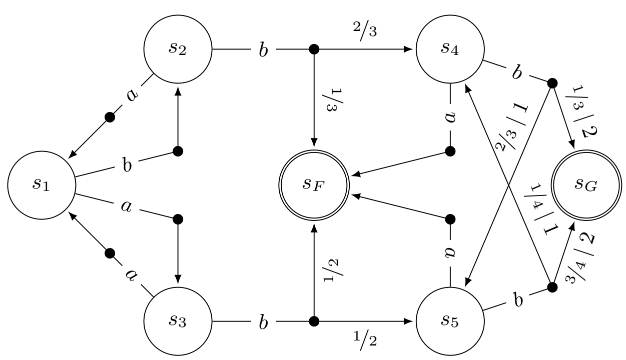
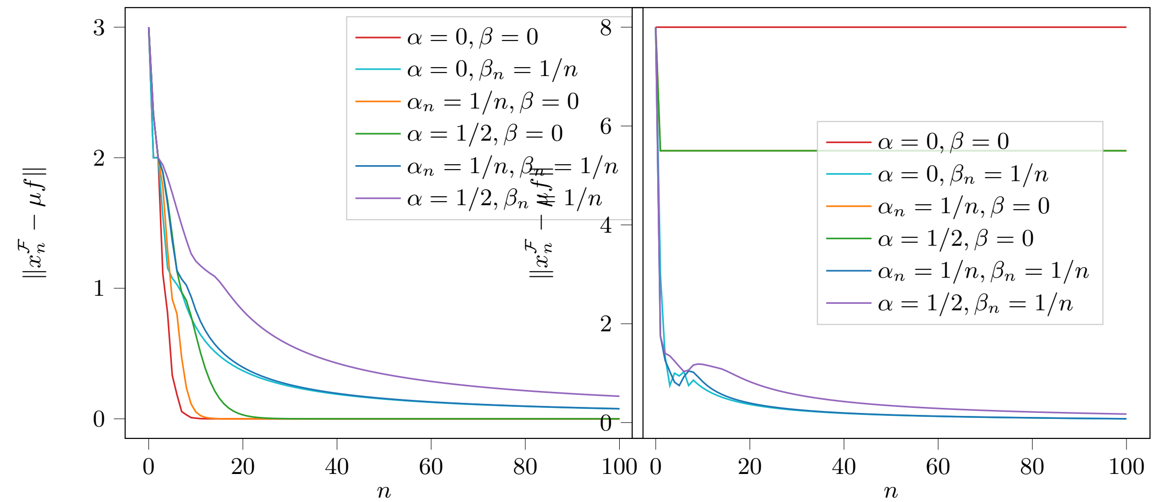


Fig. 4: An MDP with one maximal end-component (s_1, s_2, and s_3) where unspecified transition probabilities are 1 and unspecified rewards are 0.



(a) Generated sequences $(x_n^{\mathcal{F}})$ with initial value $v_0 = 0$.

(b) Generated sequences $(x_n^{\mathcal{F}})$ with initial value $v_0 = (10, 5, 4, 3, 2, 1, 0)$.

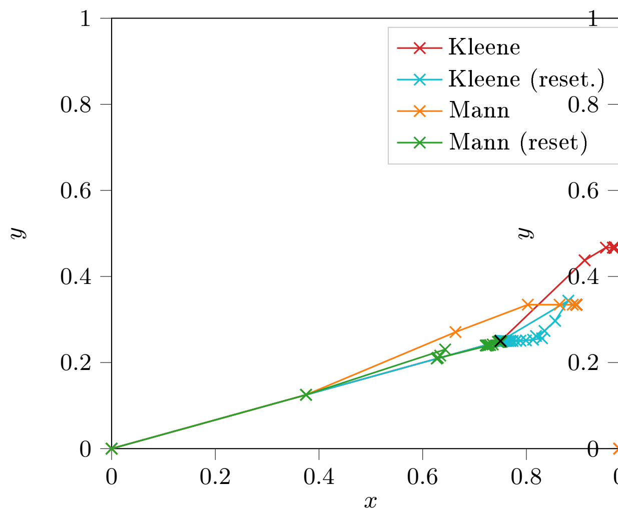
Fig. 5: Results of different iteration schemes for the exact Bellman operator of the MDP in Figure 4.

already mentioned in Section 4 that the considered undampened iterations can only guarantee convergence to the least fixpoint with resetting, i.e., starting the iteration anew from the bottom 0 with each new approximation f_n (with carefully chosen number of iterations if necessary). We see this in Figure 6 for the relatively simple monotone non-expansive function sequence

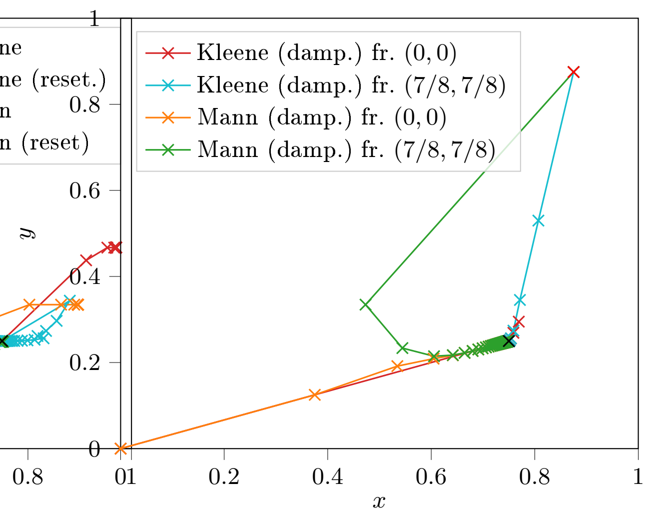
$$f_n : [0,1]^2 \to [0,1]^2, x \mapsto \max(x, (1-a) \cdot x^n + a) \tag{12}$$

where the least fixpoint $\mu f = a$ of the limit function $f(x) = \max(x, a)$ was chosen to be $(3/4, 1/2)$. It is clearly visible that undampened iteration without resetting cannot reach the least fixpoint, even when starting from 0 (red and orange lines in Figure 6a with Kleene iteration converging to $\sim (0.95, 0.5)$ and Mann iteration to $\sim (0.9, 0.35)$). Versions with resetting as well as dampened versions, on the other hand, do converge to the least fixpoint, the latter even when starting from an overapproximation such as $(7/8, 7/8)$ (cf. Figure 6b). This also serves as a visual representation of how the dampened versions approach the fixpoint.

Lastly, we compare the performance of dampened iterations to resetting iterations on randomly generated MDPs. Figure 7a shows the results of performing a dampened Mann iteration and an undampened Kleene iteration with resets on sampled (state-value)-Bellman operators of one hundred randomly generated normalised MDPs with 50 states (in equal amounts simple Markov chains, MCs with five end-component, as well as simple MDPs and MDPs with five MECs) for



(a) Generated sequences $(x_n^{\mathcal{F}})$ on the function sequence (12) without dampening.

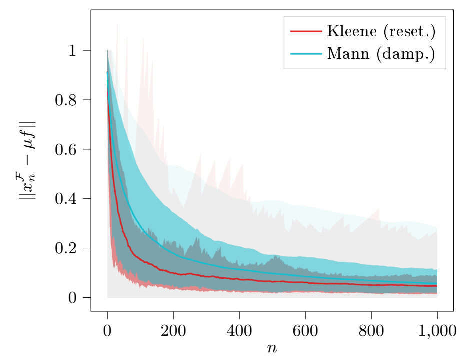


(b) Generated sequences $(x_n^{\mathcal{F}})$ on the function sequence (12) with dampening starting from $x_1 = (0, 0)$ and $x_2 = (7/8, 7/8)$ (marked as a violet cross), respectively.

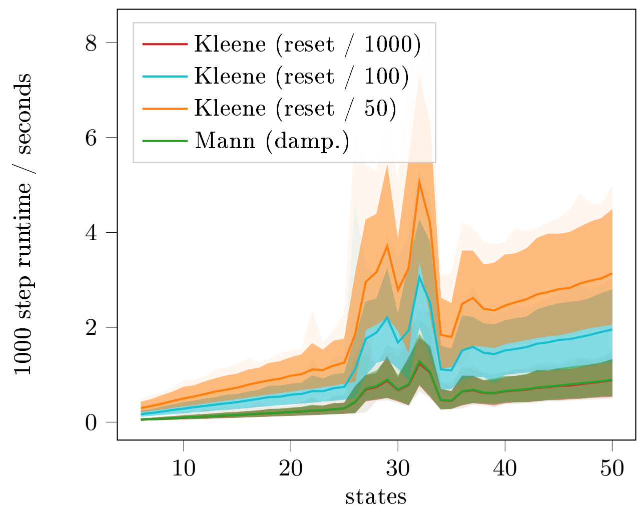
Fig. 6: Sequences $(x_n^{\mathcal{F}})$ generated by different iteration iterations on the function sequence (12). The fixpoint μf of the limit function is marked as a black cross.

which Kleene iteration on the exact Bellman operator took less than 1000 steps to stop (we stop when the update $< 10^{-6}$). The lines denoting the mean error of the value approximations suggest that Kleene iteration with resets performs better overall, but not significantly better. The areas showing the 90th percentile and maximum, on the other hand, illustrate that Kleene iteration with resets is more dependent on the quality of the currently sampled MDP (with error spikes at points where the MDP were badly sampled) while the dampened Mann iteration reduces the variance introduced by random sampling.

The main advantage of dampened iterations, when compared to resetting iterations, is their ability to reuse knowledge gained from old approximations. In particular, one step $n \mapsto n+1$ only takes one application of $T_n^{\mathcal{S}}(f_n, \cdot)$ while it takes $n+1$ applications of $T_k^{\mathcal{S}}(f_n, \cdot)$ in a Kleene iteration with resets. In practice, it might be unnecessary to compute every $x_n^{\mathcal{F}}$ in a resetting iteration. In an offline scenario, it might be sufficient to just compute $x_n^{\mathcal{F}}$ for one n (i.e., performing Kleene iteration on one approximation f_n), but reinforcement learning is most often treated in an online setting where live updates to the values are required as new samples are generated. Furthermore, even in an offline scenario, the above results showcase that the results of resetting schemes vary highly with the quality of the current approximation, whence relying on more than a single approximation f_n might produce better results. In the context of resetting schemes, this, of course, comes at a high cost of required runtime as is clearly visible in Figure 7b, comparing the runtime of a dampened Mann iteration over 1000 approximation steps with Kleene iteration on f_{1000} (which is slightly but not significantly faster; note that the corresponding red line is almost covered by the green line) and with Kleene iteration resetting every 100 or every 50 steps (resulting in 10 and 20 performed iterations), respectively, on one hundred randomly generated MDPs per number of states (again half of which were simple and again half of simple and non-simple MDPs were Markov chains). The variance in runtime for each iteration scheme was mainly a result of the number of actions in the MDP (with Markov chains resulting in the lowest runtime).



(a) Errors of resetting Kleene and dampened Mann scheme on 100 randomly generated MDPs of value 1. The lines denote the mean error, the half-opaque area the 10th to 90th percentile, and the weakly visible area the minimum to maximum error.



(b) Runtime of resetting and dampened Mann scheme on 100 randomly generated MDPs per number of states (6 – 50). The lines denote the mean runtime, the area the minimum to maximum runtime for the given scheme. Note that the red line is hidden beneath the green one due to overlap.

Fig. 7: Results on sampled Bellman operators of randomly generated MDPs.