On Policy Evaluation Algorithms in Distributional Reinforcement Learning

Julian Gerstenberg GERSTENB@MATH.UNI-FRANKFURT.DE
Institute of Mathematics
Goethe University Frankfurt
Frankfurt am Main, Germany

Ralph Neininger NEININGR@MATH.UNI-FRANKFURT.DE
Institute of Mathematics
Goethe University Frankfurt
Frankfurt am Main, Germany

Denis Spiegel SPIEGEL@MATH.UNI-FRANKFURT.DE
Institute of Mathematics
Goethe University Frankfurt
Frankfurt am Main, Germany

Abstract

We introduce a novel class of algorithms to efficiently approximate the unknown return distributions in policy evaluation problems from distributional reinforcement learning (DRL). The proposed distributional dynamic programming algorithms are suitable for underlying Markov decision processes (MDPs) having an arbitrary probabilistic reward mechanism, including continuous reward distributions with unbounded support being potentially heavy-tailed.

For a plain instance of our proposed class of algorithms we prove error bounds, both within Wasserstein and Kolmogorov–Smirnov distances. Furthermore, for return distributions having probability density functions the algorithms yield approximations for these densities; error bounds are given within supremum norm. We introduce the concept of quantile-spline discretizations to come up with algorithms showing promising results in simulation experiments.

While the performance of our algorithms can rigorously be analysed they can be seen as universal black box algorithms applicable to a large class of MDPs. We also derive new properties of probability metrics commonly used in DRL on which our quantitative analysis is based.

Keywords: distributional reinforcement learning, distributional dynamic programming, distributional Bellman operator, Markov decision process, probability metrics

1 Introduction

In Reinforcement Learning (RL) agents take actions in an environment in order to maximize a cumulative random reward, called return, with respect to its expectation. It has long been also of interest to optimize other characteristics of the probability distribution of the return than its expectation. In recent years, a systematic study of the practical benefits and theoretical foundations of optimizing the return distribution has been taken up; for a comprehensive account of this direction, called Distributional Reinforcement Learning (DRL), see the recently published book Bellemare et al. (2023). In the present paper we

focus on *Distributional Dynamic Programming* (DDP) algorithms for distributional policy evaluation, see Chapter 5 of Bellemare et al. (2023), i.e., on the construction and analysis of algorithms which yield approximations of the return distribution of a Markov decision process (MDP) when a fixed policy is being considered.

A crucial ingredient of any MDP is its reward mechanism, which specifies the rewards given to an agent for state-action-state transitions. Such a mechanism implies return distributions which are characterised as fixed-points of the Distributional Bellman Operator (DBO), we recall key results in Section 2. DDP algorithms have typically been developed and analysed for MDPs for which all reward distributions are *finitely supported*. A probability distribution μ on $\mathbb{R}$ is finitely supported if and only if it has a *representation* $\mu = \sum_{i=1}^{m} p_i \delta_{x_i}$ with $m \in \mathbb{N}$, $p_i \geq 0$ probabilities summing to 1, *support points* $x_i \in \mathbb{R}$ and δ_x the Dirac measure in $x \in \mathbb{R}$. Finitely supported distributions are also called *empirical* in DRL literature. The state of the art for DDP algorithms with finitely supported reward distributions is covered by Chapter 5 of Bellemare et al. (2023). Roughly, most of these algorithms operate by iterations of the DBO together with a projection step, which maps a distribution to a fixed-size (i.e., fixed m) finitely supported distribution, see also Bellemare et al. (2017); Maddison et al. (2017); Dabney et al. (2018); Rowland et al. (2018).

The aim of the present paper is to develop and analyse DDP algorithms that are universally applicable for a wide class of MDPs, including cases of MDPs in which (some of the) reward distributions are not finitely supported, including distributions that have unbounded support and/or are continuous and/or are heavy-tailed. Such MDPs are of interest in applications, for example, in pricing and trading in insurance and finance, see Krasheninnikova et al. (2019), Kolm and Ritter (2020) and the recent survey Hambly et al. (2023).

To be specific, our DDP algorithms are applicable to any MDP for which cumulative distribution functions (CDFs) as well as quantile functions (QFs) of its reward distributions can be evaluated efficiently, e.g., if all reward distributions are contained in well-known distribution families for which methods to evaluate CDF and QF are efficiently implemented in popular software packages, e.g., in `R` or the `Python`-package `SciPy`.

The DDP algorithms we are proposing in Section 3 also operate by iterations of the DBO together with projection steps that map arbitrary distributions to finitely supported ones. To take up an idea of Devroye and Neininger (2002), where fixed-points of related recursive distributional equations were approximated, we allow *varying projections*, i.e. projections which depend on the update step. In particular, as Devroye and Neininger (2002), we allow the sizes of the finitely supported representations to increase with each iteration. In the context of the approximation of certain perpetuities this idea has also been applied, see Knape and Neininger (2008). Furthermore, we allow projections to depend on *previously calculated approximations*, which also seems to generalise previous approaches. A general framework for our family of algorithms is presented in Section 3.

From Section 4 on we investigate the complexity and performance of our algorithms. First, in Section 4 bounds are derived in an abstract setting for our general class of DDP algorithms where we also address how to choose our varying projections, in particular how to let the number of support points grow. We quantify the quality of the calculated approximations by use of *analysis metrics*, such as Wasserstein distances. Our setting and analysis does not require the projections to be non-expansive with respect to the analysis metrics, cf. Chapters 4 and 5 in Bellemare et al. (2023) and the discussion in our Section

4.2. Later, in Section 6, which is also of independent interest, we explain how to extend these bounds to the Kolmogorov–Smirnov distance and how to obtain approximations of densities of return distributions when they exist together with uniform error bounds; we also offer a sufficient criteria for the return distributions to have densities at all and discuss the relation to nonuniform random number generation in Remark 22.

Concrete universal DDP algorithms are defined and analysed in Section 5. They are universal in the sense that they apply to the whole class of MDPs for which the CDFs and QFs of all reward distributions can be efficiently evaluated, regardless of their nature (discrete, continuous, unbounded support, heavy-tailed, etc.). Our projections are constructed such that they only require that CDFs of the reward distributions can be evaluated. We discuss the trade-off between space-time complexity and approximation quality for a plain instance of the algorithms in Section 5.1. This yields further evidence that one should let the finitely supported representation sizes increase with each iteration. For practical purpose, we suggest to use the universal DDP algorithms presented in Sections 5.2 and 5.3, which are based on bounding quantiles of convolutions, respextively approximating quantiles by linear-spline interpolations. We conclude with a small simulation study in Section 5.4 to demonstrate that our algorithms outperform simple Monte-Carlo estimation techniques significantly.

Our algorithms are reasonably applicable for small finite state-action spaces. For larger spaces methods based on Markov chains exploring the state space and temporal difference learning have been developed, see Chapter 6 of Bellemare et al. (2023) and Morimura et al. (2010a) for a popular particle smoothing approach within Markov chain techniques.

1.1 Notations

Let $\mathcal{B}(\mathbb{R})$ be the Borel σ-field on $\mathbb{R}$ and $\mathscr{P}(\mathbb{R})$ be the set of probability measures on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Let $\mathscr{P}_{\mathrm{fin}}(\mathbb{R}) \subsetneq \mathscr{P}(\mathbb{R})$ be the subset of finitely supported distributions. Every $\mu \in \mathscr{P}(\mathbb{R})$ is specified by its cumulative distribution function (CDF) $F_\mu : \mathbb{R} \to [0,1]$ as well by its quantile function (QF) $F_\mu^{-1} : (0,1) \to \mathbb{R}$ defined as

$$F_\mu(x) = \mu\big((-\infty, x]\big) \text{ and } F_\mu^{-1}(u) = \inf\{x \in \mathbb{R} \mid F_\mu(x) \geq u\},$$

related by $u \leq F_\mu(x) \Leftrightarrow F_\mu^{-1}(u) \leq x$. The set $\mathscr{P}(\mathbb{R})$ is closed under convex combinations, which transfers to the CDF: $F_{q\mu+(1-q)\nu} = qF_\mu + (1-q)F_\nu$ holds for all $q \in [0,1], \mu, \nu \in \mathscr{P}(\mathbb{R})$. Convex combinations do not transfer to the QF. A distribution $\mu \in \mathscr{P}(\mathbb{R})$ is said to be continuous if F_μ is continuous and is said to possess the *probability density function* (PDF) $f_\mu : \mathbb{R} \to [0, \infty]$ if $F_\mu(x) = \int_{-\infty}^x f_\mu(y)\mathrm{d}y$ holds for all $x \in \mathbb{R}$.

It is convenient to work with random variables: we write $\mathcal{L}(X) = \mathbb{P}[X \in \cdot\,] \in \mathscr{P}(\mathbb{R})$ for the distribution of a real-valued random variable (RV) X defined on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$. It is $\mathcal{L}(X) = \mu$ if and only if $F_\mu(x) = \mathbb{P}[X \leq x]$ for all $x \in \mathbb{R}$ and a random variable X with $\mathcal{L}(X) = \mu$ is given by $X := F_\mu^{-1}(U)$, where U is a RV continuous uniform on $(0,1)$. If it exists, let $\mathbb{E}[X] \in [-\infty, \infty]$ be the expectation of X. If $\mathbb{P}[X \geq 0] = 1$, then $\mathbb{E}[X] \in [0, \infty]$ exists. For $\alpha \in (0, \infty)$ let $\mathscr{P}_\alpha(\mathbb{R}) \subsetneq \mathscr{P}(\mathbb{R})$ be the set of distributions $\mu = \mathcal{L}(X)$ with finite α-moment $\mathbb{E}[|X|^\alpha] < \infty$. Note that $\mathscr{P}_{\mathrm{fin}}(\mathbb{R}) \subsetneq \mathscr{P}_\alpha(\mathbb{R})$ and that $\beta < \alpha$ implies $\mathscr{P}_\alpha(\mathbb{R}) \subsetneq \mathscr{P}_\beta(\mathbb{R})$. For an event $F \in \mathcal{F}$ with $\mathbb{P}[F] > 0$ let $\mathcal{L}(X|F) = \mathbb{P}[X \in \cdot\,|F] \in \mathscr{P}(\mathbb{R})$ be the conditional distribution and $\mathbb{E}[X|F] = \mathbb{E}[X1_F]/\mathbb{P}[F]$ the conditional expectation of X given F has occurred.

Finally, for sets A and I it will be convenient to write elements of A^I as $[a_i : i \in I] \in A^I$. For real numbers $x, y \in \mathbb{R}$ we write $x \vee y = \max\{x, y\}$ and $x \wedge y = \min\{x, y\}$. We use the Bachmann–Landau big-O notation.

2 Distributional Bellman operator and dynamic programming

A Markov Decision Process (MDP) with a fixed stationary policy is a tupel

$$\mathfrak{mdp} := (\mathcal{S}, \mathcal{A}, p, \pi, \gamma, r)$$

with finite set of states $\mathcal{S} \ni s, \bar{s}$, finite set of actions $\mathcal{A} \ni a$, state-action-state transition probabilities $p = [p(\bar{s}|s, a) : (s, a, \bar{s}) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}]$, in which $p(\bar{s}|s, a) \in [0, 1]$ is the probability to make a transition from s to $\bar{s}$ when performing action a, a fixed stationary policy $\pi = [\pi(a|s) : (s, a) \in \mathcal{S} \times \mathcal{A}]$ in which $\pi(a|s) \in [0, 1]$ is the probability of the agent choosing action a when in state s, a fixed discount factor $\gamma \in (0, 1)$ and $r = [r(\,\cdot\,|s, a, \bar{s}) : (s, a, \bar{s}) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}]$ an arbitrary probabilistic reward mechanism, in which

$$r(\,\cdot\,|s, a, \bar{s}) = \mathcal{L}(R_{sa\bar{s}}) \in \mathscr{P}(\mathbb{R})$$

is a probability distribution modelling the distribution of the immediate and possibly random $\mathbb{R}$-valued reward $R_{sa\bar{s}}$ given to the agent for a transition from state s to $\bar{s}$ having used action a. Let $F_{sa\bar{s}}$ be the CDF and $F^{-1}_{sa\bar{s}} : (0, 1) \to \mathbb{R}$ be the QF of $r(\,\cdot\,|s, a, \bar{s}) = \mathcal{L}(R_{sa\bar{s}})$.

The ingredients of $\mathfrak{mdp}$ give rise to the following dynamical process: at time $t = 0$ the agent is in a uniform random initial state $S^{(0)} \in \mathcal{S}$. If at time $t \in \mathbb{N}_0$ the agent is in state $S^{(t)}$, conditionally on $S^{(t)}$ and the past, the agent chooses a random action $A^{(t)}$ distributed as $\pi(\,\cdot\,|S^{(t)})$. The environment reacts with a random successor state $S^{(t+1)}$ distributed as $p(\,\cdot\,|S^{(t)}, A^{(t)})$. The agent receives an immediate real-valued random reward $R^{(t)}$ distributed as $r(\,\cdot\,|S^{(t)}, A^{(t)}, S^{(t+1)})$. The resulting stochastic process

$$(S^{(t)}, A^{(t)}, R^{(t)})_{t \in \mathbb{N}_0}$$

is called *full MDP*, and the *return random variable* is defined as

$$G^* = \sum\nolimits_{t=0}^{\infty} \gamma^t R^{(t)} = R^{(0)} + \gamma R^{(1)} + \gamma^2 R^{(2)} + \dots$$

which exists with probability one as a $\mathbb{R}$-valued random variable under the assumption

(A1) $\mathbb{E}[\log(1 \vee |R_{sa\bar{s}}|)] < \infty$ for all $s, a, \bar{s}$,

we refer to Gerstenberg et al. (2023) for a precise result. We assume (A1) holds throughout the paper, but consider stronger moment assumptions on reward distributions later, cf. Remark 4. Of interest in distributional policy evaluation is the distribution of G^* when starting the full MDP in a fixed but arbitrary given state $s \in \mathcal{S}$. The (state-)return distribution $\eta^* \in \mathscr{P}(\mathbb{R})^{\mathcal{S}}$ is the $\mathcal{S}$-indexed collection of all these conditional distributions:

$$\eta^* = [\eta^*_s : s \in \mathcal{S}] \quad \text{with} \quad \eta^*_s = \mathcal{L}(G^*|S^{(0)} = s).$$

We call η^*_s the s-th component of η^*. The return distribution η^* can rarely by described analytically and the task of distributional policy evaluation algorithms is to approximate η^*. We discuss algorithms that are applicable in practice under the following two assumptions on the ingredients of $\mathfrak{mdp}$:

(A2) $F_{sa\bar{s}}(x), F^{-1}_{sa\bar{s}}(u)$ can be calculated in constant time for all $s, a, \bar{s}$ and $x \in \mathbb{R}, u \in (0,1)$,

(A3) $p(\bar{s}|s,a), \pi(a|s) \in [0,1]$ are known for all $s, a, \bar{s}$.

Algorithms to approximate η^* can be derived from the fact that η^* is the unique fixed-point of the *Distributional Bellman Operator* (DBO), which is introduced next. In the following, let $\eta = [\eta_s : s \in \mathcal{S}] = [\mathcal{L}(G_s) : s \in \mathcal{S}] \in \mathscr{P}(\mathbb{R})^{\mathcal{S}}$ be an arbitrary $\mathcal{S}$-indexed collection of distributions and assume that the three collections of random variables $(G_s)_s, (R_{sa\bar{s}})_{sa\bar{s}}$ and the full MDP $(S^{(t)}, A^{(t)}, R^{(t)})_t$ are defined on a common probability space and are independent.

Definition 1 *The DBO is the map $\mathcal{T} : \mathscr{P}(\mathbb{R})^{\mathcal{S}} \to \mathscr{P}(\mathbb{R})^{\mathcal{S}},\ \eta \mapsto \mathcal{T}\eta = [\mathcal{T}_s\eta : s \in \mathcal{S}]$ defined by its component functions $\mathcal{T}_s : \mathscr{P}(\mathbb{R})^{\mathcal{S}} \to \mathscr{P}(\mathbb{R})$ via $\mathcal{T}_s\eta := \mathcal{L}(R^{(0)} + \gamma G_{S^{(1)}} | S^{(0)} = s)$.*

Theorem 2 *Under (A1) it holds that*

(a) η^* is the unique fixed-point of $\mathcal{T}$, that is for every $\eta \in \mathscr{P}(\mathbb{R})^{\mathcal{S}}$ it is $\eta = \eta^*$ if and only if η solves the so-called distributional Bellman equation $\mathcal{T}\eta = \eta$,

(b) for every $\eta \in \mathscr{P}(\mathbb{R})^{\mathcal{S}}$ it holds $\mathcal{T}^{\circ n}\eta \to_w \eta^*$, where $\mathcal{T}^{\circ n} = \mathcal{T} \circ \cdots \circ \mathcal{T}$ is the n-th iteration of $\mathcal{T}$ and $\to_w$ denotes (component-wise) weak convergence of distributions.

Proof In case $\mathbb{E}[|R_{sa\bar{s}}|] < \infty$ for all $s, a, \bar{s}$ this result if well-known and follows, for example, from Proposition 4.34 in Bellemare et al. (2023). Under the more general assumption (A1) it follows from Theorem 1 in Gerstenberg et al. (2023). ■

Theorem 2(b) leads to a theoretical approximation algorithm for η^*: choose an initial approximation $\eta^{(0)}$ and calculate a sequence $\eta^{(n)}, n \in \mathbb{N}$ of approximations by the inductive update rule $\eta^{(n)} = \mathcal{T}\eta^{(n-1)}$. However, this is not practical for applications, as we explain next. First, by conditioning on $\{A^{(0)} = a, S^{(1)} = \bar{s}\}$, the s-th component of $\mathcal{T}\eta$ equals

$$\mathcal{T}_s\eta = \sum\nolimits_{(a,\bar{s}) \in \mathcal{A}\times\mathcal{S}} \pi(a|s) p(\bar{s}|s,a) \mathcal{L}\left(R_{sa\bar{s}} + \gamma G_{\bar{s}}\right). \tag{1}$$

The convolutions $\mathcal{L}\left(R_{sa\bar{s}} + \gamma G_{\bar{s}}\right)$ also equal (measure theoretic) convex combinations:

$$\mathcal{L}\left(R_{sa\bar{s}} + \gamma G_{\bar{s}}\right) = \int_{\mathbb{R}} \mathcal{L}\left(R_{sa\bar{s}} + \gamma z\right) \mathrm{d}\eta_{\bar{s}}(z) = \int_{\mathbb{R}} \mathcal{L}\left(y + \gamma G_{\bar{s}}\right) \mathrm{d}r(y|s,a,\bar{s}). \tag{2}$$

Now, assume all components of η are finitely supported with $m \in \mathbb{N}$ particles, say $\eta_s = \sum_{i=1}^m w_{si}\delta_{z_{si}}$. Combining (1) and (2) gives

$$\mathcal{T}_s\eta = \sum\nolimits_{(a,\bar{s}) \in \mathcal{A}\times\mathcal{S}} \sum\nolimits_{i=1}^m \pi(a|s) p(\bar{s}|s,a) w_{\bar{s}i} \mathcal{L}\left(R_{sa\bar{s}} + \gamma z_{\bar{s}i}\right). \tag{3}$$

Classical DDP algorithms are developed and analysed under the assumption

(A2.Fin): $r(\,\cdot\,|s,a,\bar{s}) \in \mathscr{P}_{\text{fin}}(\mathbb{R})$ is finitely supported for all $s, a, \bar{s}$,

which implies (A2). If (A2.Fin) holds, that is every reward distribution is finitely supported, say of size $N \in \mathbb{N}$ with $\mathcal{L}(R_{sa\bar{s}}) = \sum_{l=1}^{N} q_{sa\bar{s}l}\delta_{r_{sa\bar{s}l}}$, formula (3) yields

$$\mathcal{T}_s\eta = \sum\nolimits_{(a,\bar{s})\in\mathcal{A}\times\mathcal{S}} \sum\nolimits_{i=1}^{m} \sum\nolimits_{l=1}^{N} \pi(a|s)p(\bar{s}|s,a)w_{\bar{s}i}q_{sa\bar{s}l}\delta_{r_{sa\bar{s}l}+\gamma z_{\bar{s}i}}, \tag{4}$$

which is finitely supported of size (at most) $m \cdot |\mathcal{A}| \cdot |\mathcal{S}| \cdot N$ and a representation can be calculated with a space-time complexity of order $\mathrm{O}(m \cdot |\mathcal{A}| \cdot |\mathcal{S}| \cdot N)$, see Algorithm 5.1 in Bellemare et al. (2023). As a consequence, if $\eta^{(0)}$ has finitely supported components of size m, inductively calculating the n-th approximation $\eta^{(n)} = \mathcal{T}\eta^{(n-1)}$ by applying (4) to $\eta = \eta^{(n-1)}$ is possible in theory, but has a space-time complexity of order $\mathrm{O}(m \cdot |\mathcal{A}|^n \cdot |\mathcal{S}|^n \cdot N^n)$, which is prohibitive for applications.

If only (A2) holds, the situation is even more complicated: (3) shows that $\mathcal{T}_s\eta$ can be an intricate mixture involving continuous distributions and/or distributions with unbounded support, in particular it is, in general, no longer finitely supported. However, (3) also shows that the CDF of $\mathcal{T}_s\eta$ at $x \in \mathbb{R}$ is given by

$$F_{\mathcal{T}_s\eta}(x) = \sum\nolimits_{(a,\bar{s})\in\mathcal{A}\times\mathcal{S}} \sum\nolimits_{i=1}^{m} \pi(a|s)p(\bar{s}|s,a)w_{\bar{s}i}F_{sa\bar{s}}\left(x - \gamma z_{\bar{s}i}\right), \tag{5}$$

which can be calculated explicitly under (A2) with a time-complexity of order $\mathrm{O}(m \cdot |\mathcal{A}| \cdot |\mathcal{S}|)$. However, calculating the CDF of even the second iterate $\mathcal{T}^{\circ 2}\eta$ is no longer possible exactly.

The idea of existing DDP algorithms that work under (A2.FIN) is to prevent the above mentioned space-time complexity blow-up by choosing a *projection map* $\Pi : \mathscr{P}(\mathbb{R}) \to \mathscr{P}(\mathbb{R})$ that maps any $\mu \in \mathscr{P}(\mathbb{R})$ to a finitely supported distribution $\Pi(\mu)$ having a *fixed-size* m, and to use the update rule $\eta^{(n)} = \bar{\Pi}\mathcal{T}\eta^{(n-1)}$, where $\bar{\Pi}$ is the component-wise extension of Π to a map $\mathscr{P}(\mathbb{R})^{\mathcal{S}} \to \mathscr{P}(\mathbb{R})^{\mathcal{S}}$. Assuming $\Pi(\mu)$ can be calculated for any finitely supported μ within controllable space-time-complexity, this leads to useful algorithms, which are discussed and analysed in Chapter 5 of Bellemare et al. (2023). Many of these algorithms are not directly applicable under (A2), as the used projections can often not easily be evaluated for distributions which are not finitely supported. Our approach is to focus on a special class of (parameterised) projections $\Pi(\cdot, \xi)$ in which $\Pi(\mu, \xi)$ depends on μ only by a controllable number of evaluations of its CDF.

Remark 3 (State-Action return distributions) *The so-called* state-action*-return distribution is defined as the $(\mathcal{S} \times \mathcal{A})$-indexed collection $\zeta^* = [\zeta^*_{sa} : (s,a) \in \mathcal{S} \times \mathcal{A}]$ with $\zeta^*_{sa} = \mathcal{L}(G^*|S^{(0)} = s, A^{(0)} = a)$. For convenience, we only discuss state-return distributions, but everything could be formulated analogously for the state-action case. This restriction is a recurring theme in the field of distributional policy evaluation, see Section 4 of Morimura et al. (2010b) or Chapter 5 of Bellemare et al. (2023).*

Remark 4 (Moment Assumptions) *Each of the following (parameterised) moment assumptions on reward distributions implies (A1):*[1]

(A1.Pα), $\alpha \in (0,\infty)$: $\mathbb{E}[|R_{sa\bar{s}}|^\alpha] < \infty$ *for all $s, a, \bar{s}$,*

1. In the notations (A1.Pα), (A1.Eλ), (A1.B) the P indicates polynomial tails, the E exponential tails and B bounded support, where α and λ specify the degree and rate respectively.

(A1.Eλ), $\lambda \in (0,\infty)$: $\mathbb{E}[\exp(\lambda|R_{sa\bar{s}}|)] < \infty$ *for all* $s, a, \bar{s}$,

(A1.B) : *there is* $[r_{\min}, r_{\max}] \subset \mathbb{R}$ *with* $\mathbb{P}\big[r_{\min} \leq R_{sa\bar{s}} \leq r_{\max}\big] = 1$ *for all* $s, a, \bar{s}$.

Condition (A1.B) implies (A1.Eλ) for every λ, (A1.Eλ) implies (A1.Pα) for every α and (A1.Pα) implies (A1). The bounded rewards case (A1.B) is frequently encountered in the (distributional) reinforcement learning literature, as well as (A1.Pα) with α = 1. All of these moment assumptions transfer to the return distributions, let $\eta_s^* = \mathcal{L}(G_s^*)$*: (A1.Pα) implies* $\mathbb{E}[|G_s^*|^\alpha] < \infty$ *for all s, (A1.Eλ) implies* $\mathbb{E}[\exp(\lambda|G_s^*|)] < \infty$ *for all s and (A1.B) via bounding interval* $[r_{\min}, r_{\max}] \subset \mathbb{R}$ *implies* $\mathbb{P}[v_{\min} \leq G_s^* \leq v_{\max}] = 1$ *for all s with bounding interval* $[v_{\min}, v_{\max}] = [\frac{r_{\min}}{1-\gamma}, \frac{r_{\max}}{1-\gamma}]$*. Also note that (A2.Fin) implies (A1.B), hence* η_s^* *has compact support in this case. In case of unbounded support, the design and analysis of DDP algorithms is more challenging.*

3 A general DDP Framework

A *parameterised projection* is a map

$$\Pi : \mathscr{P}(\mathbb{R}) \times \Xi \longrightarrow \mathscr{P}(\mathbb{R}), \quad (\mu, \xi) \longmapsto \Pi(\mu, \xi),$$

where $\Xi \ni \xi$ is a set of parameters. Parameters are chosen by a *parameter algorithm* A, that can depend on ingredients of $\mathfrak{mdp}$, and maps as

$$A : \mathscr{P}(\mathbb{R})^{\mathcal{S}} \times \mathcal{S} \times \mathbb{N} \longrightarrow \Xi, \quad (\eta, s, k) \longmapsto A(\eta, s, k).$$

Given an initial approximation $\eta^{(0)} \in \mathscr{P}(\mathbb{R})^{\mathcal{S}}$, the maps Π and A are used to calculate a sequence

$$(\eta^{(n)}, \boldsymbol{\xi}^{(n)}) \in \mathscr{P}(\mathbb{R})^{\mathcal{S}} \times \Xi^{\mathcal{S}}, \ n \in \mathbb{N} \tag{6}$$

inductively over $n \in \mathbb{N}$ as follows:

1. calculate $\boldsymbol{\xi}^{(n)} = [\xi_s^{(n)} : s \in \mathcal{S}]$ with $\xi_s^{(n)} = A(\eta^{(n-1)}, s, n)$,
2. calculate $\eta^{(n)} = [\eta_s^{(n)} : s \in \mathcal{S}]$ with $\eta_s^{(n)} = \Pi(\mathcal{T}_s \eta^{(n-1)}, \xi_s^{(n)})$.

All concrete parameterised projections we consider in this paper calculate finitely supported distributions in which the size of $\Pi(\mu, \xi) \in \mathscr{P}_{\text{fin}}(\mathbb{R})$ depends on the parameter ξ. Further, as we allow A to depend on the ingredients of $\mathfrak{mdp}$ and $\mathcal{T}$ is a function of $\mathfrak{mdp}$, the parameters $\boldsymbol{\xi}^{(n)}$ calculated in the n-th update step may depend on (properties of) $\mathcal{T}\eta^{(n-1)}$, see Section 5.3. Finally, the DDP framework considered in Chapter 5 of Bellemare et al. (2023) is included in this setting by considering Ξ to be a one-point set, we elaborate on this in Section 4.2.

Example 1 (Quantile Dynamic Programming, QDP) *Let* $M : \mathbb{N} \to \mathbb{N}$ *be a function, which is a hyper-parameter of the algorithm. Let* $\Xi = \mathbb{N}$ *and*

$$\Pi : \mathscr{P}(\mathbb{R}) \times \mathbb{N} \to \mathscr{P}(\mathbb{R}), \qquad \Pi(\mu, m) = \frac{1}{m} \sum\nolimits_{i=1}^{m} \delta_{x_i} \ \text{ with } x_i = F_\mu^{-1}\left(\frac{2i-1}{2m}\right),$$
$$A : \mathscr{P}(\mathbb{R})^{\mathcal{S}} \times \mathcal{S} \times \mathbb{N} \to \mathbb{N}, \qquad A(\eta, s, k) = M(k).$$

Let $\eta^{(0)}$ have finitely supported components. The iterative procedure to calculate $\eta^{(n)}, n \in \mathbb{N}$ as in (6) *specifies to*

$$\eta_s^{(n)} = \frac{1}{M(n)} \sum\nolimits_{i=1}^{M(n)} \delta_{x_i} \text{ with } x_i = F^{-1}_{\mathcal{T}_s\eta^{(n-1)}}\left(\frac{2i-1}{2M(n)}\right). \tag{7}$$

Following Chapter of 5 of Bellemare et al. (2023), who introduce and analyse this algorithm for $M(k) \equiv m \in \mathbb{N}$ constant and under (A2.Fin), we call this the Quantile Dynamic Programming (QDP) *algorithm. To perform* (7) *in practice, quantiles of $\mathcal{T}_s\eta^{(n-1)}$ need to be accessible, which holds under (A2.Fin) and leads to Algorithm 5.4 in Bellemare et al. (2023), which can be easily modified to work for arbitrary size functions M. By analysing the trade-off between approximation quality and time complexity, we argue to choose $M(k)$ growing exponentially in k. To be precise, choosing $M(k) = \lceil (1/\theta)^k \rceil$ for some $\theta \in [\gamma^c, 1)$, cf. Example 2.*

QDP is directly applicable only in case (A2.Fin), because then, for any finitely supported η, it is $\mathcal{T}_s\eta$ finitely supported, see (4), and hence its QF $F^{-1}_{\mathcal{T}_s\eta}$ can be explicitly evaluated. Under the more general assumption (A2), only the CDF $F_{\mathcal{T}_s\eta}$ can be evaluated, see (5), which is the CDF of an intricate mixture of continuous and/or unbounded distributions. Inverting CDFs is, in general, only possible with numerical approximation methods, which can be costly. Similar problems arise when aiming to design random number generators that are universally applicable for distributions μ in which only their CDFs can be evaluated, we refer to Section 7 of Hörmann et al. (2004) and Chapter II.2 in Devroye (1986). Note that many of these numerical inversion algorithms assume some sort of regularity on the CDF, which, in general, does not hold under (A2).

Concrete DDP algorithms applicable under (A2) are introduced and analysed in Section 5, which starts by introducing a particular parameterised projection, see Figure 2 for a visualisation. In Sections 5.1, 5.2 and 5.3 we discuss concrete parameter algorithms for this projection. The latter leads to an *approximation of the QDP algorithm applicable under (A2)*, where quantiles are approximated by *linear splines*, but this approximation is *on the level of parameters of the projection, not on the level of designing a projection.*

4 Bounding the approximation error

Let $\mathrm{d} : \mathscr{P}(\mathbb{R}) \times \mathscr{P}(\mathbb{R}) \to [0, \infty]$ be an *analysis metric* on $\mathscr{P}(\mathbb{R})$, which is allowed to assign infinite distance $\mathrm{d}(\mu, \nu) = \infty$ to pairs $\mu \neq \nu$.[2] By triangle inequality, d is a proper finite metric restricted to $\mathscr{P}_{\mathrm{d}}(\mathbb{R}) = \{\mu \in \mathscr{P}(\mathbb{R}) : \mathrm{d}(\mu, \delta_0) < \infty\} \subseteq \mathscr{P}(\mathbb{R})$.[3] Any such d is extended to a metric $\bar{\mathrm{d}}$ on $\mathscr{P}(\mathbb{R})^{\mathcal{S}}$ by $\bar{\mathrm{d}}(\eta, \eta') = \max_{s\in\mathcal{S}} \mathrm{d}(\eta_s, \eta'_s)$, which is a proper finite metric on $\mathscr{P}_{\mathrm{d}}(\mathbb{R})^{\mathcal{S}} \subseteq \mathscr{P}(\mathbb{R})^{\mathcal{S}}$. With respect to the analysis metric d, the quality of an approximation $\eta^{(n)}$ for η^* is measured by $\bar{\mathrm{d}}(\eta^{(n)}, \eta^*)$. Two popular parameterised families of analysis metrics are the following:

2. d satisfies $\mathrm{d}(\mu, \nu) = 0 \Leftrightarrow \mu = \nu$ and $\mathrm{d}(\mu, \nu) = \mathrm{d}(\nu, \mu)$ and $\mathrm{d}(\mu, \nu) \leq \mathrm{d}(\mu, \zeta) + \mathrm{d}(\zeta, \nu)$ for all $\mu, \nu, \zeta \in \mathscr{P}(\mathbb{R})$.
3. In Chapter 5 of Bellemare et al. (2023) it is further required that any $\mu \in \mathscr{P}_{\mathrm{d}}(\mathbb{R})$ has finite first moment.

The β-*Wasserstein* distances $w_\beta, \beta \in (0,\infty]$, are defined as

$$w_\beta(\mu,\nu) = \begin{cases} \int_0^1 |F_\mu^{-1}(u) - F_\nu^{-1}(u)|^\beta \mathrm{d}u, & \beta \in (0,1), \\ \left[\int_0^1 |F_\mu^{-1}(u) - F_\nu^{-1}(u)|^\beta \mathrm{d}u\right]^{1/\beta}, & \beta \in [1,\infty), \\ \sup_{u\in(0,1)} |F_\mu^{-1}(u) - F_\nu^{-1}(u)|, & \beta = \infty. \end{cases}$$

We have $\mathscr{P}_{w_\beta}(\mathbb{R}) = \mathscr{P}_\beta(\mathbb{R})$ for $\beta \in (0,\infty)$ and $\mathscr{P}_{w_\infty}(\mathbb{R})$ is the set of distributions with bounded support.

The *Birnbaum–Orlicz average* distances $\ell_\beta, \beta \in (0,\infty]$, are defined as

$$\ell_\beta(\mu,\nu) = \begin{cases} \int_{-\infty}^{\infty} |F_\mu(x) - F_\nu(x)|^\beta \mathrm{d}x, & \beta \in (0,1), \\ \left[\int_{-\infty}^{\infty} |F_\mu(x) - F_\nu(x)|^\beta \mathrm{d}x\right]^{1/\beta}, & \beta \in [1,\infty), \\ \sup_{x\in\mathbb{R}} |F_\mu(x) - F_\nu(x)|, & \beta = \infty. \end{cases}$$

It is $\mathrm{ks} := \ell_\infty$ the *Kolmogorov–Smirnov* distance, ℓ_2 is also known as *Cramér* distance and $\ell_1 = w_1$. A complete characterisation of all spaces $\mathscr{P}_{\ell_\beta}(\mathbb{R}), \beta \in (0,\infty]$, in terms of moment assumptions is challenging. However, for our purpose, it is enough to observe the following: for $\beta = 1$ we have $\mathscr{P}_{\ell_1}(\mathbb{R}) = \mathscr{P}_{w_1}(\mathbb{R}) = \mathscr{P}_1(\mathbb{R})$, in case $\beta = \infty$ it is $\mathscr{P}_{\ell_\infty}(\mathbb{R}) = \mathscr{P}_{\mathrm{ks}}(\mathbb{R}) = \mathscr{P}(\mathbb{R})$, that is ks is a metric on the whole of $\mathscr{P}(\mathbb{R})$, and for $\beta \in (1,\infty)$ we have

$$\mathscr{P}_{\frac{1}{\beta}}(\mathbb{R}) \subsetneq \mathscr{P}_{\ell_\beta}(\mathbb{R}) \subsetneq \bigcap\nolimits_{0<\varepsilon<\frac{1}{\beta}} \mathscr{P}_{\frac{1}{\beta}-\varepsilon}(\mathbb{R}), \tag{8}$$

which is proved in Proposition 25 in Appendix A.1.

4.1 The Accumulated Projection Error

The following theorem is the key tool to study $\bar{\mathrm{d}}(\eta^{(n)}, \eta^*)$, but also shows that not all of the above introduced metrics are equally admissible as analysis metrics. Let $c \in (0,\infty)$ be a fixed constant.

Theorem 5 (Theorem 4.25 of Bellemare et al. (2023)) *Let* d *be c-homogeneous, regular and convex*[4]*. Then* $\mathcal{T}$ *is a* γ^c*-contraction with respect to* $\bar{\mathrm{d}}$*, that is*

$$\bar{\mathrm{d}}\left(\mathcal{T}\eta, \mathcal{T}\eta'\right) \leq \gamma^c \cdot \bar{\mathrm{d}}(\eta,\eta') \text{ for all } \eta,\eta' \in \mathscr{P}(\mathbb{R})^{\mathcal{S}} \tag{9}$$

and, consequently, for all $\eta \in \mathscr{P}(\mathbb{R})^{\mathcal{S}}$*:*

(i) $\bar{\mathrm{d}}\left(\mathcal{T}^{\circ n}\eta, \eta^*\right) \leq \gamma^{cn} \cdot \bar{\mathrm{d}}(\eta,\eta^*)$ *for all* $n \in \mathbb{N}_0$*,*

(ii) $\bar{\mathrm{d}}(\eta,\eta^*) < \infty \implies \bar{\mathrm{d}}(\eta,\eta^*) \leq \frac{\bar{\mathrm{d}}(\eta,\mathcal{T}\eta)}{1-\gamma^c}$.

Of the above introduced metrics, the following have the properties stated in Theorem 5:

- $\mathrm{d} = w_\beta, \beta \in (0,\infty]$ with $c = \min\{1,\beta\}$,

4. See Definition 26 in Appendix A.2, also refer to Definitions 4.23-25 in Bellemare et al. (2023).

- $\mathrm{d} = \ell_\beta, \beta \in [1,\infty)$ with $c = 1/\beta$.

In particular, $\mathrm{ks} = \ell_\infty$ is *not* covered by Theorem 5.

Remark 6 *The inequalities stated in Theorem 5 are of interest in case the distances appearing in the upper bounds are finite. We have that*

$$\eta, \eta', \eta^* \in \mathscr{P}_\mathrm{d}(\mathbb{R})^\mathcal{S}, \;\; \mathcal{T}(\mathscr{P}_\mathrm{d}(\mathbb{R})^\mathcal{S}) \subseteq \mathscr{P}_\mathrm{d}(\mathbb{R})^\mathcal{S} \implies \bar{\mathrm{d}}(\eta,\eta'), \bar{\mathrm{d}}(\eta,\eta^*), \bar{\mathrm{d}}(\eta,\mathcal{T}\eta) < \infty.$$

By triangle inequality, (9) *and properties of* d *it can be shown that the moment assumption*

$$\textit{(A1.d)}: \quad \mathcal{L}(R_{sa\bar{s}}) \in \mathscr{P}_\mathrm{d}(\mathbb{R}) \textit{ for all } s, a, \bar{s}$$

implies $\mathcal{T}(\mathscr{P}_\mathrm{d}(\mathbb{R})^\mathcal{S}) \subseteq \mathscr{P}_\mathrm{d}(\mathbb{R})^\mathcal{S}$. *This alone may, in general, not be sufficient to conclude* $\eta^* \in \mathscr{P}_\mathrm{d}(\mathbb{R})^\mathcal{S}$, *but for the metrics we consider the following holds: for* $\mathrm{d} = w_\beta, \beta \in (0,\infty)$, *it is* $\mathscr{P}_{w_\beta}(\mathbb{R}) = \mathscr{P}_\beta(\mathbb{R})$, *hence (A1.*$w_\beta$*) is equivalent to (A1.P*β*) and this assumption implies* $\eta^* \in \mathscr{P}_\beta(\mathbb{R})^\mathcal{S}$, *see Remark 4. Similarly,* $\mathscr{P}_{w_\infty}(\mathbb{R})$ *is the set of distributions with bounded support, hence (A1.*w_∞*) is equivalent to (A1.B), which implies that all components of* η^* *have bounded support. Finally, we have* $\mathscr{P}_{1/\beta}(\mathbb{R}) \subset \mathscr{P}_{\ell_\beta}(\mathbb{R})$ *for* $\beta \in [1,\infty)$, *see Proposition 25, hence (A1.*ℓ_β*) is implied by (A1.P*$\frac{1}{\beta}$*), which implies* $\eta^* \in \mathscr{P}_{1/\beta}(\mathbb{R})^\mathcal{S} \subseteq \mathscr{P}_{\ell_\beta}(\mathbb{R})^\mathcal{S}$.

The Kolmogorov–Smirnov distance $\mathrm{ks} = \ell_\infty$ does not satisfy the properties stated in Theorem 5 and thus is not as easily admissible as, for example, Wasserstein distances. However, we demonstrate in Section 6 how bounds on approximation errors $\bar{\mathrm{d}}(\eta^{(n)}, \eta^*)$ using $\mathrm{d} = w_\beta$ for some $\beta \in (0,\infty)$ lead to useful bounds for $\overline{\mathrm{ks}}(\eta^{(n)}, \eta^*)$.

For now, let d be a c-homogeneous, regular and convex metric on $\mathscr{P}(\mathbb{R})$. Let $\eta^{(0)}$ be an initial approximation and let $(\eta^{(n)}, \boldsymbol{\xi}^{(n)}), n \in \mathbb{N}$ be the sequence calculated inductively as in (6). A possible approach to bound the approximation error $\bar{\mathrm{d}}(\eta^{(n)}, \eta^*)$ is as follows: define the *k-th projection error*(PE) and the *n-th accumulated projection error*(APE) by

$$\begin{aligned} \mathrm{PE}(k,\mathrm{d}) &= \bar{\mathrm{d}}(\eta^{(k)}, \mathcal{T}\eta^{(k-1)}), \\ \mathrm{APE}(n,\mathrm{d}) &= \sum\nolimits_{k=1}^{n} \gamma^{c(n-k)} \cdot \mathrm{PE}(k,\mathrm{d}). \end{aligned}$$

Proposition 7 $\bar{\mathrm{d}}\left(\eta^{(n)}, \eta^*\right) \le \mathrm{APE}(n,\mathrm{d}) + \gamma^{cn}\bar{\mathrm{d}}(\eta^{(0)}, \eta^*)$.

Proof By induction on $n \in \mathbb{N}_0$: for $n = 0$ it is obvious. Assuming the inequality holds for some $n \in \mathbb{N}_0$ and noticing the recursive formula $\mathrm{APE}(n+1,\mathrm{d}) = \mathrm{PE}(n+1,\mathrm{d}) + \gamma^c \mathrm{APE}(n,\mathrm{d})$ it follows

$$\begin{aligned} \bar{\mathrm{d}}\left(\eta^{(n+1)}, \eta^*\right) &\le \bar{\mathrm{d}}\left(\eta^{(n+1)}, \mathcal{T}\eta^{(n)}\right) + \bar{\mathrm{d}}\left(\mathcal{T}\eta^{(n)}, \eta^*\right) \\ &= \mathrm{PE}(n+1,\mathrm{d}) + \bar{\mathrm{d}}\left(\mathcal{T}\eta^{(n)}, \mathcal{T}\eta^*\right) \\ &\le \mathrm{PE}(n+1,\mathrm{d}) + \gamma^c \bar{\mathrm{d}}(\eta^{(n)}, \eta^*) \\ &\le \mathrm{PE}(n+1,\mathrm{d}) + \gamma^c \left(\mathrm{APE}(n,\mathrm{d}) + \gamma^{cn}\bar{\mathrm{d}}(\eta^{(0)}, \eta^*)\right) \\ &= \mathrm{APE}(n+1,\mathrm{d}) + \gamma^{c(n+1)}\bar{\mathrm{d}}(\eta^{(0)}, \eta^*), \end{aligned}$$

where the first inequality follows from triangle inequality, the equality from $\mathcal{T}\eta^* = \eta^*$, the second inequality from Theorem 5 and the third by induction hypothesis. ■

Bounds on projection errors yield bounds on the accumulated projection error:

Lemma 8 *Let $r \in (0,\infty), \theta \in (0,1), D > 0$. Denote with $\mathrm{Li}_s(z) = \sum_{k=1}^{\infty} k^{-s} z^k$ the polylogarithm of order s and argument z. Then the following implications hold*

$$\forall k = 1,\dots,n:\ \mathrm{PE}(k,\mathrm{d}) \le D \quad \Longrightarrow \quad \mathrm{APE}(n,\mathrm{d}) \le \frac{D}{1-\gamma^c},$$

$$\forall k = 1,\dots,n:\ \mathrm{PE}(k,\mathrm{d}) \le D \cdot k^{-r} \quad \Longrightarrow \quad \mathrm{APE}(n,\mathrm{d}) \le \frac{D \cdot \mathrm{Li}_{-r}(\gamma^c)}{\gamma^c} \cdot n^{-r},$$

$$\forall k = 1,\dots,n:\ \mathrm{PE}(k,\mathrm{d}) \le D \cdot \theta^k \quad \Longrightarrow \quad \mathrm{APE}(n,\mathrm{d}) \le \begin{cases} \frac{D}{(\gamma^c/\theta)-1} \cdot \gamma^{cn}, & \theta < \gamma^c, \\ D \cdot n \cdot \gamma^{cn}, & \theta = \gamma^c, \\ \frac{D}{1-\gamma^c/\theta} \cdot \theta^n, & \theta > \gamma^c. \end{cases}$$

Proof The first and third implications follow from bounding finite geometric series by infinite geometric series, resp. explicit evaluation in case $\theta = \gamma^c$. The second implication follows from $1/(n-k) \le (k+1)/n$ for every $k = 0,\dots,n-1$, bounding finite by infinite series and plugging in the definition of the polylogarithm. ■

Proposition 7 in combination with Lemma 8 show that if projection errors decay towards zero fast enough as $k \to \infty$, then the approximation error $\bar{\mathrm{d}}(\eta^{(n)}, \eta^*)$ also decays to zero as $n \to \infty$ with a comparable rate of decay. This allows to bound the minimal number of iterations required to calculate an approximation $\eta^{(n)}$ that is ε-close to η^* with respect to d. By investigating the trade-off between time complexity and approximation quality, we make a case for constructing DDP algorithms such that $\mathrm{PE}(k,\mathrm{d})$ decays exponentially with some rate $\theta \in [\gamma^c, 1)$, corresponding to the last (two) parts of the previous lemma, see the following Example 2 and Section 5.1.

Example 2 (Analysing QDP with respect to $\mathrm{d} = w_1$) *Let $\eta^{(n)}, n \in \mathbb{N}$ be as in Example 1 having used the size function $M : \mathbb{N} \to \mathbb{N}$, that is $\eta_s^{(n)} = \Pi(\mathcal{T}_s \eta^{(n-1)}, M(n))$ with $\Pi(\mu, m) = m^{-1} \sum_{i=1}^{m} \delta_{F_\mu^{-1}(2i-1/2m)}$. We consider the analysis metric $\mathrm{d} = w_1 = \ell_1$, which is $c = 1$-homogeneous. To analyse QDP, we note the following inequalities:*

$$w_1(\Pi(\mu,m),\mu) \le \int_0^{\frac{1}{2m}} \left(F_\mu^{-1}(1-u) - F_\mu^{-1}(u)\right) \mathrm{d}u, \quad w_1(\mu,\nu) \le w_\infty(\mu,\nu).$$

As discussed, QDP is practical in case (A2.Fin) holds, which implies that reward distributions are supported on some compact interval $[r_{\min}, r_{\max}] \subset \mathbb{R}$. Let $[v_{\min}, v_{\max}] = [r_{\min}/(1-\gamma), r_{\max}/(1-\gamma)]$. If all components of $\eta \in \mathscr{P}(\mathbb{R})^{\mathcal{S}}$ are supported on $[v_{\min}, v_{\max}]$, then all components of $\mathcal{T}\eta$ are supported on $[v_{\min}, v_{\max}]$ and the same holds for the components of η^*. Assume $\eta_s^{(0)}$ is finitely supported with $M(0)$ particles supported on $[v_{\min}, v_{\max}]$.

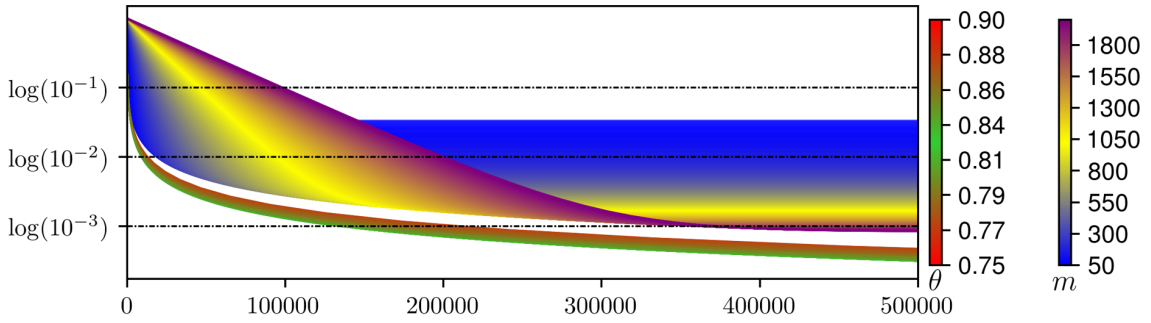

Figure 1: (Partially) overlapping curves $T \mapsto \log(e(n(T)))$ for $\gamma = 0.7$ and different size functions; $M(k) = \lceil (1/\theta)^k \rceil$ with $\theta \in [0.75, 0.9]$ and $M(k) \equiv m$ constant with $m \in \{50, 51, \dots, 2000\}$. For each choice of M the curve has a single color.

By bounding quantiles of compactly-supported distributions by the endpoints of their support, using the above inequalities and plugging in the definitions of PE, APE *as well as using Proposition 7 we obtain*

$$\mathrm{PE}(k, w_1) \leq \frac{v_{\max} - v_{\min}}{2M(k)}, \quad \mathrm{APE}(n, w_1) \leq \sum_{k=1}^{n} \gamma^{n-k} \frac{v_{\max} - v_{\min}}{2M(k)}, \quad \overline{w}_1(\eta^{(n)}, \eta^*) \leq e(n),$$

where the upper bound $e(n)$ on the n-th approximation error is

$$e(n) = \sum_{k=1}^{n} \gamma^{n-k} \frac{[v_{\max} - v_{\min}]}{2M(k)} + \gamma^n [v_{\max} - v_{\min}].^5 \tag{10}$$

Treating the sizes of $\mathcal{S}, \mathcal{A}$ as well as the number of particles of reward distributions as constants, the time complexity to calculate $\eta^{(n)}$ is of order $\mathrm{O}(\mathrm{time}(n))$ *with*

$$\mathrm{time}(n) = \sum_{k=1}^{n} M(k) \log\left(M(k)\right),$$

compare to page 137 in Bellemare et al. (2023). To visualise the effect of choosing different size functions M, let

$$n(T) = \max\{n \in \mathbb{N}_0 | \mathrm{time}(n) \leq T\}$$

be the number of iterations QDP has calculated up to (a global constant of) time $T > 0$. Thus, $e(n(T))$ is a guaranteed bound for the approximation error after execution time T. Note that for $e(n(T)) \to 0$ as $T \to \infty$ it is necessary that $M(k) \to \infty$ as $k \to \infty$. Figure 1 shows plots of functions $T \mapsto \log(e(n(T)))$ for different choices of sizes functions M: we compare $M(k) = \lceil (1/\theta)^k \rceil$ for different θ with $M(k) \equiv m$ for different m. The plot suggests to prefer the exponential choice with $\theta \approx \frac{\gamma+1}{2}$.

5. if $M(k) \equiv m \in \mathbb{N}$ is constant, $e(n)$ can be further bounded by $[v_{\max} - v_{\min}] \cdot \left[\frac{1}{2m(1-\gamma)} + \gamma^n\right]$, cf. Lemma 5.30 of Bellemare et al. (2023).

4.2 Fixed Non-Expansive Projection

In case Ξ contains one element, Π can be reduced to a function $\Pi : \mathscr{P}(\mathbb{R}) \to \mathscr{P}(\mathbb{R})$. Let $\bar{\Pi} : \mathscr{P}(\mathbb{R})^{\mathcal{S}} \to \mathscr{P}(\mathbb{R})^{\mathcal{S}}$ be the component-wise extension of Π. The n-th calculated approximation $\eta^{(n)}$ then takes the form $\eta^{(n)} = (\bar{\Pi}\mathcal{T})^{\circ n}(\eta^{(0)})$ and the parameter algorithm A as well as the sequence $\boldsymbol{\xi}^{(n)}, n \in \mathbb{N}$, can be eliminated from the formalism of Section 3. This corresponds to the DDP framework considered in Chapter 5 in Bellemare et al. (2023). Let d be c-homogeneous, regular and convex. Proposition 7 applies and leads to

$$\bar{\mathrm{d}}(\eta^{(n)}, \eta^*) \leq \gamma^{cn}\bar{\mathrm{d}}(\eta^{(0)}, \eta^*) + \sum\nolimits_{k=1}^{n} \gamma^{c(n-k)}\bar{\mathrm{d}}(\eta^{(k)}, \mathcal{T}\eta^{(k-1)}). \tag{APE-bound}$$

In case Π is a *non-expansion* with respect to d, the approximation errors $\bar{\mathrm{d}}(\eta^{(n)}, \eta^*)$ have been studied in Chapter 5 of Bellemare et al. (2023). To discuss how the bound (APE-bound) relates we make the following assumptions:

(i) Π is a *non-expansion* with respect to d, that is $\mathrm{d}(\Pi\mu, \Pi\nu) \leq \mathrm{d}(\mu, \nu)$ for all $\mu, \nu \in \mathscr{P}(\mathbb{R})$,

(ii) $\mathcal{T}(\mathscr{P}_{\mathrm{d}}(\mathbb{R})^{\mathcal{S}}) \subseteq \mathscr{P}_{\mathrm{d}}(\mathbb{R})^{\mathcal{S}}$ and $\bar{\Pi}(\mathscr{P}_{\mathrm{d}}(\mathbb{R})^{\mathcal{S}}) \subseteq \mathscr{P}_{\mathrm{d}}(\mathbb{R})^{\mathcal{S}}$,

(iii) $\bar{\Pi}\mathcal{T}$ has a fixed-point $\hat{\eta} \in \mathscr{P}_{\mathrm{d}}(\mathbb{R})^{\mathcal{S}}$.

Note that (i) implies that $\bar{\Pi}\mathcal{T}$ is a γ^c-contraction; (ii) implies that $\bar{\Pi}\mathcal{T}(\mathscr{P}_{\mathrm{d}}(\mathbb{R})^{\mathcal{S}}) \subseteq \mathscr{P}_{\mathrm{d}}(\mathbb{R})^{\mathcal{S}}$. In particular, the fixed-point $\hat{\eta}$ of $\bar{\Pi}\mathcal{T}$ is unique within $\mathscr{P}_{\mathrm{d}}(\mathbb{R})^{\mathcal{S}}$. Furthermore, (i)+(ii) imply (iii) in case the space $(\mathscr{P}_{\mathrm{d}}(\mathbb{R})^{\mathcal{S}}, \bar{\mathrm{d}})$ is complete. By triangle inequality, conditions (i)-(iii) lead to

$$\bar{\mathrm{d}}(\eta^{(n)}, \eta^*) \leq \gamma^{cn}\bar{\mathrm{d}}(\eta^{(0)}, \hat{\eta}) + \bar{\mathrm{d}}(\hat{\eta}, \eta^*). \tag{$\hat{\eta}$-bound}$$

Note that ($\hat{\eta}$-bound) makes use of assumptions (i)-(iii), while (APE-bound) does not. However, under these additional assumptions, (APE-bound) can be further bounded and is seen to be close to ($\hat{\eta}$-bound): first, for every $k \in \mathbb{N}$ it holds that

$$\begin{aligned}
\bar{\mathrm{d}}(\eta^{(k)}, \mathcal{T}\eta^{(k-1)}) &\leq \bar{\mathrm{d}}(\eta^{(k)}, \hat{\eta}) + \bar{\mathrm{d}}(\hat{\eta}, \mathcal{T}\hat{\eta}) + \bar{\mathrm{d}}(\mathcal{T}\hat{\eta}, \mathcal{T}\eta^{(k-1)}) \\
&\leq \gamma^{ck}\bar{\mathrm{d}}(\eta^{(0)}, \hat{\eta}) + \bar{\mathrm{d}}(\hat{\eta}, \mathcal{T}\hat{\eta}) + \gamma^{c}\bar{\mathrm{d}}(\hat{\eta}, \eta^{(k-1)}) \\
&\leq 2\gamma^{ck}\bar{\mathrm{d}}(\eta^{(0)}, \hat{\eta}) + \bar{\mathrm{d}}(\mathcal{T}\hat{\eta}, \hat{\eta})
\end{aligned}$$

and hence

$$\text{(APE-bound)} \leq \gamma^{cn}\bar{\mathrm{d}}(\eta^{(0)}, \eta^*) + 2n\gamma^{cn}\bar{\mathrm{d}}(\eta^{(0)}, \hat{\eta}) + \frac{\bar{\mathrm{d}}\left(\mathcal{T}\hat{\eta}, \hat{\eta}\right)}{1-\gamma^c}. \tag{11}$$

The only non-vanishing terms in the upper bounds ($\hat{\eta}$-bound), resp. (11), are $\bar{\mathrm{d}}(\hat{\eta}, \eta^*)$, resp. $\bar{\mathrm{d}}\left(\mathcal{T}\hat{\eta}, \hat{\eta}\right)/(1-\gamma^c)$. They are related by

$$\frac{\max\left\{\bar{\mathrm{d}}(\bar{\Pi}\eta^*, \eta^*), \bar{\mathrm{d}}(\mathcal{T}\hat{\eta}, \hat{\eta})\right\}}{1+\gamma^c} \leq \bar{\mathrm{d}}(\hat{\eta}, \eta^*) \leq \frac{\min\left\{\bar{\mathrm{d}}(\bar{\Pi}\eta^*, \eta^*), \bar{\mathrm{d}}(\mathcal{T}\hat{\eta}, \hat{\eta})\right\}}{1-\gamma^c}, \tag{12}$$

which follows from triangle inequality and contractive properties.[6] Finally,

$$\begin{aligned}
(\hat{\eta}\text{-bound}) \quad &\leq \quad \gamma^{cn}\bar{\mathrm{d}}(\eta^{(0)}, \hat{\eta}) \quad + \quad \frac{\bar{\mathrm{d}}\left(\bar{\Pi}\eta^*, \eta^*\right)}{1-\gamma^c} \quad , \\
\text{(APE-bound)} \quad &\leq \quad \gamma^{cn}\bar{\mathrm{d}}(\eta^{(0)}, \eta^*) \quad + \quad 2n\gamma^{cn}\bar{\mathrm{d}}(\eta^{(0)}, \hat{\eta}) \quad + \quad \frac{1+\gamma^c}{1-\gamma^c} \cdot \frac{\bar{\mathrm{d}}\left(\bar{\Pi}\eta^*, \eta^*\right)}{1-\gamma^c} \quad .
\end{aligned}$$

6. Also, (12) immediately implies $\eta^* = \hat{\eta} \Leftrightarrow \Pi\eta^* = \eta^* \Leftrightarrow \mathcal{T}\hat{\eta} = \hat{\eta}$.

The vanishing term in the final upper bound for (APE-bound) decays to zero faster than θ^n for every $\theta \in (\gamma^c, 1)$ and the non-vanishing term is the same as in ($\hat{\eta}$-bound) up to a factor $(1+\gamma^c)/(1-\gamma^c)$.

In summary, we conclude that bounding approximation errors by accumulated projection errors is a very flexible approach, that can also cover varying projections or fixed-projections that are expansive with respect to d. In the special case of fixed non-expansive projections, ($\hat{\eta}$-bound) appears to be sharper than (APE-bound), but for practical reasons the bounds are close.

Remark 9 *The QDP-Algorithm with a constant size function $M(k) \equiv m$ fits the framework of a fixed-projection, $\Pi = \Pi(\cdot, m)$. Note that this projection is an expansion with respect to w_1, that is the approximation errors cannot be bounded by* ($\hat{\eta}$-bound) *directly; cf. Lemma 5.30 and Exercise 5.20 in Bellemare et al. (2023).*

5 A class of DDP algorithms

A parameterised projection $\Pi : \mathscr{P}(\mathbb{R}) \times \Xi \to \mathscr{P}_{\text{fin}}(\mathbb{R})$ that is well-suited for DDP algorithms under (A2), closely related to *quantization schemes* of Lloyd (1982), is given as follows: The parameter space is $\Xi = \bigcup_{m\in\mathbb{N}} \Xi_m$ with

$$\Xi_m = \Big\{ \xi = (x_1, \dots, x_m, y_1, \dots, y_{m-1}) \in \mathbb{R}^{2m-1} \;\Big|\; x_1 \le y_1 \le x_2 \le y_2 \le \cdots \le y_{m-1} \le x_m \Big\}.$$

and $\Pi(\mu, \xi) \in \mathscr{P}_{\text{fin}}(\mathbb{R})$ is defined, for $\mu \in \mathscr{P}(\mathbb{R}), \xi = (x_1, \dots, x_m, y_1, \dots, y_{m-1}) \in \Xi_m$, as

$$\Pi(\mu, \xi) = \sum\nolimits_{i=1}^{m} (F_\mu(y_i) - F_\mu(y_{i-1})) \cdot \delta_{x_i}, \tag{13}$$

with the convention $y_0 = -\infty, y_m = +\infty$ and $F_\mu(-\infty) = 0, F_\mu(+\infty) = 1$, see Figure 2. In particular, $\Pi(\mu, \xi)$ depends on μ only by a finite number of evaluations F_μ, which makes it well-suited for DDP algorithms under (A2), as the CDF of $\mu = \mathcal{T}_s\eta$ can be evaluated in this case by (5). In particular, formula (5) directly yields the following:

Theorem 10 *Let $m', m \in \mathbb{N}$ and $\eta = [\eta_{\bar{s}} : \bar{s} \in \mathcal{S}] \in \mathscr{P}(\mathbb{R})^{\mathcal{S}}$ with $\eta_{\bar{s}} = \sum_{j=1}^{m'} q_{\bar{s}j}\delta_{z_{\bar{s}j}}$. Let $\xi = (x_1, \dots, x_n, y_1, \dots, y_{m-1}) \in \Xi_m$ and $s \in \mathcal{S}$. Then $\Pi(\mathcal{T}_s(\eta), \xi) = \sum_{i=1}^{m} o_i \delta_{x_i}$ with*

$$o_i = \sum\nolimits_{(a,\bar{s})\in\mathcal{A}\times\mathcal{S}} \sum\nolimits_{j=1}^{m'} \pi(a|s) \cdot p(\bar{s}|s,a) \cdot q_{\bar{s}j} \cdot \left[F_{sa\bar{s}}(y_i - \gamma z_{\bar{s}j}) - F_{sa\bar{s}}(y_{i-1} - \gamma z_{\bar{s}j}) \right].$$

That is, $\Pi(\mathcal{T}_s(\eta), \xi)$ can be calculated under (A2)+(A3) in time of order $\mathrm{O}(m' \cdot m)$.

Let $\eta^{(0)} \in \mathscr{P}_{\text{fin}}(\mathbb{R})^{\mathcal{S}}$ be an initial approximation. In order to calculate approximations $\eta^{(n)}, n \in \mathbb{N}$ using Π one has to decide for a suitable *parameter algorithm*

$$A : \mathscr{P}(\mathbb{R})^{\mathcal{S}} \times \mathcal{S} \times \mathbb{N} \;\longrightarrow\; \Xi = \cup_{m\in\mathbb{N}} \Xi_m.$$

In order to control the space-time complexity, it is reasonable to choose a *size function* $M : \mathbb{N} \to \mathbb{N}$ and to consider only parameter algorithms A such that

$$A(\cdot, \cdot, k) \in \Xi_{M(k)} \;\text{ for all }\; k \in \mathbb{N}. \tag{14}$$

Given such a parameter algorithm, the iterative procedure to calculate $\eta^{(n)}, n \in \mathbb{N}$ is thus

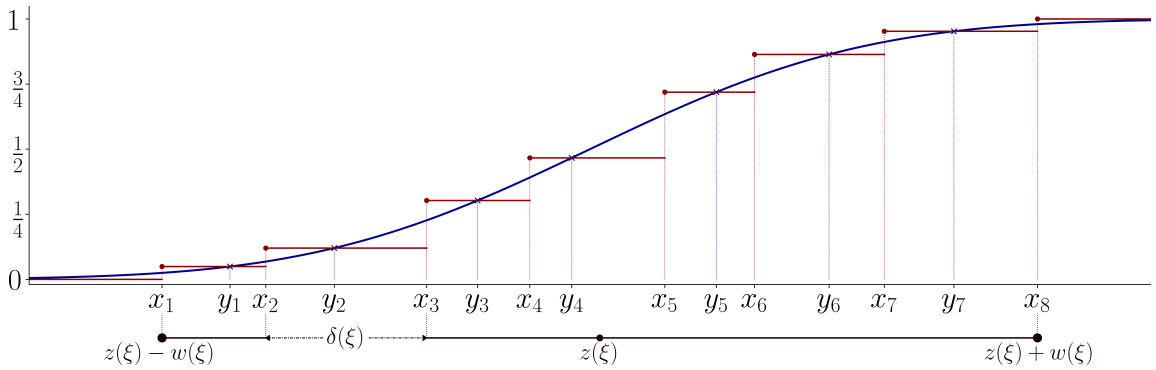

Figure 2: CDFs of μ (blue) and $\Pi(\mu, \xi)$ (red) with $\xi = (x_1, \dots, x_8, y_1, \dots, y_7) \in \Xi_8$. The support of $\Pi(\mu, \xi)$ is the compact interval $[z(\xi) - w(\xi), z(\xi) + w(\xi)] = [x_1, x_8]$ and $\delta(\xi) = \max_{2 \le i \le 8} |x_i - x_{i-1}|$.

1. Calculate $\boldsymbol{\xi}^{(n)} = [\boldsymbol{\xi}_s^{(n)} : s \in \mathcal{S}]$ with $\boldsymbol{\xi}_s^{(n)} = A(\eta^{(n-1)}, s, n) \in \Xi_{M(n)}$,
2. Calculate $\eta^{(n)} = [\eta_s^{(n)} : s \in \mathcal{S}]$ with $\eta_s^{(n)} = \Pi(\mathcal{T}_s \eta^{(n-1)}, \xi_s^{(n)})$ as in Theorem 10.

For practical purposes, A has to be such that $A(\eta^{(n-1)}, s, n)$ can be evaluated under Assumptions (A2)+(A3) within a controllable space-time complexity. In Remark 12 we discuss the topic of *optimal parameter algorithms* and demonstrate that these are, in general, *not easily available for practical purposes*. In the following subsections, we consider three different parameter algorithms A that can be applied under Assumptions (A2)+(A3).

(i) In Section 5.1 we introduce plain parameter algorithms (PPA). By analysing the trade-off between approximation quality and time complexity in detail, we collect additional evidence that constructing DDP algorithms with projection errors decaying of the order $\mathrm{O}(\theta^k)$ with $\theta \in [\gamma^c, 1)$ seems reasonable.

(ii) In Section 5.2 we introduce an adaptive version of the plain parameter algorithms (ADP), which is derived by bounding quantiles of convolutions. A partial analysis leads to a useful *black-box algorithm*, which is suggested for application in case (A1.Pα) holds for large α.

(iii) In Section 5.3 we modify the algorithm of Section 5.2 further by considering *spline interpolations of inverse quantile functions* (QSP). The emerging algorithm can be seen as an approximation of the QDP-algorithm applicable under assumptions (A2)+(A3) and is a second *black-box* algorithm, that can be used also in presence of heavy-tailed reward distributions, see assumption (A1.Pα).

The mathematical analysis of Section 5.1 relies on bounding projection errors, we present the basic tool now. Let $\xi = (x_1, \dots, x_m, y_1, \dots, y_{m-1}) \in \Xi$. Key-characteristics of ξ are

$$\delta(\xi) = \max_{2 \le i \le m} |x_i - x_{i-1}|, \quad z(\xi) = \frac{x_m + x_1}{2}, \quad w(\xi) = \frac{x_m - x_1}{2}.$$

So, for any $\mu \in \mathscr{P}(\mathbb{R})$ we have that $\Pi(\mu, \xi)$ is finitely supported on m points that lie in the compact interval $[z(\xi) - w(\xi), z(\xi) + w(\xi)]$, see Figure 2. As we deal with distributions that can have unbounded support and the analysis metrics we consider (also) depend on the tail behaviour of distributions, we need to introduce the following terms: for $\mu = \mathcal{L}(X) \in \mathscr{P}(\mathbb{R}), \xi \in \Xi$ and $\beta \in (0, \infty)$ define

$$\mathrm{tail}_{w_\beta}(\mu, \xi) = \int_0^\infty x^{\beta-1} \mathbb{P}[|X - z(\xi)| > w(\xi) + x] \, \mathrm{d}x,$$
$$\mathrm{tail}_{\ell_\beta}(\mu, \xi) = \int_0^\infty \mathbb{P}[|X - z(\xi)| > w(\xi) + x]^\beta \, \mathrm{d}x.$$

The following is proved in Appendix A.3:

Lemma 11 *Let $\mu \in \mathscr{P}(\mathbb{R}), \xi \in \Xi$. Then*

(a) $w_\beta(\Pi(\mu, \xi), \mu) \le 4 \cdot \max\left\{\delta(\xi)^{\frac{\beta}{1 \vee \beta}}, \ \mathrm{tail}_{w_\beta}(\mu, \xi)^{\frac{1}{1 \vee \beta}}\right\}$ *for all* $\beta \in (0, \infty)$,

(b) $\ell_\beta(\Pi(\mu, \xi), \mu) \le 4 \cdot \max\left\{\delta(\xi)^{\frac{1}{\beta}}, \ \mathrm{tail}_{\ell_\beta}(\mu, \xi)^{\frac{1}{\beta}}\right\}$ *for all* $\beta \in [1, \infty)$.

Remark 12 (Optimal Parameter Algorithms) *Let* d *be a c-homogeneous, regular, convex analysis metric and consider only parameter algorithms that satisfy* (14)*. Choosing a parameter algorithm A with the property*

$$A(\eta, s, k) \ \in \ \mathrm{argmin}_{\xi \in \Xi_m} \mathrm{d}\left(\Pi(\mu, \xi), \mu\right) \ \textit{with} \ (\mu, m) = (\mathcal{T}_s \eta, M(k))$$

for all η, s, k, would result in calculating approximations with minimal projections errors with respect to d*. The task to minimise* $\mathrm{d}\left(\Pi(\mu, \xi), \mu\right)$ *over $\xi \in \Xi_m$ for given $\mu \in \mathscr{P}(\mathbb{R})$ can, in general, only be solved by approximation methods: for example, consider* $\mathrm{d} = w_2$*. For $\xi = (x_1, \ldots, x_m, y_1, \ldots, y_{m-1}) \in \Xi_m$ it is*

$$w_2(\Pi(\mu, \xi), \mu) = \sum\nolimits_{i=1}^m \int_{y_{i-1}}^{y_i} (x - x_i)^2 \mathrm{d}\mu(x).$$

Minimising this expression in ξ is the topic of the groundbreaking paper Lloyd (1982), which has led to what is now known as Lloyd's algorithm*. This had motivated various other algorithms for related minimisation problems (involving distributions on higher dimensions, $\mu \in \mathscr{P}(\mathbb{R}^k)$) such as (variants of)* m-means clustering algorithms *or algorithms to construct* optimal centroidal Voronoi tessellations*. We refer to Du et al. (1999) and Kloeckner (2012), where these types of problems are discussed also for other metrics than* $\mathrm{d} = w_2$ *and in higher dimensions. Regarding algorithms, Lloyd's algorithm (and its variants) may depend on μ in a more complex way than on finitely many evaluations of the cdf F_μ. Hence, using these in our situation for $\mu = \mathcal{T}_s \eta$ under (A2), would need to apply further approximation methods. Considering the trade-off between space-time complexity and approximation quality, we leave it for future research to investigate if* parameter algorithms *of such high complexity are of any benefit for the DDP task at hand.*

Finally, note that the minimisation problem $\min_\nu w_1(\nu, \mu)$ *in which the minimum runs over all ν of the form $\nu = \frac{1}{m} \sum_{i=1}^m \delta_{x_i}$ has been solved explicitly by letting $x_i = F_\mu^{-1}((2i - 1)/2m)$, see Proposition 5.15 Bellemare et al. (2023), which was used to motivate the QDP algorithm (Example 1) in that reference.*

5.1 Plain parameter algorithm (PPA)

For each $m \in \mathbb{N}, m \geq 2, w > 0, z \in \mathbb{R}$ let

$$\xi^{\mathrm{lin}}(m, w, z) = (x_1, \ldots, x_m, y_1, \ldots, y_{m-1}) \in \Xi_m \text{ with}$$
$$x_i = z + \frac{2i - 1 - m}{m - 1} \cdot w,$$
$$y_i = z + \frac{2i - m}{m - 1} \cdot w,$$

so the interlaced points $(x_1, y_1, x_2, y_2, \ldots, y_{m-1}, x_m)$ form an evenly spaced interpolation of the compact interval $[z - w, z + w]$. Further, with $\xi = \xi^{\mathrm{lin}}(m, w, z)$ it holds that $(z(\xi), w(\xi), \delta(\xi)) = (z, w, \frac{2w}{m-1})$. In case $m = 1$ let $\xi^{\mathrm{lin}}(1, w, z) = (z) \in \Xi_1$. The plain parameter algorithm is defined as follows:

Definition 13 (PPA) *The* plain parameter algorithm *(PPA) has hyper-parameter (functions)* $M : \mathbb{N} \to \mathbb{N}$, $W : \mathbb{N} \to (0, \infty)$, $z \in \mathbb{R}$ *with* M, W *non-decreasing and is defined as* $A(\eta, s, k) := \xi^{\mathrm{lin}}(M(k), W(k), z)$.

In the following, let A be as in Definition 13 and $\eta^{(0)} \in \mathscr{P}_{\mathrm{fin}}(\mathbb{R})^{\mathcal{S}}$ with $\eta_s^{(0)} = \delta_z$ for all $s \in \mathcal{S}$. Set $M(0) := 1$. The time to calculate $\eta^{(n)}$, denoted by $\mathrm{time}(n)$, is dominated by the applications of Π, thus

$$\mathrm{time}(n) = \mathrm{O}\left(\sum_{k=1}^{n} M(k-1)M(k)\right).$$

Our goal is to analyse the trade-off between time-complexity and approximation errors with respect to metrics $\mathrm{d} = w_\beta, \beta \in (0, \infty)$ and $\mathrm{d} = \ell_\beta, \beta \in [1, \infty)$. For that, let

$$\delta(k) = \delta(\xi^{\mathrm{lin}}(M(k), W(k), z)) = \frac{2W(k)}{M(k) - 1}.$$

In case (A1.B), that is reward distributions have uniformly bounded support $[r_{\min}, r_{\max}]$, the condition

$$\left[\frac{r_{\min}}{1-\gamma}, \frac{r_{\max}}{1-\gamma}\right] \subseteq [z - W(k), z + W(k)] \tag{15}$$

implies that all components of $\eta^*, \eta^{(n)}, n \in \mathbb{N}$ are supported on $[z - W(k), z + W(k)]$ and Lemma 11 yields $\mathrm{PE}(w_\beta, k) \leq 4\delta(k)^{\frac{\beta}{1 \vee \beta}}$ resp. $\mathrm{PE}(\ell_\beta, k) \leq 4\delta(k)^{\frac{1}{\beta}}$. A natural choice for z, W to ensure (15) for all k is $z = \frac{r_{\max} + r_{\min}}{2(1-\gamma)}$ and $W(k) \equiv \frac{r_{\max} - r_{\min}}{2(1-\gamma)}$ (constant).

In presence of unbounded reward distributions, an analysis of the projection errors via Lemma 11 is still possible, but the tail bound terms need more care: in the following, O-notation is with respect to $k \to \infty$; the constant may depend on all quantities except k and concrete O-constants are presented in Appendix A.3, see Theorem 30.

Theorem 14 *(a) It is* $\mathrm{PE}(w_\beta, k) = \mathrm{O}\left(\max\left\{\delta(k)^{\frac{\beta}{1 \vee \beta}}, T(k)\right\}\right)$ *with*

$$T(k) = \begin{cases} W(k)^{-\frac{(\alpha - \beta)}{1 \vee \beta}}, & \text{if } (A1.P\alpha) \text{ holds with } \alpha > \beta, \\ \exp\left(-\frac{\lambda(1-\gamma)}{1 \vee \beta} W(k)\right), & \text{if } (A1.E\lambda) \text{ holds with } \lambda > 0, \\ 0, & \text{if } (A1.B) \text{ and } (15) \text{ hold.} \end{cases}$$

(b) It is $\mathrm{PE}(\ell_\beta, k) = \mathrm{O}\left(\max\left\{\delta(k)^{\frac{1}{\beta}},\ T(k)\right\}\right)$ *with*

$$T(k) = \begin{cases} W(k)^{-(\alpha-\frac{1}{\beta})}, & \text{if } (A1.P\alpha) \text{ holds with } \alpha > 1/\beta, \\ \exp\left(-\lambda(1-\gamma)W(k)\right), & \text{if } (A1.E\lambda) \text{ holds with } \lambda > 0, \\ 0, & \text{if } (A1.B) \text{ and } (15) \text{ hold.} \end{cases}$$

In the following, we only consider $\mathrm{d} = w_\beta$ with some fixed $\beta \in (0, \infty)$. Let

$$n^*(\varepsilon) = \min\{n \in \mathbb{N} \mid \overline{w}_\beta(\eta^{(n)}, \eta^*) \le \varepsilon\}, \ \ \varepsilon > 0.$$

The following result investigates the asymptotic behaviour of $\mathrm{time}(n^*(\varepsilon))$ as $\varepsilon \to 0$ and leads to suggest that letting the size function M increase exponentially (with rate $1/\theta, \theta \in (\gamma^c, 1)$, case (b)), is superior to letting it grow polynomial (case (a)). We write $f(n) = \Theta(g(n))$ if there are constants $C_1, C_2 > 0$ such that $C_1 g(n) \le f(n) \le C_2 g(n)$ for all $n \in \mathbb{N}$.

Theorem 15 *Assume* $(A1.P\alpha)$ *holds with* $\alpha > \beta$ *and let* $h = \frac{1 \vee \beta}{\beta} \frac{\alpha}{\alpha - \beta} > 1$.

(a) Let $r > 0$. *Then*

$$\left\{\begin{aligned} M(n) &= \Theta\left(n^{hr}\right), \\ W(n) &= \Theta\left(n^{\frac{\beta}{\alpha}hr}\right) \end{aligned}\right\} \Longrightarrow \left\{\begin{aligned} \overline{w}_\beta(\eta^{(n)}, \eta^*) &= \mathrm{O}\left(n^{-r}\right) \ \text{as } n \to \infty, \\ n^*(\varepsilon) &= \mathrm{O}\left((1/\varepsilon)^{1/r}\right) \ \text{as } \varepsilon \to 0, \\ \mathrm{time}(n) &= \mathrm{O}\left(n^{2hr+1}\right) \ \text{as } n \to \infty, \\ \mathrm{time}(n^*(\varepsilon)) &= \mathrm{O}\left((1/\varepsilon)^{2h+\frac{1}{r}}\right) \ \text{as } \varepsilon \to 0. \end{aligned}\right\}$$

(b) Let $\theta \in (\gamma^c, 1)$ *with* $c = \min\{1, \beta\}$. *Then*

$$\left\{\begin{aligned} M(n) &= \Theta\left((1/\theta)^{hn}\right), \\ W(n) &= \Theta\left((1/\theta)^{\frac{\beta}{\alpha}hn}\right) \end{aligned}\right\} \Longrightarrow \left\{\begin{aligned} \overline{w}_\beta(\eta^{(n)}, \eta^*) &= \mathrm{O}\left(\theta^n\right) \ \text{as } n \to \infty, \\ n^*(\varepsilon) &\le \frac{\log(\epsilon)}{\log(\theta)} + \mathrm{O}(1) \ \text{as } \varepsilon \to 0, \\ \mathrm{time}(n) &= \mathrm{O}\left((1/\theta)^{2hn}\right) \ \text{as } n \to \infty, \\ \mathrm{time}(n^*(\varepsilon)) &= \mathrm{O}\left((1/\varepsilon)^{2h}\right) \ \text{as } \varepsilon \to 0. \end{aligned}\right\}$$

Proof By Theorem 14(a) we have

$$\mathrm{PE}(w_\beta, k) = \mathrm{O}\left(\max\left\{\left(\frac{W(k)}{M(k)}\right)^{\frac{\beta}{1\vee\beta}},\ W(k)^{-\frac{(\alpha-\beta)}{1\vee\beta}}\right\}\right).$$

Choosing $M(k), W(k)$ as in (a), resp. (b) leads to $\mathrm{PE}(k, w_\beta) = \mathrm{O}\left(k^{-r}\right)$, resp. $\mathrm{PE}(k, w_\beta) = \mathrm{O}\left(\theta^k\right)$. Hence, by Lemma 8, $\overline{w}_\beta(\eta^{(n)}, \eta^*) = \mathrm{O}(n^{-r})$, resp. $\mathrm{O}(\theta^n)$. This immediately leads to the asymptotic bounds for $n^*(\varepsilon)$ in both cases. The time complexity bounds follow from

$$\sum_{k=1}^n (k-1)^{hr} k^{hr} = \Theta(n^{2hr+1}), \ \ \sum_{k=1}^n (1/\theta^h)^{k-1} (1/\theta^h)^k = \Theta\left((1/\theta^h)^{2n}\right),$$

both as $n \to \infty$. Plugging in the obtained asymptotic bounds for $n^*(\varepsilon)$ in the time bound yields the claim. ■

Comparing the asymptotic expressions for $\text{time}(n^*(\varepsilon))$, we find that choosing M, W as in (b), that is enforcing exponential decay with rate $\theta \in (\gamma^c, 1)$ in the projection errors, leads to a slower asymptotic growth of $\text{time}(n^*(\varepsilon))$ as $\varepsilon \to 0$ when compared to enforcing polynomial decay as in (a). This is evidence for favouring exponential decay over polynomial decay. When considering M, W as in (b) but with $\theta \in (0, \gamma^c)$, the same line of reasoning applies, but requiring $\text{time}(n^*(\varepsilon)) = \mathrm{O}((1/\varepsilon)^{h \log(\theta)/\log(\gamma^c)})$, which grows even faster then choosing M, W as in (a). The case $\theta = \gamma^c$ remains open as well as the question which $\theta \in [\gamma^c, 1)$ may be the preferred choice. Experiments, also see Example 2, suggest the heuristic $\theta \approx \frac{\gamma+1}{2}$ (when bounding with a $c = 1$ homogeneous analysis metric).

Remark 16 *Similar conclusions can be obtained when replacing the analysis metric* $\mathrm{d} = w_\beta$ *with* $\mathrm{d} = \ell_\beta, \beta \in [1, \infty)$ *and/or strengthening the moment assumption (A1.Pα) to (A1.Eλ) or even (A1.B). In all of these cases, using the corresponding parts of Theorem 14, it is possible to choose* M, W *to enforce either polynomial or exponential decay of projection errors. A verbatim analysis of* $\text{time}(n^*(\varepsilon))$ *always leads to the same conclusion: constructing algorithms such that projection errors decay exponentially with rate* $\theta \in (\gamma^c, 1)$ *seems to be preferable over polynomial decay.*

We investigated the plain parameter algorithm to collect further evidence that choosing $M(k) = \lceil (1/\theta)^k \rceil$ with $\theta \approx \frac{\gamma+1}{2}$ is a reasonable black-box approach. However, for practical purposes, the plain parameter algorithms is problematic: the k-th update step projects $\mathcal{T}_s \eta^{(k-1)}$ to a distribution supported on $[z - W(k), z + W(k)]$. If the parameters W, z are chosen poorly, it may occur that, for reasonably small k, the interval $[z - W(k), z + W(k)]$ carries only a (very) small amount of mass of $\mathcal{T}_s \eta^{(k-1)}$ which leads to a high projection error. This is (partly) resolved by a modification of the plain parameter algorithm discussed in the next section.

5.2 Adaptive Plain Parameter Algorithm

For practical purposes, we propose a modification of the plain parameter algorithm by replacing $[z - W(k), z + W(k)]$ with a compact interval that contains a *guaranteed* amount of mass of $\mathcal{T}_s \eta^{(k-1)}$. This modification is derived and justified by the following Lemma, shown in Appendix A.4:

Lemma 17 *Let* $\eta = [\eta_{\bar{s}} : \bar{s} \in \mathcal{S}] \in \mathscr{P}(\mathbb{R})^{\mathcal{S}}, \varepsilon_u \in (0, 1/2]$ *and* $s \in \mathcal{S}$. *Define*

$$x_{\min} = \min_{(a,\bar{s})} \Big\{ F^{-1}_{sa\bar{s}}\big(1 - \sqrt{1 - \varepsilon_u}\big) + \gamma \cdot F^{-1}_{\eta_{\bar{s}}}\big(1 - \sqrt{1 - \varepsilon_u}\big) \Big\}, \tag{16}$$

$$x_{\max} = \max_{(a,\bar{s})} \Big\{ F^{-1}_{sa\bar{s}}\big(\sqrt{1 - \varepsilon_u}\big) + \gamma \cdot F^{-1}_{\eta_{\bar{s}}}\big(\sqrt{1 - \varepsilon_u}\big) \Big\}, \tag{17}$$

where both min *and* max *run over all pairs* $(a, \bar{s}) \in \mathcal{A} \times \mathcal{S}$ *with* $\pi(a|s)p(\bar{s}|s, a) > 0$. *Then*

$$\max\Big\{ \mathcal{T}_s \eta(-\infty, x_{\min}),\ \mathcal{T}_s \eta(x_{\max}, \infty) \Big\} \leq \varepsilon_u \quad \textit{and} \quad \mathcal{T}_s \eta[x_{\min}, x_{\max}] \geq 1 - 2\varepsilon_u.$$

As the QF of a given finitely supported distribution can be evaluated, the values $x_{\min}$ and $x_{\max}$ can be calculated under Assumptions (A2)+(A3) for any $\eta \in \mathscr{P}_{\text{fin}}(\mathbb{R})^{\mathcal{S}}$. This leads to the following:

Definition 18 (ADP) *The* adaptive plain parameter algorithm (ADP) *has hyper-parameter functions $M : \mathbb{N} \to \mathbb{N}$, $\mathcal{E}_u : \mathbb{N} \to (0, 1/2]$ with M non-decreasing and $\mathcal{E}_u$ non-increasing and calculates $A(\eta, s, k) \in \Xi_{M(k)}$ as follows:*

1. $(m, \varepsilon_u) = (M(k), \mathcal{E}_u(k))$,
2. *Calculate $x_{\min}$ and $x_{\max}$ as in* (16) *and* (17),
3. $A(\eta, s, k) = \xi^{\text{lin}}(m, \frac{x_{\max} - x_{\min}}{2}, \frac{x_{\max} + x_{\min}}{2})$.

For practical purposes, we suggest to use (ADP) only if assumption *(A1.Pα) is satisfied with a large α*, in which case we recommend the hyper-parameter functions

$$M(k) = \lceil (1/\theta)^k \rceil, \quad \mathcal{E}_u(k) = \frac{1}{2M(k)} \quad \text{with } \theta = \frac{\gamma + 1}{2}. \tag{18}$$

In case (A1.Pα) does not hold for some large α, that is in the presence of heavy-tailed reward distributions (where we recommend to use the QSP algorithm discussed in Section 5.3), the discussion in Appendix A.5 leads to suggest that this choice of hyper-parameters may not yield high quality approximations, which is in line with the results of our controlled experiment in Section 5.4. The reasoning behind recommendation (18) is discussed in Appendix A.5.

5.3 Quantile-Spline Parameter Algorithm

Choosing a parameter algorithm A that produces evenly spaced interpolation points is not necessary to obtain high quality approximations. Aiming to approximate QDP, we now modify the adaptive plain parameter algorithm of Section 5.2 leading to a DDP algorithm that is applicable under (A2) and seems suitable also in presence of heavy-tailed reward distributions.

Recall that QDP updates $\eta \in \mathscr{P}(\mathbb{R})^{\mathcal{S}}$ using $m \in \mathbb{N}$ particles as $\tilde{\eta}_s = \frac{1}{m} \sum_{i=1}^{m} \delta_{x_i}$ with $x_i = F^{-1}_{\mathcal{T}_s \eta}\left(\frac{2i-1}{2m}\right)$. In case $\mathcal{T}_s \eta$ is continuous, we have that $\tilde{\eta}_s = \Pi(\mathcal{T}_s \eta, \tilde{\xi})$ with $\tilde{\xi} = (x_1, \dots, x_m, y_1, \dots, y_{m-1}) \in \Xi_m$ where x_i is as before and $y_i = F^{-1}_{\mathcal{T}_s \eta}\left(\frac{i}{m}\right)$. As explained above, calculating quantiles of $\mathcal{T}_s \eta$ is problematic under the assumption (A2). We suggest to replace $\tilde{\xi} \in \Xi_m$ with an *approximation of quantiles* $\xi \in \Xi_m$ that uses only a controllable number of evaluations of the cdf $F_{\mathcal{T}_s \eta}$. Note that the approximation we suggest takes place on the level of the *parameters*: errors in the quantile approximation $\xi \approx \tilde{\xi}$ may result in a non-optimal parameter, but by calculating $\Pi(\mathcal{T}_s \eta, \xi)$ the approximated quantiles are *reweighed* according to the true distribution $\mathcal{T}_s \eta$.

The suggested approximation of quantiles of $\mathcal{T}_s \eta$ is constructed as follows: first, we calculate the interval $[x_{\min}, x_{\max}] \subset \mathbb{R}$ as in (16) and (17) with $\mathcal{E}_u(n) = \frac{1}{2m}$. Thus, this interval contains at least a mass of $1 - \frac{1}{m}$ of $\mathcal{T}_s \eta$. On that interval, we perform a linear interpolation of $F_{\mathcal{T}_s \eta}$ via $m' \in \mathbb{N}$ evaluations on evenly spaced points in $[x_{\min}, x_{\max}]$. This interpolated CDF is continuous and can easily be inverted; we define ξ to contain quantiles

of this interpolated CDF. Linear interpolation and inversion can be combined in one step, which is step 4 in the following:

Definition 19 (QSP) *The hyper-parameters are* $M : \mathbb{N} \to \mathbb{N}$, $M' : \mathbb{N} \to \mathbb{N}$ *with* M, M' *non-decreasing. The* quantile-spline parameter algorithm (QSP) *calculates* $A(\eta, s, k) \in \Xi_{M(k)}$ *as follows:*

1. $(m, m') = (M(k), M'(k))$,
2. *Calculate* $x_{\min}, x_{\max}$ *as in* (16) *and* (17) *with* $\epsilon_u = \frac{1}{2m}$.
3. *Let* $z_l \leftarrow x_{\min} + (x_{\max} - x_{\min}) \cdot \frac{l}{m'+1}$ *for* $l = 0, 1, \ldots, m' + 1$,
4. *Let* $L : [0, 1] \to \mathbb{R}$ *be the linear spline through the points*[7]

$$(0, z_0)\,,\ (F_{\mathcal{T}_s\eta}(z_1), z_1)\,,\ (F_{\mathcal{T}_s\eta}(z_2), z_2)\,, \ldots,\ (F_{\mathcal{T}_s\eta}(z_{m'}), z_{m'})\,,\ (1, z_{m'+1})\,,$$

5. $A(\eta, s, k) = (x_1, \ldots, x_m, y_1, \ldots, y_{m-1})$ *with* $x_i = L\big(\frac{i-1}{m-1}\big)$ *and* $y_i = L\big(\frac{2i-1}{2m-2}\big)$.

We suggest to choose the hyper-parameter functions M, M' as follows:

$$M(k) = \lceil (1/\theta)^k \rceil,\ \ M'(k) = \lceil (1/4) \cdot (1/\theta)^k \rceil \ \text{ with } \theta = \frac{\gamma + 1}{2}. \tag{19}$$

5.4 Controlled Experiment

We consider two versions of a MDP in which return distributions are known analytically: let $\mathcal{A} = \{a\}$ have one element and $\mathcal{S} = \{1, 2, 3\}$, thus $\pi(a|s) \equiv 1$. State-action-state transitions are deterministic and circular: $p(\bar{s}|s, a) = 1 :\Leftrightarrow (s, \bar{s}) \in \{(1, 2), (2, 3), (3, 1)\}$. Besides s and $\gamma := 0.7$ the return distribution η^*_s depends only on the three reward distributions $\mathcal{L}(R_{1a2})$, $\mathcal{L}(R_{2a3})$, $\mathcal{L}(R_{3a1})$. For all $(s, \bar{s})$ with $(s, \bar{s}) \notin \{(1, 2), (2, 3), (3, 1)\}$ set $\mathcal{L}(R_{sa\bar{s}}) = \delta_0$. For the three relevant reward distributions, we consider two versions. Let $\mathrm{Norm}(\mu, \sigma^2)$ be the normal distribution with expectation $\mu \in \mathbb{R}$ and variance $\sigma^2 \in (0, \infty)$ and let $\mathrm{Cauchy}(\mu, s)$ be the Cauchy distribution with location $\mu \in \mathbb{R}$ and scale $s > 0$. It is $\mathrm{Cauchy}(\mu, s) \in \mathscr{P}_\alpha(\mathbb{R})$ for $\alpha \in (0, 1)$, but $\mathrm{Cauchy}(\mu, s) \notin \mathscr{P}_1(\mathbb{R})$, while $\mathrm{Norm}(\mu, \sigma^2) \in \mathscr{P}_\alpha(\mathbb{R})$ for all $\alpha \in (0, \infty)$.

The two versions we consider are:

$$(i)\quad \begin{pmatrix}\mathcal{L}(R_{1a2})\\ \mathcal{L}(R_{2a3})\\ \mathcal{L}(R_{3a1})\end{pmatrix} = \begin{pmatrix}\mathrm{Norm}(-3, 1)\\ \mathrm{Norm}(5, 2)\\ \mathrm{Norm}(0, 0.5)\end{pmatrix} \quad\Longrightarrow\quad \begin{pmatrix}\mathcal{L}(\eta^*_1)\\ \mathcal{L}(\eta^*_2)\\ \mathcal{L}(\eta^*_3)\end{pmatrix} = \begin{pmatrix}\mathrm{Norm}(0.761, 2.380)\\ \mathrm{Norm}(5.373, 2.816)\\ \mathrm{Norm}(0.533, 1.666)\end{pmatrix},$$

$$(ii)\quad \begin{pmatrix}\mathcal{L}(R_{1a2})\\ \mathcal{L}(R_{2a3})\\ \mathcal{L}(R_{3a1})\end{pmatrix} = \begin{pmatrix}\mathrm{Cauchy}(-3, 0.5)\\ \mathrm{Cauchy}(5, 0.1)\\ \mathrm{Cauchy}(0, 5)\end{pmatrix} \quad\Longrightarrow\quad \begin{pmatrix}\mathcal{L}(\eta^*_1)\\ \mathcal{L}(\eta^*_2)\\ \mathcal{L}(\eta^*_3)\end{pmatrix} = \begin{pmatrix}\mathrm{Cauchy}(0.761, 4.597)\\ \mathrm{Cauchy}(5.373, 5.852)\\ \mathrm{Cauchy}(0.533, 8.218)\end{pmatrix}.$$

We compare three DDP algorithms to approximate (estimate) η^*: The two blackbox algorithms of Section 5 with parameter algorithms ADP (Section 5.2) and QSP (Section 5.3) and hyper-parameter choices as in (18) and (19) and initial approximation $\eta^{(0)} = [\delta_0 : \bar{s} \in$

7. remove all pairs (p_l, z_l) for which there is $l' \in \{0, \ldots, l-1\}$ with $p_{l'} = p_l$.

alg	time(s)	size	n	$\overline{\text{ks}}$	$\overline{w_1}$	$\overline{\ell_2}$	type
ADP	40.17	38868	54	0.0003	0.0007	0.0003	numerical
QSP	34.32	33036	53	0.0002	0.0025	0.0005	numerical
MC	45.00	15588	30	0.0125	0.0296	0.0132	random
MC2	134.11	38868	30	0.0079	0.0168	0.0067	random

Figure 3: Results for mdp (i), where size is the number of stored particles, resp. the number of stored samples in MC estimation. Calculating the next approximation of ADP and QSP exceeds 45 seconds.

alg	time(s)	size	n	$\overline{\text{ks}}$	$\overline{w_1}$	$\overline{\ell_2}$	type
ADP	36.96	38868	54	0.2317	∞	0.6892	numerical
QSP	34.61	33036	53	0.0012	∞	0.0399	numerical
MC	44.99	14643	30	0.0138	∞	0.0967	random
MC2	146.22	38868	30	0.0080	∞	0.1285	random

Figure 4: Results for mdp (ii). Since Cauchy$(\mu, s) \notin \mathscr{P}_1(\mathbb{R})$, the w_1-distances are infinite. However, Cauchy$(\mu, s) \in \mathscr{P}_{1/2}(\mathbb{R}) \subsetneq \mathscr{P}_{\ell_2}(\mathbb{R})$, thus ℓ_2-distances are finite.

$\mathcal{S}$]. We let both algorithms run up to time $t_{\max} = 45$ seconds (on a standard notebook) and return the last completely calculated approximation $\eta^{(n)}$ before that time. Third, we consider a Monte-Carlo estimation (MC), for which we choose $n = 30$ fixed, approximate $\eta^* \approx \mathcal{T}^{\circ n}(\eta^{(0)})$ and estimate the components of the latter by using that, for large N, we have $\mathcal{T}_s^{\circ n}(\eta^{(0)}) \approx \frac{1}{N}\sum_{l=1}^{N} \delta_{G_{s,n-1,l}}$, where $G_{s,n-1,l}, l = 1, 2, \dots$ are iid samples of $\mathcal{T}_s^{\circ n}(\eta^{(0)}) = \mathcal{L}(\sum_{t=0}^{n-1} \gamma^t R^{(t)} | S^{(0)} = s)$. A sample of $\mathcal{T}_s^{\circ n}(\eta^{(0)})$ can be generated by simulating a trajectory of the full MDP started in s up to length n and calculating the γ-discounted sum of rewards along the way. For each s we generate N such samples, where N is the maximal amount that is possible within $t_{\max} = 45$ seconds. For comparison, we repeat the Monte Carlo procedure, but with N chosen such that $3N$ (the amount of generated samples) equals the size of the output of the ADP algorithm. The output of this second Monte Carlo estimation is (MC2).

Each procedure returns some approximation (estimation) $\eta^{(n)}$. We report $\bar{\mathrm{d}}(\eta^{(n)}, \eta^*)$ for $\mathrm{d} = \mathrm{ks}, w_1, \ell_2$ in Figures 3 and 4. For mdp (i) it is seen that both ADP and QSP outperform (MC, MC2) significantly for any analysis metric. The approximation qualities of ADP and QSP are in a comparable range. Things differ for mdp (ii): still, QSP clearly outperforms (MC, MC2) in any metric, but now ADP fails to yield useful approximations, which is to be expected: the heuristics in choosing the hyper-parameters of ADP, see (18), fail for heavily tailed distributions such as the Cauchy distributions.

6 Kolmogorov–Smirnov distance and density approximation

We discussed how to obtain bounds for $\bar{\mathrm{d}}(\eta^{(n)}, \eta^*)$ with respect to a c-homogeneous, regular and convex analysis metric d, for example $\mathrm{d} = w_\beta, \beta \in (0, \infty]$ or $\mathrm{d} = \ell_\beta, \beta \in [1, \infty)$. The Kolmogorov–Smirnov distance $\mathrm{ks} = \ell_\infty$ is not covered by these results. In Section 6.1, we show how to bound $\overline{\mathrm{ks}}(\eta^{(n)}, \eta^*)$ in terms of $\overline{w}_\beta(\eta^{(n)}, \eta^*), \beta \in (0, \infty)$ provided the components η_s^* satisfy certain *continuity properties*, such as possessing a bounded pdf f_s^*. Further, if a pdf f_s^* for η_s^* is known to exists, we suggest to construct an approximation of it by choosing $\delta > 0$ and considering $f_{s,\delta}^{(n)} = f_{\eta_s^{(n)},\delta}$, where, for any $\mu \in \mathscr{P}(\mathbb{R})$ and $\delta > 0$, the probability density $f_{\mu,\delta} : \mathbb{R} \to [0, \infty)$ is defined by

$$f_{\mu,\delta}(x) = \frac{F_\mu(x + \delta) - F_\mu(x - \delta)}{2\delta}, \quad x \in \mathbb{R}.$$

Under suitable regularity assumptions on f_s^*, we further show how the supremum distance between $f_{s,\delta}^{(n)}$ and f_s^* can be bounded in terms of $\overline{\mathrm{ks}}(\eta^{(n)}, \eta^*)$. In Section 6.2, we provide sufficient criteria for the return distribution to posses probability density functions (satisfying the needed regularity properties) that can, in principle, be checked based on the given ingredients of the mdp.

6.1 Uniform Bounds

Let $M > 0, \varrho \in (0, 1]$. A function $g : \mathbb{R} \to \mathbb{R}$ is ϱ-Hölder (continuous) with constant M if $|g(x) - g(y)| \leq M \cdot |x - y|^\varrho$ holds for all $x, y \in \mathbb{R}$. Recall, that a distribution $\nu \in \mathscr{P}(\mathbb{R})$ has pdf $f_\nu : \mathbb{R} \to [0, \infty]$ if for all $x \in \mathbb{R}$ we have $F_\nu(x) = \int_{-\infty}^x f_\nu(y)\mathrm{d}y$. If f_ν is bounded by $M > 0$, then $|F_\nu(x) - F_\nu(y)| \leq M|x - y|$ and thus F_ν is $\varrho = 1$-Hölder continuous with constant M.

Part (a) of the following Lemma is a slight generalisation of Lemma 5.1 in Fill and Janson (2002), part (b) follows from Lemma 2.7 in Knape and Neininger (2008):

Lemma 20 *Let $\mu, \nu \in \mathscr{P}(\mathbb{R})$.*

(a) If F_ν is ϱ-Hölder with constant M, then for every $\beta \in (0, \infty)$

$$\mathrm{ks}(\mu, \nu) \leq a^{-a} \cdot M^{1-a} \cdot w_\beta(\mu, \nu)^{a(1 \vee \beta)} \quad \text{with } a = \frac{\varrho}{\varrho + \beta}. \tag{20}$$

In particular, if ν has a pdf f_ν bounded by M, then (20) holds with $\varrho = 1$.

(b) If ν has a pdf f_ν that is τ-Hölder with constant C, then for every $\delta \in (0, \infty)$

$$\sup_{x \in \mathbb{R}} |f_{\mu,\delta}(x) - f_\nu(x)| \leq \frac{1}{\delta}\mathrm{ks}(\mu, \nu) + C \cdot \delta^\tau.$$

If, additionally, f_ν is bounded by M, then $\mathrm{ks}(\mu, \nu)$ in the upper bound can be further bounded by (20) with $\varrho = 1$.

Proof For (a) we note that Lemma 5.1 in Fill and Janson (2002) is the case in which ν has a density bounded by M, which they used to bound $\max_{|x-y| \leq \epsilon} |F_\nu(x) - F_\nu(y)| \leq M\varepsilon^\varrho$

with $\varrho = 1$. It is easy to see that their proof extends to general $\rho > 0$. For (b), see Lemma 2.7 in Knape and Neininger (2008). ■

By passing to supremum distances over states, Lemma 20 yields the following:

Corollary 21 *Let $\eta^{(n)} = [\eta_s^{(n)} : s \in \mathcal{S}]$ be any approximation of $\eta^* = [\eta_s^* : s \in \mathcal{S}]$.*

(a) If every $F_{\eta_s^*}$ is ϱ-Hölder with constant M, then for every $\beta \in (0, \infty)$

$$\overline{\mathrm{ks}}(\eta^{(n)}, \eta^*) \le a^{-a} \cdot M^{1-a} \cdot \overline{w}_\beta(\eta^{(n)}, \eta^*)^{a(1\vee\beta)} \quad \text{with } a = \frac{\varrho}{\varrho + \beta}. \tag{21}$$

In particular, if every η_s^* has a pdf f_s^* bounded by M, then (21) holds with $\varrho = 1$.

(b) If every η_s^* has a pdf f_s^* that is τ-Hölder with constant C, then for every $\delta \in (0, \infty)$

$$\max_{s\in\mathcal{S}} \sup_{x\in\mathbb{R}} |f_{s,\delta}^{(n)}(x) - f_s^*(x)| \le \frac{1}{\delta}\overline{\mathrm{ks}}(\eta^{(n)}, \eta^*) + C \cdot \delta^\tau. \tag{22}$$

If, additionally, every f_s^* is bounded by M, then $\overline{\mathrm{ks}}(\eta^{(n)}, \eta^*)$ in the upper bound can be further bounded by (21) with $\varrho = 1$.

Remark 22 *The approximation of densities in supremum norm has also an application in nonuniform random number generation in the context of von Neumann's rejection method; for an introduction see Section II.3 of Devroye (1986). When applying the rejection method it may not be possible to evaluate a given probability density to decide about rejection resp. acceptance of a sample, e.g., the density may only be given implicitly as the density of the fixed-point of a DBO. Then, to make the rejection method applicable it is sufficient (besides constructing an appropriate dominating density) to be able to approximate the given density to arbitrary precision in finite time as demonstrated in Devroye et al. (2000); Devroye and Neininger (2002). Since the big-*O *constants in our Corollary 21 can be made explicit, the bound (22) is sufficient for the purpose of the rejection method. Note that for perfect simulation from fixed-points of the DBO also coupling from the past algorithms may, in principle, be applicable, see Devroye and James (2011); Dadoun and Neininger (2014).*

For other potential use of Corollary 21 in the context of certain statistical functionals see Chapter 20 of van der Vaart (1998).

6.2 Sufficient criteria for existence of return densities with properties

We provide sufficient criteria for when the return distribution components possess pdfs, resp. bounded, resp. Hölder continuous pdfs. Let $s_0 \in \mathcal{S}$ and $(S^{(t)}, A^{(t)}, R^{(t)})_{t\in\mathbb{N}_0}$ be the full MDP. We write $\mathbb{P}_{s_0}$ instead of $\mathbb{P}[\,\cdot\,|S^{(0)} = s_0]$, similar $\mathbb{E}_{s_0}$. In particular, we have $\eta_{s_0}^* = \mathbb{P}_{s_0}[G \in \cdot\,]$ with $G = \sum_t \gamma^t R^{(t)}$. Further, fix an arbitrary subset $\mathcal{G} \subseteq \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ and define the $\mathbb{N}_0 \cup \{\infty\}$-valued random hitting time

$$T = \min\big\{t \in \mathbb{N}_0 \;\big|\; (S^{(t)}, A^{(t)}, S^{(t+1)}) \in \mathcal{G}\big\},$$

with $\min\emptyset := \infty$. Note that $\mathbb{E}_{s_0}[T] < \infty$ implies $\mathcal{G} \neq \emptyset$ and $\mathbb{P}_{s_0}[T \in \mathbb{N}] = 1$.

Theorem 23 *Suppose* $\mathbb{E}_{s_0}[T] < \infty$ *and that for every* $(s, a, \bar{s}) \in \mathcal{G}$ *the distribution* $r(\,\cdot\,|s, a, \bar{s})$ *has a pdf* $f_{(s,a,\bar{s})}$*. Then* $\eta^*_{s_0}$ *has pdf* $f^*_{s_0}$ *given by*

$$f^*_{s_0}(x) = \mathbb{E}_{s_0}\Big[\frac{1}{\gamma^T} f_{(S,A,\bar{S})}\left(\frac{x - Z}{\gamma^T}\right)\Big],\ x \in \mathbb{R}, \tag{23}$$

where $(S, A, \bar{S}) = (S^{(T)}, A^{(T)}, S^{(T+1)}) \in \mathcal{G}$ *and* $Z = \sum_{t\in\mathbb{N}_0\setminus\{T\}} \gamma^t R^{(t)}$*.*

The proof of Theorem 23 can be found in the Appendix A.6. From (23) one can deduce sufficient criteria for when $\eta^*_{s_0}$ has a bounded, resp. Hölder continuous pdf. Let $\Psi(z) = \mathbb{E}_{s_0}[e^{zT}], z \in \mathbb{R}$ be the moment generating function of T. Note that we may have $\Psi(z) = \infty$ and that $\Psi(z) < \infty$ for some $z > 0$ implies $\mathbb{E}_{s_0}[T] < \infty$.

Corollary 24 *In the setting of Theorem 23, let* $f^*_{s_0}$ *be the pdf of* $\eta^*_{s_0}$ *given by* (23)*.*

(a) If every $f_{(s,a,\bar{s})}, (s, a, \bar{s}) \in \mathcal{G}$ *is bounded and* $\Psi\big(\log(1/\gamma)\big) < \infty$*, then* $f^*_{s_0}$ *is bounded.*

(b) If every $f_{(s,a,\bar{s})}, (s, a, \bar{s}) \in \mathcal{G}$ *is* τ*-Hölder continuous and* $\Psi\big((\tau+1)\log(1/\gamma)\big) < \infty$*, then* $f^*_{s_0}$ *is* τ*-Hölder continuous.*

The proof of Corollary 24 is contained in Appendix A.6.

Example 3 *Consider* $\mathcal{S} = \{0, 1\}, \mathcal{A} = \{0\}, s_0 = 0$ *and, for all* $(s, a) \in \mathcal{S}\times\mathcal{A}$*, the reward distributions* $r(\,\cdot\,|s, a, 0) = \delta_0$ *and let* $r(\,\cdot\,|s, a, 1)$ *have pdf* f*. Consider* $\mathcal{G} = \{(0, 0, 1), (1, 0, 1)\}$ *and let* $\alpha = p(1|0, 0) \in (0, 1)$*. It then holds that* $\mathbb{P}_{s_0}[T = k] = \alpha(1 - \alpha)^k, k \in \mathbb{N}_0$*, that is* T *has a geometric distribution with parameter* α*. In particular,* $\mathbb{E}_{s_0}[T] < \infty$ *and hence, by Theorem 23,* $\eta^*_{s_0}$ *has pdf* $f^*_{s_0}$ *given by* (23)*, note that* $f_{(S,A,\bar{S})} = f$ *in this case. By geometric series arguments,* $\Psi(z) < \infty \Leftrightarrow e^z(1 - \alpha) < 1$ *and hence, by Corollary 24,*

- f *bounded and* $(1 - \alpha) < \gamma \implies f^*_{s_0}$ *is bounded,*
- f τ*-Hölder continuous and* $(1 - \alpha) < \gamma^{\tau+1} \implies f^*_{s_0}$ *is* τ*-Hölder continuous.*

The presented conditions for $\eta^*_{s_0}$ to have a pdf (with properties) are not necessary, as illustrated by the following example, which also shows that finding *sufficient and necessary* conditions for $\eta^*_{s_0}$ to possess a pdf (with properties) is a hard problem in general, we refer to Section 2.5 of Diaconis and Freedman (1999) for discussions of related examples.

Example 4 *Let* $\mathcal{S} = \mathcal{A} = \{0\}$*,* $r(\,\cdot\,|0, 0, 0) = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$ *(fair coin toss) and* $\gamma = \frac{1}{2}$*. Then* η^*_0 *is continuous uniform on* $[0, 2]$*, hence possesses a bounded pdf, although the reward distribution does not have a pdf. When the reward distribution is changed to* $r(\,\cdot\,|0, 0, 0) = q\delta_0 + (1 - q)\delta_1$ *with* $q \in (0, 1), q \neq \frac{1}{2}$*, then* $\eta^*_{s_0}$ *does not possess a pdf, although still having a Hölder continuous cdf.*

Acknowledgments and Disclosure of Funding

The first author was partially supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - 502386356.

Appendix A. Appendix

A.1 Characterisation of $\mathscr{P}_{\ell_\beta}(\mathbb{R})$ for $\beta \in [1,\infty)$

Proposition 25 *$\mathscr{P}_{\ell_1}(\mathbb{R}) = \mathscr{P}_1(\mathbb{R})$ and for every $\beta \in (1,\infty)$ it holds that*

$$\mathscr{P}_{\frac{1}{\beta}}(\mathbb{R}) \subsetneq \mathscr{P}_{\ell_\beta}(\mathbb{R}) \subsetneq \bigcap\nolimits_{0<\varepsilon<\frac{1}{\beta}} \mathscr{P}_{\frac{1}{\beta}-\varepsilon}(\mathbb{R}).$$

Proof Let $\mu = \mathcal{L}(X) \in \mathscr{P}(\mathbb{R})$. It is $\mathbb{E}[|X|^\alpha] = \alpha \int_0^\infty x^{\alpha-1}\mathbb{P}[|X|>x]\mathrm{d}x$ for every $\alpha \in (0,\infty)$ and $\ell_\beta(\mu,\delta_0)^\beta = \int_0^\infty \left(\mathbb{P}[X<x]^\beta + \mathbb{P}[X>x]^\beta\right)\mathrm{d}x$ for every $\beta \in [1,\infty)$. Further, for every $x>0$ it is $\mathbb{P}[X<x]^\beta + \mathbb{P}[X>x]^\beta \le \mathbb{P}[|X|>x]^\beta \le 2^{\beta-1}\left(\mathbb{P}[X<x]^\beta + \mathbb{P}[X>x]^\beta\right)$ and thus

$$\mu \in \mathscr{P}_{\ell_\beta}(\mathbb{R}) \Leftrightarrow \int_0^\infty \mathbb{P}[|X|>x]^\beta \mathrm{d}x < \infty \text{ and } \mu \in \mathscr{P}_\alpha(\mathbb{R}) \Leftrightarrow \int_0^\infty x^{\alpha-1}\mathbb{P}[|X|>x]\mathrm{d}x < \infty.$$

The conditions are equivalent in case $\beta=\alpha=1$ and $\mathscr{P}_{\ell_1}(\mathbb{R}) = \mathscr{P}_1(\mathbb{R})$ follows immediately. Now, let $\beta \in (1,\infty)$. For the first inclusion, let $\mu \in \mathscr{P}_{\frac{1}{\beta}}(\mathbb{R})$. By Markov's inequality, for every $x>0$ it is $\mathbb{P}[|X|>x] \le \mathbb{E}[|X|^{\frac{1}{\beta}}]x^{-\frac{1}{\beta}}$. Using $\beta-1\ge 0$ and $\frac{-(\beta-1)}{\beta} = \frac{1}{\beta}-1$ leads to

$$\begin{aligned}\int_0^\infty \mathbb{P}[|X|>x]^\beta \mathrm{d}x &= \int_0^\infty \mathbb{P}[|X|>x]^{\beta-1}\cdot\mathbb{P}[|X|>x]\,\mathrm{d}x \\ &\le \int_0^\infty \left(\mathbb{E}[|X|^{\frac{1}{\beta}}]x^{-\frac{1}{\beta}}\right)^{\beta-1}\cdot\mathbb{P}[|X|>x]\,\mathrm{d}x = \beta\cdot\mathbb{E}[|X|^{\frac{1}{\beta}}]^\beta,\end{aligned}$$

which is $<\infty$ by assumption. For the second inclusion, let $\mu \in \mathscr{P}_{\ell_\beta}(\mathbb{R})$ and $\varepsilon \in (0,\frac{1}{\beta})$. Using the integral formula for $\mathbb{E}[|X|^{\frac{1}{\beta}-\varepsilon}]$ and applying Hölder's inequality with $p = \frac{\beta}{\beta-1}$ and $q=\beta$ gives

$$\begin{aligned}\left(\frac{1}{\beta}-\varepsilon\right)^{-1}\cdot\mathbb{E}[|X|^{\frac{1}{\beta}-\varepsilon}] &= \int_0^\infty x^{\frac{1}{\beta}-\varepsilon-1}\cdot\mathbb{P}[|X|>x]\,\mathrm{d}x \\ &\le \left[\int_0^\infty x^{\left[\frac{1}{\beta}-\varepsilon-1\right]\frac{\beta}{\beta-1}}\mathrm{d}x\right]^{\frac{\beta-1}{\beta}}\cdot\left[\int_0^\infty \mathbb{P}[|X|>x]^\beta\mathrm{d}x\right]^{\frac{1}{\beta}}.\end{aligned}$$

The first factor is $<\infty$ because $\left[\frac{1}{\beta}-\varepsilon-1\right]\frac{\beta}{\beta-1} < -1$ and the second by assumption. To see that both inclusions are strict, note that $\int_2^\infty x^{-1}\log(x)^{-c}\mathrm{d}x < \infty$ if and only if $c>1$. Thus, if $\mu \in \mathscr{P}(\mathbb{R})$ has tail probabilities $\mathbb{P}[|X|>x] = x^{-\frac{1}{\beta}}\log(x)^{-1}$ for $x\ge 2$ it is $\mu \in \mathscr{P}_{\ell_\beta}(\mathbb{R})$, but $\mu \notin \mathscr{P}_{\frac{1}{\beta}}(\mathbb{R})$. Similar, if $\mathbb{P}[|X|>x] = x^{-\frac{1}{\beta}}\log(x)^{-\frac{1}{\beta}}$ for $x\ge 2$, it is $\mu \notin \mathscr{P}_{\ell_\beta}(\mathbb{R})$ but $\mu \in \mathscr{P}_{\frac{1}{\beta}-\varepsilon}$ for every $\varepsilon \in (0,\frac{1}{\beta})$. ∎

A.2 Definition of c-homogeneous, regular and convex metric

For $a \in \mathbb{R}$ and $\mu = \mathcal{L}(X) \in \mathscr{P}(\mathbb{R})$ let $a^\#\mu := \mathcal{L}(aX)$. Further, for $\mu,\nu \in \mathscr{P}(\mathbb{R})$ let $\mu * \nu \in \mathscr{P}(\mathbb{R})$ be the convolution.

Definition 26 *Let $c \in (0,\infty), p \in [1,\infty)$. An extended metric $\mathrm{d} : \mathscr{P}(\mathbb{R}) \times \mathscr{P}(\mathbb{R}) \to [0,\infty]$ is called*

- *c-homogeneous if for all $\gamma \in [0,1]$ and $\mu,\nu \in \mathscr{P}(\mathbb{R})$ it is $\mathrm{d}(\gamma^{\#}\mu, \gamma^{\#}\nu) = \gamma^c \mathrm{d}(\mu,\nu)$.*
- *regular if for all $\mu,\nu,\vartheta \in \mathscr{P}(\mathbb{R})$ it is $\mathrm{d}(\mu * \vartheta, \nu * \vartheta) \le \mathrm{d}(\mu,\nu)$.*
- *p-convex if for all $m \in \mathbb{N}$ and $(\alpha_1,\dots,\alpha_m) \in [0,1]^m$ with $\sum_{i=1}^m \alpha_i = 1$ and $(\mu_1,\dots,\mu_m)$, $(\nu_1,\dots,\nu_m) \in \mathscr{P}(\mathbb{R})^m$ it holds that $\mathrm{d}^p\left(\sum_{i=1}^m \alpha_i\mu_i, \sum_{i=1}^m \alpha_i\nu_i\right) \le \sum_{i=1}^m \alpha_i \mathrm{d}^p(\mu_i,\nu_i)$.*
- *convex if there exists $p \in [1,\infty)$ such that d is p-convex.*

In classical literature on probability metrics such as Rachev (1991) the properties c-homogeneous plus regular are called $(c,+)$-ideal.

A.3 Proof of Theorem 14

For $\xi = (x_1,\dots,x_m,y_1,\dots,y_{m-1}) \in \Xi_m$ define the map

$$\mathrm{pr}(\,\cdot\,|\xi) : \mathbb{R} \to \{x_1,\dots,x_m\}, \;\; \mathrm{pr}(x|\xi) = x_i :\Leftrightarrow x \in (y_{i-1}, y_i].$$

Further, the *push-forward* of a measurable map $f : \mathbb{R} \to \mathbb{R}$ is denoted by $f^{\#} : \mathscr{P}(\mathbb{R}) \to \mathscr{P}(\mathbb{R})$ defined as $f^{\#}(\mu) = \mu \circ f^{-1}$. That is, $\mu = \mathcal{L}(X) \Rightarrow f^{\#}(\mu) = \mathcal{L}(f(X))$.

Lemma 27 *For every $\mu \in \mathscr{P}(\mathbb{R})$ and $\xi = (x_1,\dots,x_m,y_1,\dots,y_{m-1}) \in \Xi_m$ it holds that*

(a) $\Pi(\mu,\xi) = \mathrm{pr}^{\#}(\mu|\xi)$

(b) $F_{\Pi(\mu,\xi)} = \sum_{i=1}^{m+1} F_\mu(y_{i-1}) 1_{[x_{i-1},x_i)}(\,\cdot\,)$,

(c) $F^{-1}_{\Pi(\mu,\xi)} = \mathrm{pr}(F_\mu^{-1}(\,\cdot\,)|\xi)$,

where we set $y_0 = x_0 = -\infty, y_m = x_{m+1} = \infty, F_\mu(-\infty) = 0, F_\mu(\infty) = 1$.

Proof (a) is obvious from the definitions and (b) follows from (a) applied to sets of the form $(-\infty,x], x \in \mathbb{R}$. (c) follows from the fact the inverse quantile function F_μ^{-1} is the unique left-continuous non-decreasing function $(0,1) \to \mathbb{R}$ such that $\mathcal{L}(F_\mu^{-1}(U)) = \mu$ holds for U continuous uniform on $(0,1)$. Now, $\mathrm{pr}(F_\mu^{-1}(U)|\xi)$ has distribution $\mathrm{pr}^{\#}(\mu|\xi)$ and $\mathrm{pr}(F_\mu^{-1}(\cdot)|\xi)$ is left-continuous and non-decreasing, because both $\mathrm{pr}(\cdot|\xi)$ and F_μ^{-1} are. ■

For $\mu = \mathcal{L}(X) \in \mathscr{P}(\mathbb{R}), z \in \mathbb{R}, w, \beta \in (0,\infty)$ let $\mathrm{tail}_{w_\beta}(\mu,z,w) = \int_0^\infty x^{\beta-1}\mathbb{P}[|X-z| > w + x]\mathrm{d}x$ and $\mathrm{tail}_{\ell_\beta}(\mu,z,w) = \int_0^\infty \mathbb{P}[|X-z| > w+x]^\beta \mathrm{d}x$. In particular, $\mathrm{tail}_{w_\beta}(\mu,\xi) = \mathrm{tail}_{w_\beta}(\mu, z(\xi), w(\xi))$, similar for $\mathrm{tail}_{\ell_\beta}$.

Proof [Proof of Lemma 11] (a) It is $w_\beta(\Pi(\mu,\xi),\mu) = \mathbb{E}[|\mathrm{pr}(X|\xi) - X|^\beta]^{\frac{1}{1\vee\beta}}$ by Lemma 27. Further, $x \in [x_1,x_m]$ implies $|\mathrm{pr}(x|\xi) - x| \le 2\delta(\xi)$. Writing $1 \equiv 1(X \in [x_1,x_m]) + 1(X <$

$x_1) + 1(X > x_m)$, using linearity of expectation and the formula $\mathbb{E}[Z^\beta] = \beta \int_0^\infty x^{\beta-1}\mathbb{P}[Z > x]\mathrm{d}x$ we have

$$\begin{aligned} w_\beta\big(\Pi(\mu,\xi),\mu\big)^{1\vee\beta} &= \mathbb{E}[|\mathrm{pr}(X|\xi) - X|^\beta] \\ &\le 2^\beta\delta(\xi)^\beta + \mathbb{E}[(x_1 - X)^\beta 1(X < x_1)] + \mathbb{E}[(X - x_m)^\beta 1(X > x_m)] \\ &= 2^\beta\delta(\xi)^\beta + \beta\int_0^\infty x^{\beta-1}\left(\mathbb{P}[X < x_1 - x] + \mathbb{P}[X > x_m + x]\right)\mathrm{d}x \\ &= 2^\beta\delta(\xi)^\beta + \beta\int_0^\infty x^{\beta-1}\mathbb{P}\left[|X - z(\xi)| > w(\xi) + x\right]\mathrm{d}x \\ &= 2^\beta\left[\delta(\xi)^\beta + \frac{\beta}{2^\beta}\mathrm{tail}_{w_\beta}(\mu,\xi)\right] \le 2^{\beta+1}\max\{\delta(\xi)^\beta, \frac{\beta}{2^\beta}\mathrm{tail}_{w_\beta}(\mu,\xi)\}. \end{aligned}$$

Taking both sides to the power of $1/(1\vee\beta)$ and using $2^{(\beta+1)/(1\vee\beta)} \le 4$ and $(\frac{\beta}{2^\beta})^{1/(1\vee\beta)} \le 1$ yields the claim.

(b) By Lemma 27 the cdf of $\Pi(\mu,\xi)$ satisfies $x \in [x_{i-1}, x_i) \Rightarrow F_{\Pi(\mu,\xi)}(x) = F_\mu(y_{i-1})$ for every $i = 1,\dots,m+1$ where $x_0 = -\infty, x_{m+1} = \infty, y_0 = -\infty, y_m = \infty$. Hence

$$\begin{aligned} \ell_\beta\big(\Pi(\mu,\xi),\mu\big)^\beta &= \int_{-\infty}^\infty |F_\mu(x) - F_{\Pi(\mu,\xi)}(x)|^\beta \mathrm{d}x \\ &= \sum_{i=1}^{m+1}\int_{x_{i-1}}^{x_i} |F_\mu(x) - F_\mu(y_{i-1})|^\beta \mathrm{d}x \\ &\le \int_{-\infty}^{x_1} F_\mu(x)^\beta \mathrm{d}x + \int_{x_m}^\infty (1 - F_\mu(x))^\beta \mathrm{d}x + \sum_{i=2}^m (x_i - x_{i-1})|F_\mu(x_i) - F_\mu(x_{i-1})|^\beta \\ &\le \int_{-\infty}^{x_1} F_\mu(x)^\beta \mathrm{d}x + \int_{x_m}^\infty (1 - F_\mu(x))^\beta \mathrm{d}x + \delta(\xi)\cdot\max_{2\le i\le m}|F_\mu(x_i) - F_\mu(x_{i-1})|^{\beta-1} \\ &\le 2\int_0^\infty \mathbb{P}\left[|X - z(\xi)| > w(\xi) + x\right]^\beta \mathrm{d}x + \delta(\xi) \le 4\max\{\delta(\xi), \mathrm{tail}_{\ell_\beta}(\mu,\xi)\}. \end{aligned}$$

Taking both sides to the power of $(1/\beta)$ yields the claim. ■

Let $\eta^{(k)}, k \in \mathbb{N}$ be calculated using (PPA) with hyper-parameter (functions) M, W, z and let $\xi^{(k)} = \xi^{\mathrm{lin}}(M(k), W(k), z)$. Assume $\eta_s^{(0)} = \delta_z$.

Lemma 28 *Let* $\mathrm{tail} = \mathrm{tail}_{w_\beta}, \beta \in (0,\infty)$ *or* $\mathrm{tail} = \mathrm{tail}_{\ell_\beta}$ *with* $\beta \in [1,\infty)$. *Then for all* $s \in \mathcal{S}$ *and* $k \in \mathbb{N}$ *it is*

$$\mathrm{tail}(\mathcal{T}_s\eta^{(k-1)}, \xi^{(k)}) \le \max_{(a,\bar s)\in\mathcal{A}\times\mathcal{S}} \mathrm{tail}\Big(\mathcal{L}(R_{sa\bar s}), (1-\gamma)z, (1-\gamma)W(k)\Big).$$

Proof Write $\eta^{(k-1)} = [\mathcal{L}(G_{k-1,\bar s}) : \bar s \in \mathcal{S}]$ with $(G_{k-1,\bar s})_{\bar s}$ independent of $(R_{sa\bar s})_{sa\bar s}$ and the full mdp $(S^{(t)}, A^{(t)}, R^{(t)})_{t\in\mathbb{N}_0}$, thus $\mathcal{L}(R^{(0)} + \gamma G_{k-1,S^{(1)}}|S^{(0)} = s) = \mathcal{T}_s\eta^{(k-1)}$. It holds that

$$\mathbb{P}[|R^{(0)} + \gamma G_{k-1,S^{(1)}} - z| > W(k) + x|S^{(0)} = s] \le \max_{a,\bar s}\mathbb{P}[|R_{sa\bar s} + \gamma G_{k-1,\bar s} - z| > W(k) + x].$$

Since $\eta_s^{(k-1)}$ is the output of the previous step of the algorithm, $G_{k-1,\bar{s}}$ is concentrated on $[z-W(k-1), z+W(k-1)] \subseteq [z-W(k), z+W(k)]$, note that we assume $W(k-1) \leq W(k)$. Writing $0 = \gamma z - \gamma z$, triangle inequality yields $|R_{sa\bar{s}} + \gamma G_{k-1,\bar{s}} - z| \leq |R_{sa\bar{s}} - (1-\gamma)z| + \gamma W(k)$. Hence for every $x > 0$ it is

$$\mathbb{P}\left[|(R_{sa\bar{s}} + \gamma G_{k-1,\bar{s}}) - z| > W(k) + x\right] \leq \mathbb{P}\left[|R_{sa\bar{s}} - (1-\gamma)z| > (1-\gamma)W(k) + x\right].$$

Combining these bounds yields the claim. ■

Let $B(\cdot,\cdot)$ be the Beta function and $\Gamma(\cdot)$ the Γ-function

Lemma 29 *Let $\mu = \mathcal{L}(X) \in \mathscr{P}(\mathbb{R})$ and $z \in \mathbb{R}, w > 0$.*

(a)

$$\begin{aligned} 0 < \beta < \alpha \quad &\Longrightarrow \quad \mathrm{tail}_{w_\beta}(\mu, z, w) \leq \frac{B(\beta, \alpha-\beta)\mathbb{E}[|X-z|^\alpha]}{w^{\alpha-\beta}} \\ \lambda > 0, \beta > 0 \quad &\Longrightarrow \quad \mathrm{tail}_{w_\beta}(\mu, z, w) \leq \frac{\Gamma(\beta)\mathbb{E}[\exp(\lambda|X-z|)]}{\lambda^\beta \exp(\lambda w)} \\ \mathbb{P}[|X-z| \leq w] = 1 \quad &\Longrightarrow \quad \mathrm{tail}_{w_\beta}(\mu, z, w) = 0. \end{aligned}$$

(b)

$$\begin{aligned} 1 \leq \frac{1}{\beta} < \alpha \quad &\Longrightarrow \quad \mathrm{tail}_{\ell_\beta}(\mu, z, w) \leq \frac{\mathbb{E}[|X-z|^\alpha]^\beta}{(\alpha\beta-1)w^{\alpha\beta-1}} \\ \lambda > 0, \beta > 0 \quad &\Longrightarrow \quad \mathrm{tail}_{\ell_\beta}(\mu, z, w) \leq \frac{\mathbb{E}[\exp(\lambda|X-z|)]^\beta}{\lambda\beta \exp(\lambda\beta w)} \\ \mathbb{P}[|X-z| \leq w] = 1 \quad &\Longrightarrow \quad \mathrm{tail}_{\ell_\beta}(\mu, z, w) = 0. \end{aligned}$$

Proof In both (a) and (b) the third implication is obvious. For the other implications, apply Markov's inequality as $\mathbb{P}[|X-z| > w + x] \leq \mathbb{E}[h(|X-z|)]/h(w+x)$ with functions $h(y) = y^\alpha$ (first implications in (a)+(b)) and $h(y) = \exp(\lambda y)$ (second implications in both (a) and (b)) and then explicitly calculate the bounding integrals. For the first implication in (a) notice that $\int_0^\infty \frac{x^{\beta-1}}{(w+x)^\alpha} \mathrm{d}x = B(\beta, \alpha-\beta) w^{-(\alpha-\beta)}$, the other integrals are more elementary. ■

Still, let $\eta^{(k)}, k \in \mathbb{N}$ be the output of PPA with parameter functions M, W, z and initial approximations $\eta_s^{(0)} = \delta_z$. In the following, let $\alpha, \lambda \in (0, \infty)$ and define

$$P(z,\alpha) = \max_{sa\bar{s}} \mathbb{E}[|R_{sa\bar{s}} - (1-\gamma)z|^\alpha] \text{ and } E(z,\lambda) = \max_{sa\bar{s}} \mathbb{E}[\exp(\lambda|R_{sa\bar{s}} - (1-\gamma)z|)].$$

That is, (A1.Pα) holds iff $P(z,\alpha) < \infty$ and (A1.Eλ) holds iff $E(z,\lambda) < \infty$. Further, set

$$\delta(k) = \frac{2W(k)}{M(k)-1}.$$

Theorem 30 *(a) It is* $\mathrm{PE}(w_\beta, k) \le 4 \cdot \max\left\{\delta(k)^{\frac{\beta}{1\vee\beta}},\, T(k)\right\}$ *with*

- $T(k) \le [B(\beta, \alpha-\beta)P(z,\alpha)]^{\frac{1}{1\vee\beta}} \cdot [(1-\gamma)W(k)]^{-\frac{\alpha-\beta}{1\vee\beta}}$ *in case* $\beta \in (0,\alpha)$,
- $T(k) \le \left[\Gamma(\beta)\lambda^{-\beta}E(z,\lambda)\right]^{\frac{1}{1\vee\beta}} \cdot \exp\left(-\frac{\lambda(1-\gamma)}{1\vee\beta}W(k)\right)$ *in case* $\beta \in (0,\infty)$,
- $T(k) = 0$ *in case* $\mathbb{P}[\frac{R_{sa\bar{s}}}{1-\gamma} \in [z-W(k), z+W(k)]] = 1$ *for all* $s, a, \bar{s}$.

(b) It is $\mathrm{PE}(\ell_\beta, k) \le 4 \cdot \max\left\{\delta(k)^{\frac{1}{\beta}},\, T(k)\right\}$ *with*

- $T(k) \le \left[(\alpha\beta-1)^{-1/\beta}P(z,\alpha)\right] \cdot [(1-\gamma)W(k)]^{-(\alpha-\frac{1}{\beta})}$ *in case* $\beta \ge 1, \beta > 1/\alpha$,
- $T(k) \le \left[(\lambda\beta)^{-1/\beta}E(z,\lambda)\right] \cdot \exp\left(-\lambda(1-\gamma)W(k)\right)$ *in case* $\beta \ge 1$,
- $T(k) = 0$ *in case* $\mathbb{P}[\frac{R_{sa\bar{s}}}{1-\gamma} \in [z-W(k), z+W(k)]] = 1$ *for all* $s, a, \bar{s}$.

Proof We consider only $\mathrm{d} = w_\beta$, the case $\mathrm{d} = \ell_\beta$ can be shown similarly. Applying Lemma 11 leads to

$$\mathrm{PE}(w_\beta, k) = \max_s w_\beta(\Pi(\mathcal{T}_s\eta^{(k-1)}, \xi^{(k)}), \mathcal{T}_s\eta^{(k-1)}) \le 4 \cdot \max\left\{\delta(k)^{\frac{\beta}{1\vee\beta}},\, T(k)\right\}$$

with $T(k) = \max_s \mathrm{tail}_{w_\beta}(\mathcal{T}_s\eta^{(k-1)}, \xi^{(k)})^{\frac{1}{1\vee\beta}}$. Applying Lemma 28 leads to

$$T(k)^{1\vee\beta} \le \max_{s,a,\bar{s}} \mathrm{tail}_{w_\beta}\Big(\mathcal{L}(R_{sa\bar{s}}), (1-\gamma)z, (1-\gamma)W(k)\Big).$$

The claimed w_β-bounds follow from Lemma 29. ∎

A.4 Proof of Lemma 17

Let $\mu = \mathcal{L}(X) \in \mathscr{P}(\mathbb{R})$ and $u \in (0,1)$. Some $x \in \mathbb{R}$ is a u-quantile of μ, that is x satisfies $\mathbb{P}[X < x] \le u \le \mathbb{P}[X \le x]$, if and only if $x \in [F_\mu^{-1}(u), F_\mu^{+1}(u)]$, where the function $F_\mu^{+1} : (0,1) \to \mathbb{R}$ is defined by $F_\mu^{+1}(u) = \sup\{x \in \mathbb{R} | \lim_{y\uparrow x} F_\mu(y) \le u\}$.

Lemma 31 *Let* $\mu, \nu \in \mathscr{P}(\mathbb{R})$ *with convolution* $\mu * \nu \in \mathscr{P}(\mathbb{R})$ *and let* $u \in (0,1)$.

$$F_{\mu*\nu}^{-1}(u) \le F_\mu^{-1}(\sqrt{u}) + F_\nu^{-1}(\sqrt{u})$$
$$F_{\mu*\nu}^{+1}(u) \ge F_\mu^{+1}(1-\sqrt{1-u}) + F_\nu^{+1}(1-\sqrt{1-u}).$$

Proof Let X, Y be independent with $\mathcal{L}(X) = \mu, \mathcal{L}(Y) = \nu$, thus $\mu * \nu = \mathcal{L}(X+Y)$. Let $x = F_\mu^{-1}(\sqrt{u})$ and $y = F_\nu^{-1}(\sqrt{u})$, thus $F_\mu(x) \ge \sqrt{u}$ and $F_\nu(y) \ge \sqrt{u}$. It follows

$$F_{\mu*\nu}(x+y) = \mathbb{P}\left[X+Y \le x+y\right] \ge \mathbb{P}[X \le x, Y \le y] = F_\mu(x)F_\nu(y) \ge \sqrt{u}\sqrt{u} = u.$$

For all $z \in \mathbb{R}$ the equivalence $F_{\mu*\nu}(z) \ge u \Leftrightarrow z \ge F_{\mu*\nu}^{-1}(u)$ holds. Applying this to $z = x+y$ gives the first claimed bound. For the second claim, note $F_{\mathcal{L}(-Z)}^{-1}(u) = -F_{\mathcal{L}(Z)}^{+1}(1-u)$ holds

for any RV Z. Applying the first bound to $(\mu', \nu') = (\mathcal{L}(-X), \mathcal{L}(-Y))$ leads, for every $v \in (0,1)$, to

$$-F^{+1}_{\mu * \nu}(1-v) = F^{-1}_{\mathcal{L}((-X)+(-Y))}(v) \leq -(F^{+1}_{\mu}(1-\sqrt{v}) + F^{+1}_{\nu}(1-\sqrt{v})).$$

Multiplying by (-1) and substituting $v = 1-u$ yields the claim. ■

Proof [Proof of Lemma 17] Let $\eta = [\mathcal{L}(G_{\bar{s}}) : \bar{s} \in \mathcal{S}]$ with $(G_{\bar{s}})_{\bar{s}}$ independent from $(R_{sa\bar{s}})_{sa\bar{s}}$. Then $\mathcal{T}_s\eta(-\infty, x_{\min}) = \sum_{a,\bar{s}} \pi(a|s)p(\bar{s}|s,a)\mathbb{P}[R_{sa\bar{s}} + \gamma G_{\bar{s}} < x_{\min}]$. For every relevant $(a,\bar{s})$, using the definition of $x_{\min}$, the fact that $F^{-1}_{\mu} \leq F^{+1}_{\mu}$ and that $F^{+1}_{\mathcal{L}(\gamma X)} = \gamma F^{+1}_{\mathcal{L}(X)}$ in combination with the previous Lemma:

$$\begin{aligned}
\mathbb{P}[R_{sa\bar{s}} + \gamma G_{\bar{s}} < x_{\min}] &\leq \mathbb{P}[R_{sa\bar{s}} + \gamma G_{\bar{s}} < F^{-1}_{sa\bar{s}}\big(1-\sqrt{1-\varepsilon_u}\big) + \gamma \cdot F^{-1}_{\eta_{\bar{s}}}\big(1-\sqrt{1-\varepsilon_u}\big)] \\
&\leq \mathbb{P}[R_{sa\bar{s}} + \gamma G_{\bar{s}} < F^{+1}_{sa\bar{s}}\big(1-\sqrt{1-\varepsilon_u}\big) + \gamma \cdot F^{+1}_{\eta_{\bar{s}}}\big(1-\sqrt{1-\varepsilon_u}\big)] \\
&\leq \mathbb{P}[R_{sa\bar{s}} + \gamma G_{\bar{s}} < F^{+1}_{\mathcal{L}(R_{sa\bar{s}}+\gamma G_{\bar{s}})}(\varepsilon_u)] \leq \varepsilon_u.
\end{aligned}$$

To show $\mathcal{T}_s\eta(x_{\max}, \infty) \leq \varepsilon_u$ proceed similar and note that for each relevant pair $(a,\bar{s})$ it is

$$\begin{aligned}
\mathbb{P}[R_{sa\bar{s}} + \gamma G_{\bar{s}} > x_{\max}] &= 1 - \mathbb{P}[R_{sa\bar{s}} + \gamma G_{\bar{s}} \leq x_{\max}] \\
&\leq 1 - \mathbb{P}[R_{sa\bar{s}} + \gamma G_{\bar{s}} \leq F^{-1}_{sa\bar{s}}\big(\sqrt{1-\varepsilon_u}\big) + \gamma \cdot F^{-1}_{\eta_{\bar{s}}}\big(\sqrt{1-\varepsilon_u}\big)] \\
&\leq 1 - \mathbb{P}[R_{sa\bar{s}} + \gamma G_{\bar{s}} \leq F^{-1}_{\mathcal{L}(R_{sa\bar{s}}+\gamma G_{\bar{s}})}(1-\varepsilon_u)] \leq 1 - (1-\varepsilon_u) = \varepsilon_u.
\end{aligned}$$

The bound $\mathcal{T}_s\eta[x_{\min}, x_{\max}] \geq 1 - 2\varepsilon_u$ follows immediately. ■

A.5 Choice of the hyper-parameter functions in (18)

The reasoning behind recommendation (18) is now explained: Let $\eta^{(n)}$ be the n-th output when using arbitrary parameter functions $M, \mathcal{E}_u$. By construction, $\eta^{(n)}_s$ is finitely supported with particles $(x_i, p_i), i = 1, \ldots, M(n)$, where the x_i's are evenly spaced with distance $|x_{i+1} - x_i|$ being equal to

$$\delta(n,s) := \delta\left(A(\eta^{(n-1)}, s, n)\right) = \frac{x_{\max}(n,s) - x_{\min}(n,s)}{M(n)-1},$$

where $x_{\min}(n,s), x_{\max}(n,s)$ are given by (16) and (17) with $\eta = \eta^{(n-1)}$ and $\epsilon_u = \mathcal{E}_u(n)$. In order to obtain high quality approximations, M and $\mathcal{E}_u$ should be chosen such that both $\delta(n,s) \to 0$ and $\mathcal{E}_u(n) \to 0$ as $n \to \infty$. The asymptotic behaviour of $\delta(n,s)$ can be bounded as follows:

Theorem 32 *Assume (A1.Pα) with $\alpha > 0$ and set $\mathcal{E}_u(0) := 1$. Then for every $s \in \mathcal{S}$ we have $\delta(n) := \max_{s \in \mathcal{S}} \delta(n,s) = \mathrm{O}\left(\frac{1}{M(n)} \sum_{k=0}^{n} \gamma^{n-k} \mathcal{E}_u(k)^{-1/\alpha}\right)$ as $n \to \infty$.*

Proof Let

$$\begin{aligned}
x_{\min}(n,s) &= \min_{(a,\bar{s})} \left[F^{-1}_{sa\bar{s}}\left(1 - \sqrt{1-\mathcal{E}_u(n)}\right) + \gamma F^{-1}_{\eta^{(n-1)}_{\bar{s}}}\left(1 - \sqrt{1-\mathcal{E}_u(n)}\right)\right] \\
x_{\max}(n,s) &= \max_{(a,\bar{s})} \left[F^{-1}_{sa\bar{s}}\left(\sqrt{1-\mathcal{E}_u(n)}\right) + \gamma F^{-1}_{\eta^{(n-1)}_{\bar{s}}}\left(\sqrt{1-\mathcal{E}_u(n)}\right)\right].
\end{aligned}$$

With $H(u) = \max_{sa\bar{s}} \left[F^{-1}_{sa\bar{s}}\left(\sqrt{1-u}\right) - F^{-1}_{sa\bar{s}}\left(1-\sqrt{1-u}\right)\right]$ it is

$$\begin{aligned} x_{\max}(n,s) - x_{\min}(n,s) & \\ &\leq H(\mathcal{E}_n(u)) + \gamma \max_{\bar{s}} \left[F^{-1}_{\eta^{(n-1)}_{\bar{s}}}\left(\sqrt{1-\mathcal{E}_u(n)}\right) - F^{-1}_{\eta^{(n-1)}_{\bar{s}}}\left(1-\sqrt{1-\mathcal{E}_u(n)}\right)\right] \\ &\leq H(\mathcal{E}_n(u)) + \gamma \max_{\bar{s}} \left[x_{\max}(n-1,\bar{s}) - x_{\min}(n-1,\bar{s})\right], \end{aligned}$$

where the second inequality holds since $\eta^{(n-1)}_{\bar{s}}$ is supported on $[x_{\min}(n-1,\bar{s}), x_{\max}(n-1,\bar{s})]$. Hence with $W(n) := \max_{s\in\mathcal{S}} \left[x_{\max}(n,s) - x_{\min}(n,s)\right]$ it is $W(n) \leq H(\mathcal{E}_n(u)) + \gamma W(n-1)$. Assuming each $\eta^{(0)}_s$ it supported on $[a,b]$ and iterating this inequality inductively down to $n=0$ yields $W(n) \leq \sum_{k=1}^n \gamma^{n-k} H(\mathcal{E}_u(k)) + \gamma^n (b-a)$ and hence

$$\delta(n,s) \leq \max_{\bar{s}\in\mathcal{S}} \delta(n,\bar{s}) \leq \frac{W(n)}{M(n)-1} \leq \frac{1}{M(n)-1} \sum_{k=1}^n \gamma^{n-k} H\left(\mathcal{E}_u(k)\right) + \gamma^n \frac{b-a}{M(n)-1}.$$

Suppose (A1.Pα). By Lemma 33 there exists $C>0$ with $F^{-1}_{sa\bar{s}}(v) - F^{-1}_{sa\bar{s}}(1-v) \leq \frac{C}{(1-v)^{1/\alpha}}$ for all $v\in(0,1)$ and $s,a,\bar{s}$. Using $1/(1-\sqrt{1-u}) \leq 2/u$ for all $u\in(0,1)$, it follows $H(u) \leq C\left(\frac{2}{u}\right)^{1/\alpha}$. Thus, setting $\mathcal{E}_u(0) := 1$, the previous display results in

$$\delta(n,s) \leq \frac{1}{M(n)-1} \sum_{k=1}^n \gamma^{n-k} C \left(\frac{2}{\mathcal{E}_u(k)}\right)^{1/\alpha} + \gamma^n \frac{b-a}{M(n)-1},$$

which is $\mathrm{O}\left(\frac{1}{M(n)} \sum_{k=0}^n \gamma^{n-k} \mathcal{E}_u(k)^{-1/\alpha}\right)$ as claimed. ∎

Now suppose (A1.Pα) holds, let $\beta < \alpha$ and set $h = \frac{1\vee\beta}{\beta} \frac{\alpha}{\alpha-\beta} > 1$. Theorem 32 shows: Choosing $M(n) = \mathrm{O}((1/\theta)^{hn})$ and $\mathcal{E}_u(n) = \mathrm{O}(\theta^{\beta h n})$ leads to $\delta(n) = \mathrm{O}\left(\theta^{\frac{1\vee\beta}{\beta} n}\right)$, similar to Theorem 15. For $\varepsilon > 0$ let $n(\varepsilon) = \min\{n\in\mathbb{N} | \max\{\mathcal{E}_u(n), \delta(n)\} \leq \varepsilon\} \in \mathbb{N}$. Using the obtained asymptotic formulas shows that $n(\varepsilon) \leq \max\{\frac{1}{\beta\cdot h}, \frac{\beta}{1\vee\beta}\} \frac{\log(1/\varepsilon)}{\log(1/\theta)} + \mathrm{O}(1)$ as $\varepsilon\to 0$. Thus, the time to calculate the $n(\varepsilon)$-th approximation is of order $\mathrm{O}\left((1/\varepsilon)^{2r(\beta,\alpha)}\right)$ asymptotically as $\varepsilon\to 0$, where $r(\beta,\alpha) = \max\{\frac{1}{\beta}, \frac{\alpha}{\alpha-\beta}\}$. The function $(0,\alpha) \ni \beta \mapsto r(\beta,\alpha)$ takes its minimum at $\beta = \frac{\alpha}{\alpha+1}$ with minimal value $r(\frac{\alpha}{\alpha+1},\alpha) = \frac{\alpha+1}{\alpha}$. Choosing $\beta = \frac{\alpha}{\alpha+1}$ leads to $h = \left(\frac{\alpha+1}{\alpha}\right)^2$ and $\beta h = \frac{1}{\beta} = \frac{\alpha+1}{\alpha}$. Thus, it is reasonable to choose $M(k) = \mathrm{O}\left((1/\theta)^{\left(\frac{\alpha+1}{\alpha}\right)^2 \cdot n}\right)$ and $\mathcal{E}_u(k) = \mathrm{O}\left(\theta^{\frac{\alpha+1}{\alpha}\cdot n}\right)$ with a time to calculate the $n(\varepsilon)$'th approximation of order $\mathrm{O}\left((1/\varepsilon)^{2\cdot\frac{\alpha+1}{\alpha}}\right)$ as $\varepsilon\to 0$. For practical purposes, it seems reasonable to suggest this method only in case $\frac{\alpha+1}{\alpha} > 1$ is small, where the lower bound 1 is attained in the limit $\alpha\to\infty$. In the limiting case, it is $M(k) = \mathrm{O}\left((1/\theta)^k\right)$ and $\mathcal{E}_u(k) = \mathrm{O}\left(1/M(k)\right)$, which leads to (18).

It remains to show:

Lemma 33 *Let $\mu = \mathcal{L}(X) \in \mathscr{P}_\alpha(\mathbb{R})$. Then for all $u\in(0,1)$ we have*

$$-\mathbb{E}[|X|^\alpha]^{1/\alpha} \cdot \frac{1}{u^{1/\alpha}} \;\leq\; F^{-1}_\mu(u) \;\leq\; \mathbb{E}[|X|^\alpha]^{1/\alpha} \cdot \frac{1}{(1-u)^{1/\alpha}}.$$

Proof Let $D := \mathbb{E}[|X|^\alpha] < \infty$. In case $D = 0$ it is $\mathbb{P}[X = 0] = 1$ and $F_\mu^{-1}(u) = 0$. Suppose $D > 0$. Markov's inequality gives $1 - F_\mu(x) = \mathbb{P}[X > x] \leq \mathbb{P}[|X| > x] \leq Dx^{-\alpha}$ for all $x > 0$, hence $F_\mu(x) \geq 1 - Dx^{-\alpha}$ for all $x > 0$. Applying this to $x = \left(\frac{1-u}{D}\right)^{-\frac{1}{\alpha}} > 0$ gives $F_\mu(x) \geq u$, hence $x \geq F_\mu^{-1}(u)$, which is the second bound. Applying the second bound to $\tilde{\mu} = \mathcal{L}(-X)$, $u = 1 - q \in (0, 1)$ and using $F^{-1}_{\mathcal{L}(-X)}(u) = -F^{+1}_{\mathcal{L}(X)}(1 - u)$ gives $F_\mu^{+1}(q) \geq -\mathbb{E}[|X|^\alpha]^{1/\alpha} \cdot \frac{1}{q^{1/\alpha}}$. The lower bound follows from this by $F_\mu^{-1}(q) \geq \limsup_{q' \uparrow q} F^{+1}(q') \geq \limsup_{q' \uparrow q} -\mathbb{E}[|X|^\alpha]^{1/\alpha} \cdot \frac{1}{(q')^{1/\alpha}} = -\mathbb{E}[|X|^\alpha]^{1/\alpha} \cdot \frac{1}{q^{1/\alpha}}$. ■

A.6 Proofs of Theorem 23 and Corollary 24

Proof [Theorem 23] The stochastic process $\boldsymbol{SA\bar{S}} = (S^{(t)}, A^{(t)}, S^{(t+1)})_{t \in \mathbb{N}_0}$ is a random variable taking values in the infinite product space $(\mathcal{S} \times \mathcal{A} \times \mathcal{S})^{\mathbb{N}_0}$, which has elements of the form $\boldsymbol{sa\bar{s}} = (s_t, a_t, \bar{s}_t)_{t \in \mathbb{N}_0}$. Let $\Upsilon := \mathbb{P}_{s_0}[\boldsymbol{SA\bar{S}} \in \cdot\,]$ be its distribution. Recall that $T = \min\{t \in \mathbb{N}_0 | (S^{(t)}, A^{(t)}, R^{(t)}) \in \mathcal{G}\} \in \mathbb{N}_0 \cup \{\infty\}$. Letting $\mathcal{H} \subseteq (\mathcal{S} \times \mathcal{A} \times \mathcal{S})^{\mathbb{N}_0}$ be the subset of sequences visiting $\mathcal{G}$ we have $\{T < \infty\} = \{\boldsymbol{SA\bar{S}} \in \mathcal{H}\}$ and hence $\Upsilon(\mathcal{H}) = 1$ by assumption. Recall that $G^* = \gamma^T R^{(T)} + Z$ with $Z = \sum_{t \in \mathbb{N}_0 \setminus \{T\}} \gamma^t R^{(t)}$ and $(S, A, \bar{S}) := (S^{(T)}, A^{(T)}, S^{(T+1)}) \in \mathcal{G}$. Let $\mathbb{P}_{\boldsymbol{sa\bar{s}}}[(T, S, A, \bar{S}, R^{(T)}, Z) \in \cdot], \boldsymbol{sa\bar{s}} \in (\mathcal{S} \times \mathcal{A} \times \mathcal{S})^{\mathbb{N}_0}$ be a regular conditional distribution of $(T, S, A, \bar{S}, R^{(T)}, Z)$ given $\boldsymbol{SA\bar{S}}$. In particular, for every measurable set $B \subseteq \mathbb{R}$ we have

$$\eta^*_{s_0}(B) = \mathbb{P}_{s_0}[G^* \in B] = \int_{\mathcal{H}} \mathbb{P}_{\boldsymbol{sa\bar{s}}}[\gamma^T R^{(T)} + Z \in B] \mathrm{d}\Upsilon(\boldsymbol{sa\bar{s}}). \tag{24}$$

For Υ-almost all sequences $\boldsymbol{sa\bar{s}} \in \mathcal{H}$ we have, under $\mathbb{P}_{\boldsymbol{sa\bar{s}}}$, that the random variables $T \in \mathbb{N}, (S, A, \bar{S}) \in \mathcal{G}$ are deterministic, $R^{(T)} \sim r(\cdot|S, A, \bar{S})$ has PDF $f_{(S,A,\bar{S})}$ and $R^{(T)}, Z$ are independent. Recall that, if $a > 0, b \in \mathbb{R}$ are constants and X has PDF f, then $aX + b$ has PDF $x \mapsto \frac{1}{a} f(\frac{x-b}{a})$. If further Y is a RV independent of X, then $aX + Y$ has PDF $x \mapsto \mathbb{E}\left[\frac{1}{a} f(\frac{x-Y}{a})\right]$, which follows from Fubini's theorem. Hence, for Υ-almost all sequences $\boldsymbol{sa\bar{s}} \in \mathcal{H}$ it holds that the distribution of $G^* = \gamma^T R^{(T)} + Z$ under $\mathbb{P}_{\boldsymbol{sa\bar{s}}}$ has density $x \mapsto \mathbb{E}_{\boldsymbol{sa\bar{s}}}[\frac{1}{\gamma^T} f_{(S,A,\bar{S})}\left(\frac{x-Z}{\gamma^T}\right)]$, where $T, (S, A, \bar{S})$ are non-random under $\mathbb{P}_{\boldsymbol{sa\bar{s}}}$ with values depending on $\boldsymbol{sa\bar{s}}$. Hence, for every measurable $B \subseteq \mathbb{R}$ it is

$$\begin{aligned}
\int_{\mathcal{H}} \mathbb{P}_{\boldsymbol{sa\bar{s}}}[\gamma^T R^{(T)} + Z \in B] \, \mathrm{d}\Upsilon(\boldsymbol{sa\bar{s}}) &= \int_{\mathcal{H}} \int_B \mathbb{E}_{\boldsymbol{sa\bar{s}}}\left[\frac{1}{\gamma^T} f_{(S,A,\bar{S})}\left(\frac{x-Z}{\gamma^T}\right)\right] \mathrm{d}x \, \mathrm{d}\Upsilon(\boldsymbol{sa\bar{s}}) \\
&= \int_B \int_{\mathcal{H}} \mathbb{E}_{\boldsymbol{sa\bar{s}}}\left[\frac{1}{\gamma^T} f_{(S,A,\bar{S})}\left(\frac{x-Z}{\gamma^T}\right)\right] \mathrm{d}\Upsilon(\boldsymbol{sa\bar{s}}) \, \mathrm{d}x \\
&= \int_B \mathbb{E}_{s_0}\left[\frac{1}{\gamma^T} f_{(S,A,\bar{S})}\left(\frac{x-Z}{\gamma^T}\right)\right] \mathrm{d}x,
\end{aligned}$$

where the second equation holds by Fubini's theorem and the third is the key property of regular conditional distributions. The claim follows in combination with (24). ■

Proof [Corollary 24] (a) Let $M > 0$ be such that $|f_{(s,a,\bar{s})}(x)| \leq M$ for all $(s, a, \bar{s}) \in \mathcal{G}, x \in \mathbb{R}$. Since $(S, A, \bar{S}) \in \mathcal{G}$, it is $|f_{(S,A,\bar{S})}(x)| \leq M$ for all $x \in \mathbb{R}$ and thus, for every $x \in \mathbb{R}$, it

follows $|f^*_{s_0}(x)| \le M \cdot \mathbb{E}_{s_0}[\gamma^{-T}] = M \cdot \Psi(\log(1/\gamma)) < \infty$. For (b), let $C > 0$ be such that $|f_{(s,a,\bar{s})}(x) - f_{(s,a,\bar{s})}(y)| \le C|x-y|^\tau$ for all $x, y \in \mathbb{R}$ and $(s,a,\bar{s}) \in \mathcal{G}$. This bound is also satisfied for the random $(S, A, \bar{S}) \in \mathcal{G}$ and thus, for all $x, y \in \mathbb{R}$,

$$\begin{aligned}
|f^*_{s_0}(x) - f^*_{s_0}(y)| &\le \mathbb{E}_{s_0}\Big[\frac{1}{\gamma^T}\Big|f_{(S,A,\bar{S})}\left(\frac{x-Z}{\gamma^T}\right) - f_{(S,A,\bar{S})}\left(\frac{y-Z}{\gamma^T}\right)\Big|\Big] \\
&\le \mathbb{E}_{s_0}\Big[\frac{1}{\gamma^T} \cdot C \cdot \Big|\frac{x-Z}{\gamma^T} - \frac{y-Z}{\gamma^T}\Big|^\tau\Big] \\
&= C \cdot \mathbb{E}_{s_0}\big[\gamma^{-(\tau+1)T}\big] \cdot |x-y|^\tau = C \cdot \Psi\big((\tau+1)\log(\gamma^{-1})\big) \cdot \ |x-y|^\tau,
\end{aligned}$$

with $C \cdot \Psi\big((\tau+1)\log(\gamma^{-1})\big) < \infty$ by assumption. ■

References

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 449–458. JMLR.org, 2017.

Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. MIT Press, 2023. `http://www.distributional-rl.org`.

Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018. ISBN 978-1-57735-800-8.

Benjamin Dadoun and Ralph Neininger. A statistical view on exchanges in Quickselect. In *ANALCO14—Meeting on Analytic Algorithmics and Combinatorics*, pages 40–51. SIAM, Philadelphia, PA, 2014. doi: 10.1137/1.9781611973204.4.

Luc Devroye. *Non-Uniform Random Variate Generation*. Springer New York, NY, 1986. doi: 10.1007/978-1-4613-8643-8.

Luc Devroye and Lancelot F. James. The double cftp method. *ACM Trans. Model. Comput. Simul.*, 21(2), feb 2011. ISSN 1049-3301. doi: 10.1145/1899396.1899398.

Luc Devroye and Ralph Neininger. Density approximation and exact simulation of random variables that are solutions of fixed-point equations. *Adv. in Appl. Probab.*, 34(2):441–468, 2002. ISSN 0001-8678. doi: 10.1239/aap/1025131226.

Luc Devroye, James Allen Fill, and Ralph Neininger. Perfect simulation from the Quicksort limit distribution. *Electron. Comm. Probab.*, 5:95–99, 2000. ISSN 1083-589X. doi: 10.1214/ECP.v5-1024.

Persi Diaconis and David Freedman. Iterated random functions. *SIAM review*, 41(1):45–76, 1999.

Qiang Du, Vance Faber, and Max Gunzburger. Centroidal voronoi tessellations: Applications and algorithms. *SIAM review*, 41(4):637–676, 1999.

James Allen Fill and Svante Janson. Quicksort asymptotics. *J. Algorithms*, 44(1):4–28, 2002. ISSN 0196-6774. doi: 10.1016/S0196-6774(02)00216-X. Analysis of algorithms.

Julian Gerstenberg, Ralph Neininger, and Denis Spiegel. On solutions of the distributional Bellman equation. *Electron. Res. Arch.*, 31(8):4459–4483, 2023. doi: 10.3934/era.2023228.

Ben Hambly, Renyuan Xu, and Huining Yang. Recent advances in reinforcement learning in finance. *Mathematical Finance*, 33(3):437–503, 2023. doi: https://doi.org/10.1111/mafi.12382.

Wolfgang Hörmann, Josef Leydold, and Gerhard Derflinger. *Automatic nonuniform random variate generation*. Springer Berlin, Heidelberg, 2004. doi: 10.1007/978-3-662-05946-3.

Benoit Kloeckner. Approximation by finitely supported measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 18(2):343–359, 2012.

Margarete Knape and Ralph Neininger. Approximating perpetuities. *Methodol. Comput. Appl. Probab.*, 10(4):507–529, 2008. ISSN 1387-5841. doi: 10.1007/s11009-007-9059-x.

Petter N Kolm and Gordon Ritter. Modern perspectives on reinforcement learning in finance. *Modern Perspectives on Reinforcement Learning in Finance (September 6, 2019). The Journal of Machine Learning in Finance*, 1(1), 2020.

Elena Krasheninnikova, Javier García, Roberto Maestre, and Fernando Fernández. Reinforcement learning for pricing strategy optimization in the insurance industry. *Engineering applications of artificial intelligence*, 80:8–19, 2019.

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

Chris Maddison, Dieterich Lawson, George Tucker, Nicolas Heess, Arnaud Doucet, Andriy Mnih, and Yee Teh. Particle value functions. In *Proceedings of the International Conference on Learning Representations (Workshop Track)*, 03 2017.

Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 799–806, Madison, WI, USA, 2010a. Omnipress. ISBN 9781605589077.

Tetsuro Morimura, Masashi Sugiyama, Hisashi Kashima, Hirotaka Hachiya, and Toshiyuki Tanaka. Nonparametric return distribution approximation for reinforcement learning. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 799–806, Madison, WI, USA, 2010b. Omnipress. ISBN 9781605589077.

Svetlozar T. Rachev. *Probability metrics and the stability of stochastic models*. Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics. John Wiley & Sons, Ltd., Chichester, 1991. ISBN 0-471-92877-1.

Mark Rowland, Marc Bellemare, Will Dabney, Remi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 29–37. PMLR, 09–11 Apr 2018.

A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998. ISBN 0-521-49603-9; 0-521-78450-6. doi: 10.1017/CBO9780511802256.