

Distributional Bellman Operators with Power-Law Horizon Weights: Contraction and Fixed-Point Analysis

Yashar Mousavi

Glasgow Caledonian University

Rashin Mousavi

Paradise Research Center

Arash Mousavi

Paradise Research Center

Mahsa Tavasoli

North Carolina Agricultural and Technical State University

Ibrahim Beklan Kucukdemiral

Glasgow Caledonian University

Afef Fekih

University of Louisiana at Lafayette

Umit Cali

`umit.cali@ntnu.no`

Norwegian University of Science and Technology

Article

Keywords: distributional reinforcement learning, Bellman operator, Wasserstein contraction, fractional calculus, power-law weighting, stochastic dominance, credit assignment

Posted Date: October 14th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11129917/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Distributional Bellman Operators with Power-Law Horizon Weights: Contraction and Fixed-Point Analysis

Yashar Mousavi¹, Rashin Mousavi², Arash Mousavi^{2,3}, Mahsa Tavasoli⁴,
Ibrahim Beklan Kucukdemiral¹, Afef Fekih⁵, Umit Cali^{6,7,8*}

¹Department of Engineering, School of Science and Engineering, Glasgow Caledonian University, Glasgow, G4 0BA, UK.

²Department of Electrical Engineering and Applied Sciences, Paradise Research Center, Jahrom, 7416813647, Iran.

³Department of Electrical Engineering, Islamic Azad University, Marvdasht, 73711-13119, Iran.

⁴Department of Applied Science & Technology, North Carolina Agricultural and Technical State University, Greensboro, NC, 27411, USA.

⁵Electrical and Computer Engineering Department, University of Louisiana at Lafayette, P.O. Box 43890, Lafayette, LA, 70504, USA.

⁶School of Physics, Engineering and Technology, University of York, York, YO10 5DD, UK.

^{7*}Department of Electric Energy, Norwegian University of Science and Technology, Trondheim, Norway.

⁸Artificial Intelligence, Artificial Intelligence, Robotics and Connected Infrastructure Research Center, American University of Sharjah, Sharjah, UAE.

*Corresponding author(s). E-mail(s): umit.cali@ntnu.no;

Abstract

Distributional reinforcement learning models the full distribution of returns and relies on the contraction property of the distributional Bellman operator for stable policy evaluation. However, it remains unclear whether this property is preserved when the standard one-step Bellman backup is replaced by a multi-horizon backup with power-law memory. This paper addresses this question by introducing and analyzing a general multi-horizon distributional Bellman operator defined by an arbitrary normalized distribution of bootstrap horizons. We prove that the proposed operator is a contraction in the supremal p -Wasserstein metric, with the exact contraction modulus $\kappa_p(w) = (\sum_{n \geq 1} w(n) \gamma^{np})^{1/p}$ holds only for the standard one-step backup. We further prove that the operator preserves the true return distribution as its unique fixed point. In addition, we show that shifting more weight toward longer horizons improves the contraction rate under first-order stochastic dominance, with a strict result for power-law horizon weights. Closed-form contraction moduli are derived for both power-law and geometric weight families. The analysis is also extended to state-dependent horizon weights, and an explicit example is constructed to show that the derived contraction modulus is the exact worst-case Lipschitz constant. The resulting modulus is further used to characterize the iteration complexity, and the theoretical results are illustrated on a tabular Markov decision process. These results establish that power-law multi-horizon Bellman backups can preserve both contraction and the correct return-distribution fixed point, providing a theoretical basis for incorporating long-memory mechanisms directly into distributional critics.

Keywords: distributional reinforcement learning; Bellman operator; Wasserstein contraction; fractional calculus; power-law weighting; stochastic dominance; credit assignment

1 Introduction

Temporal credit assignment, understood as the attribution of an observed outcome to the past actions responsible for it, remains a central obstacle in reinforcement learning (RL) whenever rewards are displaced by hundreds of steps from the decisions that caused them [1]. The dominant credit-assignment mechanism is the exponentially decaying eligibility trace of temporal-difference (TD) learning with parameter λ [2], which distributes credit geometrically. Its influence becomes numerically negligible within tens of steps for standard hyperparameter settings. This geometric forgetting has motivated a family of *power-law* credit-assignment methods built on fractional calculus, in which Mittag–Leffler memory kernels replace exponential decay so that credit is sustained over substantially longer horizons [3, 4].

A representative instance of this family is the FracTrace framework [5], which incorporates variable-order Mittag–Leffler kernels at two architectural levels. Hierarchical traces maintain multi-timescale *state* representations, while fractional eligibility traces shape *policy-side* advantages. A deliberate design principle of that framework is that the critic retains the standard one-step distributional Bellman operator, so that fractionalization is applied exclusively to the policy side. The stated justification is a contraction argument, because the standard distributional Bellman evaluation operator is a γ -contraction in the supremal p -Wasserstein metric [6], and this property underwrites the convergence guarantees of the critic. Nevertheless, the question of whether the *Bellman backup itself* could be fractionalized is explicitly left open there. The obstruction is technical and precise. The coupling argument that establishes γ -contraction of the standard operator compares the outputs of the two operators through a single immediate next-state distribution, and such an argument does not transfer directly to an operator that mixes returns over many different horizons under power-law weights [5]. The restriction to the policy side was consequently adopted as a conservative measure, so as to remain within a regime whose contraction properties were already understood.

A distinct line of work in distributional RL has meanwhile shown that multi-step operators of a related kind do contract. Tang et al. [7] establish that the distributional Retrace operator, an off-policy backup that mixes n -step distributional targets under trace coefficients derived from clipped importance ratios, is a γ -contraction in the supremal p -Wasserstein metric with the true return distribution as its unique fixed point. That result confirms, in a more general off-policy setting, that mixing distributional backups across horizons need not destroy contraction, and it is the closest structural precedent to the object studied here. It does not, however, settle the question that motivates this paper, for three reasons. First, the horizon weighting in distributional Retrace is generated endogenously by the interaction of the trace parameter with the importance ratios between the target and behavior policies, and not specified as an arbitrary exogenous distribution over horizons. Second, the contraction modulus is characterized only through the effective horizon at the two limiting cases of the trace parameter, not through a closed-form expression for a general weighting. Third, no comparison across weight families ordered by tail heaviness, and no state-dependent extension matching the per-state kernel modulation of FracTrace, is available in that setting. Whether an *arbitrary* normalized horizon-weight distribution, and in particular the power-law family that governs the kernels of FracTrace, preserves contraction with an explicit and comparable modulus is accordingly still open.

A closely related line of work in the scalar setting studies weighted horizon mixtures under a different name. Daley et al. [8] analyze *compound returns*, which are weighted averages of n -step returns, and characterize their bias and variance through the same contraction modulus that governs the underlying n -step operators, establishing that any compound return sharing the contraction modulus of a given n -step return has strictly lower variance. That analysis operates in the scalar expected-return setting and is concerned with variance and sample complexity. The object studied here is instead a mixture of *return distributions*, and the contribution is the general p -Wasserstein contraction factor, its sharp worst-case characterization, the stochastic-dominance comparative law, the power-law and polylogarithm specialization, and the state-dependent extension. The two analyses accordingly address the same weighted-mixture construction from different and complementary angles, one governing variance in the scalar case and the other governing contraction in the distributional case.

The base case for the present analysis is the classical result of Bellemare et al. [6], which shows that the one-step distributional Bellman evaluation operator is a γ -contraction in the supremal p -Wasserstein metric through a coupling of the shared next state and reward. The unconditional contraction established below is a natural extension of that same coupling technique, presented as such, and not as a new proof technique. Two further papers address different but adjacent questions. Amortila et al. [9] study the distribution of the stochastic-approximation *iterate* produced by algorithms such as TD(λ) and Q -learning under a constant step size. They show that this iterate distribution is itself contractive toward a stationary distribution whose mean recovers the true value function. That notion of “distributional” concerns the sampling variability of a scalar-valued estimator, and not the distribution of the return itself, so the two analyses are not in direct competition. Nam et al. [10] contribute a multi-step target construction, whose Sample-Replacement algorithm builds a distributional analogue of the λ -return for actor-critic methods. That construction is algorithmic and empirical, and it neither furnishes a general closed-form contraction modulus nor supports a comparison across weight families ordered by tail heaviness. The use of first-order stochastic dominance as the instrument for that comparison, characterized formally following the standard treatment of Shaked and Shanthikumar [11], appears to be new in this context.

This paper addresses this gap by studying the multi-horizon mixing operator $\mathcal{T}_w$. For a normalized weight distribution w over bootstrap horizons $n \geq 1$, $\mathcal{T}_w$ combines n -step distributional Bellman backups. Geometric weights recover the compound update used in distributional TD(λ)-type multi-step targets [7, 10], while power-law weights provide the long-memory formulation considered in this work. Our analysis goes beyond establishing contraction. It provides an exact contraction rate and shows how this rate depends on the choice of horizon weights. The main contributions of this paper are as follows:

- (i) *Unconditional contraction and fixed-point identification.* For any normalized horizon-weight distribution w and any $p \geq 1$, the operator $\mathcal{T}_w$ is a γ -contraction in the supremal p -Wasserstein metric, with explicit modulus $\kappa_p(w) = (\sum_{n \geq 1} w(n) \gamma^{np})^{1/p} \leq \gamma$. Equality holds if and only if w is the point mass at $n = 1$, which is the one-step operator, so that power-law Bellman mixing preserves contraction and never weakens the guarantee already enjoyed by the standard one-step backup. The unique fixed point is identified explicitly as the true return-distribution function Z^π , so that the contraction is to the correct object and not merely to some stationary point of the mixture.
- (ii) *Comparative contraction law.* Horizon distributions that place more weight on longer horizons, in the sense of first-order stochastic dominance, contract no more weakly, and strictly so within the power-law family itself. The proof proceeds through this dominance ordering together with a monotone integrand, and it cautions explicitly against the natural but incorrect intuition based on Jensen’s inequality and mean-preserving spreads, which points in the wrong direction.
- (iii) *Closed-form contraction moduli.* For the power-law family $w_\alpha(n) \propto n^{-(1+\alpha)}$ the identity $\kappa_p(w_\alpha)^p = \text{Li}_{1+\alpha}(\gamma^p)/\zeta(1+\alpha)$ is obtained through the polylogarithm, together with a rational closed form for the geometric family. These expressions place the learned slow-level order range of FracTrace on an explicit and directly evaluable map.
- (iv) *State-dependent extension.* For state-indexed weight families $\{w_s\}_{s \in \mathcal{S}}$, which capture a state-indexed analogue motivated by the state-modulated kernels of FracTrace, a uniform worst-case modulus $\kappa_p^{\text{sup}} = (\sup_s \sum_n w_s(n) \gamma^{np})^{1/p} \leq \gamma$ is established together with a uniform comparative law, and no renormalization-inflation constant is incurred.
- (v) *Sharpness.* The bound is attained, since for every w and every p there exists an explicit Markov decision process (MDP) and a pair of return-distribution functions that realize $\kappa_p(w)$ exactly. Consequently $\kappa_p(w)$ is the exact worst-case Lipschitz constant of $\mathcal{T}_w$ over the admissible MDP class, and the comparative ordering established above is a property of that worst case and not an artifact of the bound.
- (vi) *Iteration complexity and verification.* The modulus is converted into an explicit iteration count $k(\epsilon) = \lceil \ln(\epsilon/D) / \ln \kappa_p(w) \rceil$, and both the sharpness claim and the contraction bound are verified numerically on tabular Markov decision processes.

The remainder of the paper is organized as follows. Section 2 fixes notation, recalls the one-step operator together with its contraction property, defines the multi-horizon mixing operator, and distinguishes its summable weight family from the non-summable advantage kernel of FracTrace. Section 3 presents the main results and their proofs. Section 4 reports the numerical evaluation of the closed-form moduli, the resulting iteration complexity, and the tabular verification. Section 5 discusses the findings and concludes the paper.

2 Preliminaries and Problem Formulation

2.1 Markov Decision Processes and Return Distributions

Consider a Markov decision process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ with state space $\mathcal{S}$, action space $\mathcal{A}$, transition kernel $P : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{P}(\mathcal{S})$, reward function r satisfying the bounded-reward condition $|r(s, a)| \leq R_{\max}$ for all (s, a) , and discount factor $\gamma \in (0, 1)$. The reward function is taken to be deterministic. The coupling arguments used throughout, beginning with the proof of Proposition 3, couple identical reward prefixes and leave only the bootstrap distributions to differ, so the results extend to a common bounded stochastic reward kernel under synchronous coupling of the shared trajectory and reward realization; that extension is not carried out here. The excluded boundary case $\gamma = 0$ reduces every operator considered in this paper to a single deterministic evaluation of r , which is outside the scope of the contraction analysis that follows; its consequences for the results below are recorded once the relevant quantities have been introduced. A stationary policy $\pi : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A})$ induces the transition kernel $P^\pi(s' | s) = \sum_a \pi(a | s) P(s' | s, a)$. For clarity of exposition the policy π is treated as deterministic throughout, with $\pi(s)$ denoting the selected action. The stochastic case follows by averaging over $a \sim \pi(\cdot | s)$ at every occurrence and alters none of the arguments, because the action-averaged kernel P^π is itself a well-defined Markov kernel on $\mathcal{S}$, so that the identical-trajectory-prefix coupling on which the proofs below rely is unaffected by whether π is deterministic or stochastic. Throughout, $\mathcal{S}$ and $\mathcal{A}$ are taken to be finite or countable. This choice is fully compatible with the tabular numerical setting of Section 4.4, and it renders vacuous every measurability question that would otherwise arise from the state-conditional optimal Wasserstein couplings used in the proofs below. The general case of uncountable, standard Borel state and action spaces is not treated here; it requires explicit measurability assumptions on P , π , r , and Z , together with a measurable-selection argument for the state-conditional couplings, and is left for future work.

Distributional RL models the *return-distribution function* $Z : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{P}_p(\mathbb{R})$ in place of the scalar value function alone, where $Z(s, a)$ denotes the law of the random return $\sum_{t \geq 0} \gamma^t r(S_t, A_t)$ obtained by starting from (s, a) and following π . Here $\mathcal{P}_p(\mathbb{R})$ denotes the probability measures on $\mathbb{R}$ with finite p -th moment [6]. Under the bounded-reward condition every such return lies almost surely in $[-V_{\max}, V_{\max}]$, where $V_{\max} := R_{\max}/(1 - \gamma)$. The natural domain of the analysis is therefore the space

$$\mathcal{Z} := \left\{ Z : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{P}_p(\mathbb{R}) \mid \text{supp } Z(s, a) \subseteq [-V_{\max}, V_{\max}] \ \forall (s, a) \right\} \quad (1)$$

of return-distribution functions with uniformly bounded support.

2.2 Wasserstein Metric and the Supremal Metric

For $p \geq 1$ and $\mu, \nu \in \mathcal{P}_p(\mathbb{R})$, the p -Wasserstein distance is

$$W_p(\mu, \nu) = \left(\inf_{\xi \in \Pi(\mu, \nu)} \mathbb{E}_{(X, Y) \sim \xi} |X - Y|^p \right)^{1/p}, \quad (2)$$

where $\Pi(\mu, \nu)$ is the set of couplings of μ and ν [12]. The space $\mathcal{Z}$ is equipped with the *supremal* p -Wasserstein metric

$$\bar{d}_p(Z_1, Z_2) = \sup_{(s, a) \in \mathcal{S} \times \mathcal{A}} W_p(Z_1(s, a), Z_2(s, a)). \quad (3)$$

The pair $(\mathcal{Z}, \bar{d}_p)$ is a complete metric space [6], so that the Banach fixed-point theorem applies to any contraction mapping $\mathcal{Z} \rightarrow \mathcal{Z}$. The following property of W_p^p drives the entire analysis and is recorded first.

Lemma 1 (Convexity of W_p^p under mixtures) *Let $\{\mu_i\}_{i \in \mathcal{I}}$ and $\{\nu_i\}_{i \in \mathcal{I}}$ be families in $\mathcal{P}_p(\mathbb{R})$ indexed by a countable set $\mathcal{I}$, and let $\{\omega_i\}_{i \in \mathcal{I}}$ be nonnegative weights with $\sum_i \omega_i = 1$. Then*

$$W_p\left(\sum_i \omega_i \mu_i, \sum_i \omega_i \nu_i\right)^p \leq \sum_i \omega_i W_p(\mu_i, \nu_i)^p. \quad (4)$$

Proof For each i , let $\xi_i \in \Pi(\mu_i, \nu_i)$ be an optimal coupling attaining $W_p(\mu_i, \nu_i)$. The mixture $\xi := \sum_i \omega_i \xi_i$ is a coupling of $\sum_i \omega_i \mu_i$ and $\sum_i \omega_i \nu_i$, since its marginals are the corresponding mixtures. Hence, by the definition of W_p as an infimum over couplings,

$$W_p\left(\sum_i \omega_i \mu_i, \sum_i \omega_i \nu_i\right)^p \leq \mathbb{E}_{(X,Y) \sim \xi} |X - Y|^p = \sum_i \omega_i \mathbb{E}_{\xi_i} |X - Y|^p = \sum_i \omega_i W_p(\mu_i, \nu_i)^p. \quad \square \quad (5)$$

2.3 The One-Step Operator and Its Contraction

For a fixed policy π , the standard *one-step* distributional Bellman evaluation operator is

$$(\mathcal{T}^\pi Z)(s, a) \stackrel{d}{=} r(s, a) + \gamma Z(s', \pi(s')), \quad s' \sim P(\cdot | s, a), \quad (6)$$

where $\stackrel{d}{=}$ denotes equality in distribution and the right-hand side is understood as the law of the displayed random variable. The following classical statement is the base case that the present analysis generalizes.

Proposition 2 (Bellemare et al. [6]) *Under the bounded-reward condition, $\mathcal{T}^\pi$ maps $\mathcal{Z}$ into itself and is a γ -contraction in $\bar{d}_p$, so that $\bar{d}_p(\mathcal{T}^\pi Z_1, \mathcal{T}^\pi Z_2) \leq \gamma \bar{d}_p(Z_1, Z_2)$ for all $Z_1, Z_2 \in \mathcal{Z}$. Consequently $\mathcal{T}^\pi$ has a unique fixed point $Z^\pi \in \mathcal{Z}$, which is the return-distribution function of π .*

It is precisely the locality of the coupling underlying Proposition 2, which involves a *single* next-state distribution, that fails to extend transparently to an operator mixing many horizons. That failure constitutes the obstruction removed in this paper. The remainder of this section defines the multi-horizon operator that removes it and fixes the weight family to which the closed-form results developed later in the paper specialize.

2.4 Definition

Let $w = \{w(n)\}_{n \geq 1}$ denote a *horizon-weight distribution*, meaning a sequence with $w(n) \geq 0$ and $\sum_{n \geq 1} w(n) = 1$. Along a trajectory generated by P and π from (s, a) , with $S_t = s$ and $A_t = a$, only the first transition uses the supplied action, so that $S_{t+1} \sim P(\cdot | s, a)$, while every subsequent transition follows π , so that $S_{t+k+1} \sim P(\cdot | S_{t+k}, \pi(S_{t+k}))$ for $k \geq 1$. The n -step bootstrapped return random variable is defined by

$$G^{(n)}(s, a; Z) = r(s, a) + \sum_{k=1}^{n-1} \gamma^k R_{t+k} + \gamma^n \zeta_n, \quad \zeta_n \sim Z(S_{t+n}, \pi(S_{t+n})), \quad (7)$$

where $R_{t+k} = r(S_{t+k}, \pi(S_{t+k}))$ for $k \geq 1$ and ζ_n is a bootstrap sample drawn from the return distribution at the realized state S_{t+n} . This convention ensures that $n = 1$ reduces exactly to the one-step operator of (6), since $G^{(1)}(s, a; Z) = r(s, a) + \gamma \zeta_1$ with $S_{t+1} \sim P(\cdot | s, a)$. Using $\pi(s)$ in place of the supplied action a at the first step, as an unweighted average over $k = 0, \dots, n-1$ would otherwise do, would instead make $\mathcal{T}^{(n)}|_{n=1}$ disagree with $\mathcal{T}^\pi$ whenever $a \neq \pi(s)$. The n -step distributional operator is then $(\mathcal{T}^{(n)} Z)(s, a) := \text{Law}[G^{(n)}(s, a; Z)]$.

Definition 1 (Multi-horizon mixing operator) The *multi-horizon mixing operator* associated with w is the w -mixture of the n -step operators,

$$(\mathcal{T}_w Z)(s, a) := \sum_{n=1}^{\infty} w(n) (\mathcal{T}^{(n)} Z)(s, a), \quad (8)$$

which is the law obtained by first sampling a horizon $n \sim w$ and then sampling $G^{(n)}(s, a; Z)$.

Definition 1 is the distributional analogue of the compound update based on the λ -return, and it is worth confirming this connection explicitly now that the operator has been given a precise form. The geometric weighting $w_\lambda(n) = (1 - \lambda)\lambda^{n-1}$ is the specific instance that recovers the compound update introduced in Section 1 in its distributional TD(λ)-style form [7, 10]. The *mixture* interpretation deserves emphasis, since it prescribes sampling a horizon and then a return, and it is therefore distinct from a pathwise weighted sum of returns. The mixture is the construction for which contraction holds cleanly, and it is moreover the construction implicit in the compound backup.

2.5 Which Weight Family Is Appropriate

A point of potential confusion must be settled before the analysis proceeds, because it marks the exact place at which the object studied here diverges from the object treated by Mousavi et al. [5]. The *policy-side* advantage kernel of FracTrace is the normalized Mittag-Leffler kernel with asymptotic decay $K_\alpha(k) \sim k^{-\alpha}/\Gamma(1 - \alpha)$ for $\alpha \in (0, 1)$, and this kernel is *not* absolutely summable over k . Its non-summability is precisely the reason that a truncation-horizon framework and a renormalization-inflation constant were required there in order to bound the finite-horizon advantage error. That kernel weights the *lags of a fixed backup* instead of the *horizons of a Bellman mixture*, and it does not constitute a probability distribution over k .

The multi-horizon Bellman operator of Definition 1 requires a genuine probability distribution over horizons n , meaning a *summable* sequence normalized to unit mass. The natural power-law analogue is therefore

$$w_\alpha(n) = \frac{n^{-(1+\alpha)}}{\zeta(1+\alpha)}, \quad \alpha > 0, \quad n \geq 1, \quad (9)$$

whose exponent $1 + \alpha$ matches the $\Theta(j^{-1-\alpha})$ decay of the Grünwald-Letnikov coefficients that underlie the fractional difference operator [?], this being the summable object appearing in the trace-boundedness lemma of FracTrace, and not the non-summable $k^{-\alpha}$ advantage kernel. Conflating the two decay rates would import a divergence that does not belong to the Bellman-mixing setting. This matching of decay rates is a motivating analogy and not a formal derivation. The operator $\mathcal{T}_w$ of Definition 1 is defined directly as a mixture over an arbitrary exogenous horizon-weight distribution w . The power-law instance (9) is a discrete probability weighting chosen to mirror the tail behavior of a fractional-order kernel, and it is not an object obtained by applying a fractional derivative or difference operator to the Bellman recursion itself. The verb “fractionalize” is retained throughout, following the terminology of Mousavi et al. [5], to name the question under investigation, namely whether the same tail-decay family that governs the fractional-order kernels of FracTrace can be substituted into the Bellman backup. It should not be read as a claim that $\mathcal{T}_w$ is itself constructed through fractional calculus. All results are consequently stated for a general weighting w , and specialization to (9) occurs only where closed forms become available.

3 Main Results

This section establishes that $\mathcal{T}_w$ always contracts in Section 3.1, that heavier tails contract more tightly in Section 3.2, and that closed-form moduli are available in Section 3.3. It further shows that the entire theory extends to state-dependent weights in Section 3.4, and that the modulus is sharp and convertible into an iteration complexity in Section 3.5.

3.1 Unconditional Contraction

Proposition 3 (*n*-step operator) *Under the bounded-reward condition, for every $n \geq 1$ the operator $\mathcal{T}^{(n)}$ maps $\mathcal{Z}$ into itself and satisfies $\bar{d}_p(\mathcal{T}^{(n)}Z_1, \mathcal{T}^{(n)}Z_2) \leq \gamma^n \bar{d}_p(Z_1, Z_2)$ for all $Z_1, Z_2 \in \mathcal{Z}$.*

Proof Invariance. Fix $Z \in \mathcal{Z}$, so that $|\zeta_n| \leq V_{\max}$ almost surely. For any horizon $n \geq 1$,

$$|G^{(n)}(s, a; Z)| \leq \sum_{k=0}^{n-1} \gamma^k R_{\max} + \gamma^n V_{\max} = R_{\max} \frac{1 - \gamma^n}{1 - \gamma} + \gamma^n \frac{R_{\max}}{1 - \gamma} = \frac{R_{\max}}{1 - \gamma} = V_{\max}, \quad (10)$$

so that $\mathcal{T}^{(n)}Z(s, a)$ is supported on $[-V_{\max}, V_{\max}]$ and therefore $\mathcal{T}^{(n)}Z \in \mathcal{Z}$.

Contraction. Fix (s, a) and $n \geq 1$. Couple $G^{(n)}(s, a; Z_1)$ and $G^{(n)}(s, a; Z_2)$ by using the *same* realization of the trajectory $S_{t+1}, \dots, S_{t+n}$ and of the rewards $R_{t+1}, \dots, R_{t+n-1}$ for both, alongside the shared deterministic initial reward $r(s, a)$, and, conditionally on the realized bootstrap state $S_{t+n} = s''$, by drawing (ζ_n^1, ζ_n^2) from an optimal W_p -coupling of $Z_1(s'', \pi(s''))$ and $Z_2(s'', \pi(s''))$. Under this coupling the reward prefix $r(s, a) + \sum_{k=1}^{n-1} \gamma^k R_{t+k}$ is identical on both sides and cancels, which leaves

$$W_p(\mathcal{T}^{(n)}Z_1(s, a), \mathcal{T}^{(n)}Z_2(s, a))^p \leq \mathbb{E}[|\gamma^n(\zeta_n^1 - \zeta_n^2)|^p] \quad (11)$$

$$= \gamma^{np} \mathbb{E}_{s''}[W_p(Z_1(s'', \pi(s'')), Z_2(s'', \pi(s'')))^p] \leq \gamma^{np} \bar{d}_p(Z_1, Z_2)^p. \quad (12)$$

The first inequality holds because a specific coupling upper-bounds the optimal-coupling infimum defining W_p , the intermediate equality follows from the tower property together with the optimality of the conditional couplings, and the final inequality follows from (3). Taking the supremum over (s, a) and then the p -th root gives the claim. $\square$

Theorem 4 (Unconditional contraction of $\mathcal{T}_w$) *Let w be any horizon-weight distribution and let $p \geq 1$. Then $\mathcal{T}_w$ maps $\mathcal{Z}$ into itself and is a contraction on $(\mathcal{Z}, \bar{d}_p)$,*

$$\bar{d}_p(\mathcal{T}_w Z_1, \mathcal{T}_w Z_2) \leq \kappa_p(w) \bar{d}_p(Z_1, Z_2), \quad \kappa_p(w) := \left(\sum_{n=1}^{\infty} w(n) \gamma^{np} \right)^{1/p}, \quad (13)$$

for all $Z_1, Z_2 \in \mathcal{Z}$. Moreover $\kappa_p(w) \leq \gamma$ holds for every admissible w , with equality if and only if $w(1) = 1$. Consequently $\mathcal{T}_w$ has a unique fixed point in $\mathcal{Z}$.

Proof Invariance of $\mathcal{Z}$ follows from Proposition 3, because a w -mixture of measures supported on $[-V_{\max}, V_{\max}]$ is again supported there. By Definition 1, for each (s, a) the measures $\mathcal{T}_w Z_1(s, a)$ and $\mathcal{T}_w Z_2(s, a)$ are the w -mixtures of $\{\mathcal{T}^{(n)}Z_1(s, a)\}_n$ and $\{\mathcal{T}^{(n)}Z_2(s, a)\}_n$. Applying Lemma 1 with $\omega_n = w(n)$ and then Proposition 3 yields

$$W_p(\mathcal{T}_w Z_1(s, a), \mathcal{T}_w Z_2(s, a))^p \leq \sum_{n \geq 1} w(n) W_p(\mathcal{T}^{(n)}Z_1(s, a), \mathcal{T}^{(n)}Z_2(s, a))^p \leq \left(\sum_{n \geq 1} w(n) \gamma^{np} \right) \bar{d}_p(Z_1, Z_2)^p. \quad (14)$$

Taking the supremum over (s, a) and the p -th root yields (13). For the modulus bound, note that $n \geq 1$ and $\gamma \in (0, 1)$ together imply $\gamma^{np} \leq \gamma^p$, and hence

$$\kappa_p(w)^p = \sum_{n \geq 1} w(n) \gamma^{np} \leq \gamma^p \sum_{n \geq 1} w(n) = \gamma^p, \quad (15)$$

so that $\kappa_p(w) \leq \gamma$. For the equality condition, observe that $\gamma^p - \kappa_p(w)^p = \sum_{n \geq 1} w(n)(\gamma^p - \gamma^{np})$, where every summand is nonnegative and is strictly positive whenever $n \geq 2$ and $w(n) > 0$. The sum therefore vanishes if and only if $w(n) = 0$ for all $n \geq 2$, which is the condition $w(1) = 1$. Finally, $\kappa_p(w)^p \leq \gamma^p < 1$ shows that $\mathcal{T}_w$ is a genuine contraction, and completeness of $(\mathcal{Z}, \bar{d}_p)$ together with the Banach fixed-point theorem gives a unique fixed point. $\square$

Corollary 5 (Fixed-point preservation) *Let w be any horizon-weight distribution and let $p \geq 1$. The return-distribution function Z^π of Proposition 2 satisfies $\mathcal{T}^{(n)}Z^\pi = Z^\pi$ for every $n \geq 1$, and consequently $\mathcal{T}_w Z^\pi = Z^\pi$. The unique fixed point of $\mathcal{T}_w$ furnished by Theorem 4 is therefore Z^π itself, and not merely some stationary point of the mixture.*

Proof Fix (s, a) and $n \geq 1$, and let $\zeta_n \sim Z^\pi(S_{t+n}, \pi(S_{t+n}))$ as in (7), coupled with the actual continuation of the trajectory from S_{t+n} under π . This coupling is licensed by the Markov property, since $Z^\pi(S_{t+n}, \pi(S_{t+n}))$ is by definition the law of $\sum_{j \geq 0} \gamma^j R_{t+n+j}$ for an independent continuation under π , which has the same law as the actual future of the same trajectory conditional on S_{t+n} . Under this coupling,

$$G^{(n)}(s, a; Z^\pi) \stackrel{d}{=} r(s, a) + \sum_{k=1}^{n-1} \gamma^k R_{t+k} + \gamma^n \sum_{j \geq 0} \gamma^j R_{t+n+j} = \sum_{k \geq 0} \gamma^k R_{t+k}, \quad (16)$$

which is exactly the true return starting from (s, a) and following π throughout, whose law is $Z^\pi(s, a)$ by definition. Hence $\mathcal{T}^{(n)} Z^\pi(s, a) = \text{Law}[G^{(n)}(s, a; Z^\pi)] = Z^\pi(s, a)$ for every $n \geq 1$ and every (s, a) , so that $\mathcal{T}^{(n)} Z^\pi = Z^\pi$. By Definition 1, $\mathcal{T}_w Z^\pi = \sum_{n \geq 1} w(n) \mathcal{T}^{(n)} Z^\pi = \sum_{n \geq 1} w(n) Z^\pi = Z^\pi$, using $\sum_n w(n) = 1$. Since Theorem 4 establishes that $\mathcal{T}_w$ has a unique fixed point, that fixed point must coincide with Z^π . $\square$

Theorem 4 settles the question left open by Mousavi et al. [5], since mixing returns across many horizons under power-law weights preserves γ -contraction and the resulting modulus never exceeds that of the one-step operator. The underlying mechanism is transparent, because a longer horizon n defers bootstrapping and multiplies the residual discrepancy by $\gamma^n \leq \gamma$, so that any weight placed beyond $n = 1$ can only tighten the contraction. Corollary 5 completes the safety argument, since contraction alone only guarantees convergence to some fixed point, and the fractionalized critic would be unsafe if that fixed point were not the correct return-distribution function. The corollary rules out this possibility directly.

Remark 1 (Reconciliation with the design of FracTrace) Theorem 4 shows that the anticipated obstruction is not fatal, because a *normalized* power-law Bellman mixture does contract. This finding nevertheless does not contradict the design of FracTrace. The advantage estimator employed there is not a normalized horizon mixture but a convolution against a *non-summable* lag kernel, as discussed in Section 2.5, and restricting fractionalization to the policy side is exactly what permits the use of that heavier and non-probability kernel without requiring it to serve as a Bellman backup. The present result therefore clarifies the architecture instead of overturning it, since the critic *could* be fractionalized safely with a summable horizon-weight mixture, although doing so would couple two roles that FracTrace deliberately keeps separate.

3.2 Comparative Contraction and the Role of Tail Heaviness

Theorem 4 bounds each $\kappa_p(w)$ by γ , and the moduli are now ordered across weight families. Write $\bar{F}_w(n) := \sum_{m \geq n} w(m)$ for the upper-tail function of w . Recall that w_2 *first-order stochastically dominates* w_1 toward larger horizons, written $w_2 \succeq_{\text{st}} w_1$, whenever $\bar{F}_{w_2}(n) \geq \bar{F}_{w_1}(n)$ for every $n \geq 1$, or equivalently whenever $\mathbb{E}_{w_2}[h(n)] \geq \mathbb{E}_{w_1}[h(n)]$ for every nondecreasing bounded function h .

Theorem 6 (Comparative contraction) *Let w_1, w_2 be horizon-weight distributions with $w_2 \succeq_{\text{st}} w_1$, and let $p \geq 1$. Then*

$$\kappa_p(w_2) \leq \kappa_p(w_1). \quad (17)$$

Within the power-law family (9), the modulus $\kappa_p(w_\alpha)$ is strictly increasing in α , so that a lighter tail produces a weaker contraction. Consequently any power-law member that stochastically dominates a geometric reference w_λ contracts at least as tightly as that reference.

Proof Write $\kappa_p(w)^p = \sum_{n \geq 1} w(n) \gamma^{np} = \mathbb{E}_{n \sim w}[g(n)]$ with $g(n) := \gamma^{np}$. Because $\gamma \in (0, 1)$, the function g is strictly decreasing in n . First-order stochastic dominance $w_2 \succeq_{\text{st}} w_1$ is equivalent to the inequality $\mathbb{E}_{w_2}[g] \leq \mathbb{E}_{w_1}[g]$ for every nonincreasing g , and applying this equivalence yields $\kappa_p(w_2)^p \leq \kappa_p(w_1)^p$, which is (17).

For strict monotonicity within (9), fix $\alpha_1 < \alpha_2$. The likelihood ratio $w_{\alpha_1}(n)/w_{\alpha_2}(n) \propto n^{\alpha_2 - \alpha_1}$ is strictly increasing in n , so that w_{α_1} dominates w_{α_2} in the monotone likelihood ratio order. That ordering implies the strict first-order dominance $w_{\alpha_1} \succ_{\text{st}} w_{\alpha_2}$, and therefore $\kappa_p(w_{\alpha_1}) < \kappa_p(w_{\alpha_2})$, since g is strictly decreasing while w_{α_1} and w_{α_2} place distinct mass beyond $n = 1$. The geometric comparison is the special case $w_1 = w_\lambda$ and $w_2 = w_\alpha$ of (17). $\square$

Remark 2 (Why a convexity argument is the wrong instrument) An argument through convexity is tempting, because $g(n) = \gamma^{np}$ is convex in n , so that a more spread-out weighting would appear, by Jensen's inequality, to *increase* $\mathbb{E}[g]$ and hence to *weaken* the contraction. That conclusion is the opposite of Theorem 6. The intuition is misleading because a comparison of tail-parameter families such as (9) across α is *not* a mean-preserving spread. Heavier tails do not merely redistribute mass at a fixed mean, since they shift mass toward larger n altogether, and that shift is a stochastic-dominance comparison. For a *monotone* integrand, first-order dominance fixes the direction unambiguously, whereas the convexity argument simply does not apply. This distinction is flagged explicitly because it constitutes an easy trap, and the correct instrument is monotonicity under stochastic dominance and not convexity under spread.

The convexity of g is nevertheless not vacuous. It governs comparisons at a fixed *mean* horizon instead of under a dominance shift, and it is recorded next in that complementary role.

Proposition 7 (Contraction–cost frontier at fixed mean horizon) *Let $\gamma \in (0, 1)$ and $p \geq 1$. Let w be a horizon-weight distribution with finite mean $m := \mathbb{E}_{n \sim w}[n] < \infty$. Then*

$$\kappa_p(w) \geq \gamma^m. \quad (18)$$

Proof Write $\kappa_p(w)^p = \mathbb{E}_{n \sim w}[\gamma^{np}]$. The map $n \mapsto \gamma^{np} = e^{np \ln \gamma}$ is convex on $[1, \infty)$ for every fixed $\gamma \in (0, 1)$ and $p \geq 1$, being an exponential function of n . Jensen's inequality therefore gives $\mathbb{E}_{n \sim w}[\gamma^{np}] \geq \gamma^p \mathbb{E}_{n \sim w}[n] = \gamma^{pm}$, and taking p -th roots yields (18). $\square$

The bound (18) is not sharp when m is not an integer, because it is attained only by the deterministic horizon $N \equiv m$, and N is constrained to the admissible support $\{1, 2, 3, \dots\}$. The sharp bound among horizon-weight distributions with mean m is recorded next.

Corollary 8 (Sharp frontier for integer-valued horizons) *Let $k := \lfloor m \rfloor$ and $\theta := m - k \in [0, 1)$. Among all horizon-weight distributions with mean m , the sharp lower bound is*

$$\kappa_p(w)^p \geq (1 - \theta)\gamma^{pk} + \theta\gamma^{p(k+1)}, \quad (19)$$

attained by the two-point distribution $w(k) = 1 - \theta$, $w(k + 1) = \theta$. For integer m , $\theta = 0$ and (19) reduces to $\kappa_p(w) \geq \gamma^m$ with equality at the point mass $N \equiv m$, recovering (18).

Proof Let $L(x) := \gamma^{pk} + (\gamma^{p(k+1)} - \gamma^{pk})(x - k)$ denote the affine function agreeing with $x \mapsto \gamma^{px}$ at $x = k$ and at $x = k + 1$. Because $x \mapsto \gamma^{px}$ is convex, it lies on or below L on the interval $[k, k + 1]$ and on or above L outside that interval. In particular $\gamma^{pn} \geq L(n)$ for every integer $n \geq 1$, since no integer lies strictly between the consecutive integers k and $k + 1$. For any horizon-weight distribution w with mean m ,

$$\kappa_p(w)^p = \mathbb{E}_{n \sim w}[\gamma^{pn}] \geq \mathbb{E}_{n \sim w}[L(n)] = L(m) = (1 - \theta)\gamma^{pk} + \theta\gamma^{p(k+1)}, \quad (20)$$

using that L is affine so that $\mathbb{E}_{n \sim w}[L(n)] = L(\mathbb{E}_{n \sim w}[n]) = L(m)$. Equality holds when w places all of its mass on $\{k, k + 1\}$, since $\gamma^{pn} = L(n)$ exactly at those two points. $\square$

Remark 3 (Scope of the frontier bound) Proposition 7 and Corollary 8 caution against a natural misreading of Theorem 6, namely that dispersion or heavy-tailedness is itself a free source of tighter contraction. At a *fixed* mean bootstrap horizon m , the sharp frontier (19) shows the achievable modulus is bounded below, so that no redistribution of mass among horizons summing to the same mean can beat the two-point frontier concentrated at $\lfloor m \rfloor$ and $\lceil m \rceil$. The gains reported later in this paper are instead purchased by shifting mass toward *longer* mean horizons, exactly as the stochastic-dominance mechanism of Theorem 6 prescribes, and not by dispersion at a fixed mean. The bound is in fact vacuous for the untruncated power-law family (9) used throughout this paper, since $\mathbb{E}_{n \sim w_\alpha}[n] = \sum_{n \geq 1} n^{-\alpha} / \zeta(1 + \alpha)$ diverges for every $\alpha \in (0, 1]$. The bound γ^m is then 0 and carries no information in that regime, a divergence that reflects the idealized weighting and not the bound itself.

3.3 Closed-Form Contraction Moduli

Corollary 9 (Closed forms) *Let $p \geq 1$ and $\gamma \in (0, 1)$.*

- (i) *For the geometric family $w_\lambda(n) = (1 - \lambda)\lambda^{n-1}$ with $\lambda \in [0, 1)$,*

$$\kappa_p(w_\lambda)^p = \frac{(1 - \lambda)\gamma^p}{1 - \lambda\gamma^p}. \quad (21)$$

- (ii) *For the untruncated power-law family (9) with $\alpha > 0$,*

$$\kappa_p(w_\alpha)^p = \frac{\text{Li}_{1+\alpha}(\gamma^p)}{\zeta(1+\alpha)}, \quad \text{Li}_s(\mu) := \sum_{n=1}^{\infty} \frac{\mu^n}{n^s}, \quad (22)$$

where Li_s is the polylogarithm and ζ is the Riemann zeta function. For a finite truncation horizon J , the normalized truncated weight $w_\alpha^J(n) = n^{-(1+\alpha)} / \sum_{m=1}^J m^{-(1+\alpha)}$ gives $\kappa_p(w_\alpha^J)^p = (\sum_{n=1}^J n^{-(1+\alpha)} \gamma^{np}) / (\sum_{n=1}^J n^{-(1+\alpha)})$, which converges to (22) as $J \rightarrow \infty$.

Proof (i) Substituting the geometric weights,

$$\kappa_p(w_\lambda)^p = \sum_{n \geq 1} (1 - \lambda)\lambda^{n-1} \gamma^{np} = (1 - \lambda)\gamma^p \sum_{n \geq 1} (\lambda\gamma^p)^{n-1} = \frac{(1 - \lambda)\gamma^p}{1 - \lambda\gamma^p}, \quad (23)$$

where the geometric series converges because $\lambda\gamma^p < 1$.

- (ii) With the normalizing constant $\sum_{n \geq 1} n^{-(1+\alpha)} = \zeta(1+\alpha)$, which is finite for $\alpha > 0$,

$$\kappa_p(w_\alpha)^p = \frac{1}{\zeta(1+\alpha)} \sum_{n \geq 1} \frac{(\gamma^p)^n}{n^{1+\alpha}} = \frac{\text{Li}_{1+\alpha}(\gamma^p)}{\zeta(1+\alpha)}, \quad (24)$$

where the series converges absolutely since $\gamma^p < 1$. The truncated statement is the same computation carried out over $n \leq J$, and dominated convergence gives the stated limit. $\square$

Remark 4 (Numerical note) The polylogarithm form (22) is not merely cosmetic. For small α , which corresponds to heavy tails, the truncated sum $\kappa_p(w_\alpha^J)$ converges to (22) only slowly in J , so that a naive truncation *overstates* the modulus. The closed form is consequently the reliable route for evaluating the heavy-tail regime of interest here.

3.4 State-Dependent Weight Families

The architecture of FracTrace does not employ a fixed horizon weighting, since its kernels are *state-modulated*, so that the effective weighting at a state s is a state-indexed distribution. The entire theory survives under state dependence, provided that the weights are normalized *per state* exactly as in the mixture construction. This is the point at which the present Bellman-mixing analysis becomes simpler than the advantage-kernel analysis of FracTrace, because no renormalization-inflation constant of the form $1/[(1 - \epsilon_{\text{exp}})m_{\text{min}}]$ is incurred. Per-state normalization is intrinsic to the mixture operator and is therefore not an external correction applied to a non-summable kernel.

Let $\{w_s\}_{s \in \mathcal{S}}$ be a family of horizon-weight distributions, each satisfying $w_s(n) \geq 0$ and $\sum_{n \geq 1} w_s(n) = 1$. Define

$$(\mathcal{T}_w Z)(s, a) := \sum_{n=1}^{\infty} w_s(n) (\mathcal{T}^{(n)} Z)(s, a). \quad (25)$$

The symbol $\mathcal{T}_w$ is retained in (25) for the state-dependent operator, since (8) is recovered as the special case in which w_s does not depend on s . The subscript w denotes whichever weighting object, a single distribution or a state-indexed family of distributions, is in force in the surrounding context, the two uses are never invoked simultaneously for the same operator, and no ambiguity arises in what follows.

Theorem 10 (Uniform contraction and comparison under state dependence) *Let $\{w_s\}_{s \in \mathcal{S}}$ be a state-indexed family of horizon-weight distributions, fixed independently of Z (that is, each w_s is a measurable function of s alone and is not updated jointly with the iterate), and let $p \geq 1$. Then (25) maps $\mathcal{Z}$ into itself and is a contraction in $\bar{d}_p$ with*

uniform modulus

$$\bar{d}_p(\mathcal{T}_w Z_1, \mathcal{T}_w Z_2) \leq \kappa_p^{\sup} \bar{d}_p(Z_1, Z_2), \quad \kappa_p^{\sup} := \left(\sup_{s \in \mathcal{S}} \sum_{n \geq 1} w_s(n) \gamma^{np} \right)^{1/p} \leq \gamma. \quad (26)$$

Moreover, if $\{w_s^{(1)}\}$ and $\{w_s^{(2)}\}$ satisfy the uniform dominance $w_s^{(2)} \succeq_{\text{st}} w_s^{(1)}$ for every $s \in \mathcal{S}$, then $\kappa_p^{\sup}(w^{(2)}) \leq \kappa_p^{\sup}(w^{(1)})$. Hence $\mathcal{T}_w$ has a unique fixed point in $\mathcal{Z}$, and heavier state-wise tails yield no larger worst-case modulus. By the argument of Corollary 5, applied separately at each state, that fixed point is Z^π . Since $\mathcal{T}^{(n)} Z^\pi = Z^\pi$ for every n , it follows that $\sum_{n \geq 1} w_s(n) \mathcal{T}^{(n)} Z^\pi = Z^\pi \sum_{n \geq 1} w_s(n) = Z^\pi$ at every s , regardless of how the weighting varies across states.

Proof Invariance $\mathcal{T}_w(\mathcal{Z}) \subseteq \mathcal{Z}$ follows as in Proposition 3, since the bound established there holds for each n and is preserved under any per-state mixture. Fix (s, a) . As in the proof of Theorem 4, Lemma 1 applied with weights $w_s(n)$ together with Proposition 3 gives

$$W_p(\mathcal{T}_w Z_1(s, a), \mathcal{T}_w Z_2(s, a))^p \leq \left(\sum_{n \geq 1} w_s(n) \gamma^{np} \right) \bar{d}_p(Z_1, Z_2)^p \leq \left(\sup_{s'} \sum_{n \geq 1} w_{s'}(n) \gamma^{np} \right) \bar{d}_p(Z_1, Z_2)^p. \quad (27)$$

Taking the supremum over (s, a) and then the p -th root gives (26). The bound $\kappa_p^{\sup} \leq \gamma$ holds because $\sum_{n \geq 1} w_s(n) \gamma^{np} \leq \gamma^p$ for every s by the argument of (15), each w_s being a probability distribution over $n \geq 1$. For the comparative claim, apply Theorem 6 pointwise, so that $\sum_n w_s^{(2)}(n) \gamma^{np} \leq \sum_n w_s^{(1)}(n) \gamma^{np}$ for each s . Taking the supremum over s preserves the inequality, since $\sup_s a_s \leq \sup_s b_s$ whenever $a_s \leq b_s$ for all s . Uniqueness follows from $\kappa_p^{\sup} \leq \gamma < 1$ together with the Banach fixed-point theorem. $\square$

Remark 5 (Relation to the state modulation of FracTrace) Theorem 10 resolves, and indeed strengthens, the state-dependent extension that Mousavi et al. [5] scoped without carrying out. The anticipated obstruction, namely an inflation constant surviving renormalization, arises only for a kernel that is normalized *after* an un-normalized modulation. In the Bellman-mixing operator the weights form a probability distribution over horizons at every state by construction, so that the worst-case modulus is a plain supremum over states with no inflation factor, and the comparative law survives under uniform stochastic dominance. Should the exact un-normalized modulated kernel of FracTrace be insisted upon, however, the appropriate object is no longer a Bellman backup, and it falls under the advantage-side truncation analysis of that framework instead of the contraction theory developed here. Indexing the weight family by the state-action pair instead of by state alone, so that $\{w_{s,a}\}_{(s,a) \in \mathcal{S} \times \mathcal{A}}$ replaces $\{w_s\}_{s \in \mathcal{S}}$ throughout, changes none of the proof of Theorem 10, since the supremum in (26) is already taken over the same index set that labels the operator. This finer-grained extension is noted here and not developed separately, as it is not required by the state-indexed kernels of FracTrace.

3.5 Sharpness of the Modulus and Convergence Rate

To make the dependence of the modulus on the specific MDP explicit, write

$$L_p(M, w) := \sup_{Z_1 \neq Z_2} \frac{\bar{d}_p(\mathcal{T}_w^M Z_1, \mathcal{T}_w^M Z_2)}{\bar{d}_p(Z_1, Z_2)} \quad (28)$$

for the exact Lipschitz constant of $\mathcal{T}_w$ on a specific admissible MDP M , where $\mathcal{T}_w^M$ denotes the operator of Definition 1 built from the transition kernel and reward of M . Theorem 4 is then the statement that $\kappa_p(w)$ is a uniform contraction factor, in the sense that $L_p(M, w) \leq \kappa_p(w)$ for every admissible M . The bound is now shown to be *attained* by some particular MDP in the admissible class, so that $\kappa_p(w) = \sup_M L_p(M, w)$ is in fact the sharp worst-case modulus, and it cannot be improved by any sharper argument that remains uniform over the class. The modulus is then converted into an explicit iteration complexity, which is the form in which a contraction rate acquires practical meaning.

Proposition 11 (Attainment of the modulus) *Let w be any horizon-weight distribution and let $p \geq 1$. There exist an MDP M^* satisfying the bounded-reward condition and return-distribution functions $Z_1, Z_2 \in \mathcal{Z}$ with $Z_1 \neq Z_2$*

such that

$$\bar{d}_p(\mathcal{T}_w Z_1, \mathcal{T}_w Z_2) = \kappa_p(w) \bar{d}_p(Z_1, Z_2) \quad (29)$$

for the operator $\mathcal{T}_w = \mathcal{T}_w^{M^*}$ induced by M^* , so that $L_p(M^*, w) = \kappa_p(w)$. Since Theorem 4 already gives $L_p(M, w) \leq \kappa_p(w)$ for every admissible M , this attainment yields

$$\sup_M L_p(M, w) = \kappa_p(w). \quad (30)$$

Thus $\kappa_p(w)$ is the exact worst-case Lipschitz constant of $\mathcal{T}_w$ over the admissible MDP class, and the bound (13) is tight for every w and every p .

Proof Throughout this proof, $\mathcal{T}_w$ and $\mathcal{T}^{(n)}$ denote the operators induced by the single MDP M^* constructed below. Fix any $R_{\max} > 0$ and let $V_{\max} = R_{\max}/(1 - \gamma)$. Take the single-state and single-action MDP with $\mathcal{S} = \{s\}$ and $\mathcal{A} = \{a\}$, equipped with the deterministic self-loop $P(s | s, a) = 1$ and the identically zero reward $r(s, a) = 0$, which satisfies $|r| \leq R_{\max}$. Along the unique trajectory all rewards vanish, so that (7) reduces to $G^{(n)}(s, a; Z) = \gamma^n \zeta_n$ with $\zeta_n \sim Z(s, a)$, and hence $\mathcal{T}^{(n)} Z(s, a)$ is the law of $\gamma^n \zeta$ for $\zeta \sim Z(s, a)$.

Let $Z_1(s, a) = \delta_0$ and $Z_2(s, a) = \delta_c$ with $c := V_{\max} > 0$. Both belong to $\mathcal{Z}$, since their supports are contained in $[-V_{\max}, V_{\max}]$. Then $\mathcal{T}^{(n)} Z_1(s, a) = \delta_0$ for every n , so that $\mathcal{T}_w Z_1(s, a) = \delta_0$, whereas $\mathcal{T}^{(n)} Z_2(s, a) = \delta_{\gamma^n c}$ and therefore $\mathcal{T}_w Z_2(s, a) = \sum_{n \geq 1} w(n) \delta_{\gamma^n c}$. Because the first argument is a Dirac mass at the origin, the coupling is forced and the Wasserstein distance is computed by direct integration, which gives

$$W_p(\delta_0, \mathcal{T}_w Z_2(s, a))^p = \int |x|^p \mathcal{T}_w Z_2(s, a)(dx) = \sum_{n \geq 1} w(n) (\gamma^n c)^p = \kappa_p(w) c^p. \quad (31)$$

Since the state-action space is a singleton, $\bar{d}_p(\mathcal{T}_w Z_1, \mathcal{T}_w Z_2) = \kappa_p(w) c$ while $\bar{d}_p(Z_1, Z_2) = W_p(\delta_0, \delta_c) = c$, which establishes (29). The identity (30) then follows from Theorem 4, which bounds every $L_p(M, w)$ by $\kappa_p(w)$. $\square$

Remark 6 (Interpretation of sharpness) Proposition 11 shows that the comparative law of Theorem 6 orders *exact worst-case* Lipschitz constants, taken over the admissible MDP class, and not merely upper bounds. Statements asserting that heavier tails contract more tightly therefore cannot be reversed by a more refined analysis of that worst case, since the ordering is a property of $\sup_M L_p(M, w)$, established uniformly over the class, and not an artifact of the bound. It should nevertheless be noted that the extremal instance is degenerate by design, because it isolates the bootstrap discount from any reward or transition structure, so that $L_p(M, w)$ for a typical M can be, and empirically is, strictly smaller than $\kappa_p(w)$. The modulus is consequently attained in the worst case, although it is typically conservative on a given MDP, a gap that is quantified empirically later in this paper.

Contraction moduli acquire operational meaning through the number of applications of the operator required in order to reach a target accuracy, and that count is established next.

Corollary 12 (Iteration complexity) *Let w be a horizon-weight distribution, let $p \geq 1$, and let Z^* denote the unique fixed point of $\mathcal{T}_w$ furnished by Theorem 4, which Corollary 5 identifies as $Z^* = Z^\pi$. For any $Z_0 \in \mathcal{Z}$ define the iterates $Z_{k+1} = \mathcal{T}_w Z_k$ and set $D := \bar{d}_p(Z_0, Z^*)$. Then*

$$\bar{d}_p(Z_k, Z^*) \leq \kappa_p(w)^k D \quad \text{for all } k \geq 0, \quad (32)$$

and, for any target accuracy $\epsilon \in (0, D)$, the iteration count

$$k(\epsilon) = \left\lceil \frac{\ln(\epsilon/D)}{\ln \kappa_p(w)} \right\rceil \quad (33)$$

suffices to guarantee $\bar{d}_p(Z_k, Z^*) \leq \epsilon$. Moreover $D \leq 2V_{\max}$, so that (33) holds uniformly over initializations with D replaced by $2V_{\max}$. Finally, if $w_2 \succeq_{\text{st}} w_1$, then

$$\frac{k_2(\epsilon)}{k_1(\epsilon)} \longrightarrow \frac{\ln \kappa_p(w_1)}{\ln \kappa_p(w_2)} \leq 1 \quad \text{as } \epsilon \downarrow 0, \quad (34)$$

so that heavier-tailed horizon weightings require asymptotically no more iterations, and generically strictly fewer.

Proof Inequality (32) follows by induction from Theorem 4, since $\bar{d}_p(Z_{k+1}, Z^*) = \bar{d}_p(\mathcal{T}_w Z_k, \mathcal{T}_w Z^*) \leq \kappa_p(w) \bar{d}_p(Z_k, Z^*)$ by the fixed-point identity $\mathcal{T}_w Z^* = Z^*$. Requiring $\kappa_p(w)^k D \leq \epsilon$ and taking logarithms, where $\ln \kappa_p(w) < 0$ because $\kappa_p(w) \leq \gamma < 1$, reverses the inequality and yields $k \geq \ln(\epsilon/D)/\ln \kappa_p(w)$, which is (33). The bound $D \leq 2V_{\max}$ holds because every element of $\mathcal{Z}$ is supported in $[-V_{\max}, V_{\max}]$, so that any coupling gives $W_p \leq 2V_{\max}$. For (34), drop the ceiling and note that $k_i(\epsilon) = \ln(\epsilon/D)/\ln \kappa_p(w_i)$, so that the ratio equals $\ln \kappa_p(w_1)/\ln \kappa_p(w_2)$ exactly, while the ceiling contributes an $O(1)$ term that vanishes relative to $k_i(\epsilon) \rightarrow \infty$. By Theorem 6, $\kappa_p(w_2) \leq \kappa_p(w_1) < 1$, whence $\ln \kappa_p(w_2) \leq \ln \kappa_p(w_1) < 0$ and the displayed ratio is at most one. $\square$

Remark 7 (The rate improvement is not obtained without cost) Corollary 12 counts applications of $\mathcal{T}_w$, and a single application of the multi-horizon operator is more expensive than a single one-step backup, because evaluating (8) requires the n -step backups up to the truncation horizon J . Since $\mathcal{T}^{(n)} = (\mathcal{T}^\pi)^n$ holds for the evaluation operator, these backups can be produced by J successive one-step applications and reused across the mixture, so that the cost per application of $\mathcal{T}_w$ is $O(J)$ one-step backups. The comparison in (34) should therefore be read per operator application and not per backup. The practical gain is consequently governed by whether the reduction in $k(\epsilon)$ outweighs the factor J . No wall-clock advantage is claimed here, only the rate statement, and the cost-adjusted comparison is regarded as an open empirical question. This accounting concerns the exact, deterministic computation of the full truncated mixture, in which every n -step component up to J is evaluated and combined. A distinct computational model arises from a stochastic sample-based update, in which a single horizon $N \sim w$ is drawn and only the resulting N -step target is computed. The cost of that single target is $O(N)$, so the expected cost of the update is governed by $\mathbb{E}[N]$ and not by J . The two computational models are not interchangeable, and the $O(J)$ accounting given above applies specifically to the exact truncated mixture. Remark 3 records why this distinction matters for the untruncated power-law family, since $\mathbb{E}[N] = \infty$ for every $\alpha \in (0, 1]$ makes the sampled-target model inapplicable in that regime without truncation.

Remark 8 (The excluded boundary case $\gamma = 0$) The standing assumption $\gamma \in (0, 1)$ fixed in Section 2 excludes a genuine boundary case, and its consequences are recorded here now that the relevant quantities are available. At $\gamma = 0$, every term γ^{np} with $n \geq 1$ vanishes, so that $\kappa_p(w) = 0$ for every admissible w by (13). The equality condition of Theorem 4 accordingly holds for every w , and not uniquely at the point mass $w(1) = 1$. The strict monotonicity asserted in Theorem 6 becomes vacuous as well, since every modulus coincides at 0. The iteration formula (33) is undefined at this boundary, because $\ln \kappa_p(w)$ is undefined at $\kappa_p(w) = 0$. These degeneracies are consistent with the observation in Section 2 that $\gamma = 0$ reduces every operator considered here to a single deterministic evaluation of the reward, for which no iterative contraction argument is required.

4 Numerical Illustration

This section instantiates the closed-form moduli of Corollary 9 and locates the learned slow-level order range of FracTrace on the resulting contraction map. All tabulated and plotted values are computed directly from the closed forms (21) and (22), so that every entry is reproducible to the stated precision without recourse to simulation.

4.1 Contraction Moduli Across Weight Families

The discount factor is fixed at $\gamma = 0.99$, in agreement with the setting used throughout [5], and the contraction modulus $\kappa_p(w)$ is evaluated for $p \in \{1, 2\}$ across three weight families instantiated at eight distinct weightings. The families comprise the point mass at $n = 1$, which is the standard one-step operator and for which $\kappa_p = \gamma$ by Theorem 4, together with the geometric family w_λ evaluated at the reference $\lambda = 0.95$ used as the exponential baseline in [5]. A third family, the power-law family w_α of (9), is evaluated over a grid of orders α that spans the light-tail regime at $\alpha = 0.90$ down to and through the learned slow-level range $\alpha \in [0.08, 0.16]$. Each modulus is computed in two independent ways, namely from the closed form of Corollary 9 and by direct summation truncated at a large horizon J , and the two evaluations are cross-checked against one another. In accordance with Remark 4, the closed form is treated as authoritative in the heavy-tail regime, where truncation converges slowly.

Table 1 Contraction modulus $\kappa_p(w)$ at $\gamma = 0.99$ for $p \in \{1, 2\}$, computed from the closed forms of Corollary 9. The one-step row equals γ by Theorem 4; every other entry is strictly smaller, and the power-law entries increase strictly with α by Theorem 6.

Weight family	κ_p ($p=1$)	κ_p ($p=2$)
Point mass $w = \delta_1$ (one-step)	0.9900	0.9900
Geometric, $\lambda = 0.95$ (reference)	0.8319	0.8433
Power-law, $\alpha = 0.90$	0.9580	0.9637
Power-law, $\alpha = 0.50$	0.8696	0.9049
Power-law, $\alpha = 0.30$	0.7255	0.8147
Power-law, $\alpha = 0.16$ ¹	0.5097	0.6731
Power-law, $\alpha = 0.10$ ¹	0.3632	0.5640
Power-law, $\alpha = 0.08$ ¹	0.3041	0.5146

¹ Within FracTrace’s learned slow-level order range $\alpha \in [0.08, 0.16]$ [5].

Table 1 reports $\kappa_p(w)$ at the eight weightings for $p \in \{1, 2\}$. The one-step entry equals $\gamma = 0.99$ exactly, and by Theorem 4 every other entry is strictly smaller, while the power-law entries increase strictly with α as Theorem 6 requires. Two features deserve particular emphasis. The first is a crossover already visible in the table, since the light-tailed member at $\alpha = 0.90$ contracts at 0.9580 for $p = 1$ and is therefore *weaker* than the geometric reference at 0.8319. This confirms that the power-law family does not dominate the geometric family universally. The second is that the ordering reverses decisively inside the learned range, because at $\alpha = 0.08$ the modulus falls to 0.3041 for $p = 1$. This represents a reduction of roughly 63% relative to the geometric reference and roughly 69% relative to the one-step operator. Solving $\kappa_p(w_\alpha) = \kappa_p(w_\lambda)$ locates the crossover at $\alpha^* \approx 0.4285$ for $p = 1$ and at $\alpha^* \approx 0.3464$ for $p = 2$, so that power-law Bellman mixing out-contracts the exponential baseline precisely when $\alpha < \alpha^*$, and the learned range of FracTrace lies well inside that regime.

4.2 Contraction Map and the FracTrace Regime

Table 1 evaluates the contraction modulus at eight discrete weightings. In order to expose the full dependence on tail heaviness, and in particular to locate the boundary beyond which power-law Bellman mixing ceases to be advantageous, the power-law order is now swept continuously. Figure 1 plots $\kappa_p(w_\alpha)$ as a function of α over the range $[0.05, 1]$ at $\gamma = 0.99$, computed from the polylogarithm form (22) for $p = 1$ and $p = 2$. Two horizontal references are overlaid, namely the geometric modulus (21) at the exponential baseline $\lambda = 0.95$, and the one-step value $\kappa_p = \gamma$, which by Theorem 4 upper-bounds *every* admissible horizon weighting and therefore acts as a ceiling on the entire plot. The shaded band marks the learned slow-level order range $\alpha \in [0.08, 0.16]$ [5]. Markers placed on the two curves identify the orders tabulated in Table 1, so that the figure and the table remain mutually verifiable at those points.

Four features of Figure 1 deserve to be read off explicitly, because each corresponds to a statement proved in Section 3.

The ceiling is never violated. Both curves lie strictly below the dotted line $\kappa_p = \gamma$ across the entire sweep, which is the graphical content of the bound $\kappa_p(w) \leq \gamma$ in Theorem 4 together with its equality condition. Since every w_α places positive mass on horizons $n \geq 2$, the inequality is strict for all $\alpha > 0$. The one-step operator is therefore the *worst* contraction among all admissible horizon weightings and not the best, and no choice of α can degrade the modulus below the guarantee already enjoyed by the standard critic.

Monotonicity in the order α . Both curves increase strictly and smoothly across the plotted range, which is exactly the comparative law of Theorem 6 rendered continuously. Increasing α lightens the tail of w_α and thereby effects a first-order stochastic-dominance decrease toward shorter horizons, so that the modulus increases accordingly. The visual steepness at small α is informative in its own right, since the contraction benefit is concentrated in the heavy-tail regime and decays rapidly thereafter. The practical advantage of power-law mixing is consequently not uniform across the family and is instead purchased specifically by heavy tails.

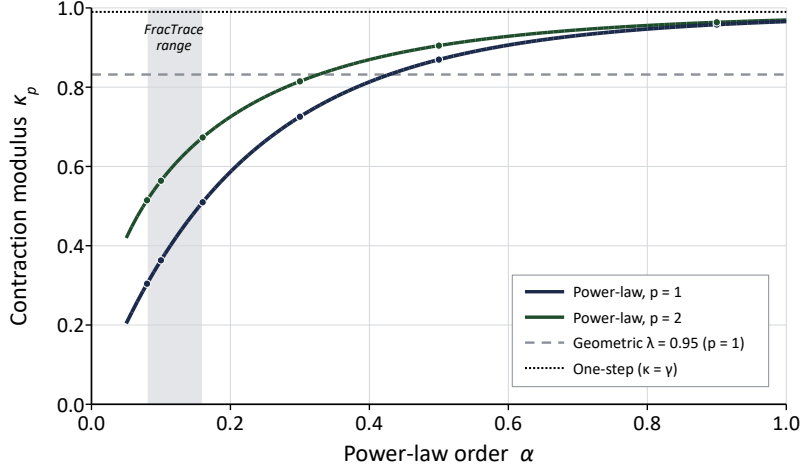


Fig. 1 Contraction modulus $\kappa_p(w_\alpha)$ versus power-law order α at $\gamma = 0.99$, for $p = 1$ (dark navy) and $p = 2$ (dark green); markers indicate the orders tabulated in Table 1. The dashed grey line is the geometric reference $\lambda = 0.95$ at $p = 1$ ($\kappa_p = 0.8319$); the dotted black line is the one-step value $\kappa_p = \gamma = 0.99$. The shaded band marks FracTrace’s learned range $\alpha \in [0.08, 0.16]$ [5]. The curves cross the geometric reference at $\alpha^* \approx 0.4285$ ($p = 1$) and $\alpha^* \approx 0.3464$ ($p = 2$).

The two asymptotic limits. As $\alpha \rightarrow \infty$ the weight w_α converges to the point mass δ_1 , and consistently $\kappa_p(w_\alpha) \rightarrow \gamma$, so that the power-law family degenerates continuously into the one-step operator and recovers the equality case of Theorem 4. In the opposite direction, as $\alpha \rightarrow 0^+$ the normalizing constant $\zeta(1 + \alpha)$ diverges while $\text{Li}_{1+\alpha}(\gamma^p)$ remains bounded, so that $\kappa_p(w_\alpha) \rightarrow 0$. The modulus therefore sweeps the full admissible interval $(0, \gamma)$ as α ranges over $(0, \infty)$, so the family is maximally expressive in this sense. Any contraction rate strictly better than the one-step rate becomes attainable through an appropriate choice of tail index. This expressiveness describes the exact, untruncated operator, and it should be read alongside Remark 3, since the limit $\alpha \rightarrow 0^+$ is achieved only by shifting mass toward arbitrarily long horizons, and for every $\alpha \in (0, 1]$ the untruncated family already has infinite mean bootstrap horizon.

Ordering in the Wasserstein order p . The curve for $p = 2$ lies above the curve for $p = 1$ at every α , and the same ordering holds for the geometric reference, where 0.8433 exceeds 0.8319. This ordering is not an artifact of the numerics but a structural fact. Writing $\kappa_p(w) = (\mathbb{E}_{n \sim w}[(\gamma^n)^p])^{1/p} = \|\gamma^n\|_{L^p(w)}$ exhibits the modulus as the $L^p(w)$ -norm of the deterministic sequence $n \mapsto \gamma^n$, and L^p norms on a probability space are nondecreasing in p by the power-mean inequality. Contraction statements in higher Wasserstein orders are therefore necessarily weaker, uniformly in the weighting. The qualitative conclusions of this paper are nevertheless unaffected, because the comparative law of Theorem 6 holds separately at each fixed p .

The crossover and its interpretation. The intersection of each power-law curve with the geometric reference constitutes the practically decisive feature of the figure. Solving $\kappa_p(w_\alpha) = \kappa_p(w_\lambda)$ numerically from the closed forms of Corollary 9 gives

$$\alpha^* \approx 0.4285 \quad (p = 1), \quad \alpha^* \approx 0.3464 \quad (p = 2), \quad (35)$$

evaluated at $\gamma = 0.99$ and $\lambda = 0.95$. For $\alpha < \alpha^*$ the power-law mixture contracts strictly faster than the exponential baseline, whereas for $\alpha > \alpha^*$ it contracts strictly more slowly. This makes precise the caveat that the power-law family does *not* dominate the geometric family universally, since Theorem 6 orders moduli only under first-order stochastic dominance. A light-tailed power law such as the member at $\alpha = 0.90$ simply fails to dominate w_λ at $\lambda = 0.95$. The figure displays that failure directly, with the marker at $\alpha = 0.90$ sitting above the dashed reference at 0.9580 against 0.8319. Theorem 6 is a sufficient ordering result and not a necessary one, and this qualification matters on the favorable side of the crossover as well. At $\alpha = 0.08$, for instance, $\mathbb{P}(N \geq 2) = 1 - 1/\zeta(1.08) \approx 0.9236$ under $w_{0.08}$, against $\mathbb{P}(N \geq 2) = \lambda = 0.95$ under the geometric reference w_λ , so that $w_{0.08}$ does not first-order stochastically dominate w_λ either, since the tail inequality required by dominance already fails at $n = 2$. Theorem 6 therefore does not, by itself, explain why $\kappa_p(w_{0.08})$ falls below $\kappa_p(w_\lambda)$ in Table 1; that specific crossover is instead a direct consequence of evaluating

Table 2 Iteration count $k(\epsilon)$ from (33) to reach $\epsilon/D = 10^{-3}$ at $\gamma = 0.99$, using the moduli of Table 1. Counts are per application of $\mathcal{T}_w$; see Remark 7 on the cost of one such application.

Weight family	$p = 1$		$p = 2$	
	κ_p	$k(\epsilon)$	κ_p	$k(\epsilon)$
Point mass $w = \delta_1$ (one-step)	0.9900	688	0.9900	688
Geometric, $\lambda = 0.95$ (reference)	0.8319	38	0.8433	41
Power-law, $\alpha = 0.30$	0.7255	22	0.8147	34
Power-law, $\alpha = 0.16$ ¹	0.5097	11	0.6731	18
Power-law, $\alpha = 0.10$ ¹	0.3632	7	0.5640	13
Power-law, $\alpha = 0.08$ ¹	0.3041	6	0.5146	11

¹ Within FracTrace’s learned slow-level order range $\alpha \in [0.08, 0.16]$ [5].

and comparing the closed forms of Corollary 9 at the two members in question, exactly as (35) does below. Theorem 6 orders moduli *within* a stochastically nested family, such as the power-law family indexed by α ; it is not claimed to order two arbitrary members drawn from different families.

The location of the learned range of FracTrace relative to (35) constitutes the substantive observation. The band $[0.08, 0.16]$ lies entirely within the favorable regime, at roughly a factor of three below α^* , and the corresponding moduli, which range from 0.3041 to 0.5097 at $p = 1$, fall well below both the geometric reference and the one-step ceiling. Two qualifications nevertheless keep this observation appropriately scoped. First, FracTrace learns those orders for its *advantage* kernel, which is a non-summable lag kernel and therefore not a Bellman mixture, as established in Section 2.5. The comparison is consequently suggestive of the regime in which power-law memory is empirically useful, and it is not a direct statement about the critic of that framework, which retains the one-step backup. Second, the crossover α^* depends on both γ and the geometric reference λ , so that (35) should be read as the crossover for the specific baseline adopted in [5] and not as a universal constant. Subject to those qualifications, the figure supports a clear design conclusion. Under a heuristic transfer of the same tail-index range to a summable horizon weighting, the resulting Bellman operator would have contracted substantially faster than either the exponential compound backup or the standard one-step operator.

All quantities plotted in Figure 1 are evaluated from (22) and not by truncated summation. As noted in Remark 4, this distinction is material precisely in the heavy-tail region occupied by the shaded band, where truncation converges slowly and systematically overstates the modulus. A truncated evaluation would therefore compress the very gap that the figure is intended to display.

4.3 Iteration Complexity

Table 2 converts the moduli of Table 1 into the iteration counts (33) required in order to reduce the initial error by three orders of magnitude, corresponding to $\epsilon/D = 10^{-3}$. This conversion supplies the operational content of the contraction map, because the modulus enters the iteration count logarithmically, so that the differences visible in Figure 1 translate into order-of-magnitude differences in the number of operator applications.

The spread across families is substantial. At $p = 1$ the one-step operator requires 688 applications, the exponential baseline requires 38, and a power-law weighting at the heavy end of the learned range requires only 6. This amounts to a reduction of two orders of magnitude relative to the one-step backup and to a factor of roughly six relative to the geometric reference. The column for $p = 2$ is uniformly larger, as required by the L^p ordering discussed in Section 4.2, although it exhibits the same qualitative pattern. Consistent with Remark 7, these numbers count operator applications and do not represent a wall-clock claim.

The count of 6 operator applications carries a further caveat that is stated here explicitly, because it is otherwise easy to overstate. Table 2 is computed from the untruncated closed form (22), and by Remark 3 the power-law family at every order reported in Table 1, $\alpha \in \{0.08, 0.10, 0.16, 0.30, 0.50, 0.90\}$, has infinite mean bootstrap horizon. A literal implementation of a single application of $\mathcal{T}_w$ by sampling $n \sim w_\alpha$ and rolling out n one-step backups therefore has infinite expected cost per application. The value 6 describes the idealized operator built from the exact, untruncated weighting, and it is not directly a claim about a

Table 3 Realized contraction ratio $\bar{d}_p(\mathcal{T}_w Z_1, \mathcal{T}_w Z_2)/\bar{d}_p(Z_1, Z_2)$ on a random tabular MDP ($|S| = 6$, $N = 512$ quantile particles, $\gamma = 0.99$), against the predicted modulus for the same truncated weight ($J = 200$). Every realized ratio respects the bound of Theorem 4.

Weight family ($J = 200$)	$p = 1$		$p = 2$	
	κ_p (bound)	realized	κ_p (bound)	realized
Geometric, $\lambda = 0.95$	0.8320	0.6525	0.8433	0.6128
Power-law, $\alpha = 0.50$	0.9181	0.7647	0.9303	0.7476
Power-law, $\alpha = 0.30$	0.8743	0.7187	0.8957	0.6919
Power-law, $\alpha = 0.16$	0.8328	0.6772	0.8626	0.6445
Power-law, $\alpha = 0.08$	0.8052	0.6503	0.8404	0.6148

finite-cost implementation. Under the same truncation $J = 200$ actually used in the numerical verification of Section 4.4, the modulus at $\alpha = 0.08$, $p = 1$ rises to $\kappa_p(w_{0.08}^{200}) = 0.8052$ (Table 3), which by (33) requires 32 operator applications instead of 6, each still costing $O(J) = O(200)$ one-step backups by Remark 7. The reduction relative to the one-step operator remains substantial at this practical truncation, a factor of roughly 21 in operator applications, although it is markedly smaller than the untruncated number suggests. The two numbers answer different questions. The value 6 is the idealized-operator count, and the value 32 is the count for the finite-horizon operator that can actually be implemented at $J = 200$.

4.4 Verification on a Tabular MDP

The moduli reported above are analytic consequences of Corollary 9. It remains to verify numerically that they govern the behavior of the operator as constructed, and to quantify how conservative the worst-case bound proves to be on a non-degenerate problem. Two complementary experiments are reported for that purpose, and both are reproducible from the simulation files described in the Data Availability Statement.

Experiment 1 concerning sharpness. The extremal construction of Proposition 11 is instantiated directly, with a single state, zero reward and a deterministic self-loop, together with $Z_1 = \delta_0$ and $Z_2 = \delta_c$, and the realized ratio $\bar{d}_p(\mathcal{T}_w Z_1, \mathcal{T}_w Z_2)/\bar{d}_p(Z_1, Z_2)$ is measured. Across all tested orders $\alpha \in \{0.08, 0.16, 0.30, 0.50\}$ and both values $p \in \{1, 2\}$, the measured ratio agrees with the predicted $\kappa_p(w)$ to machine precision, with a maximum absolute deviation of 1.1×10^{-16} , which confirms attainment.

Experiment 2 on a randomly generated instance. A randomly parameterized tabular MDP is generated with $|S| = 6$ states, a dense stochastic transition kernel and a reward table $r(s, a)$ drawn once from a Gaussian distribution and held fixed thereafter, and each return distribution is represented by $N = 512$ quantile particles. The reward table is sampled once to construct a specific finite MDP, and it is not resampled at every transition, so the realized instance satisfies the bounded-reward condition trivially, with $R_{\max}$ given by the largest magnitude in the fixed table. This construction is distinct from an unbounded per-transition Gaussian reward process, which the framework of Section 2 does not cover. Starting from two unrelated initial return-distribution functions, the operator $\mathcal{T}_w$ is applied once and the realized ratio is recorded. The $N = 512$ quantile-particle representation is the standard practical implementation of a return-distribution function in distributional RL [13], so the operator applied here is the quantile-projected operator, and not a re-verification of the exact operator on the infinite-dimensional space $\mathcal{P}_p(\mathbb{R})$. That exact operator is already verified, without any projection, by the degenerate closed-form construction of Experiment 1 and by the closed forms of Table 1. Experiment 2 should accordingly be read as an approximation study of the operator as it would actually be implemented, with the particle count N held fixed at a representative value and not varied. Table 3 reports the outcome. Because the operator is implemented with the *truncated* weight w_α^J at $J = 200$, the predicted moduli in that table are the corresponding truncated values from Corollary 9(ii), and they therefore differ, substantially so at small α , from the untruncated values of Table 1. This discrepancy is exactly the slow convergence flagged in Remark 4, and the comparison remains internally consistent because prediction and measurement employ the same weight.

Two conclusions follow from Table 3. First, every realized ratio lies strictly below the corresponding bound, as Theorem 4 requires, and no violation was observed in this instance. Second, the gap between the

realized ratio and the bound remains consistent across the five families evaluated on this instance, spanning roughly 0.15 to 0.23 in absolute terms, so that the worst-case modulus is conservative on this randomly generated MDP. This conservatism is not misleading, since it preserves the *ordering* of the families, and that ordering is precisely the property asserted by the comparative law. These observations are drawn from a single random instance, and they are reported as a consistency check, and not as an average-case or typical-case claim. A broader study across multiple random MDPs, including the distribution of the gap between the realized contraction rate and the theoretical bound, is left for future work. Such a study would not affect the bound in Theorem 4, which holds for every admissible MDP regardless of how it is generated. Together with Experiment 1, these results show that the bound is exact in the worst case, as established by Proposition 11, but conservative for the MDP considered here. Importantly, the predicted ordering of the contraction rates is preserved in both cases.

5 Discussion and Conclusion

This paper established the contraction and fixed-point properties of multi-horizon distributional Bellman operators under normalized horizon weights. For any normalized weight distribution w , the operator $\mathcal{T}_w$ is a contraction in the supremal p -Wasserstein metric with modulus $\kappa_p(w) = \left(\sum_{n \geq 1} w(n) \gamma^{np}\right)^{1/p} \leq \gamma$, while preserving the policy return distribution Z^π as its unique fixed point. Replacing the standard one-step backup with a normalized power-law mixture therefore does not weaken the contraction guarantee of distributional policy evaluation.

The analysis further characterized how the choice of horizon weights governs the contraction rate. Under first-order stochastic dominance, placing more weight on longer horizons yields a smaller contraction modulus, strictly so within the power-law family. Closed-form expressions for the power-law and geometric families make this ordering directly comparable. The same guarantees extend to state-dependent horizon weights without introducing an additional renormalization factor, and the derived modulus is in fact the exact worst-case Lipschitz constant over the considered MDP class, and not only an upper bound. The numerical results corroborate these theoretical findings and illustrate the resulting differences in contraction rate across horizon-weight families.

These results also clarify the relationship to FracTrace. The normalized power-law horizon mixture studied here is fundamentally different from the non-summable lag weighting used for policy-side advantages in that framework, since the latter does not define a probability distribution over bootstrap horizons and therefore does not form the Bellman mixture considered in this work. Contraction and fixed-point preservation accordingly do not, by themselves, require power-law memory to be confined to the policy side.

The analysis carries several limitations. The comparison between power-law and geometric weights is not universal, and cross-family ordering under first-order stochastic dominance applies only when the required dominance condition holds. For the power-law weights considered in the numerical study, the observed crossover with geometric weighting follows from direct comparison of the closed-form moduli, and not from stochastic dominance. The present results, moreover, concern the exact population-level Bellman operator and do not directly extend to off-policy, importance-weighted updates or to stochastic approximation.

A natural direction for future work is to extend the analysis to off-policy and importance-weighted settings, to investigate more general state-dependent horizon weightings, and to examine the trade-off between contraction rate and computational cost. Although longer-horizon mixtures can require fewer operator iterations, each multi-horizon update carries a higher computational cost, and whether the theoretical reduction in iteration count translates into faster learning on practical reinforcement-learning problems remains an open and important question.

Author contributions

Y.M.: Conceptualization, Supervision, Methodology. R.M.: Methodology, Formal analysis, Investigation, Writing – Original Draft. A.M.: Methodology, Formal analysis, Investigation, Writing – Original Draft.

M.T.: Writing – Review & Editing. I.B.K.: Writing – Review & Editing. A.F.: Writing – Review & Editing. U.C.: Supervision, Writing – Review & Editing.

Data availability

This manuscript does not report data generation or analysis.

Additional information

Competing interests: The authors declare no competing interests.

References

- [1] Pignatelli, E., Ferret, J., Geist, M., Mesnard, T., Hasselt, H., Pietquin, O., Toni, L.: A survey of temporal credit assignment in deep reinforcement learning. arXiv preprint arXiv:2312.01072 (2023)
- [2] Barto, A.G., Sutton, R.S., Anderson, C.W.: Looking back on the actor–critic architecture. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* **51**(1), 40–50 (2020)
- [3] Pawar, U., Telangi, K.: Fractional policy gradients: Reinforcement learning with long-term memory. arXiv preprint arXiv:2507.00073 (2025)
- [4] Gorenflo, R., Kilbas, A.A., Mainardi, F., Rogosin, S.V.: Mittag-Leffler Functions, Related Topics and Applications, 2nd edn. Springer Monographs in Mathematics. Springer, Berlin, Heidelberg (2020). <https://doi.org/10.1007/978-3-662-61550-8>
- [5] Mousavi, Y., Mousavi, R., Mousavi, A., Tavasoli, M., Kucukdemiral, I.B., Fekih, A., Cali, U.: Frac-Trace: Long-horizon credit assignment in deep reinforcement learning via fractional eligibility traces. *Neurocomputing* **706**, 134924 (2026) <https://doi.org/10.1016/j.neucom.2026.134924>
- [6] Bellemare, M.G., Dabney, W., Munos, R.: A distributional perspective on reinforcement learning. In: Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 449–458. PMLR, ??? (2017). arXiv:1707.06887. <https://proceedings.mlr.press/v70/bellemare17a.html>
- [7] Tang, Y., Munos, R., Rowland, M., Avila Pires, B., Dabney, W., Bellemare, M.G.: The nature of temporal difference errors in multi-step distributional reinforcement learning. In: Advances in Neural Information Processing Systems, vol. 35, pp. 30265–30276 (2022)
- [8] Daley, B., White, M., Machado, M.C.: Averaging n -step returns reduces variance in reinforcement learning. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 9904–9930. PMLR, ??? (2024). arXiv:2402.03903
- [9] Amortila, P., Precup, D., Panangaden, P., Bellemare, M.G.: A distributional analysis of sampling-based reinforcement learning algorithms. In: Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 108, pp. 4357–4366. PMLR, ??? (2020). arXiv:2003.12239. <https://proceedings.mlr.press/v108/amortila20a.html>
- [10] Nam, D.W., Kim, Y., Park, C.Y.: GMAC: A distributional perspective on actor-critic framework. In: Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 7927–7936. PMLR, ??? (2021). arXiv:2105.11366. <https://proceedings.mlr.press/v139/nam21a.html>

- [11] Shaked, M., Shanthikumar, J.G.: Stochastic Orders. Springer Series in Statistics. Springer, New York (2007). <https://doi.org/10.1007/978-0-387-34675-5>
- [12] Villani, C.: Optimal Transport: Old and New. Grundlehren der mathematischen Wissenschaften, vol. 338. Springer, Berlin/Heidelberg, Germany (2009)
- [13] Dabney, W., Rowland, M., Bellemare, M.G., Munos, R.: Distributional reinforcement learning with quantile regression. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, pp. 2892–2901 (2018). arXiv:1710.10044