Quotient-Categorical Representations for Bellman-Compatible Average-Reward Distributional Reinforcement Learning

Ege C. Kaya, Aliasghar Pourghani, Vijay Gupta, Abolfazl Hashemi
Elmore Family School of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47906
{kayae,apourgha,gupta869,abolfazl}@purdue.edu

Abstract

Average-reward reinforcement learning requires estimating the gain and a state-dependent bias, which solves the Poisson equation only up to an additive constant. Distributional analogues of the bias inherit this same ambiguity. We introduce a quotient-space formulation in which state-indexed laws are identified up to a common translation, together with a categorical parameterization that respects this symmetry. On this space, we define a projected average-reward distributional operator and show that it is well-defined, non-expansive in a coordinate Cramér metric, and admits fixed points. We interpret the projected fixed points through geometrically-stopped differential returns, whose categorical operators are contractive and whose fixed points converge in the Abel sense to a canonical fixed point of the projected undiscounted operator. We show that the means of the projected fixed points match the scalar bias up to an error caused by endpoint clipping of the categorical projections, and we bound this error. Finally, we establish almost-sure convergence and finite-iteration residual bounds for temporal-difference recursions with gain-centered rewards under i.i.d. and Markovian sampling. When the gain is unknown, we couple the recursion with an online gain estimator and prove the same guarantees under Markovian sampling, with a bound on the gain error.

1 Introduction

Distributional reinforcement learning (DRL) is well developed in the discounted setting, where the standard contraction arguments and Bellman-compatible return distributions support both analysis and algorithms [3, 4, 38, 17, 16]. The average-reward setting is structurally different. When the discounting is removed, the contraction geometry is lost, and fixed-policy evaluation is no longer described by a discounted return. Instead, average-reward evaluation revolves around two objects: the *gain*, which is the long-run average reward, and the *bias*, which solves the average-reward analogue of the Bellman equation, namely the Poisson equation. The bias of a state is the expected *differential return* from that state, that is, the cumulative sum of gain-centered rewards. The bias distinguishes states and policies with the same gain, captures transient differential performance, and underlies average-reward evaluation and policy improvement [34, 44, 28]. However, it is defined only up to an additive constant. This creates an obstruction for a distributional analogue, since a state-indexed family of laws on $\mathbb{R}$ has a fixed origin, yet translating these laws by the same constant should not change the object we learn. A second obstruction is that, unlike the discounted return, the differential return generally has no limiting law, since its fluctuations grow with the horizon due to the lack of discounting and its dampening effect. A distributional analogue of the bias must therefore rest on a different object, such as the geometrically stopped return.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

We therefore model state-indexed law families modulo common translations, preserving the additive-constant symmetry of the Poisson equation in a finite-dimensional categorical representation. The learned object is a Bellman-consistent family of state-indexed laws whose means recover the scalar bias up to clipping. Each law arises as a limit of geometrically stopped differential-return laws and can therefore be read as describing centered rewards accumulated over a random horizon. This reading gives the laws meaning beyond their means, for instance through tail-risk measures such as CVaR, for which we provide approximation bounds. Our objective is complementary to the average-reward distributional criterion studied by Rojas and Lee [37], which refines the average-reward criterion itself through limiting per-step reward distributions. We instead refine the bias, the Bellman-compatible object that records state-dependent differential performance. Thus, our goal is not merely to describe long-run reward statistics, but to construct a distributional analogue of the scalar bias that can be used in average-reward dynamic-programming and temporal-difference (TD) learning arguments.

Part of the challenge is algorithmic. The projected maps we obtain are non-expansive rather than contractive, so discounted DRL contraction arguments do not apply. We show that the sampled categorical updates can be written as TD / stochastic approximation (SA) recursions in the classical sense [36, 43, 46, 8, 9], then prove that their mean-field maps are non-expansive and share fixed points with the projected differential-return operator. This connects our recursion to recent SA results for non-expansive maps [10, 7].

Contributions. Our main contributions are as follows.

1. We define exact and quotient-categorical projected operators on the quotient space of state-indexed law families modulo common translation, prove that both are well-defined and non-expansive, and show that the latter admits a fixed point.
2. We interpret projected fixed points through geometrically stopped returns, whose projected operators are contractive and whose fixed points converge as $\gamma \uparrow 1$ to a distinguished fixed point of the projected undiscounted operator. We show that the means of the projected fixed points match the scalar bias up to an explicit clipping error due to categorical projections.
3. We prove almost-sure (a.s.) convergence and finite-iteration residual bounds for categorical TD recursions with gain-centered rewards under i.i.d. and Markovian sampling. When the gain is unknown, a coupled recursion with an online gain estimator attains the same Markovian guarantees, together with a bound on the gain error.

2 Related work

DRL is built around Bellman-compatible return distributions in discounted MDPs, with categorical and quantile parameterizations providing the dominant practical representations [3, 38, 17, 16]. Early theory studied the geometry, statistics, and relative advantages of categorical and quantile DRL [38, 27, 39]. Recent work has made many advancements by sharpening the statistical case for learning return distributions [49, 50], providing non-asymptotic efficiency of distributional TD [32, 33], near-minimax guarantees in the generative-model regime [40], establishing multivariate distributional dynamic programming [52], offline distributional policy evaluation guarantees [53], and convergence analyses for KL-trained categorical updates [22]. In the average-reward setting, however, most work remains scalar, from early R-learning and average-reward foundations to modern off-policy, robust, and function-approximation results [41, 1, 28, 48, 51, 47, 54]. The closest prior average-reward DRL work is [37], where the authors study long-run per-step reward and differential return distributions via quantile representations. As opposed to these works, we seek a Bellman-compatible distributional analogue of the bias, which makes the quotient symmetry induced by additive-constant non-uniqueness central rather than incidental. On the algorithmic side, our finite-iteration analysis draws on fixed-point methods for non-expansive maps [26, 29, 25, 11, 10, 7] and complements classical TD / SA analyses for contractive or linear settings [19, 45, 6, 18, 42, 31, 35, 20, 13–15].

3 Preliminaries

We work with a finite average-reward MDP [34] $(\mathcal{S}, \mathcal{A}, R, P)$, where $\mathcal{S} = \{1, \dots, m\}$ is the finite set of states and $\mathcal{A}$ is the finite set of actions, $R : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the deterministic reward function and $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel. A policy π is fixed throughout, and we focus on

its evaluation. With π fixed, the induced process is a Markov reward process with transition matrix P^π, where $P^\pi_{ij} = \sum_{a\in\mathcal{A}} P(j \mid i,a)\,\pi(a \mid i)$ [44]. Throughout the paper, we assume that the induced Markov chain $\{S_t\}$ is irreducible and aperiodic, and write μ for its unique stationary distribution. We denote by R_{ij} the finite-valued random one-step reward conditioned on $(S_t = i, S_{t+1} = j)$. We define the expected one-step reward vector by $r^\pi_i := \sum_{j\in\mathcal{S}} P^\pi_{ij}\mathbb{E}[R_{ij}]$ for all $i \in \mathcal{S}$. The (scalar) *gain* or *average reward* is then $\bar{r}^\pi := \sum_{i\in\mathcal{S}} \mu_i r^\pi_i$.

The *bias* or *relative value function* [34] $v^\pi \in \mathbb{R}^m$ is any solution of the Poisson equation [30]

$$v^\pi = r^\pi - \bar{r}^\pi \mathbf{1} + P^\pi v^\pi. \tag{1}$$

The Poisson equation is, in some sense, the analogue of the Bellman equation [5] in the average-reward setting, where the discount factor has been set to 1 and the rewards have been centered. It can equivalently be written as $(I - P^\pi)v^\pi = r^\pi - \bar{r}^\pi\mathbf{1}$. Since the null space of $(I - P^\pi)$ contains any constant vector $c\mathbf{1} \in \mathbb{R}^m$, the solutions of (1) are unique up to addition of a common constant, as the standing irreducibility assumption rules out any other null directions. This translation symmetry motivates the quotient-categorical formulation constructed in the rest of the work.

4 Quotient-categorical bias estimation

We begin our inquiry by seeking a distributional object which obeys two properties of the scalar bias: satisfying the gain-centered Poisson equation, and being defined only up to a common additive constant. Let $\mathcal{F}^\mathcal{S} := \{\eta = (\eta_i)_{i\in\mathcal{S}} : \eta_i \in \mathcal{P}_2(\mathbb{R})\}$ denote the set of state-indexed collections of probability laws on $\mathbb{R}$ with finite second moment. We compare elements of $\mathcal{F}^\mathcal{S}$ with the Cramér metric. For $\xi,\zeta \in \mathcal{P}_2(\mathbb{R})$ with CDFs F_ξ, F_ζ, let $\ell_\mathrm{C}(\xi,\zeta) := (\int_\mathbb{R}(F_\xi(x) - F_\zeta(x))^2\,\mathrm{d}x)^{1/2}$, and for $\eta,\zeta \in \mathcal{F}^\mathcal{S}$, let $\ell_{\mathrm{C},\infty}(\eta,\zeta) := \max_{i\in\mathcal{S}} \ell_\mathrm{C}(\eta_i,\zeta_i)$. For $c \in \mathbb{R}$, let $\tau_c(x) := x + c$, and define *common-translation* on $\mathcal{F}^\mathcal{S}$ by $((\tau_c)_\# \eta)_i := (\tau_c)_\#\eta_i$, where $\#$ denotes pushforward. We also define the equivalence relation $\approx$ on any pair $\eta,\zeta \in \mathcal{F}^\mathcal{S}$ by $\eta \approx \zeta \iff \exists c \in \mathbb{R}$ such that $\zeta = (\tau_c)_\#\eta$, and denote the quotient space by $\mathcal{F}^\mathcal{S}/\approx$. Let us write $B_{ij} := R_{ij} - \bar{r}^\pi$ for the centered one-step reward, and $\nu_{ij} := \mathrm{Law}(B_{ij})$. We now define the average-reward distributional operator.

Definition 1. *The* average-reward distributional operator $\mathcal{T} : \mathcal{F}^\mathcal{S} \to \mathcal{F}^\mathcal{S}$ *is defined componentwise:*

$$(\mathcal{T}\eta)_i := \sum_{j\in\mathcal{S}} P^\pi_{ij}\,(\nu_{ij} * \eta_j), \quad \textit{for all } i \in \mathcal{S}. \tag{2}$$

Proposition 2. *For every $\eta \in \mathcal{F}^\mathcal{S}$ and $c \in \mathbb{R}$,*

$$\mathcal{T}\big((\tau_c)_\#\eta\big) = (\tau_c)_\#\mathcal{T}\eta. \tag{3}$$

Consequently, $\mathcal{T}$ induces a well-defined operator $\overline{\mathcal{T}}([\eta]) := [\mathcal{T}\eta]$ on the quotient space $\mathcal{F}^\mathcal{S}/\approx$.

The quotient formulation is instructive in seeing how the gain is not part of the quotient target itself. If we center rewards by an arbitrary constant $g \in \mathbb{R}$ instead of $\bar{r}^\pi$, the resulting state-indexed collection of laws differs only by a common translation. The following proposition formalizes this.

Proposition 3. *Let $g \in \mathbb{R}$. Define the* g-centered average-reward distributional operator $(\mathcal{T}_g\eta)_i := \sum_{j\in\mathcal{S}} P^\pi_{ij}(\nu^{(g)}_{ij} * \eta_j)$, *where $\nu^{(g)}_{ij} := \mathrm{Law}(R_{ij} - g)$, so that $\mathcal{T} = \mathcal{T}_{\bar{r}^\pi}$. Then, for any $g, g' \in \mathbb{R}$ and $\eta \in \mathcal{F}^\mathcal{S}$, $[\mathcal{T}_g\eta] = [\mathcal{T}_{g'}\eta]$ in $\mathcal{F}^\mathcal{S}/\approx$.*

Guided by these observations, we introduce a categorical representation class which respects the same common-translation symmetry. Let $\Theta = \{\theta_1 < \cdots < \theta_d\} \subset \mathbb{R}$ be an ordered support with constant stride $\varsigma := \theta_{k+1} - \theta_k > 0$ for $k = 1,\ldots,d-1$.[1] Let Δ_d denote the probability simplex in $\mathbb{R}^d$. For $p \in \Delta_d^\mathcal{S}$ and $c \in \mathbb{R}$, define the representative family of laws

$$\eta^{p,c}_i := \sum_{k=1}^{d} p_{i,k}\,\delta_{\theta_k + c}, \quad \text{for all } i \in \mathcal{S}. \tag{4}$$

For each $c \in \mathbb{R}$, define the c-shifted support $\Theta_c := \{\theta + c : \theta \in \Theta\}$. The state-indexed collection of coefficients p therefore uniquely characterizes the following equivalence class of law–grid pairs:

$$\big[(\eta^{p,c}, \Theta_c)\big] := \big\{((\tau_a)_\#\eta^{p,c}, \Theta_{c+a}) : a \in \mathbb{R}\big\}. \tag{5}$$

[1] It is straightforward to use state-dependent supports and non-uniform stride, we forgo these for simplicity of exposition.

In each pair within the equivalence class, the grid is translated together with the state-indexed laws, hence the coefficients p encode the shape of each law but not its location on the real line. We abbreviate this class by $[\eta^p]$ when the accompanying grid is clear. We call this a *quotient-categorical* representation: categorical because each state law is represented on the finite grid Θ, and quotient because the common shift c is not part of the state, but only a choice of representative.

The categorical projection of a point mass $x \in [\theta_k + c, \theta_{k+1} + c]$ is

$$\Pi_{\mathrm{C}}^{\Theta_c} \delta_x = \frac{\theta_{k+1} + c - x}{\varsigma} \delta_{\theta_k + c} + \frac{x - (\theta_k + c)}{\varsigma} \delta_{\theta_{k+1} + c}. \tag{6}$$

Mass below $\theta_1 + c$ or above $\theta_d + c$ is clipped at the nearest endpoint, and the projection extends linearly to probability laws. This is the standard categorical projection [3] onto Θ_c, applied over all states. We will show that this projection is shift-equivariant: first shifting a distribution by c and then projecting onto the c-shifted categorical support is equivalent to projecting it onto the unshifted support and then shifting by c. Consequently, translating a law together with its grid leaves the coefficients of its projection unchanged.

Lemma 4 (Projection shift-equivariance)**.** *For every $\eta \in \mathcal{P}_2(\mathbb{R})$ and $c \in \mathbb{R}$,*

$$\Pi_{\mathrm{C}}^{\Theta_c}\big((\tau_c)_{\#}\eta\big) = (\tau_c)_{\#}(\Pi_{\mathrm{C}}^{\Theta}\eta). \tag{7}$$

This same fact is illustrated in Figure 1. Every representative $\eta^{p,c}$ is supported on the c-shifted grid Θ_c. For two such representatives with the same shift c, the CDF difference is constant between consecutive atoms, so the Cramér integral reduces to a finite sum of cumulative masses. For $p \in \Delta_d$, let $F_p(\theta_k) := \sum_{j=1}^{k} p_j$ for $k = 1, \dots, d-1$. We define, for $p, q \in \Delta_d$, the coordinate Cramér metric by

$$\ell_{\mathrm{C}}^{\Theta}(p, q)^2 := \varsigma \sum_{k=1}^{d-1} \big(F_p(\theta_k) - F_q(\theta_k)\big)^2. \tag{8}$$

The statewise extension of this is the so-called supremum-Cramér metric [3, 4], where for $p, q \in \Delta_d^{\mathcal{S}}$,

$$\ell_{\mathrm{C},\infty}^{\Theta}(p, q) := \max_{i \in \mathcal{S}} \ell_{\mathrm{C}}^{\Theta}(p_i, q_i). \tag{9}$$

The next proposition shows that these coordinate metrics coincide exactly with ℓ_{C} and $\ell_{\mathrm{C},\infty}$ on the corresponding representative families.

Proposition 5. *For every $p, q \in \Delta_d^{\mathcal{S}}$ and $c \in \mathbb{R}$,*

$$\ell_{\mathrm{C}}(\eta_i^{p,c}, \eta_i^{q,c}) = \ell_{\mathrm{C}}^{\Theta}(p_i, q_i), \quad \textit{for all } i \in \mathcal{S}. \tag{10}$$

Consequently,

$$\ell_{\mathrm{C},\infty}(\eta^{p,c}, \eta^{q,c}) = \ell_{\mathrm{C},\infty}^{\Theta}(p, q), \quad \textit{for all } c \in \mathbb{R}. \tag{11}$$

Proposition 5 shows that the representative families $\eta^{p,c}$ faithfully encode the relevant Cramér geometry in the coefficient vectors $p \in \Delta_d^{\mathcal{S}}$, regardless of the shift c, as long as the same c is chosen for both representatives. We now use this representation to define the projected analogue of the g-centered operator $\mathcal{T}_g$ on categorical quotient coordinates. Proposition 3 previously showed that, at the quotient level, $\mathcal{T}_g$ does not depend on the choice of g: changing g only changes the representative of the output, but lands in the same equivalence class. After categorical projection, however, the representative matters, since the projection is performed relative to a fixed shifted support Θ_c.

Definition 6. *Let $g \in \mathbb{R}$ and $p \in \Delta_d^{\mathcal{S}}$ and choose $c \in \mathbb{R}$ arbitrarily. Applying $\mathcal{T}_g$ to the representative family $\eta^{p,c}$ and projecting each component back onto Θ_c yields another categorical family on Θ_c. We define $\mathcal{G}_g(p) = q \in \Delta_d^{\mathcal{S}}$ to be the unique element that satisfies*

$$\Pi_{\mathrm{C}}^{\Theta_c}(\mathcal{T}_g \eta^{p,c})_i = \eta_i^{q,c}, \quad \textit{for all } i \in \mathcal{S}. \tag{12}$$

We note that this definition is independent of the choice of the representative shift c. Indeed, since by definition $\eta^{p,c'} = (\tau_{c'-c})_{\#}\eta^{p,c}$, the same argument as in Proposition 2 gives

$$\mathcal{T}_g \eta^{p,c'} = (\tau_{c'-c})_{\#} \mathcal{T}_g \eta^{p,c}. \tag{13}$$

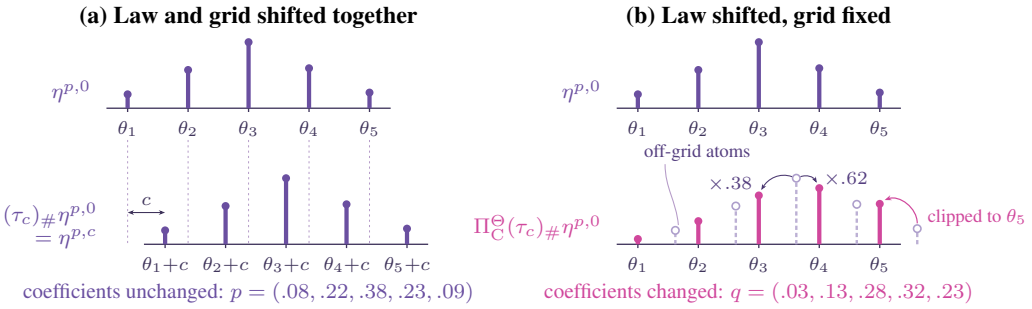


Figure 1: **(a)** Shifting a law together with its grid leaves its coefficients p unchanged (Lemma 4). **(b)** Shifting the law against a fixed grid moves its atoms off the grid. Projection then splits each atom between its two neighbors and clips mass beyond the last atom, giving $q \neq p$.

The application of Lemma 4 shows that projection onto $\Theta_{c'}$ or onto Θ_c produces categorical families with the same coefficient vector q. Therefore, $\mathcal{G}_g$ is a well-defined operator on $\Delta_d^{\mathcal{S}}$ for any $g \in \mathbb{R}$. Unlike $\mathcal{T}$, which acts on $\mathcal{F}^{\mathcal{S}}/\approx$, the projected operator acts on coefficients, because translates of the same law projected onto a fixed grid generally have different coefficients (Figure 1(b)). In particular, changing g translates the backup $\mathcal{T}_g \eta^{p,c}$ against the fixed grid Θ_c, so $\mathcal{G}_g$ depends on g even though $[\mathcal{T}_g \eta]$ does not (Proposition 3). We therefore single out $\mathcal{G} := \mathcal{G}_{\bar{r}^\pi}$ as the "correctly-centered" projected operator. As in discounted categorical TD learning [3], $\mathcal{G}$ is a projected Bellman backup, here non-expansive rather than contractive.

We now state the main theorem of this section, establishing the non-expansiveness of $\mathcal{G}_g$.

Theorem 7 (Non-expansiveness of $\mathcal{G}_g$)**.** *For every $g \in \mathbb{R}$, and every $p, q \in \Delta_d^{\mathcal{S}}$,*

$$\ell^{\Theta}_{\mathrm{C},\infty}\big(\mathcal{G}_g(p), \mathcal{G}_g(q)\big) \leq \ell^{\Theta}_{\mathrm{C},\infty}(p, q). \tag{14}$$

In particular, the correctly-centered operator $\mathcal{G}$ is non-expansive.

Proof sketch. Fix a common representative shift c and compare the representative families $\eta^{p,c}$ and $\eta^{q,c}$. For each state i, the ith component of $\mathcal{T}_g$ is a P^π_{ij}-weighted average of convolved laws $\nu^{(g)}_{ij} * \eta_j$. Convolution with a fixed law is non-expansive in the Cramér metric, and the subsequent P^π_{ij}-averaging is also non-expansive. Thus $\mathcal{T}_g$ is non-expansive in $\ell_{\mathrm{C},\infty}$. Composing with the standard non-expansive categorical projection and then using Proposition 5 yields the claim. □

Although the exact operator $\mathcal{T}_g$ need not have a fixed point in $\mathcal{F}^{\mathcal{S}}$, the projected operator $\mathcal{G}_g$ is a continuous map on the compact set $\Delta_d^{\mathcal{S}}$ and therefore always has one.

Corollary 8. *For every $g \in \mathbb{R}$, $\mathrm{fix}(\mathcal{G}_g)$ is nonempty. In particular, $\mathcal{G}$ has a fixed point.*

The full proofs of Theorem 7 and Corollary 8 are given in Appendix A.

4.1 Interpreting the projected fixed point

$\mathcal{T}_g$ need not have a fixed point, but we can still interpret the fixed points of $\mathcal{G}$ through differential returns stopped at a random time. Fix $\gamma \in (0, 1)$ and let N_γ be independent of the chain with $\mathbb{P}(N_\gamma = n) = (1-\gamma)\gamma^{n-1}$ for $n \geq 1$. For $S_0 = i$, let R_t denote the reward on the transition from S_t to S_{t+1}, let the geometrically stopped differential return be

$$Y_i^\gamma := \sum_{t=0}^{N_\gamma - 1} (R_t - \bar{r}^\pi), \tag{15}$$

and let $\eta_i^\gamma := \mathrm{Law}(Y_i^\gamma)$. Without loss of generality, we assume $0 \in \Theta$.[2] The corresponding exact and projected operators are

$$(\mathcal{T}_\gamma \eta)_i := \sum_{j \in \mathcal{S}} P^\pi_{ij}\, \nu_{ij} * \big((1-\gamma)\delta_0 + \gamma \eta_j\big), \qquad \mathcal{G}_\gamma(p) := (1-\gamma)\mathcal{G}(p^0) + \gamma \mathcal{G}(p), \tag{16}$$

[2] Since $\mathcal{G}$ does not depend on the representative shift c, any atom $\theta_\circ \in \Theta$ can serve as the origin by taking $c = -\theta_\circ$.

where $p^0 \in \Delta_d^{\mathcal{S}}$ puts unit mass on the atom $0 \in \Theta$ in every state. After each centered reward, both operators stop with probability $1-\gamma$ and continue with probability γ. Finally, let $v^\circ := \sum_{t=0}^{\infty}(P^\pi)^t\big(r^\pi - \bar{r}^\pi \mathbf{1}\big)$ denote the solution of the Poisson equation with $\mu^\top v^\circ = 0$.

Theorem 9 (Stopped returns and the Abel limit)**.** *Let $\gamma \in (0,1)$. The operators $\mathcal{T}_\gamma$ and $\mathcal{G}_\gamma$ are γ-contractive in $\ell_{\mathrm{C},\infty}$ and $\ell^{\Theta}_{\mathrm{C},\infty}$, respectively. The fixed point of $\mathcal{T}_\gamma$ is η^γ, and $\mathbb{E}[Y^\gamma] \to v^\circ$ as $\gamma \uparrow 1$. The fixed point p_γ of $\mathcal{G}_\gamma$ satisfies*

$$p_\gamma \to p^\dagger := \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \mathcal{G}^n(p^0) \in \mathrm{fix}(\mathcal{G}) \quad \textit{as } \gamma \uparrow 1. \tag{17}$$

For $\gamma < 1$, the projected fixed point p_γ therefore approximates the law of a randomly stopped differential return, and Proposition 25 bounds this approximation error in the Cramér metric, W_1, and CVaR. As $\gamma \uparrow 1$, these approximations select the fixed point $p^\dagger$ of $\mathcal{G}$. Theorem 24 also gives a closed form for p_γ and a criterion for the uniqueness of the fixed point of $\mathcal{G}$.

We next relate the means of the projected fixed points to the scalar bias. For the rest of this section, let $\Theta = \{-M, -M+\varsigma, \dots, M\}$ with $M \geq 1$. For $p \in \Delta_d^{\mathcal{S}}$, let $m(p)$ have entries $m_i(p) := \sum_{k=1}^{d} p_{i,k}\theta_k$, the mean of $\eta_i^{p,0}$. Categorical projection preserves the mean of every law supported in $[-M, M]$, so it can change means only through clipping at the endpoints. With $[x]_{-M}^{M} := \max\{-M, \min\{x, M\}\}$, this change is measured by the clipping error

$$\beta_i(p) := \sum_{j\in\mathcal{S}} P_{ij}^\pi \mathbb{E}\big[[B_{ij}+X_j]_{-M}^{M} - (B_{ij}+X_j)\big], \qquad X_j \sim \eta^{p_j,0}, \tag{18}$$

which vanishes when no mass is clipped.

Proposition 10 (Mean recovery under clipping)**.** *Every $p^\star \in \mathrm{fix}(\mathcal{G})$ satisfies*

$$(I - P^\pi)m(p^\star) = r^\pi - \bar{r}^\pi \mathbf{1} + \beta(p^\star). \tag{19}$$

Moreover, there are constants C_1, C_2 that depend on the Markov reward process, but not on the grid, such that every $p \in \Delta_d^{\mathcal{S}}$ satisfies

$$\inf_{c\in\mathbb{R}} \|m(p) - v^\pi - c\mathbf{1}\|_\infty \leq C_1\big(M^{-1} + d^{-1}\big) + C_2\sqrt{M}\,\ell^{\Theta}_{\mathrm{C},\infty}\big(p, \mathcal{G}(p)\big). \tag{20}$$

For a fixed point, the residual term vanishes and the mean error is $O(M^{-1} + d^{-1})$. The bound also turns the residual guarantees of Sections 5 and 6 into guarantees for the scalar bias. Appendix B provides the full statements and proofs, including the explicit constants for C_1 and C_2 in Proposition 26.

5 A centered stochastic approximation recursion

We now turn from the exact projected backup $\mathcal{G}$, a non-expansive map with at least one fixed point, to an SA thereof, based on sampled transitions. Customarily, SA in RL comes as a TD update, where we observe only one transition at a time and asynchronously modify only one state block.

A one-state sample is defined to be a tuple $y = (s, b, s') \in \mathcal{Y} := \mathcal{S} \times [-1, 1] \times \mathcal{S}$.[3] For $p \in \Delta_d^{\mathcal{S}}$, we define $H(p, y) = q \in \Delta_d^{\mathcal{S}}$, the one-sample backup map, by

$$q_u = \begin{cases} p_u, & u \neq s, \\ \tilde{q}, & u = s, \end{cases} \tag{21}$$

where $\tilde{q} \in \Delta_d$ is the unique coefficient vector satisfying $\Pi_{\mathrm{C}}^{\Theta}\big((\tau_b)_{\#}\eta^{p_{s'},0}\big) = \eta^{\tilde{q},0}$. The asynchronous recursion is thus the well-known stochastic Krasnosel'skiĭ–Mann (SKM) iteration [26, 29]

$$p_{k+1} = p_k + \alpha_k\big(H(p_k, Y_k) - p_k\big), \quad k \geq 0. \tag{22}$$

The next proposition rewrites the g-centered projected operator $\mathcal{G}_g$ blockwise in categorical coefficients, which is the form needed to derive asynchronous SA updates. Before the statement, we introduce a slight abuse of notation. For $u \in \Delta_d$ and $c \in \mathbb{R}$, we write $\eta^{u,c} := \sum_{k=1}^{d} u_k \delta_{\theta_k + c}$. In particular, $(\eta^{p,c})_i = \eta^{p_i,c}$ for $p \in \Delta_d^{\mathcal{S}}$.

[3] Note that rewards lying in $[0, 1]$ implies that centered rewards lie in $[-1, 1]$.

Proposition 11. *Let $g \in \mathbb{R}$, $p \in \Delta_d^{\mathcal{S}}$, and $i \in \mathcal{S}$. The ith block of $\mathcal{G}_g(p)$ is the unique vector $q_i \in \Delta_d$ such that*

$$\eta^{q_i,0} = \Pi_{\mathrm{C}}^{\Theta}\Big(\sum_{j\in\mathcal{S}} P_{ij}^{\pi}(\nu_{ij}^{(g)} * \eta^{p_j,0})\Big). \tag{23}$$

The next lemma simply states that the one-sample backup map is also a non-expansion.

Lemma 12. *For every $y \in \mathcal{Y}$ and $p, q \in \Delta_d^{\mathcal{S}}$,*

$$\ell_{\mathrm{C},\infty}^{\Theta}\big(H(p,y), H(q,y)\big) \le \ell_{\mathrm{C},\infty}^{\Theta}(p,q). \tag{24}$$

We will now work on creating an asynchronous relaxation of $\mathcal{G}$ through the use of the one-sample backup map H. Let $\rho \in \Delta(\mathcal{S})$ such that $\rho_{\min} := \min_{i\in\mathcal{S}} \rho_i > 0$. We define

$$D_\rho := \mathrm{diag}(\rho_1, \dots, \rho_m), \qquad h_\rho(p) := p + D_\rho\big(\mathcal{G}(p) - p\big). \tag{25}$$

h_ρ is the *mean-field map* of the one-sample stochastic update under state-sampling law ρ, in the standard SA sense. Of particular interest will be when $\rho = \mu$, the stationary distribution of the Markov chain under π. We have the following favorable results on this averaged map.

Proposition 13. *Suppose $Y = (S, B, S')$ is sampled such that $S \sim \rho$ and*

$$\mathbb{P}(S' = j \mid S = i) = P_{ij}^{\pi}, \qquad B \mid (S = i, S' = j) \sim \nu_{ij} \quad \textit{for all } i, j \in \mathcal{S}. \tag{26}$$

Then, $\mathbb{E}\big[H(p,Y)\big] = h_\rho(p)$ for all $p \in \Delta_d^{\mathcal{S}}$, h_ρ is non-expansive in $\ell_{\mathrm{C},\infty}^{\Theta}$, and $\mathrm{fix}(h_\rho) = \mathrm{fix}(\mathcal{G})$. In particular, the same conclusions hold with $\rho = \mu$ under the standing irreducibility assumption.

The next corollary will be useful in translating residual bounds for h_ρ into residual bounds for $\mathcal{G}$.

Corollary 14. *For every $p \in \Delta_d^{\mathcal{S}}$,*

$$\rho_{\min}\ell_{\mathrm{C},\infty}^{\Theta}\big(p, \mathcal{G}(p)\big) \le \ell_{\mathrm{C},\infty}^{\Theta}\big(p, h_\rho(p)\big) \le \ell_{\mathrm{C},\infty}^{\Theta}\big(p, \mathcal{G}(p)\big). \tag{27}$$

Proofs of Proposition 11, Lemma 12, Proposition 13, and Corollary 14 are given in Appendix C.

We state the two main SA results for the centered-reward recursion approximating $\mathcal{G}$. The i.i.d. regime is the relevant model for settings such as replay-buffer sampling or a generative simulator, where we can approximately draw independent one-step samples from a fixed sampling law. The Markovian regime captures online asynchronous TD-learning along a single trajectory of the policy-induced chain. In both settings, the quantity we control is the mean-field residual $\ell_{\mathrm{C},\infty}^{\Theta}\big(p_k, h_\rho(p_k)\big)$. This is the standard quantity in non-expansive SA [29, 11, 10, 7, 25], where fixed points need not be unique.

Theorem 15 (Centered recursion i.i.d. convergence)**.** *Suppose the samples $Y_k = (S_k, B_k, S_k')$ are i.i.d., that $S_k \sim \rho$ with $\rho_{\min} > 0$ and that*

$$\mathbb{P}(S_k' = j \mid S_k = i) = P_{ij}^{\pi}, \qquad B_k \mid (S_k = i, S_k' = j) \sim \nu_{ij} \quad \textit{for all } i, j \in \mathcal{S}. \tag{28}$$

The following two conclusions hold for the centered SKM recursion (22). First, if $\alpha_k = (k+1)^{-a}$ with $\frac{2}{3} < a \le 1$, then p_k converges a.s. to a random fixed point $p^\star \in \mathrm{fix}(h_\rho) = \mathrm{fix}(\mathcal{G})$. Second, fix any $a_1 \in \big(\frac{2}{3}, 1\big)$ and set $\varepsilon_{a_1} := (3a_1 - 2)/6$. There is an explicit two-phase step size schedule and constants $C_{\mathrm{iid}}, \kappa_{a_1} > 0$ such that, for all $k \ge 1$,

$$\mathbb{E}\Big[\ell_{\mathrm{C},\infty}^{\Theta}\big(p_k, \mathcal{G}(p_k)\big)\Big] \le \frac{C_{\mathrm{iid}}}{\rho_{\min}}(k+1)^{-1/6} \min\left\{\kappa_{a_1} \log(k+1), (k+1)^{\varepsilon_{a_1}}\right\}, \tag{29}$$

where C_{iid} depends only on a_1, $|\mathcal{S}|$, and $\mathrm{diam}_{\ell_{\mathrm{C},\infty}^{\Theta}}(\Delta_d^{\mathcal{S}}) = \sqrt{\theta_d - \theta_1}$. The full fixed-exponent rate profile is presented in Appendix C. In particular, choosing $a = \frac{2}{3} + \epsilon$ for $\epsilon > 0$ gives a.s. convergence with expected residual decaying at rate $O(k^{-1/6+\epsilon/2})$.

Theorem 16 (Centered recursion Markovian convergence)**.** *Let $(S_k)_{k\ge0}$ be the policy-induced Markov chain, which is irreducible and aperiodic with stationary distribution μ, and let $Y_k = (S_k, B_k, S_{k+1})$ be the associated centered one-sample process. The following two conclusions hold for the centered SKM recursion (22). First, if $\alpha_k = (k+1)^{-a}$ with $\frac{4}{5} < a \le 1$, then p_k converges a.s. to a random fixed point $p^\star \in \mathrm{fix}(h_\mu) = \mathrm{fix}(\mathcal{G})$. Second, fix any $a_1 \in \big(\frac{4}{5}, 1\big)$ and set*

$\varepsilon_{a_1} := (5a_1 - 4)/10$. *There is an explicit two-phase step size schedule and constants* $C_{\mathrm{mk}}, \kappa_{a_1} > 0$ *such that, for all* $k \geq 1$,

$$\mathbb{E}\Big[\ell^{\Theta}_{\mathrm{C},\infty}\big(p_k, \mathcal{G}(p_k)\big)\Big] \leq \frac{C_{\mathrm{mk}}}{\mu_{\min}}(k+1)^{-1/10} \min\left\{\kappa_{a_1} \log(k+1), (k+1)^{\varepsilon_{a_1}}\right\}, \tag{30}$$

where C_{mk} *depends on* a_1, $|\mathcal{S}|$, $\mathrm{diam}_{\ell^{\Theta}_{\mathrm{C},\infty}}(\Delta_d^{\mathcal{S}}) = \sqrt{\theta_d - \theta_1}$, *and a finite-state Poisson constant for* (Y_k), *equivalently expressable through the norm of its fundamental matrix. The full fixed-exponent rate profile is presented in Appendix C. In particular, choosing* $a = \frac{4}{5} + \epsilon$ *for* $\epsilon > 0$ *gives a.s. convergence with expected residual decaying at rate* $O(k^{-1/10+\epsilon/2})$.

Proof sketch for Theorems 15 and 16. By Lemma 12, the one-sample backup map is non-expansive. By Proposition 13 and Corollary 8, the mean-field map is non-expansive with a non-empty set of fixed points. Compactness of $\Delta_d^{\mathcal{S}}$ yields bounded iterates and bounded martingale noise, so the i.i.d. recursion falls under the bounded-noise non-expansive SA results of Bravo and Cominetti [10]. In the Markovian regime, finite $\mathcal{S}$ and $\mathcal{A}$ and the deterministic reward $R(s,a)$ make $(Y_k)_{k\geq 0}$ a finite-state irreducible and aperiodic chain, so the fixed-exponent results of Blaser and Zhang [7] apply. The two-phase residual bounds are proved separately in Appendix C. □

These bounds use a two-phase step size. A fixed supercritical exponent gives a.s. convergence and a logarithm-free residual bound but loses a polynomial factor, whereas the critical exponent gives the sharper power of k up to a logarithm. The two-phase schedule tracks the smaller of the two. Its switch point depends only on a_1 but can be very late, so over practical horizons the fixed-exponent bound applies. The powers $k^{-1/6}$ and $k^{-1/10}$ come from general non-expansive SA and are neither lower bounds nor intrinsic categorical rates. Faster instance-dependent rates may follow from anchored variance-reduced methods [24].

6 A coupled recursion for when the gain is unknown

The centered recursion isolates the non-expansive structure, but it is not directly implementable unless the gain $\bar{r}^{\pi}$ is known. In practice, we observe uncentered rewards R, not $B := R - \bar{r}^{\pi}$.

Proposition 17. *In general, replacing the centered reward* $B = R - \bar{r}^{\pi}$ *by the raw reward* R *in the one-sample backup* (21) *does not yield an SA of* $\mathcal{G}$.

The example is proved in Appendix D. We now replace centered rewards by uncentered ones, and augment the quotient-categorical update with an online gain estimate g on $\mathcal{Z} := \Delta_d^{\mathcal{S}} \times [0,1]$. We show that the augmented update is non-expansive, identify its mean-field map and its fixed points, and relate its residual to $\ell^{\Theta}_{\mathrm{C},\infty}\big(p, \mathcal{G}(p)\big)$ and the gain error $|g - \bar{r}^{\pi}|$. For $g \in \mathbb{R}$ and an uncentered sample $y = (s, r, s') \in \mathcal{S} \times [0,1] \times \mathcal{S}$, define the gain-parameterized one-sample backup map $H_g(p, y) = q$, where

$$q_u = \begin{cases} p_u, & u \neq s, \\ \tilde{q}, & u = s, \end{cases} \tag{31}$$

where $\tilde{q} \in \Delta_d$ is the unique coefficient vector satisfying $\Pi^{\Theta}_{\mathrm{C}}\big((\tau_{r-g})_{\#}\eta^{p_{s'},0}\big) = \eta^{\tilde{q},0}$. The corresponding mean-field map is defined in terms of the stationary state distribution μ as $h_{\mu}^{(g)}(p) := p + D_{\mu}(\mathcal{G}_g(p) - p)$. Throughout the section, fix $\lambda \geq \varsigma^{-1/2}$, and define the product metric

$$d_{\lambda}\big((p,g),(q,g')\big) := \ell^{\Theta}_{\mathrm{C},\infty}(p,q) + \lambda|g - g'|. \tag{32}$$

We will now use a coupled SKM recursion with the same step size, one for the categorical coefficients, and the other for the gain:

$$\begin{aligned} p_{k+1} &= p_k + \alpha_k\big(H_{g_k}(p_k, Y_k) - p_k\big), \\ g_{k+1} &= g_k + \alpha_k(R_k - g_k), \end{aligned} \tag{33}$$

where $Y_k = (S_k, R_k, S_{k+1})$. The gain recursion is a running average of the observed rewards, used to center each sampled reward. We next show that the augmented update is non-expansive in d_{λ}.

Proposition 18. *For every uncentered sample* $y = (s, r, s') \in \mathcal{S} \times [0,1] \times \mathcal{S}$, *and* $(p,g),(q,g') \in \mathcal{Z}$,

$$d_{\lambda}\big((H_g(p,y), r), (H_{g'}(q,y), r)\big) \leq d_{\lambda}\big((p,g),(q,g')\big). \tag{34}$$

The next proposition identifies the mean-field map associated with the stationary distribution μ.

Proposition 19. *Let $Y = (S, R, S')$ be a one-step sample from the stationary distribution of the policy-induced chain, i.e.,*

$$S \sim \mu, \qquad \mathbb{P}(S' = j \mid S = i) = P^{\pi}_{ij}, \qquad R \mid (S = i, S' = j) \sim \mathrm{Law}(R_{ij}) \tag{35}$$

for all $i, j \in \mathcal{S}$. Then

$$\mathbb{E}\big[\big(H_g(p, Y), R\big)\big] = \big(h^{(g)}_{\mu}(p), \bar{r}^{\pi}\big) \quad \text{for all } (p, g) \in \mathcal{Z}. \tag{36}$$

Moreover, the map $(p, g) \mapsto \big(h^{(g)}_{\mu}(p), \bar{r}^{\pi}\big)$ is non-expansive in d_λ, and $\mathrm{fix}\big((p, g) \mapsto \big(h^{(g)}_{\mu}(p), \bar{r}^{\pi}\big)\big) = \{(p, \bar{r}^{\pi}) : p \in \mathrm{fix}(\mathcal{G})\}$.

Corollary 20. *For every $(p, g) \in \mathcal{Z}$,*

$$\mu_{\min} \ell^{\Theta}_{\mathrm{C},\infty}\big(p, \mathcal{G}(p)\big) \le d_\lambda\big((p, g), \big(h^{(g)}_{\mu}(p), \bar{r}^{\pi}\big)\big) \quad \text{and} \quad |g - \bar{r}^{\pi}| \le \lambda^{-1} d_\lambda\big((p, g), \big(h^{(g)}_{\mu}(p), \bar{r}^{\pi}\big)\big). \tag{37}$$

Appendix D proves the non-expansiveness, mean-field, residual-conversion and convergence results. We now establish the non-asymptotic convergence of the coupled SKM recursion (33).

Theorem 21 (Coupled recursion Markovian convergence)**.** *Let $(S_k)_{k\ge 0}$ be the policy-induced Markov chain, which is irreducible and aperiodic with stationary distribution μ, let $Y_k = (S_k, R_k, S_{k+1})$ be the associated uncentered one-sample process, and let $\lambda \ge \varsigma^{-1/2}$. First, if $\alpha_k = (k+1)^{-a}$ with $\frac{4}{5} < a \le 1$, then the coupled SKM recursion (33) converges a.s. to a random limit $(p^\star, \bar{r}^{\pi})$ such that $p^\star \in \mathrm{fix}(\mathcal{G})$. Second, fix any $a_1 \in \big(\frac{4}{5}, 1\big)$ and set $\varepsilon_{a_1} := (5a_1 - 4)/10$. There is an explicit two-phase step size schedule and constants $C^{\mathcal{Z}}_{\mathrm{mk}}, \kappa_{a_1} > 0$ such that, for all $k \ge 1$,*

$$\begin{aligned} \mathbb{E}\big[\ell^{\Theta}_{\mathrm{C},\infty}\big(p_k, \mathcal{G}(p_k)\big)\big] &\le \frac{C^{\mathcal{Z}}_{\mathrm{mk}}}{\mu_{\min}} (k+1)^{-1/10} \min\left\{\kappa_{a_1} \log(k+1), (k+1)^{\varepsilon_{a_1}}\right\}, \\ \mathbb{E}[|g_k - \bar{r}^{\pi}|] &\le \frac{C^{\mathcal{Z}}_{\mathrm{mk}}}{\lambda} (k+1)^{-1/10} \min\left\{\kappa_{a_1} \log(k+1), (k+1)^{\varepsilon_{a_1}}\right\}, \end{aligned} \tag{38}$$

where $C^{\mathcal{Z}}_{\mathrm{mk}}$ depends on a_1, $\mathrm{diam}_{d_\lambda}(\mathcal{Z}) = \sqrt{\theta_d - \theta_1} + \lambda$, and on the corresponding finite-state Poisson constant for the uncentered chain (Y_k). The fixed-exponent choice $a = \frac{4}{5} + \epsilon$ for $\epsilon > 0$ gives a.s. convergence with expected residual and gain error decaying at rate $O(k^{-1/10+\epsilon/2})$.

The $k^{-1/10}$ rate comes from the joint guarantee, and sharper rates for g_k follow from standard scalar SA results, since the gain recursion does not depend on p (e.g., the running reward average satisfies $\mathbb{E}|g_k - \bar{r}^{\pi}| = O(k^{-1/2})$). Because $C^{\mathcal{Z}}_{\mathrm{mk}}$ grows with λ, the smallest admissible choice $\lambda = \varsigma^{-1/2}$ is preferable. Appendix E shows that unit-step synchronous backups avoid the gain at the exact quotient-law level, a structure that projection and averaging break.

7 Experimental results

We present two fixed-policy evaluation experiments: five-state MDPs validating the quotient-categorical and SA results, and a continuous-state `Pendulum-v1` experiment [12] in Appendix G, illustrating neural function approximation. The experiments directly test whether the algorithms behave according to the theoretical predictions rather than compare task performance against unrelated baselines. Both use a uniformly spaced categorical support Θ and the supremum-Cramér metric.

Five-state MDPs. One five-state MDP compares exact KM, centered SKM with i.i.d. and Markovian samples, and the coupled recursion. The second illustrates stopped-return distributions. Appendix F gives the transition matrices, rewards, grids, and stepsize schedules.

Figure 2 reports the residual $\ell^{\Theta}_{\mathrm{C},\infty}(p_k, \mathcal{G}(p_k))$ (bounded in Theorems 15, 16 and 21) and the gain error. In panel (a), both stochastic residuals decay at roughly $k^{-1/3}$ to $k^{-2/5}$ (Appendix F), well ahead of the worst-case rates $k^{-1/8}$ and $k^{-3/40}$ guaranteed for these step sizes. In panel (b), estimating the gain online costs essentially nothing relative to the known-gain recursion, consistent with Theorem 21. Panel (c) isolates centering: with $g = 0$ every sampled shift is nonnegative, the iterates approach $\mathrm{fix}(\mathcal{G}_0)$ where all mass sits at θ_d, and the residual to $\mathcal{G}$ grows instead of decaying.

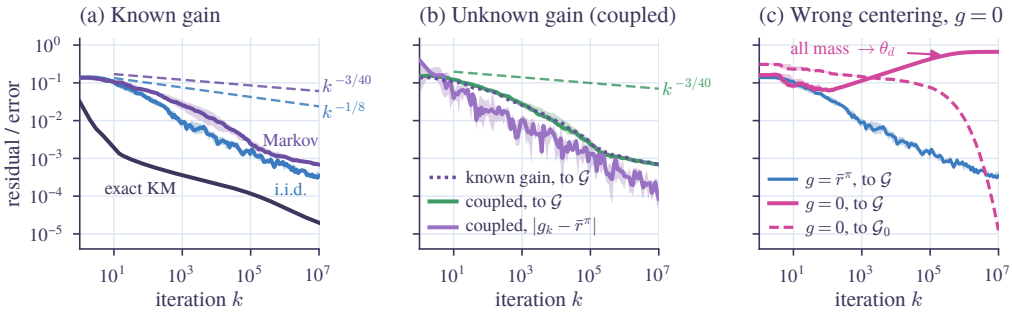


Figure 2: Five-state MDP. Residuals are $\ell^{\Theta}_{\mathrm{C},\infty}(p_k, \mathcal{G}(p_k))$ unless labeled $\mathcal{G}_0$. Stochastic runs show medians over ten seeds with interquartile bands. **(a)** Known gain. **(b)** Unknown gain, with the known-gain Markovian residual from (a) dotted. **(c)** Centering with $g = 0$, which targets $\mathcal{G}_0$. Dashed lines: theorem rates for the step sizes used, up to constants.

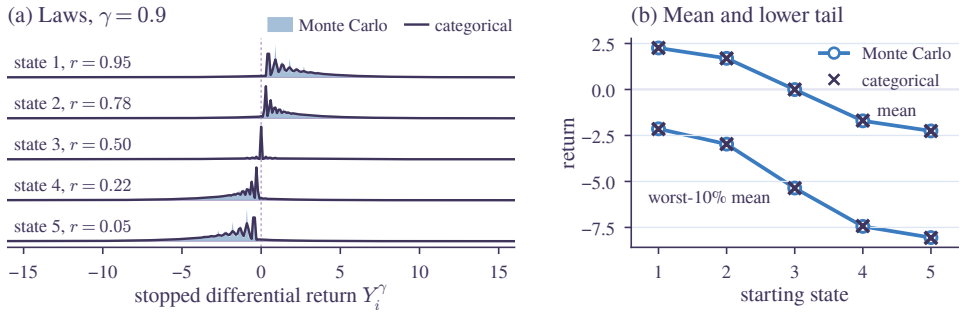


Figure 3: Stopped differential returns, $\gamma = 0.9$ (mean horizon: ten rewards). **(a)** Projected Monte Carlo laws (10^5 returns per state, shaded blue) and categorical fixed points (dark lines), each scaled to its peak. **(b)** Means and worst-10% means per starting state.

Figure 3 evaluates the target from Theorem 9. Persistent high- and low-reward regions produce distinct, skewed laws across starting states, and the categorical fixed points track the Monte Carlo laws and lower-tail summaries. The grid $[-20, 20]$ is wide enough for clipping to be negligible, since fewer than 0.03% of sampled stopped returns fall outside it.

8 Conclusion

We introduced a quotient-categorical formulation for average-reward distributional evaluation of differential returns, resolving the additive-constant ambiguity of the scalar bias at the level of state-indexed laws. This viewpoint leads to a projected operator that is well-defined and non-expansive in a coordinate Cramér metric, and to TD-style SA schemes with a.s. convergence and finite-iteration residual guarantees. When the gain is unknown, augmenting the categorical recursion with an online gain estimate preserves the same non-expansive structure and yields joint control of the projected residual and gain error. The projected fixed points arise as limits of geometrically stopped differential-return laws, and their means recover the scalar bias up to an explicit clipping error. Experiments on a five-state MDP agree with the theory, and a neural `Pendulum-v1` experiment in Appendix G illustrates it qualitatively. The same machinery should also extend to multivariate categorical representations by replacing the one-dimensional cumulative embedding and projection with their multivariate analogues, following the construction of Kaya and Hashemi [23]. Our theory covers finite-state fixed-policy evaluation with ergodic sampling and a fixed categorical grid. A reducible chain can leave states unvisited and have class-dependent gains, and without a single closed class in the state–atom kernel the projected fixed point need not be unique, although the Abel limit selects a canonical one. A stochastic anchored-selection theorem, nonlinear function approximation, and control remain open.

Acknowledgments and Disclosure of Funding

This work is supported in part by NSF DMS-2502560 and CNS-2313109.

References

[1] Jinane Abounadi, Dimitri Bertsekas, and Vivek S Borkar. Learning algorithms for Markov decision processes with average cost. *SIAM Journal on Control and Optimization*, 40(3): 681–698, 2001.

[2] Charalambos D Aliprantis and Kim C Border. *Infinite dimensional analysis: a hitchhiker's guide*. Springer, 2006.

[3] Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pages 449–458. Pmlr, 2017.

[4] Marc G Bellemare, Will Dabney, and Mark Rowland. *Distributional reinforcement learning*. MIT Press, 2023.

[5] Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

[6] Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pages 1691–1692. PMLR, 2018.

[7] Ethan Blaser and Shangtong Zhang. Asymptotic and finite sample analysis of nonexpansive stochastic approximations with Markovian noise. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 19764–19772, 2026.

[8] Vivek S Borkar. Asynchronous stochastic approximations. *SIAM Journal on Control and Optimization*, 36(3):840–851, 1998.

[9] Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 100. Springer, 2008.

[10] Mario Bravo and Roberto Cominetti. Stochastic fixed-point iterations for nonexpansive maps: Convergence and error bounds. *SIAM Journal on Control and Optimization*, 62(1):191–219, 2024.

[11] Mario Bravo, Roberto Cominetti, and Matías Pavez-Signé. Rates of convergence for inexact Krasnosel'skii–Mann iterations in Banach spaces. *Mathematical Programming*, 175(1):241–262, 2019.

[12] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI Gym, 2016.

[13] Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. *Advances in Neural Information Processing Systems*, 33:8223–8234, 2020.

[14] Zaiwei Chen, Sheng Zhang, Thinh T Doan, John-Paul Clarke, and Siva Theja Maguluri. Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning. *Automatica*, 146:110623, 2022.

[15] Zaiwei Chen, Siva T Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. A Lyapunov theory for finite-sample guarantees of Markovian stochastic approximation. *Operations Research*, 72(4):1352–1367, 2024.

[16] Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pages 1096–1105. PMLR, 2018.

[17] Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[18] Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. Finite sample analyses for TD (0) with function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[19] Peter Dayan and Terrence J Sejnowski. TD (λ) converges with probability 1. *Machine Learning*, 14(3):295–301, 1994.

[20] Kristopher De Asis, Alan Chan, Silviu Pitis, Richard Sutton, and Daniel Graves. Fixed-horizon temporal difference methods for stable reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3741–3748, 2020.

[21] Godfrey H Hardy. *Divergent Series*, volume 334. American Mathematical Society, 2024.

[22] Tyler Kastner, Mark Rowland, Yunhao Tang, Murat A Erdogdu, and Amir-massoud Farahmand. Categorical distributional reinforcement learning with Kullback-Leibler divergence: Convergence and asymptotics. In *Forty-second International Conference on Machine Learning*, 2025.

[23] Ege C. Kaya and Abolfazl Hashemi. A sharp finite-iteration theory for asynchronous categorical distributional temporal-difference learning, 2026. URL https://arxiv.org/abs/2605.06866.

[24] Ege C. Kaya, Arda Fazla, M. Berk Sahin, and Abolfazl Hashemi. A Banach-space theory of Markovian Halpern iteration for non-expansive maps, 2026. URL https://arxiv.org/abs/2608.15966.

[25] Tae-Hwa Kim and Hong-Kun Xu. Robustness of Mann's algorithm for nonexpansive mappings. *Journal of Mathematical Analysis and Applications*, 327(2):1105–1115, 2007.

[26] Mark Aleksandrovich Krasnosel'skii. Two remarks on the method of successive approximations. *Uspekhi matematicheskikh nauk*, 10(1):123–127, 1955.

[27] Clare Lyle, Marc G Bellemare, and Pablo Samuel Castro. A comparative analysis of expected and distributional reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4504–4511, 2019.

[28] Sridhar Mahadevan. Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine learning*, 22(1):159–195, 1996.

[29] W Robert Mann. Mean value methods in iteration. *Proceedings of the American Mathematical Society*, 4(3):506–510, 1953.

[30] Sean Meyn, Richard L. Tweedie, and Peter W. Glynn. *Markov Chains and Stochastic Stability*. Cambridge Mathematical Library. Cambridge University Press, 2 edition, 2009.

[31] Gandharv Patil, LA Prashanth, Dheeraj Nagaraj, and Doina Precup. Finite time analysis of temporal difference learning with linear function approximation: Tail averaging and regularisation. In *International Conference on Artificial Intelligence and Statistics*, pages 5438–5448. PMLR, 2023.

[32] Yang Peng, Liangyu Zhang, and Zhihua Zhang. Statistical efficiency of distributional temporal difference learning. *Advances in Neural Information Processing Systems*, 37:24724–24761, 2024.

[33] Yang Peng, Kaicheng Jin, Liangyu Zhang, and Zhihua Zhang. A finite sample analysis of distributional TD learning with linear function approximation. *arXiv preprint arXiv:2502.14172*, 2025.

[34] Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition, 1994. ISBN 0471619779.

[35] Guannan Qu and Adam Wierman. Finite-time analysis of asynchronous stochastic approximation and q-learning. In *Conference on learning theory*, pages 3185–3205. PMLR, 2020.

[36] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[37] Juan Sebastian Rojas and Chi-Guhn Lee. A differential perspective on distributional reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 25160–25167, 2026.

[38] Mark Rowland, Marc Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 29–37. PMLR, 2018.

[39] Mark Rowland, Robert Dadashi, Saurabh Kumar, Rémi Munos, Marc G Bellemare, and Will Dabney. Statistics and samples in distributional reinforcement learning. In *International Conference on Machine Learning*, pages 5528–5536. PMLR, 2019.

[40] Mark Rowland, Li K Wenliang, Rémi Munos, Clare Lyle, Yunhao Tang, and Will Dabney. Near-minimax-optimal distributional reinforcement learning with a generative model. *Advances in Neural Information Processing Systems*, 37:132774–132823, 2024.

[41] Anton Schwartz. A reinforcement learning method for maximizing undiscounted rewards. *Machine Learning Proceedings 1993*, 1993.

[42] Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation and TD learning. In *Conference on learning theory*, pages 2803–2830. PMLR, 2019.

[43] Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44, 1988.

[44] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[45] John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-difference learning with function approximation. *Advances in neural information processing systems*, 9, 1996.

[46] John N Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine learning*, 16(3):185–202, 1994.

[47] Sattar Vakili and Julia Olkhovskaya. Kernel-based function approximation for average reward reinforcement learning: an optimist no-regret algorithm. *Advances in Neural Information Processing Systems*, 37:25401–25425, 2024.

[48] Yi Wan, Abhishek Naik, and Richard S Sutton. Learning and planning in average-reward Markov decision processes. In *International Conference on Machine Learning*, pages 10653–10662. PMLR, 2021.

[49] Kaiwen Wang, Kevin Zhou, Runzhe Wu, Nathan Kallus, and Wen Sun. The benefits of being distributional: Small-loss bounds for reinforcement learning. *Advances in neural information processing systems*, 36:2275–2312, 2023.

[50] Kaiwen Wang, Owen Oertell, Alekh Agarwal, Nathan Kallus, and Wen Sun. More benefits of being distributional: Second-order bounds for reinforcement learning. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 51192–51213. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wang24ba.html.

[51] Yue Wang, Alvaro Velasquez, George K Atia, Ashley Prater-Bennette, and Shaofeng Zou. Model-free robust average-reward reinforcement learning. In *International Conference on Machine Learning*, pages 36431–36469. PMLR, 2023.

[52] Harley Wiltzer, Jesse Farebrother, Arthur Gretton, and Mark Rowland. Foundations of multivariate distributional reinforcement learning. *Advances in Neural Information Processing Systems*, 37:101297–101336, 2024.

[53] Runzhe Wu, Masatoshi Uehara, and Wen Sun. Distributional offline policy evaluation with predictive error guarantees. In *International Conference on Machine Learning*, pages 37685–37712. PMLR, 2023.

[54] Matthew Zurek and Yudong Chen. Span-based optimal sample complexity for weakly communicating and general average reward MDPs. *Advances in Neural Information Processing Systems*, 37:33455–33504, 2024.

Appendix Table of Contents

A Proofs of Section 4 **17**

A.1 Proof of Proposition 2 . . . 17
A.2 Proof of Proposition 3 . . . 17
A.3 Proof of Lemma 4 . . . 17
A.4 Proof of Proposition 5 . . . 17
A.5 Proof of Theorem 7 . . . 18
A.6 Proof of Corollary 8 . . . 18

B Interpretation and approximation results **19**

B.1 Formal statements . . . 19
B.2 Recursively projected returns . . . 20
B.3 Proof of Theorem 24 . . . 20
B.4 Proof of Proposition 25 . . . 21
B.5 Proof of Proposition 26 . . . 22

C Proofs of Section 5 **24**

C.1 Proof of Proposition 11 . . . 24
C.2 Proof of Lemma 12 . . . 25
C.3 Proof of Proposition 13 . . . 25
C.4 Proof of Corollary 14 . . . 25
C.5 Proof of Theorem 15 . . . 25
C.6 Proof of Theorem 16 . . . 29
C.7 Proof of Proposition 17 . . . 33

D Proofs of Section 6 **34**

D.1 Proof of Proposition 18 . . . 35
D.2 Proof of Proposition 19 . . . 35
D.3 Proof of Corollary 20 . . . 36
D.4 Proof of Theorem 21 . . . 36

E Synchronous gain-free structure **37**

F Five-state experimental details **38**

G Neural function approximation results **38**

A Proofs of Section 4

A.1 Proof of Proposition 2

Proof of Proposition 2. Fix $\eta \in \mathcal{F}^{\mathcal{S}}$, $c \in \mathbb{R}$, and $i \in \mathcal{S}$. For each j, convolution commutes with common translation:

$$\nu_{ij} * \big((\tau_c)_{\#}\eta_j\big) = (\tau_c)_{\#}(\nu_{ij} * \eta_j). \tag{39}$$

Since pushforward by the same translation is linear on mixtures,

$$\begin{aligned}\big(\mathcal{T}\big((\tau_c)_{\#}\eta\big)\big)_i &= \sum_{j\in\mathcal{S}} P^{\pi}_{ij}\, \nu_{ij} * \big((\tau_c)_{\#}\eta_j\big) \\ &= \sum_{j\in\mathcal{S}} P^{\pi}_{ij}\, (\tau_c)_{\#}(\nu_{ij} * \eta_j) \\ &= (\tau_c)_{\#} \sum_{j\in\mathcal{S}} P^{\pi}_{ij}(\nu_{ij} * \eta_j) = \big((\tau_c)_{\#}\mathcal{T}\eta\big)_i.\end{aligned} \tag{40}$$

This proves (3). Therefore, if $\eta \approx \zeta$, then $\mathcal{T}\eta \approx \mathcal{T}\zeta$, so $\overline{\mathcal{T}}([\eta]) := [\mathcal{T}\eta]$ is independent of the representative. □

A.2 Proof of Proposition 3

Proof of Proposition 3. Fix $i \in \mathcal{S}$. Since $R_{ij} - g' = (R_{ij} - g) + (g - g')$, we have

$$\nu_{ij}^{(g')} = (\tau_{g-g'})_{\#}\nu_{ij}^{(g)}. \tag{41}$$

Using the shift-equivariance of translation and convolution:

$$\nu_{ij}^{(g')} * \eta_j = \big((\tau_{g-g'})_{\#}\nu_{ij}^{(g)}\big) * \eta_j = (\tau_{g-g'})_{\#}(\nu_{ij}^{(g)} * \eta_j). \tag{42}$$

Summing over j therefore yields

$$(\mathcal{T}_{g'}\eta)_i = (\tau_{g-g'})_{\#}(\mathcal{T}_g\eta)_i. \tag{43}$$

Hence, $\mathcal{T}_{g'}\eta$ and $\mathcal{T}_g\eta$ differ by a common translation and thus determine the same equivalence class. □

A.3 Proof of Lemma 4

Proof of Lemma 4. Write $\Pi^{\Theta}_{\mathrm{C}}\eta = \sum_{k=1}^{d} p_k \delta_{\theta_k}$. The coefficients p_k are determined by the usual linear interpolation weights that split mass between neighboring support points in Θ, together with edge clipping outside the interval $[\theta_1, \theta_d]$. Replacing η by $(\tau_c)_{\#}\eta$ and the support Θ by Θ_c translates every relevant location by the same scalar c, so the interpolation weights are unchanged. Consequently, the projected coefficients are the same, and the projected law is translated by c. □

A.4 Proof of Proposition 5

Proof of Proposition 5. Fix a state $i \in \mathcal{S}$ and a common shift $c \in \mathbb{R}$. Since both $\eta_i^{p,c}$ and $\eta_i^{q,c}$ are supported on the same ordered grid Θ_c, the difference of their cdfs is constant on each interval $[\theta_k + c, \theta_{k+1} + c)$, where it equals $\sum_{j=1}^{k}(p_{i,j} - q_{i,j})$, and it vanishes outside $[\theta_1 + c, \theta_d + c)$ because both cdfs equal 0 below $\theta_1 + c$ and 1 from $\theta_d + c$ on. Therefore

$$\begin{aligned}\ell_{\mathrm{C}}(\eta_i^{p,c}, \eta_i^{q,c})^2 &= \sum_{k=1}^{d-1}(\theta_{k+1} - \theta_k)\Big(\sum_{j=1}^{k}(p_{i,j} - q_{i,j})\Big)^2 \\ &= \varsigma \sum_{k=1}^{d-1}\big(F_{p_i}(\theta_k) - F_{q_i}(\theta_k)\big)^2 = \ell^{\Theta}_{\mathrm{C}}(p_i, q_i)^2.\end{aligned} \tag{44}$$

Taking square roots proves the statewise identity, and the block-supremum identity follows by taking the maximum over the states. □

A.5 Proof of Theorem 7

We first state two lemmas.

Lemma 22. *For every $g \in \mathbb{R}$ and all $\eta, \zeta \in \mathcal{F}^{\mathcal{S}}$,*

$$\ell_{\mathrm{C},\infty}(\mathcal{T}_g\eta, \mathcal{T}_g\zeta) \le \ell_{\mathrm{C},\infty}(\eta, \zeta). \tag{45}$$

Proof. Fix $i \in \mathcal{S}$. The output law at state i is a mixture of shifted successor laws, so

$$F_{(\mathcal{T}_g\eta)_i}(x) = \sum_{j\in\mathcal{S}} P^{\pi}_{ij}\mathbb{E}\big[F_{\eta_j}(x - R_{ij} + g)\big], \tag{46}$$

and similarly for ζ. Jensen's inequality then gives

$$\ell_{\mathrm{C}}\big((\mathcal{T}_g\eta)_i, (\mathcal{T}_g\zeta)_i\big)^2 \le \sum_{j\in\mathcal{S}} P^{\pi}_{ij}\mathbb{E}\Big[\int_{\mathbb{R}} \big(F_{\eta_j}(x - R_{ij} + g) - F_{\zeta_j}(x - R_{ij} + g)\big)^2 \mathrm{d}x\Big]. \tag{47}$$

A change of variables gets rid of the translation, yielding

$$\ell_{\mathrm{C}}\big((\mathcal{T}_g\eta)_i, (\mathcal{T}_g\zeta)_i\big)^2 \le \sum_{j\in\mathcal{S}} P^{\pi}_{ij}\ell_{\mathrm{C}}(\eta_j, \zeta_j)^2 \le \ell_{\mathrm{C},\infty}(\eta, \zeta)^2. \tag{48}$$

Taking the maximum of the left-hand side over states and taking square roots proves the claim. □

Lemma 23. *For each $c \in \mathbb{R}$, the scalar categorical projection $\Pi_{\mathrm{C}}^{\Theta_c}$ is non-expansive in the Cramér metric. Consequently, its statewise product is non-expansive in $\ell_{\mathrm{C},\infty}$.*

Proof. This is the standard scalar categorical Cramér projection property, see Rowland et al. [38]. Applying it statewise and taking the supremum over the blocks yield the claim. □

Proof of Theorem 7. Fix $g \in \mathbb{R}$. To prove well-definedness, fix $p \in \Delta_d^{\mathcal{S}}$, and two representatives $\eta^{p,c}$ and $\eta^{p,c'}$. By the argument of Proposition 2, which applies verbatim to $\mathcal{T}_g$,

$$\mathcal{T}_g\eta^{p,c'} = (\tau_{c'-c})_{\#}\mathcal{T}_g\eta^{p,c}. \tag{49}$$

Since $\eta^{p,c} = (\tau_c)_{\#}\eta^{p,0}$, the preceding display with (c', c) replaced by $(c, 0)$ gives $\mathcal{T}_g\eta^{p,c} = (\tau_c)_{\#}\mathcal{T}_g\eta^{p,0}$, and Lemma 4 applied statewise yields $\Pi_{\mathrm{C}}^{\Theta_c}(\mathcal{T}_g\eta^{p,c}) = (\tau_c)_{\#}\Pi_{\mathrm{C}}^{\Theta}(\mathcal{T}_g\eta^{p,0})$. The same holds with c' in place of c, so both representatives yield the same probability coefficients.

For non-expansiveness of $\mathcal{G}_g$ fix the representative shift $c = 0$. Then

$$\eta^{\mathcal{G}_g(p),0} = \Pi_{\mathrm{C}}^{\Theta}(\mathcal{T}_g\eta^{p,0}), \qquad \eta^{\mathcal{G}_g(q),0} = \Pi_{\mathrm{C}}^{\Theta}(\mathcal{T}_g\eta^{q,0}). \tag{50}$$

Using Proposition 5 and Lemmas 22 and 23,

$$\begin{aligned} \ell_{\mathrm{C},\infty}^{\Theta}\big(\mathcal{G}_g(p), \mathcal{G}_g(q)\big) &= \ell_{\mathrm{C},\infty}(\eta^{\mathcal{G}_g(p),0}, \eta^{\mathcal{G}_g(q),0}) \\ &\le \ell_{\mathrm{C},\infty}(\mathcal{T}_g\eta^{p,0}, \mathcal{T}_g\eta^{q,0}) \\ &\le \ell_{\mathrm{C},\infty}(\eta^{p,0}, \eta^{q,0}) = \ell_{\mathrm{C},\infty}^{\Theta}(p, q). \end{aligned} \tag{51}$$

□

A.6 Proof of Corollary 8

Proof of Corollary 8. $\Delta_d^{\mathcal{S}}$ is a nonempty compact convex subset of a finite-dimensional Euclidean space. By Theorem 7, $\mathcal{G}_g$ is 1-Lipschitz in the cumulative-coordinate metric, which on this finite-dimensional set is equivalent to the Euclidean metric, so $\mathcal{G}_g$ is continuous. Brouwer's fixed-point theorem [2] therefore guarantees the existence of fixed points. □

B Interpretation and approximation results

This section collects the formal statements behind the interpretation summarized in Section 4.1. Let $C_\Theta(x)$ denote a draw from the categorical projection $\Pi_{\mathrm{C}}^{\Theta}\delta_x$. Thus, inside the grid it takes one of the two neighboring atoms with the interpolation probabilities in (6), and outside the grid it equals the nearest endpoint.

B.1 Formal statements

As in Section 4.1, assume without loss of generality that $0 \in \Theta$, let p^0 put unit mass at 0 in every state, and define $p^{n+1} = \mathcal{G}(p^n)$. Equivalently, if $Z_i^{(0)} = 0$, the recursively projected n-step return obeys

$$Z_i^{(n+1)} = C_\Theta\big(B_{iJ} + Z_J^{(n)}\big), \qquad J \sim P_{i\cdot}^{\pi}, \tag{52}$$

with a fresh successor, a reward drawn conditionally on that transition, and a fresh projection draw at each step.

Theorem 24 (Geometric target and Abel endpoint)**.** *Let $\gamma \in (0,1)$ and define*

$$(\mathcal{T}_\gamma \eta)_i := \sum_j P_{ij}^{\pi}\, \nu_{ij} * \big((1-\gamma)\delta_0 + \gamma\eta_j\big), \qquad \mathcal{G}_\gamma(p) := (1-\gamma)\mathcal{G}(p^0) + \gamma\mathcal{G}(p). \tag{53}$$

Then $\mathcal{T}_\gamma$ and $\mathcal{G}_\gamma$ are γ-contractive in $\ell_{\mathrm{C},\infty}$ and $\ell_{\mathrm{C},\infty}^{\Theta}$, respectively. The unique fixed point of $\mathcal{T}_\gamma$ is the law η^γ of the stopped return in (15), while the unique fixed point of $\mathcal{G}_\gamma$ is

$$p_\gamma = (1-\gamma)\sum_{n=1}^{\infty} \gamma^{n-1}\mathcal{G}^n(p^0). \tag{54}$$

Moreover, with $h := r^\pi - \bar{r}^\pi \mathbf{1}$,

$$\mathbb{E}[Y^\gamma] = (I - \gamma P^\pi)^{-1}h = \sum_{t=0}^{\infty} \gamma^t (P^\pi)^t h \longrightarrow v^\circ \quad \text{as } \gamma \uparrow 1, \tag{55}$$

and

$$p_\gamma \longrightarrow p^\dagger := \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \mathcal{G}^n(p^0) \in \mathrm{fix}(\mathcal{G}). \tag{56}$$

Finally, fixed points of $\mathcal{G}$ are in bijection with invariant distributions of the finite state–atom kernel in (74). In particular, the fixed point is unique exactly when that kernel has one closed communicating class.

Here, p^0 represents zero terminal continuation. After each centered reward, the trajectory stops with probability $1-\gamma$ and continues with probability γ, so $\mathbb{E}[N_\gamma] = 1/(1-\gamma)$. This target can be sampled from the same one-step transitions used by TD by drawing an independent stopping decision after each observed reward, without requiring a model of the MDP. As $\gamma \uparrow 1$, $\mathcal{G}_\gamma(p) \to \mathcal{G}(p)$ for every p, while the mean horizon diverges and the Abel limit above selects an undiscounted fixed point.

To see why the stopped differential-return laws are useful, and how they can distinguish between policies, consider two one-state reward processes. The first always gives reward $1/2$. In the second, two actions with rewards 0 and 1 are selected independently at each step with equal probability. Both have gain $1/2$ and stationary-mean-zero bias 0. After centering, however, the first stopped return is identically zero, whereas the second sums independent increments in $\{-1/2, 1/2\}$ over the geometric horizon and has variance $1/[4(1-\gamma)]$. Thus the stopped law records transient reward risk that is invisible to both the gain and the scalar bias.

For the remaining statements, take the uniform symmetric grid $\Theta = \{-M, -M+\varsigma, \dots, M\}$ and recall that $|B_{ij}| \le 1$, since rewards lie in $[0,1]$.

Proposition 25 (Stopped-return approximation)**.** *For every $\gamma \in (0,1)$,*

$$\ell_{\mathrm{C},\infty}(\eta^{p_\gamma,0}, \eta^\gamma) \le \frac{1}{1-\gamma}\left(\frac{\varsigma}{4} + \frac{\gamma^{\lfloor M \rfloor}}{1-\gamma}\right)^{1/2}. \tag{57}$$

Corresponding W_1 and CVaR bounds are given in the proof below.

For a fixed γ and desired accuracy, this bound gives a direct grid-selection rule: increase M until the geometric clipping term is negligible, then increase d to reduce the spacing $\varsigma = 2M/(d-1)$. Values of γ closer to one require wider supports because they produce longer return horizons.

Recall the means $m(p)$ and the clipping defect $\beta(p)$ from (18), and let $Z_\pi := (I - P^\pi + \mathbf{1}\mu^\top)^{-1}$ and $V_\pi := \inf_c \|v^\pi - c\mathbf{1}\|_\infty$.

Proposition 26 (Poisson consistency and grid error)**.** *Every $p^\star \in \mathrm{fix}(\mathcal{G})$ satisfies*

$$(I - P^\pi)m(p^\star) = r^\pi - \bar{r}^\pi \mathbf{1} + \beta(p^\star), \tag{58}$$

and therefore

$$\inf_c \|m(p^\star) - v^\pi - c\mathbf{1}\|_\infty \le \|Z_\pi\|_\infty \|\beta(p^\star)\|_\infty. \tag{59}$$

Writing $C_\pi := 2(V_\pi + \|Z_\pi\|_\infty) + 1$, the clipping defect obeys

$$\|\beta(p^\star)\|_\infty \le \frac{C_\pi + \varsigma}{2M\mu_{\min}} = \frac{C_\pi}{2M\mu_{\min}} + \frac{1}{(d-1)\mu_{\min}}. \tag{60}$$

More generally, every $p \in \Delta_d^{\mathcal{S}}$ satisfies

$$\inf_c \|m(p) - v^\pi - c\mathbf{1}\|_\infty \le \|Z_\pi\|_\infty \left(\sqrt{2M}\, \ell^{\Theta}_{\mathrm{C},\infty}(p, \mathcal{G}p) + \|\beta(p)\|_\infty\right). \tag{61}$$

Finally, the clipping defect of any $p \in \Delta_d^{\mathcal{S}}$ is controlled by its residual $\ell := \ell^{\Theta}_{\mathrm{C},\infty}(p, \mathcal{G}p)$:

$$\|\beta(p)\|_\infty \le \frac{C_\pi + \varsigma + 2\big(\|Z_\pi\|_\infty \sqrt{2M} + \sqrt{2/3}\, M^{3/2}\big)\ell}{2M\mu_{\min}}, \tag{62}$$

and therefore

$$\inf_c \|m(p) - v^\pi - c\mathbf{1}\|_\infty \le \|Z_\pi\|_\infty \Big(\frac{C_\pi + \varsigma}{2M\mu_{\min}} + K_M\, \ell\Big), \tag{63}$$

where $K_M := \sqrt{2M} + \big(\sqrt{2M/3} + \|Z_\pi\|_\infty \sqrt{2/M}\big)/\mu_{\min}$. Thus fixed-point mean error vanishes as $M, d \to \infty$, and so does the mean error of any iterate whose residual is $o(M^{-1/2})$.

In particular, since $\varsigma/(2M) = 1/(d-1) \le 2/d$ and, for $M \ge 1$, $K_M \le \sqrt{M}\big(\sqrt{2} + (\sqrt{2/3} + \sqrt{2}\,\|Z_\pi\|_\infty)/\mu_{\min}\big)$, Proposition 10 holds with

$$C_1 := \frac{\|Z_\pi\|_\infty}{\mu_{\min}} \max\Big\{\frac{C_\pi}{2}, 2\Big\}, \qquad C_2 := \|Z_\pi\|_\infty \Big(\sqrt{2} + \frac{\sqrt{2/3} + \sqrt{2}\,\|Z_\pi\|_\infty}{\mu_{\min}}\Big). \tag{64}$$

B.2 Recursively projected returns

Let p^0 put unit mass at 0 in every state. We prove by induction that the recursion (52) has law $\eta_i^{\mathcal{G}^n(p^0),0}$ at depth n. The claim is immediate for $n = 0$. If it holds at depth n, then conditioning on $J = j$ and using the conditional independence of the one-step reward and the fresh continuation return gives

$$\mathrm{Law}(B_{ij} + Z_j^{(n)}) = \nu_{ij} * \eta_j^{\mathcal{G}^n(p^0),0}. \tag{65}$$

The law of $C_\Theta(B_{ij} + Z_j^{(n)})$ is the categorical projection of this convolution. Averaging over j and using linearity of categorical projection therefore gives exactly the ith block of $\mathcal{G}^{n+1}(p^0)$.

B.3 Proof of Theorem 24

Proof. Conditioning on the first transition and on whether the process stops after its first centered reward gives

$$\eta_i^\gamma = \sum_j P_{ij}^\pi\, \nu_{ij} * \big((1-\gamma)\delta_0 + \gamma \eta_j^\gamma\big), \tag{66}$$

so η^γ is a fixed point of $\mathcal{T}_\gamma$. For any $\eta, \zeta \in \mathcal{F}^{\mathcal{S}}$, the cdf difference of the continuation mixtures is $\gamma(F_{\eta_j} - F_{\zeta_j})$. The same Jensen and translation argument as in Lemma 22 therefore yields

$$\ell_{\mathrm{C},\infty}(\mathcal{T}_\gamma \eta, \mathcal{T}_\gamma \zeta) \le \gamma \ell_{\mathrm{C},\infty}(\eta, \zeta). \tag{67}$$

Existence is supplied by the stopped-return law above. If $\zeta = \mathcal{T}_\gamma \zeta$, then $\ell_{\mathrm{C},\infty}(\zeta, \eta^\gamma) = \ell_{\mathrm{C},\infty}(\mathcal{T}_\gamma \zeta, \mathcal{T}_\gamma \eta^\gamma) \le \gamma\, \ell_{\mathrm{C},\infty}(\zeta, \eta^\gamma)$, so $\ell_{\mathrm{C},\infty}(\zeta, \eta^\gamma) = 0$ because $\gamma < 1$. Hence η^γ is unique.

Categorical projection is linear on mixtures, so applying it to the stopped backup gives

$$\mathcal{G}_\gamma(p) = (1-\gamma)\mathcal{G}(p^0) + \gamma \mathcal{G}(p). \tag{68}$$

Theorem 7 hence implies

$$\ell^{\Theta}_{\mathrm{C},\infty}(\mathcal{G}_\gamma p, \mathcal{G}_\gamma q) \le \gamma \ell^{\Theta}_{\mathrm{C},\infty}(p, q). \tag{69}$$

Iterating the affine map gives the unique fixed point

$$p_\gamma = (1-\gamma) \sum_{n=1}^{\infty} \gamma^{n-1} \mathcal{G}^n(p^0). \tag{70}$$

The term $\mathcal{G}^n(p^0)$ is the n-step recursively projected return proved above, and its coefficient $(1-\gamma)\gamma^{n-1}$ is precisely $\mathbb{P}(N_\gamma = n)$.

Writing $h := r^\pi - \bar{r}^\pi \mathbf{1}$, taking expectations in the stopped Bellman equation gives

$$y^\gamma = h + \gamma P y^\gamma, \qquad y^\gamma := \mathbb{E}[Y^\gamma], \tag{71}$$

and therefore $y^\gamma = (I - \gamma P^\pi)^{-1} h = \sum_{t \ge 0} \gamma^t (P^\pi)^t h$. Since $\mu^\top h = 0$, we also have $\mu^\top y^\gamma = 0$. Under irreducibility and aperiodicity, $(P^\pi)^t h$ is summable and the resolvent converges to the stationary-mean-zero Poisson solution v° as $\gamma \uparrow 1$.

To establish the endpoint statement, note that $\mathcal{G}$ is linear on the coefficient space, although its matrix is not stochastic. By the intertwining relation (75) proved below, $w = D_\mu p$ is mapped to $wQ^n = D_\mu \mathcal{G}^n(p)$ for the stochastic kernel Q. Consequently, the powers of $\mathcal{G}$ are bounded on $\Delta_d^{\mathcal{S}}$, its Cesàro averages converge by the finite-dimensional mean ergodic theorem for Q, and their limit is fixed by $\mathcal{G}$. The finite-dimensional Abel–Cesàro theorem [21] gives

$$\lim_{\gamma \uparrow 1} (1-\gamma) \sum_{n=1}^{\infty} \gamma^{n-1} \mathcal{G}^n(p^0) = \lim_{N \to \infty} \frac{1}{N} \sum_{n=1}^{N} \mathcal{G}^n(p^0) = p^\dagger \in \mathrm{fix}(\mathcal{G}). \tag{72}$$

For completeness, we now give the claimed finite Markov representation. For $1 \le l, k \le d$, define

$$K_{ij}(l,k) := \mathbb{P}\big(C_\Theta(B_{ij} + \theta_l) = \theta_k\big), \qquad P^{\leftarrow}_{ji} := \frac{\mu_i P^\pi_{ij}}{\mu_j}, \tag{73}$$

and define a kernel Q on $\mathcal{S} \times \{1, \dots, d\}$ by

$$Q((j,l),(i,k)) := P^{\leftarrow}_{ji} K_{ij}(l,k). \tag{74}$$

Both K_{ij} and $P^{\leftarrow}$ are stochastic, hence so is Q. If $w_{i,k} := \mu_i p_{i,k}$, then direct substitution gives

$$(wQ)_{i,k} = \mu_i (\mathcal{G} p)_{i,k}. \tag{75}$$

Every invariant distribution of Q has state marginal μ, because its state marginal is invariant for the irreducible time-reversed chain $P^{\leftarrow}$. Thus (75) is a bijection between invariant distributions of Q and fixed points of $\mathcal{G}$. A finite Markov kernel has a unique invariant distribution exactly when it has one closed communicating class, which proves the final claim. □

B.4 Proof of Proposition 25

Proof. Let Π denote statewise categorical projection onto $[-M, M]$. Since $\eta^{p_\gamma, 0} = \Pi \mathcal{T}_\gamma \eta^{p_\gamma, 0}$ and $\eta^\gamma = \mathcal{T}_\gamma \eta^\gamma$, contraction and non-expansiveness of projection imply

$$\begin{aligned} \ell_{\mathrm{C},\infty}(\eta^{p_\gamma,0}, \eta^\gamma) &\le \ell_{\mathrm{C},\infty}(\Pi \mathcal{T}_\gamma \eta^{p_\gamma,0}, \Pi \mathcal{T}_\gamma \eta^\gamma) + \ell_{\mathrm{C},\infty}(\Pi \eta^\gamma, \eta^\gamma) \\ &\le \gamma \ell_{\mathrm{C},\infty}(\eta^{p_\gamma,0}, \eta^\gamma) + \ell_{\mathrm{C},\infty}(\Pi \eta^\gamma, \eta^\gamma). \end{aligned} \tag{76}$$

Rearranging gives the standard contraction-perturbation bound

$$\ell_{\mathrm{C},\infty}(\eta^{p_\gamma,0}, \eta^\gamma) \le \frac{\ell_{\mathrm{C},\infty}(\Pi \eta^\gamma, \eta^\gamma)}{1-\gamma}. \tag{77}$$

For a point x inside one grid cell, a direct integration of the two cdfs in (6) gives

$$\ell_{\mathrm{C}}(\Pi\delta_x, \delta_x)^2 = \frac{(x-\theta_k)(\theta_{k+1}-x)}{\varsigma} \le \frac{\varsigma}{4}. \tag{78}$$

For $|x| > M$, endpoint clipping gives $\ell_{\mathrm{C}}(\Pi\delta_x, \delta_x)^2 = |x| - M$. Convexity under mixtures therefore yields

$$\ell_{\mathrm{C}}(\Pi\xi, \xi)^2 \le \frac{\varsigma}{4} + \mathbb{E}_\xi[(|X| - M)_+]. \tag{79}$$

Since $|Y_i^\gamma| \le bN_\gamma$,

$$\mathbb{P}(|Y_i^\gamma| > M) \le \gamma^{\lfloor M \rfloor}, \qquad \mathbb{E}[(|Y_i^\gamma| - M)_+] \le \frac{\gamma^{\lfloor M \rfloor}}{1-\gamma}. \tag{80}$$

Combining (77)–(80) proves (57).

We finally record the promised risk-metric conversion. If ξ and ζ are supported on $[-M, M]$, Cauchy–Schwarz applied to their cdfs gives

$$W_1(\xi, \zeta) = \int_{-M}^{M} |F_\xi - F_\zeta| \le \sqrt{2M}\, \ell_{\mathrm{C}}(\xi, \zeta). \tag{81}$$

Let $e_\gamma := \gamma^{\lfloor M \rfloor}/(1-\gamma)$ and let $\bar\eta_i^\gamma$ be the law of $[Y_i^\gamma]_{-M}^{M}$. Then $W_1(\eta_i^\gamma, \bar\eta_i^\gamma) \le e_\gamma$ and $\ell_{\mathrm{C}}(\eta_i^\gamma, \bar\eta_i^\gamma) \le \sqrt{e_\gamma}$. Hence, with D_γ denoting the right-hand side of (57),

$$W_1(\eta_i^{p_\gamma, 0}, \eta_i^\gamma) \le e_\gamma + \sqrt{2M}(D_\gamma + \sqrt{e_\gamma}). \tag{82}$$

For $\mathrm{CVaR}_\alpha(X) := \inf_t\{t + (1-\alpha)^{-1}\mathbb{E}[(X-t)_+]\}$, the map from a law to CVaR is $(1-\alpha)^{-1}$-Lipschitz in W_1. Dividing (82) by $1-\alpha$ therefore gives an explicit CVaR approximation bound. The corresponding downside-CVaR bound follows by applying the same statement to $-X$.

Consequently, the stopped laws support fixed-policy comparisons through quantiles, downside probabilities, and CVaR at the chosen expected horizon. □

B.5 Proof of Proposition 26

Proof. Categorical interpolation preserves the input mean inside $[-M, M]$, while endpoint projection clips it. Consequently,

$$m(\mathcal{G}p) = r^\pi - \bar r^\pi \mathbf{1} + Pm(p) + \beta(p). \tag{83}$$

At a fixed point, this identity becomes (58). Left multiplication by $\mu^\top$ gives $\mu^\top\beta(p^\star) = 0$. Since $Z_\pi(I - P^\pi) = I - \mathbf{1}\mu^\top$, comparison with the Poisson equation yields

$$m(p^\star) - v^\pi = Z_\pi\beta(p^\star) + c\mathbf{1} \tag{84}$$

for some c, proving (59).

For a general p, set $e(p) := m(p) - m(\mathcal{G}p)$. Equation (83) gives

$$(I - P^\pi)m(p) = r^\pi - \bar r^\pi \mathbf{1} + \beta(p) + e(p). \tag{85}$$

The difference of the means of two laws on $[-M, M]$ is their integrated cdf difference, so Cauchy–Schwarz gives

$$\|e(p)\|_\infty \le \sqrt{2M}\, \ell^{\Theta}_{\mathrm{C},\infty}(p, \mathcal{G}p). \tag{86}$$

Applying Z_π proves (61).

To prove the uniform clipping estimate, first note that $|\beta_i(p^\star)| \le 1$, since shifting a point in $[-M, M]$ by B_{ij} can move it beyond an endpoint by at most $|B_{ij}|$. Let $I \sim \mu$, $J \sim P_{I\cdot}^\pi$, $X \sim \eta^{p^\star_J, 0}$, $B \sim \nu_{IJ}$, and $C = C_\Theta(X + B)$, with the continuation and projection draws conditionally independent. Define

$$D(x) := x^2 - ([x]_{-M}^{M})^2 \ge 2M(|x| - M)_+. \tag{87}$$

For $x \in [-M, M]$, categorical interpolation satisfies

$$\mathbb{E}[C_\Theta(x)^2] = x^2 + V(x), \qquad 0 \le V(x) \le \varsigma \tag{88}$$

when $x = X + B$: the interpolation variance is the product of the distances to the two neighboring atoms, one of which is at most $|B|$, while both are at most ς. At a fixed point, C and X have the same distribution after averaging with $I \sim \mu$, which gives

$$\mathbb{E}[D(X+B)] = 2\mathbb{E}[BX] + \mathbb{E}[B^2] + \mathbb{E}[V(X+B)]. \tag{89}$$

The centered reward has $\mathbb{E}[B] = 0$ under the stationary transition law. Moreover, (59) and the preliminary bound $\|\beta(p^\star)\|_\infty \le 1$ allow a constant c such that

$$\max_j |m_j(p^\star) - c| \le V_\pi + \|Z_\pi\|_\infty. \tag{90}$$

Conditional independence gives

$$|\mathbb{E}[BX]| = |\mathbb{E}[B(m_J(p^\star) - c)]| \le V_\pi + \|Z_\pi\|_\infty. \tag{91}$$

Thus (89) is at most $C_\pi + \varsigma$. On the other hand,

$$\mu_i |\beta_i(p^\star)| \le \mathbb{E}[(|X+B| - M)_+ \mathbf{1}\{I = i\}] \le \frac{\mathbb{E}[D(X+B)]}{2M}. \tag{92}$$

Taking the maximum over i, using $\mu_i \ge \mu_{\min}$ and $\varsigma = 2M/(d-1)$, proves (60).

For a general p, repeat this argument with $p^\star$ replaced by p. The preliminary bound $|\beta_i(p)| \le 1$ still holds. Since C has law $\sum_i \mu_i \eta^{(\mathcal{G}p)_i,0}$ and X has law $\sum_j \mu_j \eta^{p_j,0}$, identity (89) acquires the correction $\mathbb{E}[X^2] - \mathbb{E}[C^2]$ on its right-hand side. For laws ξ, ζ on $[-M, M]$, integration by parts and Cauchy–Schwarz give

$$\big|\mathbb{E}_\xi[X^2] - \mathbb{E}_\zeta[X^2]\big| = \Big| \int_{-M}^{M} 2x\big(F_\zeta(x) - F_\xi(x)\big)\,\mathrm{d}x \Big| \le 2\sqrt{2M^3/3}\,\ell_{\mathrm{C}}(\xi, \zeta), \tag{93}$$

so the correction is at most $2\sqrt{2/3}\,M^{3/2}\ell$. In the bound on $\mathbb{E}[BX]$, (61) and $\|\beta(p)\|_\infty \le 1$ give a constant c with $\max_j |m_j(p) - c| \le V_\pi + \|Z_\pi\|_\infty(1 + \sqrt{2M}\,\ell)$. With these two changes, the argument above yields (62), and substituting it into (61) yields (63). □

C Proofs of Section 5

For the convenience of proofs, let us write $L_b : \Delta_d \to \Delta_d$ for the unique coefficient map satisfying

$$\Pi_{\mathrm{C}}^{\Theta}\big((\tau_b)_{\#}\eta^{p,0}\big) = \eta^{L_b p,0}, \qquad p \in \Delta_d, b \in \mathbb{R}. \tag{94}$$

We first fix three conventions used in all SA proofs below. First, following Bravo and Cominetti [10] and Blaser and Zhang [7], we write the recursion as $p_k = p_{k-1} + \alpha_k\big(H(p_{k-1}, Y_k) - p_{k-1}\big)$ for $k \geq 1$. This is (22) started from its first iterate $p_1 = H(p_0, Y_0)$, with indices shifted by one. Every step size $\alpha_k = (k+1)^{-a}$ used in the analysis therefore lies in $(0,1)$, the iterate p_{k-1} is measurable with respect to $Y_0, \dots, Y_{k-1}$, and, for $k \geq 2$, the shift changes the constants in the residual bounds by at most a factor of 2. The first iterate, which uses $\alpha_0 = 1$, is covered by the trivial bound $\mathrm{diam}_{\ell^{\Theta}_{\mathrm{C},\infty}}(\Delta_d^{\mathcal{S}})$ after enlarging constants.

Second, we identify $p \in \Delta_d^{\mathcal{S}}$ with $F(p) := \big(F_{p_i}(\theta_k)\big)_{i\in\mathcal{S},\, k\leq d-1}$, so that $\ell^{\Theta}_{\mathrm{C},\infty}(p,q) = \sqrt{\varsigma}\max_i \|F(p)_i - F(q)_i\|_2 =: \|F(p) - F(q)\|_{\mathrm{C},\infty}$, and let $\|\cdot\|_{\mathrm{C},2}$ denote $\sqrt{\varsigma}$ times the ambient Euclidean norm, so that $\|\cdot\|_{\mathrm{C},\infty} \leq \|\cdot\|_{\mathrm{C},2} \leq \sqrt{|\mathcal{S}|}\,\|\cdot\|_{\mathrm{C},\infty}$. The image of $\Delta_d^{\mathcal{S}}$ is $K_0^{\mathcal{S}}$ with $K_0 := \{u \in \mathbb{R}^{d-1} : 0 \leq u_1 \leq \cdots \leq u_{d-1} \leq 1\}$, a compact convex set.

Third, the SA results we cite are stated for maps defined on the whole space, and their proofs evaluate the mean-field map at auxiliary points such as $x_k - \overline{U}_k$ below, which may leave $K_0^{\mathcal{S}}$. Let Proj be the blockwise Euclidean projection onto $K_0^{\mathcal{S}}$. Euclidean projection onto a closed convex set is 1-Lipschitz, so Proj is non-expansive in $\|\cdot\|_{\mathrm{C},\infty}$. Hence $\tilde{H}(\cdot, y) := H(\mathrm{Proj}(\cdot), y)$ and $\tilde{h} := h \circ \mathrm{Proj}$, for $h \in \{h_\rho, h_\mu\}$, are non-expansive on the whole space, take values in $K_0^{\mathcal{S}}$, coincide with $H(\cdot, y)$ and h on $K_0^{\mathcal{S}}$, and satisfy $\mathrm{fix}(\tilde{h}) = \mathrm{fix}(h)$, since every fixed point lies in the range $K_0^{\mathcal{S}}$. The SKM iterates are convex combinations of points of $K_0^{\mathcal{S}}$, so replacing (H, h) by $(\tilde{H}, \tilde{h})$ changes neither the iterates nor the noise, and the bounded-range alternative of hypothesis (H4) in Bravo and Cominetti [10] holds with constant $\mathrm{diam}_{\ell^{\Theta}_{\mathrm{C},\infty}}(\Delta_d^{\mathcal{S}})$. All SA results below are applied to these extended maps. For the coupled recursion of Appendix D, the gain coordinate is extended in the same way by clipping to $[0,1]$.

C.1 Proof of Proposition 11

Proof of Proposition 11. Fix $g \in \mathbb{R}$ and $i \in \mathcal{S}$ and consider the representative shift $c = 0$. By definition of $\mathcal{G}_g$,

$$\eta_i^{\mathcal{G}_g(p),0} = \Pi_{\mathrm{C}}^{\Theta}\Big(\sum_{j\in\mathcal{S}} P_{ij}^{\pi}(\nu_{ij}^{(g)} * \eta_j^{p,0})\Big). \tag{95}$$

Because scalar categorical projection is linear on mixtures,

$$\eta_i^{\mathcal{G}_g(p),0} = \sum_{j\in\mathcal{S}} P_{ij}^{\pi}\Pi_{\mathrm{C}}^{\Theta}(\nu_{ij}^{(g)} * \eta_j^{p,0}). \tag{96}$$

Now, fix $j \in \mathcal{S}$ and let $B \sim \nu_{ij}^{(g)}$. Using the definition of L_b,

$$\Pi_{\mathrm{C}}^{\Theta}(\delta_b * \eta_j^{p,0}) = \eta^{L_b p_j,0} \qquad \text{for every } b \in \mathbb{R}. \tag{97}$$

Taking expectation over B,

$$\Pi_{\mathrm{C}}^{\Theta}(\nu_{ij}^{(g)} * \eta_j^{p,0}) = \mathbb{E}\big[\eta^{L_B p_j,0}\big]. \tag{98}$$

Since each law inside the expectation is categorical on the same support Θ, taking expectation simply averages the coefficient vectors, so

$$\mathbb{E}\big[\eta^{L_B p_j,0}\big] = \eta^{\mathbb{E}[L_B p_j],0}. \tag{99}$$

Comparing coefficients on the common support Θ shows that

$$\mathcal{G}_g(p)_i = \sum_{j\in\mathcal{S}} P_{ij}^{\pi}\mathbb{E}[L_B p_j], \qquad B \sim \nu_{ij}^{(g)}. \tag{100}$$

□

C.2 Proof of Lemma 12

Proof of Lemma 12. Let $y = (s, b, s')$. For blocks $u \neq s$, the sampled update leaves the block unchanged. For the updated block,

$$\ell_{\mathrm{C}}^{\Theta}\big(H(p,y)_s, H(q,y)_s\big) = \ell_{\mathrm{C}}^{\Theta}(L_b p_{s'}, L_b q_{s'}). \tag{101}$$

Using Proposition 5 together with projection non-expansiveness and translation invariance of ℓ_{C} gives

$$\ell_{\mathrm{C}}^{\Theta}(L_b p_{s'}, L_b q_{s'}) \leq \ell_{\mathrm{C}}^{\Theta}(p_{s'}, q_{s'}). \tag{102}$$

Taking the maximum over all state blocks proves the result. □

C.3 Proof of Proposition 13

Proof of Proposition 13. The expectation identity follows blockwise from Proposition 11. $\mathrm{fix}(\mathcal{G}) \subseteq \mathrm{fix}(h_\rho)$ follows trivially from the definition of h_ρ. Conversely, if $h_\rho(p) = p$, then $\rho_i(\mathcal{G}(p)_i - p_i) = 0$ for every $i \in \mathcal{S}$. $\rho_{\min} > 0$ implies that $\mathcal{G}(p)_i = p_i$ for every $i \in \mathcal{S}$, hence $p \in \mathrm{fix}(\mathcal{G})$.

For the non-expansiveness statement, fix $i \in \mathcal{S}$. Since

$$h_\rho(p)_i = (1-\rho_i)p_i + \rho_i \mathcal{G}(p)_i, \qquad h_\rho(q)_i = (1-\rho_i)q_i + \rho_i \mathcal{G}(q)_i, \tag{103}$$

we have for each $k = 1, \ldots, d-1$,

$$F_{h_\rho(p)_i}(\theta_k) - F_{h_\rho(q)_i}(\theta_k) = (1-\rho_i)\big(F_{p_i}(\theta_k) - F_{q_i}(\theta_k)\big) + \rho_i\big(F_{\mathcal{G}(p)_i}(\theta_k) - F_{\mathcal{G}(q)_i}(\theta_k)\big). \tag{104}$$

Through the triangle inequality with the vector of cumulative differences with $(d-1)$ elements,

$$\begin{aligned} \ell_{\mathrm{C}}^{\Theta}\big(h_\rho(p)_i, h_\rho(q)_i\big) &\leq (1-\rho_i)\ell_{\mathrm{C}}^{\Theta}(p_i, q_i) + \rho_i \ell_{\mathrm{C}}^{\Theta}\big(\mathcal{G}(p)_i, \mathcal{G}(q)_i\big) \\ &\leq \ell_{\mathrm{C},\infty}^{\Theta}(p, q), \end{aligned} \tag{105}$$

where the last inequality uses Theorem 7. Taking the maximum over $i \in \mathcal{S}$ proves non-expansiveness. □

C.4 Proof of Corollary 14

Proof of Corollary 14. For each state $i \in \mathcal{S}$,

$$h_\rho(p)_i - p_i = \rho_i\big(\mathcal{G}(p)_i - p_i\big). \tag{106}$$

Hence, for each $k = 1, \ldots, d-1$,

$$F_{h_\rho(p)_i}(\theta_k) - F_{p_i}(\theta_k) = \rho_i\big(F_{\mathcal{G}(p)_i}(\theta_k) - F_{p_i}(\theta_k)\big), \tag{107}$$

and therefore

$$\ell_{\mathrm{C}}^{\Theta}\big(p_i, h_\rho(p)_i\big) = \rho_i \ell_{\mathrm{C}}^{\Theta}\big(p_i, \mathcal{G}(p)_i\big). \tag{108}$$

Taking the maximum over i yields

$$\ell_{\mathrm{C},\infty}^{\Theta}\big(p, h_\rho(p)\big) = \max_{i \in \mathcal{S}} \rho_i \ell_{\mathrm{C}}^{\Theta}\big(p_i, \mathcal{G}(p)_i\big), \tag{109}$$

from which the stated bounds follow immediately. □

C.5 Proof of Theorem 15

We first state the following proposition.

Proposition 27 (Two-phase i.i.d. residual refinement)**.** *Under the assumptions of Theorem 15, fix* $a_1 \in \left(\frac{2}{3}, 1\right)$ *and set*

$$\varepsilon_{a_1} := \frac{3a_1 - 2}{6}, \qquad \kappa_{a_1} := \frac{2^{\varepsilon_{a_1}}}{\log 2}. \tag{110}$$

Let T *be the largest integer such that*

$$(k+1)^{\varepsilon_{a_1}} \leq \kappa_{a_1} \log(k+1), \qquad 1 \leq k \leq T. \tag{111}$$

If the centered SKM recursion uses the two-phase step size

$$\alpha_k = \begin{cases} (k+1)^{-a_1}, & 1 \le k \le T, \\ (k+1)^{-2/3}, & k > T, \end{cases} \tag{112}$$

then there exists $C_{\mathrm{iid}} > 0$ such that, for all $k \ge 1$,

$$\mathbb{E}\Big[\ell^{\Theta}_{\mathrm{C},\infty}\big(p_k, h_\rho(p_k)\big)\Big] \le C_{\mathrm{iid}}(k+1)^{-1/6} \min\Big\{\kappa_{a_1}\log(k+1), (k+1)^{\varepsilon_{a_1}}\Big\}. \tag{113}$$

Consequently,

$$\mathbb{E}\Big[\ell^{\Theta}_{\mathrm{C},\infty}\big(p_k, \mathcal{G}(p_k)\big)\Big] \le \frac{C_{\mathrm{iid}}}{\rho_{\min}}(k+1)^{-1/6} \min\Big\{\kappa_{a_1}\log(k+1), (k+1)^{\varepsilon_{a_1}}\Big\}. \tag{114}$$

Proof. Since the tail exponent in (112) is the critical value $2/3$, the a.s. convergence assertion remains the separate $a > \frac{2}{3}$ assertion in Theorem 15. The role of the two-phase schedule is to interpolate between the near-critical transient behavior of an exponent $a_1 > \frac{2}{3}$ and the critical last-iterate residual behavior.

We first record the constants used below. Let

$$D := \sqrt{\theta_d - \theta_1}, \qquad M_2 := \sqrt{|\mathcal{S}|}D, \tag{115}$$

and define

$$S_{3a_1/2} := \sum_{t=1}^{\infty}(t+1)^{-3a_1/2}, \quad \omega_{a_1} := \frac{1-2^{-a_1}}{2}, \quad \nu_{a_1} := \frac{1-2^{-a_1}}{1-a_1}, \quad \eta_{a_1} := 1-(3/4)^{1-a_1}. \tag{116}$$

Also let

$$R_{a_1} := \frac{(4/3)^{a_1}}{1-4^{-a_1}}, \tag{117}$$

$$K_{a_1} := \frac{D}{\sqrt{\pi\omega_{a_1}}} + \frac{2^{1+a_1/2}M_2}{\sqrt{\pi}}\left(\frac{S_{3a_1/2}}{\sqrt{\nu_{a_1}\eta_{a_1}}} + \frac{2^{1+a_1/2}R_{a_1}}{\sqrt{1-a_1}}\right), \tag{118}$$

and

$$B_T := \max\left\{\frac{2}{\sqrt{3}}, (32/3)^{1/6}(T+2)^{\varepsilon_{a_1}}\right\}. \tag{119}$$

The final constant can be chosen as

$$C_{\mathrm{iid}} := \max\left\{K_{a_1} + 6M_2,\ K_{a_1}B_T + \frac{2^{4+a_1/2}M_2\sqrt{6}}{3\sqrt{\pi}} + 6M_2\right\}. \tag{120}$$

The threshold in (111) is well defined. Indeed, for $x \ge 2$, the function $x^{\varepsilon_{a_1}}/\log x$ decreases until $x = \exp(1/\varepsilon_{a_1})$ and then increases to infinity. Its value at $x = 2$ is κ_{a_1}. Hence the set of integers satisfying (111) is a nonempty initial interval, and after its last element T the reverse inequality holds.

We now derive the residual estimate. Work in the cumulative-coordinate embedding associated with $\ell^{\Theta}_{\mathrm{C},\infty}$. The block-sup Cramér metric is the block-sup norm of Euclidean cumulative-coordinate vectors, and this block-sup norm is bounded by the ambient Euclidean norm. Under the i.i.d. assumptions, the noise

$$U_k := H(p_{k-1}, Y_k) - h_\rho(p_{k-1}) \tag{121}$$

is a martingale difference and satisfies $\|U_k\|_{\mathrm{C},2} \le M_2$. Define the averaged noise process

$$\overline{U}_k = (1-\alpha_k)\overline{U}_{k-1} + \alpha_k U_k, \qquad \overline{U}_0 = 0. \tag{122}$$

By orthogonality of martingale differences in the Euclidean cumulative-coordinate space and Jensen's inequality,

$$\mathbb{E}\big[\|\overline{U}_k\|_{\mathrm{C},\infty}\big] \le \mathbb{E}\big[\|\overline{U}_k\|_{\mathrm{C},2}\big] \le M_2\sqrt{\sum_{t=1}^{k}\alpha_t^2\prod_{r=t+1}^{k}(1-\alpha_r)^2}. \tag{123}$$

Let

$$v_k := \sum_{t=1}^{k} \alpha_t^2 \prod_{r=t+1}^{k} (1-\alpha_r)^2. \tag{124}$$

We claim that $v_k \le \alpha_k$ for all $k \ge 1$. Since

$$v_k = (1-\alpha_k)^2 v_{k-1} + \alpha_k^2, \tag{125}$$

it is enough by induction to verify

$$\alpha_k^{-1} - \alpha_{k-1}^{-1} \le 1. \tag{126}$$

In the first phase this follows from the mean-value theorem:

$$(k+1)^{a_1} - k^{a_1} \le a_1 k^{a_1-1} < 1. \tag{127}$$

In the second phase,

$$(k+1)^{2/3} - k^{2/3} \le \frac{2}{3} k^{-1/3} < 1. \tag{128}$$

At the transition point,

$$\alpha_{T+1}^{-1} - \alpha_T^{-1} = (T+2)^{2/3} - (T+1)^{a_1} \le (T+1)^{2/3} - (T+1)^{a_1} + \frac{2}{3}(T+1)^{-1/3} < 1, \tag{129}$$

because $a_1 > 2/3$. Therefore (123) gives

$$\mathbb{E}\big[\|\overline{U}_k\|_{\mathrm{C},\infty}\big] \le M_2 \sqrt{\alpha_k}. \tag{130}$$

Following the inexact Krasnosel'skiĭ–Mann estimate of Bravo and Cominetti [10, Theorem 2.10], with (130) and $\alpha_{t-1} \le 2^{a_1}\alpha_t$, which holds in both phases and at the transition, where the step size increases, yields

$$\mathbb{E}\Big[\ell^{\Theta}_{\mathrm{C},\infty}\big(p_k, h_\rho(p_k)\big)\Big] \le \frac{D}{\sqrt{\pi\tau_k}} + \frac{2^{1+a_1/2} M_2}{\sqrt{\pi}} \sum_{t=1}^{k-1} \frac{\alpha_t^{3/2}}{\sqrt{\tau_k - \tau_t}} + 6M_2\sqrt{\alpha_k}, \tag{131}$$

where $\tau_k := \sum_{t=1}^{k} \alpha_t(1-\alpha_t)$. It remains to bound the middle sum in the two phases.

Primary phase. Assume $1 \le k \le T$, so $\alpha_t = (t+1)^{-a_1}$. We use

$$\omega_{a_1}(k+1)^{1-a_1} \le \tau_k \le \frac{1}{1-a_1}(k+1)^{1-a_1}. \tag{132}$$

The upper bound is immediate by integral comparison. For the lower bound, $1-\alpha_t \ge 1-2^{-a_1}$ and the standard estimate $\tau_k \ge \tau_1 k^{1-a_1}$ imply the displayed inequality.

Split the middle sum at $\lfloor k/2 \rfloor$. For $t \le \lfloor k/2 \rfloor$,

$$\tau_k - \tau_t \ge \nu_{a_1}\eta_{a_1}(k+1)^{1-a_1}, \tag{133}$$

and therefore

$$\sum_{t=1}^{\lfloor k/2 \rfloor} \frac{\alpha_t^{3/2}}{\sqrt{\tau_k - \tau_t}} \le \frac{S_{3a_1/2}}{\sqrt{\nu_{a_1}\eta_{a_1}}}(k+1)^{-(1-a_1)/2}. \tag{134}$$

For $t > \lfloor k/2 \rfloor$,

$$\alpha_t^{1/2} \le 2^{a_1/2}(k+1)^{-a_1/2}. \tag{135}$$

Using the left-shifted integral comparison

$$\frac{1}{\sqrt{\tau_k - \tau_t}} \le \frac{1}{\alpha_{t+1}(1-\alpha_{t+1})} \int_{\tau_t}^{\tau_{t+1}} \frac{\mathrm{d}\tau}{\sqrt{\tau_k - \tau}}, \tag{136}$$

and the ratio bound

$$\frac{\alpha_t}{\alpha_{t+1}(1-\alpha_{t+1})} \le R_{a_1}, \tag{137}$$

we get

$$\sum_{t=\lfloor k/2 \rfloor + 1}^{k-1} \frac{\alpha_t^{3/2}}{\sqrt{\tau_k - \tau_t}} \le \frac{2^{1+a_1/2} R_{a_1}}{\sqrt{1-a_1}}(k+1)^{-(1-a_1)/2}. \tag{138}$$

Combining these estimates with (131) and (132) gives

$$\mathbb{E}\Big[\ell^{\Theta}_{\mathrm{C},\infty}\big(p_k, h_\rho(p_k)\big)\Big] \le (K_{a_1} + 6M_2)(k+1)^{-(1-a_1)/2}. \tag{139}$$

Since

$$(k+1)^{-(1-a_1)/2} = (k+1)^{-1/6}(k+1)^{\varepsilon_{a_1}}, \tag{140}$$

and (111) makes the minimum equal to $(k+1)^{\varepsilon_{a_1}}$ in this phase, (139) proves the claimed bound for $k \le T$.

Secondary phase. Now let $k > T$, so the tail stepsizes satisfy $\alpha_t = (t+1)^{-2/3}$. For any $t' \ge T$,

$$\tau_k - \tau_{t'} = \sum_{t=t'+1}^{k} \alpha_t(1-\alpha_t) \ge \frac{1}{2}\sum_{t=t'+1}^{k} (t+1)^{-2/3} \ge \frac{3}{2}\big((k+2)^{1/3} - (t'+2)^{1/3}\big). \tag{141}$$

Split the middle sum in (131) at T.

For the tail part $t > T$, we have $\alpha_t^{3/2} = (t+1)^{-1} \le \frac{4}{3}(t+2)^{-1}$. By (141) and the shift $(k,t) \mapsto (k+1, t+1)$, the tail part is at most $\frac{4}{3}$ times the following sum with $k+1$ replaced by $k+2$, which we bound for a generic k:

$$\sum_{t=T+1}^{k-1} \frac{(t+1)^{-1}}{\sqrt{\frac{3}{2}\big((k+1)^{1/3} - (t+1)^{1/3}\big)}} \le \int_T^k \frac{(x+1)^{-1}}{\sqrt{\frac{3}{2}\big((k+1)^{1/3} - (x+1)^{1/3}\big)}}\,\mathrm{d}x. \tag{142}$$

With $u = (x+1)^{1/3}$ and $c = (k+1)^{1/3}$, this integral equals

$$\sqrt{6}\int_{(T+1)^{1/3}}^{(k+1)^{1/3}} \frac{\mathrm{d}u}{u\sqrt{(k+1)^{1/3} - u}} \tag{143}$$

and is bounded by

$$\sqrt{6}\,(k+1)^{-1/6}\log(k+1). \tag{144}$$

To see the last step, write $x = ((k+1)/(T+1))^{1/3} > 1$ in the logarithmic expression obtained after integration:

$$\frac{\sqrt{x}+\sqrt{x-1}}{\sqrt{x}-\sqrt{x-1}} = 2x - 1 + 2\sqrt{x^2-x} \le 4x - 1 \le 3x^3 \le (T+1)x^3 = k+1. \tag{145}$$

Applied with $k+2$ in place of $k+1$, and using $(k+2)^{-1/6} \le (k+1)^{-1/6}$ and $\log(k+2) \le 2\log(k+1)$, this bounds the tail part by

$$\frac{8\sqrt{6}}{3}\,(k+1)^{-1/6}\log(k+1). \tag{146}$$

It remains to control the finitely many primary-phase terms $t \le T$. Evaluating the first two terms in (131) at $T+1$ and using the primary-phase estimate gives

$$\frac{D}{\sqrt{\pi\tau_{T+1}}} + \frac{2^{1+a_1/2}M_2}{\sqrt{\pi}}\sum_{t=1}^{T} \frac{\alpha_t^{3/2}}{\sqrt{\tau_{T+1}-\tau_t}} \le K_{a_1}(T+2)^{-(1-a_1)/2}. \tag{147}$$

If $k+2 \ge 8(T+3)$, then (141) with $t' = T+1$ gives

$$\tau_k - \tau_{T+1} \ge \frac{3}{4}(k+2)^{1/3} \ge \frac{3}{4}(k+1)^{1/3}, \tag{148}$$

and hence these early terms are bounded by

$$K_{a_1}\frac{2}{\sqrt{3}}(k+1)^{-1/6}. \tag{149}$$

If instead $k+2 < 8(T+3)$, then $k+1 < 8(T+3) \le \frac{32}{3}(T+2)$, so

$$K_{a_1}(T+2)^{-(1-a_1)/2} \le K_{a_1}(32/3)^{1/6}(T+2)^{\varepsilon_{a_1}}(k+1)^{-1/6}. \tag{150}$$

Thus, in both cases, the early terms are bounded by

$$K_{a_1} B_T (k+1)^{-1/6}. \tag{151}$$

Combining (146), (151), and the terminal term $6M_2\sqrt{\alpha_k} = 6M_2(k+1)^{-1/3} \le 6M_2(k+1)^{-1/6}$, we obtain

$$\mathbb{E}\Big[\ell^{\Theta}_{\mathrm{C},\infty}\big(p_k, h_\rho(p_k)\big)\Big] \le \left(K_{a_1} B_T + \frac{2^{4+a_1/2} M_2 \sqrt{6}}{3\sqrt{\pi}} + 6M_2\right)(k+1)^{-1/6}\log(k+1). \tag{152}$$

For $k > T$, the definition of the last crossing T gives

$$(k+1)^{\varepsilon_{a_1}} > \kappa_{a_1}\log(k+1), \tag{153}$$

so the minimum in the proposition is $\kappa_{a_1}\log(k+1)$. Since $\kappa_{a_1} \ge 1$, the choice (120) proves the residual bound in the secondary phase as well. The projected-operator residual follows from Corollary 14. □

Remark 28. *As a concrete example, $a_1 = 3/4$ gives $\varepsilon_{a_1} = 1/24$, and the two-phase residual bound becomes*

$$\mathbb{E}\Big[\ell^{\Theta}_{\mathrm{C},\infty}\big(p_k, \mathcal{G}(p_k)\big)\Big] \le \frac{C_{\mathrm{iid}}}{\rho_{\min}}(k+1)^{-1/6}\min\left\{\frac{2^{1/24}}{\log 2}\log(k+1), (k+1)^{1/24}\right\}. \tag{154}$$

We are now in a position to state the proof of Theorem 15.

Proof of Theorem 15. Under the theorem assumptions, the mean-field map of the recursion is h_ρ by Proposition 13. By Lemma 12 the sampled map is samplewise 1-Lipschitz, and by Proposition 13 and Corollary 8, the mean-field map is non-expansive with nonempty fixed-point set. Compactness of $\Delta_d^{\mathcal{S}}$ implies bounded iterates and bounded martingale-difference noise. The almost sure convergence result of Bravo and Cominetti [10, Theorem 2.5 and Example 2.7] therefore applies to h_ρ for every $\frac{2}{3} < a \le 1$.

For the expected residual, Bravo and Cominetti [10, Theorem 4.4] gives the following last-iterate profile for $\frac{1}{2} \le a \le 1$: there exist constants $C_1(a), C_2, C_3(a), C_4 > 0$, depending only on a, $|\mathcal{S}|$ (through the comparison of $\|\cdot\|_{\mathrm{C},\infty}$ with $\|\cdot\|_{\mathrm{C},2}$), the support diameter $\mathrm{diam}_{\ell^{\Theta}_{\mathrm{C},\infty}}(\Delta_d^{\mathcal{S}}) = \sqrt{\theta_d - \theta_1}$, and any uniform bound on $\ell^{\Theta}_{\mathrm{C},\infty}(H(p,y),p)$. Since both $H(p,y)$ and p lie in $\Delta_d^{\mathcal{S}}$, this sampled-update term is itself bounded by the same diameter, so no additional problem-dependent constant is needed in our setting. Thus

$$\mathbb{E}\big[\ell^{\Theta}_{\mathrm{C},\infty}\big(p_k, h_\rho(p_k)\big)\big] \le \begin{cases} C_1(a)(k+1)^{-(a-\frac{1}{2})}, & \frac{1}{2} \le a < \frac{2}{3}, \\ C_2\frac{\log(k+1)}{(k+1)^{1/6}}, & a = \frac{2}{3}, \\ C_3(a)(k+1)^{-\frac{1-a}{2}}, & \frac{2}{3} < a < 1, \\ C_4(\log(k+1))^{-\frac{1}{2}}, & a = 1. \end{cases} \tag{155}$$

The two-phase residual assertion stated in the theorem is Proposition 27. The projected-operator residual bound then follows from Corollary 14. □

C.6 Proof of Theorem 16

We first state a lemma and then a proposition.

Lemma 29 (Markovian inexact-KM reduction)**.** *Consider an SKM recursion $x_k = x_{k-1} + \alpha_k\big(H(x_{k-1}, Y_k) - x_{k-1}\big)$ on a compact convex subset K of a finite-dimensional space with norm $\|\cdot\|$, driven by a finite-state irreducible and aperiodic Markov chain (Y_k), with samplewise non-expansive update map H and non-expansive stationary mean-field map h, both extended to the whole space as in the conventions above. Let the step sizes be either $\alpha_k = (k+1)^{-a}$ with $\frac{4}{5} < a < 1$, or the matched two-phase schedule in Proposition 31. Let*

$$\tau_k := \sum_{t=1}^{k} \alpha_t(1-\alpha_t), \qquad \sigma(0) := 1, \qquad \sigma(u) := \min\{1, (\pi u)^{-1/2}\} \quad (u > 0). \tag{156}$$

There are constants $A_0, A_1, A_2 > 0$ such that the iterates satisfy

$$\mathbb{E}\big[\|x_k - h(x_k)\|\big] \le A_0\sigma(\tau_k)\left(1 + \sum_{t=2}^{k} \alpha_t\omega_{t-1}\right) + A_1 \sum_{t=2}^{k} \alpha_t\sigma(\tau_k - \tau_t)\omega_{t-1} + A_1\omega_k, \quad (157)$$

where the deterministic envelope ω_k obeys

$$\omega_k \le A_2\tau_k\sqrt{\alpha_{k+1}}. \quad (158)$$

Proof. Let $\xi_k := H(x_{k-1}, Y_k) - h(x_{k-1})$, let $\overline{U}_k := (1-\alpha_k)\overline{U}_{k-1} + \alpha_k\xi_k$ with $\overline{U}_0 := 0$, and set $\omega_k := \mathbb{E}\|\overline{U}_k\|$. As in the proof of Blaser and Zhang [7, Theorem 3.1], $z_k := x_k - \overline{U}_k$ satisfies

$$z_k = (1-\alpha_k)z_{k-1} + \alpha_k\Big(h(z_{k-1}) + \big(h(x_{k-1}) - h(z_{k-1})\big)\Big), \quad (159)$$

an inexact Krasnosel'skiĭ–Mann iteration for h whose errors are bounded by $\|\overline{U}_{k-1}\|$. The deterministic inexact Krasnosel'skiĭ–Mann estimate of Bravo and Cominetti [10], applied with errors $e_k = \overline{U}_{k-1}$ exactly as in that proof, combined with $\|x_k - h(x_k)\| \le \|z_k - h(z_k)\| + 2\|\overline{U}_k\|$ and taken in expectation, gives (157). This step holds for any step sizes in $(0,1)$ and uses no martingale structure, so the Markovian dependence and the schedule enter only through ω_k and τ_k. Lemma 30 below gives $\omega_k \le A\sqrt{\alpha_k}$ for both schedules. Since $\tau_k \ge \tau_1 > 0$ and $\alpha_k \le 2\alpha_{k+1}$, this implies (158). Proposition 31 below then bounds the remaining critical-tail sums in (157). □

Lemma 30 (Markovian averaged noise)**.** *In the setting of Lemma 29, let $\|\cdot\|_2$ be a Euclidean norm with $\|\cdot\| \le \|\cdot\|_2 \le c_E\|\cdot\|$, and suppose that the step sizes $\alpha_k \in (0,1)$ are non-increasing and satisfy $\alpha_k^{-1} - \alpha_{k-1}^{-1} \le \bar{a}$ for all $k \ge 2$ and some $\bar{a} < 1$. Then there is a constant A, depending only on $\bar{a}$, α_1, c_E, the diameter of K, and the fundamental matrix of (Y_k), such that $\mathbb{E}\|\overline{U}_k\|_2 \le A\sqrt{\alpha_k}$ for all $k \ge 1$.*

Both schedules in Lemma 29 qualify. For $\alpha_k = (k+1)^{-a}$, $(k+1)^a - k^a \le a < 1$. For the two-phase schedule, the first phase gives at most a_1, the transition gives 0 because $\alpha_{T+1} = \alpha_T$, and in the tail $\gamma_T^{-1}\big((k+1)^{4/5} - k^{4/5}\big) \le \frac{4}{5}\gamma_T^{-1}(T+1)^{-1/5} \le \frac{4}{5}(T+1)^{a_1-1} < a_1$. For the categorical recursion, $c_E = \sqrt{|\mathcal{S}|}$. For the coupled recursion with $\|(u,g)\|_2 := \sqrt{2}\,(\|u\|_{\mathrm{C},2}^2 + \lambda^2 g^2)^{1/2}$, $c_E = \sqrt{2|\mathcal{S}|}$.

Proof. Let Q_Y be the transition kernel of (Y_k) and π_Y its stationary law, so $h = \sum_y \pi_Y(y)H(\cdot, y)$. Since (Y_k) is finite, irreducible and aperiodic, its fundamental matrix $Z_Y := (I - Q_Y + \mathbf{1}\pi_Y^\top)^{-1}$ exists, $Z_Y - \mathbf{1}\pi_Y^\top = \sum_{n\ge 0}(Q_Y^n - \mathbf{1}\pi_Y^\top)$ has zero row sums, $\kappa_Y := \|Z_Y - \mathbf{1}\pi_Y^\top\|_\infty < \infty$, and

$$u(x,y) := \sum_{y'} \big(Z_Y - \mathbf{1}\pi_Y^\top\big)(y,y')\,H(x,y') \quad (160)$$

solves the Poisson equation $u - Q_Y u = H - h$. Because the rows of $Z_Y - \mathbf{1}\pi_Y^\top$ sum to zero and H takes values in K, $\|u\|_2 \le \kappa_Y \operatorname{diam}_2(K) =: B$, and the same holds for $Q_Y u$. Because each $H(\cdot, y')$ is non-expansive in $\|\cdot\|$, both $u(\cdot, y)$ and $Q_Y u(\cdot, y)$ are Lipschitz from $\|\cdot\|$ to $\|\cdot\|_2$ with constant $L := c_E\kappa_Y$.

With $\mathcal{F}_k := \sigma(Y_0, \dots, Y_k)$, the point x_{k-1} is $\mathcal{F}_{k-1}$-measurable, and

$$\xi_k = m_k + (r_{k-1} - r_k) + e_k, \quad (161)$$

where $m_k := u(x_{k-1}, Y_k) - (Q_Y u)(x_{k-1}, Y_{k-1})$, $r_k := (Q_Y u)(x_k, Y_k)$ and $e_k := (Q_Y u)(x_k, Y_k) - (Q_Y u)(x_{k-1}, Y_k)$. Here $\mathbb{E}[m_k \mid \mathcal{F}_{k-1}] = 0$, $\|m_k\|_2 \le 2B$, $\|r_k\|_2 \le B$, and $\|e_k\|_2 \le L\|x_k - x_{k-1}\| \le LD\alpha_k$ with D the $\|\cdot\|$-diameter of K.

Write $\overline{U}_k = \sum_{t=1}^k w_{k,t}\xi_t$ with $w_{k,t} := \alpha_t \prod_{r=t+1}^k (1-\alpha_r)$, and let $v_k := \sum_t w_{k,t}^2$ and $s_k := \sum_t w_{k,t}\alpha_t$. Then $v_k = (1-\alpha_k)^2 v_{k-1} + \alpha_k^2$ and $s_k = (1-\alpha_k)s_{k-1} + \alpha_k^2$. As in the proof of Proposition 27, $v_k \le \alpha_k$. Moreover, $s_k \le \alpha_k/(1-\bar{a})$ by induction: $s_1 = \alpha_1^2 \le \alpha_1$, and

$$\alpha_k - (1-\alpha_k)\alpha_{k-1} = \alpha_k\alpha_{k-1}\big(1 - (\alpha_k^{-1} - \alpha_{k-1}^{-1})\big) \ge (1-\bar{a})\alpha_k^2. \quad (162)$$

We bound the three parts of $\overline{U}_k$ separately.

For the martingale part, orthogonality of martingale differences gives $\mathbb{E}\|\sum_t w_{k,t} m_t\|_2^2 \le 4B^2 v_k \le 4B^2\alpha_k$.

For the Lipschitz part, $\|\sum_t w_{k,t} e_t\|_2 \le LD\, s_k \le LD\alpha_k/(1-\bar a)$.

For the telescoping part, summation by parts gives

$$\sum_{t=1}^{k} w_{k,t}(r_{t-1}-r_t) = w_{k,1}r_0 - \alpha_k r_k + \sum_{t=1}^{k-1}(w_{k,t+1}-w_{k,t})\, r_t. \tag{163}$$

Here $w_{k,1} \le s_k/\alpha_1$. For $t<k$, $w_{k,t+1}-w_{k,t} = \prod_{r=t+2}^{k}(1-\alpha_r)\,(\alpha_{t+1}-\alpha_t+\alpha_t\alpha_{t+1})$. Since $0 \le \alpha_t - \alpha_{t+1} = \alpha_t\alpha_{t+1}(\alpha_{t+1}^{-1}-\alpha_t^{-1}) \le \bar a\,\alpha_t\alpha_{t+1}$, the last factor lies in $[0,\alpha_t\alpha_{t+1}]$, and $\alpha_t \le (1+\bar a)\alpha_{t+1} \le 2\alpha_{t+1}$. Hence $\sum_{t<k}|w_{k,t+1}-w_{k,t}| \le 2s_k$, and the telescoping part has norm at most $B\big(1+(2+\alpha_1^{-1})/(1-\bar a)\big)\alpha_k$.

Combining the three parts and using $\alpha_k \le \sqrt{\alpha_k}$ proves the claim with $A := 3B + \big(LD + B(2+\alpha_1^{-1})\big)/(1-\bar a)$. □

Lemma 30 is sharper than the envelope $\omega_k = O(\tau_k\sqrt{\alpha_{k+1}})$ of Blaser and Zhang [7, Lemma C.4]. Their analysis allows unbounded iterates, so the Poisson solution can grow like τ_k along the trajectory. On the compact set K it is uniformly bounded, which removes this factor.

Proposition 31 (Two-phase Markovian residual refinement)**.** *Under the assumptions of Theorem 16, fix $a_1 \in \big(\frac{4}{5},1\big)$ and set*

$$\varepsilon_{a_1} := \frac{5a_1-4}{10}, \qquad \kappa_{a_1} := \frac{2^{\varepsilon_{a_1}}}{\log 2}. \tag{164}$$

Let T be the largest integer such that

$$(k+1)^{\varepsilon_{a_1}} \le \kappa_{a_1}\log(k+1), \qquad 1 \le k \le T. \tag{165}$$

Set $\gamma_T := (T+1)^{-a_1}(T+2)^{4/5}$, and use the two-phase step size

$$\alpha_k = \begin{cases} (k+1)^{-a_1}, & 1 \le k \le T, \\ \gamma_T (k+1)^{-4/5}, & k > T. \end{cases} \tag{166}$$

Then there exists $C_{\mathrm{mk}} > 0$ such that, for all $k \ge 1$,

$$\mathbb{E}\Big[\ell^{\Theta}_{\mathrm{C},\infty}\big(p_k, h_\mu(p_k)\big)\Big] \le C_{\mathrm{mk}}(k+1)^{-1/10}\min\Big\{\kappa_{a_1}\log(k+1), (k+1)^{\varepsilon_{a_1}}\Big\}. \tag{167}$$

Consequently,

$$\mathbb{E}\Big[\ell^{\Theta}_{\mathrm{C},\infty}\big(p_k, \mathcal{G}(p_k)\big)\Big] \le \frac{C_{\mathrm{mk}}}{\mu_{\min}}(k+1)^{-1/10}\min\Big\{\kappa_{a_1}\log(k+1), (k+1)^{\varepsilon_{a_1}}\Big\}. \tag{168}$$

Proof. The threshold is well defined by the same argument as in Proposition 27: the function $x^{\varepsilon_{a_1}}/\log x$ decreases from its value at $x=2$ until $x = \exp(1/\varepsilon_{a_1})$ and then increases to infinity. Thus the set of integers satisfying (165) is a nonempty initial interval. The factor γ_T makes the schedule continuous at the transition, since $\alpha_{T+1} = \alpha_T$, and hence non-increasing.

For $k \le T$, the recursion has only used the polynomial schedule $\alpha_t = (t+1)^{-a_1}$. The polynomial Markovian residual bound of Blaser and Zhang [7, Theorem 3.1] therefore gives

$$\mathbb{E}\Big[\ell^{\Theta}_{\mathrm{C},\infty}\big(p_k, h_\mu(p_k)\big)\Big] \le C(k+1)^{-(1-a_1)/2}. \tag{169}$$

Since

$$(k+1)^{-(1-a_1)/2} = (k+1)^{-1/10}(k+1)^{\varepsilon_{a_1}}, \tag{170}$$

and the minimum in the proposition is $(k+1)^{\varepsilon_{a_1}}$ in this phase, the desired bound follows for $k \le T$.

For $k > T$, we apply Lemma 29 in the cumulative-coordinate norm that represents the block-sup Cramér metric. Since T is fixed once a_1 is fixed, the two-phase schedule gives constants $0 < c_\tau < C_\tau < \infty$ such that, for all $k > T$,

$$c_\tau(k+1)^{1/5} \le \tau_k \le C_\tau(k+1)^{1/5}. \tag{171}$$

Indeed, the tail contribution is a constant multiple of $\sum_{t>T}(t+1)^{-4/5}$, and the finitely many transition values are absorbed into c_τ and C_τ. Combining (171) with (158) yields

$$\omega_k \le C(k+1)^{-1/5}. \tag{172}$$

Consequently,

$$\sum_{t=2}^{k} \alpha_t \omega_{t-1} \le C\left(1 + \sum_{t=T+1}^{k} (t+1)^{-1}\right) \le C\log(k+1). \tag{173}$$

The first term in (157) is therefore bounded by $C(k+1)^{-1/10}\log(k+1)$, and the final correction term (172) is smaller than this bound.

The convolution term in (157) is handled by separating the finite prefix from the critical tail. The terms with $t \le T$ contribute at most $C(k+1)^{-1/10}$ after increasing C, again using (171). The finitely many near-transition cases are absorbed into the same constant. For the tail terms $T < t < k$, (172) gives $\alpha_t\omega_{t-1} \le C(t+1)^{-1}$, and the tail lower bound gives

$$\tau_k - \tau_t \ge c\big((k+1)^{1/5} - (t+1)^{1/5}\big). \tag{174}$$

Thus

$$\sum_{t=T+1}^{k-1} \alpha_t \sigma(\tau_k - \tau_t)\omega_{t-1} \le C \sum_{t=T+1}^{k-1} \frac{(t+1)^{-1}}{\sqrt{(k+1)^{1/5} - (t+1)^{1/5}}}. \tag{175}$$

By the same integral comparison used in the i.i.d. proof, with the change of variables $u = (x+1)^{1/5}$, the last display is at most

$$C\int_{(T+1)^{1/5}}^{(k+1)^{1/5}} \frac{\mathrm{d}u}{u\sqrt{(k+1)^{1/5} - u}} \le C(k+1)^{-1/10}\log(k+1). \tag{176}$$

The omitted $t = k$ endpoint is bounded by $C(k+1)^{-1}$ and is therefore harmless. Combining these bounds in (157) gives

$$\mathbb{E}\Big[\ell^{\Theta}_{\mathrm{C},\infty}\big(p_k, h_\mu(p_k)\big)\Big] \le C(k+1)^{-1/10}\log(k+1), \qquad k > T. \tag{177}$$

For $k > T$, the definition of the largest crossing T implies

$$(k+1)^{\varepsilon_{a_1}} > \kappa_{a_1}\log(k+1), \tag{178}$$

so the minimum in the proposition is $\kappa_{a_1}\log(k+1)$. Since $\kappa_{a_1} \ge 1$, enlarging the constant proves the claimed residual bound in the secondary phase. The projected-operator residual bound follows from Corollary 14. □

We are now in a position to state the proof of Theorem 16.

Proof of Theorem 16. Because $\mathcal{S}$ and $\mathcal{A}$ are finite and the reward function $R : \mathcal{S}\times\mathcal{A} \to [0,1]$ is deterministic, the centered reward B_k is finite even after conditioning only on (S_k, S_{k+1}). Hence the feasible centered sample set

$$\{(i,b,j) \in \mathcal{S}\times\mathbb{R}\times\mathcal{S} : \mathbb{P}(S_k = i, B_k = b, S_{k+1} = j) > 0\} \tag{179}$$

is finite. The process $(Y_k)_{k\ge0}$ is Markov because the conditional law of (B_{k+1}, S_{k+2}) depends only on S_{k+1}, which is the last coordinate of $Y_k = (S_k, B_k, S_{k+1})$. It is irreducible and aperiodic on its feasible support because from any feasible triple (i,b,j) to any other feasible triple (i',b',j'), one may first follow the irreducible and aperiodic state chain from j to i' and then realize the feasible transition producing (b',j'). By Lemma 12, the sampled map is samplewise non-expansive, and by Proposition 13 and Corollary 8, the averaged operator is h_μ, which is non-expansive with non-empty fixed-point set. The hypotheses of Blaser and Zhang [7] are therefore satisfied, giving the fixed-exponent a.s. convergence assertion and the residual bound $C/\sqrt{\tau_k}$ with $\tau_k = \sum_{t=1}^{k}\alpha_t(1-\alpha_t)$. The two-phase residual assertion is Proposition 31. The projected residual bounds once again follow from Corollary 14. □

C.7 Proof of Proposition 17

Proof of Proposition 17. Consider a one-state deterministic MDP. The transition matrix is (1), the reward is almost surely r, and therefore $\bar{r}^\pi = r$. For any categorical state $p \in \Delta_d$, the centered update uses the shift $r - \bar{r}^\pi = 0$, so $\mathcal{G}(p) = L_0 p = p$. Hence, $\mathcal{G} = \mathrm{Id}$. Now replace the centered reward by the uncentered reward in the one-state sampled update. The resulting update sends p to $L_r p$. If the support has at least two atoms and $r \neq 0$, then $L_r p \neq p$ in general. For example if p is a point mass at an interior atom, translating by r and projecting back to the same support changes the cumulative coordinates. Therefore, the uncentered update is not an unbiased sample of $\mathcal{G}$, and in general it does not define an SA of the projected quotient operator. □

D Proofs of Section 6

For convenience of notation, let us write

$$\widehat{H}\big((p,g),(s,r,s')\big) := \big(H_g(p,(s,r,s')),r\big) \tag{180}$$

and

$$\hat{h}_\mu(p,g) := \big(h_\mu^{(g)}(p),\bar{r}^\pi\big). \tag{181}$$

This notation is more intuitive in making the gain estimate g explicitly an argument of the recursion.

Lemma 32. *For every $b,b' \in \mathbb{R}$ and all $p,q \in \Delta_d$,*

$$\ell_{\mathrm{C}}^{\Theta}(L_b p, L_{b'} q) \le \ell_{\mathrm{C}}^{\Theta}(p,q) + \varsigma^{-1/2}|b-b'|. \tag{182}$$

Proof. By the triangle inequality and the non-expansiveness of $L_{b'}$,

$$\ell_{\mathrm{C}}^{\Theta}(L_b p, L_{b'} q) \le \ell_{\mathrm{C}}^{\Theta}(L_b p, L_{b'} p) + \ell_{\mathrm{C}}^{\Theta}(L_{b'} p, L_{b'} q) \le \ell_{\mathrm{C}}^{\Theta}(L_b p, L_{b'} p) + \ell_{\mathrm{C}}^{\Theta}(p,q). \tag{183}$$

It therefore suffices to bound $\ell_{\mathrm{C}}^{\Theta}(L_b p, L_{b'} p)$. Write $p = \sum_{m=1}^{d} p_m e_m$, where e_m is the mth canonical basis vector. Since scalar categorical projection is linear on mixtures,

$$L_t p = \sum_{m=1}^{d} p_m L_t e_m \qquad \text{for every } t \in \mathbb{R}. \tag{184}$$

Applying triangle inequality to the vector of cumulative masses yields

$$\ell_{\mathrm{C}}^{\Theta}(L_b p, L_{b'} p) \le \sum_{m=1}^{d} p_m \ell_{\mathrm{C}}^{\Theta}(L_b e_m, L_{b'} e_m). \tag{185}$$

Fix m and define

$$v_m(t) := \big(F_{L_t e_m}(\theta_k)\big)_{k=1}^{d-1} \in \mathbb{R}^{d-1}. \tag{186}$$

Since $\ell_{\mathrm{C}}^{\Theta}(u,v) = \sqrt{\varsigma}\big\|(F_u(\theta_k) - F_v(\theta_k))_{k=1}^{d-1}\big\|_2$, it is enough to control $t \mapsto v_m(t)$. Let $x = \theta_m + t$ be the translated atom location. If $x \le \theta_1$, then $L_t e_m = e_1$ and $v_m(t) = \mathbf{1}$. If $x \ge \theta_d$, then $L_t e_m = e_d$ and $v_m(t) = 0$. If instead $x \in [\theta_k, \theta_{k+1}]$ for some $k \in \{1,\dots,d-1\}$, scalar categorical projection places mass

$$\phi_k(x) := \frac{\theta_{k+1} - x}{\varsigma} \tag{187}$$

on θ_k and mass $1 - \phi_k(x)$ on θ_{k+1}. Hence

$$v_m(t) = \big(\underbrace{0,\dots,0}_{k-1 \text{ entries}}, \phi_k(x), \underbrace{1,\dots,1}_{d-1-k \text{ entries}}\big), \tag{188}$$

so on this interval v_m is affine and

$$\frac{\mathrm{d}}{\mathrm{d}t} v_m(t) = -\frac{1}{\varsigma} e_k. \tag{189}$$

Therefore $\big\|\frac{\mathrm{d}}{\mathrm{d}t} v_m(t)\big\|_2 \le \varsigma^{-1}$ wherever the derivative exists. Integrating this bound and multiplying by $\sqrt{\varsigma}$ from the definition of $\ell_{\mathrm{C}}^{\Theta}$ gives

$$\ell_{\mathrm{C}}^{\Theta}(L_b e_m, L_{b'} e_m) \le \varsigma^{-1/2}|b-b'|. \tag{190}$$

Substituting this into (185) yields

$$\ell_{\mathrm{C}}^{\Theta}(L_b p, L_{b'} p) \le \varsigma^{-1/2}|b-b'| \tag{191}$$

□

Lemma 33. *For all $g,g' \in \mathbb{R}$ and all $p,q \in \Delta_d^{\mathcal{S}}$,*

$$\ell_{\mathrm{C},\infty}^{\Theta}\big(\mathcal{G}_g(p), \mathcal{G}_{g'}(q)\big) \le \ell_{\mathrm{C},\infty}^{\Theta}\big(p,q\big) + \varsigma^{-1/2}|g-g'|. \tag{192}$$

Proof. Fix $i \in \mathcal{S}$. By Proposition 11,

$$\mathcal{G}_g(p)_i = \sum_{j\in\mathcal{S}} P^\pi_{ij}\mathbb{E}[L_{R_{ij}-g}p_j], \qquad \mathcal{G}_{g'}(q)_i = \sum_{j\in\mathcal{S}} P^\pi_{ij}\mathbb{E}[L_{R_{ij}-g'}q_j]. \tag{193}$$

For each j, let

$$u_j := \big(F_{L_{R_{ij}-g}p_j}(\theta_k) - F_{L_{R_{ij}-g'}q_j}(\theta_k)\big)_{k=1}^{d-1} \in \mathbb{R}^{d-1}. \tag{194}$$

Since cumulative masses are linear in the coefficient vector, using triangle inequality,

$$\begin{aligned}
\ell^\Theta_{\mathrm{C}}\big(\mathcal{G}_g(p)_i, \mathcal{G}_{g'}(q)_i\big) &\le \sqrt{\varsigma}\Big\|\sum_{j\in\mathcal{S}} P^\pi_{ij}\mathbb{E}[u_j]\Big\|_2 \\
&\le \sum_{j\in\mathcal{S}} P^\pi_{ij}\mathbb{E}\big[\ell^\Theta_{\mathrm{C}}(L_{R_{ij}-g}p_j, L_{R_{ij}-g'}q_j)\big] \\
&\le \sum_{j\in\mathcal{S}} P^\pi_{ij}\mathbb{E}\big[\ell^\Theta_{\mathrm{C}}(p_j, q_j) + \varsigma^{-1/2}|g-g'|\big] \\
&\le \ell^\Theta_{\mathrm{C},\infty}(p,q) + \varsigma^{-1/2}|g-g'|.
\end{aligned} \tag{195}$$

Taking the maximum over $i \in \mathcal{S}$ proves the result. □

D.1 Proof of Proposition 18

Proof of Proposition 18. Let $y = (s, r, s')$. The second coordinate of $\widehat{H}$ is the common reward sample r, so $\lambda|r - r| = 0$. For the first coordinate, only the block s is updated, and Lemma 32 gives

$$\ell^\Theta_{\mathrm{C}}\big(H_g(p,y)_s, H_{g'}(q,y)_s\big) = \ell^\Theta_{\mathrm{C}}(L_{r-g}p_{s'}, L_{r-g'}q_{s'}) \le \ell^\Theta_{\mathrm{C}}(p_{s'}, q_{s'}) + \varsigma^{-1/2}|g-g'|. \tag{196}$$

Taking the maximum over states,

$$\ell^\Theta_{\mathrm{C},\infty}\big(H_g(p,y), H_{g'}(q,y)\big) \le \ell^\Theta_{\mathrm{C},\infty}(p,q) + \varsigma^{-1/2}|g-g'|. \tag{197}$$

Since $\lambda \ge \varsigma^{-1/2}$ by definition,

$$d_\lambda\big(\widehat{H}\big((p,g),y\big), \widehat{H}\big((q,g'),y\big)\big) \le \ell^\Theta_{\mathrm{C},\infty}(p,q) + \varsigma^{-1/2}|g-g'| \le d_\lambda\big((p,g),(q,g')\big). \tag{198}$$

□

D.2 Proof of Proposition 19

Proof of Proposition 19. The expectation identity follows from Proposition 11 applied under the stationary one-step law:

$$\mathbb{E}\big[H_g(p,Y)\big] = h^{(g)}_\mu(p), \qquad \mathbb{E}[R] = \bar{r}^\pi. \tag{199}$$

For non-expansiveness, use the same cumulative-distribution argument as in the proof of Proposition 13. Since

$$h^{(g)}_\mu(p)_i = (1-\mu_i)p_i + \mu_i\mathcal{G}_g(p)_i, \qquad h^{(g')}_\mu(q)_i = (1-\mu_i)q_i + \mu_i\mathcal{G}_{g'}(q)_i, \tag{200}$$

we have for each $k = 1, \dots, d-1$,

$$F_{h^{(g)}_\mu(p)_i}(\theta_k) - F_{h^{(g')}_\mu(q)_i}(\theta_k) = (1-\mu_i)\big(F_{p_i}(\theta_k) - F_{q_i}(\theta_k)\big) + \mu_i\big(F_{\mathcal{G}_g(p)_i}(\theta_k) - F_{\mathcal{G}_{g'}(q)_i}(\theta_k)\big). \tag{201}$$

Triangle inequality gives, for each $i \in \mathcal{S}$,

$$\begin{aligned}
\ell^\Theta_{\mathrm{C}}\big(h^{(g)}_\mu(p)_i, h^{(g')}_\mu(q)_i\big) &\le (1-\mu_i)\ell^\Theta_{\mathrm{C}}(p_i,q_i) + \mu_i\ell^\Theta_{\mathrm{C}}\big(\mathcal{G}_g(p)_i, \mathcal{G}_{g'}(q)_i\big) \\
&\le (1-\mu_i)\ell^\Theta_{\mathrm{C},\infty}(p,q) + \mu_i\big(\ell^\Theta_{\mathrm{C},\infty}(p,q) + \varsigma^{-1/2}|g-g'|\big) \\
&\le \ell^\Theta_{\mathrm{C},\infty}(p,q) + \varsigma^{-1/2}|g-g'|.
\end{aligned} \tag{202}$$

Taking the maximum over $i \in \mathcal{S}$ yields

$$\ell^\Theta_{\mathrm{C},\infty}\big(h^{(g)}_\mu(p), h^{(g')}_\mu(q)\big) \le \ell^\Theta_{\mathrm{C},\infty}(p,q) + \varsigma^{-1/2}|g-g'|, \tag{203}$$

and therefore

$$d_\lambda\big(\hat{h}_\mu(p,g), \hat{h}_\mu(q,g')\big) \le \ell^\Theta_{\mathrm{C},\infty}(p,q) + \varsigma^{-1/2}|g-g'| \le d_\lambda\big((p,g),(q,g')\big). \tag{204}$$

For the statement about fixed points, if $(p,g) \in \mathrm{fix}(\hat{h}_\mu)$, then necessarily $g = \bar{r}^\pi$. The first coordinate then reads $p = h_\mu(p)$, and $\mu_{\min} > 0$ implies $p \in \mathrm{fix}(\mathcal{G})$ by Proposition 13. The converse implication is immediate. □

D.3 Proof of Corollary 20

Proof of Corollary 20. For $i \in \mathcal{S}$,

$$\mu_i \ell_{\mathrm{C}}^{\Theta}\big(p_i, \mathcal{G}_g(p)_i\big) = \ell_{\mathrm{C}}^{\Theta}\big(p_i, h_\mu^{(g)}(p)_i\big). \tag{205}$$

Also,

$$\ell_{\mathrm{C}}^{\Theta}\big(\mathcal{G}_g(p)_i, \mathcal{G}(p)_i\big) \le \varsigma^{-1/2}|g - \bar{r}^{\pi}| \tag{206}$$

by Lemma 33 with $p = q$. Hence

$$\mu_i \ell_{\mathrm{C}}^{\Theta}\big(p_i, \mathcal{G}(p)_i\big) \le \ell_{\mathrm{C}}^{\Theta}\big(p_i, h_\mu^{(g)}(p)_i\big) + \varsigma^{-1/2}|g - \bar{r}^{\pi}|. \tag{207}$$

Taking the maximum over $i \in \mathcal{S}$,

$$\mu_{\min} \ell_{\mathrm{C},\infty}^{\Theta}\big(p, \mathcal{G}(p)\big) \le \ell_{\mathrm{C},\infty}^{\Theta}\big(p, h_\mu^{(g)}(p)\big) + \varsigma^{-1/2}|g - \bar{r}^{\pi}| \le d_\lambda\big((p,g), \hat{h}_\mu(p,g)\big), \tag{208}$$

where the last step uses $\lambda \ge \varsigma^{-1/2}$. The gain-error bound is immediate from the definition of d_λ. □

D.4 Proof of Theorem 21

Proof of Theorem 21. Let $z_k = (p_k, g_k) \in \mathcal{Z}$. The coupled SKM recursion (33) can be written as

$$z_{k+1} = z_k + \alpha_k\big(\widehat{H}(z_k, Y_k) - z_k\big). \tag{209}$$

View each categorical block p_i through its cumulative masses $(F_{p_i}(\theta_1), \dots, F_{p_i}(\theta_{d-1}))$, so that $\ell_{\mathrm{C},\infty}^{\Theta}$ becomes the block-sup norm of these cumulative-coordinate vectors. On the augmented coordinates (u, g), use the product norm $\|(u,g)\|_\lambda := \|u\|_{\mathrm{C},\infty} + \lambda|g|$. Under this identification, d_λ is the induced distance on the compact convex set $\mathcal{Z}$. By Proposition 18, the sampled map is samplewise non-expansive in this norm. By Proposition 19 and Corollary 8, the mean-field map induced by the stationary one-step law is non-expansive with a non-empty set of fixed-points. Since rewards are finite-valued, the uncentered chain $(Y_k)_{k\ge0}$ is finite-state on its feasible support. The fixed-exponent Markovian theorem of Blaser and Zhang [7] therefore applies. The final two bounds then follow from Corollary 20. The two-phase residual bound follows by applying Proposition 31's argument on the compact product space $(\mathcal{Z}, d_\lambda)$, with $\widehat{H}$ in place of H and $\hat{h}_\mu$ in place of h_μ. The only constants that change are the product-space diameter and the finite-state Poisson constant of the uncentered chain $(Y_k)_{k\ge0}$. Corollary 20 then gives the projected-operator and gain-error bounds. □

E Synchronous gain-free structure

The previous section handles raw rewards by estimating the gain online. We close with a complementary exact-law observation: if one forms a full synchronous sampled backup before choosing categorical representatives, then the centering constant disappears already at the quotient level. This is a stronger, samplewise invariance statement than the operator-level invariance of Proposition 3, but it lives in the idealized space of quotient distributions rather than in the practical fixed-grid categorical representation. To make the statement precise, let

$$\mathbf{Y} = \big((S'_i, R_i)\big)_{i\in\mathcal{S}} \tag{210}$$

be a synchronous one-step sample such that, for each $i \in \mathcal{S}$,

$$\mathbb{P}(S'_i = j) = P^\pi_{ij}, \qquad R_i \mid (S'_i = j) \sim \mathrm{Law}(R_{ij}), \tag{211}$$

with the coordinates sampled independently across i. Thus $\mathbf{Y}$ contains one sampled successor and reward for every state. For $\eta \in \mathcal{F}^{\mathcal{S}}$ and $g \in \mathbb{R}$, define the exact synchronous sample map by

$$\big(\mathcal{H}_g(\eta, \mathbf{Y})\big)_i := (\tau_{R_i - g})_\# \eta_{S'_i}, \qquad i \in \mathcal{S}. \tag{212}$$

Changing g changes each centered reward $R_i - g$ by the same additive amount. Since the update is synchronous, that same shift appears in every output block.

Proposition 34. *For every synchronous sample $\mathbf{Y}$, every $g, g' \in \mathbb{R}$, and every $\eta \in \mathcal{F}^{\mathcal{S}}$,*

$$\mathcal{H}_{g'}(\eta, \mathbf{Y}) = (\tau_{g-g'})_\# \mathcal{H}_g(\eta, \mathbf{Y}). \tag{213}$$

Consequently,

$$[\mathcal{H}_g(\eta, \mathbf{Y})] = [\mathcal{H}_{g'}(\eta, \mathbf{Y})] \quad \textit{in } \mathcal{F}^{\mathcal{S}}/\approx. \tag{214}$$

Thus the exact synchronous quotient sample is independent of the centering constant.

Proof. For every state i, $R_i - g' = (R_i - g) + (g - g')$, so $(\tau_{R_i-g'})_\# \eta_{S'_i} = (\tau_{g-g'})_\# \big((\tau_{R_i-g})_\# \eta_{S'_i}\big)$. The translation $(\tau_{g-g'})_\#$ is common to every state block, so the two output families represent the same element in the equivalence class. □

Taking the representative with $g = \bar{r}^\pi$ connects this gain-free quotient sample to the exact centered operator:

$$\mathbb{E}[\mathcal{H}_{\bar{r}^\pi}(\eta, \mathbf{Y})] = \mathcal{T}\eta. \tag{215}$$

To see this, condition on the successor state $S'_i = j$. The ith block of $\mathcal{H}_{\bar{r}^\pi}(\eta, \mathbf{Y})$ is then $(\tau_{R_i - \bar{r}^\pi})_\# \eta_j$ whose conditional law is $\nu_{ij} * \eta_j$. Averaging over j with weights P^π_{ij} gives exactly the ith component of $\mathcal{T}\eta$ from (2).

Thus synchronous exact samples provide a gain-free sample representation of the quotient-law target. This does not contradict the need for the online-gain categorical recursion above: once we choose a fixed categorical grid, projecting a translated representative can change the coefficients, so the exact quotient symmetry is not automatically preserved by a practical fixed-grid categorical update.

F Five-state experimental details

For the convergence experiment in Figure 2, we use deterministic state rewards $r = (0.90, 0.10, 0.55, 0.80, 0.05)$ and

$$P^\pi = \begin{bmatrix} .10 & .45 & .25 & .15 & .05 \\ .20 & .20 & .35 & .15 & .10 \\ .10 & .15 & .30 & .35 & .10 \\ .15 & .10 & .25 & .25 & .25 \\ .30 & .15 & .10 & .15 & .30 \end{bmatrix}. \tag{216}$$

The categorical grid is $[-5, 5]$ with $d = 321$ and $\varsigma = 0.03125$. Every method runs for 10^7 iterations over ten seeds from the uniform initialization $p_{0,i} = (1/d, \dots, 1/d)$. The i.i.d. recursions sample S_k uniformly ($\rho_i = 1/5$), the Markovian and coupled recursions start from $S_0 \sim \mu$, and the coupled recursion starts from $g_0 = 0$. The wrong-centering ablation uses i.i.d. samples and the i.i.d. schedule with $g = 0$. The exact KM exponent is 0.75. The i.i.d. recursion uses the two-phase schedule of Proposition 27 with $a_1 = 0.75$. The theoretically prescribed phase switch occurs beyond the plotted horizon. The Markovian and coupled recursions use fixed exponent 0.85. Curves show medians and bands show interquartile ranges. The empirical rates quoted for Figure 2(a) are least-squares slopes of the log median residual against $\log k$ over $k \geq 10^4$: -0.33 (i.i.d.), -0.40 (Markovian), and -0.38 (coupled). The state–atom kernel in (74) has 1605 states, one closed class of size 1557, and 48 transient singleton classes, so Theorem 24 implies that the categorical fixed point is unique.

For Figure 3, we instead use rewards $(0.95, 0.78, 0.50, 0.22, 0.05)$ and

$$P^\pi = \begin{bmatrix} .78 & .17 & .03 & .01 & .01 \\ .18 & .72 & .07 & .02 & .01 \\ .04 & .08 & .76 & .08 & .04 \\ .01 & .02 & .07 & .72 & .18 \\ .01 & .01 & .03 & .17 & .78 \end{bmatrix}. \tag{217}$$

The large self-transition probabilities keep the chain in high- or low-reward regions long enough for the transient return to depend strongly on the starting state. We use geometric continuation probability $\gamma = 0.9$, a 401-atom grid on $[-20, 20]$, and $100{,}000$ exact Monte Carlo returns per starting state. The largest empirical overflow probability is 3×10^{-4}, the largest categorical endpoint mass is 2.7×10^{-4}, and the maximum Cramér discrepancy between the recursive categorical fixed point and the projected Monte Carlo law is 0.012.

G Neural function approximation results

We consider fixed-policy evaluation on `Pendulum-v1` [12], with observations $(\cos\theta, \sin\theta, \dot\theta)$, torque $a \in [-2, 2]$, and raw reward

$$r_{\mathrm{raw}}(\theta, \dot\theta, a) = -\theta^2 - 0.1\dot\theta^2 - 0.001a^2, \tag{218}$$

normalized to $[0, 1]$ by $(r_{\mathrm{raw}} + 16.2736044)/16.2736044$. The fixed stochastic energy-based policy is

$$a_t = \mathrm{clip}\Big(K_E \dot\theta_t \cos\theta_t + K_D \dot\theta_t + \sigma\epsilon_t, -2, 2\Big), \qquad \epsilon_t \sim \mathcal{N}(0, 1), \tag{219}$$

with $K_E = 1$, $K_D = 0.2$ and $\sigma = 0.7$.

The average-reward theory concerns one Markov chain, so all data come from a single trajectory whose stored transitions are exactly that chain's transitions. We remove the 200-step time limit and, after each step, restart from the environment's initial law ($\theta \sim U[-\pi, \pi]$, $\dot\theta \sim U[-1, 1]$) with probability $1/200$, keeping the usual mean episode length. A restart is part of the stored transition and of the simulator used for the residual estimates. Without restarts the policy settles near the hanging position and the problem becomes nearly trivial. With the deterministic 200-step reset that `Pendulum-v1` uses by default [12], the stored transitions, the state distribution and the Monte Carlo gain would refer to different chains.

Each categorical critic is a two-hidden-layer ReLU network of widths $(256, 256)$ with a 2001-way softmax on the support $[-200, 200]$ ($\varsigma = 0.2$). We train three of them on the same data, centered

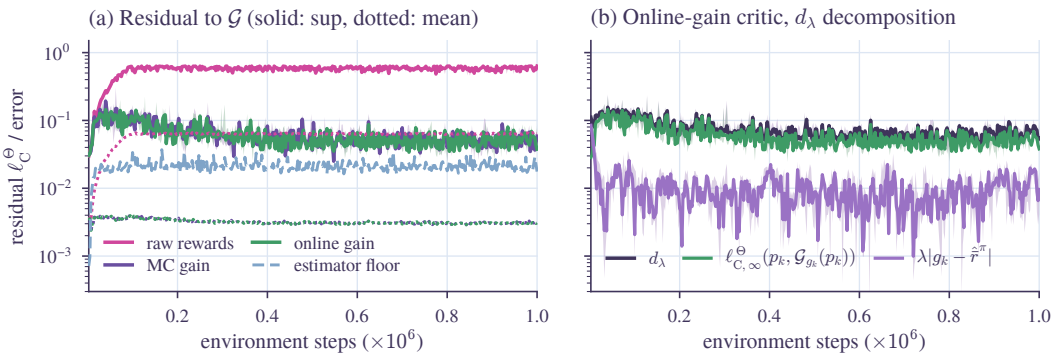


Figure 4: Continuous-state `Pendulum-v1` evaluation. Curves show medians over five seeds, and bands show interquartile ranges. **(a)** Monte Carlo estimate of the residual to the correctly centered operator $\mathcal{G}$ for the critics centered with the Monte Carlo gain, with the online gain, and with raw rewards. Solid curves take the supremum over 1000 validation states, and dotted curves the average. The dashed blue curve is the Monte Carlo error of the estimator itself. **(b)** Online-gain critic: d_λ with its categorical part $\ell^\Theta_{C,\infty}(p_k, \mathcal{G}_{g_k}(p_k))$ and gain part $\lambda|g_k - \hat{\bar{r}}^\pi|$, $\lambda = \varsigma^{-1/2}$.

with a Monte Carlo gain estimate $\hat{\bar{r}}^\pi$ from a separate 200,000-step trajectory, centered with the online gain g_t (running average with coefficient 10^{-4}, started at 0.5), and with raw rewards. A scalar average-reward TD critic serves as a reference for the categorical mean. Adam learning rates are 10^{-4} (categorical) and $3 \cdot 10^{-4}$ (scalar), the replay buffer holds 300,000 transitions, training starts after 5000 steps, and each environment step performs two updates with minibatches of 512. Target networks use Polyak coefficient 0.005. A mean-centering penalty with coefficient 10^{-4} fixes the common shift. It is a practical regularizer and not part of the analyzed recursion. Each run lasts 10^6 environment steps, and we report five seeds.

To measure convergence, every 2500 steps the residual $\ell^\Theta_C(p(s), \mathcal{G}(p)(s))$ is estimated at 1000 validation states by averaging 128 simulated one-step backups of the current critic, and reported as a supremum and an average over states. Splitting the 128 backups into two halves gives an estimate of the Monte Carlo error of this estimator, shown as the estimator floor.

Over the last 20% of training, the critics centered with the Monte Carlo and online gains reach supremum residuals of 0.04–0.07 and average residuals of 0.003, against an estimator floor of 0.02 and 0.0017, respectively. A substantial part of the remaining residual, and most of the average one, is therefore estimation noise. The raw-reward critic instead reaches 0.56–0.61. Every sampled shift is nonnegative, so all of its mass ends on the upper endpoint, the same mechanism as the $g = 0$ ablation in Figure 2(c). In panel (b), the weighted gain error falls from 0.08–0.09 to 0.009–0.014, and d_λ is dominated by its categorical part. The plateaus reflect the constant learning rates. Since Theorem 21 requires diminishing step sizes, the experiment illustrates it only qualitatively. The categorical means agree with the scalar critic ($R^2 = 0.99$), while the learned laws are wide (standard deviation about 125) with about 3% of their mass on the endpoints. This is consistent with Section 4.1, since in the undiscounted setting the spread of a projected fixed point is set by clipping, and the means remain the reliable output (Proposition 26).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: As claimed, we establish the quotient-categorical formulation, non-expansiveness results, stochastic approximation guarantees, online-gain recursion and the experimental scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The assumptions and limitations are stated in the theorem statements and conclusion.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: All theoretical claims are stated with their assumptions in the main text, and full proofs are provided in the appendix. The proof sketches in the main text refer to the corresponding appendix sections and cited SA results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper specifies the algorithms and residual quantities used in validating the theoretical predictions. The appendix details the continuous-state experiment and how residuals are estimated.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code will be provided.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental section specifies the evaluated recursions, residual metrics, centering ablation, and categorical representation. The appendix gives the setup of the neural function-approximation experiment and the MC residual estimation procedure.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The plots include variability bands across runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The experiments in the main text are lightweight fixed-policy evaluations, and the main contribution is theoretical. The neural network experiments in the appendix discuss the compute resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work is mainly theoretical and the experiments use synthetic environments and a standard simulator, with no human subjects, private data, or high-risk deployed system.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The contributions are mainly theoretical in nature and the work does not introduce a deployed system, dataset, or model with a direct societal application. We do not identify any immediate societal impacts.

    Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models, scraped datasets, or assets with misuse potential.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The only external asset used is the standard `Pendulum-v1` environment, which is properly credited through citation [12].

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release any new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.