RETIMED BELLMAN FLOWS: ESCAPING THE IMPOSSIBLE TRIANGLE OF VELOCITY BOOTSTRAPPING

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow critics learn return distributions by transporting Gaussian noise to Bellman endpoints via continuous velocity fields. While velocity bootstrapping stabilizes training by querying a successor teacher, existing methods face a structural dilemma: on straight paths, no residual-free same-time affine mapping can preserve Gaussian initial noise while maintaining an unbiased target. To overcome this limitation, we introduce Retimed Bellman Flows (ReBF). ReBF queries the teacher critic at a dynamically shifted earlier flow time, aligning intermediate student and teacher trajectories. By combining this retimed clock with fresh, decoupled noise generation, ReBF constructs a provably conditionally unbiased velocity target that preserves the Bellman fixed point and contracts under Wasserstein distances. Empirically, ReBF reduces W_1 distance to ground-truth return distributions by up to $7.7\times$ on synthetic MRPs and outperforms existing flow critics across 38 challenging OGBench and D4RL offline reinforcement learning tasks.

1 INTRODUCTION

Flow critics represent the law of a return, not only its mean, by a velocity field that transports Gaussian noise to returns, trained by flow matching (Lipman et al., 2023; Dong et al., 2026; Xu et al., 2026). The return can be a scalar, as in distributional reinforcement learning (Bellemare et al., 2017), or vector-valued, such as successor features (Appendix E.4). Unlike categorical critics such as C51, which fix a support (Bellemare et al., 2017), and quantile critics such as IQN, whose quantile function has no canonical multivariate analogue (Dabney et al., 2018a), a flow critic transports noise to returns of any dimension, and flow critics have shown strong performance against both on offline reinforcement learning benchmarks (Dong et al., 2026; Xu et al., 2026).

Training a flow-based critic via temporal-difference (TD) learning, however, introduces a fundamental mismatch between flow generation and Bellman updates. While flow matching requires learning a continuous trajectory over intermediate flow times $t \in [0, 1]$, the distributional Bellman equation strictly constrains only the marginal distribution at the endpoint $t = 1$, leaving intermediate flow dynamics unconstrained, inducing severe source inconsistency and path ambiguity across flow steps. Furthermore, directly regressing velocity vectors onto stochastically sampled successor returns suffers from high sample variance and unstable convergence. To address these issues, motivated by bootstrapping in standard value learning (Sutton, 1988), recent works (Dong et al., 2026; Xu et al., 2026) employ *velocity bootstrapping*: regressing the student's velocity field onto a target constructed from the teacher's velocity at the successor state–action pair. Under this framework, the *student* learns a generative path transporting Gaussian noise to the Bellman target return, while the *teacher* defines a path transporting noise to the next-state return X'. Consequently, the bootstrapping target crucially depends on how points along the student's intermediate trajectory are aligned with those of the teacher.

Formulating this alignment thus calls for a precise matching mechanism between the intermediate flow steps. That match should respect the distributional Bellman equation throughout the flow, which asks for three design properties (Section 3.2). (i) **Source consistency (S):** the path starts from standard Gaussian noise, the same source used at generation. (ii) **Affinity (A):** the current point is an affine image of the teacher query, so the two carry the same information given the transition. (iii) **Decoupled generation (D):** the successor return is generated with fresh noise, $\rho_1 = 0$. Affinity and decoupled generation are the road to unbiasedness: when the teacher is the exact velocity for that

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: **Comparison of flow-critic target designs** across three design properties: source consistency (**S**), affinity (**A**), and decoupled generation (**D**). On the straight path, no residual-free same-time affine lift with a time-only scale holds for every endpoint pair (Corollary 3.1); non-affine, state-dependent, and other-path queries lie outside that statement. ReBF retimes the query and generates the successor with fresh independence. With a coupling-matched teacher and exact population regression, the target is unbiased on canonical fields.

Method	Source consistency (S)	Affinity (A)	Decoupled generation (D)
Value Flows (Dong et al., 2026)	✗ $R+\gamma\tilde{X}_0 \neq X_0$	✓ $Z_t = R+\gamma X'_t$	✗
PCBF (Xu et al., 2026)	✓ at X_0	✗ contains X_0	✗
ReBF (ours)	✓ at X_0	✓ warped affine	✓

pairing, the target is conditionally unbiased. A learned teacher can still leave approximation error. As Table 1 records, neither prior method has both **(A)** and **(D)**. The obstruction is narrower than a three-way impossibility: on the straight path, source consistency and a same-time query cannot be combined in a residual-free affine lift whose scale depends only on time and discount (Corollary 3.1). This leads to the following question:

Can flow-based velocity bootstrapping be source-consistent, affine, and decoupled at the same time, and therefore unbiased?

Retimed Bellman Flows (ReBF) queries the teacher at an earlier flow time, chosen so that the contributions of source noise and successor return align after accounting for the reward and discount. This preserves the student's Gaussian starting point and Bellman endpoint while giving each intermediate student point a direct affine correspondence with a teacher query. Geometric alignment alone does not ensure an unbiased velocity target. ReBF also generates the successor return with noise independent of the student's source noise, matching the pairing under which the reused critic learns its conditional velocity. When that teacher is the coupling-matched velocity and the population regression is exact, the bootstrapped target has the correct conditional mean at each student point. On canonical fields, under fresh independence, it preserves the Bellman fixed point. A learned teacher can still leave approximation error. The time shift is determined by the discount and flow time, requiring no additional tuning parameter or teacher evaluation. We formalize these claims in Sections 4 and 5.

Our main contribution is Retimed Bellman Flows (ReBF), a framework for flow-based distributional value estimation: **(i)** On the straight source path, we prove that no residual-free same-time affine lift with a time-only scale holds for every endpoint pair; non-affine, state-dependent, and other-path methods lie outside that statement; **(ii)** ReBF queries the teacher at a dynamically shifted earlier flow time $\tau(t)$ and generates the successor under fresh independence ($\rho_1 = 0$ together with (A6)), a closed-form target with no added hyperparameter; **(iii)** we prove that this retimed shift is the unique such affine lift conjugate to the Bellman map, and that a coupling-matched teacher, exact population regression, and those independence assumptions make the target unbiased at every bootstrap weight on canonical fields, with a defect off those fields; and **(iv)** across toy MRPs, OGBench, and D4RL, ReBF improves W_1 by up to 7.7× over PCBF.

2 Related Work

Existing flow-based critics construct distributional value targets through three primary mechanisms. First, endpoint-based and sampled methods avoid pathwise velocity bootstrapping altogether: FLOQ (Agrawalla et al., 2026) regresses to point targets without modeling return distributions, Bellman Diffusion and distributional flow critics (Chen et al., 2025; Li et al., 2024) back up endpoint laws without pathwise alignment, and Temporal Difference Flows (Farebrother et al., 2025) recovers values only in expectation. Second, path-coupling frameworks leverage intermediate flow fields by incorporating successor velocities (Dong et al., 2026; Xu et al., 2026; Espinosa-Dice et al., 2025); however, they inevitably trade off path consistency against target bias. For instance, Value Flows maintains same-time affinity but violates source consistency, whereas PCBF preserves Gaussian initial noise but suffers from target bias under same-time, shared-noise queries with a self-teacher

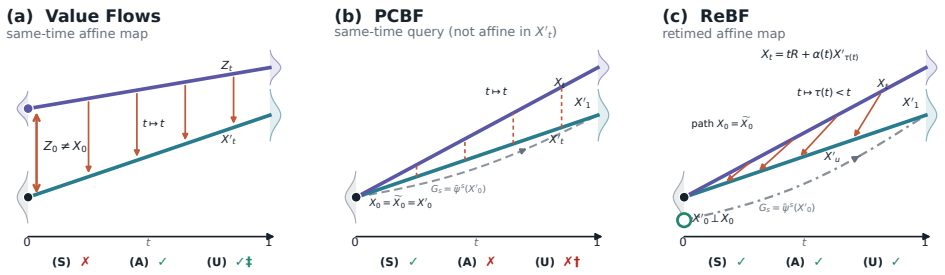


Figure 1: **Three ways to bootstrap a velocity.** (a) Value Flows: same-time affine, but $Z_0 \neq X_0$. (b) PCBF: source-consistent and same-time, with a residual in X_0. (c) ReBF: retimed query, independent noise.

critic (Theorem 5.3), as straight-path couplings do not naturally match the teacher's conditional velocity (Lipman et al., 2023; Liu et al., 2023; Tong et al., 2024; Lipman et al., 2024). Third, alternative couplings like FlowIQN (Groom et al., 2026) match quantiles within a single return rather than noise across Bellman steps, preserving Wasserstein contraction only when the update exactly realizes the underlying Bellman operator (Section 5).

3 Research Gap and Motivation

In this section, we describe the limitations of existing flow critics, state the three design properties their targets should satisfy, and show why, on the straight path, no residual-free same-time affine lift with a time-only scale holds for every endpoint pair. On canonical fields, unbiasedness is what affinity and fresh independence produce when the teacher is coupling-matched and the population regression is exact.

3.1 Preliminaries

A flow critic models the return law $\eta^\pi(s,a)$ of a fixed policy by a velocity $v_\theta(t, x \mid s, a)$, $t \in [0,1]$. Transition tuple is $H = (R, S', A', \tilde{\gamma})$, with $S' \sim P(\cdot \mid s,a)$, $A' \sim \pi(\cdot \mid S')$, and $0 \leqslant \tilde{\gamma} \leqslant \gamma$ ($\tilde{\gamma} = 0$ at termination), and the return satisfies

$$X_1 \overset{\mathrm{D}}{=} R + \tilde{\gamma} X'_1, \qquad X'_1 \mid H \sim \eta_{S',A'}, \tag{1}$$

The right-hand side is the Bellman operator $\mathcal{T}^\pi$, with unique fixed point η^π (Bellemare et al., 2017). Generation integrates v_θ from $X_0 \sim \mathcal{N}(0, I)$; the endpoint law is the critic's model of $\eta^\pi(s,a)$. Flow matching (Lipman et al., 2023; Liu et al., 2023) trains v_θ on the straight path

$$X_t = (1-t)X_0 + tX_1, \qquad X_0 \perp\!\!\!\perp X_1 \sim \eta(s,a), \quad t \sim \mathcal{U}[0,1], \tag{2}$$

by regressing $v_\theta(t, X_t \mid s, a)$ onto $X_1 - X_0$. The minimiser is the marginal velocity $v^\star(t,x) = \mathbb{E}[X_1 - X_0 \mid X_t = x]$, whose flow from $\mathcal{N}(0,I)$ reproduces $\eta(s,a)$. At $t = 0$, $X_t = X_0$ is independent of X_1, so $v^\star(0,x) = \mathbb{E}[X_1] - x$: the mean return is one evaluation of the field at $t = 0$. Substituting the Bellman return into that path gives the student trajectory and its velocity,

$$X_t = (1-t)X_0 + t\,(R + \tilde{\gamma} X'_1), \qquad Y := \dot{X}_t = R - (1-\tilde{\gamma})X_0 + \tilde{\gamma}\,(X'_1 - X_0). \tag{3}$$

A backup carries three noises: the student source $X_0 \sim \mathcal{N}(0,I)$, the successor-path start $\tilde{X}_0$, and the generation noise X'_0 with $X'_1 = \bar{\psi}^{\,1}(X'_0 \mid S', A')$. Following Xu et al. (2026) we set $\tilde{X}_0 = X_0$ and $X'_0 = \rho_1 X_0 + \sqrt{1-\rho_1^2}E$ for independent $E \sim \mathcal{N}(0,I)$ and $\rho_1 \in [0,1]$. Velocity bootstrapping replaces the sampled chord $X'_1 - X_0$ by a frozen successor critic $\bar{v}$, queried at time $u(t)$ on $X'_u = (1-u)X_0 + uX'_1$. The resulting target is u_t.

3.2 Three properties for velocity bootstrapping

Source consistency is the requirement on the path. Affinity and decoupled generation are the two construction choices. On canonical fields, conditional unbiasedness is what those two produce when the teacher is coupling-matched and the population regression is exact.

(S) Source consistency. The source path and the successor path are the straight interpolants of flow matching, Eq. (2). Flow matching requires those paths to originate from a standard Gaussian at $t = 0$, whereas the Bellman operator constrains the return distribution at $t = 1$. Absorbing immediate rewards into the initial noise forces the path to start from a shifted distribution $R + \tilde{\gamma}\tilde{X}_0$ rather than $X_0 \sim \mathcal{N}(0, I)$, which causes boundary mismatches to propagate along the generated trajectory (Xu et al., 2026).

(A) Affinity. Every student point X_t is an affine image of a successor point $X'_{u(t)}$, given by $X_t = a(t)R + \varphi(t)X'_{u(t)}$, which reduces to the Bellman map $x \mapsto R + \tilde{\gamma}x$ at $t = 1$. The teacher query is then available in closed form and visible for every noise coupling. Visibility aligns the student's input with the teacher's query. It does not match the teacher to the coupling, so affinity alone does not make the target unbiased.

(D) Decoupled generation. The successor return is generated with fresh noise, $\rho_1 = 0$, the noise–return pairing of independent flow matching. Only the generation noise X'_0 is independent of X_0; the student and successor interpolants may still share X_0.

When the noise coupling brings no additional bias. Replacing the sampled chord by a teacher always leaves an error: the teacher is not the chord. The design question is whether the noise coupling adds a further bias, and whether that addition can be kept at zero. The exact condition is that the relevant conditional residual have zero expectation. Under an exact teacher, two structural conditions are sufficient for that. They are not necessary: cancellation, restricted support, or $\kappa = 0$ can make the residual expectation vanish without both.

1. **Query visibility,** from affinity: $X'_{u(t)}$ and X_t determine each other given H. If X_t leaves the query uncertain, the student can average over mismatched teacher evaluations.
2. **Decoupled generation:** independence holds only at $\rho_1 = 0$; a value $0 < \rho_1 < 1$ is partial correlation. The self-teacher claim uses fresh independence in the sense of (A5) and (A6), which the label $\rho_1 = 0$ alone does not give, together with a teacher that is the exact conditional velocity for that pairing. On canonical fields the critic is then its own matched teacher (Lemma B.3).

Retiming $u(t) \to \tau(t)$ supplies the first condition for every coupling. It does not match the teacher; that is decoupled generation. Affinity and decoupled generation together add no coupling bias when the teacher is exact (Theorem 5.3). A learned teacher still leaves approximation error, which is separate from the bias the coupling itself adds.

3.3 Two prior paths, and the property each gives up

Value Flows (Dong et al., 2026) satisfies **(A)** and gives up **(S)** and **(D)** on its DCFM path. The published loss is a one-to-one mixture of BCFM and DCFM. BCFM matches a path that starts at raw $\mathcal{N}(0, I)$ noise, so that term is source-consistent on its own. DCFM applies the affine map $Z_t = R + \tilde{\gamma}X'_t$ at the same time, with velocity $\dot{Z}_t = \tilde{\gamma}(X'_1 - \tilde{X}_0)$ and corrected target $u_t^{\mathrm{VF}} = \tilde{\gamma}\,\bar{v}\big(t, (Z_t - R)/\tilde{\gamma} \mid S', A'\big)$. The map is invertible, so the query is visible, and with a matched teacher the corrected target is unbiased. The published DCFM loss omits $\tilde{\gamma}$ and regresses on $\bar{v}$ itself (Dong et al., 2026, Eq. 5); with a perfect teacher that target has conditional mean $\dot{Z}_t/\tilde{\gamma}$. The DCFM path ends at the Bellman target but starts at $Z_0 = R + \tilde{\gamma}\tilde{X}_0$. Generation always starts from a fresh $\mathcal{N}(0, I)$ draw, so the DCFM term near $t = 0$ is queried on inputs that term rarely saw. The field does see Gaussian inputs near $t = 0$, through BCFM; what the mixture fits is two paths with different boundary conditions at $t = 0$. Anchoring and confidence weights are added to compensate (Dong et al., 2026).

PCBF (Xu et al., 2026) keeps the source-consistent path of Eq. (3) and queries the teacher at the same time t, at X'_t, with target $R - (1 - \tilde{\gamma})X_0 + \tilde{\gamma}\,\bar{v}(t, X'_t)$ (full correction, $\kappa = 1$ in our parametrisation). The student point and the same-time chord are not affinely related:

$$X_t = tR + \tilde{\gamma}X'_t + (1 - t)(1 - \tilde{\gamma})X_0, \qquad X'_t = \frac{X_t - tR}{\tilde{\gamma}} - \frac{(1 - t)(1 - \tilde{\gamma})}{\tilde{\gamma}}\,X_0. \tag{4}$$

The query is visible only if X_0 is. If the generation noise is not fully shared ($\rho_1 < 1$), X_0 is hidden and even an exact teacher leaves a conditioning residual (Eq. (32), correlated-Gaussian model).

Decoupled generation is the endpoint $\rho_1 = 0$ of that range, not every $\rho_1 < 1$. PCBF shares the noise ($\rho_1 = 1$): for scalar returns X_0 is recoverable from (X_t, H), the query is visible, and a teacher matched to that coupling is exact (Proposition A.1(i)); the reused critic is generally not that teacher (Proposition A.1(ii)), so the target stays biased. PCBF controls the resulting bias and variance with a gain λ tuned per domain (Xu et al., 2026). So PCBF keeps **(S)** and gives up **(A)** and **(D)**. In this construction the reused critic is not the coupling-matched teacher, so the target stays biased even before further approximation error. Missing either structural condition does not by itself force that bias.

3.4 The gap: no source-consistent same-time affine link

Figure 1 and Table 1 summarise the three methods: Value Flows gives up **(S)** and **(D)** on its DCFM path, and PCBF gives up **(A)** and **(D)**. The impossibility itself does not involve the coupling. After removing tR, the student point weights (X_0, X_1') as $(1-t,\ \tilde{\gamma}t)$, while the same-time chord point weights them as $(1-t,\ t)$, so, on that straight path, no residual-free same-time affine lift with a time-only scale holds for every endpoint pair when $0 < \tilde{\gamma} < 1$ (Corollary 3.1). Affinity is recovered by reading the teacher earlier. Together with fresh independence (A6), a coupling-matched teacher, and exact population regression, that alignment is sufficient for unbiasedness on canonical fields.

Corollary 3.1 (No same-time affine design under straight paths flow)**.** *Write $X_t' = (1-t)X_0 + tX_1'$ for the straight successor path read at the same time t. The first candidate is the same-time affine transformation $X_t = tR + \alpha(t)X_t'$. For $0 < \tilde{\gamma} < 1$ and $0 < t < 1$, no scale $\alpha(t)$ depending only on $(t, \tilde{\gamma})$ makes this identity hold for every endpoint pair (X_0, X_1'), while X_t stays the straight source path of Eq. (3).*

Its proof is in Appendix B.

How we close the gap. Reading the chord at the earlier time $\tau(t) < t$ whose weights are proportional to $(1-t,\ \tilde{\gamma}t)$ removes the residual, so the query is visible under every coupling; fresh independence in the sense of (A5) and (A6), which the label $\rho_1 = 0$ alone does not give, then makes the critic, on canonical fields, its own matched teacher (Theorem 5.3(ii)). This is ReBF, developed next.

4 Retimed Bellman Flows

ReBF keeps the source-consistent path of Eq. (3), so it satisfies (S) by construction, and removes the same-time residual by reading the successor chord at a different flow time. Section 4.1 obtains affinity by retiming the teacher query, Section 4.2 writes the resulting target, and Section 5.3 adds decoupled generation and shows that the two choices together make the target unbiased.

That is what forces a warped clock, which the next paragraphs derive. The qualifier "for every endpoint pair" matters. Under shared noise ($\rho_1 = 1$) the successor endpoint is a function of X_0, and for scalar returns with a generator that is the time-one map of a unique one-dimensional flow, both X_t and X_t' are strictly increasing in X_0 given H for $t < 1$, so X_t' is a function of (X_t, H): precisely the corner where the two clocks' oracle targets coincide, and the corner PCBF occupies (Corollary B.4).

Affine warped transformation. The same source path admits the warped-clock rewrite

$$X_t = tR + \alpha(t)X'_{\tau(t)}, \qquad X'_{\tau(t)} = (1-\tau(t))X_0 + \tau(t)X_1', \tag{5}$$

where the quantity being read is a straight path in the successor endpoint. The scale $\alpha(t)$ and the successor clock $\tau(t)$ are fixed by matching Eq. (5) to the expansion of Eq. (3).

4.1 Why the successor clock is slower

Matching X_0 and X_1' coefficients after substituting Eq. (5) into Eq. (5) and expanding Eq. (3) with $X_1 = R + \tilde{\gamma}X_1'$ yields $\alpha(t)(1-\tau(t)) = 1-t$ and $\alpha(t)\tau(t) = \tilde{\gamma}t$. Adding and dividing gives the clock, and the same relation then fixes the successor query from the current point and the transition:

$$\alpha(t) = 1-(1-\tilde{\gamma})t, \qquad \tau(t) = \tfrac{\tilde{\gamma}t}{\alpha(t)}, \qquad \xi_t = X'_{\tau(t)} = \tfrac{X_t - tR}{\alpha(t)}. \tag{6}$$

For $0 < \tilde{\gamma} < 1$ and $0 < t < 1$, $\tau(t) < t$ because the source path weights X'_1 by $\tilde{\gamma}t$ while a same-time reading weights it by t. Once the transition is known, ξ_t is determined by the current regression point.

Theorem 5.2 strengthens Corollary 3.1 from an impossibility into a characterisation: among residual-free affine lifts with a time-only scale, (α, τ) is the *only* one under which the velocity update is a Bellman map. Non-affine, state-dependent, and other-path clocks lie outside that statement.

4.2 Unbiased control variate and time-varying gain

Velocity target. Differentiating the affine relation Eq. (5) by the chain rule with Eq. (6) gives the sample velocity

$$Y = \dot{X}_t = R - (1 - \tilde{\gamma})\xi_t + g(t)(X'_1 - X_0), \tag{7}$$

with gain $g(t) = \alpha(t)\tau'(t) = \tilde{\gamma}/\alpha(t)$. Replacing the sampled chord $X'_1 - X_0$ by a frozen successor teacher $\bar{v}$, with $a_t = \kappa g(t)$, defines the control variate and the general target

$$C_{\tau(t)} := \bar{v}(\tau(t), X'_{\tau(t)} \mid H) - (X'_1 - X_0), \tag{8}$$

$$u_a = Y + a_t\, C_{\tau(t)}. \tag{9}$$

At weight κ the regression target is

$$u_\kappa = Y + \kappa g(t)\, C_{\tau(t)} = R - (1 - \tilde{\gamma})\xi_t + g(t)\{(1 - \kappa)(X'_1 - X_0) + \kappa \bar{v}(\tau(t), \xi_t \mid H)\}. \tag{10}$$

With a matched teacher, $\mathbb{E}[C_{\tau(t)} \mid X_t, H] = 0$, so every κ keeps the same conditional mean (Theorem 5.3). The weight $\kappa \in [0, 1]$ trades residual bridge noise against teacher error; $\kappa = 1$ is the partial Rao–Blackwellization of Proposition 5. Replacing a sampled quantity by its conditional expectation is Rao–Blackwellization and the gain is a control-variate weight; neither is new, and we do not claim either as a contribution. The gain $g(t) = \tilde{\gamma}/\alpha(t)$ rises from $\tilde{\gamma}$ at $t = 0$ to 1 at $t = 1$, so the weight $\kappa g(t)$ varies with t, while PCBF uses a fixed gain $\lambda = \tilde{\gamma}$. For terminal backups we use $Y = R - X_0$ and set the correction to zero, avoiding the singular $\tilde{\gamma} = 0, t = 1$ inverse.

4.3 Decoupled Generation

With $\tilde{X}_0 = X_0$ already fixed, the generation noise is

$$X'_0 = \rho_1 X_0 + \sqrt{1 - \rho_1^2}\, E, \tag{11}$$

where $E \sim \mathcal{N}(0, I_d)$ is independent of (X_0, H, t), $0 \leqslant \rho_1 \leqslant 1$, and $X'_1 = \bar{\psi}^1(X'_0 \mid S', A')$. By default ReBF sets $\rho_1 = 0$: the successor is drawn independently, and the backup is still the full correction. PCBF sets $\rho_1 = 1$, tying X'_0 to the path noises. That tie can cancel a same-time bias, but it is stronger than an unbiased bootstrap needs. Section 6 records its cost. The matched teacher on the retimed path is

$$v^\star_{\rho_1}(\tau(t), b \mid H) = \mathbb{E}[X'_1 - X_0 \mid X'_{\tau(t)} = b, H]. \tag{12}$$

At $\rho_1 = 1$, when the chord is recoverable from the successor query, the two clocks share this oracle. ReBF is the retimed clock with $\rho_1 = 0$ and $\kappa = 1$ (Algorithm 1); PCBF is the same-time clock with $\rho_1 = 1$ and $\kappa = 1$. The uncorrected critic is $\kappa = 0$, whose clock is inert.

5 Theory

Definition 5.1 (Bellman velocity operator)**.** Fix π and the source $X_0 \sim \mathcal{N}(0, I_d)$. For a velocity field v, let $\mathcal{E}(v)_c$ be the law at $t = 1$ of the flow $\dot{Z}_t = v_c(Z_t, t)$, $Z_0 = X_0$: the returns the critic *generates*. For a return-law family η, let $(\mathcal{T}^\pi \eta)_c = \mathrm{Law}(R + \tilde{\gamma} X'_1 \mid c)$ with $X'_1 \sim \eta_{S',A'}$, as in Eq. (1), and let $(\mathcal{L}\eta)_c(z, t) = \mathbb{E}[X_1 - X_0 \mid (1 - t)X_0 + tX_1 = z]$ with $X_1 \sim \eta_c$ independent of X_0: the straight flow-matching field of η_c. The *Bellman velocity operator* is

$$\mathcal{C}^\pi = \mathcal{L}\,\mathcal{T}^\pi\,\mathcal{E}: \qquad v \xrightarrow{\mathcal{E}} \text{returns} \xrightarrow{\mathcal{T}^\pi} \text{backed-up returns} \xrightarrow{\mathcal{L}} \text{new field}. \tag{13}$$

A field is *canonical* if it equals $\mathcal{L}\eta$ for some η. Its *readout* $q_v(z, t) = z + (1 - t)v(z, t)$ is the endpoint a particle at z reaches if it keeps its velocity; for a canonical field and $t < 1$ this is $\mathbb{E}[X_1 \mid X_t = z]$.

Generate returns with the critic, apply the Bellman backup, and fit the straight flow of the result. Assumptions (A1)–(A9) and the full statements are in Appendix A.

On one backup, retiming is the Bellman map. Freeze one draw $h = (r, s', a', \tilde{\gamma})$ of the backup, a *branch*; on it the Bellman equation is $X_1 = r + \tilde{\gamma} X_1'$.

Theorem 5.2 (Bellman conjugacy characterises the retiming)**.** *Fix a nonterminal branch with $0 < \tilde{\gamma} < 1$, and let v' be the successor's flow-matching field under the branch's coupling. Among affine lifts $X_t = tr + \varphi(t) X'_{\vartheta(t)}$ with $\varphi(0) = 1$ and $\vartheta(0) = 0$, whose field we write $\mathfrak{B}^{\varphi,\vartheta} v$, the readout identity*

$$q_{\mathfrak{B}^{\varphi,\vartheta} v}(z,t) = r + \tilde{\gamma}\, q_v\big(\xi, \vartheta(t)\big), \qquad \xi = \frac{z - tr}{\varphi(t)}, \tag{14}$$

holds for every field v if and only if $(\varphi, \vartheta) = (\alpha, \tau)$ of Eq. (6); a same-time query forces $\tilde{\gamma} = 1$. The backed-up field is then

$$(\mathfrak{B}_{r,\tilde{\gamma}} v')(z,t) = r - (1 - \tilde{\gamma})\, \xi + g(t)\, v'(\xi, \tau), \qquad g = \tilde{\gamma}/\alpha, \tag{15}$$

which for canonical v' is $\mathcal{C}^\pi$ on that branch, and it shrinks readout differences by exactly $\tilde{\gamma}$.

In readouts the backup is $x \mapsto r + \tilde{\gamma} x$. The same-time query is not among these lifts unless $\tilde{\gamma} = 1$.

Retiming makes the sampled target unbiased for the operator. The algorithm regresses on the sampled target u_κ of Eq. (10), which replaces a fraction κ of the chord $X_1' - X_0$ by the teacher at the retimed point.

Theorem 5.3 (Centering and realisation)**.** *Assume (A2)–(A4): a coupling-matched teacher, a gain measurable given (X_t, H), and a square-integrable target.* (i) *$\mathbb{E}[C_{\tau(t)} \mid X_t, H] = 0$, so $\mathbb{E}[u_\kappa \mid X_t] = v_t^\star(X_t \mid c)$ for every gain.* (ii) *With fresh independence (A6) — the label $\rho_1 = 0$ alone does not give it — and the critic's own field as teacher, the regression target equals $\mathcal{C}^\pi v$ on canonical fields. Off those fields a defect remains; its formula is Theorem 5.3(ii).* (iii) *Training then iterates $\mathcal{C}^\pi$.*

Part (i) is the tower property: the retimed query is a function of (X_t, H), so replacing the chord by its conditional mean leaves the average unchanged. The same-time query is not, and its residual, Eq. (32), is scaled by κ.

Corollary 5.4 (Fixed point and contraction)**.** *Under (A1), $\mathcal{E}\,\mathcal{C}^\pi = \mathcal{T}^\pi \mathcal{E}$, so $\mathcal{C}^\pi$ has the unique fixed point $v^\pi = \mathcal{L}\eta^\pi$, and the returns two critics generate contract at the Bellman rate:*

$$\delta_p(\mathcal{C}^\pi v, \mathcal{C}^\pi w) \leqslant \beta_p\, \delta_p(v, w), \qquad \delta_p\big((\mathcal{C}^\pi)^n v, v^\pi\big) \leqslant \beta_p^n\, \delta_p(v, v^\pi), \tag{16}$$

with $\beta_p = \max_c \mathbb{E}[\tilde{\gamma}^p \mid c]^{1/p} \leqslant \gamma$ and δ_p the maximum W_p distance between generated return laws. A method inherits this only when its target estimates $\mathcal{C}^\pi v$, which Theorems 5.2 and 5.3 give for the retimed clock. The full statement is Proposition 5.4.

The gain trades residual bridge noise S against teacher error e. The conditional risk is minimised at $\kappa^\star = S/(S + \|e\|^2)$ when the denominator is positive, so full correction is optimal exactly for an exact teacher (Proposition 5).

6 Experiments

We structure our empirical evaluation around three core questions: whether ReBF yields more accurate return laws, whether query-time retiming and noise decoupling explain this advantage, and whether this superior fidelity translates to downstream offline control. For policy evaluation, we benchmark under fixed architectures and budgets on three analytical MRPs (BERNOULLI, SOLITAIRE, DISCRETE-MC, details in E), successor features in FOUR-ROOMS (E.4), and POINTMAZE-TELEPORT, measuring W_1 distance against ground-truth or rollout-based reference distributions. For downstream control, we evaluate ReBF across 38 OGBench and D4RL tasks and the results are shown in Table 3. Overall, ReBF achieves **state-of-the-art performance**—most prominently on long-horizon, multi-stage manipulation tasks where uncancelled target bias severely degrades traditional value estimation.

Table 2: W_1, **mean, and std. against comparators** (lower is better; mean bias near zero). Bold: best in the row; underlined: not different from the best ($p \geqslant 0.05$). std. is $100 \sum_s |\mathrm{sd}_s - \mathrm{sd}_s^\star| / \sum_s \mathrm{sd}_s^\star$ over states. †over-dispersed; ¶per-coordinate floq; ‡published Value Flows; §full maze, 2,880 points.

Env	Metric	BCFM (same-time, 1, 0)	PCBF (same-time, 1, 1)	PCBF, tuned κ (same-time, 1, $\kappa^\star$)	floq scalar	Value Flows default	**ReBF (ours)** (retimed, 0, 1)
BERNOULLI	W_1	0.0181 ±.0023	0.0493 ±.0022	0.0180 ±.0022	0.184 ±.001	0.122 ±.004	**0.0151** ±.0017
	mean	+0.38 ±.55	+0.36 ±.49	+0.35 ±.55	**−0.42** ±.35	+12.2 ±.4	−0.17 ±.44
	std.	**1.2** ±.2	7.2 ±.4	1.4 ±.2	13.2 ±.2	3.7† ±.3	1.7 ±.3
SOLITAIRE	W_1	0.147 ±.006	0.677 ±.046	0.154 ±.007	1.464 ±.010	1.334 ±.010	**0.0878** ±.0096
	mean	**−0.87** ±.36	−14.0 ±1.2	−0.94 ±.42	+12.6 ±.6	−39.0 ±.4	−1.73 ±.46
	std.	4.5 ±.4	15.8 ±1.3	4.8 ±.5	35.3 ±.3	49.4 ±.3	**2.3** ±.3
DISCRETE-MC	W_1	0.517 ±.015	0.617 ±.016	0.493 ±.018	1.950 ±.015	6.903 ±.013	**0.298** ±.008
	mean	−0.80 ±.26	+0.04 ±.50	+0.37 ±.35	**−0.04** ±.19	−53.8 ±.1	−0.04 ±.24
	std.	8.4 ±.4	11.4 ±.4	8.7 ±.4	56.3 ±.1	62.4 ±.1	**3.7** ±.2
four-rooms (16-dim SF)	W_1	0.0523 ±.0007	0.0410 ±.0010	0.0394 ±.0008	0.372 ±.001¶	0.241 ±.002	**0.0282** ±.0009
	mean	2.9 ±.1	3.0 ±.1	2.7 ±.1	9.1 ±.2¶	29.4 ±.2	**2.7** ±.1
	std.	10.6 ±.2	8.1 ±.2	7.6 ±.2	89.8 ±.1¶	20.6† ±.1	**2.8** ±.1
OGBENCH teleport§	W_1	22.6 ±.3	19.5 ±.5	19.9 ±.7	28.3 ±.4	60.4 ±.6‡	**14.6** ±1.0
	mean	+7.7 ±.6	+10.4 ±1.2	+9.9 ±1.2	**+0.4** ±.7	+40.4 ±.5‡	+4.1 ±1.2
	std.	54.2 ±.6	21.6 ±2.0	33.6 ±7.6	92.5 ±.3	92.1 ±.4‡	**19.6** ±4.6

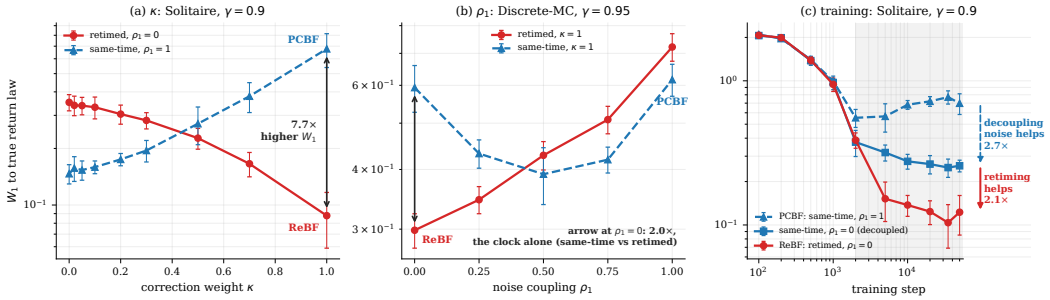


Figure 2: **Why ReBF wins.** (a) Effect of correction weight κ across clock and coupling variants. (b) Effect of noise coupling ρ_1 at full correction ($\kappa = 1$). (c) Training dynamics isolating the sequential gains of noise decoupling and clock retiming; Details in Appendix D).

Does ReBF recover the return law? It has the lowest W_1 and dispersion error on all benchmarks (Table 2). Most methods other than Value Flows get the mean nearly right, so what separates them is the shape of the law: a correct mean does not certify a correct law (Figure 3). PCBF's best κ weakens the correction toward the uncorrected BCFM, whereas ReBF is best at full correction in every environment (Figure 2a, Table 5). This is what the theory predicts: ReBF's correction is unbiased for every κ, so nothing is gained by shrinking it (Theorem 5.3), whereas PCBF's inherits the error of its mismatched teacher, scaled by κ (Proposition A.1).

Which design makes ReBF the best? Neither choice alone: ReBF needs both the retimed clock and independent noise. At full correction, retiming lowers W_1 with independent noise ($\rho_1 = 0$) but not with shared noise ($\rho_1 = 1$), in all three toys (Figures 2b and 5b). As Section 4.3 predicts, for scalar returns shared noise already makes the same-time target exact under a matched teacher, so retiming has nothing to fix there and can help only with independent noise.

Are both retiming and decouple generation necessary? Yes: clock retiming ($\tau(t)$) and independent noise coupling ($\rho_1 = 0$) act strictly synergistically. As shown in Figure 2b, retiming significantly lowers W_1 error under independent noise, whereas under shared noise ($\rho_1 = 1$), scalar-return chords are recoverable from either query, causing both clocks to target similar oracle fields. Conversely, maintaining a same-time clock under independent noise remains suboptimal. Isolating their

Table 3: **Offline RL Results.** We report performance on OGBench and D4RL. Results are averaged over 8 random seeds (4 seeds for pixel-based tasks). Bold numbers denote the values within 95% of the best performing method on each domain. Per-task results are in Table 8.

Domain	IQN	CODAC	FloQ	FQL	IQL	Value Flows	PCBF	**ReBF (Ours)**
cube-double-play (5 tasks)	42 ± 8	61 ± 6	47 ± 14	29 ± 6	6 ± 2	69 ± 4	70 ± 7	$\mathbf{76 \pm 6}$
scene-play (5 tasks)	40 ± 1	55 ± 1	$\mathbf{58 \pm 4}$	56 ± 2	28 ± 3	$\mathbf{59 \pm 4}$	$\mathbf{58 \pm 3}$	$\mathbf{60 \pm 4}$
puzzle-4x4-play (5 tasks)	27 ± 4	20 ± 18	28 ± 6	17 ± 5	7 ± 2	27 ± 4	32 ± 5	$\mathbf{36 \pm 8}$
cube-triple-play (5 tasks)	6 ± 0	2 ± 1	8 ± 3	4 ± 2	1 ± 1	$\mathbf{14 \pm 3}$	6 ± 3	12 ± 4
D4RL adroit (8 tasks)	66 ± 5	$\mathbf{69 \pm 1}$	$\mathbf{70 \pm 3}$	$\mathbf{71 \pm 4}$	$\mathbf{70}$	65 ± 2	$\mathbf{70 \pm 2}$	$\mathbf{69 \pm 1}$
visual-antmaze-teleport (5 tasks)	4 ± 2	–	–	5 ± 2	6 ± 4	13 ± 4	$\mathbf{14 \pm 4}$	13 ± 3
visual-cube-double-play (5 tasks)	1 ± 0	–	–	6 ± 1	11 ± 6	$\mathbf{13 \pm 2}$	3 ± 1	3 ± 1

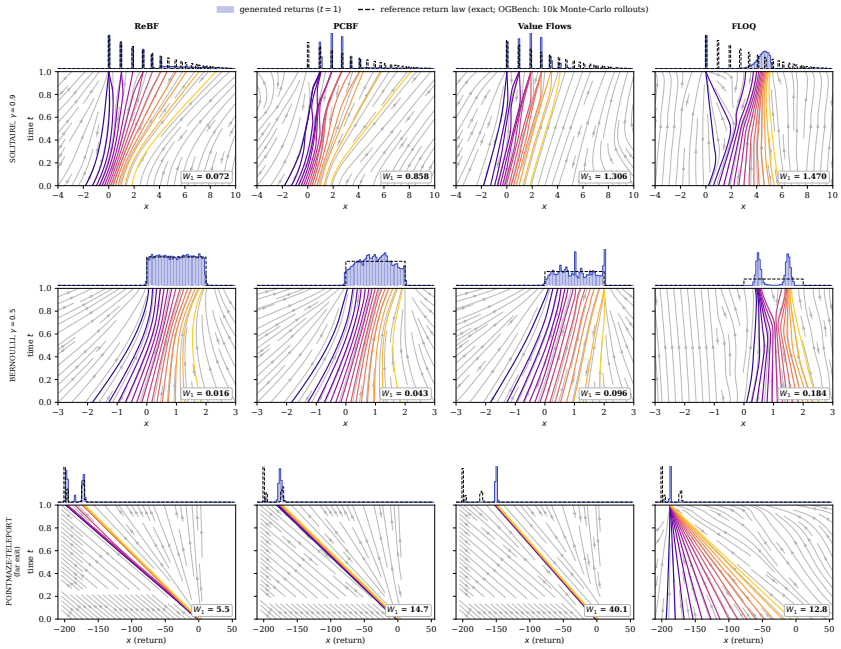


Figure 3: **Comparison of learned flow fields and generated return distributions.** Each plot shows velocity streamlines, sample trajectories from flow time $t = 0$ to $t = 1$, and the resulting terminal return distributions (purple bars) compared against ground-truth Monte-Carlo rollouts (dashed lines) for ReBF, PCBF, Value Flows, and FLOQ.

contributions on SOLITAIRE demonstrates two additive steps: decoupling noise lowers PCBF's final W_1 error by $2.7\times$, and retiming the clock provides a further $2.1\times$ drop (Figure 2c).

Why does same-time shared-noise bootstrapping plateau? Reusing a critic as its own teacher under shared noise ($\rho_1 = 1$) introduces an irreducible target bias that corrupts iterative updates. On SOLITAIRE, PCBF's error initially decreases but subsequently rises and plateaus at a biased level—a degradation shared by the retimed clock under $\rho_1 = 1$ (Figure 2c). Theoretically, a self-teacher critic under shared noise fails to match the required target field (Proposition A.1(ii)), continually feeding conditioning bias back into the Bellman backup. Independent noise ($\rho_1 = 0$) resolves this mismatch because the required teacher is precisely the marginal field the critic already learns (Lemma B.3).

Conclusion. We propose Retimed Bellman Flows (ReBF), a hyperparameter-free framework that queries the teacher critic at a dynamically shifted flow time $\tau(t)$. By restoring exact affine conjugacy with the Bellman operator, ReBF eliminates target bias without extra evaluations. Empirically, ReBF improves ground-truth return fidelity W_1 by up to $7.7\times$ on MRPs and achieves state-of-the-art performance across 38 OGBench and D4RL offline benchmarks.

AI USE STATEMENT

Large language model assistance was used in preparing this manuscript: for drafting and organizing prose, for suggesting proof strategies, and for implementing the numerical diagnostic scripts. The research questions, original research ideas, the theoretical framework, and the final form of every claim are the authors' own. The authors have verified each theorem statement and proof independently of the assistance received and have confirmed every citation against its primary source. The authors take full responsibility for all content of this paper, including any errors.

REPRODUCIBILITY STATEMENT

The analytic identity grids and the contraction moduli run on CPU and require only NumPy, SciPy and SymPy, with Matplotlib for the figures. The Gaussian column of Table 7 is Eq. (33) in closed form; its Bernoulli and Retimed columns are fixed-seed Monte Carlo estimates (2×10^7 draws, seed 0), shipped with the supplement. The toy Markov reward processes, arm definitions, training configuration, preregistered analysis protocol and per-seed scores for Section 6 are given in Appendix F; every reported W_1 is computed against the analytic return law rather than a sampled reference, and the reference-error floor for each environment is reported alongside it. Appendix A states all assumptions and Appendix B gives complete proofs.

REFERENCES

Bhavya Agrawalla, Michal Nauman, Khush Agrawal, and Aviral Kumar. floq: Training critics via flow-matching for scaling compute in value-based RL. In *International Conference on Learning Representations*, 2026. arXiv:2509.06863.

Gabriel Barth-Maron, Matthew W. Hoffman, David Budden, Will Dabney, Dan Horgan, Dhruva TB, Alistair Muldal, Nicolas Heess, and Timothy Lillicrap. Distributed distributional deterministic policy gradients. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1804.08617.

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pp. 449–458, 2017. URL https://proceedings.mlr.press/v70/bellemare17a.html.

Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. MIT Press, 2023.

Deshu Chen, Yuchen Liu, Zhijian Zhou, Chao Qu, and Yuan Qi. Unleashing flow policies with distributional critics, 2025. URL https://arxiv.org/abs/2509.23087.

Xiaoou Cheng and Jonathan Weare. The surprising efficiency of temporal difference learning for rare event prediction. In *Advances in Neural Information Processing Systems*, volume 37, pages 81257–81286, 2024.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International Conference on Machine Learning*, pp. 1096–1105, 2018a. URL https://arxiv.org/abs/1806.06923.

Will Dabney, Mark Rowland, Marc G. Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *AAAI Conference on Artificial Intelligence*, 2018b.

P. Dong, C. Zheng, Chelsea Finn, Dorsa Sadigh, and Benjamin Eysenbach. Value flows. In *International Conference on Learning Representations*, 2026.

Jingliang Duan, Yang Guan, Shengbo Eben Li, Yangang Ren, Qi Sun, and Bo Cheng. Distributional soft actor-critic: Off-policy reinforcement learning for addressing value estimation errors. *IEEE Transactions on Neural Networks and Learning Systems*, 33(11):6584–6598, 2022.

Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 5th edition, 2019.

Nicolas Espinosa-Dice, Kiante Brantley, and Wen Sun. Expressive value learning for scalable offline reinforcement learning, 2025. URL https://arxiv.org/abs/2510.08218.

Jesse Farebrother, Matteo Pirotta, Andrea Tirinzoni, Rémi Munos, Alessandro Lazaric, and Ahmed Touati. Temporal difference flows, 2025. URL https://arxiv.org/abs/2503.09817.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for deep data-driven reinforcement learning, 2020. URL https://arxiv.org/abs/2004.07219.

Michael Groom, Victor-Alexandru Darvariu, Lars Kunze, James Wilson, and Nick Hawes. Quantile-coupled flow matching for distributional reinforcement learning, 2026. URL https://arxiv.org/abs/2605.08515.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2110.06169.

Yangming Li, Chieh-Hsin Lai, Carola-Bibiane Schönlieb, Yuki Mitsufuji, and Stefano Ermon. Bellman diffusion: Generative modeling as learning a linear operator in the distribution space, 2024. URL https://arxiv.org/abs/2410.01796.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.02747.

Yaron Lipman, Marton Havasi, Peter Holderrieth, Neta Shaul, Matt Le, Brian Karrer, Ricky T. Q. Chen, David Lopez-Paz, Heli Ben-Hamu, and Itai Gat. Flow matching guide and code, 2024. URL https://arxiv.org/abs/2412.06264.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.03003.

Yecheng Jason Ma, Dinesh Jayaraman, and Osbert Bastani. Conservative offline distributional reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 19235–19247, 2021.

Thanh Nguyen-Tang, Sunil Gupta, and Svetha Venkatesh. Distributional reinforcement learning via moment matching. In *AAAI Conference on Artificial Intelligence*, pp. 9144–9152, 2021.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. OGBench: Benchmarking offline goal-conditioned RL. In *International Conference on Learning Representations*, 2025a.

Seohong Park, Qiyang Li, and Sergey Levine. Flow Q-learning. In *International Conference on Machine Learning*, pp. 48104–48127. PMLR, 2025b.

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

Richard S. Sutton, Doina Precup, and Satinder Singh. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112(1–2):181–211, 1999.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024. URL https://arxiv.org/abs/2302.00482.

Boyang Xu, Qing Zou, Siqin Yang, and Hao Yan. Path-coupled Bellman flows for distributional reinforcement learning. In *International Conference on Machine Learning*, 2026.

Appendix A states the assumptions and the full results of Section 5, and Appendix B proves them. The later appendices give the toy comparisons, environments, protocol, algorithm and limitations. The integration budget and the per-task OGBench table are an online supplement.

A Standing Assumptions and Full Statements of Section 5

Section 5 states its results briefly. This appendix gives the assumptions they use and their full statements; proofs are in Appendix B.

Standing assumptions.

(A1) *Regularity.* The condition set is finite, $0 \leqslant \tilde{\gamma} \leqslant \bar{\gamma} < 1$ and $1 \leqslant p < \infty$; rewards and return laws have uniformly finite pth moments; a field is admissible when its flow from $\mathcal{N}(0, I_d)$ is unique on $[0, 1)$ with a terminal law of finite pth moment; and exact flow matching gives $\mathcal{E}\mathcal{L} = \mathrm{id}$.

(A2) *Teacher.* For centering, the teacher is the coupling-matched oracle of Eq. (12); for realisation, it is the critic's own field, with the successor endpoint drawn from $\mathcal{E}(v)_{S',A'}$ independently of $(X_0, R, \tilde{\gamma}, t)$ given (S', A').

(A3) *Gain.* $a_t = \kappa g(t)$ with $\kappa \in [0, 1]$ measurable with respect to $\sigma(X_t, H)$; it may not inspect residual chord randomness. A gain that does can destroy centering: if $(X_1' - X_0) - D$ is $+1$ or -1 equiprobably given $\mathcal{H}$ and $\kappa = \mathbf{1}\{(X_1' - X_0) - D > 0\}$, then $\mathbb{E}[\kappa g\{D - (X_1' - X_0)\} \mid \mathcal{H}] = -g/2 \neq 0$.

(A4) *Integrability.* The regression target is square-integrable.

(A5) *Sampling and conditioning.* Given the current condition c, the sampler draws the backup $H = (R, S', A', \tilde{\gamma})$, the training time $T \sim \mathcal{U}(0, 1)$ and the source $X_0 \sim \mathcal{N}(0, I_d)$; X_0 is independent of H, and T is independent of (X_0, X_1', H) given c. Conditional expectations are regular versions, jointly measurable in time, query and branch; in particular the oracle of Eq. (12) is a version $v^{\star}_{\rho_1}(u, b \mid h)$, jointly measurable in (u, b, h), that is for each h a conditional expectation under the conditional law of (X_0, X_1') given $H = h$. At fixed (c, t) we write $\mathcal{F} = \sigma(X_t)$ and $\mathcal{H} = \sigma(X_t, H)$, as in Eq. (19).

(A6) *Independent fresh generation.* The successor endpoint satisfies $X_1' \sim \mathcal{E}(v)_{S',A'}$ and X_1' is independent of $(X_0, R, \tilde{\gamma}, T)$ given (S', A'). This holds, for example, when $\rho_1 = 0$ in Eq. (11), E is independent of (X_0, H, T), and X_1' is the time-one map of the flow of $v_{S',A'}$ applied to E. The label $\rho_1 = 0$ alone does not imply (A6).

(A7) *Bridge flow validity.* Whenever an endpoint law is read off a bridge field, that is, off the conditional-mean velocity of a pair (X_0, X_1) under any coupling, the field is admissible in the sense of (A1) and the continuity equation of the bridge marginals has a unique solution among curves of probability laws with finite pth moments. For the independent canonical lift this is the content of $\mathcal{E}\mathcal{L} = \mathrm{id}$ in (A1). Differentiation under an expectation is used only for compactly supported smooth test functions or under a displayed integrable dominating function.

(A8) *Affine clocks.* In Theorem 5.2, φ and ϑ are absolutely continuous on $[0, 1)$, ϑ takes values in $[0, 1]$, $\varphi(0) = 1$ and $\vartheta(0) = 0$; the lifted field is the velocity of the lifted path. That $\varphi > 0$ and $\vartheta < 1$ on $[0, 1)$ is a consequence, proved there, not a hypothesis.

(A9) *Terminal convention.* If $\tilde{\gamma} = 0$, the target is $Y = R - X_0$ with zero correction, and no inverse involving $\alpha(1) = 0$ is evaluated. Clock identities are stated for $0 < \tilde{\gamma} < 1$.

(A5) and (A9) describe the algorithm's sampler and terminal convention (Section 4, Appendix G) and hold throughout; each of (A6)–(A8) is invoked only where a statement cites it.

The statements concern one frozen Bellman regression update: the policy, endpoint generator, teacher and coupling sampler are fixed while taking expectations. Source seeds are independent of H, and conditional on H the dependence of X_1' on X_0 is exactly the declared coupling. Squared-error statements require finite second moments, and the W_p contraction requires finite pth moments. We do not infer these regularity conditions from finite neural-network training.

Assumption (A1), rather than finite moments alone, is what asserts $\mathcal{E}\mathcal{L} = \mathrm{id}$, and it includes atomic return laws. If X_1 has a finite atomic law and X_0 is an independent Gaussian source, then for $t < 1$ the interpolant $(1-t)X_0 + tX_1$ has a positive smooth density, the weak continuity-equation calculation of Lemma B.1 applies, and the interpolant converges to X_1 in L^p as $t \uparrow 1$. Finite atomic toy laws therefore satisfy the endpoint part once the uniqueness clause of (A1) and (A7) is imposed.

Proposition 5.4 (Fixed point and contraction; full statement). Assume (A1). *(i)* $\mathcal{E}\,\mathcal{C}^\pi = \mathcal{T}^\pi\mathcal{E}$, $\mathcal{C}^\pi\mathcal{L} = \mathcal{L}\,\mathcal{T}^\pi$ and, for every integer $n \geqslant 1$, $(\mathcal{C}^\pi)^n = \mathcal{L}(\mathcal{T}^\pi)^n\mathcal{E}$. Every fixed point of $\mathcal{C}^\pi$ is canonical and its endpoint law is fixed by $\mathcal{T}^\pi$, so $v^\pi = \mathcal{L}\eta^\pi$ is the unique fixed point.

(ii) On canonical fields put $D_p(\mathcal{L}\eta, \mathcal{L}\nu) = \max_c W_p(\eta_c, \nu_c)$; on admissible fields put $\delta_p(v, w) = \max_c W_p(\mathcal{E}(v)_c, \mathcal{E}(w)_c)$, a pseudometric only, since distinct fields can share an endpoint law. Let $\beta_p = \max_c \mathbb{E}[\tilde{\gamma}^p \mid c]^{1/p} \leqslant \bar{\gamma}$. Whenever every multiplier is at most the nominal discount γ, as for $\tilde{\gamma} \in \{0, \gamma\}$, also $\beta_p \leqslant \gamma$. For admissible v, w and $n \geqslant 1$,

$$D_p(\mathcal{C}^\pi v, \mathcal{C}^\pi w) \leqslant \beta_p\, \delta_p(v, w), \qquad D_p\big((\mathcal{C}^\pi)^n v, v^\pi\big) \leqslant \beta_p^n\, \delta_p(v, v^\pi),$$

and on canonical fields $\delta_p = D_p$. Under $\tilde{\gamma} \leqslant \bar{\gamma}$ alone the constant $\bar{\gamma}$ cannot be improved.

Theorem 5.2 (Bellman conjugacy characterises the affine retiming; full statement). Fix a nonterminal branch with $0 < \tilde{\gamma} < 1$ and its coupling, and let v' be the successor bridge's flow-matching field under that same coupling. Among affine lifts $X_t = tr + \varphi(t)X'_{\vartheta(t)}$ normalised by $\varphi(0) = 1, \vartheta(0) = 0$, with φ, ϑ absolutely continuous on $[0, 1)$ and ϑ valued in $[0, 1]$ as in (A8), read through the endpoint readout q_v of Definition 5.1:

(i) with $\xi = (z - tr)/\varphi(t)$ and $\mathfrak{B}^{\varphi,\vartheta}v$ the field of the lifted bridge, the conjugacy $q_{\mathfrak{B}^{\varphi,\vartheta}v} = r + \tilde{\gamma}\, q_v \circ (\xi, \vartheta)$ holds for every field v *if and only if* $(\varphi, \vartheta) = (\alpha, \tau)$ of Eq. (6); an interior same-time query forces $\tilde{\gamma} = 1$.

(ii) For that unique pair, with $\Phi_{r,\tilde{\gamma}}(z, t) = \big((z - tr)/\alpha(t), \tau(t)\big) =: (\xi, \tau)$, the field of the backed-up bridge is

$$(\mathfrak{B}_{r,\tilde{\gamma}}v')(z, t) = r - (1 - \tilde{\gamma})\,\xi + g(t)\, v'(\xi, \tau), \tag{17}$$

which for canonical v' is $\mathcal{C}^\pi$ on that branch.

(iii) In readout coordinates the operator is the affine Bellman map, and inherits its contraction: with $d_q(v, w) := \sup_{z,\,t<1} \|q_v - q_w\| = \sup_{z,t}(1 - t)\,\|v - w\|$,

$$q_{\mathfrak{B}_{r,\tilde{\gamma}}v}(z, t) = r + \tilde{\gamma}\, q_v\big(\Phi_{r,\tilde{\gamma}}(z, t)\big), \qquad d_q(\mathfrak{B}_{r,\tilde{\gamma}}v, \mathfrak{B}_{r,\tilde{\gamma}}w) = \tilde{\gamma}\, d_q(v, w) \tag{18}$$

the second whenever $d_q(v, w) < \infty$, as an exact equality rather than a bound.

The full statements below use $\mathcal{F} = \sigma(X_t)$ and $\mathcal{H} = \sigma(X_t, H)$. Fixing c, t,

$$A_t = R - (1 - \tilde{\gamma})\xi_t, \quad D = v^\star_{\rho_1}(\tau, \xi_t \mid H), \quad w = \mathbb{E}[Y \mid \mathcal{H}] = A_t + gD. \tag{19}$$

Here A_t is the part of the displacement the transition determines, D the matched oracle of Eq. (12) at the retimed point, and w the ideal target given everything the backup reveals. Part (i) below requires no Gaussian returns, no independent source–endpoint pair, and no deterministic rewards.

Theorem 5.3 (Conditional centering and realisation; full statement). *(i) Centering.* Suppose the teacher is Eq. (12), a_t is $\mathcal{H}$-measurable, and the target is square-integrable. For every nonterminal backup,

$$\mathbb{E}[C_{\tau(t)} \mid \mathcal{H}] = 0, \qquad \mathbb{E}[u_a \mid \mathcal{F}] = \mathbb{E}[Y \mid \mathcal{F}] = v^\star_t(X_t \mid c). \tag{20}$$

Thus every such gain has the same population regression target. Independence is a sufficient way to match an independent-CFM teacher, but it is not necessary.

(ii) Realisation. Assume (A5) and independent fresh generation (A6): $X'_1 \sim \mathcal{E}(v)_{S',A'}$, independent of $(X_0, R, \tilde{\gamma}, t)$ given (S', A'), as at $\rho_1 = 0$ in Eq. (11); and let the teacher be the critic's own field v. Then

$$\mathbb{E}[u_\kappa \mid X_t, t, c] = (\mathcal{C}^\pi v)_c(X_t, t) + \mathbb{E}\big[\kappa g\,(v - \mathcal{L}\mathcal{E}v)_{S',A'}(\xi_t, \tau) \;\big|\; X_t, t, c\big], \tag{21}$$

so on canonical fields the population regression target is $\mathcal{C}^\pi v$ for every admissible κ, and the expected squared loss is $\mathbb{E}\,\|f - \mathcal{C}^\pi v\|^2$ plus a term that does not depend on the regressor f.

(iii) Fixed point and contraction. On canonical fields the self-consistent iteration is therefore $\mathcal{C}^\pi$-iteration under any admissible gains, with fixed point v^π and rate β_p.

Proposition 5 (Exact local risk and teacher-error bias; full statement). Write the frozen teacher as $\bar{v}(\tau, \xi_t \mid H) = D + e$, so that e, measurable under $\mathcal{H}$, is exactly its error against the matched oracle, and let $S = \mathbb{E}[\|(X_1' - X_0) - D\|^2 \mid \mathcal{H}]$ be the residual bridge noise the sampled chord carries about that oracle.

(i) Local risk. For any $\mathcal{H}$-measurable κ,

$$u_\kappa - w = (1-\kappa)g\{(X_1' - X_0) - D\} + \kappa g e, \tag{22}$$

$$\mathbb{E}[\|u_\kappa - w\|^2 \mid \mathcal{H}] = g^2\{(1-\kappa)^2 S + \kappa^2 \|e\|^2\}. \tag{23}$$

The oracle risk-minimizing fraction is $\kappa^\star = S/(S + \|e\|^2)$ when the denominator is positive. So κ interpolates between the two error sources by their relative size, and $\kappa = 1$ is not a premise but the $e \to 0$ corollary: only with an exact teacher does full correction remove all within-transition bridge noise at no cost.

(ii) Teacher error and population bias. For a learned teacher with $\kappa g e \in L^2$, the population field is

$$v_\kappa(z) = v_t^\star(z) + b_t(z), \qquad b_t(z) = \mathbb{E}[\kappa g e \mid X_t = z, c], \qquad \|b_t\|_{L^2} \leqslant \|\kappa g e\|_{L^2}. \tag{24}$$

If the teacher was calibrated under another coupling $\widetilde{\rho}_1$, its error separates as

$$e = \underbrace{\bar{v} - v^\star_{\widetilde{\rho}_1}}_{\text{teacher approximation}} + \underbrace{v^\star_{\widetilde{\rho}_1} - v^\star_{\rho_1}}_{\text{coupling mismatch}}, \tag{25}$$

with all fields evaluated at the *actual* retimed query, so the norms must be measured under that query law rather than on the teacher's training distribution. Retiming removes the intrinsic query mismatch; it removes neither term in Eq. (25).

Proposition A.1 (Shared noise: exact under its matched teacher, which the reused critic is not)**.** *Let the transition be stochastic and $\rho_1 = 1$. (i) If the chord $X_1' - X_0$ is measurable with respect to the query and H, the coupling-matched teacher returns the sampled chord, so the corrected target is the sample velocity of Eq. (7) at every gain and, under (A7), the population flow reaches the true Bellman law, at either clock. In particular, although conditioning on the current bridge point as $t \to 0$ conditions on the shared source alone, so that the fitted early-time velocity is the shared-seed average of the branch maps — in one dimension with increasing branch maps, the initial velocity toward the W_2 barycenter of the Bellman branch endpoint laws — this does not fix the terminal law: branch information can re-enter through the posterior at later times. (ii) The critic the algorithm reuses is a field v whose time-one flow map is the generator ψ. If the chord is recoverable for almost every query and the flow of v is unique, v equals the matched teacher almost everywhere on the query law only if the trajectories of v are straight lines traversed at constant speed, which a flow-matching field's trajectories generally are not.*

Its proof is in Appendix B.9.

B Proofs

Proof of Corollary 3.1. The identity must hold for every endpoint pair, so the coefficients of X_0 and X_1' in $tR + \alpha(t)X_t'$ and in Eq. (3), expanded with $X_1 = R + \tilde{\gamma}X_1'$, match separately: $\alpha(1-t) = 1-t$ and $\alpha t = \tilde{\gamma} t$. For $0 < t < 1$ they force $\alpha = 1$ and $\alpha = \tilde{\gamma}$, incompatible for $0 < \tilde{\gamma} < 1$. □

B.1 Notation

Appendix shorthand. In the dense algebra below we abbreviate the sampled chord as $C := X_1' - X_0$. This is the *chord*, not the control variate: the control variate is $C_{\tau(t)} = \bar{v}(\tau(t), X'_{\tau(t)} \mid H) - C$ of Eq. (8), which always carries a subscript. The multiplier is always written $\tilde{\gamma}$, and q_v always denotes a readout (Table 4).

B.2 Reusable lemmas

Lemma B.1 (Flow matching for any coupling, including atomic endpoints)**.** *Let (X_0, X_1) be any pair with finite first moments, $Z_t = (1-t)X_0 + tX_1$, and let $v_t(z) = \mathbb{E}[X_1 - X_0 \mid Z_t = z]$ be a*

Table 4: Notation and distinctions used throughout the analysis.

Symbol	Meaning
$c = (s, a)$	Current state–action conditioning; fixed in local derivations.
$H = (R, S', A', \tilde{\gamma})$	Observed backup and its actual Bellman multiplier.
X_0, E, X_0'	Bridge source, independent auxiliary seed, and successor source.
$X_1', C = X_1' - X_0$	Successor endpoint and successor bridge displacement (the chord).
$\tau(t)$	Query clock. Same-time: $\tau(t) = t$. ReBF: Eq. (6).
$X'_{\tau(t)}, X_t$	Successor interpolant at the query clock, and current Bellman interpolant.
α, g	Affine scale and geometric gain of the retimed clock.
$\kappa, a = \kappa g$	Bootstrap fraction and effective velocity coefficient.
D, w	Coupling-matched successor prediction and $\mathbb{E}[Y \mid X_t, H]$.
e, S	Teacher error at the actual query and conditional displacement energy.
$\rho_1, \widetilde{\rho}_1$	Training-bridge coupling and teacher-reference coupling.
q_v	Readout $z + (1-t)v(z,t)$ of a field v (Definition 5.1).
β_p	Endpoint Wasserstein modulus (Proposition 5.4).
$\beta_\rho(u)$	Slope of the correlated-Gaussian teacher, Eq. (29); a function of time, not β_p.

jointly measurable version. Then $p_t = \mathrm{Law}(Z_t)$ solves $\partial_t p_t + \nabla \cdot (p_t v_t) = 0$ weakly on $[0,1)$. If the flow of v is unique and this equation has a unique solution from p_0 in the class of (A7), the flow has marginals p_t. If $X_0, X_1 \in L^p$, then $W_p(p_t, \mathrm{Law}(X_1)) \to 0$ as $t \uparrow 1$.

Proof. For $\phi \in C_c^1(\mathbb{R}^d)$, $\left\|\frac{d}{dt}\phi(Z_t)\right\| \leqslant \sup\|\nabla\phi\|\ \|X_1 - X_0\| \in L^1$, so dominated differentiation and the tower property give

$$\frac{d}{dt}\mathbb{E}[\phi(Z_t)] = \mathbb{E}[\nabla\phi(Z_t)^\top (X_1 - X_0)] = \mathbb{E}[\nabla\phi(Z_t)^\top \mathbb{E}[X_1 - X_0 \mid Z_t]] = \int \nabla\phi(z)^\top v_t(z)\, p_t(dz),$$

which is the weak continuity equation. The marginals of the flow solve the same equation from p_0, so uniqueness identifies them with p_t. Finally $W_p(p_t, \mathrm{Law}(X_1))^p \leqslant \mathbb{E}\,\|Z_t - X_1\|^p = (1-t)^p \mathbb{E}\,\|X_0 - X_1\|^p \to 0$. No density of X_1 is used, so atoms are covered. □

Lemma B.2 (Clock identities). *Let $0 < \tilde{\gamma} < 1$. The identity*

$$(1-t)x_0 + t(r + \tilde{\gamma}x_1') = tr + \alpha(t)\big\{(1-\tau(t))x_0 + \tau(t)x_1'\big\} \tag{26}$$

holds for every $(x_0, x_1') \in \mathbb{R}^d \times \mathbb{R}^d$ if and only if

$$\alpha(t) = 1 - (1-\tilde{\gamma})t, \qquad \tau(t) = \frac{\tilde{\gamma}t}{\alpha(t)}, \qquad \text{and then} \qquad g(t) := \alpha(t)\tau'(t) = \frac{\tilde{\gamma}}{\alpha(t)}.$$

Moreover $\alpha(0) = 1$, $\tau(0) = 0$, $\alpha(1) = \tilde{\gamma}$, $\tau(1) = 1$, and

$$t - \tau(t) = \frac{(1-\tilde{\gamma})\,t(1-t)}{\alpha(t)}, \qquad \max_{0\leqslant t\leqslant 1}\big\{t - \tau(t)\big\} = \frac{1-\sqrt{\tilde{\gamma}}}{1+\sqrt{\tilde{\gamma}}} \ \text{ at } \ t = \frac{1}{1+\sqrt{\tilde{\gamma}}},$$

$$g'(t) = \frac{\tilde{\gamma}(1-\tilde{\gamma})}{\alpha(t)^2} \geqslant 0, \qquad \tilde{\gamma} \leqslant g(t) \leqslant 1, \qquad \frac{\tau}{1-\tau} = \tilde{\gamma}\,\frac{t}{1-t}, \qquad g^2\,dt = \tilde{\gamma}\,d\tau,$$

$$1 - \tau = \frac{1-t}{\alpha}, \qquad (1-t)\,g = \tilde{\gamma}\,(1-\tau).$$

At fixed c and $t \leqslant 1$, the retimed query $\xi_t = (X_t - tR)/\alpha_H(t)$ is $\mathcal{H}$-measurable and $\sigma(X_t, H) = \sigma(\xi_t, H)$.

Proof. Matching the coefficients of x_0 and x_1' in Eq. (26) gives $\alpha(1-\tau) = 1 - t$ and $\alpha\tau = \tilde{\gamma}t$; adding them gives α, and dividing gives τ; conversely these values satisfy both. Then $\alpha' = -(1-\tilde{\gamma})$ and $\tau' = \tilde{\gamma}/\alpha^2$, so $g = \alpha\tau' = \tilde{\gamma}/\alpha$ and $g^2 dt = (\tilde{\gamma}^2/\alpha^2)(\alpha^2/\tilde{\gamma})\,d\tau = \tilde{\gamma}\,d\tau$. Subtraction gives $t - \tau = \{t\alpha - \tilde{\gamma}t\}/\alpha = (1-\tilde{\gamma})t(1-t)/\alpha$, whose derivative is $(1-\tilde{\gamma})\{(1-t)^2 - \tilde{\gamma}t^2\}/\alpha^2$; its only zero in $(0,1)$ is $1 - t = \sqrt{\tilde{\gamma}}\,t$, i.e. $t = 1/(1+\sqrt{\tilde{\gamma}})$, where $\alpha = \sqrt{\tilde{\gamma}}$ and the value is

$(1-\tilde{\gamma})\sqrt{\tilde{\gamma}}/\{(1+\sqrt{\tilde{\gamma}})^2\sqrt{\tilde{\gamma}}\} = (1-\sqrt{\tilde{\gamma}})/(1+\sqrt{\tilde{\gamma}})$; the endpoint values are zero, so this is the maximum, and it is $O(1-\tilde{\gamma})$ as $\tilde{\gamma} \uparrow 1$. Differentiating $g = \tilde{\gamma}/\alpha$ gives g', and $g(0) = \tilde{\gamma}$, $g(1) = 1$ give the bounds. The identity $1-\tau = (1-t)/\alpha$ follows from $\alpha - \tilde{\gamma}t = 1-t$; dividing $\tau = \tilde{\gamma}t/\alpha$ by it gives the odds identity; and multiplying it by g gives $(1-t)g = \tilde{\gamma}(1-\tau)$. Finally, given (t, H) the maps $X_t \mapsto (X_t - tR)/\alpha_H$ and $\xi \mapsto tR + \alpha_H \xi$ are mutually inverse measurable maps because $\alpha_H(t) \geqslant \tilde{\gamma} > 0$. □

Sample velocity. At fixed sampled endpoints, $X_t = tR + \alpha(t)X'_{\tau(t)}$ with $X'_u = (1-u)X_0 + uX'_1$ has the velocity of Eq. (7), $Y = \dot{X}_t = R - (1-\tilde{\gamma})\xi_t + g(t)\,C$: differentiating, $\dot{X}_t = R + \alpha' X'_\tau + \alpha\tau' \frac{d}{du}X'_u\big|_{u=\tau} = R - (1-\tilde{\gamma})\xi_t + gC$, and $\dot{X}_t = R + \tilde{\gamma}X'_1 - X_0$ directly from Eq. (3). This proves Eq. (7) directly, without substituting a model field for a sample derivative. That substitution is justified statistically by Theorem 5.3, not by the chain rule alone.

B.3 The population operator

Proof of Proposition 5.4. (i) Since $\mathcal{E}\mathcal{L} = \mathrm{id}$ by (A1), $\mathcal{E}\mathcal{C}^\pi = \mathcal{E}\mathcal{L}\mathcal{T}^\pi\mathcal{E} = \mathcal{T}^\pi\mathcal{E}$ and $\mathcal{C}^\pi\mathcal{L} = \mathcal{L}\mathcal{T}^\pi\mathcal{E}\mathcal{L} = \mathcal{L}\mathcal{T}^\pi$. If $(\mathcal{C}^\pi)^n = \mathcal{L}(\mathcal{T}^\pi)^n\mathcal{E}$ for some $n \geqslant 1$, then $(\mathcal{C}^\pi)^{n+1} = \mathcal{L}\mathcal{T}^\pi(\mathcal{E}\mathcal{L})(\mathcal{T}^\pi)^n\mathcal{E} = \mathcal{L}(\mathcal{T}^\pi)^{n+1}\mathcal{E}$; the identity fails at $n = 0$ off canonical fields, since $v \neq \mathcal{L}\mathcal{E}v$ in general. A fixed point $v = \mathcal{C}^\pi v$ lies in the range of $\mathcal{L}$, and applying $\mathcal{E}$ gives $\mathcal{E}(v) = \mathcal{T}^\pi\mathcal{E}(v)$, so $\mathcal{E}(v) = \eta^\pi$ by uniqueness of the Bellman fixed law (Banach, using the contraction below on the complete product of p-Wasserstein spaces), and $v = \mathcal{L}\mathcal{T}^\pi\eta^\pi = \mathcal{L}\eta^\pi$; conversely $\mathcal{L}\eta^\pi$ is fixed.

(ii) Fix c and use the same draw of H in two backups. Given $(S', A') = c'$, couple $X \sim \eta_{c'}$ and $\widetilde{X} \sim \nu_{c'}$ optimally in W_p, independently of $(R, \tilde{\gamma})$ given c'; this is a coupling of $(\mathcal{T}^\pi\eta)_c$ and $(\mathcal{T}^\pi\nu)_c$ in which the reward cancels, so

$$W_p^p\big((\mathcal{T}^\pi\eta)_c, (\mathcal{T}^\pi\nu)_c\big) \leqslant \mathbb{E}\big[\tilde{\gamma}^p \left\|X - \widetilde{X}\right\|^p \mid c\big] = \mathbb{E}\big[\tilde{\gamma}^p W_p^p(\eta_{S',A'}, \nu_{S',A'}) \mid c\big]$$
$$\leqslant \mathbb{E}[\tilde{\gamma}^p \mid c] \max_{c'} W_p^p(\eta_{c'}, \nu_{c'}).$$

Take the maximum over c and the pth root, and apply this to $\eta = \mathcal{E}(v)$, $\nu = \mathcal{E}(w)$; since $\mathcal{C}^\pi v = \mathcal{L}\mathcal{T}^\pi\mathcal{E}(v)$ is canonical, $D_p(\mathcal{C}^\pi v, \mathcal{C}^\pi w) = \max_c W_p((\mathcal{T}^\pi\mathcal{E}v)_c, (\mathcal{T}^\pi\mathcal{E}w)_c)$. The n-step bound follows from (i) by induction, and $\delta_p = D_p$ on canonical fields because $\mathcal{E}\mathcal{L} = \mathrm{id}$. With one condition, $R = 0$, $\tilde{\gamma} \equiv \bar{\gamma}$ and two point masses, W_p is multiplied by exactly $\bar{\gamma}$.

The short form, Eq. (16), is (i)–(ii) read in δ_p: part (i) gives $\mathcal{E}\mathcal{C}^\pi = \mathcal{T}^\pi\mathcal{E}$ and the unique fixed point, and since $\mathcal{C}^\pi v$ and $\mathcal{C}^\pi w$ are canonical, $\delta_p(\mathcal{C}^\pi v, \mathcal{C}^\pi w) = D_p(\mathcal{C}^\pi v, \mathcal{C}^\pi w)$, so part (ii) gives Eq. (16). The bound $\beta_p \leqslant \bar{\gamma}$ holds by (A1), and $\beta_p \leqslant \gamma$ whenever every multiplier is at most the nominal discount, as for $\tilde{\gamma} \in \{0, \gamma\}$. □

B.4 Conditional centering and realisation

Proof of Theorem 5.3(i). Fix c, t and a nonterminal backup. By Lemma B.2, $\tau_H(t)$ is a function of H and $\sigma(X_t, H) = \sigma(X'_{\tau_H}, H)$ with $X'_{\tau_H} = \xi_t$. Given $H = h$ the clock is the constant τ_h, and the version of Eq. (12) fixed in (A5) is, for each h, a conditional expectation under the conditional law of (X_0, X'_1) given $H = h$; evaluating it at the H-measurable clock therefore gives

$$D = v^\star_{\rho_1}(\tau_H, \xi_t \mid H) = \mathbb{E}[C \mid X'_{\tau_H}, H] = \mathbb{E}[C \mid \mathcal{H}].$$

With this teacher $C_{\tau(t)} = D - C$, so $\mathbb{E}[C_{\tau(t)} \mid \mathcal{H}] = 0$. The gain a_t is $\mathcal{H}$-measurable, and $a_t(D - C) = u_a - Y$ is integrable because the target is square-integrable and Y is integrable by (A1) (under (A3) $a_t = \kappa g$ is simply bounded by one), so, $D - C$ being integrable as well, pulling out the $\mathcal{H}$-measurable factor (Durrett, 2019) gives $\mathbb{E}[u_a - Y \mid \mathcal{H}] = \mathbb{E}[a_t(D-C) \mid \mathcal{H}] = a_t\mathbb{E}[D - C \mid \mathcal{H}] = 0$, and the tower property from $\mathcal{H}$ to $\mathcal{F}$ gives $\mathbb{E}[u_a \mid \mathcal{F}] = \mathbb{E}[Y \mid \mathcal{F}] = v^\star_t(X_t \mid c)$, the flow-matching field of the current bridge. Terminal backups use $Y = R - X_0$ with zero correction (A9), for which $u_a = Y$, so the identity holds for the whole mixture of backups. With a random training time, (A5) makes conditioning on $T = t$ reproduce the fixed-t law, so the same identities hold with $\sigma(c, T, X_T)$ and $\sigma(c, T, X_T, H)$. No Gaussian returns, independent source–endpoint pair or deterministic rewards are used. Independence is sufficient for the independent-CFM teacher to be the matched one (Lemma B.3) but not necessary for centering, which holds for every coupling with

its own matched teacher. For a dependent example take $X_1' = 2X_0$ in one dimension: at the query $X_u' = (1+u)X_0$ the chord $C = X_0$ is recoverable, so the matched oracle equals it and centering is exact despite perfect dependence. □

Lemma B.3 (At independent generation the critic is the matched teacher)**.** *Assume (A5) and (A6) at every backup. For every (S', A'), the population flow-matching field of $\eta_{S',A'} = \mathcal{E}(v)_{S',A'}$ under an independent source — the field the critic at (S', A') is trained to represent — equals the matched oracle $v^\star_{\rho_1}(\tau, \xi \mid H)$ of Eq. (12) for every branch H whose successor is (S', A'), for every return law and every transition kernel.*

Proof. Write $G = (S', A')$ and $W = (R, \tilde\gamma)$. By (A6), $\Pr(X_1' \in B \mid X_0, W, G) = \Pr(X_1' \in B \mid G) = \mathcal{E}(v)_G(B)$, and by (A5) X_0 is independent of $H = (W, G)$. Hence

$$\begin{aligned}\Pr(X_0 \in A,\ X_1' \in B \mid H) &= \mathbb{E}\big[\mathbf{1}\{X_0 \in A\} \Pr(X_1' \in B \mid X_0, W, G) \,\big|\, H\big] \\ &= \mathcal{N}(0, I_d)(A)\, \mathcal{E}(v)_G(B)\,:\end{aligned}$$

given H, the pair (X_0, X_1') is independent with the same law for every branch ending at G. Conditioning its chord on $(1-u)X_0 + uX_1'$ therefore defines, for almost every query, the same regular conditional expectation as the successor critic's own independent training bridge, which is $(\mathcal{L}\eta)_G$. □

Proof of Theorem 5.3(ii)–(iii). Under (A5) and (A6), Lemma B.3 gives $\mathbb{E}[C \mid X_t, t, c, H] = (\mathcal{L}\mathcal{E}v)_{S',A'}(\xi_t, \tau)$, the flow-matching field of the successor's generated law at the retimed query. Since κ is $\mathcal{H}$-measurable,

$$\mathbb{E}[u_\kappa \mid X_t, t, c, H] = R - (1-\tilde\gamma)\xi_t + g\big\{(1-\kappa)(\mathcal{L}\mathcal{E}v) + \kappa v\big\}_{S',A'}(\xi_t, \tau).$$

Replacing the last v by $\mathcal{L}\mathcal{E}v$ leaves $\mathbb{E}[Y \mid X_t, t, c, H]$ (Eq. (7)). By the proof of Lemma B.3, X_0 is independent of (H, X_1'), hence of the endpoint $R + \tilde\gamma X_1'$, whose law given c is $(\mathcal{T}^\pi\mathcal{E}(v))_c$; so $\mathbb{E}[Y \mid X_t, t, c]$ is the independent-source flow-matching field of that law, namely $(\mathcal{C}^\pi v)_c$. What remains is $\mathbb{E}[\kappa g(v - \mathcal{L}\mathcal{E}v)_{S',A'}(\xi_t, \tau) \mid X_t, t, c]$, which is Eq. (21). For square-integrable u_κ (A4), orthogonality of conditional expectation, $\mathbb{E}\,\|f - U\|^2 = \mathbb{E}\,\|f - M\|^2 + \mathbb{E}\,\|U - M\|^2$ for $U \in L^2$, $M = \mathbb{E}[U \mid \mathcal{G}]$ and square-integrable $\mathcal{G}$-measurable f (Durrett, 2019), applied with $U = u_\kappa$ and $\mathcal{G} = \sigma(X_t, t, c)$, gives the loss decomposition. If v is canonical the defect vanishes, and $\mathcal{C}^\pi$ maps every admissible field to a canonical one, so the population iterates follow $\mathcal{C}^\pi$ and Proposition 5.4 gives the fixed point v^π and the rate β_p. □

B.5 Bellman conjugacy

Proof of Theorem 5.2. (i) Under (A8), let $z = tr + \varphi(t)\xi$ on the interval where $\varphi \neq 0$. Differentiating the lifted path $tr + \varphi(t)X'_{\vartheta(t)}$ gives the lifted field $(\mathfrak{B}^{\varphi,\vartheta}v)(z,t) = r + \varphi'\xi + \varphi\vartheta' v(\xi, \vartheta)$, whose readout is

$$q_{\mathfrak{B}^{\varphi,\vartheta}v}(z,t) = r + \{\varphi + (1-t)\varphi'\}\xi + (1-t)\varphi\vartheta'\, v(\xi, \vartheta),$$

while the required right side is $r + \tilde\gamma q_v(\xi, \vartheta) = r + \tilde\gamma\xi + \tilde\gamma(1-\vartheta)v(\xi, \vartheta)$. Since equality must hold for every ξ and every field v, the coefficients match separately:

$$\varphi + (1-t)\varphi' = \tilde\gamma, \qquad (1-t)\,\varphi\,\vartheta' = \tilde\gamma\,(1-\vartheta) \tag{27}$$

for almost every t. The first is linear with absolutely continuous solutions $\varphi = \tilde\gamma + A(1-t)$, and $\varphi(0) = 1$ gives $A = 1 - \tilde\gamma$, i.e. $\varphi = \alpha \geqslant \tilde\gamma > 0$; in particular φ cannot vanish, since on the largest interval $[0, t^*)$ where it is nonzero it equals α, which is positive at t^*. The second then reads $\vartheta'/(1-\vartheta) = \tilde\gamma/[(1-t)\alpha]$, and since $\frac{d}{dt}\log\frac{1-t}{\alpha} = -\tilde\gamma/[(1-t)\alpha]$, the unique absolutely continuous solution with $\vartheta(0) = 0$ satisfies $1 - \vartheta = (1-t)/\alpha > 0$, i.e. $\vartheta = \tilde\gamma t/\alpha = \tau < 1$. Conversely $\alpha - (1-t)(1-\tilde\gamma) = \tilde\gamma$ and $(1-t)g = \tilde\gamma(1-\tau)$ (Lemma B.2) verify both equations. If $\vartheta(t) = t$ at an interior time, then $\tau(t) = t$, and $t - \tau(t) = (1-\tilde\gamma)t(1-t)/\alpha(t)$ (Lemma B.2) vanishes only if $\tilde\gamma = 1$: a same-time query forces $\tilde\gamma = 1$.

(ii) For (α, τ) the lifted field is $r - (1-\tilde\gamma)\xi + g\,v'(\xi, \tau)$, Eq. (17). If $v' = \mathcal{L}\eta'$ is canonical, the canonical readout of Definition 5.1 and Lemma B.2 give, with $X' \sim \eta'$ independent of X_0 and $X_\tau' = (1-\tau)X_0 + \tau X'$,

$$q_{\mathfrak{B}v'}(z,t) = r + \tilde\gamma\,\mathbb{E}[X' \mid X_\tau' = \xi] = \mathbb{E}\big[r + \tilde\gamma X' \,\big|\, (1-t)X_0 + t(r + \tilde\gamma X') = z\big],$$

because $(1-t)X_0 + t(r+\tilde{\gamma}X') = tr + \alpha X'_\tau$. This is the canonical readout of $\mathrm{Law}(r+\tilde{\gamma}X')$, and $v \mapsto q_v$ is injective for $t < 1$, so $\mathfrak{B}v'$ is the canonical field of the backed-up law: $\mathcal{C}^\pi$ on that branch.

(iii) The readout identity is the case $(\varphi, \vartheta) = (\alpha, \tau)$ of the display above. Subtracting it for v and w gives $q_{\mathfrak{B}v} - q_{\mathfrak{B}w} = \tilde{\gamma}\,(q_v - q_w) \circ \Phi_{r,\tilde{\gamma}}$ pointwise. The map $\Phi_{r,\tilde{\gamma}}$ is a bijection of $\mathbb{R}^d \times [0,1)$: the spatial map is affine with positive scale α, and τ is strictly increasing from 0 to 1 by the odds identity. The supremum therefore transfers, and $d_q(\mathfrak{B}v, \mathfrak{B}w) = \tilde{\gamma}\, d_q(v,w)$ as an equality in $[0,\infty]$, of finite quantities whenever $d_q(v,w) < \infty$.

□

Flow transfer at full gain. Fix one nonterminal branch that is deterministic given c, with reward r, successor condition c' and multiplier $0 < \tilde{\gamma} < 1$, and let an admissible field v at c', with time-one map ψ_v, serve as both teacher and endpoint generator. For every coupling ρ_1, the retimed $\kappa = 1$ target equals $(\mathfrak{B}_{r,\tilde{\gamma}}v)(X_t, t)$, so $\mathfrak{B}_{r,\tilde{\gamma}}v$ is a version of the population field: at $\kappa = 1$ the chord cancels and the target is $r - (1-\tilde{\gamma})\xi_t + g\, v(\xi_t, \tau)$ with $\xi_t = (X_t - tr)/\alpha$, a function of (X_t, t) because the branch is deterministic, and therefore its own conditional mean. Its flow is the affine pushforward of the flow of v. Let x_u solve $\dot{x}_u = v(x_u, u)$, $x_0 = x$, and put $z_t = tr + \alpha(t)x_{\tau(t)}$. Then $(z_t - tr)/\alpha = x_{\tau(t)}$ and

$$\dot{z}_t = r + \alpha' x_\tau + \alpha\tau'\, v(x_\tau, \tau) = r - (1-\tilde{\gamma})x_\tau + g\, v(x_\tau, \tau) = (\mathfrak{B}_{r,\tilde{\gamma}}v)(z_t, t)$$

by Eq. (6) and $g = \alpha\tau'$ (Lemma B.2), so z is a flow line of $\mathfrak{B}_{r,\tilde{\gamma}}v$ from $z_0 = x$. Conversely, a solution z of that equation gives $x_u = (z_{t(u)} - t(u)r)/\alpha(t(u))$, with $t(u)$ the inverse of τ, solving the equation of v; so uniqueness transfers. As $t \uparrow 1$, $z_t \to r + \tilde{\gamma}x_1 = r + \tilde{\gamma}\psi_v(x)$. The output endpoint law is therefore exactly $\mathrm{Law}(r + \tilde{\gamma}\psi_v(X_0)) = (\mathcal{T}^\pi\mathcal{E}(v))_c$, and its mean is exact, however curved the flow of v.

B.6 Fixed points, risk and teacher error

Corollary B.4 (Endpoint semiconjugacy and fixed-point correctness)**.** (a) Endpoint semiconjugacy for any coupling. *Let $\mathcal{M}^{\mathrm{or}}_{a,\chi}v$ be the population regression field at condition c formed with a source–endpoint coupling χ in which $X'_1 \mid H \sim \mathcal{E}(v)_{S',A'}$ (the coupling may correlate X'_1 with X_0), the coupling-matched teacher of χ, and a gain satisfying (A3). Assume (A4), (A5), (A9), and (A7) for the output field. Then*

$$\mathcal{E}\big(\mathcal{M}^{\mathrm{or}}_{a,\chi}v\big)_c = \big(\mathcal{T}^\pi\mathcal{E}(v)\big)_c. \tag{28}$$

For a non-independent coupling this is an endpoint statement only: it neither identifies the output field with $\mathcal{C}^\pi v$ nor gives field-level uniqueness.

(b) Fixed-point correctness. *Assume (A4), (A5), and (A7) for the actual, possibly coupled, bridge. Let the teacher be the coupling-matched oracle of Eq. (12), and let the successor endpoint satisfy $X'_1 \mid H \sim \eta^\pi_{S',A'}$ with $\eta^\pi = \mathcal{T}^\pi\eta^\pi$. At every admissible gain and every ρ_1, the retimed population update produces a field whose endpoint law is η^π: the Bellman fixed law is preserved, whatever the reward law and the transition kernel. Under independent fresh generation (A6), that field is the canonical $v^\pi = \mathcal{L}\eta^\pi$, which is therefore a fixed point; for a non-independent coupling it is the field of the coupled bridge, which in general differs from v^π while sharing its endpoint law. For the same-time update the endpoint conclusion holds if $\rho_1 = 1$ and the chord C is measurable with respect to $\sigma(X'_t, H)$ for almost every t, or $\kappa = 0$, or $\tilde{\gamma} = 1$; in the correlated-Gaussian model of Proposition B.5 these are the only cases in which the same-time field has zero residual, the residual being Eq. (32). For scalar returns a generator that is the time-one map of a unique one-dimensional flow gives the measurability for every $t < 1$, since that map is increasing, so $(1-t)x + t\psi(x)$ is injective; a deterministic generator does not give it in general.*

Proof. (a) By Theorem 5.3(i), with the conditioning on the training time justified by (A5), the population regression field is $m_c(z,t) = \mathbb{E}[X_1 - X_0 \mid X_t = z, t, c]$ with $X_1 = R + \tilde{\gamma}X'_1$ (and $X_1 = R$ on terminal branches): the flow-matching field of the actual current bridge under whatever joint law χ induces. Lemma B.1 and (A7) give $\mathcal{E}(m)_c = \mathrm{Law}(R + \tilde{\gamma}X'_1 \mid c)$, and since $X'_1 \mid H \sim \mathcal{E}(v)_{S',A'}$ this is $(\mathcal{T}^\pi\mathcal{E}(v))_c$. A coupled bridge field need not be canonical, so nothing equates the output fields themselves.

(b) Part (a), applied with η^π in the role of $\mathcal{E}(v)$, gives the endpoint law $(\mathcal{T}^\pi\eta^\pi)_c = \eta^\pi_c$; this needs $\eta^\pi = \mathcal{T}^\pi\eta^\pi$, since centering alone identifies only the one-step backed-up field. Under (A6) the matched oracle is the independent-source field of η^π (Lemma B.3), so Theorem 5.3(ii) with the canonical input v^π identifies the field with $\mathcal{L}\mathcal{T}^\pi\eta^\pi = \mathcal{L}\eta^\pi$. Coupled field equality fails in general: take one condition, $\tilde\gamma = \frac{1}{2}$, $R \sim \mathcal{N}(0, \frac{3}{4})$ independent of X_0, and $\eta^\pi = \mathcal{N}(0,1)$. The canonical flow of $\mathcal{N}(0,1)$ from an independent $\mathcal{N}(0,1)$ source has time-one map $\psi(x) = x$, so $\rho_1 = 1$ gives $X'_1 = X_0$, a recoverable chord, and $B = R + \frac{1}{2}X_0 \sim \mathcal{N}(0,1)$ with $\mathrm{Cov}(X_0, B) = \frac{1}{2}$. A standard Gaussian bridge whose source–endpoint correlation is ϱ has field slope $\{(1-t)(\varrho-1) + t(1-\varrho)\}/\{(1-t)^2 + t^2 + 2\varrho t(1-t)\}$; at $t = \frac{1}{4}$ this is $-\frac{4}{13}$ for $\varrho = \frac{1}{2}$ and $-\frac{4}{5}$ for $\varrho = 0$: the same endpoint law and a different field.

Same-time sufficiency is the argument of Proposition A.1(i), which uses neither the clock nor the randomness of the transition: if the chord is $\sigma(X'_t, H)$-measurable, the same-time oracle returns it exactly, the correction vanishes, the target is the sampled velocity, and Lemma B.1 with (A7) gives the Bellman endpoint; $\kappa = 0$ is ordinary flow matching, and at $\tilde\gamma = 1$ the two clocks agree. In the Gaussian model the residual Eq. (32) is a product of $t(1-t)$, $1-\rho_1^2$ and $1-\tilde\gamma$ over positive variances, so for $0 < t < 1$ these are its only zero cases (together with $\kappa = 0$); a nonzero field residual does not by itself preclude a correct endpoint: at full retimed gain on one deterministic branch the endpoint law is $\mathrm{Law}(r + \tilde\gamma\psi_v(X_0))$ for every admissible teacher field v, which is correct whenever v generates the correct successor law (Appendix B.5). For the scalar clause, a time-one map of a unique one-dimensional flow is increasing, so $x \mapsto (1-t)x + t\psi(x)$ is strictly increasing for $t < 1$, the query recovers x, and hence the chord $\psi(x) - x$. Determinism alone is not enough in $d \geqslant 2$: the rotation $X'_1 = -X_0$ is the time-one map of a smooth planar flow, yet at $t = \frac{1}{2}$ it gives $X'_t = 0$, so the same-time oracle returns $\mathbb{E}[X'_1 - X_0] = 0$ while $X_t = \frac{1-\tilde\gamma}{2}X_0 + \frac{1}{2}r$ reveals the chord $-2X_0$. That example fails only at one time, so it refutes a pointwise claim, not the almost-everywhere condition. □

Proof of Proposition 5. (i) Subtracting $w = A_t + gD$ from Eq. (10) gives Eq. (22). For $\mathcal{H}$-measurable κ, squaring and conditioning on $\mathcal{H}$ produces the cross term $2\kappa(1-\kappa)g^2\langle\mathbb{E}[C - D \mid \mathcal{H}], e\rangle = 0$, because $D = \mathbb{E}[C \mid \mathcal{H}]$ (proof of Theorem 5.3(i)), and the remaining terms give Eq. (23). The derivative of $(1-\kappa)^2 S + \kappa^2 \|e\|^2$ is $2\{-S + \kappa(S + \|e\|^2)\}$, so $\kappa^\star = S/(S + \|e\|^2) \in [0,1]$ is the unique minimiser when the denominator is positive; if both terms vanish, every fraction has zero conditional risk.

(ii) By Eq. (22) and the tower property, $\mathbb{E}[u_\kappa \mid X_t] - \mathbb{E}[Y \mid X_t] = \mathbb{E}[\kappa g e \mid X_t]$, and conditional Jensen, $\mathbb{E}\,\|\mathbb{E}[U \mid \mathcal{G}]\|^2 \leqslant \mathbb{E}\,\|U\|^2$ for $U \in L^2$ (Durrett, 2019), gives the L^2 bound. Eq. (25) adds and subtracts the oracle at $\widetilde\rho_1$. With a constant $\kappa \in [0,1]$ and $g \leqslant 1$, the triangle inequality gives

$$\|b_t\|_{L^2} \leqslant \kappa\left(\left\|\bar v - v^\star_{\widetilde\rho_1}\right\|_{L^2(\text{actual queries})} + \left\|v^\star_{\widetilde\rho_1} - v^\star_{\rho_1}\right\|_{L^2(\text{actual queries})}\right).$$

A field approximation error of a finite regression procedure must be added separately. If the endpoint solver changes the law of X'_1, the oracle here belongs to that approximate law, and Bellman error relative to the exact successor distribution is an additional term. □

B.7 Contraction moduli

The branchwise contraction depends on the metric. In the readout metric d_q of Theorem 5.2(iii) the modulus is exactly $\tilde\gamma$; in the velocity supremum norm it is exactly 1.

B.8 Correlated-Gaussian identities and the terminal scale error

Proposition B.5 (Correlated-Gaussian residuals and the terminal scale error of the constant-gain same-time target)**.** *Let (X_0, V) be centred standard Gaussians with correlation ρ_1, $X'_1 = \mu + \sigma V$ with $\sigma > 0$, and fix $R = r$ and $0 < \tilde\gamma \leqslant 1$. Put $P_u = (1-u)^2 + u^2\sigma^2 + 2u(1-u)\rho_1\sigma$ and $Q_t = \mathrm{Var}(X_t) = (1-t)^2 + \tilde\gamma^2 t^2\sigma^2 + 2(1-t)\tilde\gamma t\rho_1\sigma$.*

(i) Matched teacher and residuals. *The matched teacher queried at clock u, in particular at the retimed clock $u = \tau(t)$, is*

$$v^\star_{\rho_1}(u,b) = \mu + \beta_{\rho_1}(u)(b - u\mu), \qquad \beta_{\rho_1}(u) = \frac{(1-u)(\rho_1\sigma - 1) + u(\sigma^2 - \rho_1\sigma)}{(1-u)^2 + u^2\sigma^2 + 2u(1-u)\rho_1\sigma}, \tag{29}$$

and the remaining bridge uncertainty is

$$S_t = \mathrm{Var}(C \mid X_t, H) = \frac{\sigma^2\alpha(t)^2(1-\rho_1^2)}{Q_t}. \tag{30}$$

For a teacher calibrated at $\widetilde{\rho}_1$ and queried at clock $u = u(t)$,

$$\mathbb{E}[v^\star_{\widetilde{\rho}_1}(u, X'_u) - C \mid X_t = z] = \frac{\beta_{\widetilde{\rho}_1}(u)K_{BZ}(u) - K_{CZ}}{Q_t}\{z - t(r + \tilde{\gamma}\mu)\}, \tag{31}$$

where $K_{CZ} = (1-t)(\rho_1\sigma - 1) + \tilde{\gamma}t(\sigma^2 - \rho_1\sigma)$ and $K_{BZ}(u) = (1-u)(1-t) + u\tilde{\gamma}t\sigma^2 + \rho_1\sigma\{(1-u)\tilde{\gamma}t + u(1-t)\}$. The numerator vanishes at the matched retimed clock, and at the matched same-time clock it equals $t(1-t)\sigma^2(1-\rho_1^2)(1-\tilde{\gamma})/P_t$, so that

$$\mathbb{E}[v^\star_{\rho_1}(t, X'_t) - C \mid X_t = z] = \frac{t(1-t)\sigma^2(1-\rho_1^2)(1-\tilde{\gamma})}{P_tQ_t}\{z - t(r + \tilde{\gamma}\mu)\}. \tag{32}$$

(ii) Terminal scale error of the constant-gain same-time target. *In the correlated-Gaussian model above, with $0 < \tilde{\gamma} < 1$, $\rho_1 = 0$, an exact matched teacher and constant gain $\tilde{\gamma}$,*

$$\ln\frac{\sigma_{\mathrm{out}}}{\tilde{\gamma}\sigma} = \frac{-\tilde{\gamma}\ln\tilde{\gamma}}{1+\tilde{\gamma}} \qquad \textit{for every } \sigma > 0, \tag{33}$$

independently of the return scale. The error is positive: the learned law is over-dispersed. The retimed target, and the same-time target at $\rho_1 = 1$, both give exactly zero.

Proof. (i) Centre the chord, the successor query and the current point: $\bar{C} = \sigma V - X_0$, $\bar{B}_u = (1-u)X_0 + u\sigma V$, $\bar{X}_t = (1-t)X_0 + \tilde{\gamma}t\sigma V = z - t(r+\tilde{\gamma}\mu)$. Direct expansion gives $\mathrm{Cov}(\bar{C}, \bar{B}_u) = (1-u)(\rho_1\sigma - 1) + u(\sigma^2 - \rho_1\sigma)$, $\mathrm{Var}(\bar{B}_u) = P_u$, $\mathrm{Cov}(\bar{C}, \bar{X}_t) = K_{CZ}$, $\mathrm{Cov}(\bar{B}_u, \bar{X}_t) = K_{BZ}(u)$ and $\mathrm{Var}(\bar{X}_t) = Q_t$. The Gaussian regression formula $\mathbb{E}[U \mid B] = \mathbb{E}U + \mathrm{Cov}(U,B)\,\mathrm{Var}(B)^{-1}(B - \mathbb{E}B)$ gives Eq. (29). The map $(X_0, V) \mapsto (\bar{C}, \bar{X}_t)$ has determinant $-\sigma\{1-(1-\tilde{\gamma})t\} = -\sigma\alpha$ and $\det\mathrm{Cov}(X_0, V) = 1 - \rho_1^2$, so $\mathrm{Var}(\bar{C} \mid \bar{X}_t) = \det\mathrm{Cov}(\bar{C}, \bar{X}_t)/\mathrm{Var}(\bar{X}_t) = \sigma^2\alpha^2(1-\rho_1^2)/Q_t$, which is Eq. (30); at $\rho_1 = 1$ it vanishes. The teacher residual is $v^\star_{\widetilde{\rho}_1}(u, X'_u) - C = \beta_{\widetilde{\rho}_1}(u)\bar{B}_u - \bar{C}$; regressing it on $\bar{X}_t$ gives Eq. (31).

At the retimed clock $u = \tau$, $1 - \tau = (1-t)/\alpha$ and $\tau = \tilde{\gamma}t/\alpha$ (Lemma B.2; at $\tilde{\gamma} = 1$ they hold directly, with $\alpha \equiv 1$ and $\tau(t) = t$) give $\bar{B}_\tau = \bar{X}_t/\alpha$, so conditioning on X_t is conditioning on the query, and with $\widetilde{\rho}_1 = \rho_1$ the residual has zero conditional mean by the tower property; in coefficients, $\mathrm{Cov}(\bar{C}, \bar{B}_\tau) = K_{CZ}/\alpha$, $P_\tau = Q_t/\alpha^2$ and $K_{BZ}(\tau) = Q_t/\alpha$, so $\beta_{\rho_1}(\tau)K_{BZ}(\tau) - K_{CZ} = (\alpha K_{CZ}/Q_t)(Q_t/\alpha) - K_{CZ} = 0$. At the same-time clock $u = t$ with $\widetilde{\rho}_1 = \rho_1$, write $k = \mathrm{Cov}(\bar{C}, \bar{B}_t)$; then

$$\begin{aligned} k &= K_{CZ} + \Delta_1, & \Delta_1 &= (1-\tilde{\gamma})t(\sigma^2 - \rho_1\sigma), \\ K_{BZ}(t) &= P_t - \Delta_2, & \Delta_2 &= (1-\tilde{\gamma})t\{t\sigma^2 + (1-t)\rho_1\sigma\}, \end{aligned}$$

so $P_t\{\beta_{\rho_1}(t)K_{BZ}(t) - K_{CZ}\} = kK_{BZ}(t) - K_{CZ}P_t = \Delta_1P_t - k\Delta_2 = (1-\tilde{\gamma})t\,\Pi$ with

$$\begin{aligned} \Pi &= (\sigma^2 - \rho_1\sigma)P_t - k\{t\sigma^2 + (1-t)\rho_1\sigma\} \\ &= (1-t)^2\sigma^2(1-\rho_1^2) + t(1-t)\sigma^2(1-\rho_1^2) = (1-t)\sigma^2(1-\rho_1^2), \end{aligned}$$

where the t^2 terms cancel and the $(1-t)^2$ and $t(1-t)$ coefficients are collected in powers of $\rho_1\sigma$. Dividing by P_tQ_t gives Eq. (32).

(ii) Write $\tilde{\gamma}$ for the multiplier, $P_t = (1-t)^2 + t^2\sigma^2$ and $Q_t = (1-t)^2 + \tilde{\gamma}^2t^2\sigma^2$ (the $\rho_1 = 0$ values). The exact field of the current bridge is linear in the centred coordinate with slope $Q'_t/(2Q_t)$, because $\mathrm{Cov}(\dot{X}_t, X_t) = \frac{1}{2}\frac{d}{dt}\mathrm{Var}(X_t)$, and its flow has variance Q_t, ending at $\tilde{\gamma}^2\sigma^2$. Under constant same-time gain $\tilde{\gamma}$, the population field adds $\tilde{\gamma}$ times the residual of Eq. (32) at $\rho_1 = 0$, an excess slope

$$\Delta a(t) = \frac{\tilde{\gamma}(1-\tilde{\gamma})\,t(1-t)\sigma^2}{P_tQ_t},$$

and leaves the mean path unchanged. A centred linear flow multiplies its standard deviation by $\exp(\int_0^1 a)$, so $\ln(\sigma_{\text{out}}/\tilde{\gamma}\sigma) = \int_0^1 \Delta a\, dt$. With $w = t/(1-t)$, $dt = dw/(1+w)^2$, $t(1-t) = w/(1+w)^2$, $P_t = (1+\sigma^2 w^2)/(1+w)^2$ and $Q_t = (1+\tilde{\gamma}^2\sigma^2 w^2)/(1+w)^2$, and then $s = \sigma w$,

$$\int_0^1 \Delta a\, dt = \tilde{\gamma}(1-\tilde{\gamma}) \int_0^\infty \frac{s\, ds}{(1+s^2)(1+\tilde{\gamma}^2 s^2)}$$
$$= \tilde{\gamma}(1-\tilde{\gamma}) \lim_{M\to\infty} \frac{1}{2(1-\tilde{\gamma}^2)} \Big[\ln \frac{1+s^2}{1+\tilde{\gamma}^2 s^2} \Big]_0^M = \frac{\tilde{\gamma}(1-\tilde{\gamma})(-\ln\tilde{\gamma})}{1-\tilde{\gamma}^2},$$

which is $-\tilde{\gamma}\ln\tilde{\gamma}/(1+\tilde{\gamma})$, independent of σ and positive for $0 < \tilde{\gamma} < 1$. The retimed target has zero residual (part (i)), and the same-time residual carries the factor $1-\rho_1^2$, which vanishes at $\rho_1 = 1$.

□

B.9 Shared noise

Proof of Proposition A.1. (i) If the chord is measurable with respect to the query and H, the matched oracle is $D = \mathbb{E}[C \mid \text{query}, H] = C$, so the control variate vanishes and $u_a = Y = R + \tilde{\gamma}X_1' - X_0$ at every gain; the regression is ordinary flow matching onto the Bellman endpoint, and Lemma B.1 with (A7) identifies the generated law. The residual is pointwise zero, so this holds at either clock. For the early-time remark take $d = 1$, reward zero, a branch sign $S \in \{-1, 1\}$ with equal probability independent of $X_0 \sim \mathcal{N}(0,1)$, and shared-noise generators $X_1' = Sa + \sigma X_0$ with $a, \sigma > 0$. Since $X_u' = uSa + (1-u+u\sigma)X_0$ with $1-u+u\sigma > 0$, the chord is measurable with respect to (X_u', S), so the flow reaches $\frac{1}{2}\mathcal{N}(-\tilde{\gamma}a, \tilde{\gamma}^2\sigma^2) + \frac{1}{2}\mathcal{N}(\tilde{\gamma}a, \tilde{\gamma}^2\sigma^2)$, of variance $\tilde{\gamma}^2(\sigma^2 + a^2)$. At $t = 0$, however, $X_t = X_0$ carries no branch information and the fitted velocity is $\mathbb{E}[\tilde{\gamma}Sa + \tilde{\gamma}\sigma z - z] = (\tilde{\gamma}\sigma - 1)z$: the velocity of the increasing map toward $\mathcal{N}(0, \tilde{\gamma}^2\sigma^2)$, the one-dimensional W_2 barycenter of the two Bellman branch endpoint laws $\mathcal{N}(\pm\tilde{\gamma}a, \tilde{\gamma}^2\sigma^2)$, whose variance is only $\tilde{\gamma}^2\sigma^2$. Branch information re-enters through the posterior at later times.

(ii) Write $b_u(x_0) = (1-u)x_0 + u\,\psi(x_0)$ for the query built from a source x_0, where ψ is the time-one map of the reused field v. Recoverability makes the matched teacher at $(u, b_u(x_0))$ equal to $\psi(x_0) - x_0$ for almost every query. The query law is the image of (clock law) $\otimes \mathcal{N}(0, I_d)$ under $(u, x_0) \mapsto (u, b_u(x_0))$, and the clock law has a positive density on $(0,1)$. If v equals the matched teacher almost everywhere on the query law, Fubini's theorem gives, for almost every x_0, $v(b_u(x_0), u) = \psi(x_0) - x_0 = \frac{d}{du} b_u(x_0)$ for almost every u; so $u \mapsto b_u(x_0)$ is an absolutely continuous solution of $x_u = x_0 + \int_0^u v(x_s, s)\, ds$, and uniqueness of the flow (for absolutely continuous solutions; for continuous fields this is ordinary ODE uniqueness) makes it the trajectory of v, which is therefore straight at constant speed. Conversely, if the trajectory from almost every x_0 is $b_u(x_0)$, its velocity $\psi(x_0) - x_0$ equals $v(b_u(x_0), u)$ for almost every u, which is the matched teacher.

□

C Toy Results

Against a tuned comparator, on held-out seeds. Table 5 tunes κ separately for each family — the same-time shared-noise family and the retimed independent-noise family — selecting on one seed split and scoring on the other.

Against a tuned same-time comparator the retimed family wins on SOLITAIRE and DISCRETE-MC. On BERNOULLI the fifteen-seed held-out gaps do not separate ($+0.00086 \pm 0.00112$, $p = 0.455$; $+0.00141 \pm 0.00129$, $p = 0.292$), at 21–25× the evaluation floor, so the advantage is environment-dependent. BERNOULLI is also the only environment with $\gamma/(1-\gamma) = 1$, and this design does not separate smoothness from the absence of amplification. Against pinned PCBF ($\kappa = 1$) the gaps are larger (0.053 vs 0.017, 0.680 vs 0.091, 0.594 vs 0.302). In floor units the tuned comparison is 19× against 24×, 28× against 49×, and 40× against 66×.

The result the theory predicts is where each optimum sits. The retimed $\kappa^\star$ is 1 in every environment and split, so tuning gains nothing. The same-time optimum is never at $\kappa = 1$; that rung is its worst, and tuning moves it toward $\kappa = 0$. Theorem 5.3 makes the retimed correction unbiased at every

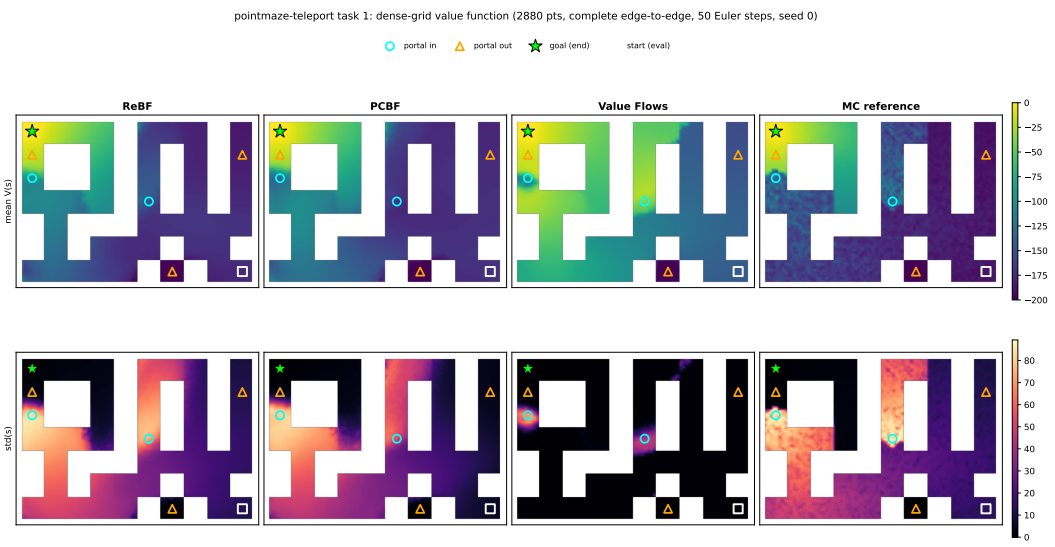


Figure 4: **Learned value and uncertainty over the whole maze**, OGBench `pointmaze-teleport` (task 1, $\gamma = 0.995$), seed 0, 50 Euler steps. Columns: ReBF, PCBF, Value Flows (the published critic) and a Monte Carlo reference; top row, the mean return $V(s)$ of the law each critic generates; bottom row, its standard deviation. Each map is evaluated at 2,880 points (8×8 per free cell, spacing 0.5) and linearly interpolated within free cells; walls are white. Markers are fixed maze geometry: portal entrances (circles), portal exits (triangles, one of them the sealed trap), the goal (star) and the "start" evaluation point (square). Against the reference, ReBF's mean map has mean absolute error 13.0 (bias +9.3) and its standard-deviation map 3.9, with spatial correlation 0.96 to the reference's and average ratio 1.05; PCBF's are 17.4 (+11.9), 6.5, 0.91 and 0.92; Value Flows' are 60.6 (+60.6), 27.8, 0.31 and 0.09, its uncertainty collapsing almost everywhere. The reference has only 30 rollouts per point, which makes its panels grainy and puts a floor of about 4.4 on the mean-map error and 1.7 on the standard-deviation error; ReBF's and PCBF's errors are two to four times these floors. One seed: a picture, not an estimate.

Table 5: Tuned same-time ($\rho_1 = 1$) versus tuned retimed ($\rho_1 = 0$), κ selected on five seeds and scored on the five held-out seeds; both split directions shown. W_1 against the exact law at 50 ODE steps, 4×10^5 draws. Every level in this table is at least $19\times$ the measured evaluation floor for its environment; the gaps on SOLITAIRE and DISCRETE-MC are significant in both split directions ($p \leqslant 0.04$ at $n = 5$), while on BERNOULLI they are 1.3–2.2$\times$ the floor at $n = 15$, which is why we call it a tie (Appendix F). Scored n is 15 on BERNOULLI and 5 elsewhere; ties are declared using the t critical value at the relevant n.

Env	Split	$\kappa^\star$ (same-time)	$\kappa^\star$ (retimed)	Ratio	p	Verdict
BERNOULLI	A|B	0.1	1	1.06×	0.455	powered null
BERNOULLI	B|A	0.02	1	1.10×	0.292	powered null
SOLITAIRE	A|B	0	1	1.62×	0.040	retimed
SOLITAIRE	B|A	0.05	1	1.89×	0.003	retimed
DISCRETE-MC	A|B	0.3	1	1.64×	0.004	retimed
DISCRETE-MC	B|A	0.3	1	1.66×	0.0003	retimed

gain. At $\rho_1 = 1$ the same-time residual of Eq. (32) vanishes, but the reused critic is generally not the teacher its coupling requires (Proposition A.1(ii)), and that error enters scaled by κ (Eq. (24)). On the full (ρ_1, κ) surface the best retimed cell is $(\rho_1 = 0, \kappa = 1)$, ReBF itself, in all three environments.

Clock and coupling interact. Holding $\kappa = 1$ and measuring the clock gap (same-time minus retimed W_1) at the two coupling extremes:

Env	gap at $\rho_1 = 0$	gap at $\rho_1 = 1$	difference	p
BERNOULLI	$+0.0408$	-0.0562	-0.0970	$< 10^{-4}$
SOLITAIRE	$+0.1623$	-0.0460	-0.2083	0.0004
DISCRETE-MC	$+0.2950$	-0.1051	-0.4001	$< 10^{-4}$

With a reused critic, shared noise settles at a biased level. On SOLITAIRE the shared-noise W_1 rises with training and PCBF then flattens (Figure 5c). From 2 000 to 50 000 steps, five seeds, PCBF moves $0.544 \to 0.751$ while ReBF falls $0.381 \to 0.126$; a rerun of the same seeds ends at 0.698 and 0.123. PCBF is about six times higher at the last checkpoint, with mean bias $-14.0\% \pm 1.2$ ($n = 10$). The arms stay together through step 1 000 and separate between 1 000 and 2 000 (-0.166 ± 0.040), so the bias appears once the critic is accurate enough for the coupling to matter. The retimed shared-noise arm rises the same way. Table 6 shows the penalty vanishes as the backup deepens. It does not scale with the horizon: at one step the bias is -14.0%, -28.5% and $-29.7 \pm 2.5\%$ at $\gamma = 0.9$, 0.99 and 0.995, while $\gamma/(1-\gamma)$ runs $9 \to 99 \to 199$. A matched teacher makes shared noise exact (Proposition A.1(i)); the reused critic is that teacher only if its flow is straight (Proposition A.1(ii)). A single deterministic branch leaves the mean exact at full retimed gain (Appendix B.5). We have no population model that also matches the measured size.

Table 6: Mean bias (%) against the exact mean, with W_1 beneath, as a function of backup depth n; SOLITAIRE, $\gamma = 0.9$, five seeds, 50 ODE steps. Standard errors are 0.6–1.4 percentage points throughout. An n-step target uses $\tilde{\gamma} = \gamma^n$ and the discounted n-step reward. W_1 here compares 4,000 learned draws against a 20,000-draw sampled reference, whose floor is 0.043 ± 0.017 (Appendix F); at $n = 20$ ReBF and PCBF differ by less than that floor's standard deviation.

Arm	$n = 1$	$n = 3$	$n = 5$	$n = 10$	$n = 20$
ReBF (retimed, ρ_1=0, κ=1)	-1.64	-3.05	-1.68	-0.25	-0.42
	0.097	0.155	0.123	0.119	0.134
PCBF (same-time, ρ_1=1, κ=1)	-14.17	-3.94	-0.45	-0.31	$+0.08$
	0.709	0.427	0.250	0.140	0.126
uncorrected (κ=0)	-1.36	-2.48	-0.97	-0.35	$+0.55$
	0.162	0.156	0.155	0.112	0.112

D ALL ENVIRONMENTS AND ALL ANALYSES

Figure 5 collects all analysis of Section 6 in every environment, with four-rooms successor features (Appendix E.4) as a fourth column. The main-text Figure 2 shows one environment for each row, and Figure 7 shows the full (ρ_1, κ) surface of the toys that rows (a) and (b) cut through (on seeds 0–4 only, so its corner cells differ from the ten-seed values in those rows, by up to 16% on BERNOULLI). The toy columns are read out at 50 ODE steps; the four-rooms column at 128, from 2×10^4 critic draws against a 4×10^5-rollout reference, and its training curves use that same instrument. The test-time budget is in Appendix I.1.

(a) Correction weight. On the toys the retimed $\rho_1 = 0$ slice falls to its minimum at $\kappa = 1$, and the same-time shared-noise family is worst there (sweep minima $\kappa = 0.02$, 0, 0.3). Green is Value Flows on its own dial $\kappa_{\mathrm{VF}} = \lambda_{\mathrm{DCFM}}/(\lambda_{\mathrm{BCFM}} + \lambda_{\mathrm{DCFM}})$, not our κ; its $\kappa_{\mathrm{VF}} = 1$ cell diverged and is omitted. On four-rooms the retimed slice falls from 0.106 at $\kappa = 0$ to 0.028 at $\kappa = 1$; the same-time family is lowest at $\kappa = 0.75$ (0.038). *(b) Noise coupling.* At $\kappa = 1$ the retimed W_1 rises with ρ_1 from its minimum at $\rho_1 = 0$, and the same-time curve is U-shaped; at $\kappa = 0.1$ the two clocks lie within 16%. On four-rooms a partially coupled same-time critic ($\rho_1 = 0.75$) ties ReBF (0.0289 against 0.0300, $p = 0.24$; held-out seeds 5–14: 0.0277 against 0.0273, $p = 0.75$), while PCBF at $\rho_1 = 1$ is higher on all 15 seeds. *(c) Training.* On SOLITAIRE the arms coincide through step 1 000; the shared-noise arms at $\rho_1 = 1$, $\kappa = 1$ then rise, so the effect tracks the coupling. On four-rooms the three arms coincide through 5 000 steps and at 20 000 read 0.028, 0.041 and 0.052 for ReBF, PCBF and the uncorrected critic.

Four-rooms maps. Figure 6 maps one seed's mean and dispersion errors over all 104 cells of four-rooms and draws its joint law at one evaluation state. Away from $w = M^\top e_{15}$, the same-time critics'

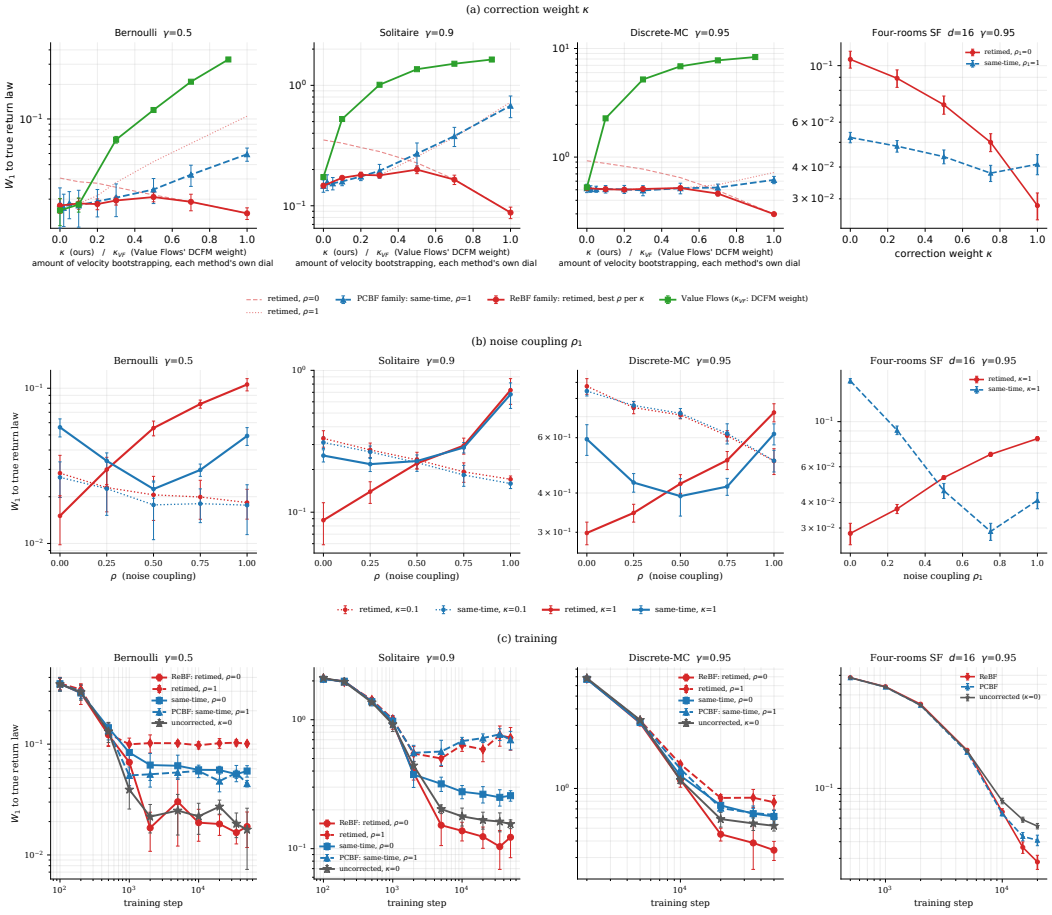


Figure 5: **Every analysis in every environment** (columns BERNOULLI, SOLITAIRE, DISCRETE-MC and four-rooms successor features, $d = 16$): (a) correction weight, (b) noise coupling, (c) training, described in Appendix D. W_1 to the exact law (4×10^5 draws) in the toy panels of rows (a) and (b), and to a 4×10^5-rollout reference in every four-rooms panel; scoring and floors in Appendix F. Error bars are ± 1 s.d., except ± 1 s.e.m. on the solid red and green curves of the toy panels of row (a). The $\kappa = 0$ arm is the uncorrected critic, shown for context. The toy training curves and Value Flows (green, row a) are scored on a coarser instrument: 4×10^3 draws against a sampled 2×10^4-draw reference, with floors of about 0.011, 0.050 and 0.075.

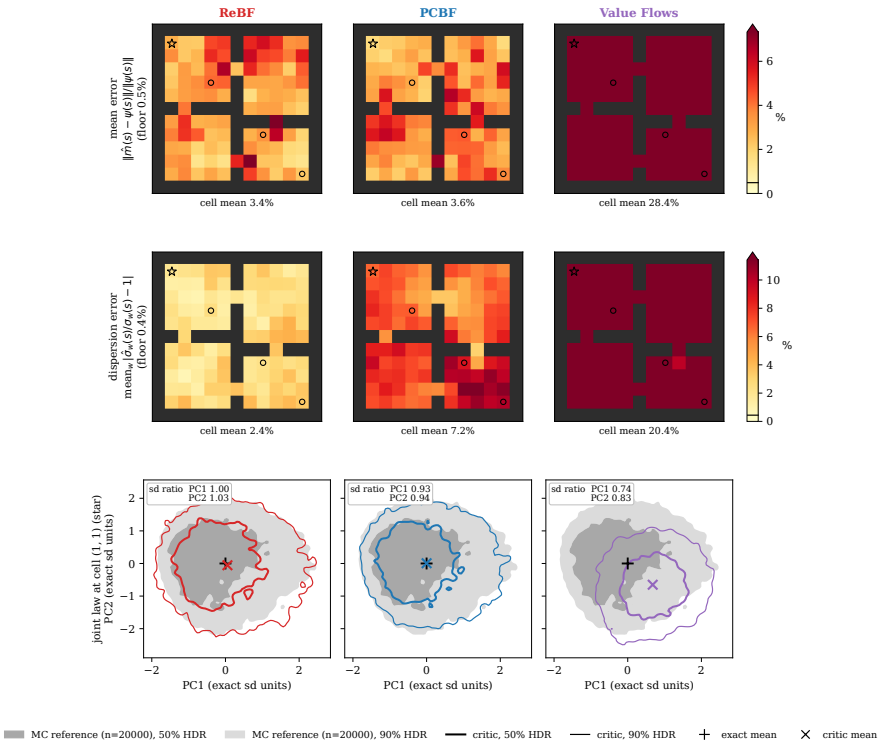


Figure 6: **Four-rooms error maps**, seed 0 ($d = 16$, $\gamma = 0.95$; columns ReBF, PCBF, Value Flows). Row 1: mean error $\|b\|/\|\psi\|$ over all 104 cells. Row 2: dispersion error, dropping the ill-conditioned direction $w = M^\top e_{15}$, so the printed cell means are not the Table 2 entries. A star marks $(1, 1)$ and circles the other evaluation states; the colour scale is set by ReBF and PCBF. Row 3: the joint law at $(1, 1)$ on its two leading principal axes.

dispersion errors have a sign. Over the other 10 directions at the four evaluation states, PCBF and the uncorrected critic BCFM are under-dispersed on 15 of 15 seeds (mean sd ratio 0.922 ± 0.003 and 0.908 ± 0.002, mean $\pm$ s.e.), and more so along $w = M^\top e_0$, the slowest room mode (horizon 18.4; 0.839 and 0.748). ReBF is at 1.000 ± 0.002, below 1 on 7 of 15 seeds, and Value Flows is over-dispersed (1.067 ± 0.003, above 1 on 15 of 15). With $w = M^\top e_{15}$ included, as in Table 2, PCBF stays below 1 (0.976) but BCFM and ReBF move above it (1.077 and 1.051).

E Environments and Their Exact Return Laws

To evaluate whether a critic learns the correct return distribution, rather than inferring it from control performance, we use Markov reward processes whose return laws can be derived exactly or computed to machine precision. They expose different kinds of stochasticity in the Bellman recursion: BERNOULLI is a single-state process whose discounted return is exactly uniform; SOLITAIRE produces a purely atomic, long-tailed return through stochastic termination; DISCRETE-MC accumulates stochasticity through repeated nearest-neighbour transitions; and a successor-feature environment, four-rooms (used in Table 2), has vector-valued returns whose mean and directional spread are exact. For each we give the process as it is implemented, derive its return law, and state the reference used for scoring (Appendix F).

E.1 BERNOULLI: ONE STATE, FAIR-COIN REWARDS

Definition. BERNOULLI (Bellemare et al., 2023, Example 2.10) has one state and one action. Every step returns to that state and pays an independent fair coin, with $\gamma = \frac{1}{2}$:

$$R_k \overset{\text{i.i.d.}}{\sim} \text{Bern}(1/2), \qquad \gamma = \tfrac{1}{2}. \tag{34}$$

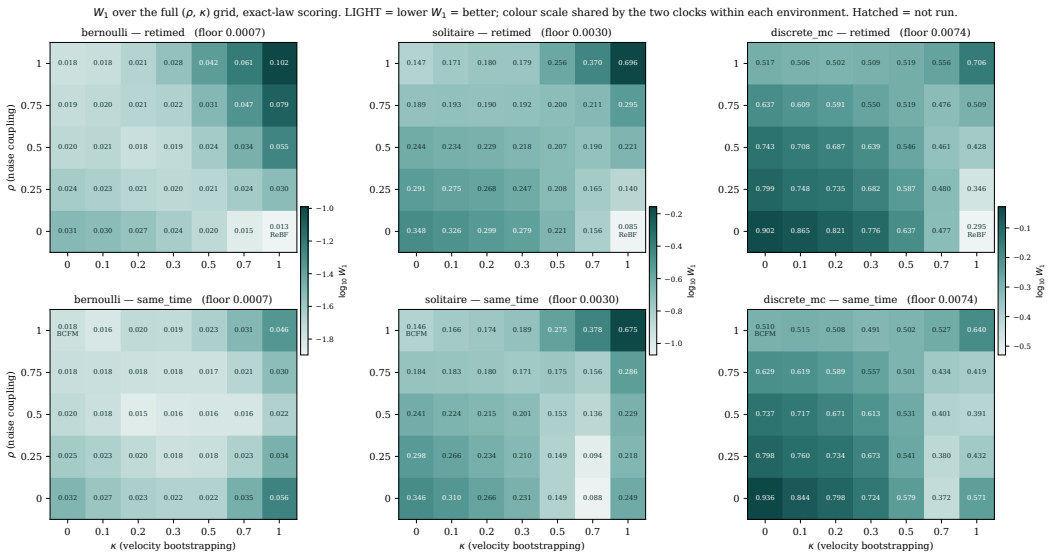


Figure 7: The (ρ_1, κ) surface: all 210 cells, seeds 0–4, W_1 to the exact law at 50 ODE steps, one colour scale per environment. On these seeds the ReBF corner $(\rho_1 = 0, \kappa = 1)$ is the lowest cell in every environment; the margin is significant only on DISCRETE-MC ($p = 0.025$). At $\kappa = 0$ the clock is inert, and the two panels agree to within 4.2% of the cell.

Training uses one-step backups, so the terminal multiplier $\tilde{\gamma} = 0$ never occurs. The critic is evaluated at the only state.

Return law.

$$G = \sum_{k \geqslant 0} \gamma^k R_k, \qquad \eta = \mathrm{Law}(G) = \mathcal{U}(0, 2). \tag{35}$$

If $G' \sim \mathcal{U}(0, 2)$ then $R + G'/2$ is the equal mixture of $\mathcal{U}(0, 1)$ and $\mathcal{U}(1, 2)$, so the uniform law is the Bellman fixed point. It is the unique fixed point because $\mathcal{T}^\pi$ contracts in W_1. The law is uniform only at $\gamma = \frac{1}{2}$; the scoring code refuses any other discount. $\mathbb{E}G = 1$ and $\mathrm{sd}(G) = 1/\sqrt{3} \approx 0.577$.

Exact reference used for scoring. The reference is the grid of $N + 1 = 200{,}001$ atoms at spacing $h = 10^{-5}$ on $[0, 2]$, whose W_1 to $\mathcal{U}(0, 2)$ is about $h/3 = 3.3 \times 10^{-6}$. The critic is read out with 4×10^5 draws and 50 Euler steps. The evaluation floor, the W_1 of that many exact draws, is 6.5×10^{-4} in Appendix F (about 10^{-3} in expectation). The coarser training-curve instrument, 4×10^3 draws against a 2×10^4-draw reference, has floor about 0.011.

E.2 SOLITAIRE

Definition. SOLITAIRE (Bellemare et al., 2023, Example 2.8) has one state. Each step rolls a fair die: a 1 pays 0 and terminates, and $2, \ldots, 6$ pay 1 and continue. With $p = 1/6$ and $q = 5/6$,

$$(R, \tilde{\gamma}) = (0, 0) \text{ with probability } p, \qquad (1, \gamma) \text{ with probability } q, \tag{36}$$

and $\gamma = 0.9$. The critic is evaluated at the only state.

Return law. Let K be the number of non-terminating rolls, $\Pr(K = k) = pq^k$. The terminating roll pays nothing, so

$$G = a_K, \qquad a_k = \frac{1 - \gamma^k}{1 - \gamma}, \qquad \eta^\pi = \sum_{k=0}^{\infty} pq^k \, \delta_{a_k}. \tag{37}$$

The atoms accumulate at $1/(1 - \gamma) = 10$, which carries no mass. The atom at 0 has mass $1/6$. The mean and standard deviation are $\mu = q/(1 - q\gamma) = 10/3$ and $\mathrm{sd}(G) = \sqrt{800/117} \approx 2.615$ (CV ≈ 0.784).

Exact reference used for scoring. The reference keeps atoms $k = 0, \dots, 1999$, renormalised; the dropped mass is negligible. W_1 is the integral of the absolute difference of distribution functions against 4×10^5 critic draws (50 Euler steps). The mean and standard deviation are the exact values above. The expected evaluation floor is 4.5×10^{-3}; the realisation quoted in Appendix F is 3.0×10^{-3}.

E.3 DISCRETE-MC: AN ABSORBING NEAREST-NEIGHBOUR CHAIN

Definition. DISCRETE-MC is a birth–death chain on $\{0, 1, \dots, 19\}$ with absorbing ends and interior states $1, \dots, 18$ (Cheng and Weare, 2024). From an interior state the walk prefers the heavier neighbour under weights $w_i = \exp(\beta \cos(4\pi(i-1)/19))$, $\beta = (n-1)/(4\pi)$. A transition into the interior pays 1; absorption pays 0 and sets $\tilde{\gamma} = 0$. Otherwise $\tilde{\gamma} = \gamma = 0.95$.

Return law. Let K be the number of rewarded steps before absorption and Q the substochastic interior block. Then

$$G(s) = x_K, \qquad x_k = \frac{1-\gamma^k}{1-\gamma}, \qquad \Pr_s(K = k) = (Q^k u)_s, \tag{38}$$

with u the one-step absorption probabilities. Absorption is certain. Means of $G(s)$ run from 1.55 at state 18 to 19.23 at state 10; per-state coefficients of variation run from 0.101 to 2.47.

Exact reference used for scoring. The reference is the atomic law from the recursion above, truncated at $k = 4000$, where the remaining mass is at most 4.7×10^{-8}. Reported W_1 averages the 18 interior states, each scored with 4×10^5 critic draws. The mean bias uses the average of $\mathbb{E}[G(s)]$, which is 12.788. The evaluation floor in Appendix F is 0.0074.

E.4 FOUR-ROOMS SUCCESSOR FEATURES ($d = 16$)

Definition. FOUR-ROOMS, the vector-valued environment of Table 2 and of the fourth column of Figure 5, is a lazy random walk on the four-rooms layout of Sutton et al. (1999) (Figure 8). The layout is the 11×11 interior of a 13×13 wall map whose rows and columns are numbered $0, \dots, 12$ from the top left. An interior wall runs down column 6 and is broken by doorways at rows 3 and 10. The left half is split by a wall along row 6 with a doorway at column 2, and the right half by a wall along row 7 with a doorway at column 9. These 17 interior wall cells leave 104 open cells: a 5×5 north-west room (rows 1–5, columns 1–5), a 6×5 north-east room (rows 1–6, columns 7–11), a 5×5 south-west room (rows 7–11, columns 1–5), a 4×5 south-east room (rows 8–11, columns 7–11), and the four single-cell doorways $(3, 6)$, $(10, 6)$, $(6, 2)$ and $(7, 9)$, written (row, column). Each doorway is the only connection between its two rooms, and the rooms form a cycle, so the north-west and south-east rooms are two doorways apart either way round. The state space $\mathcal{S}$ is the set of open cells, indexed in row-major order. From s the walk stays put with probability $p_{\text{stay}} = 0.2$ and otherwise moves to each of its four compass neighbours with probability $(1 - p_{\text{stay}})/4 = 0.2$, so it picks one of the five moves $\{\text{stay}, \text{N}, \text{S}, \text{E}, \text{W}\}$ uniformly. A move into a wall leaves it in place:

$$P(s, s') = 0.2 \text{ for each open compass neighbour } s' \text{ of } s, \qquad P(s, s) = 0.2 + 0.2\, n_{\text{wall}}(s), \tag{39}$$

where $n_{\text{wall}}(s) \in \{0, 1, 2\}$ counts the walls among the four neighbours of s, and every other entry of row s is 0. The holding probability is therefore 0.2 at 44 cells, 0.4 at 40, and 0.6 at 20, which are the 16 room corners and the 4 doorways. A move between open cells $s \to s'$ is possible exactly when $s' \to s$ is, with the same probability, so P is symmetric and doubly stochastic. Its stationary law is uniform, and it has a real orthonormal eigenbasis.

Features. P is symmetric, so it has a real orthonormal eigenbasis, computed once and stored. We keep $d = 16$ modes, log-spaced from the slowest non-constant eigenvalue $\lambda = 0.9954$ to the bottom of the spectrum, excluding the constant mode. Let V hold those eigenvectors and let M be one fixed random orthogonal matrix, shared by every arm and seed. Then

$$B = VM, \qquad \phi(s) = B^{\top}_{s,:} \in \mathbb{R}^{16}, \qquad R = \phi(S), \qquad \tilde{\gamma} = \gamma = 0.95. \tag{40}$$

It follows that $B^{\top}B = I_{16}$, that every feature averages to zero over the 104 cells, and that $0.20 \leqslant \|\phi(s)\| \leqslant 0.66$. The reward is paid in the current state. There is no termination, so the backup multiplier of Eq. (1) equals γ on every transition. Training draws the current state uniformly from all 104 cells and takes one step of P.

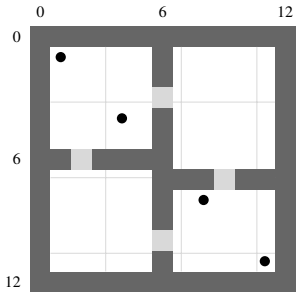


Figure 8: **The FOUR-ROOMS layout** (Appendix E.4). Walls are dark, the four doorways light grey and the other open cells white, 104 open cells in all. Dots mark the four evaluation states $(1, 1)$, $(4, 4)$, $(8, 8)$ and $(11, 11)$ (row, column; rows numbered downward).

Return law. The return is the discounted successor-feature vector

$$\Psi(s) = \sum_{k \geqslant 0} \gamma^k \phi(S_k), \quad S_0 = s,\ S_{k+1} \sim P(S_k, \cdot); \qquad \Psi(s) \overset{d}{=} \phi(s) + \gamma\, \Psi(S_1). \tag{41}$$

By the Markov property, $\Psi(S_1)$ given S_1 is an independent draw from η_{S_1}. So $\eta_s = \mathrm{Law}(\Psi(s))$ is the unique fixed point of $\eta_s = \sum_{s'} P(s, s')\,(T_s)_{\#}\eta_{s'}$ with $T_s(z) = \phi(s) + \gamma z$. It is a mixture of at most five affine images of the successors' laws (only three at the room corners and doorways), supported in the ball of radius $\max_s \|\phi(s)\| / (1 - \gamma) = 13.1$. On the three scalar environments the return is a function of a single stopping time, or a binary expansion of independent coins. Here it depends on the whole path through ϕ, and unrolling the recursion sums over infinitely many paths. We have no closed form for η_s, nor for the law of any of its one-dimensional slices, but the first two moments of every slice are exact. For $w \in \mathbb{R}^{16}$, $G_w(s) = w^\top \Psi(s)$ is the return of the scalar process on the same chain with reward $r_w = Bw$. The critic is fitted to the full vector law and never sees w, so every slice is a zero-shot readout.

Moments. The mean $\psi(s) = \mathbb{E}\Psi(s)$ and the variance of each slice $G_w = w^\top \Psi$ are exact:

$$\psi = (I - \gamma P)^{-1} B, \tag{42}$$

and, with $r_w = Bw$ and $m_w = \psi w$,

$$\sigma_w^2 = \gamma^2 (I - \gamma^2 P)^{-1} \big[P(m_w^2) - (P m_w)^2 \big]. \tag{43}$$

Exact reference used for scoring. The mean and the standard deviations are exact. The mean error is $\|b(s)\| \,/\, \|\psi(s)\|$ averaged over the four evaluation states below, where $b(s)$ is the critic's sample mean minus $\psi(s)$ from Eq. (42). The dispersion error is $\sum |\widehat{\mathrm{sd}} - \sigma_w(s)| / \sum \sigma_w(s)$ over the 44 slices below, with σ_w from Eq. (43). W_1 has no exact reference here, so unlike on the scalar environments it is scored against a *sampled* one. From each evaluation state we draw $N = 4 \times 10^5$ independent rollouts of P and truncate each at $T = 400$ steps, which moves a draw by at most $\gamma^T \max_s \|\phi(s)\| / (1 - \gamma) < 1.7 \times 10^{-8}$. As a check on the sampled references, across the fifteen per-seed references the largest of the 960 per-coordinate deviations of the sample mean from $\psi(s)$ is 3.1 standard errors, and every one of the 660 per-slice sample standard deviations lies within 1.2% of its exact σ_w. The evaluation states are the open cells at positions 0, 34, 69 and 103 of the row-major order. They are the corner $(1, 1)$ and the interior cell $(4, 4)$ of the north-west room, and the cell $(8, 8)$ beside the north wall and the corner $(11, 11)$ of the south-east room. No cell of the north-east or south-west room, and no doorway, is scored. At each evaluation state, 2×10^4 critic draws (128 Heun steps) and the reference are projected onto 11 fixed unit directions:

- four eigen-directions $w = M^\top e_k$, $k \in \{0, 5, 10, 15\}$. For these r_w is a single eigenvector with $\lambda = 0.9954$, 0.9263, 0.6531 and -0.4774 (horizons 18.4, 8.3, 2.6 and 0.69), so $m_w = r_w / (1 - \gamma\lambda)$. The first is the slowest room mode;
- three feature directions $w = e_j$, $j \in \{0, 8, 15\}$, each a fixed mixture of the 16 kept modes;
- four random directions $g / \|g\|$ with $g \sim \mathcal{N}(0, I_{16})$, drawn once.

With $n = 2 \times 10^4$ midpoint quantile levels $u_i = (i - \frac{1}{2})/n$, the score is

$$W_1^{\text{sliced}} = \frac{1}{44} \sum_s \sum_w \frac{1}{n} \sum_{i=1}^{n} \left| \hat{F}^{-1}_{w^\top \hat{X}_s}(u_i) - \hat{F}^{-1}_{w^\top \hat{\Psi}_s}(u_i) \right|, \tag{44}$$

where $\hat{X}_s$ are the critic's draws at s and $\hat{\Psi}_s$ the reference draws. $\hat{F}^{-1}$ is the empirical quantile function, with linear interpolation. The map $z \mapsto w^\top z$ is 1-Lipschitz for a unit w, so each slice lower-bounds the 16-dimensional W_1. An error orthogonal to all 11 directions is invisible to the score, which is why the exact mean and per-slice spread are reported beside it. Each training seed has its own reference draw, shared by every arm at that seed. The evaluation floor is the same score between a fresh 2×10^4-draw sample of the law and the reference, replicate j using the reference of training seed j. Over five replicates it is 0.00460 ± 0.00026 (mean $\pm$ s.e.), about 0.9% of the median σ_w. At 20,000 steps the fifteen-seed ReBF, PCBF and BCFM runs score 0.0282 ± 0.0009, 0.0410 ± 0.0010 and 0.0523 ± 0.0007 (mean $\pm$ s.e.), which is 6.1, 8.9 and 11.4 times this floor.

F Experimental Protocol

Environments. Three finite Markov reward processes with analytically known return laws, evaluated under a fixed policy: BERNOULLI ($\gamma = 0.5$), SOLITAIRE ($\gamma = 0.9$, a stochastic termination roll) and DISCRETE-MC (18 non-terminal states, $\gamma = 0.95$). Their return laws, derived in Appendix E, differ sharply in dispersion (Section 6).

Arms. Every arm is a triple (query clock, coupling ρ_1, correction weight κ) and is named by that triple; a single source file defines them, so no arm is specified twice. At $\kappa = 0$ the query clock is arithmetically inert, and we verified that the first training target is bit-identical under both clock flags; the two settings are one arm, run once for the headline comparisons; the two clock flags were trained separately as a check, and those twin runs differ by up to 4% of level.

Training and evaluation. 50k steps, batch 512, shared architecture and optimiser, 20-step Euler integration during training. Evaluation uses 50 ODE steps and 4×10^5 draws against the exact law; for DISCRETE-MC the reported W_1 averages all 18 non-terminal states. Reference-error floors under a perfect model are 6.5×10^{-4}, 3.0×10^{-3} and 7.4×10^{-3} respectively, well below every reported difference.

Selection and pairing. Tuned arms select κ on five seeds and are scored on five disjoint seeds; both split directions are reported. Comparisons are paired by seed, and ties are declared using the t critical value at the relevant sample size. Seeds share random-number streams across arms but not bit-identical trajectories: two runs with the same seed and an identical first target can differ by a few tenths of a percent after thousands of steps through floating-point accumulation alone, which bounds the resolution of any paired comparison at that scale.

The evaluation floor, and what a power control certifies. Every metric we report compares a finite sample from the learned critic against a finite reference, so it has a positive floor even for an exactly correct critic. We measure that floor rather than assume it: we draw twice from the *true* law at the sample sizes the evaluation uses and compute the same metric. For the depth sweep of Table 6 (4,000 learned draws against a 20,000-draw reference) this gives a W_1 floor of 0.0432 ± 0.0166 on SOLITAIRE at $\gamma = 0.9$. Separations are then reported as multiples of the relevant floor.

The main results are scored on a far finer instrument than the depth sweep: they use 4×10^5 learned draws against an *exact* closed-form reference, which removes reference-side noise entirely, giving floors of 6.5×10^{-4}, 3.0×10^{-3} and 7.4×10^{-3} on the three environments (each a single draw of a random quantity: the expected values are 1.0×10^{-3} and 4.5×10^{-3} on BERNOULLI and SOLITAIRE, and the DISCRETE-MC value averages states 1–4, its 18-state value being 0.006–0.009 across draws; Appendix E) — one to two orders of magnitude below the depth sweep's floor. Every headline level is at least $19\times$ its floor, and the headline gaps on SOLITAIRE and DISCRETE-MC are significant in both split directions ($p \leqslant 0.04$ at $n = 5$); the BERNOULLI gaps are 1.3–$2.2\times$ the floor at $n = 15$, and we report them as a tie. The main comparisons were therefore never near the resolution limit where they claim a winner; it is the depth sweep, run at 4,000 draws against a sampled reference, where the floor became load-bearing.

Two consequences are applied throughout. First, a claim that two arms *converge* requires the instrument to resolve the difference it asserts has vanished; where a gap falls below the floor's own standard deviation we say the arms are indistinguishable at the setting tested, and do not say that one caught up. Second, a power control certifies only the arm it is placed on. A control confirming that the *baseline* arm responds to a manipulation establishes that the cell resolves baseline-sized effects at the baseline's position in the instrument's range; it establishes nothing about an arm whose error sits an order of magnitude lower. We therefore measure the floor for the arm that *carries* the claim, at the sample size actually used.

We adopted this after discarding a 75-run sweep in which a preregistered power control passed — the baseline arm fell by $1.49\times$, paired $p = 0.022$ — while the arm whose profile was the readout sat at 0.9 to 1.15 times the measured floor. An exactly correct critic would have produced the same curve. The sweep completed, the registered rule passed, and the magnitudes, shape and error bars all looked ordinary: the failure had no symptom. Preregistration, held-out selection, paired seeds and convergence checking do not detect a power analysis aimed at the wrong arm; measuring the floor for the right one does, which is why the check is unconditional here rather than applied on suspicion.

Analytic references. Return-law references are derived from the environment dynamics rather than from any helper shipped with an environment; one such helper contained an off-by-one in the discount exponent, and results computed against it were discarded.

G Algorithm

Algorithm 1 One update of the retimed family; ReBF is $\rho_1 = 0$, $\kappa = 1$

1. Sample backup $H = (R, S', A', \tilde{\gamma})$ for c; independently draw X_0, E, t.
2. If $\tilde{\gamma} = 0$, set $X_t = (1-t)X_0 + tR$, $u = R - X_0$, and skip to step 7.
3. Set $X_0' = \rho_1 X_0 + \sqrt{1-\rho_1^2}E$ and $X_1' = \bar{\psi}^{\,1}(X_0' \mid S', A')$.
4. Set $X_t = (1-t)X_0 + t(R + \tilde{\gamma}X_1')$ and $C = X_1' - X_0$.
5. Compute α, τ, g by Eqs. (6)–(7); set $\xi = (X_t - tR)/\alpha$.
6. Query the frozen teacher and set $u = R + \tilde{\gamma}X_1' - X_0 + \kappa g(\bar{v}(\tau, \xi) - C)$.
7. Regress $v_\theta(t, X_t \mid c)$ onto $\mathrm{sg}(u)$; average target parameters.

Table 7: Terminal scale error of the same-time query with constant gain $\tilde{\gamma}$ under an exact teacher at $\rho_1 = 0$, $\gamma = 0.99$, as a percentage of the correct successor scale. Gaussian: Eq. (33) in closed form, identical for every σ. Bernoulli: the same constant-gain target on a 256-atom quantile compression of the BERNOULLI return law, estimated by fixed-seed binned Monte Carlo (2×10^7 draws, seed 0, 100 time bins, 200 state bins, 100 Euler steps, 2×10^5 evaluation draws). Retimed: the same estimator applied to the retimed target on the Bernoulli law. Its population value is exactly zero (Theorem 5.3 and Proposition B.5), so this column measures the estimator's floor, which also enters the Bernoulli column. None of these are averages over training seeds.

Backup n	$\tilde{\gamma} = 0.99^n$	Gaussian (any σ)	Bernoulli law	Retimed (floor)
1	0.9900	0.50%	0.45%	+0.09%
5	0.9510	2.48%	1.84%	+0.10%
10	0.9044	4.89%	3.50%	+0.10%
25	0.7778	11.62%	7.86%	+0.13%
50	0.6050	20.86%	13.22%	+0.19%

H Limitations

One scope statement belongs with the results before the limitations proper. As $\gamma \to 1$, the two clocks and gains converge for a fixed coupling; this does not imply convergence of the two empirical families, which also differ in ρ_1.

The analysis is for policy evaluation on small Markov reward processes under a fixed policy. Table 3 reports offline control on OGBench and D4RL; those scores are not a test of the theory. We have not studied how a flow critic should be used to extract a better policy in offline reinforcement learning. Extending the same target to other settings, including offline goal-conditioned reinforcement learning, needs further investigation. Only one of the three environments shows a large shared-noise mean bias, and we have marked which conclusions rest on it. Several results are proved in a correlated-Gaussian model rather than in general: the necessity half of the exactness set, the same-time closed form, and the scale-error identity; the last is also checked on a non-Gaussian law (Table 7). The conjugacy of Theorem 5.2 and its readout contraction (Theorem 5.2(iii)) are stated branchwise and in a specific metric; that contraction depends on the metric (Appendix B.7), and it is not a claim that same-time targets fail to contract. We prove convergence on canonical fields with independent generation (Theorem 5.3(iii)); we do not establish uniqueness or convergence of the self-consistent training map from non-canonical initial fields when the critic is its own teacher at $\kappa > 0$, nor, from any initial field, when generation is coupled ($\rho_1 > 0$), and off canonical fields we claim no field-level contraction of the self-consistent map for random transitions. The unique fixed point of $\mathcal{C}^\pi$, every fixed point of which is canonical, and its contraction in the endpoint Wasserstein pseudometric δ_p (Proposition 5.4) are unaffected. For the mean bias we measure under shared noise, the reused critic is the matched teacher only when its flow is straight (Proposition A.1(ii)). Amplification by $\gamma/(1-\gamma)$ is excluded, because the bias does not grow detectably between $\gamma = 0.99$ and 0.995 although that factor doubles, and a control with a deterministic whole backup, where the full-gain retimed map is exact (Appendix B.5), remains untested. The claim that the two clocks' trained arms converge at a fixed coupling as $\gamma \to 1$ is consistent with our $\gamma = 0.99$ measurements, but the equivalence test is void because both arms are still improving at 50k steps, so we assert no number. We did attempt a fourth discount point, at $\gamma = 0.999$, but those runs had not converged and we report no number from them. Finally, retiming does not by itself save an ODE solve — both constructions need a successor endpoint and then build their query algebraically.

I Online supplement: inference budget and per-task OGBench

The integration-budget study and the 38-task OGBench table are supplementary to the main results.

I.1 Inference-time integration budget

Figure 9 reads the same trained critics out with 1 to 200 Euler steps, so it measures the readout and not the critic. From 10 to 200 steps ReBF improves by a factor of two to four on every environment, and pinned PCBF by 1.3–2.1×. The same-time independent-noise arm, which the figure does not draw, gets *worse* on two of three ($0.030 \to 0.062$ on BERNOULLI and $0.242 \to 0.285$ on SOLITAIRE, paired $p \leqslant 0.023$; $0.662 \to 0.624$ on DISCRETE-MC, $p = 0.29$; seeds 0–4): a biased field is not rescued by integrating it more accurately, and a coarse readout hides the bias. On BERNOULLI and SOLITAIRE tuned PCBF nearly coincides with the uncorrected critic, because held-out tuning (on seeds 5–9) selects a small $\kappa^\star$ (0.02 and 0.05; 0.3 on DISCRETE-MC). On DISCRETE-MC tuned PCBF is a single point at 50 steps: the κ-ladder arms were re-scored on a shorter budget grid, and each arm is drawn only where it has data.

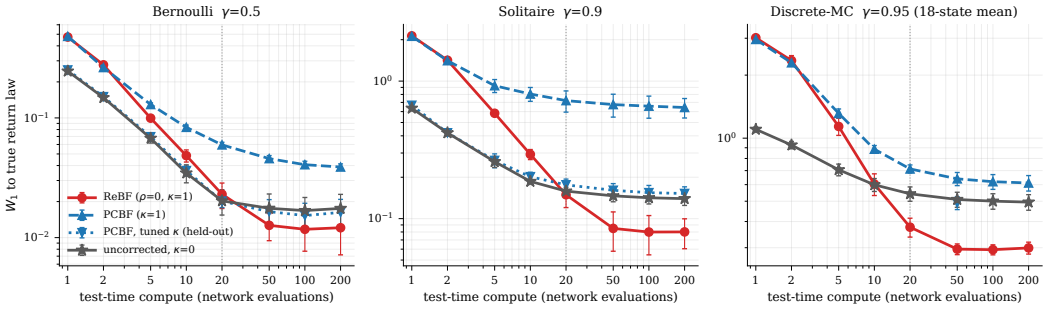


Figure 9: Test-time compute: W_1 to the exact law (4×10^5 draws) as the same trained critics are read out with 1 to 200 Euler steps (dotted line: the training resolution, 20), mean ± 1 s.d. over five paired seeds; tuned PCBF uses the $\kappa^\star$ selected on disjoint seeds. Lower is better; the x-axis is the readout budget, not training.

I.2 Per-task OGBench results

Table 8: **Full Offline RL Results.** Performance on 38 OGBench (Park et al., 2025a) and D4RL (Fu et al., 2020) tasks. The symbol $(*)$ indicates the task used for hyperparameter tuning within each domain. Results for IQN (Dabney et al., 2018a), CODAC (Ma et al., 2021), FloQ (Agrawalla et al., 2026), FQL (Park et al., 2025b), IQL (Kostrikov et al., 2022) and Value Flows (Dong et al., 2026) are from Xu et al. (2026, Table 3); PCBF results are from the supplied PCBF comparison table. ReBF results use the best evaluation checkpoint of each seed. **Bold numbers denote mean scores that are at least 95% of the highest reported mean on each task.**

	Algorithms							
Task	IQN	CODAC	FloQ	FQL	IQL	Value Flows	PCBF	**ReBF (Ours)**
cube-double-play-singletask-task1-v0	70 ± 14	80 ± 11	50 ± 24	61 ± 9	27 ± 5	$\mathbf{97 \pm 1}$	92 ± 5	$\mathbf{97 \pm 3}$
cube-double-play-singletask-task2-v0 $(*)$	24 ± 9	63 ± 4	72 ± 15	36 ± 6	1 ± 1	76 ± 7	$\mathbf{84 \pm 8}$	$\mathbf{82 \pm 9}$
cube-double-play-singletask-task3-v0	25 ± 6	66 ± 9	57 ± 14	22 ± 5	0 ± 0	73 ± 4	80 ± 7	$\mathbf{93 \pm 2}$
cube-double-play-singletask-task4-v0	10 ± 1	13 ± 2	8 ± 4	5 ± 2	0 ± 0	$\mathbf{30 \pm 5}$	$\mathbf{29 \pm 5}$	25 ± 6
cube-double-play-singletask-task5-v0	$\mathbf{81 \pm 8}$	$\mathbf{82 \pm 4}$	50 ± 11	19 ± 10	4 ± 3	69 ± 5	65 ± 8	$\mathbf{83 \pm 8}$
scene-play-singletask-task1-v0	$\mathbf{100 \pm 0}$	$\mathbf{99 \pm 0}$	$\mathbf{100 \pm 1}$	$\mathbf{100 \pm 0}$	94 ± 3	$\mathbf{99 \pm 0}$	$\mathbf{100 \pm 0}$	$\mathbf{100 \pm 0}$
scene-play-singletask-task2-v0 $(*)$	1 ± 0	85 ± 4	83 ± 10	76 ± 9	12 ± 3	$\mathbf{97 \pm 1}$	86 ± 11	$\mathbf{94 \pm 9}$
scene-play-singletask-task3-v0	94 ± 2	90 ± 3	$\mathbf{98 \pm 2}$	$\mathbf{98 \pm 1}$	32 ± 7	94 ± 2	$\mathbf{99 \pm 1}$	$\mathbf{99 \pm 1}$
scene-play-singletask-task4-v0	3 ± 1	0 ± 0	$\mathbf{9 \pm 7}$	5 ± 1	0 ± 1	7 ± 17	3 ± 3	$\mathbf{9 \pm 11}$
scene-play-singletask-task5-v0	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$
puzzle-4x4-play-singletask-task1-v0	41 ± 2	37 ± 32	$\mathbf{47 \pm 7}$	34 ± 8	12 ± 2	36 ± 3	39 ± 6	$\mathbf{46 \pm 13}$
puzzle-4x4-play-singletask-task2-v0	12 ± 4	10 ± 10	21 ± 6	16 ± 5	7 ± 4	27 ± 5	$\mathbf{28 \pm 4}$	$\mathbf{29 \pm 3}$
puzzle-4x4-play-singletask-task3-v0	$\mathbf{45 \pm 7}$	33 ± 29	36 ± 5	18 ± 5	9 ± 3	30 ± 4	38 ± 6	$\mathbf{45 \pm 10}$
puzzle-4x4-play-singletask-task4-v0 $(*)$	23 ± 2	12 ± 10	19 ± 5	11 ± 3	5 ± 2	28 ± 5	35 ± 5	$\mathbf{43 \pm 8}$
puzzle-4x4-play-singletask-task5-v0	16 ± 6	10 ± 8	16 ± 7	7 ± 3	4 ± 1	13 ± 2	$\mathbf{18 \pm 6}$	15 ± 4
cube-triple-play-singletask-task1-v0 $(*)$	29 ± 2	9 ± 5	32 ± 13	20 ± 6	4 ± 4	$\mathbf{59 \pm 12}$	25 ± 9	55 ± 15
cube-triple-play-singletask-task2-v0	0 ± 0	$\mathbf{1 \pm 0}$	0 ± 0	$\mathbf{1 \pm 2}$	0 ± 0	0 ± 0	0 ± 1	0 ± 0
cube-triple-play-singletask-task3-v0	1 ± 0	0 ± 0	$\mathbf{7 \pm 3}$	0 ± 0	0 ± 0	$\mathbf{7 \pm 3}$	3 ± 1	6 ± 3
cube-triple-play-singletask-task4-v0	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 0}$	$\mathbf{0 \pm 1}$	$\mathbf{0 \pm 0}$
cube-triple-play-singletask-task5-v0	0 ± 0	0 ± 0	0 ± 0	0 ± 0	1 ± 1	$\mathbf{2 \pm 1}$	0 ± 1	0 ± 0
pen-cloned-v1	$\mathbf{80 \pm 11}$	76 ± 2	72 ± 5	74 ± 11	$\mathbf{83}$	73 ± 5	78 ± 5	77 ± 3
pen-expert-v1	118 ± 19	$\mathbf{136 \pm 2}$	$\mathbf{140 \pm 8}$	$\mathbf{142 \pm 6}$	128	117 ± 3	$\mathbf{137 \pm 4}$	129 ± 5
door-cloned-v1	0 ± 0	0 ± 0	$\mathbf{3 \pm 2}$	2 ± 1	$\mathbf{3}$	0 ± 0	1 ± 0	1 ± 0
door-expert-v1	$\mathbf{105 \pm 0}$	$\mathbf{104 \pm 0}$	$\mathbf{104 \pm 0}$	$\mathbf{104 \pm 1}$	$\mathbf{107}$	$\mathbf{104 \pm 1}$	$\mathbf{106 \pm 0}$	$\mathbf{106 \pm 0}$
hammer-cloned-v1	0 ± 0	6 ± 0	10 ± 8	$\mathbf{11 \pm 9}$	2	1 ± 0	2 ± 1	2 ± 0
hammer-expert-v1	121 ± 7	$\mathbf{126 \pm 1}$	$\mathbf{125 \pm 2}$	$\mathbf{125 \pm 3}$	$\mathbf{129}$	$\mathbf{125 \pm 5}$	$\mathbf{126 \pm 2}$	$\mathbf{126 \pm 2}$
relocate-cloned-v1	0 ± 0	0 ± 0	0 ± 0	0 ± 0	$\mathbf{2}$	0 ± 0	0 ± 0	0 ± 0
relocate-expert-v1	103 ± 0	103 ± 2	$\mathbf{108 \pm 2}$	$\mathbf{107 \pm 1}$	$\mathbf{106}$	102 ± 2	$\mathbf{109 \pm 1}$	$\mathbf{108 \pm 1}$
visual-antmaze-teleport-navigate-singletask-task1-v0 $(*)$	2 ± 1	–	–	2 ± 1	5 ± 2	$\mathbf{10 \pm 4}$	8 ± 2	9 ± 3
visual-antmaze-teleport-navigate-singletask-task2-v0	7 ± 3	–	–	6 ± 1	10 ± 2	17 ± 5	$\mathbf{19 \pm 4}$	$\mathbf{19 \pm 4}$
visual-antmaze-teleport-navigate-singletask-task3-v0	6 ± 4	–	–	9 ± 4	7 ± 7	16 ± 3	$\mathbf{18 \pm 4}$	16 ± 2
visual-antmaze-teleport-navigate-singletask-task4-v0	4 ± 2	–	–	9 ± 1	4 ± 6	16 ± 5	$\mathbf{18 \pm 5}$	15 ± 4
visual-antmaze-teleport-navigate-singletask-task5-v0	2 ± 1	–	–	1 ± 1	2 ± 1	$\mathbf{8 \pm 2}$	$\mathbf{8 \pm 3}$	7 ± 2
visual-cube-double-play-singletask-task1-v0 $(*)$	4 ± 1	–	–	23 ± 4	$\mathbf{34 \pm 23}$	$\mathbf{35 \pm 2}$	10 ± 3	11 ± 4
visual-cube-double-play-singletask-task2-v0	0 ± 0	–	–	0 ± 0	3 ± 1	$\mathbf{4 \pm 2}$	0 ± 0	1 ± 1
visual-cube-double-play-singletask-task3-v0	0 ± 0	–	–	0 ± 0	7 ± 4	$\mathbf{11 \pm 2}$	0 ± 0	1 ± 1
visual-cube-double-play-singletask-task4-v0	0 ± 0	–	–	$\mathbf{4 \pm 2}$	2 ± 1	2 ± 1	0 ± 0	0 ± 0
visual-cube-double-play-singletask-task5-v0	1 ± 1	–	–	4 ± 1	11 ± 2	$\mathbf{13 \pm 3}$	3 ± 1	3 ± 1