MINT: Meeting-time Indicators for Truncation in Multi-Step Off-Policy RL

Seungyub Han[1], **Taehyun Cho**[1], **Dohyeong Kim**[1], **Kyungjae Lee**[2*], **Jungwoo Lee**[1*]
[1]Seoul National University, [2]Korea University
{seungyubhan,talium,rlaehgudhoho,junglee}@snu.ac.kr kyungjae_lee@korea.ac.kr

Abstract

Many multi-step off-policy temporal-difference methods handle policy mismatch through products of per-step trace coefficients, leading to a trade-off between variance control and long-horizon signal. The maximum trace length consequently becomes a sensitive hyperparameter. We propose **MINT** (Meeting-time INdicators for Truncation), which replaces these products with a binary stopping indicator: one until coupled behavior and target trajectories first meet, zero thereafter. At the true value function, the indicator's idempotence yields a variance bound governed by meeting-time survival. For practical implementation, a single sampled action provides a model-free surrogate for the stopping rule, requiring no transition model or mixing-time estimate. We derive separate sufficient conditions for contraction of the state-coupled operator and its practical surrogate under fixed-policy evaluation. On MuJoCo and DeepMind Control tasks, `MINT` matches or exceeds tuned multi-step baselines without cap-length tuning. Additional experiments on Atari 100k, RWRL, and OGBench characterize the benefits and limitations of adaptive truncation across tasks and transition dynamics.
Code: https://github.com/seungyubhan/mint

1 Introduction

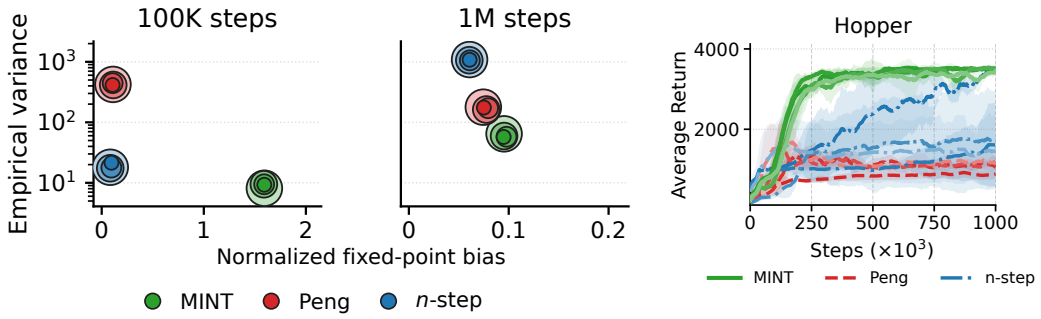
Multi-step temporal-difference (TD) learning propagates rewards over longer horizons, making it a central tool for policy evaluation and credit assignment in reinforcement learning [26, 4, 17, 40, 32]. In off-policy learning, however, data are collected by a behavior policy β that differs from the target policy π. Longer backups accumulate the effects of this mismatch, making the maximum trace length, or *cap length*, a consequential choice for learning accuracy and stability. Many off-policy multi-step methods handle this mismatch through products of per-step trace coefficients. Unclipped importance-sampling ratios can amplify variance over long traces, while clipping controls the weights but can attenuate distant contributions [29, 33]. Conservative methods such as Retrace trade signal retention for strong convergence guarantees; non-conservative methods such as Peng's $Q(\lambda)$ retain more signal but can introduce off-policy bias or require policy-alignment conditions for stability [23, 32, 14]. These trade-offs motivate a change to the trace construction itself: replacing accumulated coefficient weights with a stopping rule that adapts the backup horizon to policy mismatch.

The on-policy analysis of Duan and Wainwright [9] offers a complementary perspective by relating trace (truncation) length to the chain's mixing time. Applying this prescription off-policy requires an additional argument, and estimating the relevant mixing time can be difficult without a transition model. This motivates a sample-based alternative:

Can a coupling-based stopping rule adapt the multi-step horizon while admitting explicit variance and contraction bounds?

*Corresponding authors.

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



(a) Bias–variance trade-off.

(b) Learning curves.

Figure 1: **Cap-length sensitivity on MuJoCo Hopper (SAC).** (a) Bias and variance at 100K and 1M steps for caps $\{3, 5, 10, 20\}$ (darker shades indicate shorter caps). `MINT` stays near the observed Pareto frontier; several baseline configurations have higher variance at longer caps. (b) Learning curves for the same caps. `MINT` attains high returns across the tested values, whereas baseline performance varies more with the cap. These observations concern this task; broader comparisons and cases where the cap remains consequential are discussed in Section 5.

We introduce **MINT** (Meeting-time **IN**dicators for Truncation), which replaces trace-coefficient products with a binary stopping indicator. In the theoretical construction, behavior and target state processes are coupled, and the trace is retained until their first agreement time T. The stopping indicator is one before T and remains zero thereafter, even if the underlying processes subsequently separate. This turns the weighted multi-step backup into a randomly truncated sum of TD errors (Section 3). The key analytical property is *idempotence*: squaring a binary stopping indicator leaves it unchanged. At the true value function Q^π, this yields a variance bound expressed through meeting-time survival probabilities (Lemma 3.4). For model-free implementation, we construct an action-level stopping surrogate from the policy-density ratio at a single sampled action, requiring neither transition-kernel access nor a mixing-time estimate. The surrogate defines its own stopping process; we analyze its contraction separately from that of the state-coupled operator under fixed-policy evaluation (Sections 3.2 and 4). Figure 1 illustrates the empirical motivation on MuJoCo Hopper: `MINT` stays near the observed bias–variance frontier and attains higher returns with less sensitivity to the tested cap lengths. Across MuJoCo and DeepMind Control tasks, it matches or exceeds tuned multi-step baselines *without cap-length tuning*. Additional experiments on Atari 100k, RWRL, and OGBench, together with ablations and diagnostics, characterize the benefits and limitations of adaptive truncation across tasks and transition dynamics (Section 5 and Appendices G and H).

Contributions. Our contributions are fourfold. First, we propose a coupling-based multi-step operator and a model-free action-level surrogate that replace coefficient products with stochastic truncation (Section 3). Second, we establish a survival-based variance bound for the state-coupled estimator at Q^π (Lemma 3.4). Third, for fixed-policy evaluation, we derive separate sufficient conditions for contraction of the state-coupled operator and the expected operator of the practical surrogate (Theorems B.9 and 4.2). Fourth, we demonstrate performance matching or exceeding multi-step baselines without cap-length tuning on MuJoCo and DeepMind Control, and characterize benefits and limitations on Atari 100k, RWRL, and OGBench (Section 5 and Appendices F.5, G and H).

2 Preliminaries

We consider a Markov decision process (MDP) $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{P}_0, r, \gamma)$ with finite state and action spaces, transition kernel $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$, initial distribution $\mathcal{P}_0$, bounded reward $r : \mathcal{S} \times \mathcal{A} \to [-r_{\max}, r_{\max}]$, and discount $\gamma \in [0, 1)$ [39]. For a policy π, the Bellman operator is $\mathcal{T}^\pi Q := r + \gamma \mathcal{P}^\pi Q$, where $(\mathcal{P}^\pi Q)(s, a) := \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a),\, a' \sim \pi(\cdot|s')}[Q(s', a')]$. Trajectories $(s_t, a_t, r_t)_{t \geq 0}$ are generated by a behavior policy β, and our goal is to estimate the action-value function Q^π of a target policy $\pi \neq \beta$ from these off-policy samples [32, 40].

2.1 The Retrace Operator and the Variance Explosion Problem

A central challenge in multi-step off-policy evaluation is correcting the distributional mismatch between β and π without inflating the variance of the multi-step return. The *general Retrace operator* [29] formalizes this correction through a sequence of per-step *trace coefficients* $c_i := c(s_i, a_i)$:

$$\mathcal{R}Q \;=\; Q \;+\; \mathbb{E}_\beta\left[\sum_{t=0}^{\infty}(\gamma\lambda)^t\Big(\prod_{i=1}^{t} c_i\Big)\,\delta_t^Q\right], \qquad \delta_t^Q := r_t + \gamma\,\mathbb{E}_\pi[Q(s_{t+1},\cdot)] - Q(s_t,a_t), \quad (1)$$

with the convention $\prod_{i=1}^{0} c_i = 1$ and $\lambda \in [0,1]$ controlling the geometric decay of the trace [5]. Existing multi-step off-policy methods instantiate Equation (1) by a choice of c_i: Retrace [29] uses $c_i = \min(1, \pi(a_i \mid s_i)/\beta(a_i \mid s_i))$; Harutyunyan's $Q^\pi(\lambda)$ uses $c_i = 1$ [18]. A sufficient conservative constraint is $0 \le c_i \le \pi(a_i \mid s_i)/\beta(a_i \mid s_i)$, which guarantees contraction of the expected evaluation operator toward Q^π for fixed policies [29]. Harutyunyan's operator does not enforce this constraint. Peng's $Q(\lambda)$ instead uses a mixture of uncorrected multi-step returns; Table 1 distinguishes these constructions.

The variance explosion problem. For a finite cap L, write $Z_t := (\gamma\lambda)^t(\prod_{i=1}^{t} c_i)\delta_t^Q$. The estimator's second moment has the exact expansion

$$\mathbb{E}\left[\left(\sum_{t=0}^{L-1} Z_t\right)^2\right] = \sum_{t=0}^{L-1}\mathbb{E}[Z_t^2] + 2\sum_{0\le u<t<L}\mathbb{E}[Z_u Z_t],$$

whose diagonal terms are

$$\mathbb{E}[Z_t^2] = (\gamma\lambda)^{2t}\mathbb{E}\left[\Big(\prod_{i=1}^{t} c_i^2\Big)(\delta_t^Q)^2\right]. \quad (2)$$

The trace weights and TD errors need not be independent, and the cross terms need not vanish at an arbitrary Q. With unclipped importance-sampling coefficients, squared ratio products can amplify second moments over long traces; clipped coefficients need not exhibit this growth. Conservative choices such as Retrace keep the coefficient products bounded but can attenuate distant contributions. Non-conservative methods can retain more long-horizon signal, but their return construction may introduce off-policy bias or require additional conditions for stability [23, 33]. In practice, these trade-offs can make performance sensitive to the cap length, motivating task-specific tuning [19].

2.2 Mixing Time and Effective Horizon

For an ergodic behavior chain, let μ_β be the stationary distribution of $\mathcal{P}^\beta$. Its total-variation mixing time is

$$\tau_{\text{mix}}^\beta := \min\left\{t \ge 0 : \sup_s \left\|(\mathcal{P}^\beta)^t(s,\cdot) - \mu_\beta\right\|_{\text{TV}} \le \tfrac{1}{4}\right\},$$

which measures convergence to stationarity uniformly over initial states [24]. For $\gamma \in (0,1)$, write $H := (1-\gamma)^{-1}$ for the discount horizon. For on-policy evaluation ($\beta = \pi$), Duan and Wainwright [9] analyze multi-step kernel least-squares TD estimates. Their error bounds distinguish martingale fluctuations from a Bellman-residual contribution affected by temporal dependence. In the uniformly bounded-reward setting, consistent with our assumption $|r(s,a)| \le r_{\max}$, the look-ahead condition

$$K \;\gtrsim\; \min\{H, \tau_*^\beta\} \quad (3)$$

is sufficient for their stated finite-sample error bounds, subject to their mixing, kernel-regularity, and sample-size assumptions [9, Section 3.3.2, Eq. (29b)]. Here $\tau_*^\beta \ge 1$ denotes the minorization-based mixing parameter of the common data-generating and target chain in that analysis, distinct from the TV threshold τ_{mix}^β defined above. Using a mixing-based prescription requires estimating chain-level quantities from data [46]. Moreover, the cited on-policy result does not directly provide an off-policy guarantee when $\beta \ne \pi$. These limitations motivate a sample-based stopping rule for off-policy multi-step evaluation, developed in Section 3. A coalescent coupling of two behavior chains from different initial states bounds τ_{mix}^β. The behavior–target first agreement instead controls MINT trace survival; these two times are distinguished in Appendix B.6.

Table 1: **Multi-step off-policy operators under fixed policies.** W_t excludes $(\gamma\lambda)^t$; $\mathbb{E}[W_t^2]$ describes squared weights, not the full estimator variance.

Operator	Trace / return rule	$\mathbb{E}[W_t^2]$	Conservative?	Fixed point
Retrace [29]	$c_i = \min(1, \pi/\beta)$	$\mathbb{E}[\prod_{i=1}^{t} c_i^2]$	Yes	Q^π
Peng's $Q(\lambda)$ [23]	$c_i = 1^\dagger$	1	No	Q^{π_λ}
Uncorrected n-step	$\mathcal{N}_n^\ddagger$	$\mathbf{1}\{t < n\}$	No	β-dependent
MINT	$c_i = A_i \in \{0, 1\}$	$\Pr(T > t)$	No	Q^π

† Peng's return mixture has this TD-error representation with target $\pi_\lambda := \lambda\beta + (1-\lambda)\pi$, rather than π. ‡ $\mathcal{N}_n := (\mathcal{T}^\beta)^{n-1}\mathcal{T}^\pi$ is the ordinary fixed-horizon return ($\lambda = 1$); its fixed point is generally behavior-dependent for $n > 1$.

3 The MINT Operator

We introduce MINT, which replaces products of real-valued trace coefficients with a binary stopping rule. We first define the state-coupled operator, then establish a variance bound at Q^π, and finally describe its model-free action-level surrogate. Let S_t^β and S_t^π start from a shared initial state and evolve under a specified joint sampling scheme. At each policy-generated step, a maximal coupling of $\mathcal{P}^\beta(S_t^\beta, \cdot)$ and $\mathcal{P}^\pi(S_t^\pi, \cdot)$ specifies their next-state pair. Let $\mathcal{H}_t$ contain both state histories, past actions and rewards, and coupling randomness revealed through time t, before sampling the current behavior action. The action is sampled from $\beta(\cdot \mid S_t^\beta)$ conditionally on $\mathcal{H}_t$; write $\mathcal{G}_t = \mathcal{H}_t \vee \sigma(a_t)$. The joint construction must preserve the environment's conditional law of (r_t, S_{t+1}^β) given $\mathcal{G}_t$. One implementation samples the behavior action and transition first, then the auxiliary next state from the conditional law of the chosen coupling (Appendix B.1).

Definition 3.1 (Disagreement and stopping indicators)**.** For $i \geq 1$, define

$$A_i := \mathbf{1}\{S_i^\beta \neq S_i^\pi\}, \qquad p_i := \Pr(A_i = 1 \mid \mathcal{H}_{i-1}). \tag{4}$$

At a policy-generated step, conditional maximal coupling gives

$$p_i = \mathrm{TV}\Big(\mathcal{P}^\beta(S_{i-1}^\beta, \cdot), \mathcal{P}^\pi(S_{i-1}^\pi, \cdot)\Big).$$

When conditioning on a prescribed initial action, the first-step law is the corresponding conditional initial coupling. The conditional identity above does not assert maximality of the unconditional time-i marginals; in general,

$$\mathrm{TV}\Big(\mathcal{L}(S_i^\beta), \mathcal{L}(S_i^\pi)\Big) \leq \Pr(A_i = 1).$$

Definition 3.2 (First agreement time)**.** The first agreement time and cumulative trace mask are

$$T_{\beta,\pi} := \inf\{i \geq 1 : A_i = 0\}, \qquad W_t := \prod_{i=1}^{t} A_i, \quad W_0 := 1, \tag{5}$$

with $\inf \varnothing = \infty$. Thus

$$W_t = \mathbf{1}\{T_{\beta,\pi} > t\}. \tag{6}$$

The cumulative mask remains zero after the first agreement, even if the state processes subsequently separate. We write $s_t = S_t^\beta$ for brevity, and let $\mathbb{E}_x, \Pr_x$ refer to the joint trajectory law conditioned on the starting pair $x = (s_0, a_0)$. Replacing $\prod_{i=1}^{t} c_i$ in Equation (1) by W_t yields

$$(\mathcal{R}_{p,\lambda}Q)(x) = Q(x) + \mathbb{E}_x\left[\sum_{t\geq 0}(\gamma\lambda)^t W_t \delta_t^Q\right] = Q(x) + \mathbb{E}_x\left[\sum_{t=0}^{T_{\beta,\pi}-1}(\gamma\lambda)^t \delta_t^Q\right]. \tag{7}$$

The expectation includes environment and coupling randomness. The subscript p records the chosen disagreement process; the indicators need not be independent. As shown below, Q^π is a fixed point of this expected evaluation operator. Uniqueness and convergence follow under the separate contraction condition in Section 4.

Algorithm 1 `MINT` target (summary). Full procedure: Algorithm 2.

Require: Trajectory with behavior densities, π, $Q_{\bar{\theta}}$, γ, λ, $K_{\max} \geq 1$

1: For $1 \leq i < K_{\max}$, compute $\hat{p}_i$ by Equation (10) and independently draw $\hat{A}_i \sim \text{Bernoulli}(\hat{p}_i)$

2: $\hat{K} \leftarrow \min(K_{\max}, \hat{T}_{\beta,\pi})$

3: **return** $Q_{\bar{\theta}}(s_0, a_0) + \sum_{t=0}^{\hat{K}-1} (\gamma\lambda)^t \delta_t^{Q_{\bar{\theta}}}$

3.1 Variance at the Target Value Function

Two properties enter the variance calculation: conditional mean-zero TD errors eliminate cross terms, while idempotence turns squared binary masks into survival indicators.

Definition 3.3 (Discounted survival functional)**.** For $\gamma \in (0,1)$, $\lambda \in [0,1]$, and starting pair x, define

$$M_x(\gamma,\lambda) := \sum_{t\geq 0} (\gamma\lambda)^t \Pr_x(T_{\beta,\pi} > t) = \mathbb{E}_x\left[\frac{1-(\gamma\lambda)^{T_{\beta,\pi}}}{1-\gamma\lambda}\right] \leq \min\left\{\mathbb{E}_x[T_{\beta,\pi}], \frac{1}{1-\gamma\lambda}\right\}. \tag{8}$$

For a distribution of starting pairs, M without subscript denotes the corresponding average of M_x. The contraction analysis in Section 4 retains the starting-pair dependence and then takes a supremum.

Lemma 3.4 (Martingale property, fixed point, and variance)**.** *Under the joint sampling conditions above, let $\delta_t^{Q^\pi}$ be the TD error of Equation (1) at Q^π. Set the deterministic bound $C := \|r\|_\infty + (1+\gamma)\|Q^\pi\|_\infty < \infty$. Then, for every starting pair x:*

(i) Martingale property. *$\mathbb{E}_x[\delta_t^{Q^\pi} \mid \mathcal{G}_t] = 0$ and W_t is $\mathcal{G}_t$-measurable.*

(ii) Fixed point. *$\mathcal{R}_{p,\lambda} Q^\pi = Q^\pi$.*

(iii) Variance. *For $\hat{\Delta} := \sum_{t\geq 0} (\gamma\lambda)^t W_t \delta_t^{Q^\pi}$,*

$$\text{Var}_x(\hat{\Delta}) = \mathbb{E}_x[\hat{\Delta}^2] = \sum_{t\geq 0} (\gamma\lambda)^{2t} \mathbb{E}_x[W_t (\delta_t^{Q^\pi})^2] \leq C^2 M_x(\gamma^2, \lambda^2). \tag{9}$$

For $u < t$, the earlier TD error and both masks are measurable at $\mathcal{G}_t$, so the tower property removes the cross term. Idempotence then gives $W_t^2 = W_t$, yielding the diagonal expression in Equation (9). The full proof is in Appendix B.1. The cross-term argument also applies to other suitably adapted trace weights; the binary construction specifically expresses its diagonal weights through survival probabilities. Then, we have

$$\text{Var}_x(\hat{\Delta}) \leq C^2 \min\left\{\mathbb{E}_x[T_{\beta,\pi}], \frac{1}{1-\gamma^2\lambda^2}\right\}.$$

This bound concerns evaluation at Q^π; it does not establish variance dominance over Retrace or describe the full error of a learned critic. It offers a survival-based perspective alongside the mixing-based analysis in Section 2.2. Contraction requires the additional bounds developed in Section 4.

3.2 Practical Implementation

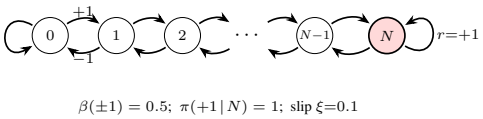
Constructing the state coupling requires transition-kernel information. The implementation instead uses the action-level keep probability

$$\hat{p}_i := 1 - \min\left(1, \frac{\pi(a_i \mid s_i)}{\beta(a_i \mid s_i)}\right). \tag{10}$$

Here β is the policy that generated the stored transition: its log-density is recorded at collection time and compared with the current π at replay. The ratio is evaluated only on sampled actions with positive behavior density. Using fresh randomness independent of subsequent rewards and transitions, draw $\hat{A}_i \sim \text{Bernoulli}(\hat{p}_i)$ and define

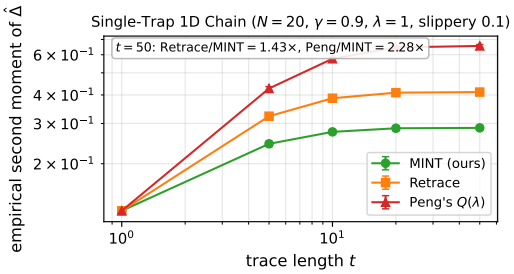
$$\hat{T}_{\beta,\pi} := \inf\{i \geq 1 : \hat{A}_i = 0\}, \qquad \hat{K} := \min(K_{\max}, \hat{T}_{\beta,\pi}), \tag{11}$$

with $\inf \varnothing = \infty$. The hard cap $K_{\max}$ is a protocol constant; the sampled stopping time selects the realized horizon within that cap. The rule retains the trace on disagreement: when $\pi = \beta$, $\hat{p}_i = 0$ and the backup is one-step. A larger $\hat{p}_i$ makes continuation more likely at that step. This differs from retaining traces in proportion to policy agreement. The expected effective horizon $K_{\text{eff}} := \mathbb{E}[\hat{K}]$ depends on the full sequence of keep probabilities and the cap.



$\beta(\pm 1) = 0.5;\ \pi(+1 \mid N) = 1;$ slip ξ=0.1

(a) Single-Trap chain MDP.



(b) Empirical second moment of $\hat{\Delta}$.

Figure 2: **Single-Trap chain: empirical second moment across trace lengths.** (a) The Single-Trap chain (N=20 states, sparse reward r=+1 at the trap state, slippery transitions ξ=0.1, behavior β uniform, target π uniform except $\pi(+1 \mid N)$=1). Reflecting boundary at s=0. (b) Empirical second moment of the multi-step Bellman-residual estimator $\hat{\Delta}$ on the Single-Trap chain over trace lengths $\{1, 5, 10, 20, 50\}$ at γ=0.9, λ=1. At t=50, the second moments of Retrace and Peng's $Q(\lambda)$ are respectively 1.43× and 2.28× that of `MINT`. Over the tested trace lengths, `MINT`'s curve levels off while both baselines continue to increase.

Relation to the state-coupled analysis. The conditional mean of the surrogate is the action-level policy TV,

$$\mathbb{E}_{a\sim\beta(\cdot|s)}[\hat{p}(s,a)] = D_s \geq d(s) := \mathrm{TV}(\mathcal{P}^{\beta}(s,\cdot), \mathcal{P}^{\pi}(s,\cdot)),$$

by data processing (Proposition A.4). This compares one-step kernels from a shared state; it does not identify the surrogate's multi-step survival law with that of the state coupling. The expected surrogate evaluation operator has its own finite-cap contraction condition (Theorem B.9 and Appendix B.8). These fixed-policy results do not establish convergence with changing replay policies or nonlinear critics. The RWRL diagnostics examine the one-step relationship and empirical backup gains (Appendices G.2 and G.4); the perturbation comparison and the distinct couplings for behavior-chain mixing and MINT survival are detailed in Lemma B.5 and Appendix B.6.

4 Theoretical Analysis

The variance lemma (Lemma 3.4) controls the estimator at Q^{π}. We next give a sufficient contraction condition for the state-coupled operator under an arbitrary behavior policy. The survival quantities in the contraction analysis depend on the starting pair, as in Definition 3.3.

Definition 4.1 (Survival-weighted policy TV)**.** Let $D_s := \mathrm{TV}(\beta(\cdot \mid s), \pi(\cdot \mid s)) \in [0,1]$, and let $\mathbb{E}_x$ and $\Pr_x$ denote expectation and probability for a starting pair $x = (s_0, a_0)$. Write $T = T_{\beta,\pi}$ and define

$$M_x := \sum_{t\geq 0} (\gamma\lambda)^t \Pr_x(T > t), \qquad N_x := \sum_{t\geq 1} (\gamma\lambda)^t \mathbb{E}_x[\mathbf{1}\{T > t\} D_{s_t}]. \tag{12}$$

Define $D^x_{\mathrm{wgt}} := N_x/(M_x - 1)$ when $M_x > 1$, and $D^x_{\mathrm{wgt}} := 0$ otherwise, and set $\bar{D}_{\mathrm{wgt}} := \sup_x D^x_{\mathrm{wgt}}$.

Theorem 4.2 (Contraction of $\mathcal{R}_{p,\lambda}$)**.** *Under the joint sampling conditions of Section 3, fix policies β, π, $\gamma \in (0,1)$, and $\lambda \in [0,1]$. Define*

$$\kappa := \sup_x \{1 - (1-\gamma)M_x + 2N_x\}. \tag{13}$$

Then, for all bounded Q, Q',

$$\|\mathcal{R}_{p,\lambda}Q - \mathcal{R}_{p,\lambda}Q'\|_{\infty} \leq \kappa \|Q - Q'\|_{\infty}. \tag{14}$$

If $\kappa < 1$, $\mathcal{R}_{p,\lambda}$ is a κ-contraction with unique fixed point Q^{π}.

For each starting pair, $1 - (1-\gamma)M_x$ is the row-mass contribution and $2N_x$ bounds the negative-weight correction. In the on-policy case $M_x = 1$ and $N_x = 0$ for all x, recovering $\kappa = \gamma$; the full proof is in Appendix B.2. This theorem concerns the state-coupled operator; the action-dependent surrogate requires its own analysis (see also Table 4). Quantitative contraction-budget simulations for $\gamma \in \{0.9, 0.99\}$ are provided in Appendix C.2. A separate exact-control result for Markov trace operators, under additional overlap and policy-improvement hypotheses, is given in Theorem B.8.

4.1 Closed-form regime and empirical validation

Under the cross-Doeblin condition [8, 24, 20] (Assumption A.3), the coupled kernels share a common component of mass ρ. Together with the conditional initial-step overlap specified in Corollary B.1, this gives the tail bound $\Pr_x(T_{\beta,\pi} > t) \le (1-\rho)^t$ uniformly over starting pairs x. Substituting this tail bound into Theorem 4.2 gives the sufficient condition $\gamma(1+2\eta) < 1$, where $\eta := \bar{D}_{\mathrm{wgt}}\lambda(1-\rho)$ (Figure 3). Larger overlap ρ offsets larger survival-weighted mismatch, while smaller λ expands the contractive region. The full analysis and contraction-budget tables are given in Appendices C and C.2, with a restart-chain illustration in Figure 9. Further empirical evaluation is presented in Section 5.

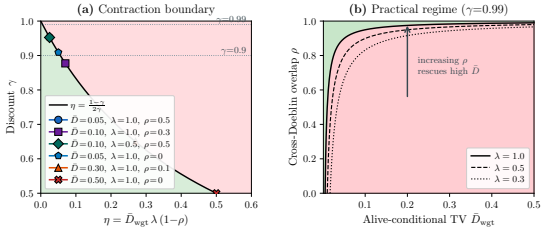


Figure 3: Contraction phase diagram (Corollary B.1); green: contraction holds, pink: no guarantee. **(a)** Boundary in (η, γ) plane with $\eta{:=}\bar{D}_{\mathrm{wgt}}\lambda(1{-}\rho)$; only the product η matters. **(b)** $(\bar{D}_{\mathrm{wgt}}, \rho)$ plane at $\gamma{=}0.99$: higher overlap ρ rescues contraction even at large $\bar{D}_{\mathrm{wgt}}$.

We evaluate estimator behavior on the Single-Trap chain (Figure 2, Appendix D), where $T_{\beta,\pi}$, $\bar{D}_{\mathrm{wgt}}$, and M are available in closed form, isolating estimator behavior from network-capacity, exploration, and replay-distribution effects. Separate high-mismatch ablations compare the action-level surrogate with horizon-matched random truncation and a symmetric log-ratio heuristic. In these stress tests, the cap reduces the retained horizon and sampled backup MSE; they do not establish a general surrogate advantage over matched random truncation (Appendix D.5).

Remark. The implemented keep rule $\hat{p}(s,a) = 1 - \min(1, \pi(a \mid s)/\beta(a \mid s))$ admits a separate contraction analysis. For fixed behavior and target policies on finite state and action spaces, a sufficient overlap condition is $\pi(a \mid s) \ge \lambda\beta(a \mid s)/(1+\lambda)$ for all (s,a), which yields a γ-contraction with unique fixed point Q^π for every finite cap $K_{\max} \ge 1$. Theorem B.9 and Corollary B.10 give the more general negative-weight condition, precise sampling hypotheses, and proof.

5 Experiments

The preceding analysis relates state-coupled contraction to survival and policy mismatch, while the toy experiments examine estimator behavior (Section 4.1 and Appendix D.5). We next evaluate the method on deep-RL benchmarks:

RQ1 How does `MINT` compare with multi-step baselines under matched experimental budgets?

RQ2 Does stochastic truncation produce a per-task adaptive trace length without estimating τ_{mix}?

RQ3 Does stochastic truncation improve learning with a matched maximum horizon and λ?

RQ4 How do target estimation and learned-critic behavior change under stochastic dynamics?

5.1 Experimental Setup

We evaluate on MuJoCo [42] and DeepMind Control [41] under SAC [15] and TD3 [13] backbones, with additional horizon diagnostics on RWRL [10] and off-policy evaluation on OGBench [31]. We also evaluate Atari 100k [22] with DrQ(ε) [47] and DER [44] backbones (Appendix F.5). In the main five-task MuJoCo comparison, all multi-step methods use a common cap of 5, with $\lambda = 0.7$ for MINT, Peng's $Q(\lambda)$, and Retrace; SAC uses fixed $\alpha = 0.2$ (Table 10). We train for 1M environment steps with five seeds per method and task, and aggregate normalized returns using the interquartile mean (IQM) with 95% stratified bootstrap confidence intervals [3]. Comparisons share the stated backbone and matched settings within each experiment; additional experiments use their separately reported caps, discounts, and entropy protocols. Full settings and normalization references are in Appendices E and F.1.

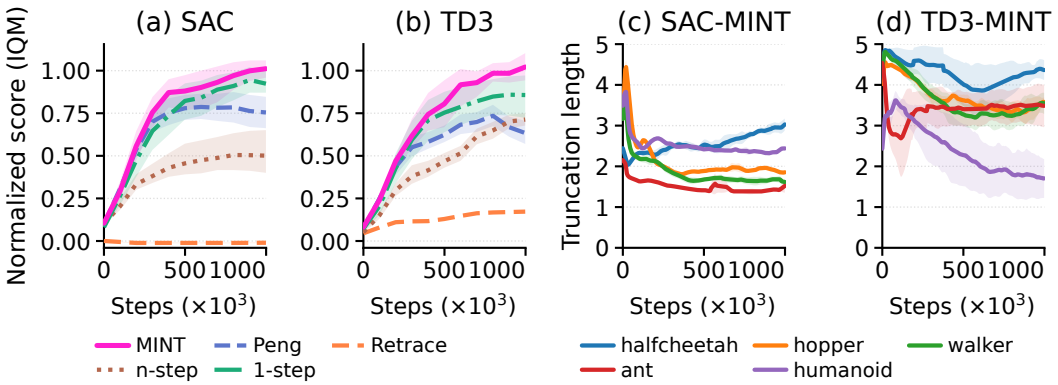


Figure 4: **Performance and adaptive truncation behavior on MuJoCo tasks.** (a–b) Aggregate normalized scores (IQM) of SAC and TD3 variants over training steps; shaded regions show 95% stratified bootstrap confidence intervals across five seeds per task. (c–d) Mean truncation lengths selected by `MINT` across five MuJoCo environments, with ± 1 standard deviation across seeds. SAC horizons shorten over training, while TD3 horizons remain closer to the cap.

5.2 Performance on Control Benchmarks

MuJoCo (dense reward). Figure 4 reports IQM-normalized scores across five MuJoCo tasks for SAC- and TD3-based methods. SAC-`MINT` achieves the strongest aggregate performance and learns fastest among the compared SAC methods. SAC-Retrace, the only conservative baseline, matches `MINT` on `humanoid-v2` and lags on the remaining four tasks. Peng's $Q(\lambda)$ and uncorrected n-step often underperform even the one-step SAC baseline. Per-task SAC and TD3 curves are provided in Figures 12 and 13.

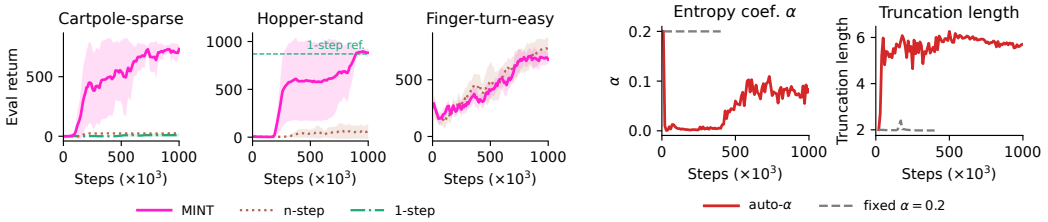
DMControl (sparse reward). We evaluate `MINT` on three sparse-reward DMControl tasks: `CartPole-Balance-Sparse`, `CartPole-Swingup-Sparse`, and `Ball-in-Cup-Catch`. On `CartPole-Balance-Sparse` (Figure 15a in Appendix F.2), `MINT` outperforms fixed-horizon baselines, achieving faster learning and higher final returns across the displayed λ values. Return estimation alone does not address tasks in which reward discovery is the bottleneck, such as the tested `Acrobot-Swingup-Sparse` setting (Appendix F.3 and Section 5.4).

Discrete-action benchmark (Atari 100k). On Atari 100k [22], direct integration into DrQ with per-step hard target synchronization diverges. A longer synchronization period stabilizes Q-values, but removing the intermediate $V - Q$ terms further improves returns on `Seaquest` and `Qbert` with unchanged trace lengths (three seeds). This implicates these terms in the remaining performance loss under that protocol. Separate, partial-seed Boltzmann-target results improve on the greedy-target variant in selected games; reduced overestimation is a possible explanation, not an isolated mechanism (Appendix F.5).

Answer to RQ1. The main MuJoCo comparison uses $K_{\max} = 5$, $\lambda = 0.7$, and fixed SAC entropy, and matches or exceeds the reported multi-step baselines. The sparse DMControl comparisons use their separately stated configurations; the common-configuration claim is restricted to the main MuJoCo comparison. On Atari 100k, the completed greedy-target comparisons underperform their baselines; partial Boltzmann results do not establish a benchmark-wide gain (Section 5.4).

5.3 Adaptive Truncation and Ablations

Adaptive horizon over training. With the common cap $K_{\max} = 5$, the effective horizon $\hat{K} = \min(K_{\max}, \hat{T}_{\beta,\pi})$ adapts within $[1, 5]$ without per-task cap tuning. The right two panels of Figure 4 track $\hat{K}$ for SAC and TD3 across five MuJoCo tasks. Under SAC, $\hat{K}$ starts at task-dependent values ($\sim$2–4.5) and contracts to $\sim$1.5–2.5 by 1M steps, consistent with increasing agreement between target and stored behavior policies. Under TD3, $\hat{K}$ remains closer to the cap ($\sim$3–5). These measurements show task- and backbone-dependent horizons without estimating τ_{mix}.



(a) Learning curves.

(b) Auto-α: α and $\hat{K}$ vs. steps.

Figure 5: **DMC ablation diagnostics.** (a) Learning curves: three seeds, 95% bootstrap confidence intervals; the horizontal 1-step reference on `Hopper-Stand` is its final return at 1M steps. (b) Entropy coefficient and effective horizon on `CartPole-Swingup-Sparse` (L=10); fixed-α shown for comparison.

Cap-length robustness. Figure 1b, panel (b) of Figure 1, compares `Hopper` learning curves for caps $\{3, 5, 10, 20\}$. Uncorrected n-step and Peng's $Q(\lambda)$ degrade at longer caps, while `MINT` is essentially flat across the tested values on this task. When $\hat{T}_{\beta,\pi} < K_{\max}$, the stopping rule selects the retained prefix before the cap binds. Additional MuJoCo comparisons are provided in Appendix F.

To test whether the cap determines the horizon under stochastic dynamics, we use RWRL `walker_walk` with actuator-noise levels $\sigma \in \{0, 0.1, 0.3\}$, fixed $\alpha = 0.05$, and 200k steps. Table 2 shows that the retained horizon remains well below the expanded cap and most traces stop early. This cap robustness is conditional: smaller caps limit return in the separate auto-α study (Table 19). Detailed diagnostics are in Appendices G, G.2 and G.4.

Table 2: **RWRL cap ablation.** Mean ± std. across three seeds; actuator-noise levels pooled.

$K_{\max}$	K_{eff}	Stops before cap (%)
5	3.9 ± 0.1	36.7 ± 3.3
50	4.8 ± 0.8	99.5 ± 0.7

Truncation kernel ablation (DMControl). We compare SAC-`MINT`, SAC-multistep with a matched maximum horizon and λ but no stochastic trace truncation, and vanilla SAC on five DMControl tasks (three seeds, 1M steps, automatic entropy tuning). SAC-`MINT` outperforms SAC-multistep on `CartPole-Swingup-Sparse` (719 ± 46 vs. 24) and `Hopper-Stand` (891 ± 24 vs. 57 ± 90), while both learn on `Finger-Turn-Easy` (Figure 5a). These are returns averaged over the final 100k steps; the reported spreads are seed standard deviations. The comparison supports adding stochastic truncation under this protocol; matched-random-truncation controls are analyzed separately in Appendix D.5.

In the `CartPole-Swingup-Sparse` entropy diagnostic ($L = 10$), auto-α reaches $\hat{K} \approx 5.7$ and learns successfully, while fixed $\alpha = 0.2$ gives short traces ($\hat{K} \approx 2$) and fails to learn (Figure 5b). Full settings, the five-task results, and entropy diagnostics are in Appendix F.3.

5.4 Evaluation under Stochastic Dynamics

Frozen-critic targets. We evaluate fixed-policy targets on deterministic and stochastic OGBench AntMaze datasets, including `navigate` and `stitch` variants. A fitted Gaussian behavior model defines target-policy displacements h in behavior standard deviations; all operators share a frozen approximate Monte-Carlo critic and finite-rollout reference. At $h = 0.5$, `MINT` has lower RMSE than Retrace on all four datasets (Table 3). Its intrinsic sampling variance is lower than a shuffled-$\hat{K}$ control in seven of the eight reported $h \in \{0.5, 1\}$ cells (Tables 3 and 24). This control preserves the marginal truncation-length distribution while removing state matching. A one-step baseline attains the lowest RMSE in most cells (Appendix H.1).

Learned-critic evaluation. The target-level advantage does not uniformly translate to learned-critic error. In fitted-Q evaluation at 300k gradient steps, Retrace remains more accurate on the stochastic maze at $h = 1$ and $h = 2$. At $h = 2$, `MINT` has two unstable seeds out of five on the stochastic maze, while the shuffled-$\hat{K}$ control has none. These results delimit the benefit of state-dependent truncation under large mismatch and multimodal dynamics. Full protocols, per-cell results, and the instability

criterion are in Appendix H and Table 25; additional transition and empirical backup-gain diagnostics are in Appendices F.2, G.2 and G.4.

Table 3: **OGBench frozen-critic evaluation ($h = 0.5$).** Approximate Monte-Carlo critic; one paired run per cell over 100 stratified evaluation states. Large: deterministic; Teleport: stochastic. Lower is better; bold marks the lower value within each metric pair.

	RMSE ↓		Intrinsic variance ↓	
Dataset	MINT	Retrace	MINT	Shuffled-$\hat{K}$
Large–navigate	**6.62**	7.76	2.03	**1.69**
Large–stitch	**2.99**	8.85	**0.49**	1.01
Teleport–navigate	**3.82**	6.09	**0.33**	1.42
Teleport–stitch	**2.93**	6.79	**1.16**	1.91

Limitations and Assumptions The state-coupled contraction condition of Theorem 4.2 and the action-level surrogate condition of Theorem B.9 are separate sufficient conditions. The surrogate theorem covers fixed-policy expected evaluation on finite state and action spaces; it does not establish convergence with changing policies and nonlinear critics. Under stochastic dynamics the target bias stays below Retrace's at small policy divergence (Appendix H), but at large divergence the state-matched truncation can destabilize a learned critic on multimodal kernels; we state this as a limitation (Appendix J). Atari diagnostics expose target-update sensitivity and performance loss from intermediate $V - Q$ terms; whether Boltzmann averaging mitigates overestimation remains a hypothesis. Full discussion: Appendix J.

6 Related Work

Retrace constrains trace coefficients to guarantee fixed-policy evaluation for arbitrary behavior policies [29]. Harutyunyan's $Q^\pi(\lambda)$ retains unweighted TD-error traces, with contraction requiring additional policy-mismatch conditions [18]; setting $A_i \equiv 1$ in our expected operator recovers it. Peng's $Q(\lambda)$ mixes uncorrected returns with a generally behavior-dependent fixed-policy limit, although it admits control convergence when the behavior policy suitably tracks a greedy policy [23]. MINT retains Q^π as a fixed point, with separate sufficient contraction conditions for its state-coupled operator and practical surrogate under fixed-policy evaluation (Theorems B.9 and 4.2).

Rowland et al. [33] analyze trade-offs among update variance, fixed-point bias, and contraction rate; Daley et al. [7] unify per-decision and trajectory-aware traces under a multistep operator. MINT's practical keep probability $h(s, a) = 1 - \min(1, \pi/\beta)$, the complement of Retrace's coefficient excluding λ, drives continuation by disagreement and yields one-step on-policy backups. Idempotence expresses the state-coupled estimator's variance bound at Q^π through survival (Lemma 3.4), without implying variance dominance over Retrace. Importance Resampling resamples replay transitions [34], whereas MINT truncates the given trajectory.

Mixing informs policy-learning analyses [16, 1, 30, 45], including average-reward methods without oracle mixing-time knowledge [38]. For on-policy multi-step evaluation, Duan and Wainwright [9] relate sufficient look-ahead to temporal dependence and model misspecification (Section 2.2); this does not directly give off-policy guarantees. MINT instead uses policy-mismatch-dependent stopping for discounted TD evaluation, motivated by coupling [27, 25, 20]. Behavior-chain mixing uses a same-kernel coalescent coupling; behavior–target first agreement controls state-coupled trace survival (Appendix B.6). The model-free surrogate defines a separate stopping process without transition-kernel access or mixing-time estimation.

7 Conclusion

MINT treats trace truncation as part of the off-policy estimator, using policy mismatch to select the realized backup horizon within a fixed cap. Binary stopping indicators yield a survival-based variance bound for the state-coupled estimator at Q^π, while the state-coupled operator and practical action-level surrogate admit separate sufficient conditions for contraction under fixed-policy evaluation. The model-free implementation requires neither transition-kernel access nor mixing-time estimation.

Acknowledgments and Disclosure of Funding

This work is in part supported by the National Research Foundation of Korea (NRF, RS-2024-00451435(20%), RS-2024-00413957(10%), RS-2026-25547694(20%)), Institute of Information & communications Technology Planning & Evaluation (IITP, RS-2025-02305453(10%), RS-2025-02273157(10%), RS-2025-25442149(15%) RS-2021-II211343(15%)) grant funded by the Ministry of Science and ICT (MSIT), Institute of New Media and Communications(INMAC), and the BK21 FOUR program of the Education, Artificial Intelligence Graduate School Program (Seoul National University), and Research Program for Future ICT Pioneers, Seoul National University in 2026.

References

[1] Y. Abbasi-Yadkori, P. Bartlett, K. Bhatia, N. Lazic, C. Szepesvari, and G. Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning*, pages 3692–3702. PMLR, 2019.

[2] J. Achiam. Spinning up in deep reinforcement learning, 2018.

[3] R. Agarwal, M. C. Machado, P. S. Castro, and M. G. Bellemare. Deep reinforcement learning at the edge of the statistical precipice. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[4] K. D. Asis and R. S. Sutton. Per-decision multi-step temporal difference learning with control variates. In *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, pages 786–794. AUAI Press, 2018.

[5] D. P. Bertsekas and S. Ioffe. Temporal differences-based policy iteration and applications in neuro-dynamic programming. *Lab. for Info. and Decision Systems Report LIDS-P-2349, MIT, Cambridge, MA*, 14:8, 1996.

[6] P. S. Castro, S. Moitra, C. Gelada, S. Kumar, and M. G. Bellemare. Dopamine: A Research Framework for Deep Reinforcement Learning. *arXiv preprint arXiv:1812.06110*, 2018. URL http://arxiv.org/abs/1812.06110.

[7] B. Daley, M. White, C. Amato, and M. C. Machado. Trajectory-aware eligibility traces for off-policy reinforcement learning. In *International Conference on Machine Learning*, pages 6818–6835. PMLR, 2023.

[8] W. Doeblin. Exposé de la théorie des chaînes simples constantes de Markoff à un nombre fini d'états. *Revue Mathématique de l'Union Interbalkanique*, 2:77–105, 1938.

[9] Y. Duan and M. J. Wainwright. Policy evaluation from a single path: Multi-step methods, mixing and mis-specification. *arXiv preprint arXiv:2211.03899*, 2022. URL https://arxiv.org/abs/2211.03899.

[10] G. Dulac-Arnold, N. Levine, D. J. Mankowitz, J. Li, C. Paduraru, S. Gowal, and T. Hester. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, 110(9):2419–2468, 2021. doi: 10.1007/s10994-021-05961-4.

[11] Farama Foundation. Minari: A dataset interface for offline reinforcement learning. https://github.com/Farama-Foundation/Minari, 2022.

[12] J. Fu, A. Kumar, O. Nachum, G. Tucker, and S. Levine. D4rl: Datasets for deep data-driven reinforcement learning. *arXiv preprint arXiv:2004.07219*, 2020.

[13] S. Fujimoto, H. van Hoof, and D. Meger. Addressing function approximation error in actor-critic methods. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1587–1596, 2018.

[14] M. Geist, B. Scherrer, et al. Off-policy learning with eligibility traces: a survey. *J. Mach. Learn. Res.*, 15(1):289–333, 2014.

[15] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1861–1870. PMLR, 2018.

[16] B. Hao, N. Lazic, Y. Abbasi-Yadkori, P. Joulani, and C. Szepesvari. Provably efficient adaptive approximate policy iteration. *arXiv preprint arXiv:2002.03069*, 2020.

[17] A. Harutyunyan. *Beyond Single-Step Temporal Difference Learning*. PhD thesis, Ph. D. thesis, Vrije Universiteit Brussel (VUB), 2018.

[18] A. Harutyunyan, M. G. Bellemare, T. Stepleton, and R. Munos. Q(λ) with off-policy corrections. In *Proceedings of the 27th International Conference on Algorithmic Learning Theory*, volume 9925 of *Lecture Notes in Computer Science*, pages 305–320. Springer, 2016.

[19] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, and D. Silver. Rainbow: Combining improvements in deep reinforcement learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[20] J. E. Johndrow and J. C. Mattingly. Coupling and decoupling to bound an approximating markov chain. *arXiv preprint arXiv:1706.02040*, 2017.

[21] J. E. Johndrow and J. C. Mattingly. Error bounds for approximations of markov chains used in bayesian sampling. *arXiv preprint arXiv:1711.05382*, 2017.

[22] L. Kaiser, M. Babaeizadeh, P. Milos, B. Osinski, R. H. Campbell, K. Czechowski, D. Erhan, C. Finn, P. Kozakowski, S. Levine, A. Mohiuddin, R. Sepassi, G. Tucker, and H. Michalewski. Model-based reinforcement learning for atari. In *International Conference on Learning Representations*, 2020.

[23] T. Kozuno, Y. Tang, M. Rowland, R. Munos, S. Kapturowski, W. Dabney, M. Valko, and D. Abel. Revisiting peng's q (λ) for modern reinforcement learning. In *International Conference on Machine Learning*, pages 5794–5804. PMLR, 2021.

[24] D. A. Levin and Y. Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[25] T. Lindvall. *Lectures on the Coupling Method*. Wiley, New York, 1992.

[26] A. R. Mahmood, H. Yu, and R. S. Sutton. Multi-step off-policy learning without importance sampling ratios. *arXiv preprint arXiv:1702.03006*, 2017.

[27] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

[28] A. Y. Mitrophanov. Sensitivity and convergence of uniformly ergodic markov chains. *Journal of Applied Probability*, 42(4):1003–1014, 2005.

[29] R. Munos, T. Stepleton, A. Harutyunyan, and M. G. Bellemare. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[30] G. Neu and J. Olkhovskaya. Online learning in mdps with linear function approximation and bandit feedback. *Advances in Neural Information Processing Systems*, 34:10407–10417, 2021.

[31] S. Park, K. Frans, B. Eysenbach, and S. Levine. OGBench: Benchmarking offline goal-conditioned RL. In *International Conference on Learning Representations*, 2025. arXiv:2410.20092.

[32] D. Precup, R. S. Sutton, and S. Dasgupta. Off-policy temporal-difference learning with function approximation. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 417–424. Morgan Kaufmann, 2001.

[33] M. Rowland, W. Dabney, and R. Munos. Adaptive trade-offs in off-policy learning. In *International Conference on Artificial Intelligence and Statistics*, pages 34–44. PMLR, 2020.

[34] M. Schlegel, W. Chung, D. Graves, J. Qian, and M. White. Importance resampling for off-policy prediction. *Advances in Neural Information Processing Systems*, 32, 2019.

[35] M. Schwarzer, A. Anand, R. Goel, R. D. Hjelm, A. Courville, and P. Bachman. Data-efficient reinforcement learning with self-predictive representations. In *International Conference on Learning Representations*, 2021.

[36] M. Schwarzer, J. S. O. Ceron, A. Courville, M. G. Bellemare, R. Agarwal, and P. S. Castro. Bigger, better, faster: Human-level Atari with human-level efficiency. In *Proceedings of the 40th International Conference on Machine Learning*, 2023.

[37] E. Solan and N. Vieille. Perturbed markov chains. *Journal of applied probability*, 40(1): 107–122, 2003.

[38] W. A. Suttle, A. Bedi, B. Patel, B. M. Sadler, A. Koppel, and D. Manocha. Beyond exponentially fast mixing in average-reward reinforcement learning via multi-level monte carlo actor-critic. In *International Conference on Machine Learning*, pages 33240–33267. PMLR, 2023.

[39] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[40] R. S. Sutton, D. Precup, and S. Singh. Intra-option learning about temporally abstract actions. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pages 556–564. Morgan Kaufmann, 1998.

[41] Y. Tassa, Y. Doron, A. Muldal, T. Erez, Y. Li, D. de Las Casas, D. Budden, A. Abdolmaleki, J. Merel, A. Lefrancq, T. Lillicrap, and M. Riedmiller. Deepmind control suite, 2018.

[42] E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.

[43] H. van Hasselt. Double q-learning. *Advances in Neural Information Processing Systems*, 23, 2010.

[44] H. van Hasselt, M. Hessel, and J. Aslanides. When to use parametric models in reinforcement learning? In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[45] C.-Y. Wei, M. J. Jahromi, H. Luo, and R. Jain. Learning infinite-horizon average-reward mdps with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 3007–3015. PMLR, 2021.

[46] G. Wolfer and A. Kontorovich. Estimating the mixing time of ergodic markov chains. In *Conference on Learning Theory*, pages 3120–3159. PMLR, 2019.

[47] D. Yarats, I. Kostrikov, and R. Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In *International Conference on Learning Representations*, 2021.

[48] W. Ye, S. Liu, T. Kurutach, P. Abbeel, and Y. Gao. Mastering Atari games with limited data. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

A Notation and Background

This appendix collects the standard definitions used throughout the paper: total variation (TV) distance, the coupling lemma, maximal coupling, and uniform ergodicity. These are the analytic primitives behind the meeting-time analysis of Lemma B.2 and the contraction analysis of Theorem 4.2.

A.1 Total Variation Distance

For any two probability distributions ν_1 and ν_2 over $\mathcal{S}$, the total variation (TV) distance $\|\cdot\|_{\mathrm{TV}}$ is

$$\|\nu_1 - \nu_2\|_{\mathrm{TV}} \coloneqq \frac{1}{2}\sum_{s\in\mathcal{S}} |\nu_1(s) - \nu_2(s)| = \max_{A\subset\mathcal{S}} |\nu_1(A) - \nu_2(A)|,$$

where the norm $\|\cdot\|$ corresponds to the L_1 metric. By definition, the TV distance takes values in $[0, 1]$.

A.2 Coupling Lemma and Maximal Coupling

Lemma A.1 (Coupling Lemma)**.** *Let ν_1 and ν_2 be two probability distributions over a finite space $\mathcal{S}$. For every coupling (X, Y) of ν_1 and ν_2,*

$$\mathrm{P}(X \neq Y) \geq \|\nu_1 - \nu_2\|_{\mathrm{TV}}.$$

There exists a coupling attaining equality, called a maximal coupling:

$$\mathrm{P}(X \neq Y) = \|\nu_1 - \nu_2\|_{\mathrm{TV}} = 1 - \sum_{s\in\mathcal{S}} \min(\nu_1(s), \nu_2(s)). \tag{15}$$

Maximal coupling minimizes the probability of disagreement $\mathrm{P}(X \neq Y)$ among all possible couplings of ν_1 and ν_2. Throughout the paper we use the construction Equation (15) to relate the meeting time of two policy-induced chains to TV distances between their transition kernels.

A.3 Uniform Ergodicity

We say the MDP is *uniformly ergodic* with parameter α_{erg} if there exist a constant $\alpha_{\mathrm{erg}} \in (0, 1)$ and $C_{\mathrm{erg}} > 0$ such that

$$\max_{s\in\mathcal{S}} \left\| \mathcal{P}^{\beta(t)}(s, \cdot) - \mu_\beta \right\|_{\mathrm{TV}} \leq C_{\mathrm{erg}}(1 - \alpha_{\mathrm{erg}})^t$$

for all $t \in \mathbb{N}$. Uniform ergodicity is a standard mild assumption in the analysis of perturbed Markov chains [28, 37, 21].

A.4 Data-Processing Inequality for Total Variation

Lemma (TV contraction under Markov kernel). Let P and Q be two probability measures over a measurable space $(\mathcal{X}, \mathcal{F})$, and let $f : \mathcal{X} \to \mathcal{Y}$ be a stochastic map, i.e., a Markov kernel such that $f(\cdot \mid x)$ is a probability distribution over $\mathcal{Y}$ for each $x \in \mathcal{X}$. Then the push-forward measures P_Y, Q_Y on $\mathcal{Y}$ defined by

$$P_Y(B) = \int_{\mathcal{X}} f(B \mid x)\, dP(x), \quad Q_Y(B) = \int_{\mathcal{X}} f(B \mid x)\, dQ(x)$$

satisfy

$$\mathrm{TV}(P_Y, Q_Y) \leq \mathrm{TV}(P, Q).$$

Proof. Let $\mu = P - Q$, so $\mu(\mathcal{X}) = 0$. For any measurable $B \subseteq \mathcal{Y}$, the function $g_B(x) = f(B \mid x)$ lies in $[0, 1]$. The variational form $\sup_{0\leq g\leq 1} |\int g\, d\mu| = \mathrm{TV}(P, Q)$ therefore gives

$$\mathrm{TV}(P_Y, Q_Y) = \sup_B \left| \int g_B\, d(P - Q) \right| \leq \mathrm{TV}(P, Q).$$

Interpretation. Applying a stochastic transformation can only reduce, not increase, the distinguishability of two distributions. In RL, this explains why the discrepancy between next-state distributions under two policies is bounded by their action-level difference:

$$\mathrm{TV}\left(\sum_a \pi(a \mid s)\mathcal{P}(\cdot \mid s,a),\ \sum_a \beta(a \mid s)\mathcal{P}(\cdot \mid s,a)\right) \le \mathrm{TV}(\pi(\cdot \mid s), \beta(\cdot \mid s)).$$

This principle underlies the use of action-level disagreement probabilities $\hat{p}_i$ in `MINT` as a proxy for policy divergence propagated through the dynamics.

A.5 Standing Assumptions

The proofs in this appendix rely on the following standing assumptions, recalled here for reference. Assumption A.3 is restated in the main text (Section 4); we add here the kernel-level TV bound used by several of the supporting lemmas.

Assumption A.2 (Kernel-level TV bound)**.** For the fixed policy pair, define the kernel discrepancy by

$$d := \sup_{s \in \mathcal{S}} d(s), \qquad d(s) := \|\mathcal{P}^{\beta}(s,\cdot) - \mathcal{P}^{\pi}(s,\cdot)\|_{\mathrm{TV}}.$$

Thus $d \in [0,1]$ and data processing gives $d \le D := \sup_s \mathrm{TV}(\beta(\cdot \mid s), \pi(\cdot \mid s))$. This comparison concerns the actual supremum defined here; an arbitrary looser numerical upper bound on d need not itself be at most D.

Assumption A.3 (Cross-Doeblin, restatement of Assumption A.3)**.** There exists $\rho \in (0, 1-d)$ such that for every pair of states $(s,s') \in \mathcal{S} \times \mathcal{S}$ and every policy pair (β, π),

$$\left\|\mathcal{P}^{\beta}(s,\cdot) - \mathcal{P}^{\pi}(s',\cdot)\right\|_{\mathrm{TV}} \le 1-\rho.$$

The cross-Doeblin condition requires the existence of a uniform overlap $\rho \in (0, 1-d)$ such that $\|\mathcal{P}^{\beta}(s,\cdot) - \mathcal{P}^{\pi}(s',\cdot)\|_{\mathrm{TV}} \le 1-\rho$ for all state pairs and all policies. Combined with the maximal coupling Equation (15), it ensures that conditional on disagreement at step t, the coupled processes meet at step $t+1$ with probability at least ρ, which is what produces the geometric meeting-time tail used in Corollary B.1.

Where each assumption is used. Table 4 distinguishes the state-coupled analysis from the finite-state, fixed-policy surrogate theorem. The latter uses no deterministic-dynamics or cross-Doeblin hypothesis, but requires its own χ_λ sufficient condition. The one-step DPI proposition is not a transfer theorem for contraction.

Table 4: Assumption usage by result. A check mark means the hypothesis is invoked in the stated analysis. State-coupled and action-level surrogate results have separate sufficient conditions.

Result	finite $\mathcal{S}$	deterministic $\mathcal{P}$	cross-Doeblin (Assumption A.3)	kernel TV d (Assumption A.2)	other
Lemma 3.4 (variance)	–	–	–	–	adapted joint law, bounded TD error
Theorem 4.2 (contraction)	–	–	–	–	state coupling, $\kappa < 1$ in Equation (13)
Corollary B.1 (closed form)	–	–	✓	✓	–
Lemma B.2 (disagreement)	–	–	✓	✓	–
Theorem B.7 (state horizon)	–	–	✓	–	initial-step overlap for $1/\rho$ bound
Lemma B.5 (perturbation)	–	–	–	✓	Assumption B.3, stationary laws; explicit $b_\pi(t)$
Proposition B.6 (behavior mixing)	–	–	–	–	same-kernel coalescent coupling, stationary law
Corollary C.3 (averaged error)	–	–	–	–	pointwise bound, any starting distribution
Theorem B.9 (surrogate)	✓	–	–	–	fixed policies, independent masks, $\kappa_{\chi,L} < 1$
Theorem B.8 (online control)	✓	–	–	–	Markov traces, (A1)–(A5) of Appendix B.7
Proposition A.4 (one-step link)	–	–	–	–	data processing

A.6 Cross-Doeblin and Two TV Quantities

Two TV quantities, two roles. By the data-processing inequality (Appendix A.4), the kernel-level TV bound d from Assumption A.2 satisfies $d \le D := \sup_s D_s$. Both quantities measure off-policy discrepancy but at different "levels":

- d enters the cross-Doeblin assumption $\rho \in (0, 1-d)$, controlling the kernel-level meeting rate.

- $\bar{D}_{\mathrm{wgt}}$ enters the contraction condition Equation (24), controlling the per-step weight deviation in the operator analysis.

D_s is action-level policy mismatch, while $\bar{D}_{\mathrm{wgt}}$ weights it by the state-coupled survival process. The implemented surrogate has its own survival process; its contraction correction weights $\chi_\lambda(s)$ instead (Theorem B.9). These weighted quantities need not coincide.

Justification of the action-level approximation. For transitions from a shared state s, define $d(s) = \mathrm{TV}(\mathcal{P}^\beta(\cdot \mid s), \mathcal{P}^\pi(\cdot \mid s))$. Data processing bounds $d(s)$ by D_s. The implementation samples an action-dependent keep probability whose mean is D_s, as follows; this does not identify it with the disagreement probability of a coupled pair at different states.

(i) Always valid. Since $\min\{1, \pi/\beta\} \in [0, 1]$, we have $\hat{p}_i \in [0, 1]$ regardless of the relative magnitudes of $\pi(a_i \mid s_i)$ and $\beta(a_i \mid s_i)$, including the unbounded-ratio regime where $\pi/\beta > 1$.

(ii) Unbiased estimator of the action-level policy TV. Taking the expectation over $a_i \sim \beta(\cdot \mid s_i)$ and using the variational identity for total variation,

$$\mathbb{E}_{a\sim\beta}\left[\min\left\{1, \frac{\pi(a \mid s)}{\beta(a \mid s)}\right\}\right] \;=\; \sum_a \min\{\beta(a \mid s),\, \pi(a \mid s)\} \;=\; 1 - \mathrm{TV}\big(\beta(\cdot \mid s), \pi(\cdot \mid s)\big), \tag{16}$$

$$\mathbb{E}_{a\sim\beta}[\hat{p}_i \mid s_i] \;=\; \mathrm{TV}\big(\beta(\cdot \mid s_i), \pi(\cdot \mid s_i)\big) \;=\; D_{s_i}. \tag{17}$$

Thus, $\hat{p}_i$ is an unbiased sample-level estimate of D_{s_i}. Its conditional mean, rather than each realized sample, upper-bounds $d(s_i)$, the one-step kernel TV from a shared state.

(iii) Fully model-free. $\hat{p}_i$ depends only on the policy ratio at the sampled action and never queries the environment transition kernel, so it is computable in any actor-critic implementation (cf. Algorithm 1).

(iv) One-step data processing under arbitrary dynamics.

Proposition A.4 (One-step surrogate and kernel TV)**.** *For an arbitrary transition kernel, let* $d(s) := \mathrm{TV}(\mathcal{P}^\beta(\cdot \mid s), \mathcal{P}^\pi(\cdot \mid s))$. *Then*

$$\mathbb{E}_{a\sim\beta(\cdot|s)}[\hat{p}(s, a)] = D_s \geq d(s).$$

If the next states are maximally coupled from a shared state s, their one-step disagreement probability is $d(s)$. Equality $D_s = d(s)$ holds, in particular, for deterministic dynamics with an injective action-to-next-state map.

Proof. The expectation identity is proved in (ii). Both next-state laws are the push-forwards of the two action laws through the common kernel $a \mapsto \mathcal{P}(\cdot \mid s, a)$, so data processing gives $d(s) \leq D_s$. Maximal coupling identifies $d(s)$ with one-step disagreement from that shared state. An injective deterministic map preserves TV, giving the stated sufficient equality case. □

This proposition is an expectation-level, one-step statement. It neither gives a samplewise upper bound on kernel disagreement nor orders the multi-step stopping times once the two state processes differ. The action-dependent surrogate contraction theorem is Theorem B.9; it uses the surrogate's own survival quantities and the negative-mass fraction χ_λ.

B Proofs

B.1 Proof of the Variance Lemma (Lemma 3.4)

Joint sampling and filtration. The behavior action is drawn from $\beta(\cdot \mid s_t)$ conditionally on the pre-action history $\mathcal{H}_t$. The enlarged history $\mathcal{G}_t = \mathcal{H}_t \vee \sigma(a_t)$ contains both coupled state histories, all past rewards and actions, and coupling randomness already revealed. It contains no information from the next reward or transition beyond that allowed by the environment law. We require

$$\mathcal{L}(r_t, s_{t+1} \mid \mathcal{G}_t) = \mathcal{L}(r_t, s_{t+1} \mid s_t, a_t).$$

This compatibility is part of the construction, not a consequence of marginal coupling alone. For example, let $\Gamma_t(u, v)$ be a maximal coupling of $\mathcal{P}^\beta(s_t, \cdot)$ and $\mathcal{P}^\pi(S_t^\pi, \cdot)$, given $\mathcal{H}_t$. First draw $a_t \sim \beta(\cdot \mid s_t)$ and the behavior reward and next state u from the environment. Then draw the auxiliary next state v from $\Gamma_t(u, v)/\mathcal{P}^\beta(s_t, u)$ on positive-denominator rows. Averaging over a_t recovers Γ_t and both policy-induced state marginals; the behavior transition law given $\mathcal{G}_t$ is unchanged. Zero-denominator rows can be assigned any conditional law, since they have zero probability under a policy-generated step. If the initial pair x is prescribed, use the induced conditional initial coupling, with a valid behavior transition on any null-support extension. All $\Pr_x(T > t)$ refer to that specified initialization.

Martingale and fixed point. Let $\delta_t = \delta_t^{Q^\pi}$, $W_t = \prod_{i=1}^t A_i = \mathbf{1}\{T > t\}$, and $\alpha = \gamma\lambda$. The Bellman identity and the conditional environment law give

$$\mathbb{E}_x[\delta_t \mid \mathcal{G}_t] = (\mathcal{T}^\pi Q^\pi)(s_t, a_t) - Q^\pi(s_t, a_t) = 0.$$

The mask W_t is $\mathcal{G}_t$-measurable, hence $\mathbb{E}_x[W_t\delta_t] = 0$. The deterministic bound $|\delta_t| \le C := \|r\|_\infty + (1+\gamma)\|Q^\pi\|_\infty$ and $\alpha < 1$ justify exchanging the sum and expectation. Thus $\mathbb{E}_x[\hat{\Delta}] = 0$ and $\mathcal{R}_{p,\lambda}Q^\pi = Q^\pi$.

Cross terms and idempotence. For $u < t$, W_u, W_t, δ_u are $\mathcal{G}_t$-measurable. Therefore

$$\mathbb{E}_x[W_u W_t \delta_u \delta_t] = \mathbb{E}_x[W_u W_t \delta_u \mathbb{E}_x[\delta_t \mid \mathcal{G}_t]] = 0.$$

This orthogonality applies to other adapted weights as well. For the binary masks, $W_t^2 = W_t$ additionally yields

$$\begin{aligned} \mathrm{Var}_x(\hat{\Delta}) = \mathbb{E}_x[\hat{\Delta}^2] &= \sum_{t\ge 0} \alpha^{2t}\mathbb{E}_x[W_t\delta_t^2] \\ &\le C^2 \sum_{t\ge 0} (\gamma^2\lambda^2)^t \Pr_x(T > t) = C^2 M_x(\gamma^2, \lambda^2). \end{aligned} \tag{18}$$

One can first expand finite sums and then pass to the limit: $|\hat{\Delta}| \le C/(1-\alpha)$ gives dominated convergence. Averaging over a specified starting distribution gives the corresponding $C^2 M(\gamma^2, \lambda^2)$ bound. No claim of universal variance dominance over clipped trace methods follows from this calculation.

B.2 Proof of Theorem 4.2

The TV average in Definition 4.1 is discounted and survival-weighted for each starting pair; the supremum is over starting pairs only, not over individual times. By definition,

$$N_x = D^x_{\mathrm{wgt}}(M_x - 1) \le \bar{D}_{\mathrm{wgt}}(M_x - 1), \qquad \bar{D}_{\mathrm{wgt}} \le \sup_s D_s =: D \le 1.$$

Fix a starting pair x and use the joint sampling conditions of Section 3. Set $W_t = \prod_{i=1}^t A_i$, $W_0 = 1$. The arrival-state mask A_t, W_{t-1}, and s_t are known before a_t is sampled; thus they are $\mathcal{H}_t$-measurable. The conditional action law is $\beta(\cdot \mid s_t)$. We retain A_t in this calculation instead of replacing it by a conditional mean taken with respect to a different filtration.

Signed coefficients. For $q = Q - Q'$, expanding the TD-error sum and telescoping gives

$$(\mathcal{R}_{p,\lambda}Q - \mathcal{R}_{p,\lambda}Q')(x) = \sum_{t\ge 1} \gamma^t\lambda^{t-1}\mathbb{E}_x\left[W_{t-1}\sum_b \big(\pi(b \mid s_t) - \lambda A_t \beta(b \mid s_t)\big) q(s_t, b)\right].$$

The conditional environment law justifies the shift of the bootstrap term to the next visited state; the pre-action conditioning justifies averaging $q(s_t, a_t)$ using β. Let $w_{x;y,b}$ be the resulting signed coefficients and define

$$b_x := \sum_{y,b} |w_{x;y,b}|. \tag{19}$$

For any $a \in \{0, 1\}$,

$$\begin{aligned}\sum_b |\pi(b \mid s) - \lambda a \beta(b \mid s)| &= 1 - \lambda a + 2 \sum_b [\lambda a \beta(b \mid s) - \pi(b \mid s)]_+ \\ &\leq 1 - \lambda a + 2\lambda a D_s.\end{aligned} \tag{20}$$

Indeed, $\lambda a \leq 1$ and $[\lambda a \beta - \pi]_+ \leq \lambda a [\beta - \pi]_+$, whose action sum is $\lambda a D_s$. Consequently,

$$b_x \leq M_{1,x} + 2\Delta_{\lambda,x}, \tag{21}$$

where

$$M_{1,x} := \sum_{t \geq 1} \gamma^t \lambda^{t-1} \mathbb{E}_x[W_{t-1}(1 - \lambda A_t)], \qquad \Delta_{\lambda,x} := \sum_{t \geq 1} (\gamma\lambda)^t \mathbb{E}_x[W_t D_{s_t}] = N_x.$$

The survival identity $W_t = W_{t-1} A_t$ gives

$$M_{1,x} = \gamma M_x - (M_x - 1) = 1 - (1 - \gamma) M_x.$$

Combining the two terms proves

$$b_x \leq \kappa_x := 1 - (1 - \gamma) M_x + 2N_x \leq 1 - (1 - \gamma) M_x + 2\bar{D}_{\text{wgt}}(M_x - 1). \tag{22}$$

At $\lambda = 0$ the same formula follows directly from $\mathcal{R}_{p,0} = \mathcal{T}^\pi$; equivalently, use the $t = 1$ term in the telescoping expression. Consequently, for all bounded Q, Q' and every starting pair x,

$$|\mathcal{R}_{p,\lambda} Q(x) - \mathcal{R}_{p,\lambda} Q'(x)| \leq \kappa_x \|Q - Q'\|_\infty. \tag{23}$$

Taking the supremum over x proves the claimed Lipschitz bound. The fixed-point identity of Lemma 3.4 gives uniqueness and convergence when this supremum is below one.

Since $1 \leq M_x \leq (1 - \gamma\lambda)^{-1}$ uniformly,

$$\sup_x \{1 - (1 - \gamma) M_x + 2N_x\} < 1 \quad \Longleftrightarrow \quad \sup_x \frac{2N_x}{M_x} < 1 - \gamma. \tag{24}$$

The uniform bounds on M_x preserve a strictly positive margin in both directions. Substituting $N_x \leq \bar{D}_{\text{wgt}}(M_x - 1)$ gives the sufficient condition

$$\sup_x \left\{ 2\bar{D}_{\text{wgt}} \frac{M_x - 1}{M_x} \right\} < 1 - \gamma. \tag{25}$$

This uses the survival-weighted identity above without requiring a bound on the TV average at each individual time. A value of M averaged over an initial distribution cannot replace the supremum in Equation (13). This proof concerns the state-coupled operator; replacing p_i by $\hat{p}_i$ is not used. The action-dependent surrogate requires its own negative-weight bound, developed in Appendix B.8.

Connection to the λ-dependent effective horizon. For every starting pair,

$$\begin{aligned}M_x &\leq K_\lambda^x := \min\left(\mathbb{E}_x[T], \frac{1}{1 - \gamma\lambda} \right), \\ K_\lambda &:= \sup_x K_\lambda^x \leq \frac{1}{1 - \gamma\lambda}.\end{aligned} \tag{26}$$

This follows from the pointwise inequality $\sum_{t=0}^{T-1} (\gamma\lambda)^t \leq \min(T, (1 - \gamma\lambda)^{-1})$. Since $(u - 1)/u$ is increasing for $u \geq 1$, a uniform sufficient condition is

$$2\bar{D}_{\text{wgt}} \frac{K_\lambda - 1}{K_\lambda} < 1 - \gamma. \tag{27}$$

The scalar K_λ here is a uniform bound over starting pairs, not an initial-distribution average. When the survival law is the same for all starting pairs, the formulas reduce to the corresponding scalar expressions. □

B.3 Closed-Form Corollary and Its Proof

Corollary B.1 (Geometric meeting time — closed-form contraction)**.** *Suppose Assumption A.3 holds with parameter $\rho \in (0, 1-d)$. Also require $\Pr_x(A_1 = 0) \geq \rho$ uniformly over the prescribed starting pairs. This initial-step condition is needed when conditioning on the behavior's initial action; it is not implied by overlap of policy-averaged kernels alone. Then, for every starting pair x, $\Pr_x(T_{\beta,\pi} > t) \leq (1-\rho)^t$, so $M_x \leq [1-\gamma\lambda(1-\rho)]^{-1}$ uniformly. The uniform TV-based sufficient condition of Theorem 4.2 is ensured by the single-product condition*

$$\gamma\big(1 + 2\bar{D}_{\mathrm{wgt}}\,\lambda(1-\rho)\big) \;<\; 1.$$

None of the factors $\bar{D}_{\mathrm{wgt}}$, λ, $(1-\rho)$ needs to be small individually; only their product must be controlled.

Assumption A.3 asserts $\|\mathcal{P}^{\beta}(s,\cdot) - \mathcal{P}^{\pi}(s',\cdot)\|_{\mathrm{TV}} \leq 1-\rho$ uniformly over state pairs. Conditional maximal coupling gives a meeting probability of at least ρ at subsequent policy-generated transitions. Combining this with the explicit initial-step condition and iterating gives $\Pr_x(T_{\beta,\pi} > t) \leq (1-\rho)^t$ for every x. Geometric summation gives

$$M_x \;\leq\; \sum_{t\geq 0}(\gamma\lambda(1-\rho))^t \;=\; \frac{1}{1-\gamma\lambda(1-\rho)}.$$

Writing $U = [1-\gamma\lambda(1-\rho)]^{-1}$, we have $(M_x - 1)/M_x \leq (U-1)/U = \gamma\lambda(1-\rho)$ uniformly in x. Substitution into Equation (25) gives the sufficient condition

$$\gamma\big(1 + 2\bar{D}_{\mathrm{wgt}}\lambda(1-\rho)\big) \;<\; 1.$$

□

B.4 Time-averaged disagreement bound

Lemma B.2 (Time-averaged disagreement; nearby chains)**.** *Consider a co-adapted coupling of the β and π state processes, with conditional maximal coupling at each policy-generated transition. Write $\mathcal{F}_k$ for the joint history through time k before the next transition, $A_k = \mathbf{1}\{S_k^{\beta} \neq S_k^{\pi}\}$, and $u_k = \mathbb{E}[A_k]$. Under Assumptions A.2 and A.3, set $\eta = \rho + d \in (0,1)$. For any initial joint law at time 1 and every integer $t \geq 1$,*

$$\frac{1}{t}\sum_{k=1}^{t} u_k \leq \frac{d}{\eta} + \frac{1-(1-\eta)^t}{\eta t}\left(u_1 - \frac{d}{\eta}\right). \tag{28}$$

Proof. Conditional maximal coupling and the same-state kernel bound give

$$\Pr(A_{k+1} = 1 \mid \mathcal{F}_k) \leq d(1-A_k) + (1-\rho)A_k.$$

Indeed, on $A_k = 0$ both kernels start at the same state, whereas on $A_k = 1$ the cross-state overlap is at least ρ. Taking expectations yields the scalar recursion

$$u_{k+1} \leq d + (1-\rho-d)u_k. \tag{29}$$

Let $r = 1-\eta \in (0,1)$ and $q_* = d/\eta$, so $d + rq_* = q_*$. Subtracting q_* and iterating the inequality, using $r \geq 0$, gives

$$u_k - q_* \leq r^{k-1}(u_1 - q_*).$$

This remains valid when $u_1 < q_*$. Summing for $k = 1, \ldots, t$ and using $\sum_{j=0}^{t-1} r^j = (1-r^t)/\eta$ proves Equation (28). □

This recursion gives the expectation bound associated with the nearby-chain analysis of Johndrow and Mattingly [20, Theorem 9], without a Poisson-equation construction. Its implication is

$$\limsup_{t\to\infty} \frac{1}{t}\sum_{k=1}^{t}\mathbb{E}[A_k] \leq \frac{d}{\rho+d}.$$

The correction in the displayed upper bound has magnitude at most $|u_1 - q_*|/(\eta t)$, hence is $O(t^{-1})$ for fixed (d,ρ). This does not assert convergence of the actual average to q_*, or monotone decay: the upper bound increases toward q_* if $u_1 < q_*$. The first-agreement survival bound is proved separately in Theorem B.7; it is not a consequence of the Cesaro bound.

B.5 Perturbed Markov chain bound

Assumption B.3 (One-step Doeblin condition for the behavior kernel)**.** For the fixed behavior kernel, there exists $a_\beta \in (0,1)$ such that

$$\delta(\mathcal{P}^\beta) := \sup_{s,s'} \|\mathcal{P}^\beta(s,\cdot) - \mathcal{P}^\beta(s',\cdot)\|_{\mathrm{TV}} \le 1 - a_\beta. \tag{30}$$

This is the one-step Doeblin condition in Johndrow and Mattingly [20, Assumption 1]. It implies

$$\|\nu\mathcal{P}^\beta - \nu'\mathcal{P}^\beta\|_{\mathrm{TV}} \le (1-a_\beta)\|\nu - \nu'\|_{\mathrm{TV}}$$

for probability laws ν, ν'. It is stronger than merely specifying an exponential envelope with arbitrary C_{erg} as in Appendix A; a_β is not identified with that envelope's α_{erg}. The assumption is local to the perturbation results below and is not an additional hypothesis of Theorem 4.2.

Lemma B.4 (Stationary perturbation bound)**.** *Under Assumptions A.2 and B.3, let μ_β and μ_π be stationary laws of $\mathcal{P}^\beta$ and $\mathcal{P}^\pi$, respectively. Then*

$$\|\mu_\beta - \mu_\pi\|_{\mathrm{TV}} \le \min\{1, d/a_\beta\}.$$

Proof. Stationarity, the triangle inequality, and the one-step contraction give

$$\begin{aligned}\|\mu_\beta - \mu_\pi\|_{\mathrm{TV}} &\le \|(\mu_\beta - \mu_\pi)\mathcal{P}^\beta\|_{\mathrm{TV}} + \|\mu_\pi(\mathcal{P}^\beta - \mathcal{P}^\pi)\|_{\mathrm{TV}} \\ &\le (1-a_\beta)\|\mu_\beta - \mu_\pi\|_{\mathrm{TV}} + d.\end{aligned}$$

Rearrange and use that TV is at most one. No mixing rate for π is needed for this stationary comparison, provided its stationary law exists. □

Lemma B.5 (Perturbed Markov chain TV bound)**.** *Under the hypotheses of Lemma B.4, define $b_\pi(t) = \sup_s \|(\mathcal{P}^\pi)^t(s,\cdot) - \mu_\pi\|_{\mathrm{TV}}$. For every initial state s and integer $t \ge 0$,*

$$\begin{aligned}\|(\mathcal{P}^\beta)^t(s,\cdot) - \mu_\beta\|_{\mathrm{TV}} \le \|(\mathcal{P}^\beta)^t(s,\cdot) - (\mathcal{P}^\pi)^t(s,\cdot)\|_{\mathrm{TV}} \\ + b_\pi(t) + \min\{1, d/a_\beta\}.\end{aligned} \tag{31}$$

If, additionally, π satisfies its own uniform-ergodicity bound $b_\pi(t) \le C_\pi(1-a_\pi)^t$, this envelope may replace $b_\pi(t)$ on the right.

Proof. Insert $(\mathcal{P}^\pi)^t(s,\cdot)$ and μ_π between $(\mathcal{P}^\beta)^t(s,\cdot)$ and μ_β, apply the triangle inequality, and use Lemma B.4. □

The target-chain mixing term is explicit: behavior-chain ergodicity alone does not supply the same rate for π. For example, if $d < a_\beta/2$, the triangle inequality implies

$$\delta(\mathcal{P}^\pi) \le \delta(\mathcal{P}^\beta) + 2d \le 1 - (a_\beta - 2d),$$

so one may take $C_\pi = 1$ and $a_\pi = a_\beta - 2d$. Alternatively, the all-policy cross-Doeblin condition of Assumption A.3, applied separately to (β,β) and (π,π), supplies one-step constants $a_\beta = a_\pi = \rho$ when that stronger condition is imposed.

For a coupling of the two different kernels, the first term in Equation (31) is at most $\Pr(S_t^\beta \ne S_t^\pi)$. This is a time-$t$ disagreement probability, not generally bounded above by $\Pr(T_{\beta,\pi} > t)$: the chains can separate after their first agreement. Thus this perturbation decomposition does not by itself turn the MINT stopping time into a behavior-chain mixing-time bound.

B.6 Behavior-chain mixing and off-policy truncation: two couplings

Behavior–behavior coupling for estimating $\tau_{\mathrm{mix}}^\beta$. Let (X_t, Y_t) be a co-adapted coupling of two copies of $\mathcal{P}^\beta$, started at arbitrary states x, y. Choose the coupling to be coalescent: once the states agree, use the same transition for both copies thereafter. Its coalescence time is

$$\tau_c^\beta := \inf\{t \ge 0 : X_t = Y_t\}.$$

Unlike an off-policy coupling, both marginals here have the same kernel. Fix such a coupling for each initial pair and set $m_\beta = \sup_{x,y} \mathbb{E}_{x,y}[\tau_c^\beta]$.

Proposition B.6 (Behavior-chain mixing bound)**.** *Suppose $\mathcal{P}^\beta$ has stationary law μ_β. For the coalescent couplings above and every integer $t \geq 0$,*

$$\sup_x \|(\mathcal{P}^\beta)^t(x,\cdot) - \mu_\beta\|_{\mathrm{TV}} \leq \sup_{x,y} \Pr_{x,y}(\tau_c^\beta > t). \tag{32}$$

If $m_\beta < \infty$, then

$$\tau_{\mathrm{mix}}^\beta \leq \lceil 4m_\beta \rceil. \tag{33}$$

Proof. Coalescence implies $\{X_t \neq Y_t\} = \{\tau_c^\beta > t\}$. The coupling inequality bounds $\|(\mathcal{P}^\beta)^t(x,\cdot) - (\mathcal{P}^\beta)^t(y,\cdot)\|_{\mathrm{TV}}$ by this tail. Integrate the second initial state against μ_β and use stationarity to obtain Equation (32). For $t \geq 1$, Markov's inequality bounds its right-hand side by m_β/t. Taking $t = \lceil 4m_\beta \rceil$ proves Equation (33); if $m_\beta = 0$, all initial states already agree and the claim is immediate. □

This is the same-kernel coupling route to mixing-time bounds [24, Section 5.2]. It preserves the purpose of using two chains to bound the behavior policy's mixing time. It is an analytic upper bound; estimating the supremum or its tail from trajectories requires additional statistical control, and is not supplied by an observed MINT stopping time.

Behavior–target coupling for the MINT survival horizon. Return to the distinct-kernel processes of Definition 3.2, with $T_{\beta,\pi} = \inf\{t \geq 1 : S_t^\beta = S_t^\pi\}$ and starting pair $x = (s_0, a_0)$. Here the cumulative mask is killed at first agreement, while the state processes may later separate. This stopping time controls the state-coupled trace and is distinct from τ_c^β.

Theorem B.7 (Expected state-coupled truncation horizon)**.** *Suppose the conditional meeting probability at each transition after time 1, on survival, is at least $\rho > 0$; conditional maximal coupling under Assumption A.3 ensures this. Write $u_{1,x} = \Pr_x(T_{\beta,\pi} > 1)$. Then, for every integer $t \geq 1$,*

$$\Pr_x(T_{\beta,\pi} > t) \leq u_{1,x}(1-\rho)^{t-1}, \qquad \mathbb{E}_x[T_{\beta,\pi}] \leq 1 + \frac{u_{1,x}}{\rho}. \tag{34}$$

If the initial conditional coupling also satisfies $u_{1,x} \leq 1 - \rho$, these simplify to

$$\Pr_x(T_{\beta,\pi} > t) \leq (1-\rho)^t \quad (t \geq 0), \qquad \mathbb{E}_x[T_{\beta,\pi}] \leq \rho^{-1}.$$

For a deterministic integer cap $L \geq 1$, define $K_{\mathrm{state}} = \min\{L, T_{\beta,\pi}\}$. Under this additional initial-step condition,

$$\mathbb{E}_x[K_{\mathrm{state}}] \leq \min\{L, \rho^{-1}\}, \qquad M_x(\gamma,\lambda) \leq \frac{1}{1-\gamma\lambda(1-\rho)}.$$

Proof. For $t \geq 1$, $\{T_{\beta,\pi} > t\}$ is measurable with respect to the joint history through time t. The conditional meeting bound gives

$$\Pr_x(T_{\beta,\pi} > t+1) = \mathbb{E}_x[\mathbf{1}\{T_{\beta,\pi} > t\} \Pr(A_{t+1} = 1 \mid \mathcal{F}_t)] \leq (1-\rho) \Pr_x(T_{\beta,\pi} > t).$$

Iterate from time 1. The tail-sum identity gives $\mathbb{E}_x[T_{\beta,\pi}] = \sum_{t\geq 0} \Pr_x(T_{\beta,\pi} > t) \leq 1 + u_{1,x}/\rho$. If the initial-step condition holds, the geometric tail starts at time 0. Since $K_{\mathrm{state}} \leq L$ and $K_{\mathrm{state}} \leq T_{\beta,\pi}$, its expectation satisfies the stated cap bound. Finally, sum the discounted tail to bound M_x. □

The initial-step condition matters when conditioning on a prescribed action a_0: overlap of policy-averaged kernels alone need not imply overlap of that conditional first transition. Without it, the same argument still gives

$$\mathbb{E}_x[K_{\mathrm{state}}] \leq \min\{L, 1 + u_{1,x}/\rho\}, \qquad M_x(\gamma,\lambda) \leq 1 + \frac{\gamma\lambda u_{1,x}}{1-\gamma\lambda(1-\rho)}.$$

These bounds concern the state coupling. They do not bound the implemented action-level surrogate's stopping time, whose survival law and contraction are analyzed separately in Appendix B.8.

Why the two times cannot be substituted. Even when $\beta = \pi$, a common-state synchronous coupling has $T_{\beta,\pi} = 1$ while the behavior chain can mix slowly. For example, with

$$\mathcal{P}^\beta = \mathcal{P}^\pi = \begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix}, \qquad 0 < \varepsilon < \tfrac{1}{2},$$

the stationary law is uniform and $\|(\mathcal{P}^\beta)^t(0,\cdot) - \mu_\beta\|_{\mathrm{TV}} = \frac{1}{2}(1-2\varepsilon)^t$. Thus $\tau^\beta_{\mathrm{mix}}$ grows without bound as $\varepsilon \downarrow 0$ although this same-start first agreement remains 1. The valid mixing bound uses the worst-initial-pair β–β coalescence time in Proposition B.6; the MINT horizon uses the β–π survival time in Theorem B.7.

B.7 Online Control with Markov Traces

This subsection distinguishes Markovianity of a coupled pair from a Markov trace on the original state–action space. The online result below concerns the latter class of expected operators. It is a conditional control result, separate from the state-coupled evaluation theorem in Theorem 4.2.

State coupling on an augmented space. For fixed β, π, let $Z_t = (S^\beta_t, S^\pi_t)$ be a Markov coupling of the policy-induced state kernels, with joint transition kernel $K_{\beta,\pi}$ on $\mathcal{S} \times \mathcal{S}$. Its marginals are $\mathcal{P}^\beta(\cdot \mid S^\beta_t)$ and $\mathcal{P}^\pi(\cdot \mid S^\pi_t)$; maximal coupling is one such construction. For $z' = (s'_\beta, s'_\pi)$, define the killed kernel

$$\widetilde{K}_{\beta,\pi}(z, z') := K_{\beta,\pi}(z, z')\mathbf{1}\{s'_\beta \neq s'_\pi\}. \tag{35}$$

If $T = \inf\{t \geq 1 : S^\beta_t = S^\pi_t\}$, then

$$\begin{aligned} \Pr_z(T > t) &= (\widetilde{K}^t_{\beta,\pi}\mathbf{1})(z), \\ \sum_{t\geq 0}(\gamma\lambda)^t \Pr_z(T > t) &= \big[(I - \gamma\lambda\widetilde{K}_{\beta,\pi})^{-1}\mathbf{1}\big](z). \end{aligned} \tag{36}$$

Killing here terminates the trace at the first meeting; it does not require the two original chains to remain together after meeting. Thus the coupled construction can be Markov on the augmented space even when its survival cannot be described by a single keep function on the behavior state–action space. In particular, under maximal coupling the next disagreement probability depends on both coordinates through $\mathrm{TV}(\mathcal{P}^\beta(\cdot \mid s_\beta), \mathcal{P}^\pi(\cdot \mid s_\pi))$. Discarding the twin state need not preserve a Markov trace representation. An online theorem for the augmented construction and a projected update of $Q(s,a)$ would require an additional operator argument; Equation (36) alone is not that argument.

Markov-trace setup for the control result. Let $\mathcal{S}, \mathcal{A}$ be finite, rewards be bounded, $\gamma \in (0,1)$, and $\lambda \in [0,1]$. At control iteration n, fix β_n, π_n and a keep function $g_n : \mathcal{S} \times \mathcal{A} \to [0,1]$. These functions may be selected using Q_n, but are held fixed when applying that iteration's operator to any input Q. Given a behavior trajectory including its rewards, draw independent masks $B^{(n)}_i \sim \mathrm{Bernoulli}(g_n(s_i, a_i))$ with fresh randomness. Write $W^{(n)}_t = \prod_{i=1}^t B^{(n)}_i$ and $W^{(n)}_0 = 1$. The masks do not inspect the following transition or reward. Define

$$\mathcal{R}_n Q(x) := Q(x) + \mathbb{E}_{\beta_n, x}\left[\sum_{t\geq 0}(\gamma\lambda)^t W^{(n)}_t \delta^{Q,\pi_n}_t\right], \tag{37}$$

where $\delta^{Q,\pi_n}_t = r_t + \gamma\mathbb{E}_{b\sim\pi_n(\cdot|s_{t+1})}Q(s_{t+1}, b) - Q(s_t, a_t)$. The index n denotes a control iteration, whereas t denotes time within its behavior trajectory. This is an uncapped population operator. The implemented action-level rule is the case $g_n = h_n$ from Equation (49); the coupled-state indicators of Theorem 4.2 are not identified with g_n by notation alone.

Theorem B.8 (Exact control convergence with Markov traces)**.** *Under the Markov-trace setup above, consider $Q_{n+1} = \mathcal{R}_n Q_n$. Assume:*

(A1) Evaluation fixed point. *$\mathcal{R}_n Q^{\pi_n} = Q^{\pi_n}$ for every n. This identity follows from the specified operator, as shown below.*

(A2) Monotone policy improvement. $\mathcal{T}^{\pi_{n+1}}Q_{n+1} \geq \mathcal{T}^{\pi_n}Q_{n+1}$ *componentwise.*

(A3) Initialization. $\mathcal{T}^{\pi_0}Q_0 \geq Q_0$ *and* $Q_0 \leq Q^*$ *componentwise.*

(A4) Coefficient overlap. $\lambda\beta_n(a \mid s)g_n(s,a) \leq \pi_n(a \mid s)$ *for every* (s,a) *and every* n.

(A5) Asymptotic greediness. *For the Bellman optimality operator* $\mathcal{T}$*,*

$$\varepsilon_n := \|\mathcal{T}Q_n - \mathcal{T}^{\pi_n}Q_n\|_\infty \longrightarrow 0.$$

This requirement is separate from (A2).

Then $Q_n \uparrow Q^*$ *componentwise and* $\|Q_n - Q^*\|_\infty \to 0$.

Proof. **Markov factorization and the evaluation fixed point.** For a keep function g and behavior policy β, define

$$C_{\beta,g}((s,a),(s',b)) := \mathcal{P}(s' \mid s,a)\beta(b \mid s')g(s',b). \tag{38}$$

Write $C_n = C_{\beta_n,g_n}$. Since $C_n \geq 0$ and $C_n\mathbf{1} \leq \mathbf{1}$, the resolvent

$$\mathcal{N}_n := (I - \gamma\lambda C_n)^{-1} = \sum_{j\geq 0}(\gamma\lambda C_n)^j$$

exists and satisfies $\mathcal{N}_n \geq I$ entrywise and $\mathbf{1} \leq \mathcal{N}_n\mathbf{1} \leq (1-\gamma\lambda)^{-1}\mathbf{1}$. Conditional mask independence and the Markov property give

$$\mathbb{E}_{\beta_n,x}[W_t^{(n)}\delta_t^{Q,\pi_n}] = \big[C_n^t(\mathcal{T}^{\pi_n}Q - Q)\big](x). \tag{39}$$

The bounded geometric sum permits exchanging expectation and summation, hence

$$\mathcal{R}_nQ = Q + \mathcal{N}_n(\mathcal{T}^{\pi_n}Q - Q). \tag{40}$$

This proves (A1) directly. In fact, invertibility of $\mathcal{N}_n$ makes Q^{π_n} the unique fixed point even before (A4) is used.

Matrix order and non-negativity. We use only

$$\mathcal{N}_n(I - \gamma\lambda C_n) = (I - \gamma\lambda C_n)\mathcal{N}_n = I.$$

Thus $\mathcal{N}_n$ commutes with C_n. No commutation with $\mathcal{P}^{\pi_n}$, or between operators from different control iterations, is assumed. The linear part of Equation (40) is

$$\begin{aligned}\mathcal{R}_nQ - \mathcal{R}_nQ' &= H_n(Q - Q'),\\ H_n &:= I + \mathcal{N}_n(\gamma\mathcal{P}^{\pi_n} - I)\\ &= \gamma\mathcal{N}_n(\mathcal{P}^{\pi_n} - \lambda C_n).\end{aligned} \tag{41}$$

Indeed, $I = \mathcal{N}_n(I - \gamma\lambda C_n)$ gives this identity without moving $\mathcal{P}^{\pi_n}$ across $\mathcal{N}_n$. Under (A4),

$$(\mathcal{P}^{\pi_n} - \lambda C_n)((s,a),(s',b)) = \mathcal{P}(s' \mid s,a)[\pi_n(b \mid s') - \lambda\beta_n(b \mid s')g_n(s',b)] \geq 0.$$

Consequently $H_n \geq 0$, so $\mathcal{R}_n$ preserves componentwise order. Also,

$$\begin{aligned}H_n\mathbf{1} &= \gamma\mathcal{N}_n\mathbf{1} - \gamma\lambda\mathcal{N}_nC_n\mathbf{1}\\ &= \mathbf{1} - (1-\gamma)\mathcal{N}_n\mathbf{1} \leq \gamma\mathbf{1}.\end{aligned} \tag{42}$$

Therefore

$$\|H_n\|_\infty \leq \gamma, \tag{43}$$

and the fixed-point identity yields

$$\|Q_{n+1} - Q^{\pi_n}\|_\infty \leq \gamma\|Q_n - Q^{\pi_n}\|_\infty. \tag{44}$$

Monotonicity of the iterates. Let $b_n := \mathcal{T}^{\pi_n}Q_n - Q_n$. The residual after the update with the same target policy is

$$\begin{aligned}\mathcal{T}^{\pi_n}Q_{n+1} - Q_{n+1} &= b_n + (\gamma\mathcal{P}^{\pi_n} - I)\mathcal{N}_nb_n\\ &= \gamma(\mathcal{P}^{\pi_n} - \lambda C_n)\mathcal{N}_nb_n.\end{aligned}$$

Here $\mathcal{N}_n - I = \gamma\lambda C_n \mathcal{N}_n$ is used with $\mathcal{N}_n$ remaining on the right. In particular, this expression is not obtained by commuting $\mathcal{P}^{\pi_n}$ and $\mathcal{N}_n$. Conditions (A2) and (A4) give

$$b_{n+1} \geq \gamma(\mathcal{P}^{\pi_n} - \lambda C_n)\mathcal{N}_n b_n \geq 0 \tag{45}$$

whenever $b_n \geq 0$. Since $b_0 \geq 0$ by (A3), induction proves $b_n \geq 0$ for every n, and

$$Q_{n+1} - Q_n = \mathcal{N}_n b_n \geq 0. \tag{46}$$

Upper bound and optimality. Because $\mathcal{T}^{\pi_n} Q^* \leq \mathcal{T} Q^* = Q^*$ and $\mathcal{N}_n \geq 0$, Equation (40) implies $\mathcal{R}_n Q^* \leq Q^*$. Order preservation then gives

$$Q_n \leq Q^* \quad \Longrightarrow \quad Q_{n+1} = \mathcal{R}_n Q_n \leq \mathcal{R}_n Q^* \leq Q^*.$$

Together with (A3), this bounds the monotone sequence above by Q^*. Finiteness of $\mathcal{S} \times \mathcal{A}$ gives convergence to some $Q_\infty \leq Q^*$ in supremum norm, and hence $\|Q_{n+1} - Q_n\|_\infty \to 0$. Since $\mathcal{N}_n \geq I$ entrywise and $b_n \geq 0$,

$$0 \leq b_n \leq \mathcal{N}_n b_n = Q_{n+1} - Q_n, \qquad \|b_n\|_\infty \to 0.$$

Now apply (A5):

$$\|\mathcal{T} Q_n - Q_n\|_\infty \leq \varepsilon_n + \|b_n\|_\infty \longrightarrow 0. \tag{47}$$

The γ-contraction of $\mathcal{T}$ yields

$$\|Q_n - Q^*\|_\infty \leq \frac{\|\mathcal{T} Q_n - Q_n\|_\infty}{1 - \gamma} \longrightarrow 0,$$

so $Q_\infty = Q^*$. □

Roles of the assumptions and the surrogate specialization. As in Munos et al. [29, Theorem 2], (A2) and (A5) perform distinct tasks: monotone policy improvement preserves the residual sign, whereas asymptotic greediness rules out convergence to a nonoptimal policy value. Choosing every π_n greedy with respect to Q_n satisfies (A5) with $\varepsilon_n = 0$ and also supplies (A2) at the next update. The Markov-trace setup supplies (A1), while (A3) initializes monotonicity. Condition (A4) is a quantitative coefficient-overlap condition, not merely a common-support assumption: it makes both H_n and the residual propagation matrix in Equation (45) nonnegative.

For $g_n = h_n = 1 - \min(1, \pi_n/\beta_n)$ on the support of β_n, $\beta_n h_n = [\beta_n - \pi_n]_+$. Thus (A4) is equivalent to

$$\pi_n(a \mid s) \geq \frac{\lambda}{1+\lambda} \beta_n(a \mid s) \quad \text{for all } (s, a) \text{ with } \beta_n(a \mid s) > 0. \tag{48}$$

Indeed, where $\beta_n \leq \pi_n$ the coefficient constraint is automatic; otherwise it is $\lambda(\beta_n - \pi_n) \leq \pi_n$, which rearranges to Equation (48). At $\lambda = 0$ the condition is automatic. This equivalence concerns the action-level surrogate; it is not a re-expression of a general paired-state survival indicator. Unlike the fixed-policy χ_λ condition of Theorem B.9, (A4) is used here to obtain the order preservation needed for control.

Deep-RL extensions: entropy, KL constraints, and smoothing. Entropy-regularized actors, including SAC [15], and KL-based policy constraints motivate considering policy pairs with substantial overlap. To invoke this theorem for the surrogate, however, a quantitative bound such as $\inf_{s,a:\beta_n(a|s)>0} \pi_n(a \mid s)/\beta_n(a \mid s) \geq \lambda/(1+\lambda)$ must hold; full support, entropy regularization, or a finite KL bound alone is not used as a proof of (A4). This discussion concerns possible extensions of the operator analysis rather than an automatic guarantee for actor–critic training.

For deterministic-policy methods such as TD3, introducing explicitly specified smoothed policies can make density ratios available. An application to those smoothed policies would still require the coefficient-overlap and policy-improvement hypotheses above. Extending the result to the deterministic limit requires a separate continuity argument and is not established by the present theorem. Likewise, entropy-regularized optimal control requires the corresponding soft Bellman operator and policy-improvement conditions if its optimum is the object of interest.

Scope of the control result. The theorem concerns exact, uncapped, finite-dimensional updates of Equation (37). It does not assert that arbitrary state-coupled traces admit this factorization on (s, a), or that augmenting the pair state alone establishes the same control result after projection. Finite-cap evaluation retains terminal bootstrap terms as in Theorem B.9; a finite-cap control extension requires its own corresponding argument. Sample-based updates, nonlinear critics, and simultaneously learned actors are further extensions. These distinctions preserve the state-coupled evaluation theorem as a separate main result.

B.8 Contraction of the Action-Level Surrogate

This subsection analyzes the implemented action-level stopping rule directly, separately from the state-coupled operator of Theorem 4.2. The behavior policy may be arbitrary; contraction requires the sufficient condition below. No transition-kernel coupling or cross-Doeblin hypothesis is used.

Setup and survival quantities. Let $\mathcal{S}, \mathcal{A}$ be finite, rewards be bounded, and β, π be fixed policies. Fix $\gamma \in (0, 1)$, $\lambda \in [0, 1]$, and an integer hard cap $L = K_{\max} \geq 1$. Write $\alpha = \gamma\lambda$ and

$$h(s, a) := 1 - \min\left(1, \frac{\pi(a \mid s)}{\beta(a \mid s)}\right). \tag{49}$$

The ratio is needed only where $\beta(a \mid s) > 0$; set $h = 0$ otherwise. Thus $\beta(a \mid s)h(s, a) = [\beta(a \mid s) - \pi(a \mid s)]_+$ everywhere. Given a behavior trajectory, including its rewards, draw independent masks $\hat{A}_i \sim \mathrm{Bernoulli}(h(s_i, a_i))$ using fresh randomness. In particular, the mask at i does not inspect the next transition or reward. Define $W_t = \prod_{i=1}^{t} \hat{A}_i$, $W_0 = 1$, and $\hat{T} = \inf\{i \geq 1 : \hat{A}_i = 0\}$, with $\inf \varnothing = \infty$. The expected surrogate operator is

$$\mathcal{R}^{\mathrm{sur}}_{\lambda,L} Q(x) := Q(x) + \mathbb{E}_x\left[\sum_{t=0}^{L-1} \alpha^t W_t \delta_t^Q\right], \qquad x = (s_0, a_0). \tag{50}$$

The cap in this definition is exactly the implementation's $\min(L, \hat{T})$; it is not an uncapped approximation.

Let $D_s = \mathrm{TV}(\beta(\cdot \mid s), \pi(\cdot \mid s))$. Define the negative-coefficient mass and its normalized fraction by

$$e_\lambda(s) := \sum_a [\lambda\beta(a \mid s)h(s, a) - \pi(a \mid s)]_+ = \sum_a [\lambda\beta(a \mid s) - (1 + \lambda)\pi(a \mid s)]_+, \tag{51}$$

$$\chi_\lambda(s) := \begin{cases} e_\lambda(s)/(\lambda D_s), & \lambda D_s > 0, \\ 0, & \lambda D_s = 0. \end{cases}$$

Since $0 \leq e_\lambda(s) \leq \lambda D_s$, we have $\chi_\lambda(s) \in [0, 1]$. This fraction measures the negative-weight correction left by the action-dependent keep rule; it is generally different from the policy TV D_s.

For every starting pair x, define

$$\widehat{M}_L^x := \sum_{t=0}^{L-1} \alpha^t \Pr_x(\hat{T} > t), \tag{52}$$

$$\bar{\chi}^x_{\lambda,L} := \frac{\sum_{t=1}^{L-1} \alpha^t \mathbb{E}_x[W_t \chi_\lambda(s_t)]}{\widehat{M}_L^x - 1},$$

where the second quantity is 0 if its denominator vanishes. All survival probabilities here belong to the surrogate's own stopping process. In particular, neither $\widehat{M}_L^x$ nor $\bar{\chi}^x_{\lambda,L}$ is inferred by substituting $\hat{p}_i$ into the state-level coupling analysis.

Theorem B.9 (Finite-cap surrogate contraction)**.** *Under the setup above, $\mathcal{R}^{\mathrm{sur}}_{\lambda,L}$ has the unique fixed point Q^π. Moreover, for all Q, Q', its supremum-norm Lipschitz modulus is bounded by*

$$\kappa_{\chi,L} := \sup_x \left\{1 - (1 - \gamma)\widehat{M}_L^x + 2\bar{\chi}^x_{\lambda,L}(\widehat{M}_L^x - 1)\right\}. \tag{53}$$

Consequently, if $\kappa_{\chi,L} < 1$, the exact iteration $Q_{n+1} = \mathcal{R}^{\mathrm{sur}}_{\lambda,L} Q_n$ satisfies

$$\|Q_n - Q^\pi\|_\infty \leq \kappa^n_{\chi,L} \|Q_0 - Q^\pi\|_\infty.$$

The condition $\kappa_{\chi,L} < 1$ is sufficient, not necessary.

Proof. **Expected operator and fixed point.** On state–action pairs define the nonnegative matrix

$$C((s,a),(s',b)) := \mathcal{P}(s' \mid s,a)\beta(b \mid s')h(s',b), \qquad S_L := \sum_{j=0}^{L-1}(\alpha C)^j, \quad S_0 := 0. \tag{54}$$

Because $C\mathbf{1} \le \mathbf{1}$ and $\alpha < 1$, $\|\alpha C\|_\infty < 1$. Conditional mask independence gives $\mathbb{E}[W_t \mid \text{trajectory}] = \prod_{i=1}^t h(s_i,a_i)$. The Markov property and the conditional mean of the TD residual therefore yield

$$\mathbb{E}_x[W_t\delta_t^Q] = \big(C^t(\mathcal{T}^\pi Q - Q)\big)(x), \qquad \mathcal{R}^{\mathrm{sur}}_{\lambda,L}Q = Q + S_L(\mathcal{T}^\pi Q - Q).$$

The factorization $S_L = (I - (\alpha C)^L)(I - \alpha C)^{-1}$ shows that S_L is invertible. Hence $\mathcal{R}^{\mathrm{sur}}_{\lambda,L}Q = Q$ if and only if $\mathcal{T}^\pi Q = Q$, proving uniqueness of the fixed point Q^π.

Finite-cap boundary. The linear part of the affine operator is

$$\begin{aligned} H_L &:= I + S_L(\gamma\mathcal{P}^\pi - I) \\ &= \gamma S_{L-1}(\mathcal{P}^\pi - \lambda C) + \gamma(\alpha C)^{L-1}\mathcal{P}^\pi. \end{aligned} \tag{55}$$

This follows by expanding the finite sums; no commutativity between C and $\mathcal{P}^\pi$ is used. The last term is the terminal bootstrap contribution, and S_{L-1}, rather than S_L, multiplies the signed interior contribution. Also,

$$S_L\mathbf{1} = (\widehat{M}_L^x)_x, \qquad H_L\mathbf{1} = \mathbf{1} - (1-\gamma)S_L\mathbf{1}. \tag{56}$$

Negative mass. Let $E_\lambda := (\lambda C - \mathcal{P}^\pi)_+$, with positive part and absolute values taken entrywise. Then

$$|\mathcal{P}^\pi - \lambda C| = \mathcal{P}^\pi - \lambda C + 2E_\lambda.$$

All other factors in Equation (55) are nonnegative, so

$$|H_L|\mathbf{1} \le \mathbf{1} - (1-\gamma)S_L\mathbf{1} + 2\gamma S_{L-1}E_\lambda\mathbf{1}. \tag{57}$$

The state-level mass of E_λ is exactly e_λ from Equation (51). If $\widetilde{\chi}(s,a) := \chi_\lambda(s)$, then

$$(E_\lambda\mathbf{1})(s,a) = \sum_{s'}\mathcal{P}(s' \mid s,a)e_\lambda(s') = \lambda(C\widetilde{\chi})(s,a),$$

where $\sum_b \beta(b \mid s')h(s',b) = D_{s'}$ was used. Therefore

$$\begin{aligned} \gamma(S_{L-1}E_\lambda\mathbf{1})(x) &= \sum_{t=1}^{L-1}\alpha^t(C^t\widetilde{\chi})(x) \\ &= \sum_{t=1}^{L-1}\alpha^t\mathbb{E}_x[W_t\chi_\lambda(s_t)] \\ &= \bar{\chi}^x_{\lambda,L}(\widehat{M}_L^x - 1). \end{aligned}$$

Substituting into Equation (57) and taking the maximum row absolute sum proves $\|H_L\|_\infty \le \kappa_{\chi,L}$. Since $\mathcal{R}^{\mathrm{sur}}_{\lambda,L}Q - \mathcal{R}^{\mathrm{sur}}_{\lambda,L}Q' = H_L(Q - Q')$, contraction and the stated geometric convergence follow when $\kappa_{\chi,L} < 1$. □

Corollary B.10 (A sufficient overlap condition)**.** *If $\pi(a \mid s) \ge \lambda\beta(a \mid s)/(1+\lambda)$ for all (s,a), then $\mathcal{R}^{\mathrm{sur}}_{\lambda,L}$ is a γ-contraction for every $L \ge 1$.*

Proof. The condition is equivalent to $\lambda\beta h \le \pi$, so $e_\lambda = 0$ and $E_\lambda = 0$. Equation (55) then gives $H_L \ge 0$. Since $S_L\mathbf{1} \ge \mathbf{1}$, Equation (56) gives $\|H_L\|_\infty \le \gamma$. This is the pointwise overlap condition appearing in Equation (48), used here solely for fixed-policy evaluation. □

Boundary cases and scope. At $L = 1$ or $\lambda = 0$ the surrogate operator reduces to $\mathcal{T}^\pi$. For an infinite cap, the same proof uses $S = (I - \alpha C)^{-1}$ in place of both S_L and S_{L-1}, and the survival sums extend to infinity. A bound at or above 1 is inconclusive: exact finite-dimensional iteration converges from every initial value if and only if $\rho(H_L) < 1$, which need not require $\|H_L\|_\infty < 1$. These are population results for fixed policies. Finite-sample concentration, sample-based stochastic approximation, changing policies, and nonlinear critic projection require additional analysis. In particular, averaging the complete action-dependent backup in a minibatch does not replace $h(s,a)$ by its state mean D_s inside that backup.

Why a contraction condition is needed. Consider two states $0, 1$ and two deterministic actions: *stay* preserves the state and *flip* switches it. Let β choose each action with probability $1/2$, let π always flip, and set rewards to zero. Then $h(s, \text{stay}) = 1$ and $h(s, \text{flip}) = 0$. For $v(0) = 1, v(1) = -1$, define $Q_0(s, a) = v(f(s, a))$, where f is the deterministic next-state map. We have $CQ_0 = Q_0/2$ and $\mathcal{P}^\pi Q_0 = -Q_0$, hence

$$H_L Q_0 = \mu_L Q_0, \qquad \mu_L = 1 - (1 + \gamma) \sum_{t=0}^{L-1} (\gamma\lambda/2)^t.$$

At $\gamma = 0.99$, $\lambda = 0.7$, $L = 5$, this eigenvalue is $\mu_L \approx -2.02993$, so exact expected iteration can diverge even with a finite cap. This example does not satisfy the new sufficient contraction condition; it explains why that condition cannot be omitted. It makes no claim about small-step stochastic updates or the empirical SAC/TD3 training runs.

B.9 A Common Conservative TV Condition

Corollary B.11 (A common sufficient condition)**.** *Let $D = \sup_s \mathrm{TV}(\beta(\cdot \mid s), \pi(\cdot \mid s))$. Under the respective fixed-policy hypotheses of Theorems B.9 and 4.2, the condition*

$$\gamma(1 + 2\lambda D) < 1 \tag{58}$$

ensures contraction of each expected evaluation operator toward Q^π; for the surrogate this holds for every finite cap $L \geq 1$.

Proof. Set $\varepsilon_0 = 1 - \gamma - 2\gamma\lambda D > 0$. For the state-coupled operator, $N_x \leq D(M_x - 1)$ and $M_x - 1 \leq \gamma\lambda M_x$ give

$$1 - (1 - \gamma)M_x + 2N_x \leq 1 - \varepsilon_0 M_x \leq 1 - \varepsilon_0 < 1$$

uniformly over x. For the surrogate, $e_\lambda(s) \leq \lambda D_s \leq \lambda D$, so $E_\lambda \mathbf{1} \leq \lambda D \mathbf{1}$. Since $S_{L-1}\mathbf{1} \leq S_L \mathbf{1}$, Equation (57) gives

$$|H_L|\mathbf{1} \leq \mathbf{1} - \varepsilon_0 S_L \mathbf{1} \leq (1 - \varepsilon_0)\mathbf{1}.$$

The separate fixed-point results complete both claims. □

For $\lambda > 0$, Equation (58) requires $D < (1 - \gamma)/(2\gamma\lambda)$. At the main MuJoCo setting $(\gamma, \lambda) = (0.99, 0.7)$, this is $D < 0.00722$ uniformly over states. It is a restrictive shared baseline, not a certificate for the experimental regime. The survival-dependent state condition and the surrogate's negative-mass condition can be less restrictive; their relative restrictiveness depends on their stopping distributions and actionwise policy ratios. The common condition neither identifies the two stopping processes nor transfers one theorem to the other operator.

C Extended Theoretical Discussion

Reading the phase diagram. Figure 3 (a) shows that the sufficient-condition boundary $\gamma(1+2\eta) < 1$ collapses to a curve in the (η, γ) plane: settings with different mismatch, trace decay, and coupling overlap are equivalent whenever they induce the same $\eta := \bar{D}_{\mathrm{wgt}}\lambda(1 - \rho)$. Panel (b) visualizes the corresponding $(\bar{D}_{\mathrm{wgt}}, \rho)$ plane at $\gamma = 0.99$ for three values of λ.

C.1 Four Regimes of Theorem 4.2

For the scalar examples below, suppose the survival law is the same for all starting pairs and write $M = M_x$. The uniform modulus reduces to $\kappa = 1 - (1 - \gamma)M + 2\bar{D}_{\mathrm{wgt}}(M - 1)$, and the uniform TV-based sufficient condition Equation (25) reduces to $2\bar{D}_{\mathrm{wgt}}(M - 1)/M < 1 - \gamma$. For nonuniform survival laws, the supremum over starting pairs must be retained. The main text discusses Regime 3 (geometric meeting); the other three regimes are recorded here for completeness.

Regime 1 — On-policy limit ($\bar{D}_{\mathrm{wgt}} \to 0$). For every starting pair, $N_x \leq \bar{D}_{\mathrm{wgt}}(M_x - 1) \leq \bar{D}_{\mathrm{wgt}}\gamma\lambda/(1 - \gamma\lambda)$, so the correction vanishes uniformly as $\bar{D}_{\mathrm{wgt}} \to 0$ for fixed γ, λ. At $\bar{D}_{\mathrm{wgt}} = 0$, $\kappa_x = 1 - (1 - \gamma)M_x \leq \gamma$ for every x. In the on-policy case, $M_x = 1$ and $N_x = 0$ for all x, recovering $\kappa = \gamma$.

Regime 2 — Deterministic meeting at $T = 1$. $\Pr(T > 0) = 1$, $\Pr(T > t) = 0$ for $t \geq 1$, so $M = 1$ and $\kappa \leq \gamma$. *One-step Bellman is recovered for any* $\bar{D}_{\text{wgt}}$, reflecting that with immediate meeting there is no off-policy correction to apply.

Regime 3 — Geometric meeting time, $T \sim \text{Geom}(q)$. $\Pr(T > t) = (1-q)^t$, so $M = 1/\big(1 - \gamma\lambda(1-q)\big)$ and $(M-1)/M = \gamma\lambda(1-q)$. Substituting into Equation (25) gives the closed-form sufficient condition

$$\boxed{\gamma\big(1 + 2\bar{D}_{\text{wgt}}\lambda(1-q)\big) \;<\; 1.}$$

The contractive regime widens as the per-step meeting probability q grows: at $q = 1$ we recover Regime 2; at $q = 0$ we recover Regime 4 below. This case is realized when the coupled processes mix at a uniform geometric rate, which is exactly the situation of Corollary B.1 (with $q = \rho$, the cross-Doeblin parameter).

Regime 4 — Worst case ($T = \infty$). The two coupled trajectories never meet. $\Pr(T > t) = 1$ for all t, so $M = 1/(1-\gamma\lambda)$ and $(M-1)/M = \gamma\lambda$. The condition becomes

$$\boxed{\gamma\big(1 + 2\bar{D}_{\text{wgt}}\lambda\big) \;<\; 1.}$$

At $\bar{D}_{\text{wgt}} = 1, \lambda = 1$ this gives $\gamma < 1/3$, the most pessimistic case. But at $\bar{D}_{\text{wgt}} = 0.1, \lambda = 1$ it admits $\gamma < 1/1.2 \approx 0.833$; at $\bar{D}_{\text{wgt}} = 0.01$ it admits $\gamma < 0.98$. Even in the never-meeting worst case, modest control over $\bar{D}_{\text{wgt}}$ rescues high discount factors.

C.2 Practical Regime Corollaries

Theorem 4.2 holds for any $\gamma \in (0, 1)$, but the worst-case quantitative bound becomes vacuous when γ is close to 1 unless $\bar{D}_{\text{wgt}}$ is correspondingly small. In this subsection we specialize to the moderate-discount regime $\gamma \leq 0.9$ and give two clean sufficient conditions: a worst-case (ℓ_∞) form and an averaged (L^1) form. Both follow from the pointwise bound of Theorem 4.2. The averaged result is an $L^\infty \to L^1$ one-step error bound; the averaging step requires neither cross-Doeblin nor ergodicity and is not an iteration-convergence argument.

Untruncated TD-error reference. Setting $A_i \equiv 1$ in Equation (7) gives $Q + (I - \gamma\lambda\mathcal{P}^\beta)^{-1}(\mathcal{T}^\pi Q - Q)$, the $Q^\pi(\lambda)$ evaluation operator of Harutyunyan et al. [18]. It retains Q^π as its fixed point. This is distinct from Peng's mixture-of-returns operator in Table 1. The no-survival reference $M_x = (1-\gamma\lambda)^{-1}$ gives the sufficient bound $\gamma(1 + 2\bar{D}_{\text{wgt}}\lambda) < 1$ from the current analysis. Comparisons with this expression concern sufficient bounds; they do not identify a convergence region for Peng's algorithm.

Corollary C.1 (Practical worst-case contraction at moderate discount)**.** *Adopt the operator setup of Theorem 4.2. Let* $K_\lambda = \sup_x \min(\mathbb{E}_x[T_{\beta,\pi}], (1-\gamma\lambda)^{-1})$ *be the uniform effective-horizon bound in Equation (26). Fix* $\gamma \in (0, 1)$.

- ***(A) Free-trace regime.*** *If* $\bar{D}_{\text{wgt}} \leq \frac{1-\gamma}{2}$*, then* $\kappa < 1$ *for* every $K_\lambda \geq 1$ *(equivalently, every meeting-time distribution and every* $\lambda \in [0, 1]$*).*
- ***(B) Bounded-trace regime.*** *If* $\bar{D}_{\text{wgt}} > \frac{1-\gamma}{2}$*, then* $\kappa < 1$ *whenever*

$$K_\lambda \;<\; \frac{2\bar{D}_{\text{wgt}}}{2\bar{D}_{\text{wgt}} - (1-\gamma)} \;=:\; K_\lambda^{\max}(\gamma, \bar{D}_{\text{wgt}}). \tag{59}$$

  In particular, choosing $\lambda < \lambda^*(\gamma, \bar{D}_{\text{wgt}}) := \frac{1-\gamma}{2\gamma\bar{D}_{\text{wgt}}}$ *forces* $\frac{1}{1-\gamma\lambda} < K_\lambda^{\max}$ *and therefore guarantees contraction for every meeting-time distribution.*

Proof. For every x, Equation (22) gives $\kappa_x \leq 1 - (1-\gamma)M_x + 2\bar{D}_{\text{wgt}}(M_x - 1)$. In Regime (A), the coefficient $2\bar{D}_{\text{wgt}} - (1-\gamma)$ of M_x is non-positive, so $M_x \geq 1$ yields $\kappa_x \leq \gamma$ uniformly. In Regime (B), this coefficient is positive, so $M_x \leq K_\lambda$ yields a uniform upper bound strictly below 1 when $K_\lambda < K_\lambda^{\max}$. The λ^* statement follows from $(1-\gamma\lambda^*)^{-1} = K_\lambda^{\max}$. □

Table 5: Contraction budget at $\gamma = 0.9$ (Corollary C.1). $\lambda^* = (1-\gamma)/(2\gamma\bar{D}_{\text{wgt}})$ is the critical λ below which contraction holds for any meeting-time distribution. $K_\lambda^{\max}$ is the maximum effective horizon preserving $\kappa < 1$ at the indicated λ.

$\bar{D}_{\text{wgt}}$	Regime	λ^*	$K_\lambda^{\max}$ (λ=1)	$K_\lambda^{\max}$ (λ=0.5)
0.001	(A)	any	∞	∞
0.05	(A)	any	∞	∞
0.10	(B)	0.556	2.00	∞
0.30	(B)	0.185	1.20	1.20
0.50	(B)	0.111	1.11	1.11
1.00	(B)	0.0556	1.05	1.05

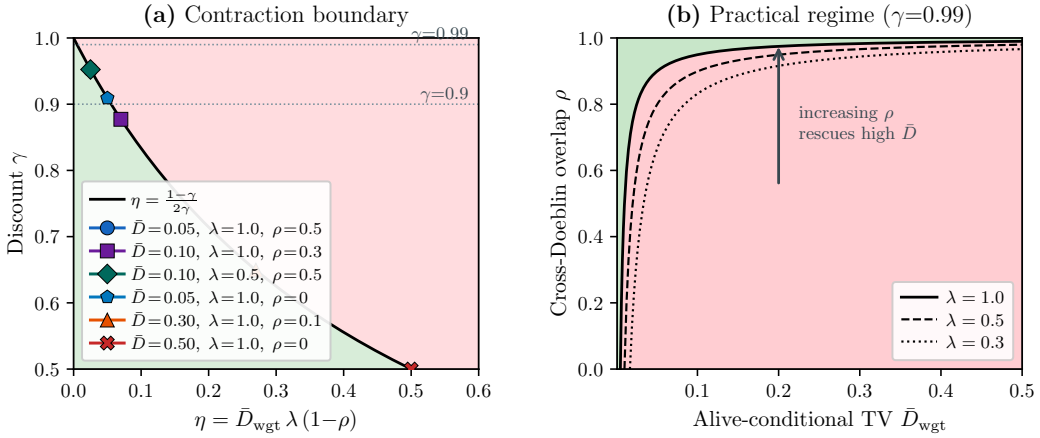


Figure 6: Contraction phase diagram of Corollary B.1. **(a)** The contraction condition $\gamma(1+2\eta) < 1$ defines a single boundary in the (η, γ) plane, where the effective mismatch $\eta := \bar{D}_{\text{wgt}}\,\lambda\,(1-\rho)$ combines all three factors. Markers show representative scenarios; only their product η determines the critical γ. **(b)** At γ=0.99, the $(\bar{D}_{\text{wgt}}, \rho)$ plane for three λ values. Higher cross-Doeblin overlap ρ rescues contraction even at large $\bar{D}_{\text{wgt}}$; reducing λ further expands the contractive region.

Remark C.2 (Specialization to $\gamma = 0.9$). At $\gamma = 0.9$ the free-trace threshold is $\bar{D}_{\text{wgt}} \le 0.05$. In the bounded-trace regime: $\bar{D}_{\text{wgt}} = 0.10$ admits $K_\lambda^{\max} = 2.00$ (effective two-step traces); $\bar{D}_{\text{wgt}} = 0.50$ admits $K_\lambda^{\max} \approx 1.11$ (only marginally longer than one-step). For deep RL with $\gamma = 0.9$ and on-policy-like data ($\bar{D}_{\text{wgt}} \lesssim 0.05$), contraction is unconditional—no tuning of λ or truncation length needed.

Corollary C.3 (Averaged one-step error bound)**.** *Adopt the state-coupled operator setup and the pointwise Lipschitz bound of Theorem 4.2, without requiring its uniform contraction condition. Fix a probability distribution ν over starting pairs $x = (s, a)$, and write*

$$\bar{M} := \mathbb{E}_{x\sim\nu}[M_x], \qquad \|f\|_{L^1(\nu)} := \mathbb{E}_{x\sim\nu}[|f(x)|].$$

Then for every bounded Q,

$$\|\mathcal{R}_{p,\lambda}Q - Q^\pi\|_{L^1(\nu)} \le \kappa^\star \|Q - Q^\pi\|_\infty, \tag{60}$$

where

$$\kappa^\star := 1 - (1-\gamma)\bar{M} + 2\bar{D}_{\text{wgt}}(\bar{M} - 1). \tag{61}$$

The coefficient is below one in either of the following regimes:

- *If $\bar{D}_{\text{wgt}} \le (1-\gamma)/2$, then $\kappa^\star \le \gamma < 1$ for every admissible $\bar{M} \ge 1$.*
- *If $\bar{D}_{\text{wgt}} > (1-\gamma)/2$, then $\kappa^\star < 1$ whenever $\bar{M} < 2\bar{D}_{\text{wgt}}/[2\bar{D}_{\text{wgt}} - (1-\gamma)]$.*

No cross-Doeblin or mixing assumption is needed for this averaging step. If the behavior chain has a stationary law μ_β, one may choose $\nu = \mu_\beta \otimes \beta$ as a special case.

Proof. The fixed-point identity and the pointwise bound of Theorem 4.2 give

$$|\mathcal{R}_{p,\lambda}Q(x) - Q^\pi(x)| \le \left[1 - (1-\gamma)M_x + 2\bar{D}_{\mathrm{wgt}}(M_x - 1)\right] \|Q - Q^\pi\|_\infty.$$

Integrating with respect to ν gives Equation (60). The two regimes follow by considering the sign of $2\bar{D}_{\mathrm{wgt}} - (1-\gamma)$ and using $\bar{M} \ge 1$. □

This is an $L^\infty \to L^1(\nu)$ one-step error bound. Even if $\kappa^\star < 1$, it is not an $L^1(\nu)$ contraction, since the right-hand side uses the supremum norm. Consequently it does not by itself prove convergence of repeated updates; the uniform sufficient condition of Theorem 4.2 supplies that separate guarantee.

Where the deep-RL runs sit. The free-trace threshold $(1-\gamma)/2$ is 0.005 at $\gamma = 0.99$. The RWRL runs report mean action-level keep probabilities between 0.70 and 0.99 depending on the arm, but these statistics do not identify the state-coupled survival quantities or certify the χ_λ condition of Theorem B.9. The measured $\hat{\kappa} = 0.84 \pm 0.13$ at $K_{\max} = 5$, below 1 on 90% of probes (Appendix G.4), is empirical evidence about visited-distribution backup gains, not an estimate of $\kappa^\star$ or a certificate of $\kappa_{\chi,L} < 1$.

D Extended Toy-Chain Validation

D.1 Diagnostic study on tabular CliffWalking

Although Assumptions A.2 and A.3 (kernel-level TV bound d and cross-Doeblin parameter ρ) may appear restrictive, they are minimal for quantifying kernel-level discrepancies relevant to mixing behavior. Our diagnostic study on the tabular `CliffWalking` environment (Figure 7 (a)) illustrates this clearly: even slight policy deviations induce near-deterministic branching and absorbing transitions, driving state-wise TV distances close to 1 and collapsing cross-state overlap. As shown in Figure 7 (b), both the state-wise discrepancies $d(s)$ and the overlap matrix $\rho(s, s')$ between an optimal policy and a uniformly random policy exhibit extreme mismatch across most states—reflecting structural properties of the MDP rather than policy proximity. Figure 7 (c) evaluates the RHS of Lemma B.2 at several illustrative (d, ρ) pairs. These curves bound time-averaged expected disagreement under their stated hypotheses; they do not verify a positive uniform ρ for CliffWalking or imply a meeting-time tail. The latter requires the separate conditional overlap assumptions in Theorem B.7.

D.2 Single-Trap chain refinements

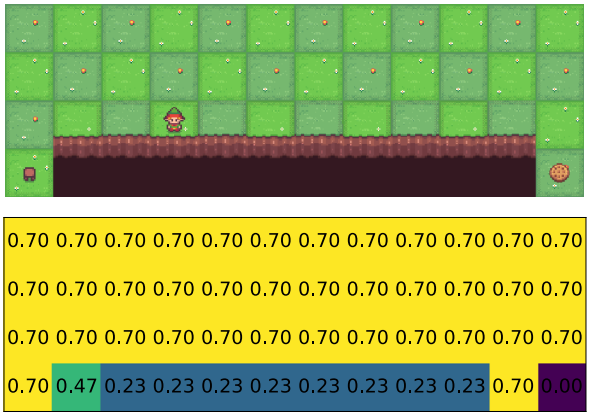
We extend the Single-Trap chain validation of the main text with two additional refinements: data-processing gap (refinement (ii)) and cross-Doeblin restart (refinement (iii)). Setup follows the main text: Single-Trap chain with states $\{0, \dots, N\}$, reflective boundaries, two actions $\{-1, +1\}$, β uniform, π identical to β except $\pi(+1 \mid N) = 1$ at the trap state. The per-state action TV is $D_s = 0$ everywhere except $D_N = 0.5$, so the worst-case bound is $\bar{D} = 0.5$ regardless of dynamics.

D.3 Refinement (ii): Data-processing gap on slippery transitions

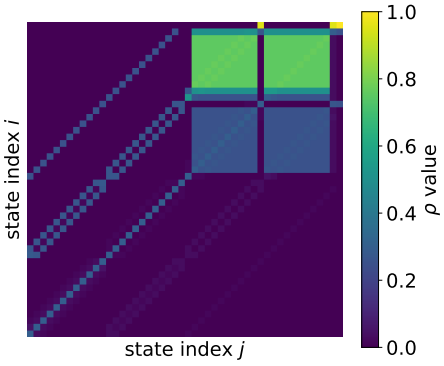
With slippery transitions (intended action with probability $1 - \xi$, a uniform random action otherwise), the action-level TV $\bar{D} = 0.5$ is unchanged, while the one-step kernel TV at the trap shrinks to 0.25 at $\xi = 0.5$. The plotted kernel-weighted diagnostic drops from 0.11 to 0.06, while the policy-TV diagnostic remains around 0.12 (Figure 8). This illustrates a one-step data-processing gap. Kernel TV cannot replace policy TV in the correction term of Theorem 4.2: that proof bounds signed coefficients on state–action values before any push-forward through the dynamics. The weighted kernel quantity is a diagnostic, not a contraction certificate. The implemented surrogate has a separate χ_λ correction (Theorem B.9).

D.4 Refinement (iii): Cross-Doeblin restart and the no-survival reference

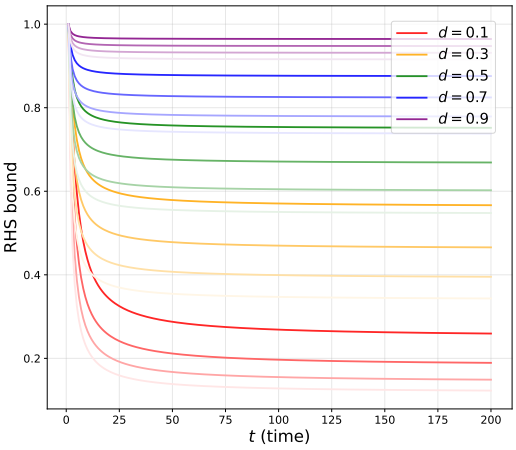
We add a uniform-restart component (with prob q_{rst}, transition is uniform over all states; with prob $1 - q_{\mathrm{rst}}$, slippery dynamics) to obtain $\rho \approx q_{\mathrm{rst}}$ exactly (Figure 9 (a)). We compare `MINT`'s contraction LHS $\gamma(1 + 2\bar{D}_{\mathrm{wgt}}\lambda(1-\rho))$ against the no-survival reference LHS $\gamma(1 + 2\bar{D}_{\mathrm{wgt}}\lambda)$ (the $\rho = 0$



(a)



(b)



(c)

Figure 7: **Diagnostic study on tabular CliffWalking.** (Panels enlarged relative to the submitted version.) **(a-top)** Environment visualization for `CliffWalking`. **(a-bottom)** State-wise total variation distance $d(s)$ between an optimal policy and a uniformly random policy, illustrating pronounced off-policy discrepancies near the cliff region. **(b)** Pairwise overlap matrix $\rho(s, s')$ between transition kernels induced by the two policies; near-zero overlap across most state pairs highlights a strong structural mismatch. **(c)** Time-averaged expected-disagreement bounds from Lemma B.2 for illustrative (d, ρ) pairs; these curves do not certify a meeting-time bound for the displayed environment.

specialization of this sufficient bound). Table 6 reports the measured region ratio $1/(1-\rho)$ alongside both LHS values. The decisive operating point is $q_{\text{rst}} = 0.20$: `MINT`'s LHS drops to $0.985 < 1$ (contraction), while the reference LHS is $1.006 > 1$ (no contraction guarantee). The empirical region ratio matches the theoretical $1/(1-\rho)$ to three decimal places across $\rho \in \{0.1, 0.3, 0.5\}$, providing a quantitative comparison of these sufficient bounds.

Summary of refinements. The diagnostics have separate roles. Survival weighting gives the plotted policy-TV value $\bar{D}_{\text{wgt}} \approx 0.1$, compared with the worst-case action TV $\bar{D} = 0.5$. Slippery dynamics give a smaller kernel-TV diagnostic, $\bar{p}_{\text{wgt}} \approx 0.06$. For any common nonnegative normalized weights ω_s, DPI implies $\sum_s \omega_s d(s) \leq \sum_s \omega_s D_s$, but this does not justify replacing $\bar{D}_{\text{wgt}}$ by $\bar{p}_{\text{wgt}}$ in Theorem 4.2. Cross-Doeblin overlap separately changes the geometric-tail sufficient bound by the factor $1-\rho$; the displayed threshold ratio $1/(1-\rho)$ compares those two state-coupled sufficient bounds with the same policy-TV parameter. These observations do not compose into a multiplicative contraction improvement and do not transfer a state-coupled guarantee to the implemented surrogate.

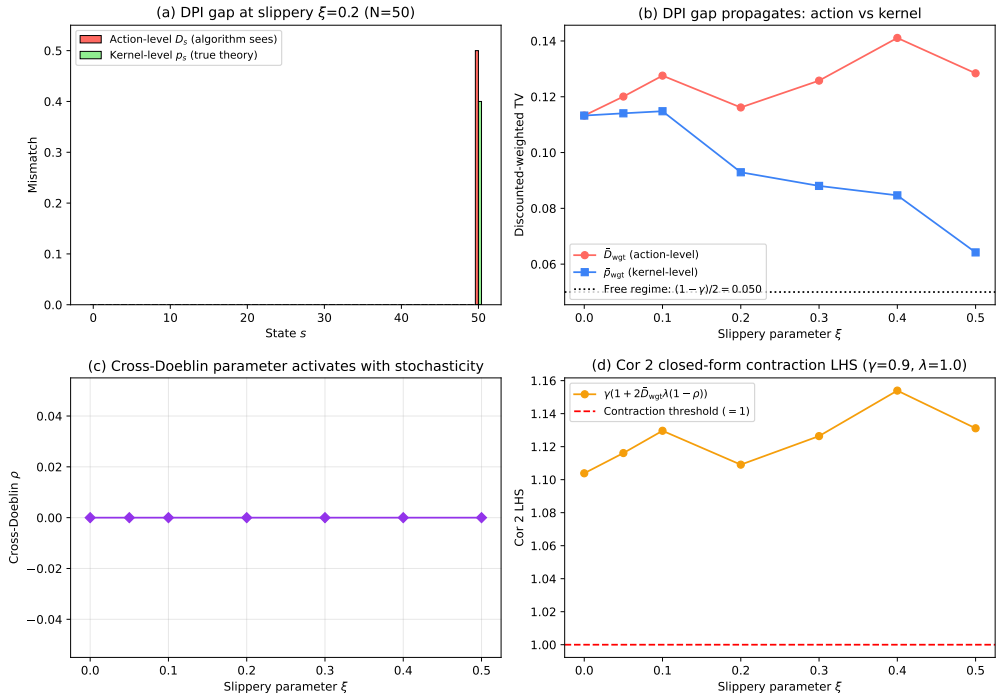


Figure 8: Stage B: data-processing gap on the slippery Single-Trap chain (N=50, γ=0.9, λ=1, slippery ξ varied). **(a)** Action-level D_s (red) is unchanged by slippery dynamics, while kernel-level p_s (green) shrinks at the trap state. **(b)** The plotted policy-TV and kernel-TV weighted diagnostics separate with ξ; the smaller kernel diagnostic is not substituted into the contraction condition. **(c)** Cross-Doeblin $\rho = 0$ for the chain MDP without restart. **(d)** Cor 2 LHS exceeds 1 throughout: this MDP is outside Corollary B.1's applicable regime; restart dynamics (Stage B-2) restore $\rho > 0$.

Table 6: Stage B-2: `MINT` and the no-survival reference bound on the restart chain (N=30, γ=0.9, λ=1, ξ=0.1). The contraction condition is $\gamma(1 + 2\bar{D}_{\text{wgt}}\lambda(1-\rho)) < 1$ for `MINT` and the ρ=0 specialization for the reference. The decisive row is $q_{\text{rst}} = 0.20$: `MINT` contracts (✓) while the reference bound is inconclusive.

q_{rst}	measured ρ	`MINT` LHS	Reference LHS	$1/(1-\rho)$
0.00	0.000	1.101	1.101	1.000
0.05	0.050	1.064	1.072	1.053
0.10	0.100	1.023	1.036	1.111
0.20	**0.200**	**0.985** ✓	1.006	1.250
0.30	0.300	0.957 ✓	0.982 ✓	1.429
0.50	0.500	0.930 ✓	0.960 ✓	2.000

D.5 High-Mismatch Ablations on the Single-Trap Chain

We compare stopping rules in a controlled high-mismatch stress test, using the single-trap dynamics but different policy pairs from the localized-mismatch refinements in Figures 8 and 9. The policies below differ at every state. Results from these settings are not pooled with, or compared in absolute scale to, the localized-mismatch experiments.

Setup and measured quantities. We use $N = 20$, $\gamma = 0.9$, $\lambda = 1$, slippery transitions with probability 0.1, and a 50-step evaluation window. The representative setting is

$$\beta(+1 \mid s) = 0.05, \qquad \pi(+1 \mid s) = 0.95, \qquad D_s = 0.90 \quad \text{for every } s.$$

An additional comparison uses $\beta(+1 \mid s) = 0.005$ and $\pi(+1 \mid s) = 0.995$, giving $D_s = 0.99$ at every state. Within each setting, all variants use the same pool of behavior trajectories: 20 replicates

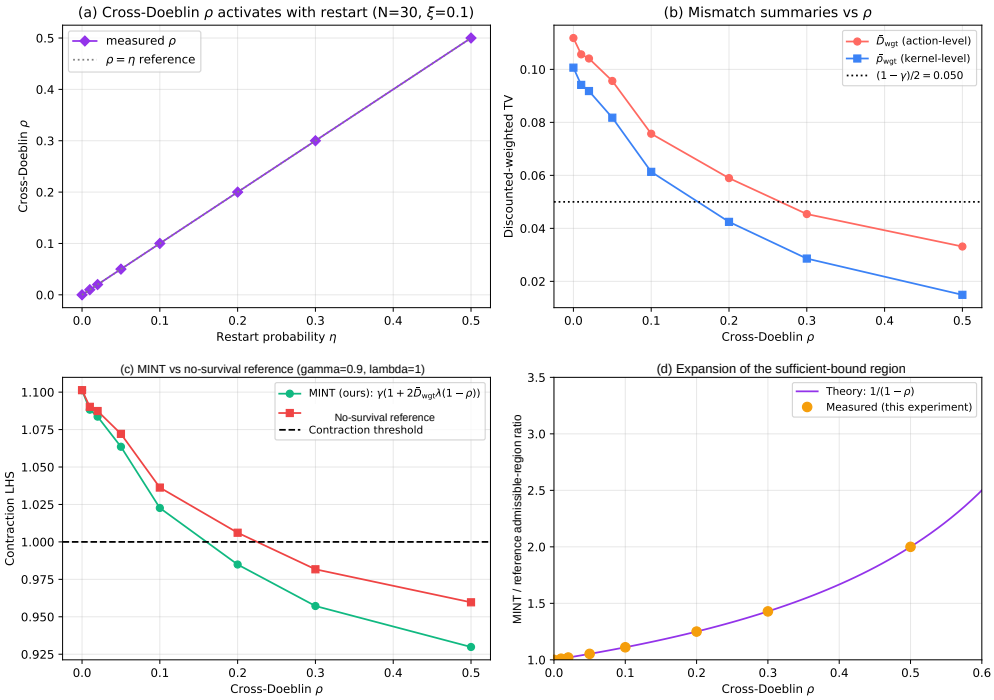


Figure 9: Stage B-2: cross-Doeblin activation and `MINT` super-region on the restart chain (N=30, γ=0.9, ξ=0.1). **(a)** Measured ρ matches restart probability q_{rst} exactly. **(b)** The plotted policy-TV and kernel-TV diagnostics decrease with ρ. Only the policy-TV quantity enters the stated contraction bound. **(c)** `MINT`'s contraction LHS (green) drops below 1 at $q_{\mathrm{rst}} = 0.20$, while the no-survival reference LHS (red) does not until $q_{\mathrm{rst}} = 0.30$. **(d)** Empirical region ratio matches the theoretical $1/(1-\rho)$ to three decimal places.

of 20,000 trajectories. The anchor is always retained, $W_0 = 1$, and $W_t = \prod_{i=1}^{t} B_i$ for continuation indicators B_i. We report the mean retained horizon and the sampling variance of

$$\widehat{\Delta}_t = \sum_{u=0}^{t-1} (\gamma\lambda)^u W_u \delta_u^{Q^\pi}$$

at $t \in \{1, 5, 10, 20, 50\}$. The variance plots concern evaluation at Q^π; the separate perturbed-critic experiment below measures sampled backup MSE.

Stopping rules and controls.

- **Action-level surrogate:** draw fresh $B_i \sim \mathrm{Bernoulli}(h(s_i, a_i))$, where $h = 1 - \min(1, \pi/\beta)$ is the keep probability in Equation (10). Set $\hat{T} = \inf\{i \geq 1 : B_i = 0\}$. With no additional cap we retain $\min(50, \hat{T})$ terms; the capped variant retains $\min(10, \hat{T})$ terms.
- **Constant-p controls:** draw independent $B_i \sim \mathrm{Bernoulli}(p)$, with p denoting keep probability. The sweep uses $p \in \{0.1, 0.5, 0.9\}$. Separate rate-matched controls choose p so that $\sum_{u=0}^{49} p^u$ matches either the surrogate or the state-coupled reference horizon. The two matching targets are distinguished in the legends.
- **State-coupled reference ("oracle" in the legends):** use the first-agreement survival mask from the coupled state processes. This is a reference requiring state-coupling information, not a minimum-MSE estimator or an upper bound on practical performance.
- **Symmetric log-ratio heuristic ("symmetric TV" in the legends):** use

$$h_{\mathrm{sym}}(s, a) = \mathrm{clip}\left(\tfrac{1}{2} \left|\log \pi(a \mid s) - \log \beta(a \mid s)\right|, 0, 1\right) \tag{62}$$

as the keep probability. This heuristic is not an exact TV estimator, and the surrogate contraction theorem is not asserted for this alternative rule.

A: Comparison with horizon-matched random truncation. Figure 10 (a,d) includes controls matched separately to the surrogate and to the state-coupled reference. In these state-independent policy-mismatch settings, the surrogate does not show a consistent advantage over its horizon-matched constant-p control across variance and backup MSE. Thus these tests do not support a general claim that the surrogate improves on random truncation at a matched horizon. They also do not establish equivalence of the estimators: the surrogate mask depends on the sampled action, whereas the constant-p mask is independent of that action. A setting with constant D_s provides limited evidence about benefits from adapting to statewise variation in policy mismatch.

B: Saturation of the symmetric heuristic. For the opposing policies here, $\frac{1}{2}|\log(\pi/\beta)| = \frac{1}{2}\log 19 \approx 1.472$ at $D_s = 0.90$, and $\frac{1}{2}\log 199 \approx 2.647$ at $D_s = 0.99$, for either action. Hence Equation (62) is identically one: the symmetric heuristic retains the full 50-step window. In the representative setting its variance is approximately 0.160, versus 0.129 for the surrogate, and its backup MSE is higher (Figure 10 (b,d)). The variance curves are closer at $D_s = 0.99$ (Figure 11), but a universal within-5% claim is not supported. This comparison diagnoses saturation of this particular heuristic, rather than proving that one clipping form is uniformly superior.

C: The cap reduces horizon and sampled backup MSE. For a state-independent pair $\beta(+1) = b$, $\pi(+1) = q > b$, the surrogate has

$$h(+1) = 0, \qquad h(-1) = \frac{q-b}{1-b}, \qquad \mathbb{E}_\beta[h \mid s] = q - b = D.$$

Conditional on the pre-action history, the mean keep probability is therefore always D. Iterating the survival recursion gives

$$\Pr(\hat{T} > u) = D^u, \qquad \mathbb{E}[\min(L, \hat{T})] = \sum_{u=0}^{L-1} D^u = \frac{1 - D^L}{1 - D}. \tag{63}$$

At $D = 0.90$, the analytic horizons are 9.948 for $L = 50$ and 6.513 for $L = 10$, consistent with the measured 10.0 and 6.52. At $D = 0.99$, they are 39.499 and 9.562, consistent with 39.5 and 9.57. The probabilities of reaching the cap are $\Pr(\hat{T} \geq 10) = D^9$, namely 38.7% and 91.4%, consistent with the insets. These marginal survival identities do not make the surrogate independent of the trajectory.

For backup MSE, set $Q_0 = Q^\pi + \xi$ with $\xi \sim \mathcal{N}(0, I)$ over state–action pairs. Each of 20 perturbation seeds fixes one ξ shared across the trajectory replicates and variants. We measure

$$\mathrm{MSE}_{\mathrm{backup}} = \mathbb{E}\Big[\big(\widehat{\mathcal{R}}Q_0(s_0, a_0) - Q^\pi(s_0, a_0)\big)^2\Big], \tag{64}$$

where $\widehat{\mathcal{R}}Q_0$ is the sampled finite-window TD-error backup. The table reports mean $\pm$ standard deviation across perturbation seeds. Figure panels (a)–(c) show ± 1 standard deviation of the variance estimates across 20 trajectory replicates; panel (d) shows bootstrap percentile 95% confidence intervals for the mean across the 20 perturbation seeds. The perturbation scale in panel (d) is $\sigma = 1$, distinct from the slippery-transition probability.

Table 7: High-mismatch surrogate and cap comparison at a 50-step evaluation window. Backup MSE is mean $\pm$ SD over 20 perturbation seeds; variance is measured at Q^π. No additional cap means that the trace is still limited by the 50-step window.

D_s	Variant	K_{eff}	Variance	Backup MSE
0.90	Surrogate, no additional cap	10.0	0.129	13.8 ± 12.1
	Surrogate, cap 10	6.52	0.128	9.5 ± 6.5
0.99	Surrogate, no additional cap	39.5	0.143	40.8 ± 46.5
	Surrogate, cap 10	9.57	0.141	16.1 ± 12.6

In the representative $D_s = 0.90$ case, cap 10 lowers mean backup MSE from 13.8 to 9.5 (31.2%), while variance changes only from 0.129 to 0.128 (about 0.8%). In the $D_s = 0.99$ comparison, it lowers mean backup MSE from 40.8 to 16.1 (60.5%), while variance changes from 0.143 to 0.141

(about 1.4%). The mean MSE decrease is accompanied by lower MSE in 16 of the 20 reported perturbation seeds in the latter setting, with a median paired capped/uncapped ratio of 0.49. These observations support the cap's role in controlling the sampled backup under the tested mismatch and perturbation distribution, not a claim of large variance reduction or an optimal choice of cap.

The state-coupled reference remains long: its measured horizons are 40.4 and 44.5, with backup MSE 29.2 ± 32.4 and 41.7 ± 48.3 at $D_s = 0.90$ and 0.99, respectively. Thus the cap does not match the surrogate horizon to the reference horizon. The smaller MSE of a much shorter capped backup does not establish an ordering between the state-coupled and surrogate contraction conditions. More aggressive constant-p truncation can yield still smaller MSE in this perturbation experiment, as shown in panel (d); the experiment does not measure the credit-assignment benefit of long backups in learned-critic training.

Scope of the MSE comparison. For a fixed Q_0 and starting pair, the sampled MSE decomposes as

$$\mathbb{E}[(\widehat{\mathcal{R}}Q_0 - Q^\pi)^2] = (\mathcal{R}Q_0 - Q^\pi)^2 + \mathrm{Var}(\widehat{\mathcal{R}}Q_0), \qquad \mathcal{R}Q_0 = \mathbb{E}[\widehat{\mathcal{R}}Q_0].$$

It therefore combines expected-backup error with sampling variance. A reduction in this quantity does not prove sup-norm contraction of the expected operator or convergence of repeated updates. These are stress tests of fixed-policy estimators; they do not supply a contraction certificate for the deep-RL runs. Only the new $\sigma = 1$ perturbation results are used here; no claim of robustness across perturbation scales is made from these figures.

Additional comparison at $D_s = 0.99$. Figure 11 uses the more extreme opposing policy pair. It produces a much longer surrogate horizon and a larger gap between capped and uncapped backup MSE, while retaining the same qualitative conclusions about the matched-random comparison and cap-induced variance change. The comparisons are within this policy setting; they do not identify the absolute variance or MSE of different policy settings with one another.

Reproducibility. The rerun and plotting entry points are `scripts/single_trap_ablation.py` and `paper_figures/ablation_single_trap.py`. The $D_s = 0.99$ run is identified by `outputs/single_trap_ablation/mismatch_b0.005_q0.995.json`; the policy pair stated above identifies the representative $D_s = 0.90$ setting. Trajectory-length curves and stopping-time histograms are taken from these high-mismatch reruns, rather than from the localized-mismatch refinement data.

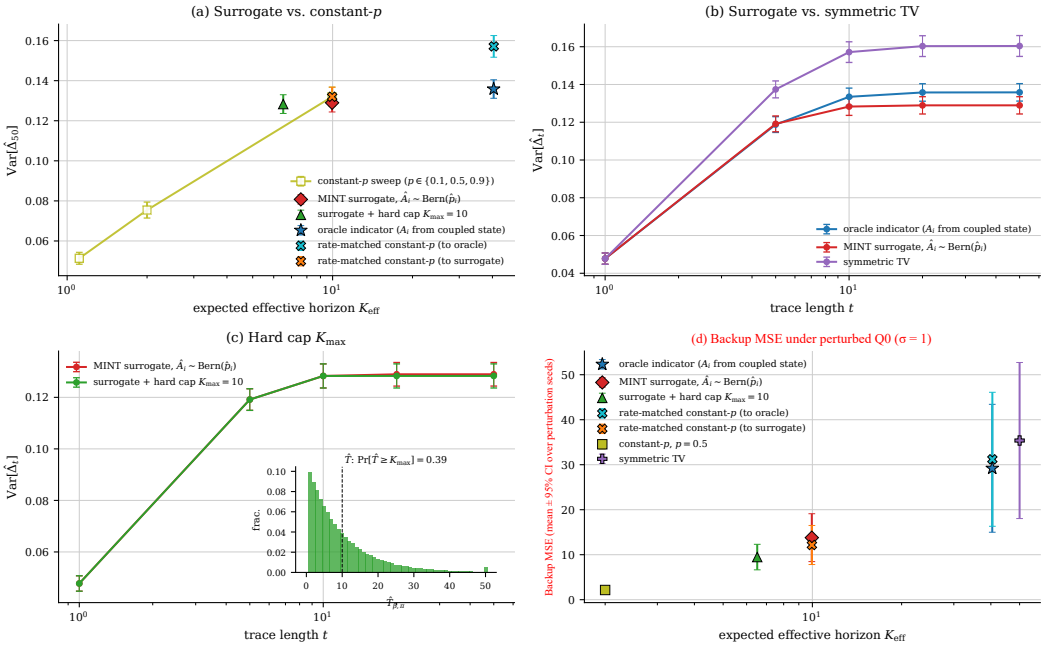


Figure 10: **Representative high-mismatch ablation:** $D_s = 0.90$**.** Policies are $\beta(+1 \mid s) = 0.05$ and $\pi(+1 \mid s) = 0.95$ at every state. **(a)** Variance versus mean retained horizon, including random controls matched separately to the surrogate and state-coupled reference. **(b)** The symmetric log-ratio heuristic saturates and retains all 50 steps; the surrogate has lower variance here. **(c)** Cap 10 reduces the measured horizon from 10.0 to 6.52 with little change in variance. The inset displays the right-censored stopping time $\min(\hat{T}, 50)$; the final bin includes $\hat{T} \geq 50$, and about 39% of traces reach cap 10. **(d)** Sampled backup MSE at $Q_0 = Q^\pi + \xi$, $\xi \sim \mathcal{N}(0, I)$; the cap lowers the mean from 13.8 to 9.5. Error bars: (a)–(c), ± 1 SD across 20 trajectory replicates; (d), bootstrap percentile 95% CI for the mean across 20 perturbation seeds. These MSEs are not expected-operator contraction measurements.

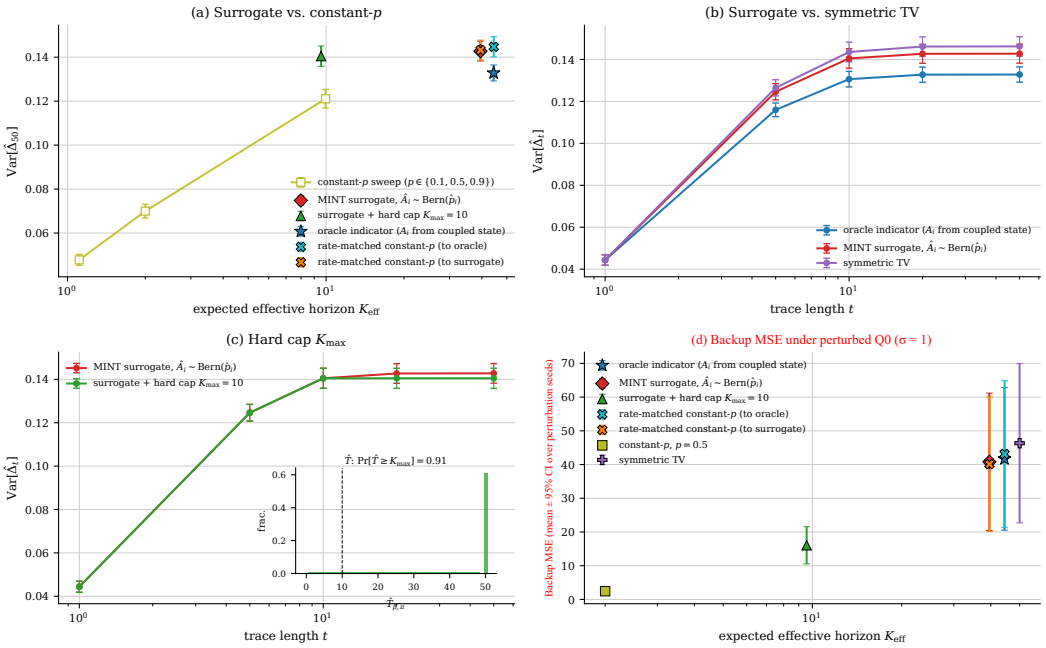


Figure 11: **Additional high-mismatch comparison:** $D_s = 0.99$**.** Policies are $\beta(+1 \mid s) = 0.005$ and $\pi(+1 \mid s) = 0.995$ at every state; panel definitions and error-bar conventions match Figure 10. Cap 10 reduces the measured surrogate horizon from 39.5 to 9.57 and mean backup MSE from 40.8 to 16.1, while variance changes only from 0.143 to 0.141. About 91% of traces reach cap 10. The inset is right-censored at 50, so the final bin includes $\hat{T} \geq 50$. The state-coupled reference has horizon 44.5 and is not a minimum-MSE benchmark. Neither the matched-random controls nor the closeness of two variance curves establishes a general surrogate advantage or a within-5% bound for the symmetric heuristic.

Algorithm 2 `MINT` multi-step target construction — ordinary-residual form matching Algorithm 1; SAC/TD3 substitutions are specified below.

Require: target policy π, current critic Q_θ, target critic $Q_{\bar{\theta}}$, sampled trajectory $(s_0, a_0, r_0, s_1, a_1, r_1, \ldots, s_H)$ from replay $\mathcal{B}$ together with the behavior log-densities $\log \beta(a_i \mid s_i)$ recorded at collection time (so β is the data-generating policy of that trajectory, not the current actor), discount γ, decay $\lambda \in [0, 1]$, hard cap $K_{\max}$ (a protocol constant, fixed to 5 in the main continuous-control comparisons).

Ensure: multi-step target $\hat{y}$ for (s_0, a_0).

[Disagreement estimate] For $i = 1, \ldots, K_{\max} - 1$, compute

$$\hat{p}_i \leftarrow 1 - \min\big(1,\ \pi(a_i \mid s_i)/\beta(a_i \mid s_i)\big) \quad \text{(action-level surrogate, Equation (10)).}$$

[Sample stochastic mask] Draw $\hat{A}_i \sim \text{Bernoulli}(\hat{p}_i)$ independently for $i = 1, \ldots, K_{\max} - 1$ using fresh randomness.

[Effective horizon] Let $\hat{T} \leftarrow \inf\{1 \le i < K_{\max} : \hat{A}_i = 0\}$, with $\inf \varnothing = \infty$; set $\hat{K} \leftarrow \min(K_{\max}, \hat{T})$.

[Trace-mask product] For $t = 0, 1, \ldots, \hat{K} - 1$, define $w_t \leftarrow (\gamma\lambda)^t \prod_{i=1}^{t} \hat{A}_i$, so $w_t = (\gamma\lambda)^t$ until step $\hat{T} - 1$ and $w_t = 0$ afterwards.

[Bellman residuals] For $t = 0, \ldots, \hat{K} - 1$, compute the Bellman residual δ_t^Q (cf. Equation (1))

$$\delta_t \leftarrow r_t + \gamma\, \mathbb{E}_{a' \sim \pi(\cdot \mid s_{t+1})} Q_{\bar{\theta}}(s_{t+1}, a') - Q_{\bar{\theta}}(s_t, a_t).$$

[Multi-step target]

$$\hat{y} \leftarrow Q_{\bar{\theta}}(s_0, a_0) + \sum_{t=0}^{\hat{K}-1} w_t\, \delta_t.$$

[Critic update] Take a gradient step on $(Q_\theta(s_0, a_0) - \hat{y})^2$; refresh $Q_{\bar{\theta}}$ and update π as in the base actor–critic algorithm.

E Implementation Details

In this section, we describe the full implementation details of `MINT`. Following the standard practice in off-policy RL, we use the PyTorch version of the implementations in OpenAI SpinningUp [2].

Experimental Setup We compare `MINT` with four baseline methods, a conventional one-step method, uncorrected multi-step method, Retrace [29] and Peng's $Q(\lambda)$ [23]. Given a randomly sampled trajectory $(s_0, a_0, r_0, s_1, a_1, r_1, s_2, \cdots)$, where Q_{θ^-} denotes the target Q-function, and $\tilde{a}_\phi(s_i)$ is a sample from $\pi_\phi(\cdot|s_i)$. The detailed targets for the Q-function of all algorithms are described in Table 8. Multi-step targets train the critic at the sampled anchor (s_0, a_0). The actor retains its backbone objective at replay states using fresh actor actions; for SAC this maximizes the critic value minus the log-policy term, as specified in Equation (70).

Training and evaluation. For all algorithms, we use $[256, 256]$-sized multi-layer perceptrons (MLPs) for all neural networks. We train with 1M environment steps for OpenAI MuJoCo and evaluate the agent every 1000 steps using a deterministic policy in 10 episodes.

Finite-cap target definitions and backward recursion. All formulas below use a fixed target critic $Q_{\bar{\theta}}$ during target construction. Write $q_i = Q_{\bar{\theta}}(s_i, a_i)$ and $v_i = \mathbb{E}_{a \sim \pi(\cdot|s_i)} Q_{\bar{\theta}}(s_i, a)$. With twin target critics, the fixed function used here can be $Q_{\bar{\theta}} = \min_{j=1,2} Q_{\bar{\theta}_j}$. The stored actions a_i and fresh target-policy actions in v_i have different roles. In particular, v_i is an expectation at s_i, not a maximization over the current state.

Let $L \ge 1$ be the available deterministic cap and let c_i exclude λ, as in Equation (1). Set

$$b_0 = 1, \qquad b_t = \prod_{j=1}^{t} c_j, \qquad \delta_i = r_i + \gamma v_{i+1} - q_i.$$

The finite-cap TD-error target is

$$Y_L(c) = q_0 + \sum_{i=0}^{L-1} (\gamma\lambda)^i b_i \delta_i. \tag{65}$$

For MINT, $c_i = \hat{A}_i$, so $b_i = \mathbf{1}\{\hat{T} > i\}$ and terms after $\hat{K} - 1$, with $\hat{K} = \min(L, \hat{T})$, vanish. Thus Equation (65) is exactly the target in the main-text Algorithm 1 and the full Algorithm 2; it includes the anchor q_0 and uses λ exactly once per continuation step.

An equivalent backward computation is

$$\begin{aligned} G_{L-1} &= r_{L-1} + \gamma v_L, \\ G_i &= r_i + \gamma v_{i+1} + \gamma\lambda c_{i+1}(G_{i+1} - q_{i+1}), \qquad i = L-2, \dots, 0, \\ Y_L(c) &= G_0. \end{aligned} \tag{66}$$

To verify the identity, subtract q_i from the recursion and substitute backward: each continuation contributes a factor $\gamma\lambda c_{i+1}$, giving Equation (65). For MINT one may start instead at $G_{\hat{K}-1} = r_{\hat{K}-1} + \gamma v_{\hat{K}}$ and recurse only over the retained prefix. The boundary uses a bootstrap at $s_{\hat{K}}$ even when the trace ends because of the cap. There is no subtraction of an unsampled q_L at the boundary.

For comparison with reward-return formulas, define $\tilde{c}_i = \lambda c_i$ for $1 \le i < L$ and impose $\tilde{c}_L = 0$. Expanding Equation (65) gives

$$Y_L(c) = \sum_{i=0}^{L-1} \gamma^i \left(\prod_{j=1}^{i} \tilde{c}_j \right) \left[r_i + \gamma\big(v_{i+1} - \tilde{c}_{i+1} q_{i+1}\big) \right]. \tag{67}$$

The term $\tilde{c}_L q_L$ is defined to be zero and need not be evaluated. This terminal convention prevents an extra subtraction at a deterministic cap; it does not remove the final value bootstrap. At $L = 1$, every row below reduces to $r_0 + \gamma v_1$.

Table 8: Consistent finite-cap targets for the ordinary Bellman residual used in the main text. Here L is deterministic; MINT's realized prefix length is $\hat{K} \le L$. The maximum-entropy substitutions are specified below.

Method	Target or continuation rule
One-step	$r_0 + \gamma v_1$.
Uncorrected L-step	$\sum_{i=0}^{L-1} \gamma^i r_i + \gamma^L v_L$.
Retrace	$Y_L(c)$ with $c_i = \min\{1, \pi(a_i \mid s_i)/\beta(a_i \mid s_i)\}$; use Equation (66).
Peng's $Q(\lambda)$	$P_{L-1} = r_{L-1} + \gamma v_L$; $P_i = r_i + \gamma[(1-\lambda)v_{i+1} + \lambda P_{i+1}]$ for $i = L-2, \dots, 0$; return P_0.
MINT	$Y_L(c)$ with $c_i = \hat{A}_i$, $\hat{A}_i \sim \text{Bernoulli}(h(s_i, a_i))$; equivalently $q_0 + \sum_{i=0}^{\hat{K}-1} (\gamma\lambda)^i \delta_i$.

Peng's row is a finite mixture of ordinary multi-step returns [23], not the $c_i \equiv 1$ specialization of Equation (65) with target π. Its expanded form is

$$P_0 = \sum_{i=0}^{L-1} (\gamma\lambda)^i [r_i + \gamma(1-\lambda)v_{i+1}] + (\gamma\lambda)^L v_L. \tag{68}$$

The last term is essential: at $\lambda = 1$ this is the uncorrected L-step return including $\gamma^L v_L$; at $\lambda = 0$ it is one-step. By contrast, $c_i \equiv 1$ in Equation (65) is the finite-cap unweighted TD-error trace with target π.

SAC soft residual and actor update. For SAC [15], replace v_i in the TD-error constructions by

$$v_i^{\text{soft}} = \mathbb{E}_{a \sim \pi_\phi(\cdot \mid s_i)} \left[\min_{j=1,2} Q_{\bar{\theta}_j}(s_i, a) - \alpha_{\text{ent}} \log \pi_\phi(a \mid s_i) \right], \tag{69}$$

and set $\delta_i^{\text{soft}} = r_i + \gamma v_{i+1}^{\text{soft}} - q_i$, with $q_i = \min_j Q_{\bar{\theta}_j}(s_i, a_i)$. An action drawn freshly from π_ϕ gives the usual Monte Carlo estimate of this bootstrap. MINT and Retrace then use the same sum and backward recursion with this soft residual. The stopping rule and cap boundary are unchanged. This is an explicitly entropy-regularized extension; the main-text evaluation results concern the ordinary Bellman residual and Q^π stated there.

For a full maximum-entropy version of the reward-return baselines, intermediate entropy terms must also be included. With $e_i = -\alpha_{\text{ent}} \log \pi_\phi(a_i \mid s_i)$ for stored actions $1 \le i < L$, the corresponding uncorrected soft return is

$$\sum_{i=0}^{L-1} \gamma^i r_i + \sum_{i=1}^{L-1} \gamma^i e_i + \gamma^L v_L^{\text{soft}}.$$

The finite mixture of these soft returns has boundary $P_{L-1}^{\text{soft}} = r_{L-1} + \gamma v_L^{\text{soft}}$ and recursion

$$P_i^{\text{soft}} = r_i + \gamma \left[(1-\lambda) v_{i+1}^{\text{soft}} + \lambda (P_{i+1}^{\text{soft}} + e_{i+1})\right] \quad (i < L-1).$$

These formulas specify full soft-return versions; inserting a soft value only at the end of an ordinary reward sum defines a different target. No entropy of the fixed starting action is added to a Q-target.

The SAC actor is optimized by gradient ascent on

$$\mathbb{E}_{s\sim\mathcal{B},\, a\sim\pi_\phi(\cdot|s)} \left[\min_j Q_{\theta_j}(s,a) - \alpha_{\text{ent}} \log \pi_\phi(a \mid s) \right], \tag{70}$$

using online critics and reparameterized actor actions. The multi-step critic target is detached during critic regression; it is not substituted for the actor's differentiable critic objective. Here α_{ent} is the SAC entropy coefficient, distinct from the discount–trace product $\gamma\lambda$.

TD3 bootstrap and the policy used in the ratio. For TD3 [13], the bootstrap has no entropy term and uses the clipped double target critic at a smoothed target action:

$$\tilde{a}_i = \text{clip}_{\mathcal{A}}\big(\mu_{\bar{\phi}}(s_i) + \text{clip}(\epsilon_i, -c, c)\big), \qquad \epsilon_i \sim \mathcal{N}(0, \sigma_{\text{targ}}^2 I),$$
$$v_i^{\text{TD3}} = \mathbb{E}[\min_j Q_{\bar{\theta}_j}(s_i, \tilde{a}_i)].$$

For a ratio-based trace, the target action law denoted by π must also be specified. If it is the smoothed law above, the ratio must refer to that law, including noise and action clipping; the formula does not justify using an unclipped Gaussian density in its place. If a different stochastic density proxy is used for stopping, it defines a distinct stopping policy and must be distinguished from the bootstrap policy. A deterministic target actor has no ordinary Lebesgue density to insert into a continuous-action ratio. Thus the sampling law and the density used for stopping must be checked in the implementation; no undocumented density proxy is inferred from the TD3 bootstrap. These distinctions do not change the algebraic target identities above.

For true terminal transitions set the subsequent bootstrap and continuation to zero. A deterministic trace cap retains the bootstrap at its last observed next state; a time-limit truncation should not be treated as a true terminal transition solely because the recording window ended.

Methods and Hyperparameters. We use two one-step RL algorithms, SAC and TD3, for the multi-step extension.

1. **Twin-Delayed Deep Deterministic Policy Gradient (TD3).** TD3 [13] adopts the same training pipeline and neural network architecture as DDPG, but introduces several improvements to address overestimation bias in Q-learning. TD3 uses two critic networks, $Q_{\theta_1}(s,a)$ and $Q_{\theta_2}(s,a)$, with independent parameter sets θ_1 and θ_2. This twin-critic design follows double Q-learning [43].
2. **Soft Actor-Critic (SAC).** SAC [15] adopts the same training pipeline and architecture as DDPG/TD3 but introduces a maximum-entropy objective. SAC also maintains two critic networks to reduce overestimation bias.

We adopt the default hyperparameters from the OpenAI SpinningUp codebase. The cap length denotes the upper limit of the sub-trajectory length for the baseline algorithms (uncorrected n-step, Retrace, PQL). Detailed values are reported below.

Table 9: TD3 Hyperparameters

Hyperparameter	**Value**
Actor learning rate	1×10^{-3}
Critic learning rate	1×10^{-3}
Batch size	100
Replay buffer size	1×10^{6}
Discount factor γ	0.99
Polyak averaging coefficient (τ)	0.995
Target policy noise (stddev)	0.2
Target noise clip	0.5
Policy update delay (frequency)	2 steps
Exploration noise (initial stddev)	0.1
Action range	[-1, 1]
Start steps (before training begins)	10000
Max episode length	1000
cap length	5
λ	0.7

Table 10: SAC Hyperparameters (MuJoCo main experiments). The entropy coefficient is *fixed* at α=0.2 (the SpinningUp default) for `MINT` and all multi-step baselines; no automatic entropy tuning is used in these runs. The DMC ablation and RWRL experiments use their separately stated entropy settings; the fixed-α statement here is restricted to the main MuJoCo SAC comparison.

Hyperparameter	**Value**
Actor learning rate	1×10^{-3}
Critic learning rate	1×10^{-3}
Entropy coefficient α	0.2 (fixed; no automatic tuning)
Batch size	100
Replay buffer size	1×10^{6}
Discount factor γ	0.99
Polyak averaging coefficient (τ)	0.995
Target update interval	Every 1 step
Start steps (before policy used)	10000
Action range	[-1, 1]
Max episode length	1000
cap length	5
λ	0.7

Algorithm. The detailed multi-step target construction is given as Algorithm 1 in the main text (Section 3.2). It is a drop-in replacement for the multi-step target in any actor–critic backbone (SAC, TD3); the trace construction computes $\hat{p}_i$, draws the keep masks, and selects the retained prefix. Its critic target is the finite-cap residual sum in Equation (65), with the explicit SAC or TD3 bootstrap substitution above.

Experimental Details. We implement `MINT` and other baselines in PyTorch on top of the standard evaluation protocol of off-policy RL algorithms in Google Dopamine [6]. Code is available at `https://github.com/seungyubhan/mint`.

Numerical treatment of the density ratio. With tanh-squashed Gaussian actors, evaluating $\pi(a_i \mid s_i)/\beta(a_i \mid s_i)$ at a stored action requires three precautions. (i) *Stored behavior density.* $\log \beta(a_i \mid s_i)$ is computed and stored at collection time (during the uniform warm-up it is the log-density of the uniform box distribution), so β is always the data-generating policy of that transition and no density model of the drifting replay distribution is needed (Algorithm 2). (ii) *Squash correction.* The log-density of a stored action under π is evaluated by inverting the squash, $u = \operatorname{atanh}(a/a_{\max})$ with the argument clipped to $[-1 + 10^{-6},\, 1 - 10^{-6}]$ to guard the saturated boundary, and adding the tanh

log-Jacobian; the policy log-standard-deviation is clamped to $[-20, 2]$. (iii) *Log-space ratio.* The clipped ratio is formed as $\min(1, \pi/\beta) = \exp\big(\min(0, \log\pi - \log\beta)\big)$: clamping the log-ratio at 0 is exactly the minimum against 1 and removes overflow, while underflow of the exponential for very negative log-ratios yields $\hat{p}_i \to 1$, the correct limit rather than a NaN. $\hat{p}_i$ is finally clamped to $[0, 1]$ and the anchor step $t = 0$ is always kept. These steps keep the ratio stable when π becomes sharply peaked relative to β (the saturation regime of Appendix G.1).

F Additional Experimental Results

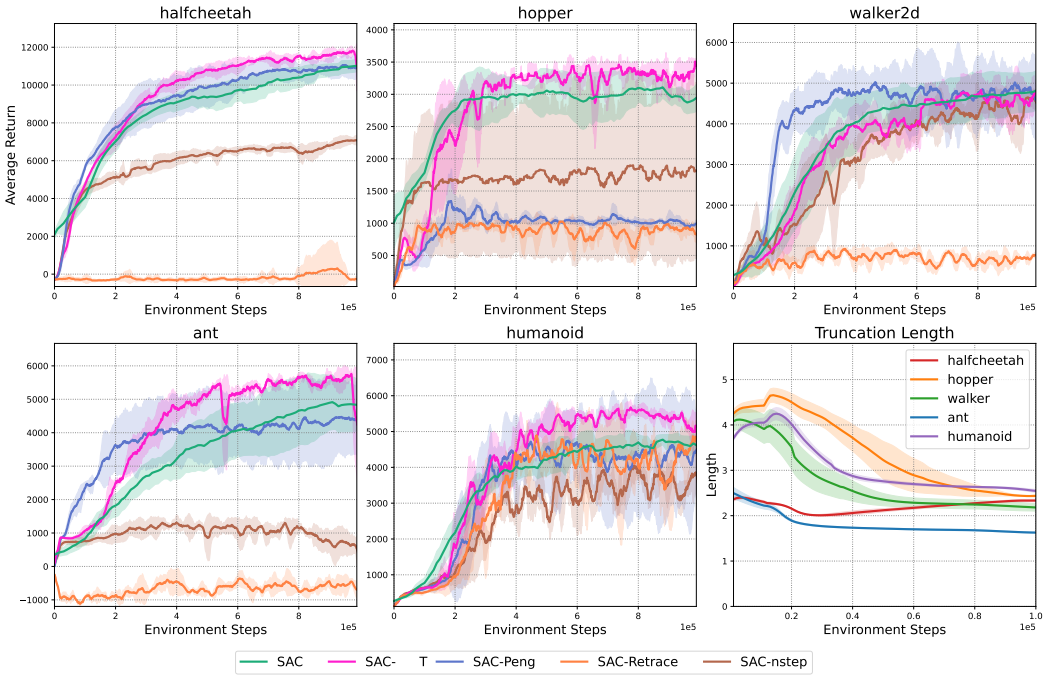


Figure 12: **Performance with stochastic truncation.** Across five MuJoCo benchmarks, `MINT` consistently outperforms multi-step baselines and achieves faster convergence. For baseline comparisons, we follow the convention of [23] and fix the cap length to $n = 5$ for all multi-step methods.

F.1 Evaluation results by the protocol of IQM

We conducted experiments following the evaluation protocol proposed in [3]. This protocol emphasizes robust evaluation through interquartile mean (IQM) scores and normalized performance.

The main MuJoCo comparison uses the five-seed-per-method-and-task protocol stated in Appendix I, with scores aggregated after min–max normalization. Additional experiments report their own seed counts. The steps reported in the table correspond to 1,000K environment interactions. Metric computation was conducted using the official codebase of [3]. For min–max normalization, expert and random scores were taken from the Minari extension [11] of D4RL [12]; the Humanoid benchmark follows Minari scores, while the remaining tasks use D4RL values.

Across both SAC- and TD3-based experiments, `MINT` exhibits faster learning curves than existing multi-step methods and baseline algorithms. Notably, only `MINT` surpasses the expert score within 1,000K steps, whereas prior methods typically require up to 3,000K steps to achieve comparable performance.

F.2 DMControl Sparse-Reward Diagnostics

Figure 15 (left) shows the sparse-reward `CartPole-Balance` comparison. The right panel reports normalized next-state distances on `Hopper` (median $\approx$0.36, 90th percentile $\approx$0.64) as a descriptive

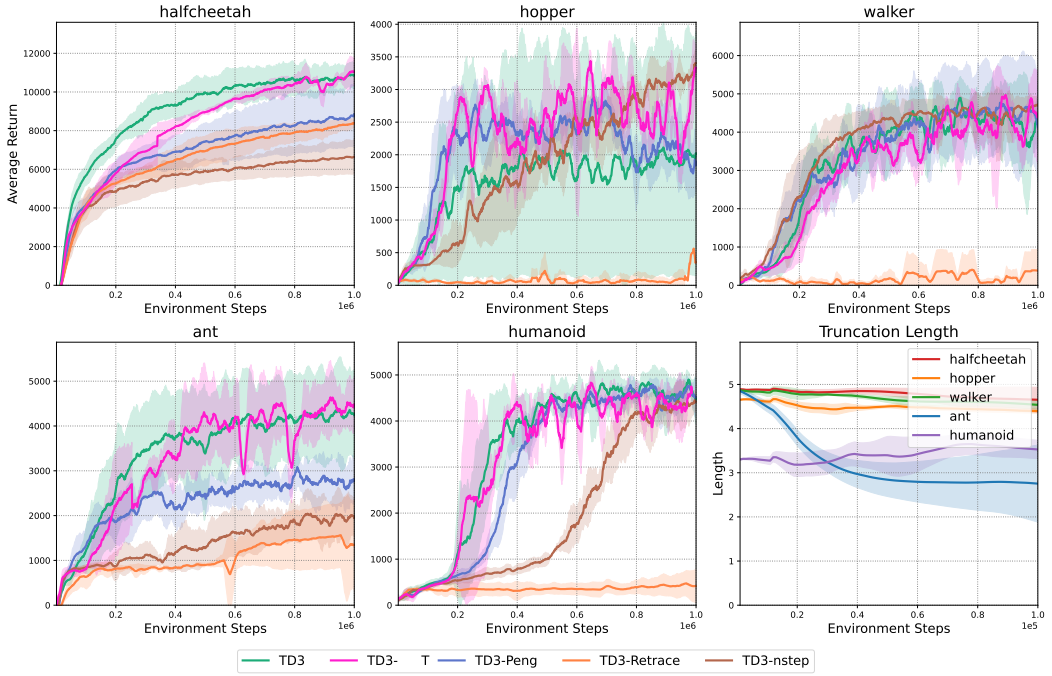


Figure 13: **Evaluation of TD3 variants over OpenAI MuJoCo environments.**

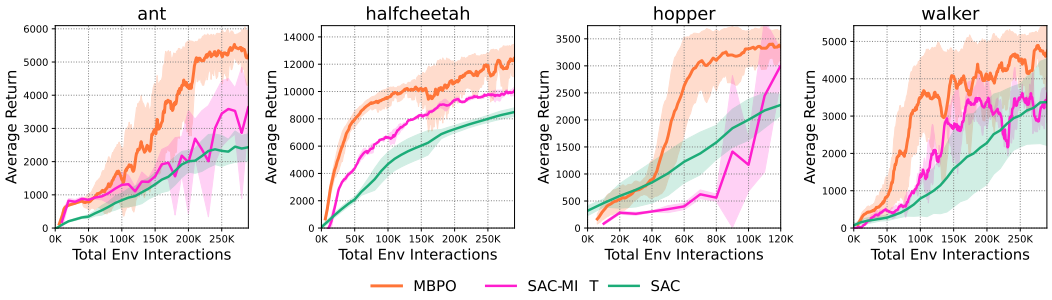


Figure 14: **Comparison of SAC, SAC-MINT, and MBPO across four MuJoCo benchmarks.** We use different multi-step configurations per environment: *halfcheetah* and *walker* use a standard multi-step setting (cap length $= 5$, $\lambda = 0.7$), whereas *ant* and *hopper* follow the longer-horizon model-based configuration inspired by MBPO (cap length $= 25$, $\lambda = 0.1$). Across all tasks, MINT consistently accelerates learning and improves sample efficiency by adaptively adjusting its effective horizon, often matching or approaching MBPO despite being entirely model-free.

transition-similarity diagnostic, rather than a verification of a TV or contraction condition. The one-step surrogate comparison is in Appendix G.2, and the empirical backup-gain probes are in Appendix G.4.

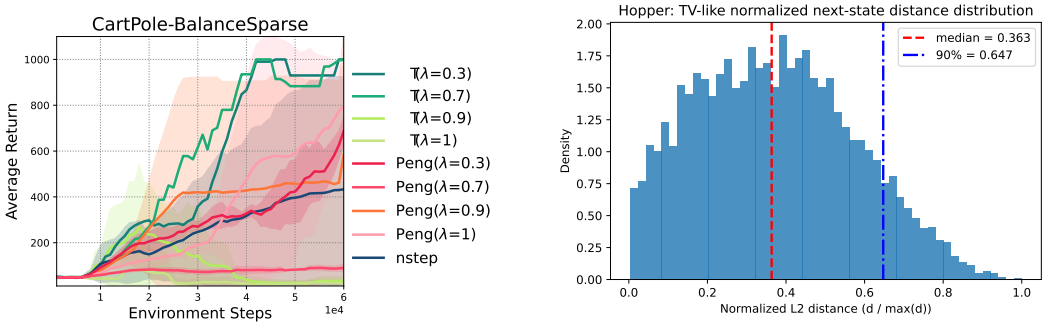
F.3 Ablation: Truncation Kernel vs. Multi-Step Returns

The experiments above demonstrate that MINT outperforms fixed-horizon multi-step methods. A natural question is whether this gain stems from the stochastic truncation rule itself, or merely from extending the backup horizon. We isolate this by comparing three variants under a unified SAC backbone on DMControl tasks:

1. **SAC-MINT**: multi-step targets with the stochastic stopping time $\hat{T}_{\beta,\pi}$ and learned entropy temperature (auto-α).

Table 11: **Reference scores for min–max normalization.** Random and expert performance values are taken from D4RL [12] and Minari [11].

Environment	Random Score	Expert Score
Hopper	−20.27	3234.3
HalfCheetah	−280.18	12135.0
Walker2d	1.63	4592.3
Ant	−325.6	3879.7
Humanoid	78.85	9024.95



(a) (b)

Figure 15: **(Left)** Sparse-reward `CartPole-Balance`: `MINT` outperforms fixed-horizon baselines across the displayed λ values. **(Right)** Normalized next-state distances on `Hopper` (median $\approx$0.36, 90th percentile $\approx$0.64), used as descriptive dynamics diagnostics; they do not certify the surrogate contraction condition of Theorem B.9.

2. **SAC-multistep**: the corresponding multi-step configuration with the same horizon L and λ, but without stochastic trace truncation. The TD-error trace and Peng return mixture are distinguished in Table 8.
3. **Vanilla SAC**: one-step TD with auto-α.

All methods share architecture, replay ratio, optimizer, and hyperparameters; only the backup operator differs (Appendix E).

Setup. We evaluate on five DMControl tasks spanning sparse and dense reward regimes: `CartPole-Swingup-Sparse` (sparse, primary discriminator), `Hopper-Stand` (dense, with a published SAC reference), `Finger-Turn-Easy` (sparse-ish, both methods learn), `Acrobot-Swingup-Sparse` (hard sparse, beyond SAC-class), and `Ball-in-Cup-Catch` (saturating, non-discriminative). Each condition uses 3 random seeds for 1M environment steps; we report the mean evaluation return over the last 100k steps (mean_{10}). Seed variance is high on the sparse `CartPole-Swingup-Sparse` task: a 5-seed re-run of SAC-`MINT` at L=10 gives 617 ± 113, which preserves the ordering against SAC-multistep (24) and vanilla SAC (9) but with a wider spread than the 3-seed L=5 figure in Table 12. SAC-`MINT` uses L=5, λ=0.9 on CartPole and L=10, λ=0.9 on the rest; γ=0.99 on dense `Hopper-Stand` and γ=0.999 on the sparse tasks (matched to vanilla SAC's per-task defaults).

Results: discriminative tasks. Table 12 reports headline returns. On `CartPole-Swingup-Sparse`, SAC-`MINT` achieves $\mathbf{719 \pm 46}$, a $\mathbf{30}\times$ improvement over SAC-multistep (24) and $\mathbf{80}\times$ over vanilla SAC (9 ± 16). On `Hopper-Stand`, SAC-`MINT` reaches $\mathbf{891 \pm 24}$, matching the published SAC reference of $\sim$870 [15], while SAC-multistep collapses to 57 ± 90 — a $\mathbf{16}\times$ gap under identical hyperparameters. On `Finger-Turn-Easy` both multi-step variants solve the task (686 ± 26 vs. 767 ± 62). Their similar returns do not establish that truncation is rare: under the disagreement-as-keep rule, a small policy-TV gap instead gives a small mean keep probability.

Table 12: **DMControl ablation.** Mean evaluation return at 1M environment steps ($\text{mean}_{10} \pm$ std, 3 seeds). Bold marks the best per row. SAC-MINT improves over uncorrected multi-step on CartPole and Hopper; both learn on Finger, saturate on Ball-in-Cup, and fail on the tested Acrobot setting.

Task	SAC-MINT	SAC-multistep	Vanilla SAC
Discriminative			
CartPole-Swingup-Sparse	$\mathbf{719 \pm 46}$	24	9 ± 16
Hopper-Stand	$\mathbf{891 \pm 24}$	57 ± 90	—
Finger-Turn-Easy	686 ± 26	$\mathbf{767 \pm 62}$	—
Scope (non-discriminative)			
Ball-in-Cup-Catch	976 ± 2	—	976 ± 1
Acrobot-Swingup-Sparse	≤ 1	—	—

Scope tasks. `Acrobot-Swingup-Sparse` is unsolved by all SAC-class baselines we tested at 1M steps ($\text{mean}_{10} \leq 1$ for both MINT and multi-step), consistent with this task requiring world-model planning rather than a better backup operator; this sets a hard-exploration boundary outside the scope of trace truncation. On `Ball-in-Cup-Catch` both methods saturate at $\sim$976 alongside vanilla SAC (976 ± 1), confirming the task is non-discriminative. These observations delimit the empirical credit-assignment setting. The operator theorem does not condition its validity on reward discovery or establish convergence of the learned critic.

Interpretation. The reported CartPole and Hopper comparisons show a benefit of the stochastic-truncation configuration under the tested training protocol. They do not follow from Lemma 3.4, which concerns the estimator at Q^π, rather than bias or learning dynamics of a changing critic. In particular, $\beta \approx \pi$ gives $\mathbb{E}_\beta[h \mid s] = D_s \approx 0$, hence frequent stopping and short traces. The similar Finger returns do not establish a causal mechanism without direct trace diagnostics.

Auto-α on sparse-reward tasks. This DMC ablation uses automatic entropy tuning, whereas the main MuJoCo SAC comparisons use fixed $\alpha = 0.2$ for all methods (Table 10). RWRL and the other additional experiments specify their own entropy protocols. Within this ablation we observe that on the sparse-reward `CartPole-Swingup-Sparse` task automatic entropy tuning succeeds where fixed $\alpha = 0.2$ fails (Figure 5b, L=10). With fixed $\alpha = 0.2$, the reported diagnostic shows a diffuse policy before reward is observed, a short effective horizon $\hat{K} \approx 2$, and failed learning. Short traces correspond to frequent zero keep masks, not to $h \approx 1$. Policy entropy by itself does not determine the action-level TV gap to the stored behavior policy. Under auto-α, α settles at 0.06–0.10 once reward is observed, the policy concentrates, $\hat{K}$ rises to $\sim$5.7 out of L=10, and learning succeeds. The increase in effective horizon is an observed property of this auto-α protocol. It cannot be deduced from policy concentration alone: continuation depends on the statewise discrepancy between the target and stored behavior policies. The fixed-α MuJoCo runs constitute a separate protocol. For dense tasks (`Hopper-Stand`) the same regime emerges naturally; for hard-exploration sparse tasks (`Acrobot`), α collapses too quickly before any reward is observed — a separate failure mode unrelated to the truncation operator itself.

F.4 Additional Early-Training Diagnostics on Atari

We report IQM results on `Pong` and `Breakout` at 500K agent steps (update horizon $= 5$, $\lambda = 1$). C51-MINT exhibits higher normalized scores and reduced variance compared to C51, Rainbow, and DQN under this distributional-learning protocol. These results are separate from the DrQ and DER Atari 100k experiments below and do not establish stability under their target-update schedules.

F.5 Discrete-Action Benchmark: Atari 100k

We evaluate MINT on the Atari 100k benchmark [22] (26 games, 100K env-step budget) on top of two backbones: DrQ(ε) [47] and DER [44]. Direct integration into DrQ exposes sensitivity to target-network updates. Increasing the hard synchronization period stabilizes Q-values, but an ablation that removes intermediate $V - Q$ terms while preserving trace lengths also improves returns

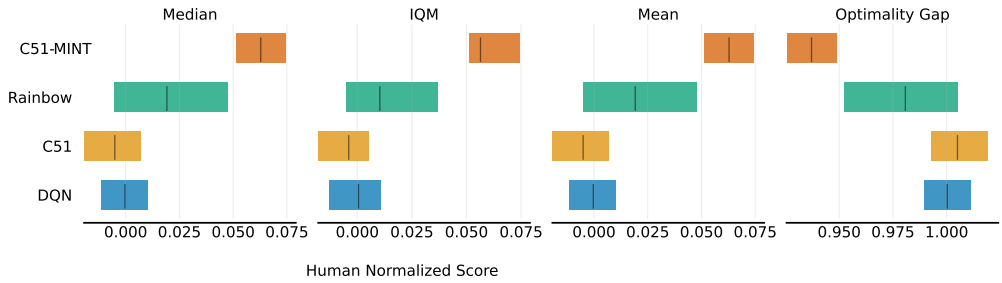


Figure 16: IQM comparison on `Pong` and `Breakout` after 500K agent steps (*update horizon* $= 5$, $\lambda = 1$). Our C51-`MINT` operator achieves clearly higher normalized scores than C51, Rainbow, and DQN, while also exhibiting lower variance.

on two games. We report these controlled observations separately from possible error-amplification mechanisms and from preliminary Boltzmann-target comparisons on the DER backbone.

Setup. We distinguish the DrQ hard-sync diagnostics from the aggregate comparisons and the DER target-policy study. The DrQ baseline uses 10-step returns and hard target synchronization every step; the controlled correction ablation changes MINT's synchronization period to 100 and uses three seeds each on `Seaquest` and `Qbert`. The completed aggregate comparisons use five seeds on each of 26 games (130 trials per algorithm); the reported MINT aggregate variants use Polyak target updates (τ_{Polyak}=0.005), $K_{\max}$=10, and λ=0.7, rather than the per-step hard-sync setting. The DER-based comparisons use a scalar-Q backbone with a data-efficient CNN, dueling head, NoisyLinear, and Double DQN; the fixed-horizon DER baseline uses n=20. We use γ=0.99 and an ε-greedy behavior policy, annealed from 1.0 to 0.01 over the first 5K steps. The Boltzmann-target DER+`MINT` results use the Polyak protocol and are preliminary (1–2 seeds on most games), not a completed five-seed comparison. Backbone, target-update schedule, and seed count therefore differ across these experiment groups. Comparison is against published baselines: DER [44], DrQ(ε) [47], SPR [35], BBF [36], and EfficientZero [48]. Aggregate metrics follow the `rliable` protocol [3] with stratified bootstrap CIs (10K resamples).

Intermediate correction terms. Let Q denote a fixed target critic and $V^\pi(s) = \sum_a \pi(a \mid s)Q(s,a)$. With consistent critic and augmentation evaluations of repeated states, the target for a realized trace length K can be written as

$$
\begin{aligned}
Y_{\text{MINT}} &= G_K^\lambda + \sum_{i=1}^{K-1} (\gamma\lambda)^i \Delta_i, \\
G_K^\lambda &= (1-\lambda) \sum_{n=1}^{K-1} \lambda^{n-1} G_n + \lambda^{K-1} G_K, \\
G_n &= \sum_{t=0}^{n-1} \gamma^t r_t + \gamma^n V^\pi(s_n), \qquad \Delta_i = V^\pi(s_i) - Q(s_i, a_i).
\end{aligned}
\tag{71}
$$

Terminal transitions zero the corresponding bootstrap. Thus G_K^λ is a truncated λ-return, reducing to a K-step return at $\lambda = 1$; the intermediate Δ_i terms distinguish the TD-error target from that return. For a greedy target evaluated with the same Q-values, $\Delta_i = \max_a Q(s_i, a) - Q(s_i, a_i) \geq 0$, and the disagreement-as-keep rule retains non-greedy actions. Nonnegativity alone does not establish overestimation relative to true values or an amplification factor above one.

Observed instability under per-step hard synchronization. With target parameters copied from the online network every step, Double DQN loses its selection–evaluation separation when both evaluations use the same inputs and randomness. DrQ's ordinary 10-step target contains a single terminal bootstrap weighted by $\gamma^{10} \approx 0.904$, whereas MINT also includes the intermediate terms in Equation (71). The later diagnostic sweep records divergence in all 36 tested runs. The retained intermediate terms were positive in 100% of the measured cases; their reported magnitudes relative

to the Q-value scale were 13–22% at $\lambda = 0.7$ and up to 50% at $\lambda = 1$. Early diagnostics over the first 100 updates recorded Q-value growth of approximately 6–11% per update. These observations support a feedback-amplification explanation under this protocol; they do not establish that ordinary n-step structure alone guarantees stability.

An earlier, separate six-game pilot also found divergence on five games with per-step hard synchronization (final Q-loss 10^3–10^{14} and Q-values reaching 10^9–10^{14}). In the Polyak aggregate runs, observed final losses remained within 0.02–1.39 across 26 games. These are observations under different update schedules, not a stability guarantee or evidence that stability is independent of target-policy choice.

Controlled ablation after stabilizing Q-values. Increasing the hard synchronization period to 100 stabilized Q-values at the DrQ scale, but returns remained below the DrQ baseline. Removing only the intermediate Δ_i terms improved returns on both tested games (Table 13); trace lengths were unchanged. This implicates these terms in the remaining performance loss under the stabilized protocol, rather than attributing that loss solely to short traces. The modified target exceeds DrQ on `Seaquest`, but remains far below it on `Qbert`; this is not evidence of a benchmark-wide improvement. Term removal is a diagnostic modification and does not generally preserve MINT's off-policy fixed-point guarantee.

Table 13: **DrQ correction-term ablation on Atari 100k.** Reported mean returns over three seeds per game. Both MINT variants use hard synchronization period 100 and unchanged trace lengths; the DrQ baseline retains its original update schedule. The final column removes only the intermediate $V - Q$ terms.

Game	DrQ baseline	MINT	MINT without intermediate $V - Q$
Seaquest	354	242	562
Qbert	2118	454	611

Possible pathways and unresolved causes. At an exact Q^π, the TD-error operator retains its fixed-point property; this does not imply harmless corrections during learning with an approximate critic. Max-selection bias and action-dependent estimation error could make the intermediate terms excessive. Independently augmenting repeated evaluations of the same state also adds discrepancies to the consistent-evaluation decomposition above; at hard-sync period 10, changing the augmentation protocol changed whether divergence occurred. This shows augmentation sensitivity, but does not isolate its contribution to the return loss at period 100. The relative roles of augmentation noise and action-dependent bias remain unresolved. A small final exploration rate does not by itself show that replay is nearly on-policy, since stored actions were collected by earlier policies. Likewise, the source of the `Seaquest` gain after term removal—truncation, λ mixing, or the changed target-update period—has not been separated.

Target-policy and trace diagnostics. The keep mask is $\hat{A}_i \sim \mathrm{Bernoulli}(h(s_i, a_i))$. The inequality $\beta(a_i \mid s_i) > \pi(a_i \mid s_i)$ makes the keep probability positive; it does not make the realized mask deterministically one. If π is deterministic greedy and β is ε-greedy with the same greedy action g, over m actions with a uniform exploration component, then

$$h(s, g) = 0, \qquad h(s, a) = 1 \ (a \neq g), \qquad \mathbb{E}_{a\sim\beta}[h(s, a)] = \varepsilon(1 - 1/m).$$

At small ε, most sampled actions are greedy and stop the trace; the rare non-greedy actions keep it. If the greedy actions differ, the mean keep probability is instead $1 - \varepsilon/m$, so a greedy target alone does not imply a short trace. This distinction matters for stored behavior policies and changing target networks. The reported near-one-step diagnostic is an observation for its specified policy/noise protocol, rather than a consequence of $h \approx 1$ or of Lemma 3.4. The relevant survival is that of the surrogate, $\Pr(\hat{T}_{\beta,\pi} > t)$.

Exploratory Boltzmann targets. We replace the ε-greedy π with the Q-aware Boltzmann form $\pi(a \mid s) = \mathrm{softmax}(Q(s, a)/\tau)$, keeping the disagreement-as-keep rule $h = 1 - \min(1, \pi/\beta)$ unchanged. For a unique maximizer of Q, $\tau \to 0$ gives a deterministic greedy target; its trace length depends on alignment with the behavior policy. As $\tau \to \infty$, the target becomes uniform,

which does not imply absence of truncation: if the behavior is also uniform, $h = 0$ and the trace stops immediately. Temperature alone does not determine the survival law. Figure 17 (b) reports the synthetic NoisyLinear-behavior diagnostic: the effective trace length is approximately 1 at $\tau = 0$, 10 at $\tau = 0.1$, 17 at $\tau = 0.5$, and 20 for larger temperatures. These are protocol-specific diagnostic values with cap 20, distinct from the deployed Atari cap 10. We select $\tau = 0.1$ using an exploratory four-temperature sweep on `Pong` and `Freeway`; this is a tuned target-policy choice, not a necessary condition established by the theory.

Hypothesis: reduced estimation-error amplification. For a fixed critic, Boltzmann averaging gives

$$V_\tau(s) = \sum_a \pi_\tau(a \mid s) Q(s,a) \leq \max_a Q(s,a).$$

Relative to a hard-max target, it therefore lowers the bootstrap value and the intermediate $V - Q$ terms for the same state–action sequence. One possible explanation for the preliminary gains is attenuation of overestimation-driven feedback. This inequality does not establish lower error relative to true values: Boltzmann also changes the evaluated policy and the continuation masks, and these effects have not been isolated. Policy averaging is distinct from hard copying or Polyak averaging of network parameters; the reported Boltzmann runs use Polyak updates. Neither these results nor the synthetic trace diagnostic establish that Boltzmann targets are necessary for MINT.

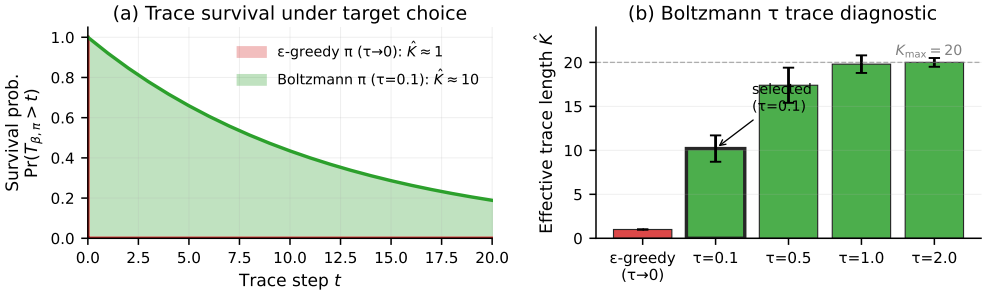


Figure 17: **Target-policy and trace diagnostics.** (a) Reported surrogate survival $\Pr(\hat{T}_{\beta,\pi} > t)$ under the diagnostic's greedy and Boltzmann target choices. (b) Synthetic NoisyLinear-behavior diagnostic with cap 20: the reported effective lengths at $\tau \in \{0, 0.1, 0.5, 1.0, 2.0\}$ are approximately $\{1, 10, 17, 20, 20\}$. These observations concern the diagnostic's behavior/noise protocol; they are not a general implication of target temperature. The deployed Atari cap is 10.

Exploratory ε-greedy target softening. A natural alternative is to soften π by raising the target ε (denoted target_eps), keeping π in the same family as β. We tested target_eps$\in\{0.0, 0.3\}$ on eight games with five seeds each (Table 14). Increasing target_eps to 0.3 raises the reported means on `MsPacman` (691→972), `PrivateEye` (4→60), `Kangaroo`, and `Jamesbond`, but lowers `Freeway` (2.4→0.8, −67%). A uniform exploration component and Boltzmann averaging assign probabilities differently across the Q-landscape, but these comparisons do not isolate which difference drives performance. The partial Boltzmann results improve on the greedy-target variant in some games and lose in others; they do not establish a general resolution of the observed trade-off.

Results: aggregate metrics. Table 15 reports human-normalized scores (HNS) aggregated over 26 games × 5 seeds. With the ε-greedy target, both `MINT` (DrQ(ε) backbone, IQM 0.066) and DER+`MINT` (ε^π=0, IQM 0.084) underperform the DER baseline (IQM 0.192 ours, 0.183 published). The partial Boltzmann sweep ($\tau = 0.1$) reports gains over the $\varepsilon^\pi = 0$ variant of 33% (`KungFuMaster`), 46% (`MsPacman`), 50% (`UpNDown`), 53% (`Gopher`), 72% (`RoadRunner`), and 166% (`Kangaroo`); reported means on `Krull`, `Qbert`, `Pong`, and `Breakout` are similar. Selected raw scores are in Table 14; Figure 19 displays only the completed greedy-target comparisons. These preliminary differences are against the greedy-target MINT variant, not gains over DER or DrQ. The partial sweep also reports losses on `Hero`, `Seaquest`, and `Jamesbond`. Full five-seed Boltzmann aggregate metrics are unavailable, so Table 15 leaves that row unestimated.

Summary. The DrQ diagnostics establish target-update sensitivity and a two-game return improvement from removing intermediate $V - Q$ terms at unchanged trace lengths. They do not identify

Table 14: **Target-policy comparison** on eight Atari 100k games (raw final evaluation). The ε-greedy variants use five seeds and show mixed effects of target softening. Boltzmann results are preliminary (1–2 seeds); differences do not identify a causal mechanism.

Game	**ε-greedy, ε^π=0.0**	**ε-greedy, ε^π=0.3**	**DER (ours)**	**Boltzmann τ=0.1 (ours)**
Freeway	2.4 ± 3.1	0.8 ± 0.8	28.3	**21.3** †
Frostbite	297 ± 133	238 ± 17	**1898**	—
Hero	$\mathbf{2372 \pm 1190}$	2033 ± 1178	6736	1494 †
Jamesbond	61 ± 29	77 ± 42	272	29 †
Kangaroo	210 ± 180	276 ± 266	950	558 †
MsPacman	691 ± 322	972 ± 80	1271	1007 †
PrivateEye	4 ± 17	60 ± 42	67	0 †
Qbert	637 ± 59	651 ± 137	1753	734 †

† Boltzmann τ=0.1 partial sweep (1–2 seeds at the time of writing); full 5-seed numbers will replace these after the in-flight sweep completes.

Table 15: **Atari 100k aggregate metrics.** Completed in-house comparisons use 26 games $\times$ 5 seeds (`rliable`, 10K stratified bootstrap); published baselines are listed separately. Higher is better for Median/IQM/Mean; lower is better for Optimality Gap (OG). Bold marks the best completed in-house result. The Boltzmann $\tau = 0.1$ sweep has only partial seeds, and no aggregate estimate is reported.

Algorithm	**Median**	**IQM**	**Mean**	**OG (↓)**
Ours, ε-greedy target (π greedy)				
`MINT` (DrQ(ε) backbone)	0.039	0.066	0.244	0.811
DER + `MINT` (ε^π=0)	0.056	0.084	0.239	0.812
DER (ours, baseline)	**0.180**	**0.192**	**0.317**	**0.703**
Ours, Boltzmann target π=softmax(Q/τ), τ=0.1 (in-flight, partial seeds)				
DER + `MINT` (Boltzmann τ=0.1)	*not estimated; partial scores in Table 14*			
Published baselines				
DER [44]	0.161	0.183	0.285	0.730
DrQ(ε) [47]	0.247	0.224	0.354	0.696
SPR [35]	0.396	0.337	0.616	0.577
BBF [36]	0.917	0.717	1.045	0.342
EfficientZero [48]	1.090	1.020	1.943	0.367

how much of the remaining error comes from augmentation noise or action-dependent bias. Separate partial-seed Boltzmann results show selected gains and losses relative to greedy-target MINT; reduced overestimation is a hypothesis, and a benchmark-wide advantage remains unestablished. The synthetic trace measurements describe their own policy/noise protocol and do not identify the cause of the return differences.

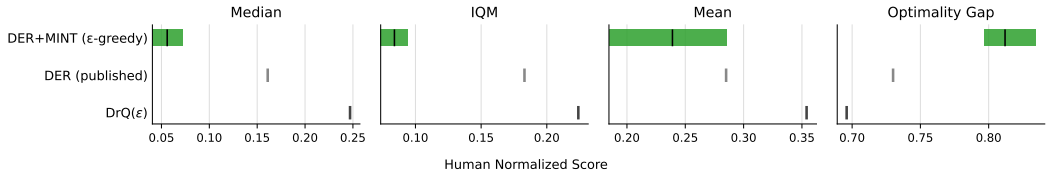


Figure 18: **Aggregate metrics (focused).** DER+MINT (ε-greedy) vs. published point estimates for DER and DrQ(ε) (from Schwarzer et al. 36). Green bars show the reported 95% bootstrap CIs. Boltzmann results are omitted because the full five-seed comparison is incomplete; partial raw scores are in Table 14.

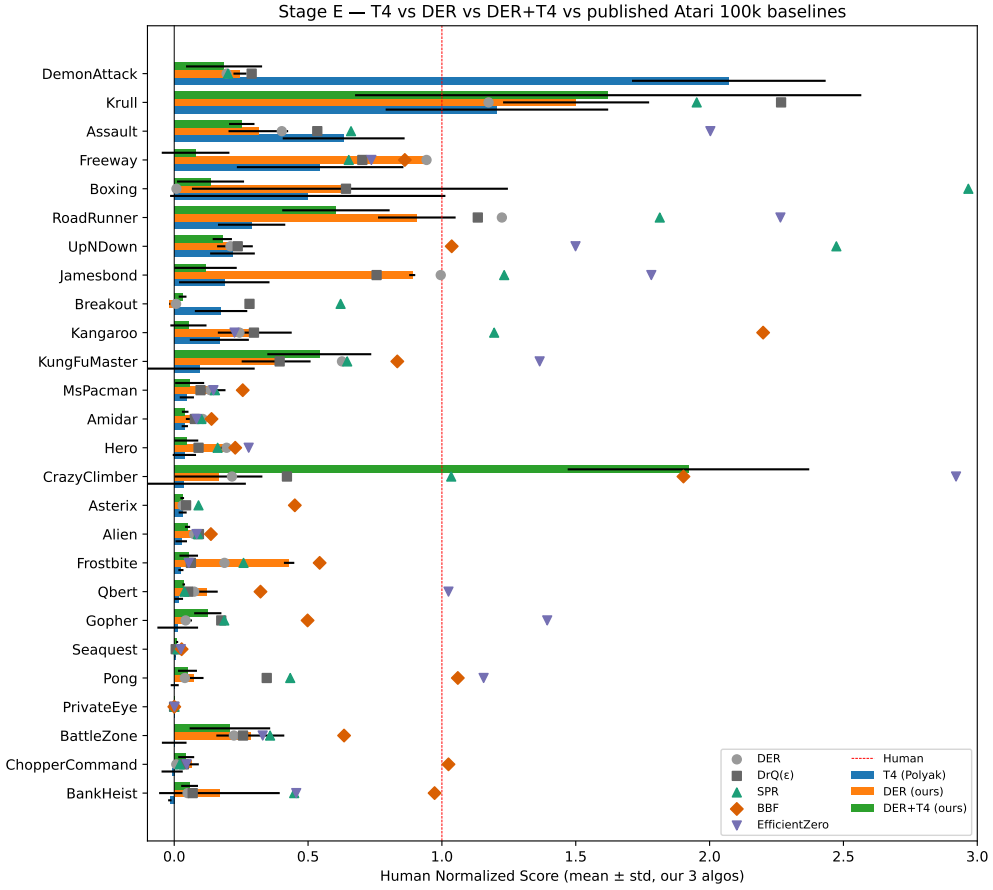


Figure 19: **Atari 100k per-game HNS** (mean $\pm$ std, 5 seeds, ε-greedy target variants only). Bars: MINT (Polyak), DER (ours), DER+MINT. Markers: published baselines. Dashed red line: Human (HNS $= 1$). The ε-greedy DER+MINT stays below DER on most games; the Boltzmann $\tau = 0.1$ variant is previewed in Table 14, with its per-game figure pending the full 5-seed sweep. Horizons up to 20 belong to the separate synthetic diagnostic; the deployed Atari cap is 10.

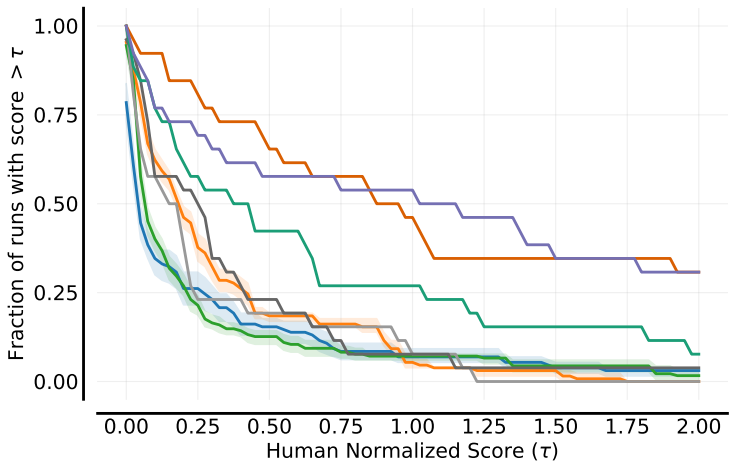


Figure 20: **Atari 100k performance profile** (`rliable`, 26 games $\times$ 5 seeds, greedy-target variants). $\hat{F}(\tau)$ is the fraction of (game, seed) pairs reaching at least τ HNS. The displayed MINT profiles lie below DER across $[0, 2]$, so the aggregate deficit is not confined to a few outlying scores. No Boltzmann profile is reported because the full five-seed comparison is incomplete.

G Cap, Signal, and Contraction in the Deep-RL Regime

This appendix reports experiments added during the NeurIPS discussion phase to answer four questions raised by the reviewers: whether the hard cap $K_{\max}$, rather than the disagreement signal, determines the effective horizon (Appendix G.1); how the action-level surrogate $\hat{p}_i$ compares with a one-step ϵ-tolerant matching diagnostic (Appendix G.2); how sensitive `MINT` is to λ and to the cap (Appendix G.3); and whether the deployed operator is contractive in the continuous-state, stochastic regime that the experiments use (Appendix G.4).

Common setup. We use RWRL `walker_walk` [10], the Real-World RL Challenge variant of the DeepMind Control task, with a SAC backbone. RWRL provides a benchmark-defined actuator-noise channel: Gaussian noise of scale $\sigma \in \{0.0, 0.1, 0.3\}$ is added to the executed action inside the environment, so that $\mathcal{P}(s' \mid s, a)$ is non-degenerate at a specified level; we use this published mechanism rather than a perturbation of our own. All other settings follow the main table ($\lambda = 0.7$, $\gamma = 0.99$, Table 10); only the cap $K_{\max} \in \{5, 10, 50\}$ and the entropy coefficient vary. Two arms are run. The *fixed-*α arm uses $\alpha = 0.05$, 3 seeds, and 200k environment steps; the main-table value $\alpha = 0.2$ does not learn on this DMControl-derived task (return $\approx$27 after 80k steps), whose per-step reward is roughly $3.5\times$ smaller than Hopper's, so α is rescaled to the reward scale, exactly the fixed-α failure mode that Figure 5b predicts. The *auto-*α arm learns α and runs 5 seeds per $(\sigma, K_{\max})$ cell (15 trials per cap) for 780k steps. These runs are a K_{eff} diagnostic on `MINT` itself—the only baseline included is uncorrected n-step at the same cap—not a re-run of the main comparison. All statistics are read at a common checkpoint for every trial.

What is and is not measured. Two quantities that the questions above might suggest are deliberately not reported, and we state the substitutes precisely. (i) We do not measure the multi-step state-coupling meeting time $T_{\beta,\pi}$ of Equation (4) on continuous control, nor a correlation between it and $\hat{K}$. The measured state-level object is a calibrated one-step ϵ-tolerant matching diagnostic, denoted $\widehat{m}_\epsilon(s)$ and compared with action-level TV in Appendix G.2; it is not identified with exact kernel TV $d(s)$; the retained-horizon and early-stopping evidence is summarized in the main text (Table 2). (ii) We do not run an uncapped ($K_{\max} = \infty$) arm; the largest tested cap is 50. Neither substitute establishes the multi-step coupling guarantee for the deployed estimator; they provide evidence about the selected horizon and an empirical association between action-level TV and the matching diagnostic.

G.1 Additional Cap and Signal Diagnostics

Table 2 in the main text reports the fixed-α retained horizons and early-stopping fractions. Here we give the corresponding normalized horizons and mean keep probabilities (Table 16), followed by the separate auto-α study.

Table 16: Additional diagnostics for the fixed-α RWRL comparison in Table 2: $\alpha = 0.05$, 200k steps, three seeds, noise levels pooled.

$K_{\max}$	$K_{\text{eff}}/K_{\max}$	$\mathbb{E}_\beta[\hat{p}]$
5	0.79	0.85
50	0.10	0.70

Auto-α diagnostic: when K_{eff} does track the cap. The auto-α arm shows the regime in which cap tracking occurs, and why. Table 17 reports K_{eff} and the fraction of traces stopping early over training. For $K_{\max} = 50$, K_{eff} rises during warm-up (while β is near-uniform the density ratio is not small), peaks near 100k steps, and declines to 25% below its peak by 780k steps, with three quarters of traces stopping early. For $K_{\max} \in \{5, 10\}$ it settles at $\approx$97% of the cap. The cause is a temperature collapse: the learned α converged to 0.002–0.014 in every arm (last column), far below the fixed 0.2 of the main table. This makes π so peaked that, for actions drawn from the more diffuse β, $\log(\pi/\beta)$ falls to ≈ -38; $\min(1, \pi/\beta)$ is then numerically zero and $\hat{p}$ saturates at $\approx$0.98. A saturated $\hat{p}$ almost never signals agreement, so traces run to the cap. The fixed-α arm confirms the diagnosis on the same task: $\mathbb{E}_\beta[\hat{p}]$ falls from 0.98 to 0.70 and K_{eff} from 21.9 to $\approx$5 at $K_{\max} = 50$

with the cap unchanged. The saturated surrogate, not the cap, was holding the traces open; a flat K_{eff} under a collapsed temperature therefore reflects saturation rather than a genuinely long meeting time, and this is why the main experiments fix α. We pool over σ because its effect on K_{eff} was smaller than the run-to-run spread, so we do not claim a trend with the noise level. A repeat of the 45-trial grid on different hardware with the same seeds reproduced every cell to within 0.1 in K_{eff} and 0.6 percentage points in the early-stopping fraction.

Table 17: Auto-α arm: K_{eff} and $\Pr[\hat{K} < K_{\max}]$ at 20k $\to$ 100k $\to$ 780k steps (15 trials per cap, σ pooled), and the learned α at the end of training (5-seed mean).

$K_{\max}$	K_{eff}	$\Pr[\hat{K} < K_{\max}]$	learned α
5	$4.34 \to 4.86 \to 4.85$	$23.9\% \to 5.5\% \to 5.7\%$	0.0023
10	$8.45 \to 9.49 \to 9.69$	$27.2\% \to 9.3\% \to 5.9\%$	0.0082
50	$23.98 \to \mathbf{29.25} \to \mathbf{21.86}$	$76.7\% \to 61.8\% \to \mathbf{75.9}\%$	0.0058–0.0138

What the cap costs. Sweeping $K_{\max}$ also locates the budget below which the cap stops being harmless (Table 19). Reducing $K_{\max}$ from 50 to 10 lowers K_{eff} by 56% but final return by only 7%; reducing it from 10 to 5 halves K_{eff} again and lowers return by 60%. This is the expected signature of $\hat{K} = \min(K_{\max}, \hat{T}_{\beta,\pi})$: while the budget exceeds the task's meeting time the cap is nearly free, and once it binds strongly, performance falls. The fixed-α arm reproduces the direction at its two caps (603 ± 110 at $K_{\max} = 50$ versus 25 ± 3 at $K_{\max} = 5$; 200k steps, 3 seeds), so the effect is not an artifact of the collapsed temperature. Two consequences for the main text follow. First, the cap-robustness observed on MuJoCo `Hopper` (Figure 1b) is a property of tasks whose meeting time is short relative to the cap; on `walker_walk` the meeting time exceeds 5 and the main-table budget binds. Second, $K_{\max} = 5$ is a deliberately conservative, parity-driven choice, so where the budget is not the binding constraint the main table understates `MINT`'s headroom over fixed-horizon baselines.

Table 18 breaks the $K_{\max} \in \{10, 50\}$ returns down by noise level: the agent learns a strong policy at every σ and performance is stable up to $\sigma = 0.3$, with the gain from $K_{\max} = 10$ to 50 marginal, so the cap effect is not driven by any single noise level.

Table 18: Eval return on RWRL `walker_walk` by transition-noise level (auto-α arm, 780k steps, mean $\pm$ s.d. over 5 seeds; max return 1000).

σ	$K_{\max} = 10$	$K_{\max} = 50$
0.0	938.3 ± 39.7	963.2 ± 6.0
0.1	878.1 ± 104.7	967.6 ± 6.7
0.3	865.7 ± 79.3	963.7 ± 3.5
pooled ($n = 15$)	894.0 ± 80.4	964.8 ± 5.5

Table 19: Return as a function of the cap (auto-α arm, 780k steps, 15 trials per cell; the only experiment with a $K_{\max} = 10$ condition).

$K_{\max}$	K_{eff}	ΔK_{eff}	return	Δ return
50	21.86	—	965	—
10	9.69	-56%	894	$\mathbf{-7}\%$
5	4.85	-50%	358	$\mathbf{-60}\%$

G.2 The Action-Level Surrogate and a One-Step Matching Diagnostic

Proposition A.4 gives the exact one-step relation $\mathbb{E}_\beta[\hat{p}_i \mid s] = D_s \geq d(s)$, where $d(s) = \text{TV}(\mathcal{P}^\beta(\cdot \mid s), \mathcal{P}^\pi(\cdot \mid s))$. RWRL's action-noise channel factorizes as $a \mapsto \tilde{a} = \text{clip}(a+\varepsilon) \mapsto s' = F(s, \tilde{a})$, with the same Markov maps for both policies, so the relation applies to the exact resulting distributions. Closeness of next-state coordinates alone does not establish TV overlap.

On the $K_{\max} = 50$ auto-α arm (seed 0), action-level D_s is evaluated by numerical quadrature of the two Gaussian policies. The simulator-based comparison uses a calibrated ϵ-tolerant matching statistic

$\widehat{m}_\epsilon(s)$, with a $\beta = \pi$ calibration control, at 250 replay states across four checkpoints. We report an ϵ-tolerant matching diagnostic, rather than an estimate of exact kernel TV. Its observed ordering relative to action-level TV is an empirical diagnostic comparison, not a verification of DPI.

The measured ordering $\mathbb{E}_\beta[\hat{p}] \geq \widehat{m}_\epsilon(s)$ holds at all 250 retained states. The reported Spearman correlations are 0.95 across the arm/checkpoint cells and 0.86 across states (Table 20), with level gaps 0.23–0.40. The diagnostic excludes states whose calibration floor exceeds 2%, so this selection may affect the observed correlations and gaps. These are directional measurements on one task and one seed; they neither identify $d(s)$ nor order multi-step stopping times or certify the χ_λ contraction condition.

Table 20: Action-level TV versus the ϵ-tolerant matching diagnostic on RWRL `walker_walk` ($K_{\max} = 50$ arm, seed 0) at four checkpoints. $\mathbb{E}_\beta[\hat{p}] \geq \widehat{m}_\epsilon(s)$ at all 250 retained states; reported Spearman correlations are 0.95 across cells and 0.86 across states. The matching statistic is not an estimate of exact kernel TV.

checkpoint	$\mathbb{E}_\beta[\hat{p}]$	matching diagnostic $\widehat{m}_\epsilon(s)$
29%	0.981	0.755
57%	0.874	0.473
71%	0.909	0.655
100%	0.735	0.365

G.3 Sensitivity to λ

λ is not a `MINT`-specific knob: the main table runs `MINT`, Peng's $Q(\lambda)$, and Retrace at the same $\lambda = 0.7$, chosen to match the Peng's $Q(\lambda)$ protocol rather than to optimize our operator. It is nevertheless a genuine hyperparameter, setting the $(\gamma\lambda)^t$ weight on every term of the backup. Table 21 reports the full sweep on MuJoCo `Ant-v2` and `Hopper-v2` (fixed $\alpha = 0.2$, 1M steps, final return averaged over the last five evaluations, mean $\pm$ s.d. across seeds) without selecting favorable settings. Return decreases with λ on `Ant-v2` and is comparatively stable on `Hopper-v2` (about 9% variation over the range); from $\lambda = 0.7$ to 0.9 the change is -12.9% on `Ant-v2` and $+3.5\%$ on `Hopper-v2`, with overlapping mean $\pm$ s.d. ranges in both cases. Two qualifications apply: $\lambda = 0.7$ is not the best setting in this sweep (`Ant-v2` prefers 0.1–0.3), consistent with its having been fixed for protocol matching; and because the baselines were not swept, we make no claim about how the gap between methods changes at other λ. Kozuno et al. [23] likewise find Peng's $Q(\lambda)$ sensitive to λ on DMControl (their Appendix J.8), with the best overall performance near $\lambda = 0.7$, so λ sensitivity is not unique to `MINT`. In the reported runs the cap is fixed at 5 independently of λ (Equation (11)), so λ acts only through the $(\gamma\lambda)^t$ weighting.

Table 21: λ sweep for SAC-`MINT` (fixed $\alpha = 0.2$, $K_{\max} = 5$, 1M steps; mean $\pm$ s.d. across 2–3 seeds per cell). The main-table value is $\lambda = 0.7$.

task	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.7$	$\lambda = 0.9$
`Ant-v2`	5812 ± 81	5991 ± 92	5466 ± 98	3922 ± 553	3417 ± 1594
`Hopper-v2`	3453 ± 8	3291 ± 60	3280 ± 3	3151 ± 161	3262 ± 137

G.4 Empirical Gain of the Deployed Operator

The next-state-distance and ϵ-tolerant matching diagnostics do not certify the sufficient condition of Theorem B.9. We therefore probe the implemented backup empirically. Subtracting backups applied to two value functions cancels the reward and entropy terms and leaves a linear map on $\Delta = Q - Q\prime$; its measured gain over the visited distribution is

$$\hat{\kappa} \;=\; \frac{\|\hat{\mathcal{T}}Q - \hat{\mathcal{T}}Q'\|}{\|Q - Q'\|},$$

which we estimate by fitted power iteration over the visited distribution. This includes the action-level surrogate and hard cap as deployed, at the reported coefficients. It is a fitted, visited-distribution diagnostic composed with critic projection. It is neither the mixed-norm state-level bound $\kappa^\star$ of Corollary C.3 nor the global surrogate upper bound $\kappa_{\chi,L}$ of Theorem B.9.

Setup. RWRL `walker_walk` with injected transition noise ($\sigma \in \{0.1, 0.3\}$), 100k steps, 3 seeds, and 60 probes across training checkpoints of the auto-α arm. We deliberately use this arm because $\hat{p}_i$ saturates there and $\hat{K}$ sits near the cap, the least favorable observed regime for the operator. The policy pairs were trained at $(\lambda, \gamma) = (0.9, 0.999)$ and probed at the deployed coefficients; λ and γ enter only the operator, not the policies. The transition operator is evaluated on shared samples, so no density model of the kernel is needed.

Table 22: Empirical backup gain $\hat{\kappa}$ on RWRL `walker_walk` with transition noise. **Top:** operating point (pooled over cap, noise, seed). **Middle:** deployed coefficients by cap. **Bottom:** comparison at $\lambda = 0$ with the one-step Bellman reference γ. These are fitted probe measurements, not certificates of the global surrogate sufficient condition.

coefficients (λ, γ)	mean $\hat{\kappa}$	median	s.d.	$\hat{\kappa} < 1$
(0.7, 0.99) (deployed, Section 5)	**0.90**	0.88	0.13	**78**%
(0.9, 0.999) (training coefficients)	1.23	1.11	0.44	27%
$K_{\max}$ at (0.7, 0.99)	mean $\hat{\kappa}$	$\hat{\kappa} < 1$	$\bar{\hat{K}}$	$\Pr[\hat{K} = K_{\max}]$
5	**0.84**	**90**%	4.68	88%
50	0.96	67%	7.75	61%
$\lambda = 0$ gate: γ	$\hat{\kappa}$	$\hat{\kappa}/\gamma$		
0.99	0.978 ± 0.058	0.988		
0.999	0.979 ± 0.064	0.980		

Results. At the deployed coefficients ($\lambda = 0.7$, $\gamma = 0.99$, $K_{\max} = 5$), the mean empirical gain is $\hat{\kappa} = 0.84 \pm 0.13$, below 1 on 90% of probes (Table 22, middle block), at checkpoints where the trace reaches the cap 88% of the time. The tighter cap gives a smaller measured gain. The same policies at $(\lambda, \gamma) = (0.9, 0.999)$ give $\hat{\kappa} = 1.23$, below 1 on 27% of probes. These comparisons distinguish observed probe behavior without certifying a global operator norm. The backup-difference implementation is checked by a unit test; at $\lambda = 0$, the measured gain is within about 2% of the one-step Bellman reference γ (bottom block).

Scope. $\hat{\kappa}$ uses a finite anchor sample, includes critic projection, and comes from one task. The dominant reported error is the power-iteration fit (median relative error $\approx$14%), not captured by the seed s.d. These probes are empirical stability diagnostics rather than estimates of $\kappa_{\chi,L}$ or proofs of convergence for the deep-RL runs. We do not claim a trend with training progress or noise level.

H Stochastic Dynamics and Stitching: Off-Policy Evaluation on OGBench

MINT's central claim concerns the multi-step *target*, and its fixed point is Q^π (Lemma 3.4). We therefore also evaluate it as a pure off-policy *evaluation* operator. Fixing the target policy removes the control-loop confounds (entropy tuning, exploration, replay non-stationarity) and provides a finite-rollout Monte Carlo reference for Q^π, so target errors are measured against that reference rather than inferred from returns. The experiment jointly addresses two reviewer questions: robustness to stochastic dynamics and to data that must be stitched across trajectories.

Setup. We use OGBench AntMaze [31], whose observation is the full simulator state, so any dataset state can be restored and the target policy rolled out from it. The grid crosses stochasticity against stitching: `antmaze-large` (deterministic) and `antmaze-teleport` (stochastic), each with the `navigate` (1000-step trajectories) and `stitch` (200-step trajectories) datasets. The stochasticity is the benchmark's own: entering a teleporter moves the agent to one of three exits uniformly at random, a discrete state jump that makes the kernel genuinely multimodal, unlike the unimodal action noise of Appendix G. We use $\gamma = 0.99$, window length $L = K_{\max} = 50$, $\lambda = 0.9$, and OGBench's single-task reward relabeling. The behavior policy β is a state-conditioned Gaussian fit to each dataset. The target policy displaces its mean by h behavior standard deviations along a fixed unit direction d in action space, $\mu_\pi(s) = \mu_\beta(s) + h\,\sigma_\beta(s)\,d$ with $\sigma_\pi = \sigma_\beta$, so $h = 0$ recovers $\pi = \beta$ exactly and h measures policy divergence in units of behavior standard deviations. Operators: MINT, Retrace(λ), and a *shuffled-*$\hat{K}$ control that permutes MINT's truncation patterns across the batch, preserving the marginal distribution of $\hat{K}$ exactly while destroying its match to each state's own π/β mismatch. Every operator sees identical states, windows, and critic values, so all comparisons are paired at the level of individual evaluation states.

The ratios use a fitted behavior model, and reported errors are relative to finite-rollout Monte Carlo references. Behavior-model misspecification and Monte Carlo uncertainty contribute additional errors beyond the ideal fixed-policy operator analysis. In particular, $h = 0$ matches the target to the fitted behavior model; it does not establish equality with the true data-generating policy.

H.1 Frozen Critic: Properties of the Operator Alone

With the critic fixed to the finite-rollout Monte Carlo estimate, the comparison holds that approximate critic constant across target operators. Table 3 in the main text reports RMSE and intrinsic sampling variance at $h = 0.5$; Table 23 adds bias (mean signed error) and the paired t-statistic of the per-state absolute-error difference between MINT and Retrace. Table 24 reports all metrics at $h = 1$. Intrinsic variance is measured with the approximate frozen critic; the exact-Q^π bound of Lemma 3.4 is not directly a guarantee for this measurement.

Table 23: Additional frozen-critic statistics at $h = 0.5$ for the comparison in Table 3. t: paired t-statistic of the per-state absolute-error difference, MINT vs. Retrace.

Dataset	Bias: MINT	Bias: Retrace	t
large-navigate (det.)	0.35	2.77	2.5
large-stitch (det.)	0.38	3.81	4.0
teleport-navigate (stoch.)	0.16	2.14	2.3
teleport-stitch (stoch.)	0.22	2.60	3.2

Paired comparisons and error metrics. Section 5.4 summarizes the target-level comparison; Tables 3, 23 and 24 give the per-cell measurements. Bias is a mean signed error and permits cancellation across states, so we also report RMSE: the MINT–Retrace ratios correspond to $2.1\times$–$13.8\times$ reductions in bias magnitude and the more modest $1.04\times$–$2.96\times$ reductions in RMSE. At $h = 0.5$ the paired per-state absolute-error comparison has $t = 2.3$–4.0; at $h = 1$ it is within sampling noise in three of the four cells ($t = 0.6$–2.2). We do not claim that the advantage grows with divergence: at $h = 2$ the ordering reverses, because Retrace's coefficient $\lambda \min(1, \pi/\beta)$ itself truncates more aggressively as π moves away from β. Across the full 16-cell comparison, MINT has lower $\mathrm{Var}_{\mathrm{int}}$ than the shuffled-$\hat{K}$ control in 15 cells; `large-navigate` at $h = 0.5$ is the exception. The control complements the rate-matched comparison of Appendix D.5.

Table 24: OGBench frozen-critic evaluation at $h = 1$ (one paired run per cell, 100 stratified evaluation states). h: target-policy displacement in behavior standard deviations; t: paired t-statistic of the per-state absolute-error difference, `MINT` vs. Retrace. M = `MINT`, R = Retrace, shuf = shuffled-$\hat{K}$ control. Results at $h = 0.5$ are split between Tables 3 and 23.

dataset	h	bias M	bias R	RMSE M	RMSE R	t	$\mathrm{Var}_{\mathrm{int}}$ M	$\mathrm{Var}_{\mathrm{int}}$ shuf
large-navigate (det.)	1.0	0.58	1.60	4.78	5.31	1.6	2.79	4.96
large-stitch (det.)	1.0	−0.33	1.80	2.78	5.32	2.2	1.45	4.08
teleport-navigate (stoch.)	1.0	−0.39	0.81	3.52	3.66	0.6	2.26	2.60
teleport-stitch (stoch.)	1.0	0.47	1.63	3.83	4.63	1.0	2.21	3.37

Stitching. The `stitch` datasets contain 200-step rather than 1000-step trajectories, so a fixed window crosses trajectory boundaries more often. Nevertheless, at $h = 0.5$ `MINT`'s paired advantage over Retrace is larger on `stitch` ($t = 4.0$ and 3.2) than on the corresponding `navigate` datasets ($t = 2.5$ and 2.3): the correction does not degrade when trajectories must be stitched.

Protocol and caveat. We use 100 evaluation states per cell, stratified by distance to a goal-reaching event, because uniformly sampled states have nearly constant Q^π on these mazes and would make the comparison uninformative. The reference $Q^\pi(s_0, a_0)$ uses 100 Monte-Carlo rollouts per state; frozen-critic values along each window use 30 rollouts per state and are shared across all operators; stochastic operators are averaged over 32 draws per state. Each cell is a single paired run; the seed-replicated version is the learned-critic study below. A one-step baseline attains the lowest measured RMSE in most cells of this approximate-frozen-critic comparison; this is an empirical result rather than an optimality claim based on an exact critic. This panel is therefore a comparison among off-policy *corrections*, not evidence that `MINT` dominates every target estimator.

H.2 Learned Critic: Fitted-Q Evaluation and a Limitation

As a complementary experiment we run the same operators inside fitted-Q evaluation, learning the critic from scratch (5 seeds) and scoring it against the same Monte-Carlo Q^π. The ranking depends on the optimization horizon: at 100k gradient steps Retrace is better on every dataset, and by 300k steps the ordering on the deterministic maze reverses—at $h = 1$ `MINT` truncates to $\hat{K} \approx 1.6$, propagates value more slowly, and needs roughly 200k updates to stabilize, but then reaches lower error on the deterministic maze. Evaluations before convergence therefore conflate propagation speed with asymptotic estimation quality. Table 25 reports RMSE at 300k steps; we give the median alongside the mean because late-training instability in two cells makes the mean alone misleading. A seed is classified as unstable when its final error is at least twice its best error during training.

Table 25: Learned-critic (fitted-Q evaluation) RMSE against the Monte-Carlo Q^π at 300k gradient steps, 5 seeds (one Retrace cell has 4). Bold marks the best median in each cell.

dataset	h	operator	median	mean ± s.d.	unstable seeds
large-navigate (det.)	1.0	`MINT`	17.15	17.15 ± 0.55	0/5
		Retrace	19.92	19.92 ± 0.53	0/4
		shuffled-$\hat{K}$	**16.62**	16.62 ± 0.56	0/5
teleport-navigate (stoch.)	1.0	`MINT`	17.55	17.20 ± 1.63	0/5
		Retrace	**12.04**	12.04 ± 0.43	0/5
		shuffled-$\hat{K}$	16.98	16.17 ± 1.98	0/5
large-navigate (det.)	2.0	`MINT`	20.72	21.06 ± 1.06	0/5
		Retrace	**17.66**	25.86 ± 16.53	**1/5**
		shuffled-$\hat{K}$	20.70	19.67 ± 1.57	0/5
teleport-navigate (stoch.)	2.0	`MINT`	27.00	64.18 ± 51.82	**2/5**
		Retrace	**15.44**	15.47 ± 0.58	0/5
		shuffled-$\hat{K}$	21.64	21.04 ± 2.43	0/5

Bias and value propagation. The stochastic-maze mean RMSE at $h = 1$ is 12.04 for Retrace and 17.20 for `MINT`. The ranking across optimization horizons described above is consistent with a

bias–propagation trade-off: a shorter effective horizon can reduce bias per backup but slow value propagation, and which effect dominates depends on critic convergence. The main-text overview is in Section 5.4.

Instability at large divergence under stochastic dynamics. At $h = 2$ both corrected operators show late-training instability. On the stochastic maze all `MINT` seeds reach RMSE 15–20 mid-training before degrading, two of five severely (final RMSE 85 and 157); one Retrace seed on the deterministic maze ends at RMSE 59 against a median of 17.66. The instability is therefore not unique to `MINT`, but it is more frequent and more severe for `MINT` under stochastic dynamics. The shuffled-$\hat{K}$ control localizes the source: it samples truncation lengths from `MINT`'s marginal distribution but removes their state matching, and it is stable in both $h = 2$ cells (0/5 unstable seeds, s.d. 1.57 and 2.43). The instability thus arises not from short stochastic truncation per se but from coupling the truncation length to the state: when $\hat{p}$ is large and the kernel is multimodal, this coupling can correlate the horizon with the value being estimated. We record this as a limitation of `MINT` at large policy divergence under stochastic dynamics (Section 5.4 and Appendix J); whether learning-rate or target-network schedules mitigate it is left open. Finally, shuffled-$\hat{K}$ performs at least as well as `MINT` in every learned-critic cell. Together with the frozen-critic variance comparison above, adaptive state matching demonstrably reduces the target's *intrinsic* variance, but we do not yet observe that reduction consistently in end-to-end learned-critic error.

I Compute and Reproducibility

Hardware. All experiments were run on a single workstation with the following configuration: Ubuntu 24.04.1 LTS (kernel 6.8.0); 2× NVIDIA RTX 5000 Ada Generation GPUs (32 GB GDDR6 each, driver 560.35.03, CUDA 12.6); an AMD Ryzen Threadripper PRO 7985WX CPU (64 cores / 128 threads); and 502 GB of system RAM. Each training run is pinned to a single GPU, and we run two parallel seeds at a time.

Compute breakdown. Estimated aggregate GPU time for the listed experimental components is reported below. GPU-hours sum device time across runs; simultaneous execution on two GPUs reduces elapsed wall-clock time, not this aggregate.

- **MuJoCo main** (5 tasks × 2 backbones × 5 methods × 5 seeds × 1M steps): 250 single-GPU runs at approximately 6 hours per run give **1,500 GPU-hours** (about 750 elapsed hours if both GPUs are continuously occupied).
- **Cap-length sweep** ($\geq$ 3 tasks × 4 caps × 4 baselines × 5 seeds × 1M steps): $\approx$ **360** GPU-hours.
- **DMControl sparse-reward** (1–2 tasks × 5 baselines × 5 seeds × 1M steps): $\approx$ **120** GPU-hours.
- **DMControl truncation-kernel ablation** (Appendix F.3; 5 tasks × 3 methods × 3 seeds × 1M steps, plus a 5-seed re-run of `CartPole-Swingup-Sparse`): $\approx$ **100** GPU-hours.
- **Ablations** (clipped surrogate $\hat{p}_i$, clipping form, hard cap $K_{\max}$): $\approx$ **80** GPU-hours.
- **Toy-chain variance experiments**: $<$ 1 hour on CPU only.

The subtotal of the listed GPU components is approximately **2,160 GPU-hours** ($1{,}500 + 360 + 120 + 100 + 80$), excluding the separately reported CPU-only toy experiments. This is a subtotal for the listed components, rather than an accounting of every additional benchmark or preliminary run.

Reproducibility. All evaluation protocols follow OpenAI SpinningUp [2] unless noted otherwise; full hyperparameters are listed in Tables 9 and 10 and the Q-target update rules in Table 8.

J Limitations and Assumptions (Full Discussion)

We list the conditions under which `MINT`'s guarantees may degrade, and indicate which we expect to tighten in future work.

Theory: discount factor. The geometric state-coupling sufficient condition of Corollary B.1 becomes restrictive as γ approaches 1. It concerns the state-coupled operator; the implemented action-level surrogate is governed instead by the sufficient condition of Theorem B.9. Empirical success at $\gamma = 0.99$ does not verify either global condition.

Theory: structural assumptions. Corollary B.1 requires cross-Doeblin overlap, which is not directly verified on continuous-control benchmarks. The next-state-distance diagnostic and the one-step comparison of Appendix G.2 do not estimate a uniform ρ. The separate surrogate theorem assumes finite state and action spaces and fixed policies; a finite-sample analysis and extensions to continuous-state, changing-policy learning remain future work.

Algorithm: surrogate $\hat{p}_i$. The surrogate obeys $\mathbb{E}_\beta[\hat{p}_i \mid s_i] = D_{s_i} \geq d(s_i)$ for one-step transitions from a shared state. This controls its mean at that state but does not order the complete state-coupling and surrogate stopping times. The action-dependent contraction correction is χ_λ, not D_s alone (Theorem B.9).

Stochastic dynamics. The one-step DPI relation of Proposition A.4 and the finite-cap surrogate theorem of Theorem B.9 allow stochastic dynamics. Their conclusions are different: the former compares one-step TV quantities, while the latter gives a conditional bound for an expected evaluation operator. Neither implies that stochastic dynamics merely loosen an otherwise unchanged state-level contraction guarantee. The RWRL and OGBench experiments (Appendices G and H) provide empirical evidence on stochastic kernels and on learned-critic limitations.

Large policy divergence under stochastic dynamics. In learned-critic off-policy evaluation on OGBench AntMaze (Appendix H.2), `MINT` is unstable in two of five seeds when the target policy is displaced by two behavior standard deviations on the stochastic `teleport` maze, and Retrace is more accurate there and at the same displacement on the deterministic maze. A control that keeps `MINT`'s marginal distribution of truncation lengths but removes their state matching is stable, so the instability is attributable to coupling the horizon to the state when $\hat{p}$ is large and the kernel is multimodal. `MINT`'s bias advantage is established at small divergence; at large divergence under stochastic dynamics it is not, and the end-to-end benefit of the adaptive signal remains to be demonstrated there.

Atari: target updates and approximate-value corrections. The DrQ diagnostics in Appendix F.5 show divergence under per-step hard target synchronization and stable Q-values after increasing the synchronization period to 100. Stability alone does not recover baseline performance: removing the intermediate $V - Q$ terms with unchanged trace lengths improves returns on `Seaquest` and `Qbert` (three seeds per game). These results implicate the terms under the tested protocol, but do not identify the relative contributions of max-selection bias, action-dependent estimation error, and augmentation noise, or explain why the modified target exceeds DrQ on `Seaquest`. Fixed-point guarantees at Q^π do not ensure benign learning dynamics with an approximate critic. Separate Boltzmann-target results use Polyak updates and partial seeds; their gains over the greedy-target variant neither establish a benchmark-wide advantage nor make Boltzmann targets necessary. Boltzmann averaging may attenuate overestimation-driven amplification, but it also changes trace continuation, and these effects have not been isolated.

Scope: credit assignment, not exploration. `MINT` modifies multi-step return estimation but does not address exploration. On exploration-bottleneck sparse tasks where the base off-policy algorithm rarely discovers reward in the first place (e.g., `Acrobot-Swingup-Sparse`, manipulator-stacking suite), trace-truncation has no signal to adapt to and `MINT` performs comparably to the underlying SAC baseline. Combining `MINT` with intrinsic motivation, demonstration-based bootstrapping, or other exploration mechanisms is complementary and left to future work.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction describe the MINT operator and its model-free action-level surrogate (Section 3), the state-coupled variance bound at Q^π (Lemma 3.4), separate sufficient contraction conditions for the two operators (Theorems B.9 and 4.2), and the empirical evaluation (Section 5 and Appendices F.5, G and H). The theoretical scope and implementation limitations are stated separately (Section 5.4 and Appendix J).

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The main-text and appendix limitations (Section 5.4 and Appendix J) distinguish the state-coupled sufficient condition from the finite-state, fixed-policy surrogate result. They identify the unproved sample-based, changing-policy, and nonlinear-critic extensions, and discuss the surrogate's one-step DPI relation, stochastic-dynamics limitations, and credit-assignment versus exploration scope.

   Guidelines:

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

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: The results state their hypotheses separately: Lemma 3.4, Theorems B.7 to B.9 and 4.2, and Corollary B.10. The surrogate theorem is summarized in Theorem B.9 and proved in Appendix B.8; the other proofs are in Appendix B. The assumption table (Table 4) separates the state-level and surrogate analyses.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We provide algorithm pseudocode (Algorithm 1), full hyperparameter tables (Appendix E), environment specifications including task names and reward structures (Section 5), network architectures, optimizer choices, and training schedules (Appendix E). Anonymized code with run scripts is included in the supplementary material.

   Guidelines:

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

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: Anonymized code with run scripts and configuration files is provided in the supplementary material, sufficient to reproduce all main experimental results on the publicly available DeepMind Control Suite (MuJoCo) benchmarks. A README with exact commands for each experiment is included.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: The main experimental setup is described in Section 5, with full hyperparameters (learning rates, batch sizes, replay ratios, network widths, λ values), optimizer choices (Adam), network architectures, and training schedules detailed in Appendix E.

   Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: Aggregate IQM scores use the stated stratified-bootstrap protocol [3]; error-bar definitions are figure-specific. For example, the adaptive-horizon panels report ± 1 standard deviation, whereas aggregate IQM intervals use the reported confidence level. The main MuJoCo comparison uses 5 seeds per method and task; the DMC truncation-kernel ablation uses 3 seeds with a separate 5-seed CartPole re-run. Additional experiments give their seed counts and distinguish mean $\pm$ standard deviation, Monte-Carlo draws, and paired tests in their own captions and protocols.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Appendix I reports hardware, the stated per-run duration, and an aggregate GPU-hour subtotal for the listed components, with CPU-only toy time separate. This subtotal is not presented as a complete ledger of all preliminary or additional-benchmark runs.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

Justification: The authors have reviewed the NeurIPS Code of Ethics and confirm that this research conforms to it in every respect. The work involves standard RL benchmark evaluation with no human subjects, no private data, and no foreseeable dual-use concerns.

Guidelines:

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

10. **Broader impacts**

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

Answer: [N/A]

Justification: This work is foundational research on off-policy reinforcement learning algorithms evaluated on standard physics-simulation benchmarks. It does not target any specific application domain and we do not foresee specific positive or negative societal impacts beyond those generally associated with incremental progress in RL methodology.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: We do not release pre-trained models, scraped datasets, or other artifacts that pose a high risk for misuse; the released code targets standard RL benchmarks (DeepMind Control Suite) and does not produce deployable policies.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: All used assets are properly cited: DeepMind Control Suite [41], MuJoCo physics engine [42], and baseline algorithm implementations. License information (Apache 2.0 for DMControl, Apache 2.0 for MuJoCo) and version details are provided in Appendix E.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: We release our `MINT` implementation together with documentation (README with installation instructions, configuration files, and run scripts) in the supplementary material. The code includes inline documentation and will be released under an open-source license upon publication.

    Guidelines:

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

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

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

Answer: [N/A]

Justification: This work does not involve crowdsourcing or research with human subjects.

Guidelines:

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

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

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

Answer: [N/A]

Justification: This work does not involve research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: LLMs are not part of the core methodology of this work; their use was limited to writing and editing assistance, which does not require declaration per the NeurIPS 2026 LLM policy.

Guidelines:

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