AVERAGE-REWARD POLICY OPTIMIZATION VIA DISCOUNTING: ANYTIME AND GLOBAL LINEAR CONVERGENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

We study average-reward policy optimization via discounting in finite Markov decision processes where every stationary policy induces one recurrent class, possibly with transient states. With a known model and exact policy evaluation, Discount-Guided Policy Gradient achieves an anytime $O(1/n)$ optimality gap after n policy updates, with nondecreasing average rewards, from any initialization and for any fixed positive discounted step size. The proof combines local gradient domination with an explicit bound on entry into a near-optimal region. For rational model data, Blackwell Policy Mirror Ascent achieves global R-linear convergence of the average-reward gap from every full-support initialization through a multiplicative error bound at an explicit Blackwell discount. Neither guarantee requires a finite global stationary-distribution mismatch coefficient.

1 INTRODUCTION

For a decision process that runs indefinitely, average reward measures the performance sustained over time. It is therefore a natural objective when there is no prescribed stopping horizon. Policy-gradient methods approach this objective by directly adjusting the policy, but their convergence theory must account for how those adjustments change the long-run state distribution. In finite discounted Markov decision processes (MDPs), global convergence guarantees are available for several policy parameterizations and update rules (Mei et al., 2020; Agarwal et al., 2021; Xiao, 2022). Average reward poses an additional difficulty: states that matter under a better policy may receive no stationary weight under the current one.

This difficulty persists under a familiar structural assumption. Suppose every stationary policy induces exactly one recurrent class, possibly with transient states. This *all-policy unichain* condition makes each policy's average reward independent of the initial state (Puterman, 1994), but permits zero stationary probabilities. By comparison, Kumar et al. (2025) analyze MDPs in which every policy induces an irreducible and aperiodic chain. Their global gradient-domination argument uses a stationary-distribution mismatch coefficient C_{PL}, the largest ratio of an optimal policy's stationary probability to that of any policy at the same state. This coefficient can be infinite under the unichain assumption when a state is recurrent under an optimal policy but transient under another.

The obstruction can involve coordinated changes across states. A policy may remain in a moderately rewarding recurrent class while leaving a better route transient. Improving that route requires changes at states where the average-reward gradient provides no signal. Our experimental constructions exhibit suboptimal local maxima with zero first-order stationarity gap, ruling out any finite global gradient-domination constant; Appendix D.2 establishes this analytically. Section 5 illustrates the resulting difficulty for direct average-reward policy gradient and the progress enabled by discounted policy updates.

Discounting retains information about transient states: a full-support initial distribution gives positive discounted visitation weights everywhere. Discounted optimization can therefore propose improvements that direct average-reward gradients may overlook. Transferring this progress requires controlling both optimization error and the discrepancy between objectives. Increasing the discount toward one reduces the discrepancy but worsens the available discounted optimization bounds, so convergence at each fixed discount is insufficient on its own.

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

Our first approach lets the discount increase toward one. A uniform approximation bound ensures that sufficiently accurate discounted candidates reach a near-optimal average-reward region. Within this region, we establish gradient domination even when a global mismatch coefficient is infinite. Around any optimal policy, continuity preserves a positive fraction of its stationary mass on its recurrent states, yielding the required local distribution comparison. Compactness gives a common near-optimal region where gradient domination holds.

Discount-Guided Policy Gradient (DGPG) combines these two facts: it tests discounted candidates by their average reward and performs direct average-reward updates in every epoch. The method requires neither a stopping horizon nor knowledge of when it reaches the near-optimal region. Discounted candidates ensure that it reaches this region, and monotone average-reward updates keep it there. The direct updates then yield an inverse-linear rate in the total number of updates.

Our second approach keeps one discount fixed. Blackwell optimality connects discounted and average-reward optimal policies, and rational model data permit an explicit discount above the Blackwell threshold (Blackwell, 1962; Grand-Clément & Petrik, 2023). However, agreement between optimizers does not quantify the quality of intermediate policies, and the additive approximation leaves a nonzero error at a fixed discount. We prove a multiplicative bound that controls the average-reward gap by the discounted gap. Blackwell Policy Mirror Ascent (BPMA) then inherits a geometric average-reward rate from discounted mirror ascent. DGPG varies the discount without a Blackwell threshold; BPMA uses rational model data to select a fixed discount supporting error transfer throughout optimization.

Contributions. For finite MDPs satisfying the all-policy unichain condition, with known transitions and rewards and exact policy evaluations, we establish the following results.

- **Regularity with transient states.** We prove a uniform discounted-to-average approximation bound, global smoothness of the average-reward objective, and gradient domination near its optimal set (Section 2). These properties hold without a finite global stationary-distribution mismatch coefficient.
- **Anytime inverse-linear convergence.** From any initial policy, DGPG produces nondecreasing average rewards and a global $O(1/n)$ objective-gap bound after n policy updates, for any fixed positive discounted step size. We give explicit constants and a bound on the number of updates needed to reach the near-optimal region; the method does not require the regularity constants as inputs (Theorem 1).
- **Global linear convergence at a fixed discount.** Under the additional rational-model assumption, our multiplicative error bound transfers discounted mirror-ascent convergence to BPMA. Its average-reward gap has an explicit geometric bound from every full-support initial policy (Theorem 2).

1.1 Related work

Average-reward policy optimization. The policy-gradient theorem covers both discounted and average-reward objectives (Sutton et al., 2000). For ergodic tabular MDPs, Kumar et al. (2025) established finite-time global convergence of policy gradient with an $O(1/T)$ rate after T iterations. Lee & Ryu (2026) developed a policy-gradient theorem for general multichain MDPs and analyzed α-clipped policy mirror ascent, obtaining approximate optimality relative to positive policies, which assign positive probability to every action at every state. Under our all-policy unichain assumption, continuity implies that the supremum over positive policies equals the unrestricted optimum. Their analysis treats general multichain MDPs through positive policies and clipping, whereas our guarantees exploit unichain structure and require neither a clipping parameter nor a finite global stationary-distribution mismatch coefficient.

Discounted policy optimization. Global convergence analyses for tabular discounted MDPs include the softmax policy-gradient results of Mei et al. (2020) and the treatment of policy parameterization and distribution mismatch by Agarwal et al. (2021). Two results supply the discounted optimization guarantees used here. Liu et al. (2025) proved convergence of projected policy gradient with any fixed positive step size under exact policy evaluation. Xiao (2022) established linear convergence of policy mirror descent with geometrically increasing step sizes, without requiring

entropy regularization of the objective. We use the former for DGPG and the latter for BPMA. These guarantees concern a fixed discounted objective; our analysis relates their optimization errors to average reward, either along a sequence of discounts approaching one or at one suitably chosen discount.

Blackwell optimality and discounted reductions. Blackwell (1962) established the existence of stationary policies that are discounted optimal for every discount sufficiently close to one; such policies are also average-reward optimal. Puterman (1994) gives a standard treatment of these optimality criteria. Grand-Clément & Petrik (2023) derived an explicit upper bound on the Blackwell discount factor from rational model data, above which discounted-optimal policies are Blackwell and average-reward optimal. We use this threshold to choose BPMA's discount. To transfer convergence rates, we additionally bound the average-reward error of approximate discounted policies. Our multiplicative error bound provides this quantitative link, allowing the discounted mirror-ascent rate to yield a global geometric average-reward guarantee.

Organization. Section 2 introduces the setting and regularity results. Sections 3 and 4 present DGPG and BPMA, respectively, and Section 5 reports numerical experiments. Complete proofs and experimental details appear in the appendices.

2 Setting and regularity

Let $M = (\mathcal{S}, \mathcal{A}, P, r)$ be a finite Markov decision process (MDP), with $S = |\mathcal{S}|$, $A = |\mathcal{A}|$, transition probabilities $P(s' \mid s, a)$, and rewards $r(s, a) \in [0, 1]$. Vectors indexed by states are column vectors and appear in bold; matrices are also bold, while scalar components are not. Write $\mathbf{1}$ for the all-ones vector and $\mathbf{I}$ for the identity matrix, with dimensions determined by context. For a finite set $\mathcal{X}$, its probability simplex is

$$\Delta(\mathcal{X}) = \{\mathbf{p} \in \mathbb{R}^{|\mathcal{X}|}_{\geq 0} : \mathbf{1}^\top \mathbf{p} = 1\}.$$

A stationary randomized policy assigns $\pi(\cdot \mid s) \in \Delta(\mathcal{A})$ to each state. We identify π with its stacked action probabilities, so the direct policy class is

$$\Pi = \prod_{s \in \mathcal{S}} \Delta(\mathcal{A}) \subset \mathbb{R}^{SA}.$$

For vectors, $\langle \cdot, \cdot \rangle$ is the Euclidean inner product and $\| \cdot \|_p$, $p \in \{1, 2, \infty\}$, denotes the usual ℓ_p norm. For matrices, $\| \cdot \|_p$ denotes the induced operator norm. We use Euclidean projection Proj_Π and write $\mathrm{supp}(\mathbf{p}) = \{i : p_i > 0\}$ for the support of a probability vector, with the same convention for $\pi(\cdot \mid s)$.

A policy induces the transition matrix and reward vector

$$[\mathbf{P}_\pi]_{ss'} = \sum_a \pi(a \mid s) P(s' \mid s, a), \qquad [\mathbf{r}_\pi]_s = \sum_a \pi(a \mid s) r(s, a).$$

Write $\mathbf{P}_{s,a} = (P(s' \mid s, a))_{s' \in \mathcal{S}}$ for the transition *row* corresponding to (s, a). We work in the exact model-based setting: the model is known and objectives and policy gradients are evaluated exactly.

Assumption 1 (All-policy unichain)**.** *For every $\pi \in \Pi$, the chain induced by $\mathbf{P}_\pi$ has exactly one recurrent class, possibly together with transient states.*

Objectives and optimality gaps. Under Assumption 1, $\mathbf{P}_\pi$ has a unique stationary distribution $\mathbf{d}^\pi \in \Delta(\mathcal{S})$. Its average reward is independent of the initial state and equals

$$J_{\mathrm{av}}(\pi) = (\mathbf{d}^\pi)^\top \mathbf{r}_\pi.$$

Fix the uniform initial distribution $\boldsymbol{\mu} = S^{-1}\mathbf{1}$. For $\gamma \in (0, 1)$, define the discounted value vector and normalized scalar objective by

$$\mathbf{V}^\pi_\gamma = (\mathbf{I} - \gamma \mathbf{P}_\pi)^{-1} \mathbf{r}_\pi, \qquad J_\gamma(\pi) = (1 - \gamma) \boldsymbol{\mu}^\top \mathbf{V}^\pi_\gamma. \tag{1}$$

In particular, $V^\pi_\gamma(s) = [\mathbf{V}^\pi_\gamma]_s$; the same component convention applies to other state vectors. Both J_γ and J_{av} lie in $[0, 1]$. Let $\mathbf{V}^\star_\gamma$ be the componentwise optimal discounted value vector and let $\pi^\star_\gamma$ be a policy attaining it. Define

$$J^\star_\gamma = \max_{\pi \in \Pi} J_\gamma(\pi) = (1 - \gamma) \boldsymbol{\mu}^\top \mathbf{V}^\star_\gamma, \qquad J^\star_{\mathrm{av}} = \max_{\pi \in \Pi} J_{\mathrm{av}}(\pi).$$

The corresponding optimality gaps are

$$\Delta_\gamma(\pi) = J_\gamma^\star - J_\gamma(\pi), \qquad \Delta_{\mathrm{av}}(\pi) = J_{\mathrm{av}}^\star - J_{\mathrm{av}}(\pi). \tag{2}$$

The normalized discounted state-occupancy distribution is

$$\mathbf{d}_{\gamma,\mu}^\pi = (1-\gamma)(\mathbf{I} - \gamma \mathbf{P}_\pi^\top)^{-1}\boldsymbol{\mu}. \tag{3}$$

Its initial-state contribution ensures $d_{\gamma,\mu}^\pi(s) \geq (1-\gamma)\mu(s) > 0$, even when $d^\pi(s) = 0$. Discounted policy gradients therefore retain positive visitation weights at states that are transient under the current policy; this motivates using discounted proposals for average-reward optimization.

From discounted accuracy to average reward. The first lemma bounds the error incurred by optimizing J_γ in place of J_{av}.

Lemma 1 (Uniform discounted-to-average bridge)**.** *Under Assumption 1, there is a finite problem-dependent H such that, for every $\pi \in \Pi$ and $\gamma \in (0,1)$,*

$$|J_\gamma(\pi) - J_{\mathrm{av}}(\pi)| \leq H(1-\gamma). \tag{4}$$

Consequently, every $q \in \Pi$ satisfies

$$\Delta_{\mathrm{av}}(q) \leq \Delta_\gamma(q) + 2H(1-\gamma). \tag{5}$$

The same H works for any initial distribution in place of the uniform $\boldsymbol{\mu}$.

The uniform bound follows from compactness and the unichain resolvent decomposition; Appendix A.2 gives the proof. Thus a sufficiently accurate discounted proposal, at a discount sufficiently close to one, has small average-reward gap.

Smoothness and local gradient domination. Gradients are taken with respect to the stacked policy probabilities, using a smooth extension to a neighborhood of Π as specified in Appendix A.1. Define the first-order stationarity gap

$$\mathcal{F}(\pi) = \max_{\pi' \in \Pi} \langle \nabla J_{\mathrm{av}}(\pi), \pi' - \pi \rangle .$$

Global smoothness makes backtracking well defined from every policy. Near the average-optimal set, gradient domination further converts first-order progress into an objective-gap bound.

Lemma 2 (Average-reward regularity)**.** *Under Assumption 1, ∇J_{av} is L_{av}-Lipschitz on Π in the Euclidean norm for some finite $L_{\mathrm{av}} > 0$. There is also $\varepsilon_{\mathrm{loc}} > 0$ such that, on*

$$\Pi_{\mathrm{loc}} = \{\pi \in \Pi : \Delta_{\mathrm{av}}(\pi) \leq \varepsilon_{\mathrm{loc}}\},$$

we have

$$\Delta_{\mathrm{av}}(\pi) \leq 2\mathcal{F}(\pi). \tag{6}$$

Smoothness follows from the smooth dependence of the stationary distribution on policy parameters. The local bound uses the performance-difference and mismatch argument underlying analyses such as Kumar et al. (2025); Appendix A proves the unichain version needed here. No finite global mismatch coefficient is assumed. Since Π_{loc} is an objective superlevel set, average-reward improvements cannot leave it.

3 Discount-Guided Policy Gradient (DGPG)

Discount-Guided Policy Gradient (DGPG) alternates discounted candidates with direct average-reward improvement. Each epoch has three operations: run discounted projected policy gradient (PPG) from the current policy, accept the resulting proposal only if it does not decrease J_{av}, and take monotone average-reward PPG steps. The discounted candidates ensure progress from arbitrary initialization; the average-reward steps exploit Lemma 2 after the current policy reaches Π_{loc}.

The method is *anytime*: it requires no stopping horizon and outputs the current policy after every update. The algorithm does not test whether the current policy has reached Π_{loc}. Both types of updates are performed in every epoch.

3.1 UPDATES AND EPOCH SCHEDULE

Discounted candidates. Fix any discounted step size $\eta > 0$. In epoch $m \geq 1$, let γ_m be the discount and $q \in \Pi$ the current candidate. A discounted step is

$$q^+ = \mathrm{Proj}_\Pi\left(q + \frac{\eta}{1-\gamma_m}\nabla J_{\gamma_m}(q)\right). \tag{7}$$

The factor $(1-\gamma_m)^{-1}$ makes (7) PPG on the *unnormalized* return $\boldsymbol{\mu}^\top \mathbf{V}^\pi_{\gamma_m}$ with the same fixed step η in every epoch, as in Liu et al. (2025).

Theorem 15 of Liu et al. (2025) implies that discounted PPG on $\boldsymbol{\mu}^\top \mathbf{V}^\pi_\gamma$, with any fixed $\eta > 0$, uniform $\boldsymbol{\mu}$, and any initialization, produces iterates π_k satisfying, for $k \geq 1$,

$$\Delta_\gamma(\pi_k) \leq \frac{C_\eta}{k(1-\gamma)}, \qquad C_\eta := S\left(1 + \frac{S(2+5A)}{\eta}\right). \tag{8}$$

Appendix B.2 gives the normalization leading to (8).

Average-reward improvement (AveragePG). From π, start with $\alpha = 1$, form

$$\pi^+(\alpha) = \mathrm{Proj}_\Pi(\pi + \alpha \nabla J_{\mathrm{av}}(\pi)),$$

and halve α until

$$J_{\mathrm{av}}(\pi^+(\alpha)) \geq J_{\mathrm{av}}(\pi) + \frac{1}{2\alpha}\|\pi^+(\alpha) - \pi\|_2^2. \tag{9}$$

Define $\mathrm{AveragePG}(\pi)$ to be the first policy $\pi^+(\alpha)$ accepted by this backtracking procedure, which restarts from $\alpha = 1$ at each call. Global smoothness in Lemma 2 guarantees termination from every policy; Appendix B.1 gives the argument. This update is monotone and requires no smoothness constant as an input. Any fixed positive initial trial step preserves the convergence guarantee with adjusted constants; Appendix B.1 gives the modification.

Balancing the two sources of error. Let D_m and A_m be the discounted and average-reward update budgets in epoch m, respectively, and let q_m be the terminal discounted candidate. Combining (8) after D_m steps with Lemma 1 gives

$$\Delta_{\mathrm{av}}(q_m) \leq \underbrace{\frac{C_\eta}{D_m(1-\gamma_m)}}_{\text{optimization error}} + \underbrace{2H(1-\gamma_m)}_{\text{discount approximation error}}. \tag{10}$$

Choosing $1-\gamma_m$ proportional to $D_m^{-1/2}$ balances the decay of the two terms without knowing H. We use

$$D_m = 2^{m+3}, \qquad A_m = mD_m, \qquad \gamma_m = 1 - 2^{-m/2}, \qquad m \geq 1. \tag{11}$$

The average-reward budget $A_m = mD_m$ makes the fraction of average-reward updates tend to one, preserving their eventual $O(1/n)$ rate in the total update count.

The current policy is unchanged during the discounted steps until the final candidate is tested. The acceptance decision precedes the output at that update. One update is either a discounted PPG step or a completed call to AVERAGEPG; candidate comparisons and backtracking trials do not increase n. Our guarantees concern this update count rather than the total number of objective evaluations.

3.2 CONVERGENCE GUARANTEE

DGPG achieves the optimal average reward from every initial policy, for any fixed discounted step size $\eta > 0$. We first quantify the number of updates sufficient to reach the near-optimal set Π_{loc}, then establish an $O(1/n)$ bound on the average-reward gap.

Let H, $\varepsilon_{\mathrm{loc}}$, and L_{av} be the constants from Section 2. We may take $0 < \varepsilon_{\mathrm{loc}} \leq 1$ without loss of generality. Define

$$C_{\mathrm{d},\eta} := 2H + \frac{C_\eta}{8}, \qquad C_{\mathrm{av}} := 192S\big(L_{\mathrm{av}} + \max\{1, 2L_{\mathrm{av}}\}\big)^2. \tag{12}$$

Algorithm 1 Discount-Guided Policy Gradient (DGPG)

Require: Initial policy $\widehat{\pi}_0 \in \Pi$; fixed discounted step $\eta > 0$
Ensure: Policy $\widehat{\pi}_n$ after each update $n \geq 1$
$\widehat{\pi} \leftarrow \widehat{\pi}_0$; $n \leftarrow 0$
for $m = 1, 2, \ldots$ **do**
  Set (D_m, A_m, γ_m) according to (11); $q \leftarrow \widehat{\pi}$
  for $j = 1, \ldots, D_m$ **do**
    Update q by the discounted PPG step (7)
    if $j = D_m$ **and** $J_{\mathrm{av}}(q) \geq J_{\mathrm{av}}(\widehat{\pi})$ **then**
      $\widehat{\pi} \leftarrow q$
    end if
    $n \leftarrow n + 1$; $\widehat{\pi}_n \leftarrow \widehat{\pi}$; **output** $\widehat{\pi}_n$
  end for
  for $j = 1, \ldots, A_m$ **do**
    $\widehat{\pi} \leftarrow \textsc{AveragePG}(\widehat{\pi})$
    $n \leftarrow n + 1$; $\widehat{\pi}_n \leftarrow \widehat{\pi}$; **output** $\widehat{\pi}_n$
  end for
end for

The discounted candidate q_m produced in epoch m satisfies

$$\Delta_{\mathrm{av}}(q_m) \leq C_{\mathrm{d},\eta}\, 2^{-m/2}. \tag{13}$$

The current policy satisfies the same bound after the candidate acceptance test; see Appendix B.2.

Define the epoch and update indices

$$m_\eta := \max\left\{1, \left\lceil 2\log_2 \frac{C_{\mathrm{d},\eta}}{\varepsilon_{\mathrm{loc}}} \right\rceil\right\}, \qquad n_\eta := m_\eta\, 2^{m_\eta+3}. \tag{14}$$

For a fixed MDP, the η subscripts emphasize dependence on the discounted step size.

Theorem 1 (Anytime inverse-linear convergence)**.** *Suppose Assumption 1 holds. Run Algorithm 1 from any initial policy $\widehat{\pi}_0 \in \Pi$ with any fixed $\eta > 0$, and let $\widehat{\pi}_n$ denote its output after n updates. With C_{av} and n_η defined in* (12) *and* (14)*, the sequence $J_{\mathrm{av}}(\widehat{\pi}_n)$ is nondecreasing and every output with $n \geq n_\eta$ belongs to Π_{loc}. Moreover,*

$$\Delta_{\mathrm{av}}(\widehat{\pi}_n) \leq \min\left\{1, \frac{2\max\{n_\eta, C_{\mathrm{av}}\}}{n}\right\}, \qquad n \geq 1, \tag{15}$$

$$\Delta_{\mathrm{av}}(\widehat{\pi}_n) \leq \frac{C_{\mathrm{av}}}{n - n_\eta + 3}, \qquad n \geq n_\eta. \tag{16}$$

The proof is given in Appendix B.3.

Interpretation. For each fixed MDP and fixed $\eta > 0$, (15) gives $\Delta_{\mathrm{av}}(\widehat{\pi}_n) = O(1/n)$. The index n_η bounds the number of updates needed to reach Π_{loc}; the algorithm may reach this set earlier. Its parameter dependence is

$$n_\eta = \widetilde{O}\left(\frac{(H + S + S^2A/\eta)^2}{\varepsilon_{\mathrm{loc}}^2}\right). \tag{17}$$

Here $\widetilde{O}$ suppresses a logarithmic factor in $(H + S + S^2A/\eta)/\varepsilon_{\mathrm{loc}}$. The discounted step size affects this bound through $C_{\mathrm{d},\eta}$. Once $n \geq 2n_\eta$, (16) gives

$$\Delta_{\mathrm{av}}(\widehat{\pi}_n) \leq \frac{2C_{\mathrm{av}}}{n} = O\left(\frac{S(1 + L_{\mathrm{av}})^2}{n}\right). \tag{18}$$

Thus the dependence on η enters the guaranteed time needed to reach this regime. The constants H, $\varepsilon_{\mathrm{loc}}$, and L_{av} depend on the MDP, and none is required by DGPG.

An additional bound from discounted candidates. Independently of the local regularity constants, the candidate-accuracy argument gives

$$\Delta_{\mathrm{av}}(\widehat{\pi}_n) \leq \max\{4, 4\sqrt{2}\, C_{\mathrm{d},\eta}\}\sqrt{\frac{\log_2 n}{n}}, \qquad n \geq 16. \tag{19}$$

This bound can be useful before the policy reaches Π_{loc}, because its constant does not involve $\varepsilon_{\mathrm{loc}}$ or L_{av}. The theorem's $O(1/n)$ bound gives the stronger asymptotic conclusion. Appendix B.4 proves (19).

4 Global linear convergence via Blackwell discounting

Section 3 obtains an $O(1/n)$ average-reward gap by varying the discount. We now obtain geometric decay using a single fixed discount. The bridge in Lemma 1 leaves an additive error $2H(1-\gamma)$ at fixed γ. Our key step is a multiplicative bound $\Delta_{\mathrm{av}}(\pi) \le K_\gamma \Delta_\gamma(\pi)$ at a discount above the Blackwell threshold.

A deterministic policy is *Blackwell-optimal* if it is discounted optimal for every discount sufficiently close to one. Let γ_{BW} be the infimum of thresholds above which every discounted-optimal deterministic policy is Blackwell-optimal. The following additional assumption makes such a discount explicitly computable.

Assumption 2 (Rational model)**.** *There is a common integer denominator $m \ge 1$ such that*

$$P(s' \mid s,a) = \frac{n_{sas'}}{m}, \qquad r(s,a) = \frac{q_{sa}}{m}, \tag{20}$$

where $n_{sas'}$ and q_{sa} are integers. Let $R_{\mathrm{num}} = \max_{s,a} |q_{sa}|$.

For rational data with distinct denominators, m can be any common denominator. Theorem 4.6 of Grand-Clément & Petrik (2023) gives $\gamma_{\mathrm{BW}} < \bar\gamma < 1$ for

$$\bar\gamma := 1 - \frac{1}{2n_{\mathrm{poly}}^{n_{\mathrm{poly}}/2+2}(L_M+1)^{n_{\mathrm{poly}}}}, \qquad n_{\mathrm{poly}} := 2S-1, \qquad L_M := 2SR_{\mathrm{num}}m^{2S}4^S. \tag{21}$$

Every deterministic $\bar\gamma$-discounted-optimal policy is therefore Blackwell-optimal and average-reward optimal. Appendix C.1 records the specialization of the cited bound.

4.1 Blackwell Policy Mirror Ascent

Blackwell Policy Mirror Ascent (BPMA) applies discounted policy mirror ascent at the fixed discount $\gamma = \bar\gamma$. For any $\gamma \in (0,1)$, the discounted action values are

$$Q_\gamma^\pi(s,a) = r(s,a) + \gamma \mathbf{P}_{s,a}\mathbf{V}_\gamma^\pi. \tag{22}$$

Given a full-support initial policy π_0, meaning $\pi_0(a \mid s) > 0$ for every (s,a), and any $\eta_0 > 0$, the updates are

$$\pi_{k+1}(a \mid s) = \frac{\pi_k(a \mid s)\exp\{\eta_k Q_\gamma^{\pi_k}(s,a)\}}{\sum_{b\in\mathcal{A}} \pi_k(b \mid s)\exp\{\eta_k Q_\gamma^{\pi_k}(s,b)\}}, \qquad \eta_k = \eta_0\gamma^{-k}. \tag{23}$$

This is the entropic policy mirror-ascent update covered by Xiao (2022), with a geometrically increasing step size. BPMA uses discounted action values to generate its policies; average-reward evaluation is needed only if one wishes to monitor their average reward.

Algorithm 2 Blackwell Policy Mirror Ascent (BPMA)

Require: Rational MDP satisfying Assumption 2; $\eta_0 > 0$; full-support $\pi_0 \in \Pi$
1: Set $\gamma \leftarrow \bar\gamma$ according to (21)
2: **for** $k = 0, 1, 2, \ldots$ **do**
3: Evaluate $Q_\gamma^{\pi_k}$ using (22)
4: Compute π_{k+1} using (23)
5: **end for**

4.2 Convergence guarantee

We first relate discounted and average-reward gaps. Lemma 2 and compactness of Π give the finite constant

$$L_J := \max_{\pi\in\Pi} \|\nabla J_{\mathrm{av}}(\pi)\|_\infty.$$

Consequently,

$$|J_{\mathrm{av}}(\pi) - J_{\mathrm{av}}(\pi')| \le L_J \|\pi - \pi'\|_1, \qquad \pi, \pi' \in \Pi. \tag{24}$$

Define the discounted Bellman loss and, when at least one loss is positive, its smallest positive value:

$$\ell_\gamma(s,a) := V_\gamma^\star(s) - r(s,a) - \gamma \mathbf{P}_{s,a}\mathbf{V}_\gamma^\star, \qquad \zeta_\gamma := \min_{s,a:\, \ell_\gamma(s,a)>0} \ell_\gamma(s,a). \tag{25}$$

Finiteness of the action and state spaces ensures $\zeta_\gamma > 0$ whenever this set is nonempty; no separate action-gap assumption is needed.

Lemma 3 (Discounted-to-average error bound)**.** *Suppose Assumptions 1 and 2 hold. At $\gamma = \bar{\gamma}$ from* (21)*, if a positive Bellman loss exists, then every $\pi \in \Pi$ satisfies*

$$\Delta_{\mathrm{av}}(\pi) \le K_\gamma \Delta_\gamma(\pi), \qquad K_\gamma := \frac{2L_J S}{(1-\gamma)\zeta_\gamma}. \tag{26}$$

If all Bellman losses vanish, every policy is average-reward optimal.

The proof is given in Appendix C.2. Unlike the additive bridge in Section 2, this bound transfers arbitrarily small discounted errors to average reward at a fixed discount.

Let $p_{\min} := \min_{s,a} \pi_0(a \mid s) > 0$. At any fixed $\gamma \in (0,1)$, Theorem 10 of Xiao (2022), specialized to (23), gives

$$\Delta_\gamma(\pi_k) \le \left(1 + \frac{(1-\gamma)\log(1/p_{\min})}{\eta_0 \gamma}\right)\left(1 - \frac{1-\gamma}{S}\right)^k. \tag{27}$$

Appendix C.3 verifies the step-size condition and normalization and retains the sharper initialization-dependent bound.

Theorem 2 (Global linear average-reward convergence)**.** *Suppose Assumptions 1 and 2 hold. Run Algorithm 2 from any full-support policy π_0 with any $\eta_0 > 0$. If a positive Bellman loss exists at $\bar{\gamma}$, define*

$$C_{\bar{\gamma}} := \frac{2L_J S}{(1-\bar{\gamma})\zeta_{\bar{\gamma}}}\left(1 + \frac{(1-\bar{\gamma})\log(1/p_{\min})}{\eta_0 \bar{\gamma}}\right). \tag{28}$$

Then, for every $k \ge 0$,

$$\Delta_{\mathrm{av}}(\pi_k) \le C_{\bar{\gamma}}\left(1 - \frac{1-\bar{\gamma}}{S}\right)^k \le C_{\bar{\gamma}} \exp\left(-\frac{1-\bar{\gamma}}{S}k\right). \tag{29}$$

If all Bellman losses vanish, $\Delta_{\mathrm{av}}(\pi_k) = 0$ for every $k \ge 0$.

The proof is given in Appendix C.3.

In the first case of the theorem, for $0 < \epsilon < C_{\bar{\gamma}}$, the following number of updates is sufficient for $\Delta_{\mathrm{av}}(\pi_k) \le \epsilon$:

$$k \ge \left\lceil \frac{S}{1-\bar{\gamma}} \log \frac{C_{\bar{\gamma}}}{\epsilon} \right\rceil. \tag{30}$$

If $\epsilon \ge C_{\bar{\gamma}}$, the initial policy already satisfies the guarantee. For each fixed MDP, initialization, and η_0, the required update count is $O(\log(1/\epsilon))$ as $\epsilon \downarrow 0$.

Interpretation. The average-reward gap has a geometric envelope from the initial iterate, for every full-support initialization. This is global R-linear convergence; it does not assert contraction of consecutive average-reward gaps. For uniform π_0, the initialization term simplifies to $\log(1/p_{\min}) = \log A$. The base step η_0 affects the prefactor, while the displayed contraction factor depends on S and $\bar{\gamma}$. The constants L_J and $\zeta_{\bar{\gamma}}$ are needed only for the analysis. This guarantee assumes exact arithmetic and exact policy evaluation. The explicit discount (21) can be extremely close to one, making both the guaranteed update count and the required numerical precision large; the linear rate alone does not imply a practical computational advantage.

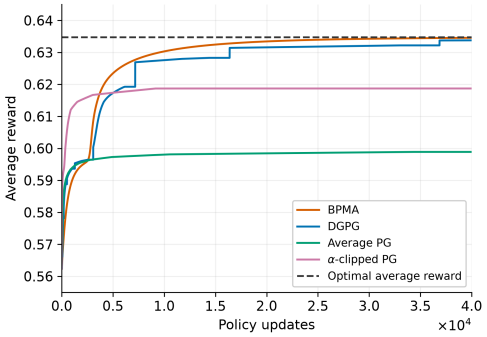


Figure 1: Experiment 1 (long route): original-scale average reward versus policy updates.

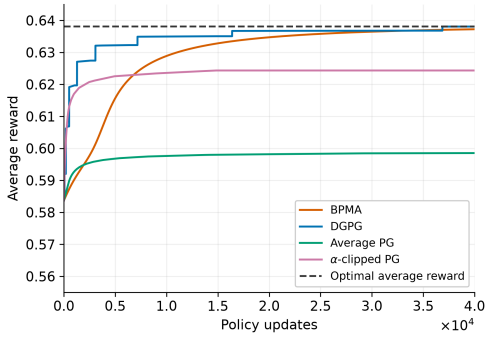


Figure 2: Experiment 2 (direct route): original-scale average reward versus policy updates.

5 Experiments

We compare BPMA, DGPG, projected average PG, and the α-clipped method of Lee & Ryu (2026) on two rational MDPs with 30 states and 20 actions. Both satisfy the all-policy unichain assumption and have suboptimal local maxima with transient states. Experiment 1 has a long route to the higher-reward region; Experiment 2 has a direct route. Reward shaping produces contrasting numerical Blackwell-threshold estimates, approximately 0.8093 and 0.2600, respectively. These estimates describe the instances and are not inputs to either algorithm. Figure 3 in Appendix D illustrates the MDPs; that appendix also specifies their transitions, rewards, and evaluation protocol.

All methods start uniformly and perform $40{,}000$ updates. BPMA uses $\bar{\gamma}$ from (21) and $\eta_k = \eta_0 \bar{\gamma}^{-k}$, with $\eta_0 = 0.02$ and 0.1 in Experiments 1 and 2. Centered Bellman evaluation stores $1-\bar{\gamma}$ separately to avoid explicitly forming the discount at working precision. The discount-dependent corrections are below this precision, so the curves are numerically indistinguishable from the $\gamma \uparrow 1$ limit; Appendix D.3 reports the precision checks. DGPG uses $\eta = \eta_0$ and initializes backtracking at η_0 rather than one, as permitted by Appendix B.1. The direct baselines use step η_0; the clipping floor is 0.005 in Experiment 1 and 0.01 in Experiment 2. Comparisons count policy updates, not running times.

Figures 1–2 report average rewards on the original reward scale. In Experiment 1, BPMA reaches gap 10^{-3} after $22{,}717$ updates, compared with $38{,}702$ for DGPG, and finishes with the smaller gap (1.98×10^{-4} versus 9.84×10^{-4}). In Experiment 2, DGPG reaches the optimum to numerical precision by update $36{,}864$, while BPMA finishes with gap 8.41×10^{-4}. Direct average PG remains suboptimal in both instances. The clipped method reaches its constrained optimum, leaving a nonzero unrestricted gap. The relative performance of BPMA and DGPG therefore depends on the instance; a global linear guarantee does not imply a uniform practical advantage.

6 Conclusion

We established two routes from discounted policy optimization to average-reward convergence in finite MDPs with transient states. Under the all-policy unichain assumption, DGPG achieves an anytime global $O(1/n)$ gap bound from any initial policy, with any fixed positive discounted step size. Under an additional rational-model assumption, BPMA achieves a global geometric average-reward bound from every full-support initial policy at one explicit Blackwell discount. Both results avoid a finite global stationary-distribution mismatch assumption. The analysis combines local average-reward regularity with quantitative bounds relating discounted and average-reward errors. Extending these exact-model guarantees to sampled updates and obtaining less conservative discount choices are natural directions for future work.

REPRODUCIBILITY STATEMENT

Sections 2–4 specify the assumptions and algorithmic updates, with proofs in Appendices A–C. Appendix D gives the rational MDPs, initialization, step sizes, numerical precision, and complete evaluation protocol. The accompanying code reproduces the experiments and figures, including centered finite-discount BPMA evaluation, comparisons across working precisions, and higher-precision Bellman checks. The experiments are deterministic and use no sampling or external datasets.

AI USE STATEMENT

Generative AI tools assisted with literature retrieval, drafting and revising manuscript text, reorganizing and simplifying proofs, deriving and checking intermediate mathematical bounds, adapting experimental code, independently verifying numerical results, and regenerating figures. They also assisted with interpreting numerical results and preparing LaTeX. The authors are responsible for independently verifying the mathematical arguments, citations, code, and numerical claims, and take responsibility for the final content.

ETHICS STATEMENT

This work develops convergence theory and evaluates algorithms on synthetic MDPs. It involves no human participants, personal data, or deployment in real-world decision systems.

REFERENCES

Alekh Agarwal, Sham M. Kakade, Jason D. Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22(98):1–76, 2021.

David Blackwell. Discrete dynamic programming. *The Annals of Mathematical Statistics*, 33(2): 719–726, 1962.

Julien Grand-Clément and Marek Petrik. Reducing blackwell and average optimality to discounted MDPs via the blackwell discount factor. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2302.00036.

Navdeep Kumar, Yashaswini Murthy, Itai Shufaro, Kfir Y. Levy, R. Srikant, and Shie Mannor. On the global convergence of policy gradient in average reward markov decision processes. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=2PRpcmJecX.

Jongmin Lee and Ernest K. Ryu. Policy gradient algorithms in average-reward multichain MDPs. *arXiv preprint arXiv:2602.18003*, 2026. URL https://arxiv.org/abs/2602.18003.

Jiacai Liu, Wenye Li, Dachao Lin, Ke Wei, and Zhihua Zhang. On the convergence of projected policy gradient for any constant step sizes. *Journal of Machine Learning Research*, 26(193):1–35, 2025. URL https://jmlr.org/papers/v26/24-1530.html.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvári, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 6820–6829, 2020.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 1994.

Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, volume 12, 2000.

Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022. URL https://jmlr.org/papers/v23/22-0056.html.

A PROOFS FOR SETTING AND REGULARITY

This appendix proves the two regularity lemmas in Section 2. We establish the smoothness and local gradient-domination claims together, then use continuity of the stationary distribution to prove the uniform discounted-to-average approximation.

A.1 AVERAGE-REWARD REGULARITY

Lemma 2 (restated). *Under Assumption 1, ∇J_{av} is L_{av}-Lipschitz on Π in the Euclidean norm for some finite $L_{\mathrm{av}} > 0$. There is also $\varepsilon_{\mathrm{loc}} > 0$ such that, on*

$$\Pi_{\mathrm{loc}} = \{\pi \in \Pi : \Delta_{\mathrm{av}}(\pi) \le \varepsilon_{\mathrm{loc}}\},$$

we have

$$\Delta_{\mathrm{av}}(\pi) \le 2\mathcal{F}(\pi).$$

Proof. **Global smoothness.** Under Assumption 1, eigenvalue one of $\mathbf{P}_\pi$ is simple. The matrix

$$\mathbf{G}_\pi = \mathbf{I} - \mathbf{P}_\pi^\top + \mathbf{1}\mathbf{1}^\top$$

is therefore invertible, and the stationary distribution is $\mathbf{d}^\pi = \mathbf{G}_\pi^{-1}\mathbf{1}$. Since $\mathbf{P}_\pi$ and $\mathbf{r}_\pi$ are affine in the direct policy parameters, this formula extends $J_{\mathrm{av}}(\pi) = (\mathbf{G}_\pi^{-1}\mathbf{1})^\top \mathbf{r}_\pi$ smoothly to an open neighborhood of Π. The Hessian has bounded Euclidean operator norm on the compact policy simplex. Convexity of Π thus gives a finite $L_{\mathrm{av}} > 0$ such that

$$\|\nabla J_{\mathrm{av}}(\pi) - \nabla J_{\mathrm{av}}(\pi')\|_2 \le L_{\mathrm{av}}\|\pi - \pi'\|_2, \qquad \pi, \pi' \in \Pi.$$

Two smooth extensions that agree on Π have gradients differing only by rowwise constants. They therefore give the same feasible directional derivatives and Euclidean projected updates.

Local gradient domination. Let $\Pi_{\mathrm{av}}^\star := \arg\max_{\pi\in\Pi} J_{\mathrm{av}}(\pi)$. For each $\bar\pi \in \Pi_{\mathrm{av}}^\star$, continuity of the stationary distribution gives a relatively open neighborhood on which

$$d^\pi(s) \ge \tfrac{1}{2} d^{\bar\pi}(s) \qquad \text{for every } s \in \mathrm{supp}(\mathbf{d}^{\bar\pi}).$$

Let $\mathcal{U}$ be the union of these neighborhoods over all $\bar\pi \in \Pi_{\mathrm{av}}^\star$. It is relatively open and contains $\Pi_{\mathrm{av}}^\star$; every $\pi \in \mathcal{U}$ admits an optimal $\bar\pi$ with

$$\max_{s\in\mathrm{supp}(\mathbf{d}^{\bar\pi})} \frac{d^{\bar\pi}(s)}{d^\pi(s)} \le 2. \tag{31}$$

The bias vector $\mathbf{h}^\pi$, normalized by $(\mathbf{d}^\pi)^\top \mathbf{h}^\pi = 0$, is the unique solution of the Poisson equation

$$J_{\mathrm{av}}(\pi)\mathbf{1} + \mathbf{h}^\pi = \mathbf{r}_\pi + \mathbf{P}_\pi \mathbf{h}^\pi. \tag{32}$$

Define the average-reward advantage

$$A_{\mathrm{av}}^\pi(s,a) := r(s,a) - J_{\mathrm{av}}(\pi) + \mathbf{P}_{s,a}\mathbf{h}^\pi - h^\pi(s).$$

The Poisson equation gives $\sum_a \pi(a \mid s) A_{\mathrm{av}}^\pi(s,a) = 0$, so $\max_a A_{\mathrm{av}}^\pi(s,a) \ge 0$ at every state. The performance-difference and policy-gradient identities are

$$J_{\mathrm{av}}(\pi') - J_{\mathrm{av}}(\pi) = \sum_s d^{\pi'}(s) \sum_a \pi'(a \mid s) A_{\mathrm{av}}^\pi(s,a),$$

$$\langle \nabla J_{\mathrm{av}}(\pi), \pi' - \pi \rangle = \sum_s d^\pi(s) \sum_a [\pi'(a \mid s) - \pi(a \mid s)] A_{\mathrm{av}}^\pi(s,a).$$

Maximizing the second identity over π' yields $\mathcal{F}(\pi) = \sum_s d^\pi(s) \max_a A_{\mathrm{av}}^\pi(s,a)$. Applying the first identity with $\pi' = \bar\pi$ and then (31) gives

$$\Delta_{\mathrm{av}}(\pi) \le \sum_s d^{\bar\pi}(s) \max_a A_{\mathrm{av}}^\pi(s,a) \le 2 \sum_s d^\pi(s) \max_a A_{\mathrm{av}}^\pi(s,a) = 2\mathcal{F}(\pi).$$

If $\Pi \setminus \mathcal{U}$ is nonempty, its compactness gives $\max_{\pi\in\Pi\setminus\mathcal{U}} J_{\mathrm{av}}(\pi) < J_{\mathrm{av}}^\star$. Choose $\varepsilon_{\mathrm{loc}} > 0$ smaller than this strict gap; if the complement is empty, any positive choice suffices. Then $\Pi_{\mathrm{loc}} \subseteq \mathcal{U}$, proving (6). Its superlevel-set definition also ensures that average-reward improvements cannot leave Π_{loc}. □

A.2 Uniform discounted-to-average approximation

Lemma 1 (restated). *Under Assumption 1, there is a finite problem-dependent H such that, for every $\pi \in \Pi$ and $\gamma \in (0,1)$,*

$$|J_\gamma(\pi) - J_{\mathrm{av}}(\pi)| \le H(1-\gamma).$$

Consequently, every $q \in \Pi$ satisfies

$$\Delta_{\mathrm{av}}(q) \le \Delta_\gamma(q) + 2H(1-\gamma).$$

The same H works for any initial distribution in place of the uniform $\boldsymbol{\mu}$.

Proof. Put $\mathbf{E}_\pi = \mathbf{1}(\mathbf{d}^\pi)^\top$. The resolvent decomposition is

$$(\mathbf{I} - \gamma \mathbf{P}_\pi)^{-1} = \frac{1}{1-\gamma}\mathbf{E}_\pi + \mathbf{R}_{\gamma,\pi}, \qquad \mathbf{R}_{\gamma,\pi} = [\mathbf{I} - \gamma(\mathbf{P}_\pi - \mathbf{E}_\pi)]^{-1}(\mathbf{I} - \mathbf{E}_\pi).$$

The matrix in brackets is invertible for every $(\pi,\gamma) \in \Pi \times [0,1]$: the eigenvalues of $\mathbf{P}_\pi - \mathbf{E}_\pi$ lie in the closed unit disk and none equals one. The stationary-distribution representation above makes $\mathbf{E}_\pi$ continuous in π. Continuity on the compact set $\Pi \times [0,1]$ therefore gives

$$H := \sup_{\pi \in \Pi,\, \gamma \in [0,1]} \|\mathbf{R}_{\gamma,\pi}\|_\infty < \infty.$$

Since $\|\mathbf{r}_\pi\|_\infty \le 1$ and every initial distribution satisfies $\|\boldsymbol{\mu}\|_1 = 1$,

$$\left|\boldsymbol{\mu}^\top \mathbf{R}_{\gamma,\pi}\mathbf{r}_\pi\right| \le H.$$

Multiplying the resolvent decomposition by $(1-\gamma)\boldsymbol{\mu}^\top$ on the left and $\mathbf{r}_\pi$ on the right proves (4). Taking maxima in the uniform bound gives $|J^\star_\gamma - J^\star_{\mathrm{av}}| \le H(1-\gamma)$. Hence

$$\Delta_{\mathrm{av}}(q) - \Delta_\gamma(q) = (J^\star_{\mathrm{av}} - J^\star_\gamma) + (J_\gamma(q) - J_{\mathrm{av}}(q)) \le 2H(1-\gamma),$$

which proves (5). □

B Proofs for anytime inverse-linear convergence

This appendix proves the guarantees in Section 3, using the regularity lemmas proved in Appendix A.

B.1 Backtracking and convergence in the near-optimal set

Global smoothness controls every backtracking trial; local gradient domination is needed only after entry. Define

$$\alpha_{\min} := \min\left\{1, \frac{1}{2L_{\mathrm{av}}}\right\}, \qquad \kappa_{\mathrm{loc}} := \frac{1}{32S(L_{\mathrm{av}} + \alpha_{\min}^{-1})^2}. \tag{33}$$

These constants satisfy $C_{\mathrm{av}} = 6/\kappa_{\mathrm{loc}}$.

Lemma 4 (Recursion in the near-optimal set)**.** *Let a_0 be the gap at entry into Π_{loc}, and let a_k be the gap immediately after the kth subsequent average-reward update. Discounted candidates may be accepted between these updates. For every $k \ge 0$,*

$$a_k - a_{k+1} \ge \kappa_{\mathrm{loc}} a_{k+1}^2, \tag{34}$$

and

$$a_k \le \frac{2}{\kappa_{\mathrm{loc}}(k+1)}. \tag{35}$$

Proof. For a backtracking trial from $x \in \Pi$, put $y = \mathrm{Proj}_\Pi(x + \alpha \nabla J_{\mathrm{av}}(x))$. Projection optimality and global smoothness give

$$\langle \nabla J_{\mathrm{av}}(x), y - x\rangle \ge \alpha^{-1}\|y-x\|_2^2, \qquad J_{\mathrm{av}}(y) - J_{\mathrm{av}}(x) \ge \left(\alpha^{-1} - \frac{L_{\mathrm{av}}}{2}\right)\|y-x\|_2^2.$$

Thus every $\alpha \le 1/L_{\mathrm{av}}$ passes (9), even if the trial leaves Π_{loc}. Halving from one terminates with $\alpha \in [\alpha_{\min}, 1]$.

Let x_k and y_k be the policies immediately before and after the $(k+1)$st average-reward update after entry, and write $u_k = y_k - x_k$. Acceptance gives

$$J_{\mathrm{av}}(y_k) - J_{\mathrm{av}}(x_k) \ge \tfrac{1}{2}\|u_k\|_2^2. \tag{36}$$

For every $\pi' \in \Pi$, projection optimality also gives

$$\langle \nabla J_{\mathrm{av}}(x_k), \pi' - y_k \rangle \le \alpha_k^{-1} \langle u_k, \pi' - y_k \rangle.$$

Using Lipschitz continuity of the gradient and the Euclidean diameter bound $\mathrm{diam}(\Pi) \le 2\sqrt{S}$, we obtain

$$\mathcal{F}(y_k) \le 2\sqrt{S}\,(L_{\mathrm{av}} + \alpha_{\min}^{-1})\|u_k\|_2.$$

Monotonicity keeps $y_k \in \Pi_{\mathrm{loc}}$. Consequently,

$$a_{k+1} = \Delta_{\mathrm{av}}(y_k) \le 2\mathcal{F}(y_k) \le 4\sqrt{S}\,(L_{\mathrm{av}} + \alpha_{\min}^{-1})\|u_k\|_2.$$

Intervening accepted candidates can only decrease the gap, so $\Delta_{\mathrm{av}}(x_k) \le a_k$. Squaring the last display and using (36) proves

$$a_k - a_{k+1} \ge J_{\mathrm{av}}(y_k) - J_{\mathrm{av}}(x_k) \ge \kappa_{\mathrm{loc}} a_{k+1}^2.$$

For (35), set $b_k = 2/[\kappa_{\mathrm{loc}}(k+1)]$. Bounded rewards give $a_0 \le 1 \le 2/\kappa_{\mathrm{loc}} = b_0$, and

$$b_{k+1} + \kappa_{\mathrm{loc}} b_{k+1}^2 - b_k = \frac{2k}{\kappa_{\mathrm{loc}}(k+1)(k+2)^2} \ge 0.$$

Since $t \mapsto t + \kappa_{\mathrm{loc}} t^2$ is increasing for $t \ge 0$, (34) proves $a_k \le b_k$ by induction. □

Other backtracking initializations. Suppose each call starts from a fixed $\bar{\alpha} > 0$. Halving then accepts a step in $[\min\{\bar{\alpha}, (2L_{\mathrm{av}})^{-1}\}, \bar{\alpha}]$. The preceding proof applies with improvement at least $\|u_k\|_2^2/(2\max\{1, \bar{\alpha}\})$. Consequently, Theorem 1 remains valid with the same n_η and with C_{av} replaced by

$$C_{\mathrm{av}}(\bar{\alpha}) := 192 S \max\{1, \bar{\alpha}\}\big(L_{\mathrm{av}} + \max\{1, \bar{\alpha}^{-1}, 2L_{\mathrm{av}}\}\big)^2.$$

Indeed, this gives the recursion constant $6/C_{\mathrm{av}}(\bar{\alpha})$, and $C_{\mathrm{av}}(\bar{\alpha})/3 \ge 1$ supplies the induction base in (35). The counting argument below is unchanged. At $\bar{\alpha} = 1$ this recovers the stated constant. This extension concerns exact evaluation and the exact acceptance condition (9).

B.2 Discounted candidate accuracy

The candidate q_m in epoch m satisfies $\Delta_{\mathrm{av}}(q_m) \le C_{\mathrm{d},\eta} 2^{-m/2}$, and the same bound holds for the current policy after the candidate acceptance test.

Proof of (13). First, uniform $\boldsymbol{\mu}$ gives $\min_s \mu(s) = 1/S$ and $\|\mathbf{d}_{\gamma,\mu}^{\pi_\gamma^\star}/\boldsymbol{\mu}\|_\infty \le S$ for any discounted-optimal policy $\pi_\gamma^\star$. Substituting these bounds into Theorem 15 of Liu et al. (2025) and multiplying its unnormalized value-gap bound by $1-\gamma$ gives (8).

Apply (8) with $k = D_m$ and $\gamma = \gamma_m$, then use Lemma 1:

$$\Delta_{\mathrm{av}}(q_m) \le \frac{C_\eta}{D_m(1-\gamma_m)} + 2H(1-\gamma_m) = C_{\mathrm{d},\eta} 2^{-m/2}.$$

The acceptance test either accepts q_m or keeps a policy with no smaller average reward, proving the same bound for the current policy. □

B.3 ANYTIME INVERSE-LINEAR CONVERGENCE

Theorem 1 (restated). *Suppose Assumption 1 holds. Run Algorithm 1 from any initial policy $\widehat{\pi}_0 \in \Pi$ with any fixed $\eta > 0$, and let $\widehat{\pi}_n$ denote its output after n updates. With C_{av} and n_η defined in (12) and (14), the sequence $J_{\mathrm{av}}(\widehat{\pi}_n)$ is nondecreasing and every output with $n \geq n_\eta$ belongs to Π_{loc}. Moreover,*

$$\Delta_{\mathrm{av}}(\widehat{\pi}_n) \leq \min\left\{1, \frac{2\max\{n_\eta, C_{\mathrm{av}}\}}{n}\right\}, \qquad n \geq 1,$$
$$\Delta_{\mathrm{av}}(\widehat{\pi}_n) \leq \frac{C_{\mathrm{av}}}{n - n_\eta + 3}, \qquad n \geq n_\eta.$$

Proof. **Monotonicity and the time to reach Π_{loc}.** The candidate acceptance test and every completed AVERAGEPG update make the output objective nondecreasing. The numbers of updates through epoch M and through the discounted steps of epoch m are

$$T_M = \sum_{j=1}^{M}(D_j + A_j) = M2^{M+4}, \qquad N_m = T_{m-1} + D_m = m2^{m+3}, \tag{37}$$

with $T_0 = 0$. At $n_\eta = N_{m_\eta}$, the output includes the epoch-m_η candidate acceptance decision. By (13) and (14),

$$\Delta_{\mathrm{av}}(\widehat{\pi}_{n_\eta}) \leq C_{\mathrm{d},\eta} 2^{-m_\eta/2} \leq \varepsilon_{\mathrm{loc}}.$$

Monotonicity keeps every subsequent output in Π_{loc}.

For the order bound on n_η, put $b := \max\{1, C_{\mathrm{d},\eta}/\varepsilon_{\mathrm{loc}}\}$. The definition of m_η gives $m_\eta \leq 1 + 2\log_2 b$ and $2^{m_\eta} \leq 2b^2$. Hence

$$n_\eta \leq 16b^2(1 + 2\log_2 b). \tag{38}$$

Since $S, A \geq 1$ and $0 < \varepsilon_{\mathrm{loc}} \leq 1$, we have $b = O((H + S + S^2A/\eta)/\varepsilon_{\mathrm{loc}})$. Substitution proves (17), with a single logarithmic factor.

Counting average-reward updates. Let $K(n)$ count average-reward updates after n_η and through update n. We claim

$$K(n) \geq \frac{n - n_\eta}{3}, \qquad n \geq n_\eta. \tag{39}$$

Starting immediately after $n_\eta = N_{m_\eta}$, pair the A_j average-reward updates of epoch j with the following D_{j+1} discounted updates, for $j = m_\eta, m_\eta + 1, \ldots$. The schedule gives

$$D_{j+1} = 2D_j \leq 2A_j.$$

Thus at least one third of each complete pair consists of average-reward updates. The same is true of every prefix of a pair: before the discounted updates begin the fraction is one, and it is at least $A_j/(A_j + D_{j+1}) \geq 1/3$ thereafter. Summing over complete pairs and the final partial pair proves (39); at $n = n_\eta$, both sides are zero.

Inverse-linear convergence. Apply Lemma 4 with $a_0 = \Delta_{\mathrm{av}}(\widehat{\pi}_{n_\eta})$. Between average-reward updates, accepted candidates can only decrease the gap. Combining (35) and (39) gives

$$\Delta_{\mathrm{av}}(\widehat{\pi}_n) \leq \frac{2/\kappa_{\mathrm{loc}}}{K(n) + 1} \leq \frac{6/\kappa_{\mathrm{loc}}}{n - n_\eta + 3} = \frac{C_{\mathrm{av}}}{n - n_\eta + 3}, \qquad n \geq n_\eta.$$

This proves (16). For $n \geq 2n_\eta$, it implies $\Delta_{\mathrm{av}}(\widehat{\pi}_n) \leq 2C_{\mathrm{av}}/n$. For $1 \leq n < 2n_\eta$, bounded rewards give $\Delta_{\mathrm{av}}(\widehat{\pi}_n) \leq 1 \leq 2n_\eta/n$. Combining the two ranges with the universal bound $\Delta_{\mathrm{av}} \leq 1$ proves (15). All constants are finite for a fixed MDP and fixed $\eta > 0$, so the gap is $O(1/n)$. □

B.4 Additional bound from discounted candidates

Proof of (19). For $m \geq 2$ and $T_{m-1} < n \leq T_m$, the preceding epoch's candidate acceptance test and monotonicity give $\Delta_{\mathrm{av}}(\widehat{\pi}_n) \leq C_{\mathrm{d},\eta} 2^{-(m-1)/2}$. Because $n \leq m2^{m+4}$ and $m \leq \log_2 n$,

$$C_{\mathrm{d},\eta} 2^{-(m-1)/2} \leq 4\sqrt{2} C_{\mathrm{d},\eta} \sqrt{\frac{m}{n}} \leq 4\sqrt{2} C_{\mathrm{d},\eta} \sqrt{\frac{\log_2 n}{n}}.$$

For $16 \leq n \leq T_1 = 32$, use $0 \leq \Delta_{\mathrm{av}} \leq 1$ and $4\sqrt{\log_2 n / n} \geq 1$. This proves (19) for every $n \geq 16$. □

C Proofs for global linear convergence

C.1 Choosing the Blackwell discount

Apply Theorem 4.6 of Grand-Clément & Petrik (2023) with $N = n_{\mathrm{poly}}$, $r_\infty = R_{\mathrm{num}}$, and $L = L_M$. Its bound gives $\gamma_{\mathrm{BW}} < \bar{\gamma} < 1$. Thus every deterministic policy that is discounted optimal at $\bar{\gamma}$ is Blackwell-optimal and consequently average-reward optimal.

C.2 Proof of the discounted-to-average error bound

Lemma 3 (restated). *Suppose Assumptions 1 and 2 hold. At $\gamma = \bar{\gamma}$ from* (21)*, if a positive Bellman loss exists, then every $\pi \in \Pi$ satisfies*

$$\Delta_{\mathrm{av}}(\pi) \leq K_\gamma \Delta_\gamma(\pi), \qquad K_\gamma := \frac{2L_J S}{(1-\gamma)\zeta_\gamma}.$$

If all Bellman losses vanish, every policy is average-reward optimal.

Proof. Fix $\gamma = \bar{\gamma}$ and define the discounted-optimal action sets

$$\mathcal{A}^\star_\gamma(s) := \{a \in \mathcal{A} : \ell_\gamma(s,a) = 0\}.$$

These sets are nonempty by the Bellman optimality equation. Every deterministic policy choosing only actions in these sets is discounted optimal and, by Appendix C.1, average-reward optimal.

We first extend this conclusion to randomized policies. Let π be supported on $\mathcal{A}^\star_\gamma(s)$ at every state. The Poisson equation (32) implies that, for each s, some $a_s \in \operatorname{supp} \pi(\cdot \mid s)$ satisfies

$$r(s,a_s) + \mathbf{P}_{s,a_s} \mathbf{h}^\pi \leq J_{\mathrm{av}}(\pi) + h^\pi(s).$$

The deterministic policy $\pi^d(s) = a_s$ is average-reward optimal. Multiplying this inequality by its stationary distribution and summing cancels the bias terms, giving

$$J^\star_{\mathrm{av}} = J_{\mathrm{av}}(\pi^d) \leq J_{\mathrm{av}}(\pi) \leq J^\star_{\mathrm{av}}.$$

Therefore every policy supported on the discounted-optimal actions is average-reward optimal. This also proves the claim when all Bellman losses vanish.

Now suppose a positive loss exists, and let $\pi \in \Pi$ be arbitrary. At each state, move all probability assigned to actions outside $\mathcal{A}^\star_\gamma(s)$ to one action in that set. Denote the resulting policy by $\widetilde{\pi}$. Then $J_{\mathrm{av}}(\widetilde{\pi}) = J^\star_{\mathrm{av}}$, and its ℓ_1 distance from π is twice the total probability moved.

The policy Bellman equation and the definition of discounted occupancy give

$$\begin{aligned} \Delta_\gamma(\pi) &= \sum_{s,a} d^\pi_{\gamma,\mu}(s) \pi(a \mid s) \ell_\gamma(s,a) \\ &\geq \frac{(1-\gamma)\zeta_\gamma}{S} \sum_s \sum_{a \notin \mathcal{A}^\star_\gamma(s)} \pi(a \mid s) \\ &= \frac{(1-\gamma)\zeta_\gamma}{2S} \|\pi - \widetilde{\pi}\|_1, \end{aligned} \tag{40}$$

where the inequality uses $d^{\pi}_{\gamma,\mu}(s) \geq (1-\gamma)/S$. Applying the Lipschitz bound (24) yields

$$\begin{aligned}\Delta_{\mathrm{av}}(\pi) &= J_{\mathrm{av}}(\widetilde{\pi}) - J_{\mathrm{av}}(\pi) \\ &\leq L_J \|\pi - \widetilde{\pi}\|_1 \leq \frac{2L_J S}{(1-\gamma)\zeta_\gamma}\Delta_\gamma(\pi).\end{aligned}$$

This proves the lemma. □

C.3 Proof of global linear convergence

Theorem 2 (restated). *Suppose Assumptions 1 and 2 hold. Run Algorithm 2 from any full-support policy π_0 with any $\eta_0 > 0$. If a positive Bellman loss exists at $\bar{\gamma}$, define*

$$C_{\bar{\gamma}} := \frac{2L_J S}{(1-\bar{\gamma})\zeta_{\bar{\gamma}}}\left(1 + \frac{(1-\bar{\gamma})\log(1/p_{\min})}{\eta_0 \bar{\gamma}}\right).$$

Then, for every $k \geq 0$,

$$\Delta_{\mathrm{av}}(\pi_k) \leq C_{\bar{\gamma}}\left(1 - \frac{1-\bar{\gamma}}{S}\right)^k \leq C_{\bar{\gamma}} \exp\left(-\frac{1-\bar{\gamma}}{S}k\right).$$

If all Bellman losses vanish, $\Delta_{\mathrm{av}}(\pi_k) = 0$ for every $k \geq 0$.

Proof. We first apply Xiao's discounted convergence theorem, then use Lemma 3 to obtain the average-reward bound.

Fix $\gamma \in (0,1)$ and a discounted-optimal policy $\pi^\star_\gamma$. For probability vectors $\mathbf{p}, \mathbf{q}$ with $q_a > 0$, write $D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{q}) := \sum_{a:p_a>0} p_a \log(p_a/q_a)$. Define

$$\vartheta_\mu := \frac{1}{1-\gamma}\left\|\frac{\mathbf{d}^{\pi^\star_\gamma}_{\gamma,\mu}}{\boldsymbol{\mu}}\right\|_\infty, \tag{41}$$

$$D^\star_0 := \sum_s d^{\pi^\star_\gamma}_{\gamma,\mu}(s) D_{\mathrm{KL}}\left(\pi^\star_\gamma(\cdot \mid s) \,\middle\|\, \pi_0(\cdot \mid s)\right), \tag{42}$$

where vector division is componentwise. Since $\pi_0(a \mid s) \geq p_{\min}$ and $\sum_{a:p_a>0} p_a \log p_a \leq 0$, each divergence is at most $\log(1/p_{\min})$. Hence $D^\star_0 \leq \log(1/p_{\min})$.

The discounted occupancy is a probability distribution, and $\mu(s) = 1/S$. Therefore

$$\frac{1}{1-\gamma} \leq \vartheta_\mu \leq \frac{S}{1-\gamma}, \qquad \frac{\eta_{k+1}}{\eta_k} = \frac{1}{\gamma} \geq \frac{\vartheta_\mu}{\vartheta_\mu - 1}.$$

The mirror update preserves full support, so the initialization and step-size conditions of Theorem 10 of Xiao (2022) hold.

Apply that theorem with the entropic mirror map, evaluation distribution $\boldsymbol{\mu}$, and costs $1-r$. This cost transformation produces the same policy updates, because the resulting action values differ from the negated reward action values by a constant. Multiplying the cost-gap bound by $1-\gamma$ gives

$$\begin{aligned}\Delta_\gamma(\pi_k) &\leq \left[\Delta_\gamma(\pi_0) + \frac{(1-\gamma)D^\star_0}{\eta_0\gamma}\right]\left(1 - \frac{1}{\vartheta_\mu}\right)^k \\ &\leq \left[1 + \frac{(1-\gamma)\log(1/p_{\min})}{\eta_0\gamma}\right]\left(1 - \frac{1-\gamma}{S}\right)^k,\end{aligned} \tag{43}$$

where the second inequality uses $\Delta_\gamma(\pi_0) \leq 1$ and the bounds above. The first inequality retains the sharper initialization-dependent guarantee.

Set $\gamma = \bar{\gamma}$. When a positive Bellman loss exists, Lemma 3 gives

$$\Delta_{\mathrm{av}}(\pi_k) \leq K_{\bar{\gamma}}\Delta_{\bar{\gamma}}(\pi_k) \leq C_{\bar{\gamma}}\left(1 - \frac{1-\bar{\gamma}}{S}\right)^k.$$

The exponential bound follows from $1 - x \leq e^{-x}$. When all Bellman losses vanish, Lemma 3 gives $\Delta_{\mathrm{av}}(\pi_k) = 0$ for every k. This proves the theorem. □

The sufficient update count (30) follows by solving $C_{\bar{\gamma}} \exp(-(1-\bar{\gamma})k/S) \leq \epsilon$ for an integer $k \geq 0$.

D EXPERIMENTAL DETAILS

This appendix specifies the two 30-state, 20-action experiments in Section 5. The MDPs differ in route length, number of symmetric tail states, and reward transformation. Both runs use known-model evaluations without trajectory sampling.

D.1 MDP CONSTRUCTION

States and actions. Both MDPs have $S = 30$ states and $A = 20$ actions $a \in \{0, \dots, 19\}$. Write $q_a = a/19$ and $p_a = 0.03 + 0.91q_a$. The states are a hub H, a bad-loop state B, a reward/reset state R, optional gates $G_1, \dots, G_{L-1}$, and M symmetric tail states $T_1, \dots, T_M$. Experiment 1 uses $(L, M) = (8, 20)$; Experiment 2 uses $(L, M) = (1, 27)$ and has no gates. For the state ordering in the code, set $H = 0$, $B = 1$, $G_i = i + 1$, $R = L + 1$, and $T_j = L + 1 + j$. Define $X = G_1$ for $L > 1$, $X = R$ for $L = 1$, and $G_L = R$. The route to the higher-reward region leaves H, crosses $L - 1$ gates, and reaches R; the competing route enters the H–B loop.

Transition kernel. It is convenient to specify each transition row as a distribution rather than list its nonzero entries individually. Let δ_s be the point mass at state s, and define

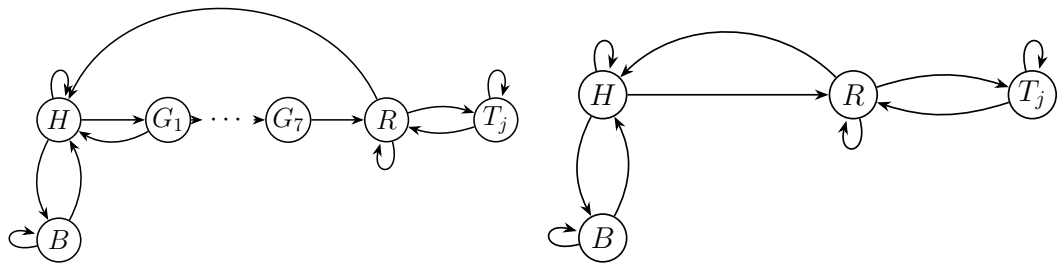
$$U_T = \frac{1}{M}\sum_{j=1}^{M} \delta_{T_j}, \qquad c_a = (0.975 - 0.18q_a)q_a, \qquad b_a = 0.018 + 0.04(1 - q_a).$$

Then the complete transition kernel is

$$\begin{aligned}
P(\cdot \mid H, a) &= 0.02\delta_H + 0.98(1 - q_a)\delta_B + 0.98q_a\delta_X, \\
P(\cdot \mid B, a) &= 0.18\delta_H + 0.82\delta_B, \\
P(\cdot \mid G_i, a) &= (1 - p_a)\delta_H + p_a\delta_{G_{i+1}} \qquad (1 \le i < L), \\
P(\cdot \mid R, a) &= (1 - 0.18q_a - c_a)\delta_H + c_a\delta_R + 0.18q_aU_T, \\
P(\cdot \mid T_j, a) &= b_a\delta_R + (1 - b_a)\delta_{T_j} \qquad (1 \le j \le M).
\end{aligned}$$

For $L = 1$, the gate row is omitted. The term U_T means that the probability $0.18q_a$ of leaving R for a tail is divided equally among all M tail states. Every non-hub state has a positive-probability return path to H under every action. Hence every stationary policy has one recurrent class, as assumed in the theory.

Figure 3 shows the transition structure; the probabilities are specified above and the original rewards below.



(a) Experiment 1: $(L, M) = (8, 20)$ **(b) Experiment 2:** $(L, M) = (1, 27)$

Figure 3: Transition structure of the two MDPs; arrows indicate possible transitions. In (a), the route to the higher-reward region passes through seven gates $G_1, \dots, G_7$ before reaching R; each gate advances with probability p_a and otherwise returns to H (the return arrow from G_1 is representative). In (b), H connects directly to R without gates. In each panel, T_j represents one of the M symmetric tail states.

Original rewards. The rewards displayed in the figures and tables are based on

$$\widetilde{r}(H,a) = 0.48 + 0.004(1-q_a), \qquad \widetilde{r}(B,a) = 0.62,$$
$$\widetilde{r}(G_i,a) = 0.48 + 0.002q_a, \qquad \widetilde{r}(R,a) = 0.57 + 0.003q_a,$$
$$\widetilde{r}(T_j,a) = 0.535 + 0.11q_a.$$

The gate-reward row is absent from Experiment 2. All displayed decimal constants are interpreted as exact rational numbers when the MDP is constructed.

D.2 Reward transformation and instance properties

The algorithms optimize a shaped and normalized reward, which changes discounted values but preserves average-reward policy rankings:

$$r_i(s,a) = \frac{\widetilde{r}(s,a) + \sum_{s'} P(s' \mid s,a)\phi_i(s') - \phi_i(s) - \ell_i}{w_i}. \tag{44}$$

Here $i \in \{1,2\}$ indexes the experiment. The complete parameters are

$$\begin{aligned}
(\ell_1, w_1) &= (-0.43, 1.608),\\
\phi_1(H) = 0, \quad \phi_1(B) &= -\tfrac{1}{500},\\
(\phi_1(G_1),\dots,\phi_1(G_7)) &= (\tfrac{1}{400}, \tfrac{1}{200}, \tfrac{1}{100}, \tfrac{1}{50}, \tfrac{1}{20}, \tfrac{1}{10}, \tfrac{2}{5}),\\
\phi_1(R) = 1, \quad \phi_1(T_j) &= \tfrac{9}{2};\\
(\ell_2, w_2) &= (-\tfrac{1279}{500}, \tfrac{24813}{2000}),\\
\phi_2(H) = 0, \quad \phi_2(B) = -\tfrac{1}{10}, \quad \phi_2(R) &= -\tfrac{31}{10}, \quad \phi_2(T_j) = 48.
\end{aligned}$$

The potential term has zero stationary mean, and $w_i > 0$. Consequently, if $J_{\mathrm{av},i}(\pi)$ is the average reward under r_i, the plotted original-scale reward is $\widetilde{J}_{\mathrm{av},i}(\pi) = \ell_i + w_i J_{\mathrm{av},i}(\pi)$; original-scale optimality gaps are w_i times the normalized gaps. The normalized rewards attain both 0 and 1. Their exact rational encodings have common denominators $290{,}244{,}000$ and $8{,}957{,}493{,}000$, respectively, with $R_{\mathrm{num}} = m$ in both cases.

The deterministic all-a_0 policy, denoted $\pi^{(0)}$, is a non-strict suboptimal local maximum. At this policy only H and B are recurrent, so direct average-reward gradients vanish on the route and tail states. Write $g_0 := \widetilde{J}_{\mathrm{av},i}(\pi^{(0)}) \approx 0.5988965517$ for its original-scale reward, which is the same in both MDPs.

To verify local maximality, use the original rewards and normalize the bias by $\widetilde{h}^{\pi^{(0)}}(H) = 0$. The Poisson equations give $\widetilde{h}^{\pi^{(0)}}(B) > 0$ and $\widetilde{h}^{\pi^{(0)}}(X) < 0$: the reward at B exceeds g_0, whereas the rewards along the route under a_0 are below g_0. Thus every $a \neq a_0$ has strictly negative advantage at H,

$$\widetilde{A}^{\pi^{(0)}}(H,a) = q_a\big[-0.004 + 0.98\big(\widetilde{h}^{\pi^{(0)}}(X) - \widetilde{h}^{\pi^{(0)}}(B)\big)\big] < 0.$$

For a nearby policy π, let σ agree with π except that it selects a_0 at H. Its recurrent class is still $\{H,B\}$, and its average reward is g_0, since all actions at B are identical. By continuity of the bias, the advantages $\widetilde{A}^{\sigma}(H,a)$ remain negative for $a \neq a_0$. The performance-difference identity then gives

$$\widetilde{J}_{\mathrm{av},i}(\pi) - g_0 = d^{\pi}(H) \sum_a \pi(a \mid H)\widetilde{A}^{\sigma}(H,a) \leq 0.$$

The maximum is non-strict because changing only the transient-state actions leaves the average reward unchanged. The reward transformation preserves this conclusion. To verify strict suboptimality, consider the policy selecting a_{19} outside B. Exact rational policy evaluation gives original-scale rewards exceeding 0.634 and 0.638 in Experiments 1 and 2, respectively, whereas $g_0 = 2171/3625 < 0.599$.

Failure of global gradient domination. Since $\pi^{(0)}$ is a suboptimal local maximum of the smooth objective on the convex policy simplex, its feasible directional derivatives are nonpositive. Thus

$$\mathcal{F}(\pi^{(0)}) = 0, \qquad \Delta_{\mathrm{av}}(\pi^{(0)}) > 0,$$

so no finite C can satisfy $\Delta_{\mathrm{av}}(\pi) \le C\mathcal{F}(\pi)$ globally. The stationary-distribution coefficient used by Kumar et al. (2025) is also infinite here. Indeed, $d^{\pi^{(0)}}(R) = 0$, whereas every average-optimal policy has $d^{\pi^\star}(R) > 0$: a policy with zero stationary mass at R must have recurrent class $\{H, B\}$ and reward g_0. Consequently, the ratio $d^{\pi^\star}(R)/d^{\pi^{(0)}}(R)$ is infinite. These are structural properties of the MDPs, independent of numerical trajectories. Table 1 reports the reference optima and numerical Blackwell-threshold estimates.

Table 1: Instance properties. Rewards and gaps use $\widetilde{r}$; Blackwell thresholds are numerical estimates for r_i.

Instance	(L, M)	Estimated γ_{BW}	Optimal reward	Gap at all-a_0
Experiment 1	$(8, 20)$	0.8093197829	0.6347643488	0.0358677971
Experiment 2	$(1, 27)$	0.2599755919	0.6380879223	0.0391913705

We obtain the reference average rewards by policy iteration, evaluating each policy through the Poisson equation with $h(H) = 0$. For the threshold estimates, we examine the policy selecting a_{19} outside B (whose actions are identical), locate the last observed negative-to-positive crossing of its (H, a_0) discounted Bellman loss by grid search and bisection, and check all Bellman losses numerically above that crossing. These values are numerical diagnostics, not a proof of the exact thresholds. The accompanying verification script checks the rational encodings, reference rewards, and clipped optima.

D.3 Optimization protocol

Initialization and symmetry. All methods start from $\pi_0(a \mid s) = 1/20$ and perform 40,000 updates without early stopping. Symmetry preserves identical policy rows on the tail states, represented by one state T for evaluation. The reduced systems have 11 and 4 states; both full MDPs and their discount certificates still use $S = 30$. The uniform initial distribution on the full model assigns mass $1/30$ to each non-tail state and $M/30$ to T. For Euclidean gradient updates, the tail-row gradient in the reduced model is divided by M, and the squared displacement in the line search counts that row M times. Mirror updates use action values directly and require no such division.

Table 2: Hyperparameters. BPMA uses the prescribed discount (21); the numerical threshold estimates in Table 1 are not used by either algorithm.

Instance	Base step η_0	Clipping floor α
Experiment 1	0.02	0.005
Experiment 2	0.1	0.01

Updates. Within each instance, suppress the index i on the normalized objective and write $J_{\mathrm{av}} = J_{\mathrm{av},i}$. Define the differential action value

$$Q^\pi_{\mathrm{av}}(s, a) = r_i(s, a) - J_{\mathrm{av}}(\pi) + \mathbf{P}_{s,a}\mathbf{h}^\pi,$$

where $\mathbf{h}^\pi$ solves (32) for r_i. The array $d^\pi(s)Q^\pi_{\mathrm{av}}(s, a)$ represents the direct-policy gradient up to rowwise constants, which do not affect Euclidean projected updates. We compare four methods:

- **BPMA:** Apply (23) at $\gamma = \bar{\gamma}$ from (21), with $\eta_k = \eta_0\bar{\gamma}^{-k}$. Store log weights, add the discounted action scores times η_k, and normalize each row. The centered evaluation below avoids the large common component of the discounted values.
- **DGPG:** Use (11) and (7) with $\eta = \eta_0$. Test a discounted proposal only after its phase is complete, and record the accepted policy at that update index. The experimental AveragePG starts backtracking at η_0 and halves the trial step until (9) holds with 10^{-14} subtracted from its right-hand side. Under exact acceptance, Appendix B.1 covers the changed initial step; the tolerance is a numerical relaxation. Line-search trials and candidate comparisons do not increase the update count.
- **Average PG:** Apply $\mathrm{Proj}_\Pi(\pi + \eta_0\nabla J_{\mathrm{av}}(\pi))$.

- α**-clipped PG:** Following Lee & Ryu (2026), project each row $\pi(\cdot \mid s) + \eta_0 Q^\pi_{\mathrm{av}}(s, \cdot)$ onto $\{\mathbf{p} \in \Delta(\mathcal{A}) : p_a \geq \alpha\}$. This update omits the stationary factor in the direct gradient.

DGPG and the direct baselines use float64. Their stationary-distribution solver sets entries of magnitude below 10^{-13} to zero, clips negative entries to zero, and renormalizes before computing average rewards and gradients.

Finite-discount BPMA evaluation. The original rational encodings give

$$1 - \bar{\gamma}_1 \approx 7.500394563217 \times 10^{-31685}, \qquad 1 - \bar{\gamma}_2 \approx 2.759884257260 \times 10^{-37045}. \tag{45}$$

A direct float64 representation rounds both discounts to one. Instead, store the positive number $\epsilon_\gamma := 1 - \bar{\gamma}$ separately using arbitrary-precision arithmetic. For each policy, define

$$u_\pi(s) := V^\pi_{\bar{\gamma}}(s) - V^\pi_{\bar{\gamma}}(H), \qquad c_\pi := \epsilon_\gamma V^\pi_{\bar{\gamma}}(H).$$

Let $\mathbf{u}_\pi$ be the column vector with components $u_\pi(s)$. The discounted Bellman equations are equivalently

$$c_\pi \mathbf{1} + (\mathbf{I} - \mathbf{P}_\pi)\mathbf{u}_\pi + \epsilon_\gamma \mathbf{P}_\pi \mathbf{u}_\pi = \mathbf{r}_\pi, \qquad u_\pi(H) = 0. \tag{46}$$

Solving this centered system avoids the common value term of order $1/\epsilon_\gamma$. The scores used in the mirror update are

$$q_\pi(s, a) := r_i(s, a) + \mathbf{P}_{s,a}\mathbf{u}_\pi - \epsilon_\gamma \mathbf{P}_{s,a}\mathbf{u}_\pi. \tag{47}$$

Since $Q^\pi_{\bar{\gamma}}(s, a) = q_\pi(s, a) + \bar{\gamma} V^\pi_{\bar{\gamma}}(H)$, the common term cancels in (23). Thus (46)–(47) evaluate the finite-discount update without substituting average-reward bias scores. The prescribed step is evaluated through $\log(\eta_k/\eta_0) = -k \log(1 - \epsilon_\gamma)$, using routines that preserve small increments.

Complete BPMA runs use 256-bit significands and are repeated with 384 bits, approximately 77 and 116 decimal digits. The maximum reward discrepancy over all updates and both instances is below 1.1×10^{-75}. At updates $0, 1, 100, 1000, 10000, 40000$, we also form $\bar{\gamma} < 1$ explicitly and independently solve the uncentered discounted Bellman equations at the recorded policy, renormalized at reference precision. These checks use 105,638 bits for Experiment 1 and 123,443 bits for Experiment 2, enough to distinguish $\bar{\gamma}$ from one with more than 100 additional decimal digits. The largest discrepancy in action-value differences is below 4.4×10^{-76}, and the largest absolute Bellman residual is below 3.2×10^{-116}.

The finite-discount corrections to order-one scores are below the working precision. Moreover, the relative step increase over $40{,}000$ updates is only about 3.00×10^{-31680} and 1.10×10^{-37040}. Consequently, the plotted curves cannot resolve the difference from the $\gamma \uparrow 1$ limiting trajectory, and the prescribed step growth provides no visible acceleration over this horizon. The precision comparisons and residual checks assess numerical accuracy; they are not rigorous error enclosures for the full trajectory.

D.4 DETAILED NUMERICAL RESULTS

For the clipping constraint $\pi(a \mid s) \geq \alpha$, write $\pi(a \mid s) = \alpha + (1 - A\alpha)\nu(a \mid s)$ with $\nu \in \Pi$. Policy iteration on the induced model

$$P_\alpha(\cdot \mid s, a) = \alpha \sum_b P(\cdot \mid s, b) + (1 - A\alpha)P(\cdot \mid s, a),$$
$$\widetilde{r}_\alpha(s, a) = \alpha \sum_b \widetilde{r}(s, b) + (1 - A\alpha)\widetilde{r}(s, a)$$

gives constrained optimal rewards 0.6187293766 and 0.6243535101. The clipped trajectories reach these values to displayed precision; their remaining unrestricted gaps reflect the clipping constraint.

Table 3 reports results after $40{,}000$ updates. Here k_ε is the first update with original-scale optimality gap at most ε; a dash means that threshold was not reached. A displayed zero denotes zero to the reported numerical precision. The certificate chooses a discount close to one in both instances, so this comparison does not isolate the effect of the Blackwell threshold.

Table 3: Original-scale rewards and gaps after $40{,}000$ updates. BPMA uses the prescribed finite discount in (45).

Method	Final reward	Final gap	$k_{10^{-2}}$	$k_{10^{-3}}$	$k_{10^{-4}}$
Experiment 1					
BPMA	0.6345663	1.9805×10^{-4}	$5{,}488$	$22{,}717$	–
DGPG	0.6337798	9.8450×10^{-4}	$7{,}168$	$38{,}702$	–
Average PG	0.5988966	3.5868×10^{-2}	–	–	–
α-clipped PG	0.6187294	1.6035×10^{-2}	–	–	–
Experiment 2					
BPMA	0.6372471	8.4079×10^{-4}	$9{,}376$	$37{,}206$	–
DGPG	0.6380879	0	$3{,}072$	$36{,}864$	$36{,}864$
Average PG	0.5985334	3.9555×10^{-2}	–	–	–
α-clipped PG	0.6243535	1.3734×10^{-2}	–	–	–