SHARED ACTORS NEED NOT SHARE CRITICS: EFFECTS OF VALUE MISMATCH IN PARALLEL REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

When an actor-critic policy is trained in parallel across multiple environments in order to achieve strong generalization capability, implementations commonly use one critic throughout training by default. Yet different environments can assign different expected returns to the same state input to the critic. A critic without environment information must then reconcile distinct value targets, systematically shifting the sampled advantages within individual environments. Using illustrative bandit models with multiple environments and a common optimal arm, we characterize how this value mismatch redistributes sampled policy updates, reinforcing unhelpful actions while attenuating or even reversing useful ones. The oracle learning processes using no baseline, the shared value, or the value specific to the sampled environment have the same mean logit update at a fixed policy and converge to the same optimal policy, yet their realized learning paths can differ sharply under stochasticity. The analysis motivates a minimal intervention: give only a logged environment index to the critic so that it can separate the value targets. Controlled CartPole and MuJoCo experiments expose the predicted shifted values, advantages, and performance gaps. In the more complex BipedalWalker and Procgen settings, the same intervention yields more stable learning and higher returns. Across all 16 Procgen games, the multihead conditional critic improves aggregate normalized return on 600 unseen levels per game by 40.8%. In conclusion, we present the theory identifying value mismatch as a direct mechanism through which critic sharing can degrade stochastic learning dynamics, not captured by scalar estimator variance alone, and conduct the experiments showing that conditioning on an index is broadly effective in parallel actor-critic RL.

1 INTRODUCTION

Policy gradient methods optimize expected return, $J(\theta) = \mathbb{E}_{\tau\sim\pi_\theta}[G(\tau)]$, by updating a stochastic policy π_θ from sampled trajectories (Williams, 1992; Sutton & Barto, 2018). A sampled update is proportional to $(G_t - B_t)\nabla_\theta \log \pi_\theta(a_t \mid s_t)$, where G_t is the sampled return and B_t is a baseline independent of the sampled action. For actor-cirtic RL (Konda & Tsitsiklis, 1999), the actor chooses actions; a learned critic supplies the reference baseline against which their returns are judged. Subtracting a baseline leaves the expected gradient unchanged at a fixed policy, and its role is classically understood through variance reduction (Weaver & Tao, 2001; Greensmith et al., 2004). But training is a closed loop: each sample changes the policy that generates the next sample. Preserving the expected gradient does not preserve the law of the online learning process. Under stochastic policy gradient, baselines can change the signs and strengths of these updates, producing different learning paths despite the same local expected direction (Chung et al., 2021; Mei et al., 2021; 2022).

One common goal of RL is to train a single generalizable policy across environments or levels, and existing solutions include procedural levels (Cobbe et al., 2020), dynamics randomization (Peng et al., 2018), and curricula such as PLR, ACCEL, and PATH (Jiang et al., 2021; Parker-Holder et al., 2022; Liu & Chen, 2026). These methods all apply actor–critic algorithms such as PPO (Schulman et al., 2017) to sample rollouts from batches of environments **in parallel** and perform stochastic policy gradient updates. Common implementations also share a single value function across sampled environments to calculate advantages (Cobbe et al., 2020; Jiang et al., 2021; Raileanu & Fergus,

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.

2021). The key problem is that given one same state, the value is always different among environments. For example, two walkers can have the same local observation but different terrain ahead. Such differences lead to different future rewards and, consequently, different value predictions for the same observation, and the shared critic will average over it.

Given one state s and one policy π, the value in environment z is $V_z^\pi(s) := \mathbb{E}_\pi[G \mid S = s, Z = z]$. We refer to $V_z^\pi(s)$ as **conditional** value baseline over z. Without environment information, the **shared** value baseline is the marginal

$$\bar{V}^\pi(s) := \mathbb{E}_\pi[G \mid S = s] = \sum_z q_z^\pi(s) V_z^\pi(s), \qquad q_z^\pi(s) := \mathbb{P}_\pi(Z = z \mid S = s). \tag{1}$$

Although the default choice of the shared value baseline leaves the expected policy gradient unchanged compared with the conditional one, we will show how the shared critic systematically miscenter advantages of states within each environment under stochastic updates. We call $e_z^\pi(s) := V_z^\pi(s) - \bar{V}^\pi(s)$ the *value mismatch*. Positive mismatch can reinforce inferior actions; negative mismatch can attenuate or reverse useful updates. Figure 1 previews the resulting learning dynamics in the bandit setting. The three processes use the same initial actor and environment distribution; they differ only in how the sampled return is centered.

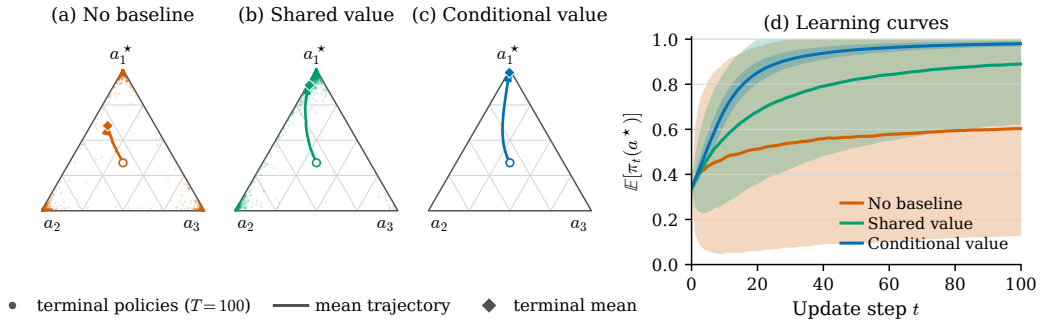

Figure 1: **Same destination, different sampled paths.** One softmax actor learns two three-arm bandits with a common optimal arm, following each update by Algorithm 1. The panels compare no baseline, the shared value, and the sampled environment's value. (a–c) Policies after 100 updates and mean trajectories. (d) Mean optimal-arm probability ± 1 s.d. across 4,000 paired runs. Rewards are $(1, 0.9, 0.8)$ and $(0.4, 0.2, 0)$, $q_E = q_H = 1/2$, $\eta = 2$, and $\pi_0 = (0.34, 0.33, 0.33)$. All three processes converge to the optimal corner (Proposition 1), despite their finite time separation.

Providing the critic with additional information has architectural precedents (Pinto et al., 2018; Yue et al., 2024), which empirically demonstrate the benefits of giving the critic access to privileged information. Our contribution is to explain why *simply* separating values changes the learning process and benefits the learning. First, we identify how a shared value reinforces suboptimal samples and attenuates or reverses useful ones. In illustrative bandits, all three baselines have the same expected update at a fixed policy and converge to the same optimum, yet conditional centering eventually turns every sample into progress while sharing produces reversed updates infinitely often under the stated mismatch condition. Second, this account motivates a deliberately minimal intervention: give only an arbitrary logged environment index to the critic. In our experiments, controlled CartPole and MuJoCo experiments exhibit value and advantage signatures consistent with the mechanism, and the conditioned critic consistently outperforms the shared one. In the more complex BipedalWalker and Procgen settings, conditioning also yields substantial empirical gains: the two designs of conditioned critics raise the final mean return on unseen BipedalWalker terrains from 90.7 (shared critic) to 155.6 and 190.4, and improve the aggregate normalized Procgen return on unseen levels over the shared critic by 21.2% and 40.8%, respectively. These results show that correcting value mismatch can be useful across several forms of parallel environment training.

2 Problem Setting: One Policy, Many Environments

An explanatory bandit model. We isolate the effect of critic sharing in the smallest model (bandit as one single-state Markov Decision Process) that retains a value that depends on the environment.

Let $\mathcal{Z}$ be a finite collection of environments and $\mathcal{A}$ a finite set of $K \geq 2$ arms. At round t, an environment $Z_t \sim q$ is drawn independently, where $q_z > 0$ and $\sum_z q_z = 1$. A single softmax actor $\pi_t(a) \propto \exp \theta_t(a)$ samples $a_t \sim \pi_t$ without observing Z_t. Environment z assigns a deterministic reward $r_z(a)$ to arm a; this finite reward table is bounded. We assume one common strict optimal arm: $r_z(a^*) > r_z(i)$ for every z and $i \neq a^*$. Thus a sampled update that decreases $\pi(a^*)$ cannot be explained by conflicting objectives across environments.

For a policy π, define $V_z^\pi := \mathbb{E}_{a\sim\pi}[r_z(a)] = \sum_a \pi(a) r_z(a)$ and $\bar{V}^\pi := \mathbb{E}_{Z\sim q}[V_Z^\pi] = \sum_z q_z V_z^\pi$. We compare three oracle baselines: $B_t^0 = 0$, $B_t^{\text{shared}} = \bar{V}^{\pi_t}$, and $B_t^{\text{cond}} = V_{Z_t}^{\pi_t}$.

Algorithm 1 Softmax Stochastic Policy Gradient across Environments

Require: distribution q, rewards $\{r_z\}$, step size η, baseline $B \in \{B^0, B^{\text{shared}}, B^{\text{cond}}\}$

1: **for** $t = 0, 1, 2, \ldots$ **do**
2: draw $Z_t \sim q$ and $a_t \sim \pi_{\theta_t}$
3: observe $G_t = r_{Z_t}(a_t)$ and evaluate the baseline B_t
4: $\theta_{t+1} \leftarrow \theta_t + \eta\,(G_t - B_t)(\mathbf{e}_{a_t} - \pi_{\theta_t})$

Algorithm 1 describes one stochastic update by sampling one arm at each step, using the full softmax score $\nabla_\theta \log \pi_\theta(a_t) = \mathbf{e}_{a_t} - \pi_\theta$. The scalar $G_t - B_t$ decides whether the sampled action is encouraged or discouraged; the score vector specifies how its probability changes. We establish a common destination before comparing the realized paths toward it.

Proposition 1 (Asymptotic analysis)**.** *Consider Algorithm 1 with finite initial logits, the common strict optimal arm defined above, and any fixed learning rate $\eta \in (0, \infty)$. For each $B \in \{B^0, B^{\text{shared}}, B^{\text{cond}}\}$,*

$$\mathbb{P}\Big(\lim_{t\to\infty} \pi_t^B(a^*) = 1\Big) = 1. \tag{2}$$

Moreover, given $\bar{r}(a) := \sum_z q_z r_z(a)$, for each such B, almost surely there exist finite random constants C_B and T_B such that, for every integer $T \geq \max\{T_B, 2\}$,

$$\frac{1}{T}\sum_{t=0}^{T-1}\left[\bar{r}(a^*) - \sum_a \pi_t^B(a)\bar{r}(a)\right] \leq C_B \frac{\log T}{T}. \tag{3}$$

Proposition 1 serves two roles. First, from any initialization with finite logits and for any fixed finite $\eta > 0$, all three processes reach the same optimal policy, and each has an $O(\log T/T)$ upper bound on time-averaged suboptimality. Their separation under finite training budgets must therefore come from how they travel, not from their final destination. Second, convergence and the infinite exploration established in its proof ensure that the regimes in Propositions 2 and 3 are reached almost surely. We adapt the proof strategy of Robertson et al. (2025) and provide the complete proof in Appendix D.

3 Baselines change the online update dynamics

At a fixed policy, every baseline independent of the sampled action yields the same expected update. Yet Algorithm 1 applies one sampled update, whose magnitude and even sign can depend on the baseline. This changes the policy and hence subsequent sampling, producing different stochastic paths despite a common mean update at matched policies. Variance summarizes the aggregate dispersion of these updates. For unbiased estimators with the same mean, reducing variance tightens the standard smooth SGD bound on expected improvement after one step (Bottou et al., 2018). It does not record which branches are reinforced: even the exact minimum variance baseline can approach the optimum more slowly and less steadily over a finite horizon than conditional value centering (Appendix C). We therefore track how individual samples change subsequent sampling.

3.1 One environment: an eventual ratchet

A value baseline is a moving bar: only rewards above the current value reinforce the sampled arm. With two arms, either sample moves the policy toward the better arm; without a baseline, a positive

reward reinforces a sampled suboptimal arm. The same distinction eventually holds with any finite number of arms. In one environment, shared and conditional values coincide; we write π^V and π^0 for the policy with value baseline and without baseline.

Proposition 2 (Eventual ratchet versus persistent drawdowns)**.** *Consider a deterministic finite bandit with finite initial logits, a unique optimal arm a^*, and any fixed $\eta > 0$. Under the oracle value baseline,*

$$\mathbb{P}\big(\exists T < \infty : \ \pi^V_{t+1}(a^*) > \pi^V_t(a^*) \text{ for every } t \geq T\big) = 1. \tag{4}$$

If at least one suboptimal arm has positive reward, the process without a baseline satisfies

$$\mathbb{P}\big(\pi^0_{t+1}(a^*) < \pi^0_t(a^*) \text{ infinitely often}\big) = 1. \tag{5}$$

Both processes converge to the best arm, but only value centering eventually turns every sample into progress. The distinction is between a path that eventually stops stepping backward and one that continues to do so, not between convergence and failure to converge. Appendix E gives the thresholds and proof.

3.2 Multiple environments: the shared offset

For the state $S = s$, a shared value marginalizes over hidden environment identity: $\bar{V}^\pi(s) = \sum_z q^\pi_z(s) V^\pi_z(s)$, where $q^\pi_z(s) = \mathbb{P}_\pi(Z = z \mid S = s)$. Define the value mismatch $e^\pi_z(s) := V^\pi_z(s) - \bar{V}^\pi(s)$. For the same sampled return G,

$$A^{\text{cond}} := G - V^\pi_z(s), \qquad A^{\text{shared}} := G - \bar{V}^\pi(s) = A^{\text{cond}} + e^\pi_z(s).$$

Thus every sampled of state s from environment z is shifted by the same offset $e^\pi_z(s)$. If its value is above the shared average, the baseline is too low and inferior actions can receive positive advantages. If its value is below the average, the baseline is too high and even optimal actions can receive negative advantages. Conditioning centers each return against its own environment's value instead.

Proposition 3 (Conditional ratchet versus persistent shared drawdowns)**.** *Consider Algorithm 1 with finite initial logits and any fixed $\eta > 0$. Under the oracle conditional baseline,*

$$\mathbb{P}\Big(\exists T < \infty : \ \pi^{B^{\text{cond}}}_{t+1}(a^*) > \pi^{B^{\text{cond}}}_t(a^*) \text{ for every } t \geq T\Big) = 1. \tag{6}$$

If $r_z(a^*)$ is not identical across environments, the shared baseline satisfies

$$\mathbb{P}\Big(\pi^{B^{\text{shared}}}_{t+1}(a^*) < \pi^{B^{\text{shared}}}_t(a^*) \text{ infinitely often}\Big) = 1. \tag{7}$$

Conditional centering eventually turns every sample into progress. Sharing reaches the same optimum but keeps discouraging optimal samples from environments below the average. These drawdowns recur infinitely often while their magnitudes vanish: recurrence does not contradict convergence. We provide exact proof of Proposition 3 in Appendix F.

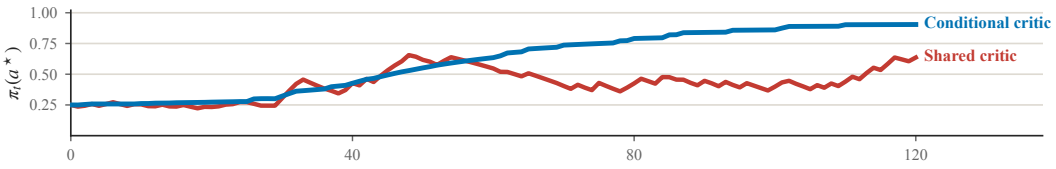

Figure 2: A sample trajectory of Algorithm 1 under the shared and the conditional value baseline ($q_E = q_H = \frac{1}{2}$, $r_E = (1, 0.8, 0.6)$, $r_H = (0.4, 0.2, 0)$, $\eta = 0.5$, $\pi_0 = (0.25, 0.45, 0.3)$). The conditional path increases at every update after an early transient, while the shared path repeatedly steps backward.

Larger mismatch can trigger reversals earlier in policy improvement; at a fixed policy, a larger η amplifies each reversed logit update. These changes affect subsequent training and can compound over a finite training horizon (Figure 8, Appendix F). Additionally, a shared baseline can distort learning even without changing the sign of the advantage. In difficult environments, the shared value is always higher than the conditional value, thus reducing the advantage of the optimal arm. This can slow learning even if suboptimal arms receive larger penalties (Lemma 11).

From oracle baselines to learned critics. In practice the baseline is learned, but more data alone does not remove a conflict between value targets. A critic trained with squared error and observing S but not Z has population target $\mathbb{E}[G \mid S] = \bar{V}^{\pi}(S)$; conditioning instead targets $\mathbb{E}[G \mid S, Z] = V_Z^{\pi}(S)$ (Lemma 10). The shared critic can therefore fit its pooled data well while remaining too optimistic for one environment and too pessimistic for another. Modern PPO commonly uses Generalized Advantage Estimation (GAE) (Schulman et al., 2016; 2017). GAE propagates and mixes the value mismatch offsets along a rollout rather than generally eliminating them (Appendix F). Figure 15 (Appendix H.1) provides one realistic example to illustrate the resulting value, advantage, and policy trajectories on PPO with learned tabular critics.

3.3 Method: Condition Only the Critic

The analysis prescribes a minimal intervention on the critic: give only the critic a logged categorical environment index z, leaving the actor and sampling process unchanged. The index conveys no explicit geometry, ordering, difficulty, or physical parameters; it identifies which value target to fit. We use two standard designs: **FiLM** (Perez et al., 2018) applies an affine scale and shift to shared critic features, while a **multihead** critic uses a shared encoder and one value head per environment (Hessel et al., 2019). FiLM adds learned index embeddings and shared affine generators; multihead adds one linear readout per environment (Appendix G.1). Neither requires additional rollouts or updates. Both retain shared structure while allowing values to differ across environments.

4 Experiments

The experiments are organized around three roles. CartPole (Barto et al., 1983) provides a controlled, end-to-end identification of the value mismatch mechanism, from conflicting value targets to shifted advantages and the resulting learning behavior. MuJoCo tests whether the predicted advantage structure persists with continuous states and actions under variations of hidden dynamics. BipedalWalker (Brockman et al., 2016) and Procgen (Cobbe et al., 2020) then evaluate the practical value and scalability of critic conditioning across large numbers of environments. All use PPO with GAE (Schulman et al., 2016; 2017). Appendix G gives all experiments details.

4.1 CartPole

We design controlled CartPole using two logged levels with the same observation space and reward function but different gravity (Figure 3). In the heterogeneous setting their gravities are $g \in \{10, 50\}$; an identical control assigns $g = 10$ to both level identities. Gravity is absent from the observation. We compare a shared critic, a multihead critic routed by the true index, the same multihead architecture with every sample routed to head 0, and a shared critic with one learned scalar bias per level. All start with identical value predictions.

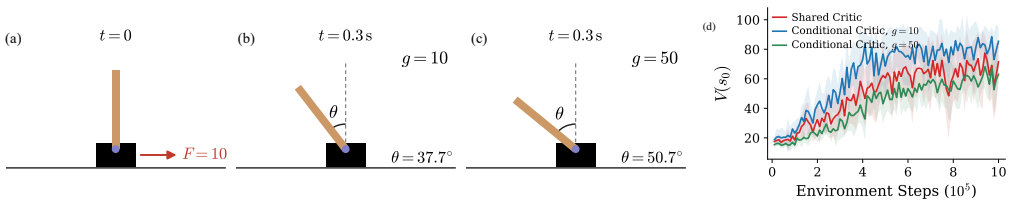

Figure 3: **One observation, two futures.** (a–c) From the same initial state and push ($F = 10$), 0.3 s of dynamics tilts the pole to $37.7°$ under $g = 10$ and $50.7°$ under $g = 50$. (d) Learned values of the initial state: multihead means separate by gravity (blue: $g = 10$; green: $g = 50$), while the shared mean (red) settles between them, consistent with value mismatch. Curves show mean $\pm$ 1 s.d. over 20 seeds.

With identical dynamics, the variants show no systematic separation (Figure 4, top). Heterogeneous gravity exposes the symptom: the shared critic's hard level value loss plateaus and its mean advantage remains negative, while return is lower and more variable across seeds. From middle to late

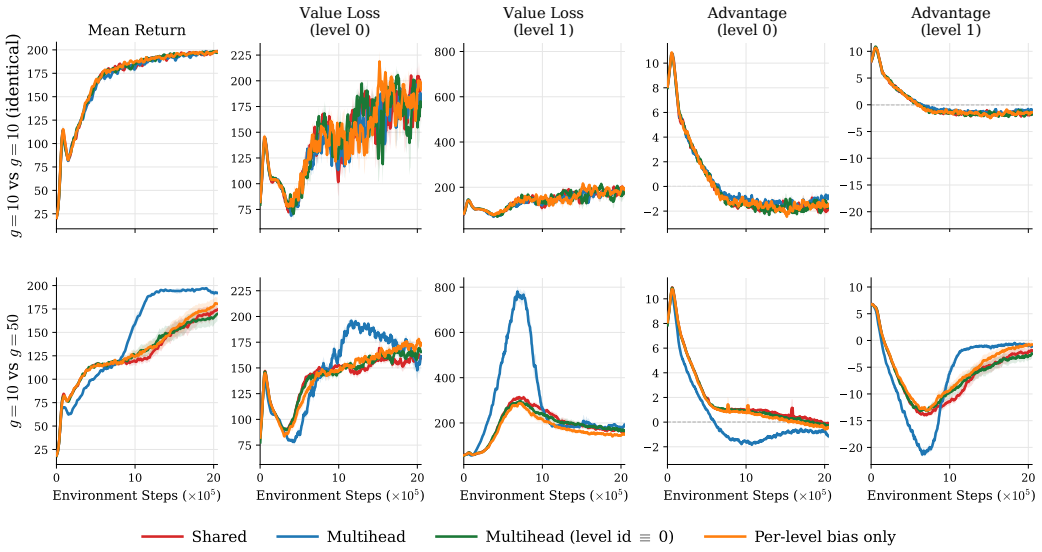

Figure 4: **CartPole mechanism controls.** Mean $\pm$ 1 s.e. over 20 seeds. Top: identical gravities; bottom: $g \in \{10, 50\}$. Columns show mean return, levelwise value losses, and mean GAE advantages. Informative multihead routing separates from the controls under heterogeneous dynamics: after an initial hard level fitting transient, its advantage recenters and return approaches 200. Constant index routing tracks the shared critic; a per level scalar bias provides only partial correction.

training, sharing shifts advantages upward on the easy level and downward on the hard level relative to multihead, matching the predicted direction of value mismatch.

Conditioning does not supply an oracle value. Each head must learn from samples of its own level during training. In this actor critic network, this learning transient is visible: the multihead critic initially incurs a larger hard level fitting loss and a more negative advantage. The contrast emerges once the hard head catches up. Its mean advantage recenters and return approaches 200, whereas the shared critic retains its loss plateau and lower return.

The remaining controls identify what makes this conditional correction useful. Routing every sample to head 0 closely tracks the shared critic: extra heads without informative level routing do not reproduce the gain. One learned scalar bias per level recovers only part of it, indicating that a state dependent conditional baseline matters.

4.2 MuJoCo

For HalfCheetah, Hopper, and Walker2d (Todorov et al., 2012), we vary body mass over ten levels (0.5 to 2.5) without explicitly providing it in the observation, while keeping episodes at a fixed length; only the FiLM critic receives the level index. FiLM has higher mean return on all three tasks (Figure 5, right middle), with the largest gains on HalfCheetah (roughly 2600 vs. 2050) and Walker2d (2450 vs. 2000); Hopper shows a smaller gap in the end, but the conditional critic still dominates the shared one during training. Remarkably, the advantage heatmaps show the predicted structure: shared critic leads to a more polarized distribution of advantages across environments, whereas conditioning largely mitigates this effect (Figure 5, right bottom).

4.3 BipedalWalker

BipedalWalker uses a training set of 100 pinned terrains dominated by hard terrains: training uses 100 pinned terrains (10 standard, 90 Hardcore) for 60M steps; evaluation uses 100 unseen terrains with the same split. Figure 17 (Appendix H.2) gives a constructed example of identical local observations with different terrain beyond the lidar range, which leads to value mismatch during training.

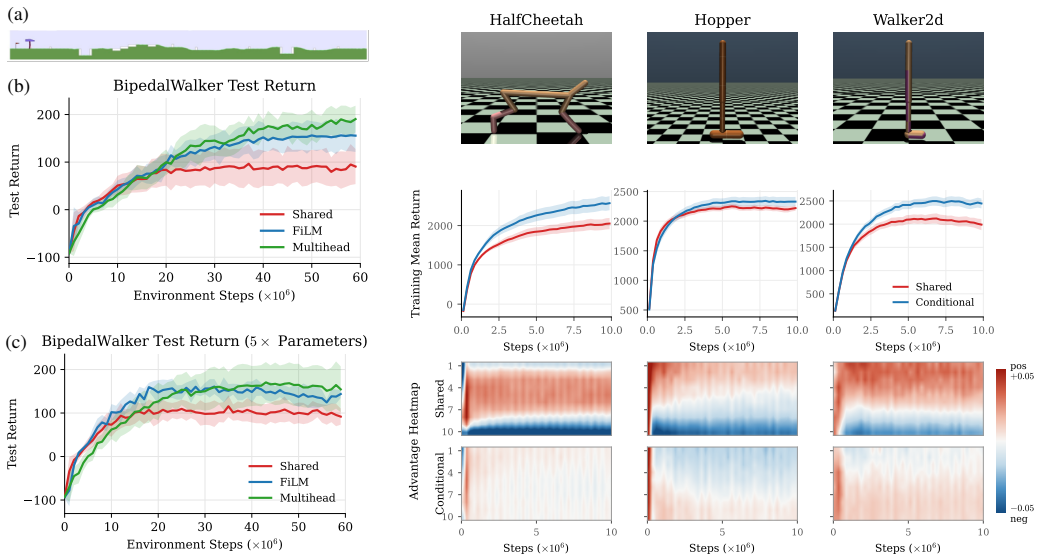

Figure 5: Continuous control benchmarks. **Left:** BipedalWalker with 100 pinned terrains: **(a)** terrain sample, **(b)** test return on 100 unseen terrains, and **(c)** the control with $5\times$ as many parameters (Section 4.5); curves show mean $\pm$ 1 s.d. over 10 seeds. **Right:** MuJoCo with ten body mass levels not explicitly provided in the observation (10M steps). Columns are HalfCheetah, Hopper, and Walker2d; the middle row shows training return, mean $\pm$ 1 s.d. over 30 seeds, and the bottom row shows continuous means across seeds of GAE advantages for each of 10 levels during training (light to heavy; red positive).

The shared critic plateaus below 100, while FiLM reaches roughly 150 and multihead 165–190, with tighter seed bands (Figure 5, left b). Both conditioning architectures escape the same plateau.

4.4 Procgen

Each of 16 Procgen games is trained separately on 200 pinned levels for 25M steps; evaluation uses 600 unseen levels. Value mismatch can arise in pixel games where we confirm that identical current visual inputs can have different conditional values (Appendix H.2). Figure 6 shows evaluation curves on the pinned levels for eight games. Figure 12 (Appendix G.5) gives the full 16 game evaluations. Both conditioned critics improve aggregate performance. Their aggregate normalized unseen-level returns improve by 21.2% and 40.8%, respectively (Table 1). Table 8 (Appendix G.5) summarizes both training and testing evaluations.

Conditional value fitting is not a persistent bottleneck. The early fitting transient in the controlled CartPole study does not persist systematically on the larger benchmarks. We monitor the value loss through the training. On BipedalWalker, the initially higher multihead value loss quickly falls to a comparable line to the shared one in training; on all three MuJoCo tasks, FiLM remains below the shared critic throughout most of training; and on Procgen, both FiLM and multihead finish below the shared critic in majority of games (Figures 13 and 14, Appendix G.6). Thus, the conditional value baseline generally can achieve accurate value estimation, supporting more effective policy learning.

4.5 Ablations

Conditional critic works not because of larger value network capacity. Conditioning changes the parameter count, although the increase is modest in several settings (Appendix Table 2). More decisively, on BipedalWalker we widen each critic architecture to roughly five times its original size while holding its conditioning mechanism fixed. The enlarged shared critic still plateaus, whereas

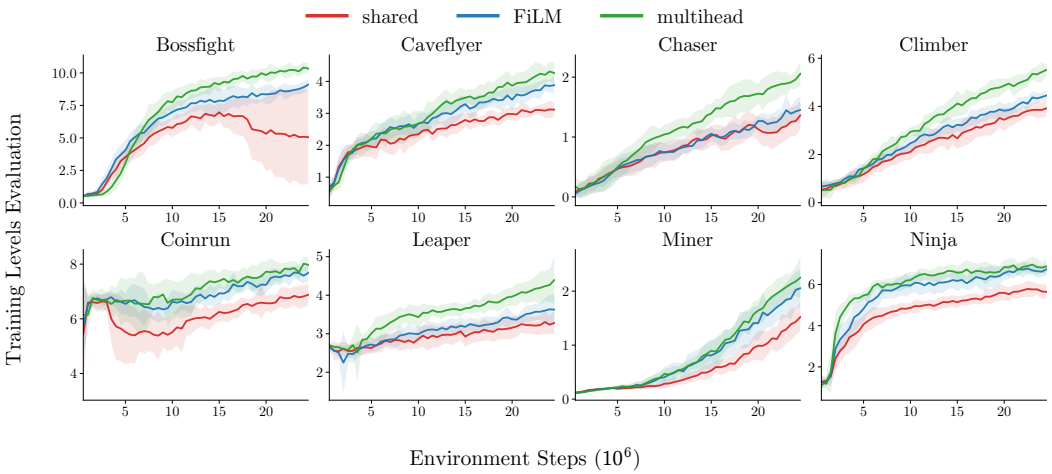

Figure 6: **Procgen evaluation dynamics.** An illustrative subset of eight games: policy evaluation return on the 200 pinned levels over 25M training steps, mean $\pm$ 1 s.d. over 10 seeds. Shared (red), FiLM (blue), and multihead (green) use the same actor architecture and sampling protocol. Unseen-level evaluation is reported separately in Table 1.

FiLM and multihead remain near 150 or above (Figure 5(c)). Parameter count alone therefore does not reproduce the conditioning gain.

Target normalization and other conditioning. PopArt achieves strong multitask performance with a multihead architecture, identifying per-game reward normalization as the key ingredient (van Hasselt et al., 2016; Hessel et al., 2019); in our single task setting, however, with the same reward scale across environments, the benefit comes from conditioning itself. PPO in MuJoCo, BipedalWalker, and Procgen already uses the same global reward scaling and advantage normalization. Adding PopArt further normalizes targets for each level, yet reduces the aggregate Procgen testing gain of multihead from 40.8% to 33.3% (Table 1, rightmost column). With reward scales matched across levels, conditioning itself drives the improvement in learning.

Additionally, we fit shared and conditional critics to identical training data and compare their advantages on the same policy update batch. Conditioning produces a markedly different advantage distribution, consistent with our theoretical account of how value mismatch reshapes sampled policy updates (Appendix H.3).

5 Related Work

Baselines and stochastic dynamics. Beyond classical estimator variance analyses (Weaver & Tao, 2001; Greensmith et al., 2004), recent work studies baseline placement, committal behavior, and update aggressiveness (Chung et al., 2021; Mei et al., 2021; 2022), alongside softmax convergence guarantees (Mei et al., 2023; Robertson et al., 2025). We identify value mismatch as a systematic placement error across environments: one shared value baseline can reinforce bad actions and suppress useful actions.

Conditioned values. Conditional critics are established tools (Schaul et al., 2015; Pinto et al., 2018), with theory covering unbiased gradients, variance, and aliasing (Baisero & Amato, 2022; Li et al., 2024; Lambrechts et al., 2025). AACC supplies simulator factors to the critic (Yue et al., 2024); PAMDP uses profile conditioning (Yang et al., 2026). Our arbitrary index instead isolates how separating values in discrete environments redistributes sampled updates along the learning path.

PopArt addresses heterogeneous return scales (van Hasselt et al., 2016; Hessel et al., 2019); its multitask value outputs are a direct architectural precedent. Widely separated scales can create

Table 1: Procgen, 16 games: final evaluation return on 600 unseen levels (25M steps; mean $\pm$ 1 s.d. over 10 seeds per method). Normalized returns per run are computed by dividing the average test return per run for each environment by the corresponding average test return of the shared critic baseline over all runs. **Bold** = best in the row.

	Evaluation return (600 unseen levels)			
Game	Shared	FiLM	Multihead	+PopArt
Bigfish	1.87 $\pm$0.17	2.88 $\pm$0.57	2.99 $\pm$0.61	**3.56 $\pm$1.27**
Bossfight	4.44 $\pm$2.56	7.44 $\pm$0.51	**8.97 $\pm$0.29**	8.11 $\pm$0.82
Caveflyer	1.08 $\pm$0.16	1.90 $\pm$0.23	**2.65 $\pm$0.43**	1.87 $\pm$0.31
Chaser	0.95 $\pm$0.26	0.94 $\pm$0.19	**1.32 $\pm$0.35**	0.75 $\pm$0.17
Climber	1.40 $\pm$0.19	1.74 $\pm$0.27	**2.75 $\pm$0.22**	2.13 $\pm$0.27
Coinrun	5.76 $\pm$0.29	6.35 $\pm$0.28	**6.80 $\pm$0.37**	6.46 $\pm$0.29
Dodgeball	0.98 $\pm$0.05	0.91 $\pm$0.14	1.10 $\pm$0.20	**1.32 $\pm$0.18**
Fruitbot	7.39 $\pm$1.58	**8.95 $\pm$0.70**	8.14 $\pm$1.24	8.31 $\pm$1.01
Heist	0.24 $\pm$0.06	**0.28 $\pm$0.05**	0.20 $\pm$0.03	0.24 $\pm$0.06
Jumper	2.23 $\pm$0.13	**2.36 $\pm$0.15**	2.32 $\pm$0.12	2.35 $\pm$0.62
Leaper	2.81 $\pm$0.36	3.02 $\pm$0.24	**3.25 $\pm$0.46**	3.09 $\pm$0.28
Maze	1.23 $\pm$0.13	1.33 $\pm$0.13	1.49 $\pm$0.18	**1.49 $\pm$0.15**
Miner	0.53 $\pm$0.12	0.58 $\pm$0.08	**0.80 $\pm$0.14**	0.77 $\pm$0.13
Ninja	3.39 $\pm$0.21	3.98 $\pm$0.15	**4.39 $\pm$0.29**	4.10 $\pm$0.41
Plunder	2.37 $\pm$0.33	2.38 $\pm$0.18	2.28 $\pm$0.18	**2.42 $\pm$0.29**
Starpilot	7.96 $\pm$1.06	10.22 $\pm$1.11	13.10 $\pm$1.46	**14.98 $\pm$2.35**
Normalized return (%)	100.0 $\pm$19.8	121.2 $\pm$27.9	**140.8 $\pm$47.1**	133.3 $\pm$41.7

large value mismatch, and we isolate the subtler case of all environments with comparable reward scales. Target normalization and environment conditioning therefore address distinct aspects of interference. We provide a comprehensive review of related work in Appendix A.

6 Discussion, Limitations, and Conclusion

Limitations. The theory isolates a pathwise mechanism in deterministic bandits with oracle baselines, a fixed mixture, and a common optimal arm. It does not establish a uniform ordering of expected finite-horizon return or a general MDP theorem. However, our theory reflects a mechanism present in practical RL: a shared baseline changes the advantages of sampled trajectories with high or ordinary returns, thereby shaping their contributions to policy updates in a bad way. On the other hand, index conditioning assumes recurring training environments with stable identifiers. New or continuously varying training environments may require simulator parameters or inferred context. Appendix B gives complete limitations and implications.

Value fitting and mismatch severity. Conditioning is most useful when environment identity explains value differences that the shared critic cannot resolve. With finite data, estimating separate components can be harder when some states in environments are rarely sampled. Mismatch depends on value differences across environments, not the number of environments. Many environments with similar values may cause little mismatch, while two with very different values can cause substantial mismatch. A small average mismatch can also hide a large mismatch in a rarely sampled environment, leading to more difficult RL training in practice.

Conclusion. We introduced value mismatch as a direct cause of worse stochastic learning dynamics when one critic is shared across environments. We give intuitive and clear theoretical supports to show the effect of value mismatch. Because the mechanism depends on value differences, it motivates a simple intervention: condition only the critic on the environment index. Across four benchmark families, this intervention improves learning while preserving one shared actor. A promising direction is to identify where mismatch is large and condition the critic only there, combining accurate conditional value fitting and correct advantage assignment.

AI USE STATEMENT

Generative AI tools assisted with manuscript organization and editing, proof checking and exposition, experimental code design. We do not use AI tools for generating research ideas, proposing mathematical methodology, and drafting the paper. The authors are responsible for verifying the AI-assisted material and take responsibility for the final text, claims, code, and artifacts.

ETHICS STATEMENT

This work studies optimization in simulated control and game benchmarks and does not involve human participants or deployment in a safety critical system. Improved training performance does not by itself establish the safety, fairness, or robustness of a downstream application; those properties require evaluation in its intended setting.

REPRODUCIBILITY STATEMENT

Section 2 specifies the theoretical assumptions and sampled update. Appendices C–F provide the proofs and auxiliary results. Appendix G gives architectures, parameter counts, hyperparameters, environment pools, seed counts, evaluation protocols, and aggregation rules. Appendix H describes the additional mechanism diagnostics. Figure and table captions specify the uncertainty convention used for each result. We leave the experiment code in the supplementary material.

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. URL https://arxiv.org/abs/1908.00261.

Andrea Baisero and Christopher Amato. Unbiased asymmetric reinforcement learning under partial observability. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 44–52, 2022. URL https://www.ifaamas.org/Proceedings/aamas2022/pdfs/p44.pdf.

Andrew G. Barto, Richard S. Sutton, and Charles W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-13(5):834–846, 1983.

Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. arXiv preprint arXiv:1606.01540, 2016. URL https://arxiv.org/abs/1606.01540.

Lonnie Chrisman. Reinforcement learning with perceptual aliasing: The perceptual distinctions approach. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, pp. 183–188. AAAI Press, 1992. URL https://aaai.org/Papers/AAAI/1992/AAAI92-029.pdf.

Wesley Chung, Valentin Thomas, Marlos C Machado, and Nicolas Le Roux. Beyond variance reduction: Understanding the true impact of baselines on policy optimization. In *ICML*, 2021.

Karl Cobbe, Chris Hesse, Jacob Hilton, and John Schulman. Leveraging procedural generation to benchmark reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2048–2056. PMLR, 2020. URL https://proceedings.mlr.press/v119/cobbe20a.html.

Karl Cobbe, Jacob Hilton, Oleg Klimov, and John Schulman. Phasic policy gradient. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2020–2027. PMLR, 2021. URL https://proceedings.mlr.press/v139/cobbe21a.html.

Daniel Ebi, Damien Ernst, Klemens Böhm, and Gaspard Lambrechts. Informed asymmetric actor-critic: Leveraging privileged signals beyond full-state access. In *Proceedings of the 43rd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2026. URL https://icml.cc/virtual/2026/poster/66074.

Lasse Espeholt, Hubert Soyer, Remi Munos, Karen Simonyan, Vlad Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. IMPALA: Scalable distributed deep-RL with importance weighted actor-learner architectures. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1407–1416. PMLR, 2018. URL https://proceedings.mlr.press/v80/espeholt18a.html.

Evan Greensmith, Peter L. Bartlett, and Jonathan Baxter. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5:1471–1530, 2004. URL https://www.jmlr.org/papers/v5/greensmith04a.html.

Shixiang Gu, Timothy Lillicrap, Zoubin Ghahramani, Richard E. Turner, and Sergey Levine. Q-Prop: Sample-efficient policy gradient with an off-policy critic. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1611.02247.

Matteo Hessel, Hubert Soyer, Lasse Espeholt, Wojciech Czarnecki, Simon Schmitt, and Hado van Hasselt. Multi-task deep reinforcement learning with PopArt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3796–3803, 2019. doi: 10.1609/aaai.v33i01.33013796. URL https://ojs.aaai.org/index.php/AAAI/article/view/4266.

Edward S. Hu, James Springer, Oleh Rybkin, and Dinesh Jayaraman. Privileged sensing scaffolds reinforcement learning. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0b334c6419042e761051d82fc621679c-Abstract-Conference.html.

Minqi Jiang, Edward Grefenstette, and Tim Rocktäschel. Prioritized level replay. In *International Conference on Machine Learning (ICML)*, 2021.

Vijay Konda and John Tsitsiklis. Actor-critic algorithms. In S. Solla, T. Leen, and K. Müller (eds.), *Advances in Neural Information Processing Systems*, volume 12. MIT Press, 1999. URL https://proceedings.neurips.cc/paper_files/paper/1999/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

Gaspard Lambrechts, Damien Ernst, and Aditya Mahajan. A theoretical justification for asymmetric actor-critic algorithms. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 32375–32405. PMLR, 2025. URL https://proceedings.mlr.press/v267/lambrechts25a.html.

Gen Li, Yuting Wei, Yuejie Chi, and Yuxin Chen. Softmax policy gradient methods can take exponential time to converge, 2022. URL https://arxiv.org/abs/2102.11270.

Jinqiu Li, Enmin Zhao, Tong Wei, Junliang Xing, and Shiming Xiang. Dual critic reinforcement learning under partial observability. In *Advances in Neural Information Processing Systems*, volume 37, pp. 116676–116704, 2024. doi: 10.52202/079017-3704. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d399b67fa017f0f7670102c88507720c-Abstract-Conference.html.

Hao Liu, Yihao Feng, Yi Mao, Dengyong Zhou, Jian Peng, and Qiang Liu. Action-dependent control variates for policy optimization via Stein identity. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1710.11198.

Zhenya Liu and Yuxin Chen. Active curriculum refinement for reinforcement learning. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026. URL https://icml.cc/virtual/2026/poster/65627.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvári, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *ICML*, 2020.

Jincheng Mei, Bo Dai, Chenjun Xiao, Csaba Szepesvári, and Dale Schuurmans. Understanding the effect of stochasticity in policy optimization. In *NeurIPS*, 2021.

Jincheng Mei, Wesley Chung, Valentin Thomas, Bo Dai, Csaba Szepesvári, and Dale Schuurmans. The role of baselines in policy gradient optimization. In *NeurIPS*, 2022.

Jincheng Mei, Zixin Zhong, Bo Dai, Alekh Agarwal, Csaba Szepesvári, and Dale Schuurmans. Stochastic gradient succeeds for bandits. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24325–24360. PMLR, 2023. URL https://proceedings.mlr.press/v202/mei23a.html.

Seungyong Moon, JunYeong Lee, and Hyun Oh Song. Rethinking value function learning for generalization in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/e19ab2dde2e60cf68d1ded18c38938f4-Abstract-Conference.html.

Jack Parker-Holder, Minqi Jiang, Michael Dennis, Mikayel Samvelyan, Jakob Foerster, Edward Grefenstette, and Tim Rocktäschel. Evolving curricula with regret-based environment design. In *International Conference on Machine Learning (ICML)*, 2022.

Xue Bin Peng, Marcin Andrychowicz, Wojciech Zaremba, and Pieter Abbeel. Sim-to-real transfer of robotic control with dynamics randomization. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3803–3810, 2018. doi: 10.1109/ICRA.2018.8460528. URL https://arxiv.org/abs/1710.06537.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *AAAI Conference on Artificial Intelligence*, 2018. URL https://arxiv.org/abs/1709.07871.

Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. Asymmetric actor critic for image-based robot learning. In *Robotics: Science and Systems (RSS)*, 2018. doi: 10.15607/RSS.2018.XIV.008. URL https://www.roboticsproceedings.org/rss14/p08.html.

Roberta Raileanu and Rob Fergus. Decoupling value and policy for generalization in reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*. PMLR, 2021. URL https://arxiv.org/abs/2102.10330.

Kate Rakelly, Aurick Zhou, Chelsea Finn, Sergey Levine, and Deirdre Quillen. Efficient off-policy meta-reinforcement learning via probabilistic context variables. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5331–5340. PMLR, 2019. URL https://proceedings.mlr.press/v97/rakelly19a.html.

Samuel Robertson, Thang Chu, Bo Dai, Dale Schuurmans, Csaba Szepesvári, and Jincheng Mei. REINFORCE converges to optimal policies with any learning rate. In *Advances in Neural Information Processing Systems*, volume 38, pp. 16010–16053, 2025. doi: 10.52202/085713-0541. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/17a84369cf6e2803101294b5d74dd92c-Abstract-Conference.html.

Tom Schaul, Daniel Horgan, Karol Gregor, and David Silver. Universal value function approximators. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 1312–1320. PMLR, 2015. URL http://proceedings.mlr.press/v37/schaul15.html.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1506.02438.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017. URL https://arxiv.org/abs/1707.06347.

Satinder P. Singh, Tommi Jaakkola, and Michael I. Jordan. Reinforcement learning with soft state aggregation. In *Advances in Neural Information Processing Systems 7*, 1994. URL https://proceedings.neurips.cc/paper_files/paper/1994/hash/287e03db1d99e0ec2edb90d079e142f3-Abstract.html.

Richard S Sutton and Andrew G Barto. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 23–30, 2017. doi: 10.1109/IROS.2017.8202133. URL https://ieeexplore.ieee.org/document/8202133.

Emanuel Todorov, Tom Erez, and Yuval Tassa. MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5026–5033, 2012.

George Tucker, Surya Bhupatiraju, Shixiang Gu, Richard E. Turner, Zoubin Ghahramani, and Sergey Levine. The mirage of action-dependent baselines in reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*. PMLR, 2018. URL https://arxiv.org/abs/1802.10031.

Hado van Hasselt, Arthur Guez, Matteo Hessel, Volodymyr Mnih, and David Silver. Learning values across many orders of magnitude. In *Advances in Neural Information Processing Systems*, 2016. URL https://arxiv.org/abs/1602.07714.

Lex Weaver and Nigel Tao. The optimal reward baseline for gradient-based reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 538–545, 2001. URL https://arxiv.org/abs/1301.2315.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

Cathy Wu, Aravind Rajeswaran, Yan Duan, Vikash Kumar, Alexandre M. Bayen, Sham Kakade, Igor Mordatch, and Pieter Abbeel. Variance reduction for policy gradient with action-dependent factorized baselines. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1803.07246.

Zhe Yang, Yi Huang, Si Chen, Xiaoting Wu, Jingyu Yao, and Junlan Feng. PAMDP: Interact to persona alignment via a partially observable markov decision process. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=tNWZVoVPzZ.

Wangyang Yue, Yuan Zhou, Xiaochuan Zhang, Yuchen Hua, Minne Li, Zunlin Fan, Zhiyuan Wang, and Guang Kou. Asymmetric actor-critic for adapting to changing environments in reinforcement learning. In *Artificial Neural Networks and Machine Learning – ICANN 2024*, volume 15019 of *Lecture Notes in Computer Science*, pp. 325–339. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72341-4_22. URL https://doi.org/10.1007/978-3-031-72341-4_22.

APPENDIX CONTENTS

The appendix contains the full proofs, experimental details, additional figures, and discussion supporting the main text.

A **Extended related work** . . . 16

B **Extended discussion and limitations** . . . 17

B.1 Statistical sharing versus target separation . . . 17

B.2 Mismatch severity . . . 17

B.3 Scope of the theory . . . 17

B.4 Scope of the method . . . 17

B.5 Implications and future directions . . . 17

C **What variance can and cannot explain** . . . 18

D **Proof of Proposition 1: convergence and rate** . . . 20

E **Proof of Proposition 2: the single-environment ratchet** . . . 29

F **Value mismatch proofs and GAE derivation** . . . 29

G **Full experimental details and results** . . . 35

G.1 Architectures and parameter counts . . . 35

G.2 CartPole: protocol and mechanism interpretation . . . 36

G.3 MuJoCo: protocol and effect sizes . . . 37

G.4 BipedalWalker and conditioning controls . . . 38

G.5 Procgen: protocol, tables, and evaluation curves . . . 39

G.6 Value loss diagnostics . . . 42

H **Additional mechanism evidence and controls** . . . 43

H.1 A minimal example with learned critics . . . 44

H.2 Identical visible inputs and different futures . . . 44

H.3 Critic choice on a fixed rollout . . . 45

A Extended related work

What do baselines do? The classical account is variance reduction of the policy gradient estimator (Williams, 1992; Weaver & Tao, 2001; Greensmith et al., 2004), refined by action-dependent and Stein control variates (Gu et al., 2017; Liu et al., 2018; Wu et al., 2018). On common benchmarks, Tucker et al. (2018) found no variance reduction from learned action-dependent baselines beyond a state-only baseline, tracing reported gains to implementation differences. In single environments, baseline placement controls sampled signs and committal behavior (Chung et al., 2021). Related work studies update aggressiveness (Mei et al., 2021; 2022), softmax convergence (Mei et al., 2020; Li et al., 2022; Agarwal et al., 2021; Mei et al., 2023), and REINFORCE convergence at any fixed learning rate (Robertson et al., 2025). We build on this process-level view: sharing a value creates a systematic placement error across environments, reintroducing rival reinforcement and suppressing useful samples despite using a value baseline.

Multitask scaling and environment variation. PopArt normalizes value targets while preserving unnormalized predictions (van Hasselt et al., 2016). Multitask PopArt combines a shared policy with value outputs indexed by task (Hessel et al., 2019), a direct architectural precedent. Widely separated task return scales make value mismatch especially visible when their critic lacks task information. We isolate the subtler case of different continuation values at comparable reward scales. Target normalization controls magnitude; conditioning separates environment-specific values. They address complementary aspects of interference.

Procgen and related benchmarks train one agent across sets of levels (Cobbe et al., 2020). PLR, ACCEL, and PATH adapt the level distribution (Jiang et al., 2021; Parker-Holder et al., 2022; Liu & Chen, 2026); domain randomization varies appearance or simulator dynamics (Tobin et al., 2017; Peng et al., 2018). IDAAC and analyses of value learning study decoupling and generalization under level variation (Raileanu & Fergus, 2021; Moon et al., 2022), while PPG separates policy and value optimization phases and increases value function sample reuse (Cobbe et al., 2021). These approaches change sampling, representation, or optimization. Our intervention changes only the critic's access to a logged environment index while retaining one actor and the same training budget.

Conditional and asymmetric critics. Conditioned values and information available only to the critic are established designs (Schaul et al., 2015; Pinto et al., 2018; Hu et al., 2024). Under partial observability, Baisero & Amato (2022) establish a conditional value identity and an unbiased asymmetric policy gradient; DCRL combines a critic using only history with one using both history and state to study a variance tradeoff (Li et al., 2024); and IAAC treats general privileged signals and studies expected gradient validity and informativeness (Ebi et al., 2026). In complementary theory for finite training horizons, privileged critics remove agent-state aliasing terms from linear actor–critic bounds (Lambrechts et al., 2025). These works explain validity, variance, or aliasing benefits. We instead study how environment-dependent value offsets redistribute realized updates even when the expected direction at a fixed policy is unchanged.

AACC encodes continuous simulator factors for the critic and studies adaptation under changing dynamics (Yue et al., 2024); it also relates conditioned values to their marginal over observations. PAMDP uses a dual critic conditioned on profiles for persona alignment (Yang et al., 2026). Our arbitrary categorical index carries no physical parameters, ordering, or profile semantics. Pooling values for inputs with different futures is a structured instance of perceptual aliasing (Chrisman, 1992; Singh et al., 1994); we study how the resulting offsets change sampled learning paths across parallel RL benchmarks.

B Extended discussion and limitations

B.1 Statistical sharing versus target separation

Conditioning is most useful when environment identity explains substantial variation in value. At the population level, the conditional critic classes contain the shared critic as a special case, so their optimal squared prediction error cannot be larger under the same data distribution. With finite data, however, estimating environment-dependent components can increase estimation error when values are already similar or some environments are sampled infrequently. The practical tradeoff is therefore between statistical sharing and separation of conflicting value targets. FiLM and multihead retain a shared representation while allowing the final prediction to separate. Empirically, the early transient in the compact CartPole study does not become a persistent fitting bottleneck on the larger benchmarks, as shown by the value loss curves in Appendix G.6.

B.2 Mismatch severity

The number of environments alone does not determine mismatch severity. At a state s visible to the critic, the average squared mismatch is $\mathbb{E}[(e_Z^\pi(s))^2 \mid S = s] = \mathrm{Var}(V_Z^\pi(s) \mid S = s)$: many tightly clustered environments may be benign, whereas two separated values can suffice. This average can also hide a rarely sampled environment with a large pointwise offset, producing infrequent but large miscentered updates.

B.3 Scope of the theory

The propositions do not establish a uniform finite-horizon ordering in expected optimal-arm probability or return. Instead, under their stated conditions, they isolate a clean pathwise mechanism: shared centering produces recurrent reversed updates, whereas conditional centering eventually rules them out. We deliberately study illustrative deterministic bandits sampled across a fixed set of environments. First, their baselines are *oracle*: in practice, both the environment-specific values and their shared marginal must be learned from data. Second, deterministic rewards make the branch signs pathwise; stochastic returns preserve the offset identity but not every realized sign. Third, the common optimal arm removes genuine conflict between environments, and the fixed environment mixture removes policy-dependent changes in environment sampling. Deep RL adds state visitation, function approximation, learned critics, GAE, and PPO transformations, so the experiments test whether the explanatory mechanism transfers rather than verify a general MDP convergence theorem. The mechanism itself is broader: it can arise whenever parallel environments produce different continuation values for inputs that the critic maps to the same representation.

B.4 Scope of the method

Our experiments use recurring environments with stable logged identifiers. This is a minimal diagnostic intervention, not a universal conditioning scheme: continuous or unseen variants may instead require simulator parameters or a learned or inferred representation (Rakelly et al., 2019). The actor never receives the identifier, so deployment remains unchanged. Conditioning addresses value mismatch, but not genuine task conflict, critic estimation error, or every possible cost of privileged information.

B.5 Implications and future directions

We introduced value mismatch as a direct cause of poorer sampled learning dynamics when one critic is shared across environments. Because the mechanism depends on value differences rather than a particular domain, it motivates a simple intervention: condition only the critic on the environment index. Across four benchmark families, this intervention improves learning while preserving one shared actor. A promising direction is to identify where mismatch is large and condition the critic only there, combining accurate conditional value fitting with statistical sharing elsewhere.

C What variance can and cannot explain

Classical baseline theory asks how an action-independent baseline reduces the variance of a policy gradient estimator (Weaver & Tao, 2001; Greensmith et al., 2004). This view gives a valid aggregate guarantee: within a fixed information set, a score-weighted baseline minimizes the trace of the gradient estimator covariance and maximizes the standard smoothness lower bound. It does not identify which sampled combinations of environment and arm carry the update, however, or how those branches change future sampling. We first state the variance guarantee and then show that its exact optimum can produce a slower and less stable process over a finite horizon than the ordinary conditional value baseline used in the main text.

Fix a policy parameter θ. All expectations and covariances below are under the current on-policy joint law of (S, Z, A, G), where $A \mid (S, Z) \sim \pi_\theta(\cdot \mid S)$. Let $J(\theta)$ denote the policy objective and assume the policy gradient identity $\mathbb{E}[G\psi] = \nabla J(\theta)$, where $\psi := \nabla_\theta \log \pi_\theta(A \mid S)$ and $w := \|\psi\|_2^2$. Let Y denote the information given to the baseline: $Y_s := S$ for a shared baseline and $Y_c := (S, Z)$ for a conditional baseline. Because the actor does not receive Z, $\mathbb{E}[\psi \mid Y_s] = \mathbb{E}[\psi \mid Y_c] = 0$.

For a baseline $B(Y)$, define $\widehat{g}_B := (G - B(Y))\psi$. If J is L_J-smooth, the ascent lemma gives

$$\mathbb{E}[J(\theta + \eta\widehat{g}_B)] \geq J(\theta) + \eta\|\nabla J(\theta)\|_2^2 - \frac{L_J\eta^2}{2}\mathbb{E}\big[w(G - B(Y))^2\big].$$

We denote the right-hand side by $\mathrm{LB}_Y(B; \eta)$ and write $\mathrm{LB}_s := \mathrm{LB}_{Y_s}$ and $\mathrm{LB}_c := \mathrm{LB}_{Y_c}$.

Proposition 4 (Richer baseline information improves the variance certificate)**.** *Suppose $\mathbb{E}[wG^2] < \infty$ and $0 < \mathbb{E}[w \mid Y] < \infty$ almost surely. Among all baselines measurable with respect to Y and satisfying $\mathbb{E}[wB(Y)^2] < \infty$, the unique minimizer up to almost-sure equality is*

$$B_Y^{\mathrm{mv}}(Y) = \frac{\mathbb{E}[wG \mid Y]}{\mathbb{E}[w \mid Y]}.$$

Equivalently, B_Y^{mv} minimizes $\mathbb{E}[w(G - B(Y))^2]$, minimizes the covariance trace $\operatorname{tr}\operatorname{Cov}(\widehat{g}_B)$, and maximizes $\mathrm{LB}_Y(B; \eta)$ for every $\eta > 0$. Let B_s^{mv} and B_c^{mv} denote the optima under Y_s and Y_c. Then

$$\mathrm{LB}_c(B_c^{\mathrm{mv}}; \eta) - \mathrm{LB}_s(B_s^{\mathrm{mv}}; \eta) = \frac{L_J\eta^2}{2}\,\mathbb{E}\Big[w\big(B_c^{\mathrm{mv}}(S, Z) - B_s^{\mathrm{mv}}(S)\big)^2\Big].$$

Proof. The baseline has zero mean score contribution, so every admissible B gives $\mathbb{E}[\widehat{g}_B] = \nabla J(\theta)$. Conditional Cauchy–Schwarz gives

$$\mathbb{E}[w \mid Y](B_Y^{\mathrm{mv}})^2 = \frac{\mathbb{E}[wG \mid Y]^2}{\mathbb{E}[w \mid Y]} \leq \mathbb{E}[wG^2 \mid Y],$$

so B_Y^{mv} is admissible. Moreover, $\mathbb{E}[w(G - B_Y^{\mathrm{mv}}) \mid Y] = 0$. Completing the square conditionally gives

$$\mathbb{E}[w(G - B)^2 \mid Y] = \mathbb{E}[w(G - B_Y^{\mathrm{mv}})^2 \mid Y] + \mathbb{E}[w \mid Y](B - B_Y^{\mathrm{mv}})^2.$$

This proves optimality and uniqueness. Because Y_c refines Y_s, the same orthogonality gives

$$\mathbb{E}[w(G - B_s^{\mathrm{mv}})^2] = \mathbb{E}[w(G - B_c^{\mathrm{mv}})^2] + \mathbb{E}[w(B_c^{\mathrm{mv}} - B_s^{\mathrm{mv}})^2],$$

which proves the gap identity. Finally, all admissible estimators have the same mean, and therefore

$$\operatorname{tr}\operatorname{Cov}(\widehat{g}_B) = \mathbb{E}[w(G - B)^2] - \|\nabla J(\theta)\|_2^2.$$

□

Relation to value mismatch. Consider the softmax policy with one logit per arm in Algorithm 1, two equally likely environments $Z \in \{E, H\}$, and two arms (a^*, i) with $r_z = (d_z, 0)$, where $d_E, d_H > 0$. Write $p := \pi(a^*)$. The squared score norms are $w(a^*) = 2(1 - p)^2$ and $w(i) = 2p^2$. Put $\bar{d} := (d_E + d_H)/2$. The average objective is $J(\theta) = \bar{d}p$, and its Hessian with respect to the two logits is

$$\nabla^2 J(\theta) = \bar{d}\,p(1 - p)(1 - 2p)\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

Consequently,

$$\sup_{\theta} \|\nabla^2 J(\theta)\|_{\mathrm{op}} = 2\bar{d} \max_{p\in[0,1]} |p(1-p)(1-2p)| = \frac{\bar{d}}{3\sqrt{3}}.$$

Thus $L_J = \bar{d}/(3\sqrt{3})$ is the smallest global Euclidean smoothness constant for this parameterization. Proposition 4 then gives

$$B_c^{\mathrm{mv}}(z) = (1-p)d_z, \qquad B_s^{\mathrm{mv}} = (1-p)\bar{d}.$$

The exact gain from the richer information set is

$$\mathrm{LB}_c(B_c^{\mathrm{mv}};\eta) - \mathrm{LB}_s(B_s^{\mathrm{mv}};\eta) = \frac{\bar{d}\eta^2}{12\sqrt{3}}\, p(1-p)^3(d_E - d_H)^2.$$

Here $V_z^\pi = pd_z$, $\bar{V}^\pi = p\bar{d}$, and $e_z^\pi := V_z^\pi - \bar{V}^\pi$. Hence the same difference equals $\frac{\bar{d}\eta^2}{3\sqrt{3}}((1-p)^3/p)\mathbb{E}[(e_Z^\pi)^2]$. At a fixed policy and data distribution, the variance certificate therefore improves quadratically with the value mismatch.

The exact baseline that minimizes the covariance trace is not generally the ordinary value prediction used by actor–critic methods:

$$B_Y^{\mathrm{mv}} = \mathbb{E}[G \mid Y] + \frac{\mathrm{Cov}(w, G \mid Y)}{\mathbb{E}[w \mid Y]}.$$

The correction appears because $w = \|\mathbf{e}_A - \pi\|_2^2$ gives more weight to actions with larger score norms. An ordinary value critic ignores this action dependent weight and minimizes the simpler prediction error $\mathbb{E}[(G - B(Y))^2]$. It is therefore an unweighted surrogate for the exact minimizer of the covariance trace, but it still has a precise variance interpretation. Define $B^{\mathrm{shared}}(S) := \mathbb{E}[G \mid S]$, $B^{\mathrm{cond}}(S,Z) := V_Z^\pi(S) := \mathbb{E}[G \mid S, Z]$, and $e_Z^\pi(S) := V_Z^\pi(S) - B^{\mathrm{shared}}(S)$. Conditional expectation gives

$$\mathbb{E}[(G - B^{\mathrm{shared}}(S))^2] = \mathbb{E}[(G - B^{\mathrm{cond}}(S,Z))^2] + \mathbb{E}[(e_Z^\pi(S))^2].$$

Moreover, $w \le 2$ for this parameterization, so

$$\mathbb{E}[w(G-B)^2] \le 2\mathbb{E}[(G-B)^2].$$

Within either information set, ordinary value prediction therefore minimizes a valid upper bound on the second moment in the smoothness certificate. Replacing the shared value by the conditional value tightens the resulting lower bound by exactly $L_J\eta^2\mathbb{E}[(e_Z^\pi(S))^2]$. Larger mismatch thus strengthens the conventional variance argument for conditioning. This aggregate certificate, however, still does not reveal how the update is allocated across sampled branches or which stochastic process follows a better path over a finite horizon.

A numerical counterexample. Figure 7 uses two equally likely environments with three arms (a^*, a_2, a_3), rewards $r_E = (1, 0.8, 0)$ and $r_H = (0.5, 0.4, 0)$, common initialization $\pi_0^{\mathrm{mv}} = \pi_0^{\mathrm{cond}} = (0.5, 0.4, 0.1)$, and $\eta = 10$. We run two copies of the score update in Algorithm 1. Each copy recomputes its own baseline from its current policy at every round:

$$B_t^{\mathrm{mv}}(z) = \frac{\sum_a \pi_t^{\mathrm{mv}}(a) w_t^{\mathrm{mv}}(a) r_z(a)}{\sum_a \pi_t^{\mathrm{mv}}(a) w_t^{\mathrm{mv}}(a)}, \qquad B_t^{\mathrm{cond}}(z) = V_z^{\pi_t^{\mathrm{cond}}} = \sum_a \pi_t^{\mathrm{cond}}(a) r_z(a),$$

where $w_t^{\mathrm{mv}}(a) := \|\mathbf{e}_a - \pi_t^{\mathrm{mv}}\|_2^2$. With one logit per arm, the first baseline exactly minimizes the covariance trace; the second is the ordinary conditional value.

In environment E, their initial values are 0.704 and 0.820. The covariance trace minimizer therefore lies below $r_E(a_2) = 0.8$ and reinforces a sample of a_2, whereas the conditional value lies above it and suppresses the same sample. That branch sends $\pi(a^*)$ from 0.5 to 0.279 and 0.547, respectively. Yet the initial covariance trace is smaller under that baseline: 0.0423 versus 0.0472.

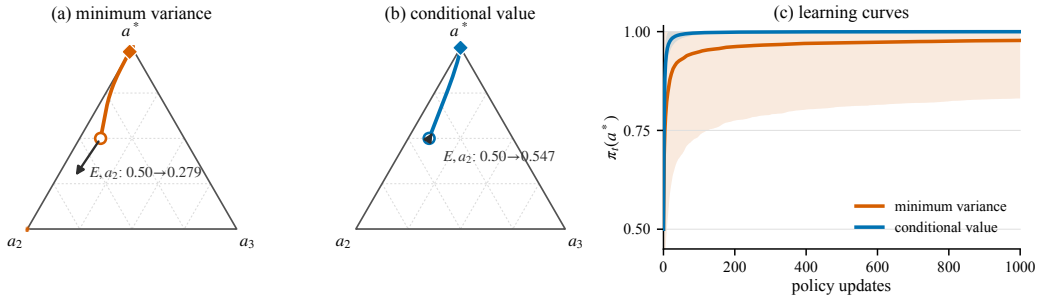

Figure 7: Minimizing the covariance trace need not produce a better process over a finite horizon. The two environments share the same optimal arm and are sampled equally; $r_H = 0.5r_E$, $\pi_0 = (0.5, 0.4, 0.1)$, and $\eta = 10$. (a,b) Terminal policies at $T = 1{,}000$ (light points; $4{,}000$ of $20{,}000$ shown) and ensemble mean trajectories (solid curves) from the common initialization (circle) to the terminal mean (diamond). The arrows show the same sampled branch at the initial policy: sampling a_2 in the easy environment E moves away from a^* under the baseline minimizing trace covariance but toward a^* under the conditional value. The initial covariance traces are 0.0423 and 0.0472, respectively. (c) Optimal arm probability over all $20{,}000$ trajectories (mean $\pm$ 1 s.d.; common random numbers). Both baselines are recomputed from their own current policies at every update. This compares the means at a fixed horizon and does not claim pathwise or asymptotic dominance.

Over $20{,}000$ trajectories with common random numbers, after $1{,}000$ updates the process using the covariance trace minimizer reaches 0.9777 ± 0.1468, while the conditional value process reaches 0.9998 ± 0.0004 (mean $\pm$ 1 s.d.). The conditional ensemble mean crosses 0.99 after 38 updates; the mean under the covariance trace minimizer remains below 0.99 through update $1{,}000$. The second arm remains the policy winner in 2.2% of runs using that baseline and in none of the conditional value runs. This is an ensemble comparison rather than pathwise dominance, but it shows that minimizing the local covariance trace can create a substantially longer and less stable committal transient.

The example exposes what the variance summary discards. At initialization, score weighting reduces the relative weight of the most probable optimal arm and more than doubles the relative weight of the rare arm with zero reward, pulling the covariance trace minimizer below $r_E(a_2)$. Baseline placement then changes which sampled arm is reinforced, and that change feeds back into future action sampling. The two processes still share the same asymptotic destination. Indeed, after averaging over Z, each has effective arm means $\bar{r}(a) - c_t^B$, where $c_t^B := \sum_z q_z B_t^B(z)$ is bounded and predictable; hence Lemma 8 applies. The example complements Chung et al. (2021), whose stronger convergence counterexample uses a natural policy gradient process with three arms. Variance remains a useful aggregate certificate, but it cannot explain the branch allocation that governs the sampled learning path.

D Proof of Proposition 1

This section gives a complete convergence proof for the predictable reward laws that depend on the environment and are induced by Algorithm 1. The proof architecture is adapted from the exploration, barrier, and elimination arguments of Robertson et al. (2025), including their use of Freedman's inequality to prove divergence. All ingredients needed here, including the rate for tail averages, are stated and proved directly in this appendix.

Proof convention. The shaded *Comment* boxes mark the steps at which we adapt the proof architecture of Robertson et al. (2025) to our multi-environment common shift setting. The intervening probabilistic argument is reproduced self-contained. "Rivals" refer to suboptimal arms or actions in the body of proof.

Proof idea. The argument has three main steps. First, after averaging over the sampled environment, every baseline induces effective arm means of the form $\bar{r}(a) - c_t$, where the same predictable

scalar c_t is subtracted from every arm. This shift changes the sampled noise but leaves every pairwise arm gap fixed, and all three baselines induce the same conditional mean update. Second, conditional Borel–Cantelli gives infinite exploration. The optimal logit has positive mean drift; infinite exploration makes its cumulative drift diverge, and stopped concentration prevents the noise from canceling that drift, so the optimal logit tends to $+\infty$. Third, a barrier argument reaches an arbitrarily large simultaneous gap and then traps the process in a region where every rival has strictly negative mean drift. Continued exploration makes each accumulated negative drift diverge, driving every rival logit to $-\infty$. Consequently, $\pi_t(a^*) \to 1$ almost surely for all three baselines. Finally, the same compensator argument yields, without requiring c_t to settle, a pathwise $O((1+\log N)/N)$ bound on average suboptimality over a random tail after the transient; it is neither a last iterate nor an expected rate.

Effective reward law. Let $\mathcal{F}_t$ contain the complete history before (Z_t, a_t) is sampled, so π_t is $\mathcal{F}_t$-measurable. Because Algorithm 1 samples Z_t independently from q and the actor does not observe Z_t, conditionally on $\mathcal{F}_t$ the draws $Z_t \sim q$ and $a_t \sim \pi_t$ are independent. Define

$$Y_t^B := r_{Z_t}(a_t) - B_t^B, \qquad \bar{r}(a) := \sum_z q_z r_z(a), \qquad \bar{V}^{\pi_t} := \sum_a \pi_t(a)\bar{r}(a).$$

Also put $R := \max_{z,a} |r_z(a)| < \infty$.

Lemma 5 (Exact reduction to a common shift)**.** *For $B \in \{0, \mathrm{shared}, \mathrm{cond}\}$, the conditional law of Y_t^B given $(\mathcal{F}_t, a_t = a)$ is predictable, supported on $[-2R, 2R]$, and has mean*

$$m_t^B(a) := \mathbb{E}[Y_t^B \mid \mathcal{F}_t, a_t = a] = \bar{r}(a) - c_t^B, \quad c_t^0 = 0, \quad c_t^{\mathrm{shared}} = c_t^{\mathrm{cond}} = \bar{V}^{\pi_t}. \tag{8}$$

Consequently, every pairwise gap is fixed, $m_t^B(a) - m_t^B(i) = \bar{r}(a) - \bar{r}(i)$, and

$$\mathbb{E}[\Delta\theta_t(a) \mid \mathcal{F}_t] = \eta\pi_t(a)\big(\bar{r}(a) - \bar{V}^{\pi_t}\big). \tag{9}$$

Proof. The zero and shared cases follow directly from their definitions. For the conditional baseline, conditional independence of Z_t and a_t gives

$$\mathbb{E}[Y_t^{\mathrm{cond}} \mid \mathcal{F}_t, a_t = a] = \sum_z q_z\big(r_z(a) - V_z^{\pi_t}\big) = \bar{r}(a) - \bar{V}^{\pi_t}.$$

Comment. The conditional oracle baseline is not a common shift on a realized environment branch. Independence of Z_t and a_t makes its mean conditional on the arm exactly the same predictable common shift after averaging over the environment index.

Substitution into the score update proves equation 9. Every oracle value is a convex combination of the finite reward table, so $|Y_t^B| \le 2R$ in all three cases. □

Throughout the remainder of the proof, write $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot \mid \mathcal{F}_t]$ and $\mathrm{Var}_t(\cdot) := \mathrm{Var}(\cdot \mid \mathcal{F}_t)$; for a stopping time τ, $\mathbb{P}_\tau(\cdot) := \mathbb{P}(\cdot \mid \mathcal{F}_\tau)$.

We first record a stopped concentration lemma that is uniform over time. It is a direct Freedman–Bernstein argument, included so that the buffers used below have explicit probabilities and do not depend on an external bandit result.

Lemma 6 (Stopped Freedman bound under drift dominance)**.** *Let τ be a stopping time finite almost surely and let $(X_t)_{t\ge\tau}$ be an adapted process taking real values with*

$$X_{t+1} - X_t = s_t + \xi_{t+1}, \qquad \mathbb{E}[\xi_{t+1} \mid \mathcal{F}_t] = 0, \qquad |\xi_{t+1}| \le b,$$

where $0 \le b < \infty$ is deterministic. Suppose that, for some deterministic $0 \le c < \infty$ with $b + c > 0$, one of the following two conditions holds at every active step:

$$s_t = \nu_t \ge 0, \qquad \mathbb{E}[\xi_{t+1}^2 \mid \mathcal{F}_t] \le c\nu_t, \tag{10}$$

$$s_t = -\nu_t \le 0, \qquad \mathbb{E}[\xi_{t+1}^2 \mid \mathcal{F}_t] \le c\nu_t, \tag{11}$$

where $\nu_t \geq 0$ is predictable. For $q \in (0,1)$, define

$$D_q(b,c) := 2(b+c)\log(1/q). \tag{12}$$

Conditionally on $\mathcal{F}_\tau$, under equation 10,

$$\mathbb{P}_\tau\left(\inf_{n\geq\tau}(X_n - X_\tau) < -D_q(b,c)\right) \leq q, \tag{13}$$

whereas under equation 11,

$$\mathbb{P}_\tau\left(\sup_{n\geq\tau}(X_n - X_\tau) > D_q(b,c)\right) \leq q. \tag{14}$$

Moreover, if $\Lambda_n := \sum_{t=\tau}^{n-1}\nu_t \to \infty$, then $X_n \to +\infty$ under equation 10 and $X_n \to -\infty$ under equation 11, almost surely. More precisely, writing $M_n := \sum_{t=\tau}^{n-1}\xi_{t+1}$, almost surely there is a finite random index N such that

$$|M_n| \leq \frac{1}{2}\Lambda_n \qquad (n \geq N). \tag{15}$$

The conclusions remain valid up to a further stopping time after multiplying every increment by the predictable indicator that this stopping time has not yet occurred.

Proof. For $0 < \lambda < 3/b$, the conditional Bernstein bound for the martingale difference is

$$\mathbb{E}_t[e^{\lambda\xi_{t+1}}] \leq \exp\left(\frac{\lambda^2}{2(1-\lambda b/3)}\mathbb{E}_t[\xi_{t+1}^2]\right), \tag{16}$$

and the same inequality holds for $-\xi_{t+1}$. Take $\lambda := 1/[2(b+c)]$. Then $\lambda b \leq 1/2$, $\lambda c \leq 1/2$, and, writing $\psi_b(\lambda) := \lambda^2/[2(1-\lambda b/3)]$,

$$\psi_b(\lambda)c \leq \frac{3}{10}\lambda < \frac{1}{2}\lambda. \tag{17}$$

For $M_n := \sum_{t=\tau}^{n-1}\xi_{t+1}$ and $\Lambda_n := \sum_{t=\tau}^{n-1}\nu_t$, the exponential process

$$\exp\left(\lambda M_n - \psi_b(\lambda)\sum_{t=\tau}^{n-1}\mathbb{E}_t[\xi_{t+1}^2]\right)$$

is a nonnegative supermartingale starting at one; so is the process with M_n replaced by $-M_n$. Ville's inequality and equation 17 therefore give, for every $x > 0$,

$$\mathbb{P}_\tau(\exists n \geq \tau : M_n \geq \Lambda_n + x) \leq e^{-\lambda x}, \qquad \mathbb{P}_\tau(\exists n \geq \tau : -M_n \geq \Lambda_n + x) \leq e^{-\lambda x}. \tag{18}$$

Since $X_n - X_\tau = \Lambda_n + M_n$ in the positive case and $X_n - X_\tau = -\Lambda_n + M_n$ in the negative case, taking $x = \lambda^{-1}\log(1/q)$ proves both equation 13 and equation 14.

For the divergence statements, equation 17 leaves the deterministic slack $\kappa := \lambda/2 - \psi_b(\lambda)c > 0$. Ville's inequality gives, for each sign,

$$\mathbb{P}_\tau(\exists n : \Lambda_n \geq u,\ M_n \geq \Lambda_n/2) \leq e^{-\kappa u}, \qquad \mathbb{P}_\tau(\exists n : \Lambda_n \geq u,\ -M_n \geq \Lambda_n/2) \leq e^{-\kappa u}$$

for every $u > 0$. Apply both bounds at $u = 1, 2, \ldots$. Their failure probabilities are summable. By Borel–Cantelli, almost surely only finitely many integer levels u admit an index n with $\Lambda_n \geq u$ and either displayed deviation. On $\Lambda_n \to \infty$, such a deviation at arbitrarily late indices would cross arbitrarily large integer levels and contradict this conclusion. Thus eventually $|M_n| < \Lambda_n/2$. Hence $\Lambda_n + M_n \to +\infty$ in the positive case and $-\Lambda_n + M_n \to -\infty$ in the negative case. Multiplying by a predictable stopping indicator preserves all conditional mean, variance, and increment bounds, proving the final assertion. □

Comment. For $G(\pi) := \mathrm{Diag}(\pi) - \pi\pi^\top$, every predictable common shift satisfies

$$G(\pi_t)(\mu - c_t\mathbf{1}) = G(\pi_t)\mu, \qquad G(\pi_t)\mathbf{1} = 0.$$

Thus the nonsettling offset changes the martingale noise law but never the mean direction or an arm gap. The coordinate formulas and offset-uniform variance bounds are derived below wherever they are used.

Lemma 7 (Elimination for general K under a predictable common shift)**.** *Assume $K \geq 2$, $0 < \eta < \infty$, and*

$$\theta_{t+1} = \theta_t + \eta Y_t(\mathbf{e}_{a_t} - \pi_t), \qquad \pi_t = \mathrm{softmax}(\theta_t), \qquad |Y_t| \leq C, \tag{19}$$

where $\mathbf{e}_a$ is the ath standard basis vector, $a_t \mid \mathcal{F}_t \sim \pi_t$, and

$$\mathbb{E}[Y_t \mid \mathcal{F}_t, a_t = a] = \mu(a) - c_t. \tag{20}$$

Here μ is fixed, c_t is predictable, and μ has the unique maximizer a^*. Suppose, almost surely, that

$$\theta_t(a^*) \to +\infty, \qquad \sum_{t=0}^{\infty} \pi_t(i) = \infty \quad (i \neq a^*). \tag{21}$$

Then

$$\theta_t(i) \to -\infty \quad (i \neq a^*), \qquad \pi_t(a^*) \to 1 \quad \textit{almost surely}. \tag{22}$$

Proof. Let $\mathcal{R} := \{1, \ldots, K\} \setminus \{a^*\}$ and define the fixed optimal gaps

$$d_i := \mu(a^*) - \mu(i), \qquad \delta := \min_{i \in \mathcal{R}} d_i > 0, \qquad L := \max_{i \in \mathcal{R}} d_i < \infty.$$

Comment. Robertson et al. organize elimination by reward tiers so that ties and multiple optimal arms are allowed. Here the shared optimum is strict, so one simultaneous optimal–rival cone controls every rival without a tier induction. The proof below defines the required stopping times and good events explicitly.

Write $p_t^* := \pi_t(a^*)$ and $\bar{d}_t := \sum_{i \in \mathcal{R}} \pi_t(i) d_i$. The common shift cancels exactly, giving

$$m_t(a^*) := \mathbb{E}_t[\Delta\theta_t(a^*)] = \eta p_t^* \bar{d}_t, \tag{23}$$

$$m_t(i) := \mathbb{E}_t[\Delta\theta_t(i)] = \eta\pi_t(i)(\bar{d}_t - d_i). \tag{24}$$

The score form and $|Y_t| \leq C$ also give, for every arm a,

$$|\Delta\theta_t(a)| \leq B_0 := \eta C, \tag{25}$$

$$\mathbb{E}_t|\Delta\theta_t(a)| \leq 2\eta C\pi_t(a)(1 - \pi_t(a)),$$

$$\mathrm{Var}_t(\Delta\theta_t(a)) \leq \eta^2 C^2 \pi_t(a)(1 - \pi_t(a)). \tag{26}$$

Step 1: arbitrary simultaneous gaps are reached. We claim that, for every deterministic $G > 0$, almost surely there is a finite time T_G such that

$$\theta_{T_G}(a^*) - \max_{i \in \mathcal{R}} \theta_{T_G}(i) \geq G. \tag{27}$$

Fix $G > 0$ and $q \in (0, 1)$. Choose a deterministic $M > 1$ satisfying

$$M \geq \frac{2(L - \delta)}{\delta}, \qquad \frac{M^2}{1 + 2M} \geq \frac{4C}{\delta}, \tag{28}$$

and set

$$U_2 := B_0 + \log((K - 1)M), \qquad U_1 := U_2 + G. \tag{29}$$

From equation 23 and equation 26,

$$m_t(a^*) \geq \eta\delta p_t^*(1 - p_t^*) \geq 0, \qquad \mathrm{Var}_t(\Delta\theta_t(a^*)) \leq c_* m_t(a^*), \quad c_* := \frac{\eta C^2}{\delta}. \tag{30}$$

The centered optimal increment has magnitude at most $b := 2B_0$. Define $D_*(q) := D_q(b, c_*)$ and the stopping time

$$\tau := \inf\{t : \theta_t(a^*) \geq U_1 + D_*(q)\}. \tag{31}$$

It is finite almost surely by equation 21. Lemma 6, restarted at τ, shows that

$$\mathcal{E}_q := \{\theta_t(a^*) \geq U_1 \text{ for every } t \geq \tau\} \tag{32}$$

has conditional probability at least $1 - q$.

Define $\Phi_t := \sum_{i \in \mathcal{R}} [\theta_t(i)]_+$, where $[x]_+ := \max\{x, 0\}$. Consider a time $t \geq \tau$ at which $\theta_t(a^*) \geq U_1$ and at least one rival logit is at least U_2. Partition

$$\mathcal{B}_t := \{i \in \mathcal{R} : \theta_t(i) \geq B_0\}, \qquad \mathcal{C}_t := \mathcal{R} \setminus \mathcal{B}_t,$$

and write $P_B := \pi_t(\mathcal{B}_t)$ and $P_C := \pi_t(\mathcal{C}_t)$. For $i \in \mathcal{B}_t$, a step of magnitude at most B_0 cannot cross below zero, so the expected increment of $[\theta_t(i)]_+$ equals $m_t(i)$. For $i \in \mathcal{C}_t$, the 1-Lipschitz property of $[\cdot]_+$ and equation 26 give an expected increase of at most $2\eta C \pi_t(i)$. Furthermore, equation 24 gives

$$\begin{aligned} \frac{1}{\eta} \sum_{i \in \mathcal{B}_t} m_t(i) &= -(1 - P_B) \sum_{i \in \mathcal{B}_t} \pi_t(i) d_i + P_B \sum_{i \in \mathcal{C}_t} \pi_t(i) d_i \\ &\leq -\delta p_t^* P_B + (L - \delta) P_B P_C. \end{aligned} \tag{33}$$

If $P_C = 0$, this is nonpositive. Otherwise, at least one rival is above U_2, whereas all logits in $\mathcal{C}_t$ are below B_0. Hence

$$\frac{P_B}{P_C} \geq M, \qquad \frac{p_t^*}{P_C} \geq M. \tag{34}$$

Indeed, each denominator has at most $K - 1$ terms, and $\theta_t(a^*) \geq U_1 > U_2$. If $x := p_t^*/P_C$ and $y := P_B/P_C$, normalization gives $P_C = (1 + x + y)^{-1}$ and therefore

$$\frac{p_t^* P_B}{P_C} = \frac{xy}{1 + x + y} \geq \frac{M^2}{1 + 2M} \geq \frac{4C}{\delta}. \tag{35}$$

The first condition in equation 28 also gives $(L - \delta) P_B P_C \leq (\delta/2) p_t^* P_B$. Combining these bounds yields

$$\mathbb{E}_t[\Phi_{t+1} - \Phi_t] \leq \eta \left[-\delta p_t^* P_B + (L - \delta) P_B P_C + 2C P_C\right] \leq 0. \tag{36}$$

Let ν be the first $t \geq \tau$ at which either $\theta_t(a^*) < U_1$ or every rival logit is below U_2. The drift bound in equation 36 shows that $\Phi_{t \wedge \nu} = \sum_{i \in \mathcal{R}} [\theta_{t \wedge \nu}(i)]_+$ is a nonnegative supermartingale. To handle its random starting value formally, restrict to each event $\{\tau = n, \Phi_n \leq m\} \in \mathcal{F}_n$ for integers n, m and then take their countable union. The convergence theorem for nonnegative supermartingales shows that the stopped process has a finite limit almost surely. On $\mathcal{E}_q$, the first stopping condition never occurs. If the second occurs, equation 29 gives equation 27; if it never occurs, the positive parts of all rival logits are bounded while $\theta_t(a^*) \to +\infty$, and equation 27 again holds at a finite later time. Thus the probability of never attaining the gap is at most q. Letting $q \downarrow 0$ proves the claim.

Step 2: the process is eventually trapped in a cone where every rival has strictly negative drift. Choose

$$\rho \in \left(\max\left\{1 - \frac{\delta}{L}, \frac{1}{K}\right\}, 1\right), \qquad \gamma := \delta - L(1 - \rho) > 0. \tag{37}$$

Whenever $p_t^* \geq \rho$, equation 23 and equation 24 imply

$$m_t(a^*) \geq 0, \qquad m_t(i) \leq -\eta\gamma\pi_t(i) \quad (i \in \mathcal{R}). \tag{38}$$

Together with equation 26, these give

$$\mathrm{Var}_t(\Delta\theta_t(a^*)) \leq c_* m_t(a^*), \qquad \mathrm{Var}_t(\Delta\theta_t(i)) \leq c_R[-m_t(i)], \quad c_R := \frac{\eta C^2}{\gamma}. \tag{39}$$

Fix $\varepsilon \in (0, 1)$, put $q := \varepsilon/K$, and define

$$D_* := D_q(b, c_*), \qquad D_R := D_q(b, c_R), \qquad g_\rho := \log \frac{(K - 1)\rho}{1 - \rho}, \qquad G_\varepsilon := g_\rho + D_* + D_R. \tag{40}$$

By Step 1, the first time T_ε such that

$$\theta_{T_\varepsilon}(a^*) - \theta_{T_\varepsilon}(i) \geq G_\varepsilon \qquad (i \in \mathcal{R}) \tag{41}$$

is a stopping time finite almost surely. This inequality implies $p^*_{T_\varepsilon} \geq \rho$. Let $\sigma := \inf\{t \geq T_\varepsilon : p^*_t < \rho\}$ and define the stopped increments $\widetilde{\Delta\theta_t(a)} := \mathbf{1}\{t < \sigma\}\Delta\theta_t(a)$. The indicator is $\mathcal{F}_t$-measurable, and every active update lies in the cone, so both equation 38 and equation 39 apply to the stopped process. Lemma 6, restarted at T_ε, and a union bound over all K arms give an event $\mathcal{G}_\varepsilon$ of conditional probability at least $1-\varepsilon$ on which, simultaneously for all $n \geq T_\varepsilon$ and all rivals i,

$$\theta_{n\wedge\sigma}(a^*) \geq \theta_{T_\varepsilon}(a^*) - D_*, \qquad \theta_{n\wedge\sigma}(i) \leq \theta_{T_\varepsilon}(i) + D_R. \tag{42}$$

If $\sigma < \infty$ on this event, then equation 41 and equation 42 imply

$$\theta_\sigma(a^*) - \theta_\sigma(i) \geq g_\rho \quad (i \in \mathcal{R}),$$

and hence

$$p^*_\sigma = \frac{1}{1 + \sum_{i\in\mathcal{R}} e^{\theta_\sigma(i)-\theta_\sigma(a^*)}} \geq \frac{1}{1+(K-1)e^{-g_\rho}} = \rho,$$

contradicting the definition of σ. Therefore, with probability at least $1-\varepsilon$, the process never leaves the cone after T_ε. If $\mathcal{A}$ denotes the event that there is no finite time after which the process remains in the cone, then $\mathbb{P}(\mathcal{A}) \leq \varepsilon$ for every $\varepsilon > 0$. Thus

$$\mathbb{P}(\exists T < \infty : \ p^*_t \geq \rho \text{ for every } t \geq T) = 1. \tag{43}$$

Step 3: eliminate every rival from deterministic restart times. The last entrance time in equation 43 need not be a stopping time. For each deterministic integer $n \geq 0$, define instead

$$\sigma_n := \inf\{t \geq n : p^*_t < \rho\}.$$

Fix $i \in \mathcal{R}$ and use the increments $\mathbf{1}\{t < \sigma_n\}\Delta\theta_t(i)$ from time n. On $\{\sigma_n = \infty\}$, equation 38 and equation 39 give the case of negative drift in Lemma 6, and its compensator satisfies

$$\sum_{t=n}^{\infty}[-m_t(i)] \geq \eta\gamma \sum_{t=n}^{\infty} \pi_t(i) = \infty$$

by equation 21. Hence $\theta_t(i) \to -\infty$ on $\{\sigma_n = \infty\}$, almost surely. The event in equation 43 agrees almost surely with $\bigcup_{n\geq 0}\{\sigma_n = \infty\}$, whose probability is one; the countable union therefore proves the limit for this rival. There are finitely many rivals, so all limits hold simultaneously. Together with $\theta_t(a^*) \to +\infty$, every margin between the optimal arm and a rival diverges and $\pi_t(a^*) \to 1$. □

The preceding rank free barrier plays the role of the tier elimination step in earlier proofs for softmax bandits. We can now prove the abstract convergence statement needed by Proposition 1.

Lemma 8 (Bounded predictable common shifts)**.** *Let $(\mathcal{F}_t)_{t\geq 0}$ be a filtration, let θ_t be $\mathcal{F}_t$-measurable, and put $\pi_t = \mathrm{softmax}(\theta_t)$. Suppose c_t is $\mathcal{F}_t$-measurable, (a_t, Y_t) and hence θ_{t+1} are $\mathcal{F}_{t+1}$-measurable, $a_t \mid \mathcal{F}_t \sim \pi_t$, and the conditional law of Y_t given $(\mathcal{F}_t, a_t = a)$ is an $\mathcal{F}_t$-measurable kernel. Consider*

$$\theta_{t+1} = \theta_t + \eta Y_t(\mathbf{e}_{a_t} - \pi_t), \qquad |Y_t| \leq C < \infty, \tag{44}$$

where

$$\mathbb{E}[Y_t \mid \mathcal{F}_t, a_t = a] = \mu(a) - c_t. \tag{45}$$

Here $\mu \in \mathbb{R}^K$ is fixed, c_t is a predictable scalar, and μ has a unique maximizer a^*. For every fixed finite $\eta > 0$ and finite initial logits,

$$\pi_t(a^*) \longrightarrow 1 \qquad \textit{almost surely}. \tag{46}$$

Proof. Put $p_t := \pi_t(a^*)$. The case $K = 1$ is immediate, so suppose $K \geq 2$.

Conservation and moment bounds. The score coordinates sum to zero, so $\sum_a \theta_t(a) = \sum_a \theta_0(a)$ pathwise. The predictable shift cancels from the drift:

$$\mathbb{E}_t[\Delta\theta_t(a)] = \eta\pi_t(a)\left(\mu(a) - \sum_j \pi_t(j)\mu(j)\right), \tag{47}$$

$$\begin{aligned} |\Delta\theta_t(a)| &\leq \eta C, \\ \mathbb{E}_t|\Delta\theta_t(a)| &\leq 2\eta C\pi_t(a)(1-\pi_t(a)), \\ \mathrm{Var}_t(\Delta\theta_t(a)) &\leq \eta^2 C^2 \pi_t(a)(1-\pi_t(a)). \end{aligned} \tag{48}$$

Exploration of all arms. Lévy's conditional Borel–Cantelli lemma gives, simultaneously for all arms,

$$\{a_t = a \text{ only finitely often}\} = \left\{\sum_t \pi_t(a) < \infty\right\} \quad \text{almost surely.} \tag{49}$$

Suppose an arm a were sampled only finitely often. After its last sample, $|\Delta\theta_t(a)| \le \eta C \pi_t(a)$, so equation 49 implies that $\theta_t(a)$ converges to a finite value. At the same time $\pi_t(a) \to 0$, hence $\max_j \theta_t(j) \to +\infty$. Logit conservation then forces some arm b to have $\liminf_t \theta_t(b) = -\infty$. Every finitely sampled arm has a finite logit limit by the preceding argument, so b is sampled infinitely often. As there are finitely many arms, it is enough to fix a deterministic pair (a, b) with these properties.

Conditional Borel–Cantelli gives $\sum_t \pi_t(b) = \infty$ and $\sum_t \pi_t(a) < \infty$. On the event $\{\theta_t(b) \le \theta_t(a)\}$, we have $\pi_t(b) \le \pi_t(a)$; hence

$$\sum_t \mathbb{P}(a_t = b,\ \theta_t(b) \le \theta_t(a) \mid \mathcal{F}_t) \le \sum_t \pi_t(a) < \infty.$$

Only finitely many such samples occur. Fix $\varepsilon > 0$. Choose T after the last sample of a and the last sample of b made while $\theta_t(b) \le \theta_t(a)$, and so late that $\eta C \sum_{t \ge T} \pi_t(a) < \varepsilon$. Since $\liminf_t[\theta_t(b) - \theta_t(a)] = -\infty$, choose $u \ge T$ with this margin below -2ε. Until the margin crosses zero, neither a nor b is sampled. A sample of any other arm can increase it by at most $\eta C \pi_t(a)$, so the total possible increase after u is less than ε. The margin can never cross zero, contradicting the fact that b is sampled infinitely often but eventually never while its margin is nonpositive. Thus every arm is sampled infinitely often.

Divergence of the optimal logit. Let $\Delta_* := \min_{i \ne a^*}[\mu(a^*) - \mu(i)] > 0$. From equation 47 and equation 48,

$$\mathbb{E}_t[\Delta\theta_t(a^*)] \ge \eta \Delta_* p_t(1 - p_t), \tag{50}$$

$$\mathrm{Var}_t(\Delta\theta_t(a^*)) \le \frac{\eta C^2}{\Delta_*} \mathbb{E}_t[\Delta\theta_t(a^*)]. \tag{51}$$

Exploration and equation 49 imply $\sum_t p_t = \sum_t (1 - p_t) = \infty$. If eventually $p_t \le 1/2$, then $p_t(1 - p_t) \ge p_t/2$; if eventually $p_t \ge 1/2$, then $p_t(1 - p_t) \ge (1 - p_t)/2$. Otherwise there are infinitely many upcrossings of $1/2$. The aggregate log odds

$$\log \frac{p_t}{1 - p_t} = \theta_t(a^*) - \log \sum_{i \ne a^*} e^{\theta_t(i)}$$

changes by at most $2\eta C$ in one step. An upcrossing therefore lands in the compact interval $[1/2, e^{2\eta C}/(1 + e^{2\eta C})]$, on which $p(1 - p)$ is bounded away from zero. In all three cases,

$$\sum_t p_t(1 - p_t) = \infty. \tag{52}$$

Apply Lemma 6 to the optimal coordinate from time zero. Its compensator diverges by equation 50 and equation 52, its variance is self-bounded by equation 51, and its centered increments have magnitude at most $2\eta C$. The divergence conclusion of the lemma gives $\theta_t(a^*) \to +\infty$ almost surely.

The equivalence in equation 49 now gives $\sum_t \pi_t(i) = \infty$ for every rival. All hypotheses of Lemma 7 hold, so every rival logit tends to $-\infty$ and equation 46 follows. □

Comment. The rate proof in Theorem E.1 of Appendix E in Robertson et al. (2025) is driven by the optimal-logit compensator, not by stationarity of every realized reward law. Because offset cancellation makes it exactly equal to the fixed-μ stationary bandit compensator, the same scalar mechanism can be proved directly without waiting for c_t to converge. The following argument also replaces the auxiliary deterministic recursion cited in that appendix by a direct exponential potential calculation.

Corollary 9 (Rate for tail averages under a common shift)**.** *Under Lemma 8, suppose $K \ge 2$, and define*

$$\delta := \min_{i \ne a^*}[\mu(a^*) - \mu(i)] > 0, \qquad L := \max_{i \ne a^*}[\mu(a^*) - \mu(i)], \qquad \kappa := \frac{\eta\delta}{4}.$$

On an event of probability one, there is a finite random integer τ such that, simultaneously for every integer $T > \tau$,

$$\frac{1}{T-\tau}\sum_{t=\tau}^{T-1}\left(\mu(a^*) - \sum_a \pi_t(a)\mu(a)\right) \le \frac{L}{\kappa(T-\tau)}\log(1+(K-1)(e^\kappa-1)(T-\tau))\,. \tag{53}$$

Consequently, the left-hand side is $O((1+\log(1+T-\tau))/(T-\tau))$ pathwise. For $K = 1$, it is identically zero.

Proof. The case $K = 1$ was separated in the statement, so assume $K \ge 2$. Softmax and the score update are invariant under adding the same constant to every logit. Subtracting $\theta_0(a^*)\mathbf{1}$ from the entire logit trajectory, we may therefore assume without loss of generality that $\theta_0(a^*) = 0$.

Put

$$p_t := \pi_t(a^*), \qquad \varepsilon_t := 1 - p_t, \qquad g_t := \mathbb{E}_t[\Delta\theta_t(a^*)],$$

and define the cumulative conditional drift and its martingale remainder by

$$\Xi_t := \sum_{s=0}^{t-1} g_s, \qquad M_t := \theta_t(a^*) - \Xi_t.$$

The proof of Lemma 8 gives

$$g_t \ge \eta\delta p_t\varepsilon_t, \qquad \mathrm{Var}_t(\Delta\theta_t(a^*)) \le \frac{\eta C^2}{\delta} g_t, \tag{54}$$

and equation 52 implies $\Xi_t \to \infty$. Applying the quantitative conclusion equation 15 of Lemma 6 from time zero therefore shows that, almost surely,

$$M_t \ge -\frac{1}{2}\Xi_t \qquad \text{for all sufficiently large } t. \tag{55}$$

The convergence and elimination parts of Lemma 8 prove $p_t \to 1$ and $\theta_t(i) \to -\infty$ for every $i \ne a^*$. Hence, on the same event of probability one, choose a finite random integer τ such that, for every $t \ge \tau$, equation 55 holds, $p_t \ge 1/2$, and $\theta_t(i) \le 0$ for all $i \ne a^*$. Define

$$S_t := \sum_{s=\tau}^{t-1} \varepsilon_s, \qquad t \ge \tau.$$

Since $g_s \ge 0$, equation 54 yields

$$\Xi_t \ge \sum_{s=\tau}^{t-1} g_s \ge \frac{\eta\delta}{2} S_t.$$

Using $\theta_0(a^*) = 0$ and equation 55, we obtain

$$\theta_t(a^*) = \Xi_t + M_t \ge \frac{1}{2}\Xi_t \ge \kappa S_t. \tag{56}$$

The softmax odds identity and $\theta_t(i) \le 0$ now give

$$\begin{aligned}\varepsilon_t \le \frac{\varepsilon_t}{p_t} &= \sum_{i\ne a^*} \exp\{\theta_t(i) - \theta_t(a^*)\} \\ &\le (K-1)e^{-\theta_t(a^*)} \le (K-1)e^{-\kappa S_t}.\end{aligned} \tag{57}$$

Let $U_t := e^{\kappa S_t}$. For $x \in [0,1]$, convexity gives the chord bound $e^{\kappa x} - 1 \le x(e^\kappa - 1)$. Thus

$$U_{t+1} - U_t = U_t(e^{\kappa\varepsilon_t} - 1) \le (e^\kappa - 1)U_t\varepsilon_t \le (K-1)(e^\kappa - 1).$$

Since $U_\tau = 1$, summing from τ to $T-1$ yields

$$S_T \le \frac{1}{\kappa}\log(1+(K-1)(e^\kappa-1)(T-\tau))\,. \tag{58}$$

Finally,

$$\mu(a^*) - \sum_a \pi_t(a)\mu(a) = \sum_{i \neq a^*} \pi_t(i)[\mu(a^*) - \mu(i)] \leq L\varepsilon_t.$$

Summing this inequality and applying equation 58 gives

$$\sum_{t=\tau}^{T-1} \left(\mu(a^*) - \sum_a \pi_t(a)\mu(a) \right) \leq \frac{L}{\kappa} \log(1 + (K-1)(e^\kappa - 1)(T - \tau)) .$$

Dividing by $T - \tau$ proves equation 53. The stated order follows from $1 + \alpha N \leq (1 + \alpha)(1 + N)$ for $N \geq 1$, with $\alpha := (K-1)(e^\kappa - 1)$. □

Comment. The final multi-environment bandit specialization depends only on the pooled mean vector.

Proof of Proposition 1. The common strict optimum implies

$$\bar{r}(a^*) - \bar{r}(i) = \sum_z q_z \big(r_z(a^*) - r_z(i)\big) > 0 \qquad (i \neq a^*).$$

By Lemma 5, each of the three baseline processes satisfies the assumptions of Lemma 8 with $C = 2R$ and $\mu = \bar{r}$. This proves the almost-sure convergence conclusion. Equation 9 records their common conditional mean update.

For each baseline process, Corollary 9 gives an almost surely finite random integer τ_B. Define

$$\delta_{\bar{r}} := \min_{i \neq a^*} \big[\bar{r}(a^*) - \bar{r}(i)\big] > 0, \qquad L_{\bar{r}} := \max_{i \neq a^*} \big[\bar{r}(a^*) - \bar{r}(i)\big],$$

$$\kappa_{\bar{r}} := \frac{\eta \delta_{\bar{r}}}{4}, \qquad \alpha_{\bar{r}} := (K-1)\big(e^{\kappa_{\bar{r}}} - 1\big),$$

and

$$\mathcal{E}_t^B := \bar{r}(a^*) - \sum_a \pi_t^B(a)\bar{r}(a).$$

For every integer $T > \tau_B$, putting $N := T - \tau_B$, the corollary gives

$$\frac{1}{N} \sum_{t=\tau_B}^{T-1} \mathcal{E}_t^B \leq \frac{L_{\bar{r}}}{\kappa_{\bar{r}} N} \log(1 + \alpha_{\bar{r}} N) .$$

The sum before τ_B

$$C_B^{\mathrm{pre}} := \sum_{t=0}^{\tau_B - 1} \mathcal{E}_t^B$$

is finite almost surely. Hence, for every $T > \tau_B$,

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathcal{E}_t^B \leq \frac{C_B^{\mathrm{pre}}}{T} + \frac{L_{\bar{r}}}{\kappa_{\bar{r}} T} \log(1 + \alpha_{\bar{r}}(T - \tau_B)) .$$

Set

$$T_B := \max\{\tau_B + 1, 2\}$$

and

$$C_B := \frac{C_B^{\mathrm{pre}}}{\log 2} + \frac{L_{\bar{r}}}{\kappa_{\bar{r}}} \left(1 + \frac{\log(1 + \alpha_{\bar{r}})}{\log 2} \right).$$

Both are finite almost surely. For every integer $T \geq T_B$,

$$\log(1 + \alpha_{\bar{r}}(T - \tau_B)) \leq \log(1 + \alpha_{\bar{r}}) + \log T,$$

and therefore

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathcal{E}_t^B \leq C_B \frac{\log T}{T}.$$

This proves equation 3. The entrance time may differ across baseline processes, but the same deterministic bound applies after entrance as a function of the tail length. The argument does not require the shared or conditional offset to converge. □

E Proof of Proposition 2

Proof. Let $p := \pi(a^*)$, $\pi_i := \pi(i)$, and $x_i := \theta(a^*) - \theta(i)$ for $i \neq a^*$. Finite logits and increments bounded over a single step imply that every softmax probability at finite time is strictly positive. Write $r_* = r(a^*)$, $r_{(2)} = \max_{i \neq a^*} r(i)$, $r_- = \min_a r(a)$, and define the finite threshold

$$\bar{p} := \max\left\{\frac{1}{2}, \frac{r_{(2)} - r_-}{r_* - r_-}\right\} < 1.$$

Value baseline. Suppose $p > \bar{p}$. Then $p > 1/2$, so $p > \pi_i$ for every rival, and

$$V^\pi \geq p r_* + (1-p) r_- > r_{(2)}.$$

At any finite time, softmax has full support and a^* is uniquely optimal, so also $V^\pi < r_*$. Consequently $A(a^*) = r_* - V^\pi > 0$ and $A(i) = r(i) - V^\pi < 0$ for every rival. If a^* is sampled, then, for every $i \neq a^*$,

$$\Delta x_i = \eta A(a^*)\big(1 - p + \pi_i\big) > 0. \tag{59}$$

If rival j is sampled, then

$$\Delta x_j = -\eta A(j)\big(1 + p - \pi_j\big) > 0, \qquad \Delta x_i = \eta A(j)(\pi_i - p) > 0 \quad (i \neq j). \tag{60}$$

Thus every possible branch strictly increases every margin between the optimal arm and a rival. Since

$$\pi(a^*) = \left(1 + \sum_{i \neq a^*} e^{-x_i}\right)^{-1},$$

it strictly increases as well. The set $\{p > \bar{p}\}$ is therefore invariant under future updates. Proposition 1 gives $p_t \to 1$ almost surely, so its first entrance time τ_V is almost surely finite, proving equation 4. For $K = 2$, the two displays reduce to the two familiar positive margin increments under value centering.

No baseline. Define the nonempty set of rivals with positive rewards $\mathcal{I}_+ := \{i \neq a^* : r(i) > 0\}$. Let $M := \sum_{i \neq a^*} x_i$. For a sampled arm a, direct summation of the score update gives the exact identity

$$\Delta M = K\eta r(a)\big(\mathbf{1}\{a = a^*\} - p\big). \tag{61}$$

Hence every draw $a = i \in \mathcal{I}_+$ gives $\Delta M = -K\eta r(i) p < 0$. Conditional on the history before the action, its probability is $\sum_{i \in \mathcal{I}_+} \pi_t(i) > 0$ at every finite time.

Again by Proposition 1, almost surely there is a finite time after which $p_t > 1/2$. Thereafter, if $i \in \mathcal{I}_+$ is sampled, then

$$\Delta x_i = -\eta r(i)(1 + p_t - \pi_t(i)) < 0, \qquad \Delta x_j = \eta r(i)(\pi_t(j) - p_t) < 0 \quad (j \neq i).$$

All margins between the optimal arm and a rival, and therefore p_t, strictly decrease. The exploration argument in the proof of Lemma 8 shows that softmax REINFORCE with bounded updates samples every arm infinitely often almost surely. Thus, on the intersection of the convergence and exploration events, which has probability one, every fixed $i \in \mathcal{I}_+$ produces infinitely many strict drawdowns. This proves the second claim and equation 5. □

F Proofs for Section 3.2

Lemma 10 (Population targets of shared and conditional critics)**.** *Assume $G \in L^2$. Among square-integrable critics that observe S but not Z, the minimizer of the population squared error is*

$$\bar{V}^\pi(S) := \mathbb{E}[G \mid S] = \sum_z q_z^\pi(S) V_z^\pi(S) \quad \textit{almost surely.}$$

Among critics that also observe Z, the corresponding minimizer is $V_Z^\pi(S) := \mathbb{E}[G \mid S, Z]$. Their optimal population risks differ by

$$\mathbb{E}\big[(V_Z^\pi(S) - \bar{V}^\pi(S))^2\big].$$

Values outside the support of the data distribution are unconstrained.

Proof. For any shared critic $\widehat{V}(S)$, conditional expectation gives

$$\mathbb{E}[(\widehat{V}(S) - G)^2] = \mathbb{E}[(\widehat{V}(S) - \mathbb{E}[G \mid S])^2] + \mathbb{E}[(G - \mathbb{E}[G \mid S])^2].$$

The first term is uniquely minimized almost surely by $\widehat{V}(S) = \bar{V}^{\pi}(S)$. Conditioning instead on (S, Z) gives the conditional target $V_Z^{\pi}(S)$. Finally, applying the same orthogonal decomposition to $V_Z^{\pi}(S)$ and $\bar{V}^{\pi}(S)$ gives the stated risk gap. □

Proof of Proposition 3. Write $\pi_t^s := \pi_t^{B^{\text{shared}}}$ and $\pi_t^c := \pi_t^{B^{\text{cond}}}$, with $p_t^B := \pi_t^B(a^*)$. For the two processes, define

$$A_{s,t}(z,a) := r_z(a) - \bar{V}^{\pi_t^s}, \qquad A_{c,t}(z,a) := r_z(a) - V_z^{\pi_t^c},$$

where $V_z^{\pi} := \sum_a \pi(a) r_z(a)$ and $\bar{V}^{\pi} := \sum_z q_z V_z^{\pi}$. These are exactly the centered residuals induced by the shared and conditional oracle baselines defined in Section 2.

For the conditional process, put

$$r_z^{(2)} := \max_{i \neq a^*} r_z(i), \qquad r_z^- := \min_a r_z(a), \qquad \bar{p}_c := \max_z \max\left\{\frac{1}{2}, \frac{r_z^{(2)} - r_z^-}{r_z(a^*) - r_z^-}\right\} < 1.$$

For the shared process, put

$$\bar{r}_* := \sum_z q_z r_z(a^*), \qquad \mathcal{Z}_- := \{z : r_z(a^*) < \bar{r}_*\}, \qquad q_- := \sum_{z \in \mathcal{Z}_-} q_z,$$

and abbreviate $e_z^\star := r_z(a^*) - \bar{r}_*$ and $g_z(i) := r_z(a^*) - r_z(i)$.

Conditional ratchet. Let $p := p_t^c$. If $p > \bar{p}_c$, then, simultaneously for every environment z, $p > 1/2$ and

$$V_z^{\pi_t^c} \geq p\, r_z(a^*) + (1-p) r_z^- > r_z^{(2)}.$$

Thus $A_{c,t}(z, a^*) > 0 > A_{c,t}(z, i)$ for every rival. The branch calculations in equation 59 and equation 60, applied with the rewards of the sampled environment, show that every possible combination of environment and action strictly increases every margin between the optimal arm and a rival. Hence p strictly increases and the region $\{p > \bar{p}_c\}$ is invariant under future updates. Proposition 1 gives $p_t^c \to 1$ almost surely, so the first entrance time τ_c is finite almost surely. This proves equation 6.

Limiting shared signs. Along the process with the shared baseline, Proposition 1 yields $\pi_t^s(a^*) \to 1$ almost surely. Consequently,

$$\bar{V}^{\pi_t^s} \longrightarrow \bar{r}_* := \sum_z q_z r_z(a^*).$$

For each $z \in \mathcal{Z}_-$ define $\delta_z := \bar{r}_* - r_z(a^*) > 0$. Then

$$A_{s,t}(z, a^*) = r_z(a^*) - \bar{V}^{\pi_t^s} \longrightarrow -\delta_z \qquad \text{almost surely.}$$

More generally, for every rival i,

$$A_{s,t}(z,i) = r_z(i) - \bar{V}^{\pi_t^s} \longrightarrow r_z(i) - \bar{r}_* = e_z^\star - g_z(i).$$

These limits give the signs of the optimal arm in hard environments and rival arms in easy environments discussed in Section 3.2.

Mismatch severity. Let $\bar{r}(a) := \sum_z q_z r_z(a)$ and define the suboptimality in average reward $\mathcal{E}_t^s := \bar{r}_* - \sum_a \pi_t^s(a)\bar{r}(a) = \bar{r}_* - \bar{V}^{\pi_t^s} \geq 0$. For a hard environment z, the identity $A_{s,t}(z, a^*) = \mathcal{E}_t^s - \delta_z$ is exact. Hence its update to the optimal arm flips sign precisely when $\mathcal{E}_t^s < \delta_z$, and after the flip its residual magnitude is $\delta_z - \mathcal{E}_t^s$. A larger limiting mismatch therefore moves the boundary at which the sign flips to a larger remaining error in average reward and increases the drawdown magnitude at a fixed policy. Likewise, for an easy environment let $e_z^+ := r_z(a^*) - \bar{r}_* > 0$. Every rival i has the exact shared residual $A_{s,t}(z,i) = e_z^+ - g_z(i) + \mathcal{E}_t^s$; increasing e_z^+ lifts more rivals above zero and strengthens every promoted branch. These identities order onset by policy quality $\mathcal{E}_t^s$: along the same trajectory under the shared baseline, a larger δ_z threshold is crossed no later than a smaller

one. They do not, by themselves, order hitting times across separately trained processes. For a fixed set $\mathcal{Z}_-$, severity changes onset and magnitude, not the limiting trigger frequency q_- proved below; changing the offsets can, of course, change which environments belong to $\mathcal{Z}_-$.

Persistent shared drawdowns. Fix $z \in \mathcal{Z}_-$. For every $\epsilon \in (0, \delta_z)$, there is an almost surely finite time $T_{z,\epsilon}$ such that $A_{s,t}(z, a^*) \le -(\delta_z - \epsilon)$ for all $t \ge T_{z,\epsilon}$. On a round with $(Z_t, a_t) = (z, a^*)$, the softmax policy gradient update gives, for every $i \ne a^*$,

$$\begin{aligned}
&\big[\theta^s_{t+1}(a^*) - \theta^s_{t+1}(i)\big] - \big[\theta^s_t(a^*) - \theta^s_t(i)\big] \\
&\quad = \eta A_{s,t}(z, a^*)\big[(1 - \pi^s_t(a^*)) - (-\pi^s_t(i))\big] \\
&\quad = \eta A_{s,t}(z, a^*)\big(1 - \pi^s_t(a^*) + \pi^s_t(i)\big) \\
&\quad \le -\eta(\delta_z - \epsilon)\big(1 - \pi^s_t(a^*) + \pi^s_t(i)\big) < 0.
\end{aligned}$$

The optimal probability can be written as

$$\pi^s_t(a^*) = \left(1 + \sum_{i \ne a^*} \exp\{-[\theta^s_t(a^*) - \theta^s_t(i)]\}\right)^{-1},$$

which is strictly increasing in every margin between the optimal arm and a rival. Their simultaneous strict decrease therefore implies $\pi^s_{t+1}(a^*) < \pi^s_t(a^*)$.

Let $E^z_t := \{Z_t = z, a_t = a^*\}$. With $\mathcal{F}_t$ denoting the history before the environment and arm are sampled (so $E^z_t \in \mathcal{F}_{t+1}$),

$$\mathbb{P}(E^z_t \mid \mathcal{F}_t) = q_z \pi^s_t(a^*) \longrightarrow q_z > 0 \qquad \text{almost surely.}$$

Thus $\sum_t \mathbb{P}(E^z_t \mid \mathcal{F}_t) = \infty$ almost surely, and Lévy's conditional Borel–Cantelli lemma gives $\mathbb{P}(E^z_t \text{ i.o.}) = 1$. Every sufficiently late occurrence is a strict drawdown.

For the frequency statement, let $J_t := \mathbf{1}\{Z_t \in \mathcal{Z}_-,\ a_t = a^*\}$ and $q_- := \sum_{z \in \mathcal{Z}_-} q_z$. Then

$$\mathbb{E}[J_t \mid \mathcal{F}_t] = q_- \pi^s_t(a^*) \longrightarrow q_-.$$

The differences $J_t - \mathbb{E}[J_t \mid \mathcal{F}_t]$ are bounded martingale differences, so the martingale strong law and Cesàro convergence imply

$$\frac{1}{n} \sum_{t=0}^{n-1} J_t \longrightarrow q_- \qquad \text{almost surely.}$$

Because $\mathcal{Z}_-$ is finite, the common time $T_- := \max_{z \in \mathcal{Z}_-} T_{z,\delta_z/2}$ is almost surely finite. After T_-, every event counted by J_t is a strict drawdown. Conversely, if $Z_t \notin \mathcal{Z}_-$ and $a_t = a^*$, then $A_{s,t}(Z_t, a^*) = r_{Z_t}(a^*) - \bar{r}_* + \mathcal{E}^s_t > 0$ at every finite time, so that branch increases the optimal probability. Let $I^{\downarrow}_t := \mathbf{1}\{\pi^s_{t+1}(a^*) < \pi^s_t(a^*)\}$ and $R_t := \mathbf{1}\{a_t \ne a^*\}$. For every $t \ge T_-$,

$$J_t \le I^{\downarrow}_t \le J_t + R_t.$$

The same martingale strong law gives $n^{-1} \sum_{t<n} R_t \to 0$ almost surely because $\mathbb{E}[R_t \mid \mathcal{F}_t] = 1 - \pi^s_t(a^*) \to 0$. Therefore

$$\frac{1}{n} \sum_{t=0}^{n-1} I^{\downarrow}_t \longrightarrow q_- > 0 \qquad \text{almost surely.}$$

This exact frequency statement implies the infinitely often conclusion in equation 7. □

Branch comparisons for the same sample. Fix an interior policy π, write $p := \pi(a^*)$, and define each environment value $V_z := \sum_a \pi(a) r_z(a)$. Write $\bar{V}^\pi := \sum_z q_z V_z$ for the shared value, and let E and H denote environments satisfying $V_E > \bar{V}^\pi > V_H$. Condition on the history before the update and compare the conditional and shared baselines on the same sampled tuple (Z, a, G), where $G = r_Z(a)$ in the deterministic bandit. Writing $e^\pi_Z := V_Z - \bar{V}^\pi$, the score update with one logit per arm gives the following exact identity when both baselines are evaluated at the same policy and on the same sample:

$$\Delta\theta^{\text{shared}} - \Delta\theta^{\text{cond}} = \eta e^\pi_Z (\mathbf{e}_a - \pi). \tag{62}$$

This comparison couples the two baselines for one branch; after their policies diverge, it does not equate their subsequent updates. For $x_i := \theta(a^*) - \theta(i)$, an update using baseline B gives

$$\Delta x_i(B) = \eta(G - B)\big(\mathbf{1}\{a = a^*\} - \mathbf{1}\{a = i\} - p + \pi(i)\big).$$

On the event in the easy environment where $Z = E$, $a = i \neq a^*$, and $\bar{V}^\pi < G < V_E$, this becomes

$$\Delta x_i^c = \eta(V_E - G)(1 + p - \pi(i)) > 0,$$
$$\Delta x_i^s = -\eta(G - \bar{V}^\pi)(1 + p - \pi(i)) < 0.$$

Thus the same rival sample corrects the policy under conditional centering but moves it in the wrong direction under shared centering.

On the event in the hard environment where $Z = H$, $a = a^*$, and $G > V_H$, it gives

$$\Delta x_i^c = \eta(G - V_H)(1 - p + \pi(i)) > 0,$$
$$\Delta x_i^s = \eta(G - \bar{V}^\pi)(1 - p + \pi(i)).$$

Their difference is $\eta(\bar{V}^\pi - V_H)(1 - p + \pi(i))$. If $V_H < G < \bar{V}^\pi$, the second line is negative, so shared centering reverses an otherwise helpful update to the optimal arm. If $G > \bar{V}^\pi$, both updates have the correct sign, but their magnitude ratio is $(G - \bar{V}^\pi)/(G - V_H) \in (0, 1)$, so shared centering attenuates the helpful update.

Effect of the learning rate. For a fixed sampled branch, the logit discrepancy created by replacing the conditional value with the shared value is proportional to η. The resulting policy trajectory is nonlinear, so this local fact does not imply a general ordering of returns across learning rates. Figure 8 shows the corresponding effect at the same finite budget in the instance from Figure 1.

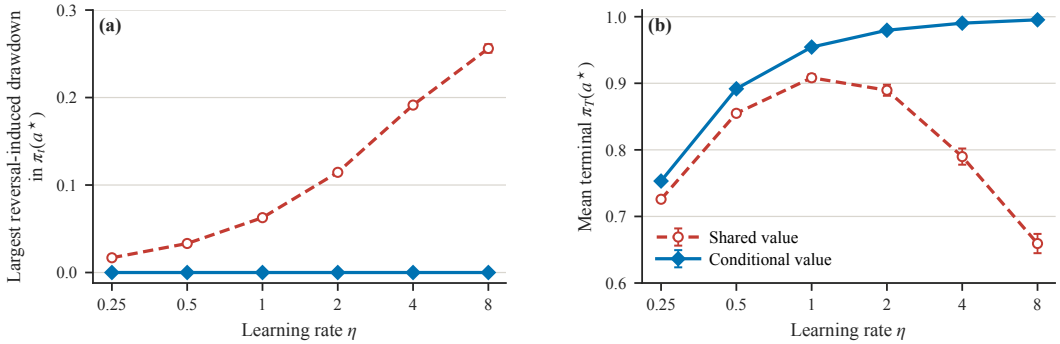

Figure 8: **At a fixed horizon, larger steps amplify the separation caused by reversed updates in the instance from Figure 1.** For each learning rate, the shared and conditional value baselines reuse the same paired random streams. **(a)** Mean across 4,000 runs of the largest decrease in a single step in $\pi_t(a^*)$ on steps where $G - \bar{V}^{\pi_t}$ and $G - V_{Z_t}^{\pi_t}$ have opposite signs; the conditional reference is zero by construction. **(b)** Mean terminal probability of the optimal arm at $T = 100$. Error bars are 95% confidence intervals. In this fixed instance, larger η magnifies reversed branches and widens the separation at this fixed budget; the plot is not a general monotonicity claim about learning rate or return.

Lemma 11 (Attenuation in a hard environment before a sign flip)**.** *Fix a deterministic bandit with $K \geq 2$ arms, unique optimal arm a^*, gaps $\Delta_i := r(a^*) - r(i) > 0$, and learning rate $\eta > 0$. There exists $\varepsilon_0 > 0$, depending only on η and the reward gaps, with the following property. Write an interior frozen policy as $p_* := \pi(a^*) = 1 - \varepsilon$ and $\pi(i) = \varepsilon w_i$, where every sum over i below ranges over $i \neq a^*$ and $\sum_i w_i = 1$. If $0 < \varepsilon < \varepsilon_0$, let $V := \sum_a \pi(a) r(a)$ and compare the conditional bar V with a shared bar $V + d$ in a hard environment, where $0 < d < r(a^*) - V$. Every sampled branch still moves probability toward a^* under both bars, and their expected logit updates are identical. Let $\pi_B^+(a^*)$ denote the probability of the optimal arm after one update using baseline B and a sampled arm $a \sim \pi$. Then*

$$\mathbb{E}_{a\sim\pi}\left[\pi_{V+d}^+(a^*)\right] < \mathbb{E}_{a\sim\pi}\left[\pi_V^+(a^*)\right].$$

Proof. Put $\bar{\Delta}_w := \sum_i w_i \Delta_i$. The conditional advantages are $A_* := r(a^*) - V = \varepsilon \bar{\Delta}_w$ and $A_i := r(i) - V = -\Delta_i + \varepsilon \bar{\Delta}_w$. Thus, after reducing ε_0 if necessary, $A_* > 0 > A_i$ for every rival and $p_* > 1/2 > \pi_i$. Since $0 < d < A_*$, raising the bar by d preserves all signs; the corresponding score directions strictly increase every optimal–rival margin. Moreover, on a sampled arm a the difference between the two logit updates is $-\eta d(\mathbf{e}_a - \pi)$, whose expectation is zero because $\sum_a \pi(a)(\mathbf{e}_a - \pi) = 0$.

It remains to compare the nonlinear probabilities. Interpolate the bar as $V + \lambda$, $0 \le \lambda \le d$, and let $H(\lambda)$ be the expected optimal probability after the update. On the optimal branch use direction $u_* = \mathbf{e}_{a^*} - \pi$; on a branch for rival i use $u_i = \pi - \mathbf{e}_i$. If $F_a(t) := [\mathrm{softmax}(\theta + t u_a)]_{a^*}$, then

$$H(\lambda) = p_* F_*(\eta(A_* - \lambda)) + \sum_i \varepsilon w_i F_i(\eta(-A_i + \lambda)) \,.$$

The exact odds between the optimal arm and each rival give, uniformly over the vector w of rival masses and $\lambda \in [0, A_*]$,

$$p_* F_*'(\eta(A_* - \lambda)) = \varepsilon^2 \left(1 + \sum_i w_i^2\right) + O(\varepsilon^3),$$
$$\sum_i \varepsilon w_i F_i'(\eta(-A_i + \lambda)) = \varepsilon^2 \sum_i w_i \left[2 w_i e^{-2\eta\Delta_i} + (1 - w_i) e^{-\eta\Delta_i}\right] + O(\varepsilon^3).$$

For completeness, these expansions follow by dividing the probability of each arm after the update by that of the optimal arm. On the optimal branch the rival-i odds are

$$\frac{\varepsilon w_i}{1 - \varepsilon} \exp\{-t\varepsilon(1 + w_i)\},$$

where $t = O(\varepsilon)$; on rival-i's branch, its own odds acquire $e^{-2\eta\Delta_i} + O(\varepsilon)$ and every other rival's odds acquire $e^{-\eta\Delta_i} + O(\varepsilon)$.

Differentiating H now yields

$$H'(\lambda) = \eta[-\varepsilon^2 \Gamma(w) + R_\varepsilon(w, \lambda)], \qquad R_\varepsilon(w, \lambda) = O(\varepsilon^3),$$

where

$$\Gamma(w) = \sum_i \left[2 w_i^2 (1 - e^{-2\eta\Delta_i}) + w_i(1 - w_i)(1 - e^{-\eta\Delta_i})\right] .$$

With $\Delta_{\min} := \min_i \Delta_i$, $\Gamma(w) \ge g_0 := 1 - e^{-\eta\Delta_{\min}} > 0$, uniformly in w. To make the remainder uniform explicit, put $\ell := \lambda/\varepsilon$; then $0 \le \ell \le \bar{\Delta}_w \le \Delta_{\max}$. The odds expressions above extend to analytic functions of (ε, w, ℓ) on a compact set, with denominators uniformly bounded away from zero. After factoring out ε^2, the remaining coefficients are uniformly continuously differentiable there. A uniform first-order Taylor bound therefore gives constants $C_0 \in (0, \infty)$ and $\varepsilon_1 > 0$, depending only on K, η, and the gaps, such that $|R_\varepsilon(w, \lambda)| \le C_0 \varepsilon^3$ for all w, $0 \le \lambda \le A_*$, and $0 < \varepsilon \le \varepsilon_1$. Taking, for example,

$$\varepsilon_0 \le \min\left\{\varepsilon_1, \frac{1}{2}, \frac{\Delta_{\min}}{2\Delta_{\max}}, \frac{g_0}{2C_0}\right\}, \qquad \Delta_{\max} := \max_i \Delta_i,$$

makes every rival advantage negative and gives $H'(\lambda) \le -(\eta g_0/2)\varepsilon^2 < 0$ throughout $[0, d]$. Integrating gives $H(d) < H(0)$. When rival mass is small, the gain from strengthening all rare branches with negative reinforcement is therefore smaller than the loss from weakening the frequent optimal branch, even before any sign reversal. □

Large ε in hard environments. The attenuation mechanism in Lemma 11 need not be confined to a nearly deterministic policy. Consider the following instance with three arms and two environments:

$$q_H = \frac{1}{20}, \quad r_H = (1, 0, 0), \qquad q_E = \frac{19}{20}, \quad r_E = (1, \tfrac{19}{20}, \tfrac{19}{20}),$$

with $\eta = 3$ and $\pi_0 = (1/2, 1/4, 1/4)$, so the initial rival mass is already $\varepsilon = 1/2$. In the hard environment, conditional centering gives optimal/rival residuals $(0.5, -0.5)$, whereas sharing gives

$(0.04875, -0.95125)$. Sharing therefore buys stronger negative reinforcement by removing most of the much larger positive update to the optimal arm.

This comparison does not rely on a sign reversal: whenever $p = \pi(a^*) \geq 1/2$, the shared optimal residual remains positive and every rival residual remains negative; each branch preserves this condition. Figure 9 shows that the conditional curve remains visibly above the shared curve over the displayed horizon. Thus the attenuation mechanism can hold when ε is bounded away from zero, not only arbitrarily close to the optimum.

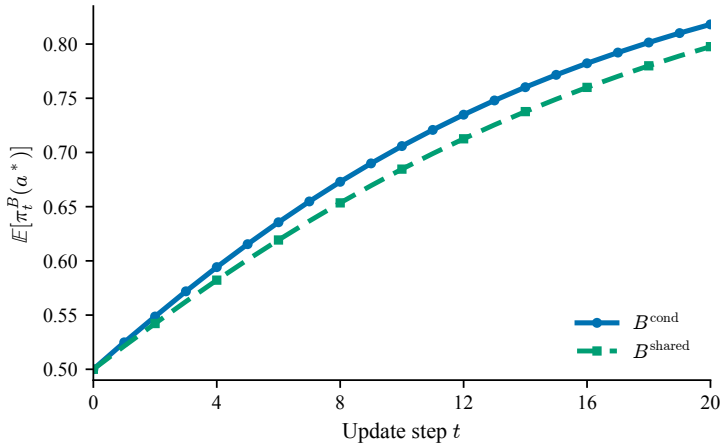

Figure 9: Separation after finitely many steps without sign reversal. Algorithm 1 with $K = 3$, $q_H = 1/20$, $q_E = 19/20$, $r_H = (1, 0, 0)$, $r_E = (1, 19/20, 19/20)$, $\eta = 3$, and $\pi_0 = (1/2, 1/4, 1/4)$. Curves are Monte Carlo means of $\pi_t^B(a^*)$ over 400,000 paired trajectories.

Derivation and closed form of the GAE mismatch recursion. Generalized Advantage Estimation (GAE) is a TD(λ) estimator built from bootstrapped temporal difference residuals (Schulman et al., 2016). Consider one rollout segment $(s_u, r_u, s_{u+1})_{u=t}^{T-1}$ from a fixed environment z, and put $n := T - t$. Let $m_u \in \{0, 1\}$ be the continuation mask after transition u: it is zero when that transition terminates the episode and one otherwise. Along this rollout, write

$$V_u^{\text{shared}} := V^{\text{shared}}(s_u), \qquad V_u^{\text{cond}} := V^{\text{cond}}(s_u, z), \qquad e_u := V_u^{\text{cond}} - V_u^{\text{shared}}.$$

For each $u \in \{t, \dots, T-1\}$ and $B \in \{\text{shared}, \text{cond}\}$, define

$$\delta_u^B = r_u + \gamma m_u V_{u+1}^B - V_u^B, \qquad \widehat{A}_u^{B,\lambda} = \sum_{\ell=0}^{T-u-1} (\gamma\lambda)^\ell M_{u,\ell} \delta_{u+\ell}^B,$$

where $M_{u,0} := 1$ and $M_{u,k} := \prod_{j=0}^{k-1} m_{u+j}$ for $k \geq 1$. Then

$$\delta_u^{\text{shared}} - \delta_u^{\text{cond}} = e_u - \gamma m_u e_{u+1}.$$

Define the GAE mismatch by $D_u^\lambda := \widehat{A}_u^{\text{shared},\lambda} - \widehat{A}_u^{\text{cond},\lambda}$ for $u < T$, with $D_T^\lambda := 0$ at the rollout boundary. The standard backward GAE recursion then gives, for $u = t, \dots, T-1$,

$$D_u^\lambda = e_u - \gamma m_u e_{u+1} + \gamma\lambda m_u D_{u+1}^\lambda, \qquad D_T^\lambda = 0. \tag{63}$$

Unrolling this recursion, or equivalently substituting into the finite sum and reindexing its second term, gives

$$D_t^\lambda = \sum_{\ell=0}^{n-1} (\gamma\lambda)^\ell M_{t,\ell} e_{t+\ell} - \sum_{k=1}^{n} \gamma(\gamma\lambda)^{k-1} M_{t,k} e_{t+k}.$$

Collecting the coefficient of each e_{t+k} yields the closed form

$$D_t^\lambda = e_t - \gamma(1-\lambda) \sum_{k=1}^{n-1} (\gamma\lambda)^{k-1} M_{t,k} e_{t+k} - \gamma(\gamma\lambda)^{n-1} M_{t,n} e_{t+n}.$$

For a single TD step ($\lambda = 0$), the recursion reduces to $D_t^0 = e_t - \gamma m_t e_{t+1}$. For $\lambda = 1$, it telescopes to $D_t^1 = e_t - \gamma^n M_{t,n} e_T$. At a genuine terminal endpoint, $m_{T-1} = 0$ and hence $M_{t,n} = 0$; at a bootstrapped nonterminal truncation, $m_{T-1} = 1$ and the final mismatch term remains. The identity compares the same rollout, rewards, evaluation states, and masks before batchwise advantage normalization and before entering PPO's clipped objective. It shows that GAE linearly filters critic mismatch and therefore does not erase it in general, although particular offset sequences can cancel. The environment index is fixed along the trajectory; if values vary over time, the stage is included in the state label.

G Full experimental details

G.1 Conditioning architectures: FiLM, multihead, and PopArt

Every main comparison changes only the value function; the diagnostic in Figure 10 that also conditions the actor is the sole exception. On the MLP benchmarks (CartPole, the MuJoCo suite, BipedalWalker) all trunks have two hidden layers of width 64 with $\tanh$. CartPole shares one trunk between actor and critic, and its reported conditional critic uses a multihead value readout. The continuous control environments use separate actor and critic trunks, and FiLM modulates the critic trunk only. Continuous control actors output a Gaussian mean from the actor trunk with a learned $\log \sigma$ independent of state (initialized to 0); hidden layers use orthogonal initialization, with gain 0.01 on the policy mean head and 1.0 on the value head. Writing $h(s) \in \mathbb{R}^{64}$ for the critic trunk features, the three MLP critics are

$$\begin{aligned} V_{\mathrm{sh}}(s) &= w^\top h(s) + b, \\ V_{\mathrm{FiLM}}(s,z) &= w^\top \big(h(s) \odot (1+\gamma_z) + \beta_z\big) + b, \qquad \gamma_z = W_\gamma c_z + u_\gamma, \;\; \beta_z = W_\beta c_z + u_\beta, \\ V_{\mathrm{mh}}(s,z) &= w_z^\top h(s) + b_z, \end{aligned}$$

where $c_z \in \mathbb{R}^d$ is a learned embedding for each level. The reference count in Table 2 uses the minimal choice $d = 1$ for CartPole; the evaluated FiLM critics use $d = 4$ on the MuJoCo bodies and $d = 16$ on BipedalWalker, while the Procgen implementation below also uses $d = 16$. The matrices $W_\gamma, W_\beta \colon \mathbb{R}^d \to \mathbb{R}^{64}$ with biases u_γ, u_β are shared generators (Perez et al., 2018). The multihead critic replaces the scalar readout by one readout row per level — the multitask value function of Hessel et al. (2019) with levels playing the role of tasks. The PopArt variant (van Hasselt et al., 2016; Hessel et al., 2019) augments the multihead critic with adaptive normalization of the value targets for each head: running first and second moments (μ_z, ν_z) track head z's target distribution, the head predicts a normalized value, $V = \sigma_z \cdot \mathrm{out} + \mu_z$, the value loss is computed in normalized space, and row z is rescaled whenever the statistics change so the denormalized output is preserved. The statistics are buffers, not parameters, so the parameter count equals that of the multihead critic.

Identical initialization. Every conditioned critic is initialized to coincide *exactly* with the shared critic at step 0: $W_\gamma, W_\beta, u_\gamma, u_\beta$ are initialized to zero, so $\gamma_z = \beta_z = 0$ and $V_{\mathrm{FiLM}}(s,z) = V_{\mathrm{sh}}(s)$; every row (w_z, b_z) of the multihead readout is initialized to the same values, so all levels return the same value. Any difference between variants is therefore produced by learning, not by a different initialization. The CartPole scalar bias control likewise starts from $\beta_0 = \beta_1 = 0$.

The actor does not see z in the proposed intervention. In every main comparison, z is an arbitrary identity label rather than a vector of physical or procedural environment parameters. It is a privileged signal available only to the critic during training; the actor is architecturally identical to the baseline's, and at deployment only the actor runs, so conditioning the critic costs nothing at test time. FiLM learns a lookup embedding for this label and multihead uses it only to select a value head; neither receives a structured descriptor from which environment dynamics could be inferred directly.

Procgen critics. On Procgen all variants share Procgen's large CNN trunk in the IMPALA style (Espeholt et al., 2018; Cobbe et al., 2020) with a final embedding $\phi(s)$ of dimension 256 and a categorical policy head over the 15 actions; the policy head always consumes the unmodulated $\phi(s)$.

The value heads are

$$V_{\mathrm{sh}}(s) = w^\top \phi(s) + b, \qquad w \in \mathbb{R}^{256},$$
$$V_{\mathrm{FiLM}}(s, z) = w^\top \big(\phi(s) + \gamma_z \odot \phi(s) + \beta_z\big) + b, \qquad c_z \in \mathbb{R}^{16},$$
$$V_{\mathrm{mh}}(s, z) = w_z^\top \phi(s) + b_z, \qquad W \in \mathbb{R}^{200\times 256}.$$

Here $\gamma_z = W_\gamma c_z + u_\gamma$ and $\beta_z = W_\beta c_z + u_\beta$. FiLM's cost for each level is the 16 embedding entries; $W_\gamma, W_\beta \colon \mathbb{R}^{16} \to \mathbb{R}^{256}$ and their biases are initialized to zero as above. The rows of the multihead critic are initialized as copies of the shared head; its cost for each level is 257 free parameters.

FiLM versus multihead. The two architectures parameterize the same object — the correction $e_z^\pi(s)$ for each level identified in Section 3.2 — through opposite statistical tradeoffs, and this is the point of running both. Each multihead readout row is an independent vector fit from *only that level's data*, whereas FiLM routes an embedding with low dimension through a generator and readout fit from *all levels jointly*. For L levels and hidden width H, FiLM adds $Ld + 2H(d+1)$ parameters, while multihead replaces one value head with $H+1$ parameters by L such heads and therefore adds $(L-1)(H+1)$ parameters relative to the shared model. With 10 levels multihead is the cheaper of the two; with the 100 levels of BipedalWalker it becomes the more expensive one ($+54.3\%$ vs. $+31.9\%$). Table 2 gives the corresponding counts for each architecture; the coverage paragraph below states which combinations were evaluated.

Table 2: Total trainable actor and critic parameter counts for the reference combinations of architecture and benchmark. Parentheses report increases relative to the shared model. FiLM retains the shared value head and adds a level embedding table and two affine generators; multihead replaces the shared value head by one head per level. The CartPole FiLM reference uses the minimal embedding dimension $d = 1$, and the Procgen counts use the large PLR CNN (Jiang et al., 2021). PopArt has the same trainable count as multihead. The table includes parameterizations not used in the main comparisons; the coverage paragraph below states which variants were evaluated.

Environment	obs / act	levels	shared	FiLM	multihead
CartPole	4 / 2	2	4,675	4,933 (+5.5%)	4,740 (+1.4%)
BipedalWalker	24 / 4	100	11,849	15,625 (+31.9%)	18,284 (+54.3%)
Walker2d	17 / 6	10	11,085	11,765 (+6.1%)	11,670 (+5.3%)
Hopper	11 / 3	10	10,119	10,799 (+6.7%)	10,704 (+5.8%)
HalfCheetah	17 / 6	10	11,085	11,765 (+6.1%)	11,670 (+5.3%)
Procgen	$64^2\times3$ / 15	200	626,256	638,160 (+1.9%)	677,399 (+8.2%)

PopArt configurations. On the MLP benchmarks the moment EMA rate is $\beta = 3 \times 10^{-3}$. On Procgen, where each estimate for a level sees roughly $1/200$ of the batch, the statistics update once per minibatch with rate $\beta = 3 \times 10^{-4}$ and the standard debiasing correction, with three stabilizers: a variance floor $\sigma_{\min} = 10^{-2}$ (PopArt with a single head uses 10^{-4}); a head's statistics update only when it has at least 8 samples in the minibatch; and normalization stays off (statistics still accumulating) for the first 50 updates — without this warmup the first update produced gradient spikes of order 10^4 from early returns.

Coverage. The experiments actually reported are as follows. The CartPole mechanism study reports a shared critic, a multihead critic, a multihead critic given a constant index, and a critic with scalar biases. The main MuJoCo return comparison reports shared and FiLM critics; the value loss diagnostic in Figure 13 additionally reports multihead. BipedalWalker additionally reports multihead, and Procgen reports shared, FiLM, multihead, and multihead+PopArt.

G.2 CartPole

CartPole (Barto et al., 1983) is the classic cart–pole balancing task: a 4-dimensional observation, two discrete actions, reward $+1$ per step while the pole stays up. The mechanism study retains two distinct logged level identities in both settings. In the identical control, $(g_0, g_1) = (10, 10)$; in the heterogeneous setting, $(g_0, g_1) = (10, 50)$. Every other parameter (force magnitude, pole mass, pole length, cart mass) remains at its default. Gravity is not part of the observation, so the two levels cannot be distinguished from the current state alone. Episodes end when the pole falls (maximum 200 steps), so episode lengths can differ between levels.

Train and evaluation use the *same* two levels: the question is the learning signal, not generalization to unseen levels. Returns are averaged within each level before averaging the two levels with equal weight.

Figure 3 in the main text illustrates the hidden dynamics difference and the learned values at a matched initial state; Figure 4 reports the controlled learning comparison.

Mechanism variants. The shared critic is $V_{\text{sh}}(s) = w^\top h(s) + b$, and the multihead critic is $V_{\text{mh}}(s, z) = w_z^\top h(s) + b_z$. The constant index control uses the same multihead parameterization but routes every sample to head 0. The scalar bias control uses

$$V_{\text{bias}}(s, z) = V_{\text{sh}}(s) + \beta_z, \qquad \beta_0 = \beta_1 = 0 \text{ at initialization.}$$

Thus the two multihead variants have 4,740 trainable parameters, the shared critic has 4,675, and the critic with a scalar bias has 4,677. The actor, shared trunk, optimizer, data, and PPO pipeline are otherwise unchanged.

Diagnostics. At every logged update and for each level, value loss is the mean squared difference between the GAE return target and the value prediction. Advantage is the mean raw GAE estimate over transitions from that level. We compute these levelwise quantities before PPO clipping. Figure 4 averages each metric across 20 seeds, with shading denoting one standard error. Table 3 lists the full configuration.

Reading the mechanism. In the heterogeneous row of Figure 4, the multihead hard-level loss initially peaks near 780 and its mean raw advantage reaches roughly -22. These landmarks show the fitting transient of the compact joint actor–critic network: each head learns from its own level while the shared representation changes. Once the hard head catches up, its advantage recenters and return approaches 200. The shared and constant-index controls remain lower and more variable; scalar biases recover only part of the gap, supporting the need for a state dependent correction. The mid-to-late upward shift on level 0 and downward shift on level 1 match the signs predicted by value mismatch.

Table 3: CartPole mechanism study hyperparameters (20 seeds per variant and setting).

Parameter	Value
Algorithm	PPO (Schulman et al., 2017)
Parallel environments	16
Rollout length	128
Total environment steps	2,048,000 (1,000 updates)
PPO epochs / minibatches	4 / 4
Clip parameter	0.2
Learning rate	3×10^{-4} (Adam, $\epsilon = 10^{-5}$)
Discount γ / GAE λ	0.99 / 0.95 (Schulman et al., 2016)
Value loss coefficient	0.5
Entropy coefficient	0.01
Max gradient norm	0.5
Advantage normalization	off (all variants)
Hidden width	64
Conditional readout	multihead with two value heads
Gravity pairs	$(10, 10)$ and $(10, 50)$
Seeds	1–20

G.3 MuJoCo suite: Walker2d, Hopper, HalfCheetah

The three standard MuJoCo locomotion bodies (Todorov et al., 2012) (Gym v4; Brockman et al., 2016) use an identical protocol so they can be compared directly:

	Walker2d-v4	Hopper-v4	HalfCheetah-v4
Observation dim	17	11	17
Action dim	6	3	6
Episode length	1000 (fixed)	1000 (fixed)	1000 (fixed)
Mass multiplier	$[0.5, 2.5]$	$[0.5, 2.5]$	$[0.5, 2.5]$

Levels. Ten levels differ in *body mass only*: every `body_mass` entry of the model is multiplied by a scale factor for each level. The factors are drawn from a distribution uniform in log space on $[0.5, 2.5]$ and sorted, using the *run seed* as the RNG seed. Consequently the shared and conditional runs of a given seed face the identical collection (the comparison is properly paired), while different seeds see different collections (the result is not an artifact of one particular grid). Level 0 is always the lightest and level 9 the heaviest; the mass is not explicitly provided in the observation.

Train / test. Reported numbers are training returns on the ten training environments, averaged within each level over the last 25% of training and then across levels with equal weight. The advantage heatmaps for individual levels in Figure 5 plot the mean sampled GAE advantage for each level, with rows ordered from lightest to heaviest level.

Reading the results. The mean curves in Figure 5 end near 2600 versus 2050 on HalfCheetah and 2450 versus 2000 on Walker2d for FiLM and shared critics, respectively. These are approximate visual landmarks, not additional estimates under a different aggregation rule. Hopper has a smaller final gap, with the conditional mean above shared through training. The accompanying heatmaps show the shared critic's positive shift on light levels and negative shift on heavy levels; conditioning largely removes this ordering.

Normalization. Both methods use running observation statistics and global reward scaling by the running standard deviation of discounted returns, as in BipedalWalker (Appendix G.4). Advantages are normalized before the PPO policy loss; reported episodic returns are raw, unscaled rewards.

Table 4: MuJoCo suite hyperparameters (30 seeds per method and body).

Parameter	Value
Algorithm	PPO
Levels	10
Environments per level	2 (20 parallel)
Rollout length	256 (5,120 transitions per update)
Total environment steps	10,000,000
PPO epochs / minibatches	10 / 8
Clip parameter	0.2
Learning rate	3×10^{-4} (Adam, $\epsilon = 10^{-5}$)
Discount γ / GAE λ	0.99 / 0.95
Value loss coefficient	0.5
Entropy coefficient	0.0
Max gradient norm	0.5
Advantage normalization	on
Hidden width	64
Level embedding dimension	4
Seeds	1–30

G.4 BipedalWalker

BipedalWalker (Brockman et al., 2016) is planar bipedal locomotion over procedurally generated terrain: a 24-dimensional lidar/proprioception observation and 4 continuous torque actions. The training collection contains 100 *pinned* terrains: 10 `BipedalWalker-v3` terrains (levels 0–9) and 90 `BipedalWalkerHardcore-v3` terrains (levels 10–99) — a collection dominated by hard terrains with stumps, pits, and stairs. Terrain is generated from the environment RNG at reset, so each environment is reseeded with its own fixed terrain seed before *every* reset and replays the same terrain for the whole run. Terrain identity is not explicitly included in the observation vector.

Train / test. Training uses the 100 pinned terrains, one parallel environment each. Testing uses 100 *unseen* terrains generated with seed offset 500 — a fresh set with the same 10/90 normal/hardcore split that never appears in training — evaluated every 1M steps, one episode per terrain, reporting average returns.

Normalization. All methods use running observation mean/std and reward scaling by the running standard deviation of discounted returns (a single global scalar, as in VecNormalize). Advantages are normalized before the PPO policy loss. Reported episode returns are always raw, unscaled rewards.

Table 5: BipedalWalker hyperparameters (10 seeds per method).

Parameter	Value
Algorithm	PPO
Parallel environments	100 (one per terrain)
Rollout length	256
Total environment steps	60,000,000
PPO epochs / minibatches	10 / 8
Clip parameter	0.2
Learning rate	3×10^{-4} (Adam, $\epsilon = 10^{-5}$)
Discount γ / GAE λ	0.99 / 0.95
Value loss coefficient	0.5
Entropy coefficient	0.0
Max gradient norm	0.5
Advantage normalization	on
Hidden width	64
Level embedding dimension	16
Evaluation interval	1,000,000 steps
Seeds	1–10

Capacity control. Figure 5(c) repeats the full comparison with each value network widened while its conditioning mechanism is held fixed, giving every critic variant roughly five times as many trainable parameters; all other settings are identical. The shared critic plateaus at the same level, and the conditioned critics again reach returns of roughly 150 and above. Increasing every critic's capacity therefore does not remove the conditioning gap in this comparison.

Other conditioning choices. Figure 10 compares two less successful ways to use the environment index. Assigning a separate value network to every terrain removes the shared value representation and learns substantially more slowly than FiLM and multihead. Supplying the index to the actor as well as the multihead critic instead collapses both training return and return on unseen terrains. Because unseen terrains carry no valid index, we evaluate this variant that conditions the actor by running the actor with each training index in turn and averaging the resulting returns. The successful intervention on this benchmark is therefore deliberately asymmetric: the actor and critic representation remain shared, while only value prediction is conditioned on the environment.

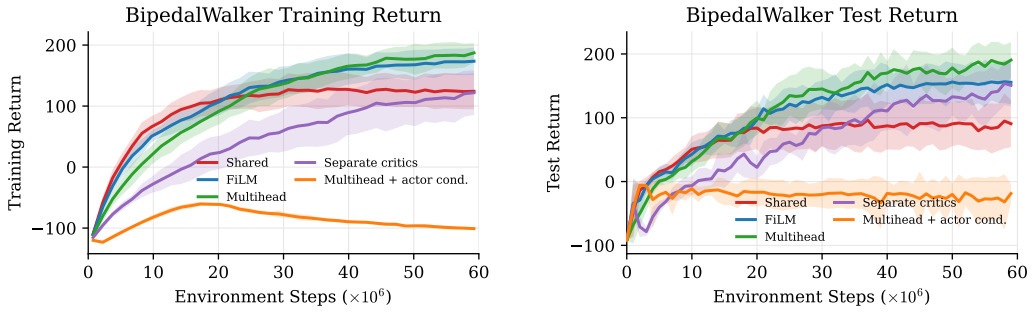

Figure 10: Alternative conditioning choices on BipedalWalker. Training return on the 100 pinned terrains (left) and return on 100 unseen terrains (right), mean $\pm$ 1 s.d. over 10 seeds. Separate critics learn more slowly but approach FiLM's final test mean; also conditioning the actor causes both returns to collapse in this setting.

G.5 PROCGEN

Procgen (Cobbe et al., 2020) is a suite of 16 procedurally generated arcade games with $64 \times 64 \times 3$ observations and a common 15-action discrete space; a level seed controls the layout, assets, entity locations and spawn times, and other details specific to each game. Each game is trained separately. Instead of sampling levels freely, every run trains on a fixed collection of $L = 200$ *pinned* levels under a hybrid difficulty split: parallel environment i is permanently assigned level

seed i, with environments 0–99 drawn from the `easy` distribution and 100–199 from `hard`. The assignment never changes during training, so level identity is exactly the environment index and statistics for each level can be logged at every update. Levels are visited with uniform weighting (one parallel copy each); we refer to this as the uniform level sampling reference. Rewards are scaled by the running standard deviation of discounted returns; observations are only scaled by $1/255$. Advantages are computed with GAE and normalized per update batch before the policy loss. Figure 11 gives a visual overview of the full benchmark.

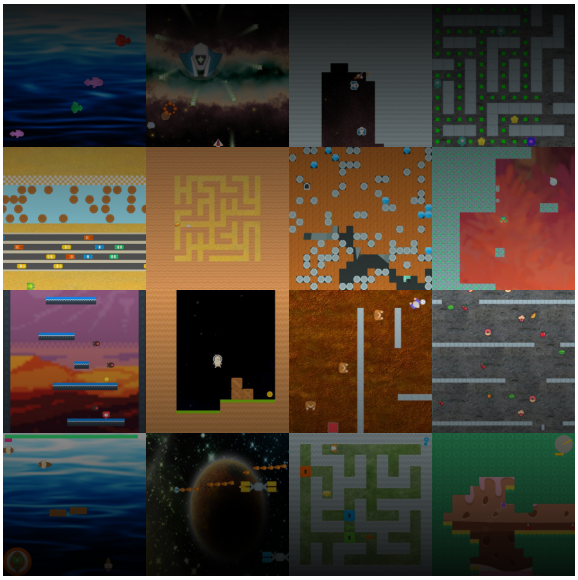

Figure 11: Example observations from all 16 Procgen games: Bigfish, Bossfight, Caveflyer, Chaser, Climber, Coinrun, Dodgeball, Fruitbot, Heist, Jumper, Leaper, Maze, Miner, Ninja, Plunder, and Starpilot. Each game generates many visually and structurally distinct levels while retaining its own mechanics. Screenshots are adapted from the official Procgen release (Cobbe et al., 2020).

Table 6: Procgen PPO hyperparameters (per game and method).

Parameter	Value
Parallel environments	200 (one per pinned level)
Rollout length	256 steps per environment
Batch size per update	51,200
PPO epochs / minibatches	3 / 8
Clip parameter	0.2
Learning rate	5×10^{-4} (Adam, $\epsilon = 10^{-5}$)
Discount γ / GAE λ	0.999 / 0.95
Entropy coefficient	0.01
Value loss coefficient	0.5
Max gradient norm	0.5
Total environment steps	25M (≈ 488 updates)
Trunk	Large CNN in the IMPALA style, $\phi(s) \in \mathbb{R}^{256}$
Level embedding dimension (FiLM)	16
Advantage normalization	on (per update batch)
Evaluation interval	5×10^5 steps
Seeds per method	10

Evaluation protocol. Every 5×10^5 environment steps, a separate, freshly reset evaluation copy of all 200 pinned levels is rolled out with the deterministic argmax policy for exactly one episode per level; the reported evaluation return is the unweighted mean over the 200 episode returns (easy/hard splits average levels 0–99 and 100–199). Because every level contributes exactly one episode, this metric weights levels equally and is unaffected by bias caused by episode length in running averages

collected under the policy. A run's final evaluation return on pinned levels is the mean of its last 4 evaluations. We use *evaluation return on the pinned training levels* for this metric; Figures 6 and 12 and Table 7 use it. The training curve reports the running mean return of the last 100 finished episodes under the stochastic behavior policy. The rows for individual games in Table 7 report raw, unscaled episodic scores. Normalized evaluation returns in its last row divide each run's final evaluation return on pinned levels by the mean final evaluation return of the shared critic under uniform level sampling over all its runs, in percent. The last row aggregates the resulting 16×10 scores for each method and reports their mean $\pm$ 1 s.d.

Evaluation on unseen levels. At the final checkpoint, we evaluate the deterministic policy for one episode on each of 600 unseen levels per game: seeds 1000–1299 under both the easy and hard modes. Levels receive equal weight. Table 1 reports mean $\pm$ 1 s.d. over the same 10 training seeds; its normalized row divides each score for a game and seed by that game's shared critic mean and aggregates the resulting 16×10 normalized scores.

Table 7: Procgen, 16 games: final evaluation return on the 200 pinned training levels (25M steps; mean $\pm$ 1 s.d. over 10 seeds per method). Per-game rows show raw scores. The last row divides each run's score by its game's shared-critic mean on those same pinned levels, multiplies by 100, and reports mean $\pm$ s.d. across the 160 game–seed entries. **Bold** = best in the row.

	Final evaluation return			
Game	shared	FiLM	multihead	+ PopArt
Bigfish	5.89 ±0.64	6.77 ±0.58	7.36 ±0.69	**7.44 ±1.86**
Bossfight	5.11 ±3.56	8.86 ±0.27	**10.13 ±0.24**	9.11 ±0.59
Caveflyer	3.11 ±0.17	3.83 ±0.21	**4.18 ±0.21**	3.03 ±0.39
Chaser	1.27 ±0.21	1.43 ±0.17	**1.96 ±0.15**	1.33 ±0.16
Climber	3.87 ±0.27	4.39 ±0.15	**5.29 ±0.18**	3.51 ±0.34
Coinrun	6.83 ±0.37	7.66 ±0.20	**7.86 ±0.10**	7.50 ±0.36
Dodgeball	3.08 ±0.20	3.42 ±0.15	**3.66 ±0.12**	3.57 ±0.25
Fruitbot	13.35 ±1.01	**13.70 ±0.32**	13.34 ±0.45	13.66 ±0.62
Heist	4.14 ±0.30	**4.27 ±0.22**	4.12 ±0.26	3.29 ±0.38
Jumper	5.44 ±0.22	**5.75 ±0.12**	5.41 ±0.10	3.88 ±1.11
Leaper	3.26 ±0.26	3.61 ±0.31	**4.08 ±0.37**	3.59 ±0.20
Maze	5.21 ±0.23	5.45 ±0.25	**5.96 ±0.15**	5.08 ±0.32
Miner	1.41 ±0.25	1.94 ±0.32	**2.11 ±0.26**	2.07 ±0.35
Ninja	5.71 ±0.30	6.67 ±0.16	**6.73 ±0.31**	6.73 ±0.68
Plunder	3.46 ±0.60	**3.67 ±0.38**	2.63 ±0.29	2.70 ±0.38
Starpilot	14.44 ±1.50	16.61 ±1.10	17.77 ±2.04	**19.16 ±1.69**
Normalized training return (%)	100.0 ±19.0	116.5 ±19.0	**124.2 ±28.3**	110.0 ±29.7

Table 8: **Procgen aggregate normalized return (%).** Within each evaluation split, each game's returns are divided by its shared-critic across-seed mean and multiplied by 100. Entries summarize 16 games $\times$ 10 seeds: mean $\pm$ descriptive s.d. across the 160 game–seed entries, not an s.e. of the aggregate mean. Full per-game results are in Tables 1 and 7.

Evaluation levels	Shared	FiLM	Multihead	Multihead + PopArt
Pinned train (200)	100.0 ± 19.0	116.5 ± 19.0	124.2 ± 28.3	110.0 ± 29.7
Unseen test (600)	100.0 ± 19.8	121.2 ± 27.9	140.8 ± 47.1	133.3 ± 41.7

Reading the evaluation curves. Figure 6 in the main text highlights eight games; Figure 12 reports all 16 games. Both evaluate the pinned training levels in separate deterministic rollouts, rather than reporting returns from stochastic training rollouts. Evaluation on unseen levels is reported in the main text's Table 1.

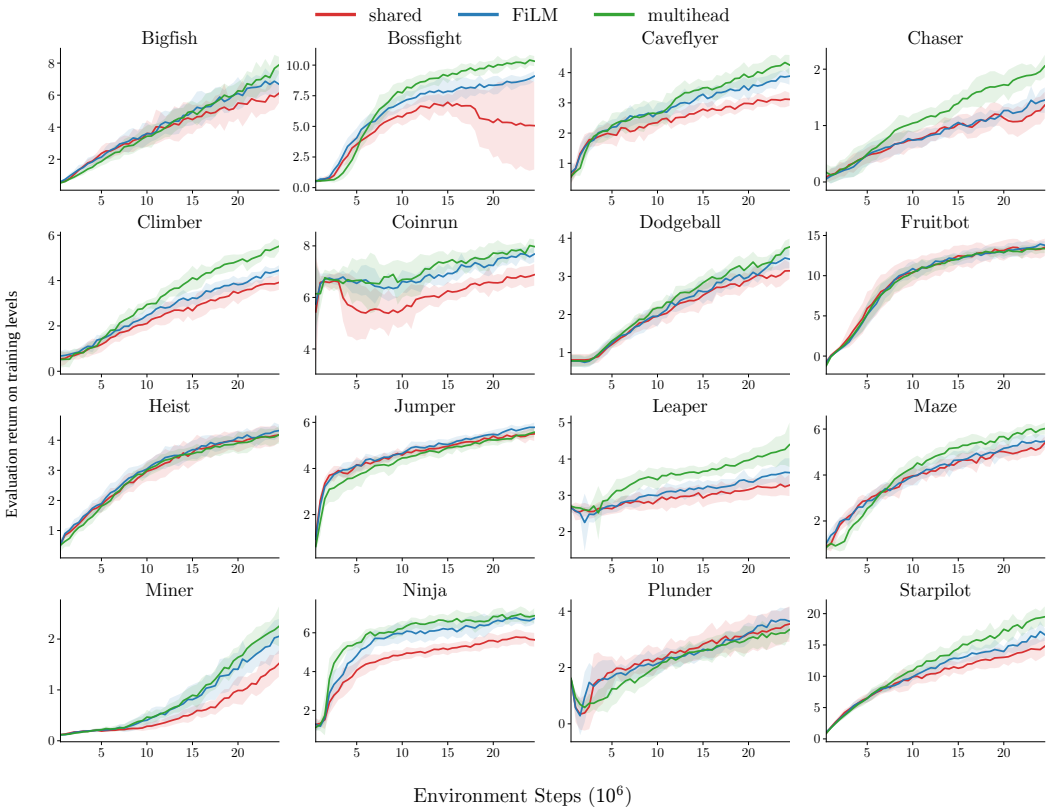

Figure 12: Complete Procgen learning curves for all 16 games. Evaluation return on the 200 pinned training levels over 25M environment steps, shown as mean $\pm$ 1 s.d. over 10 seeds for the shared, FiLM, and multihead critics. Figure 6 shows an illustrative subset of eight games.

G.6 Value loss diagnostics

We report the clipped PPO value objective throughout training to test whether critic conditioning creates a sustained fitting cost beyond the early CartPole transient. At each minibatch this is one half of the larger of the two squared errors from the unclipped and clipped value predictions, averaged over PPO epochs and minibatches. The loss is measured on each method's own training stream and therefore diagnoses the optimization burden encountered by that method. Across the larger benchmarks, conditioning does not impose a systematic penalty: the early multihead excess closes on BipedalWalker, conditioned losses are lower through most of training on all three MuJoCo tasks, and their ordering is game dependent but often favorable on Procgen.

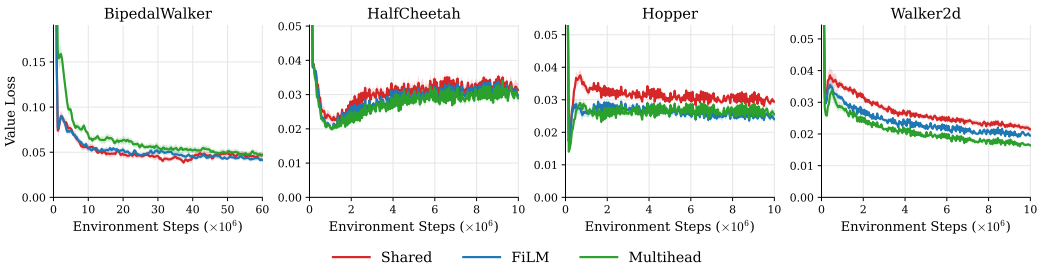

Figure 13: **Value loss diagnostics beyond CartPole.** The clipped PPO value objective on BipedalWalker and the three MuJoCo tasks; each panel uses its own vertical scale. On BipedalWalker, FiLM reaches a slightly lower loss late in training and the early multihead excess closes to a comparable level. On HalfCheetah, Hopper, and Walker2d, the conditioned critics remain below the shared critic through most of training. Curves show mean $\pm$ 1 s.d. over 10 seeds on BipedalWalker and 30 seeds for each MuJoCo task.

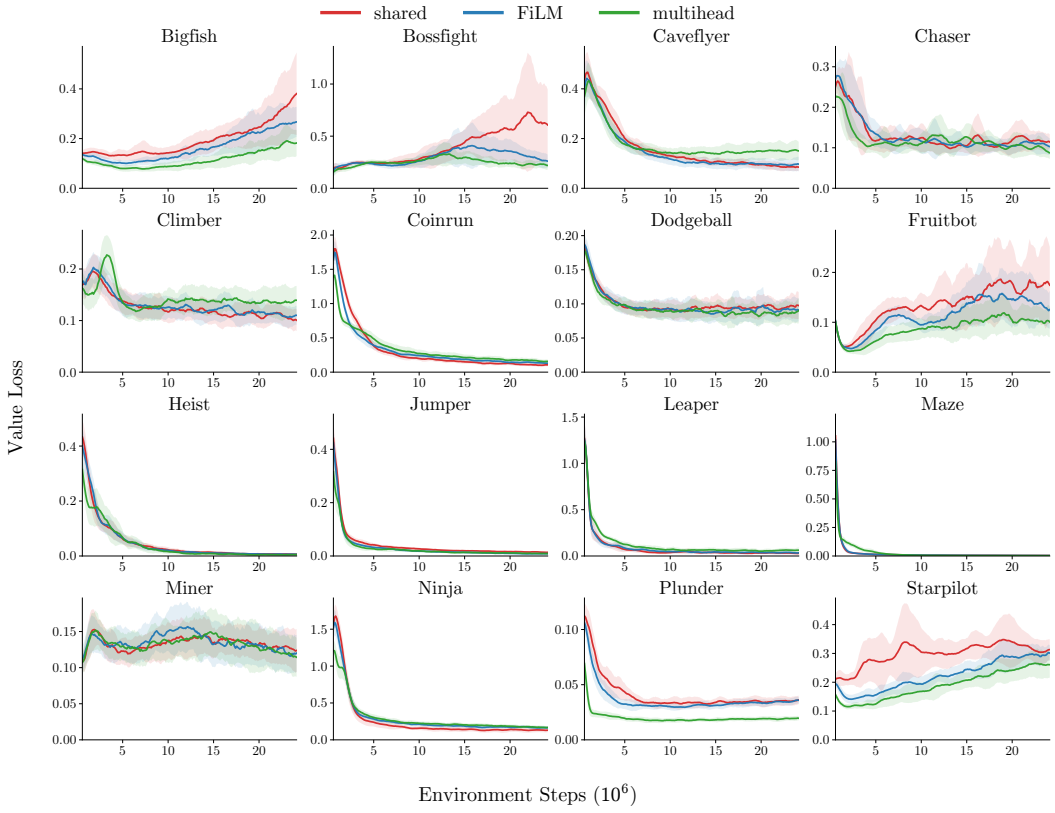

Figure 14: **Procgen value loss trajectories.** The clipped PPO value objective over 25M environment steps for all 16 games; each panel uses its own vertical scale. In late training, both the FiLM and multihead mean curves lie below the mean for the shared critic in 11 games. Curves show mean $\pm$ 1 s.d. over 10 seeds. The ordering varies by task, and conditioning does not increase the loss uniformly or create a systematic bottleneck in value fitting.

H Additional mechanism evidence and controls

This appendix complements the oracle theory with a learned-critic example, matched-observation illustrations, and fixed-rollout and capacity controls. Measured Procgen values are distinguished from the constructed BipedalWalker scene.

H.1 A MINIMAL EXAMPLE WITH LEARNED CRITICS

Figure 15 follows a shared actor across two single-state environments. As the shared critic approaches the pooled value, its mean advantages become positive in the easy environment and negative in the hard one. Fitted conditional critics remove these late-training mean sign reversals; panel E shows the resulting policy paths.

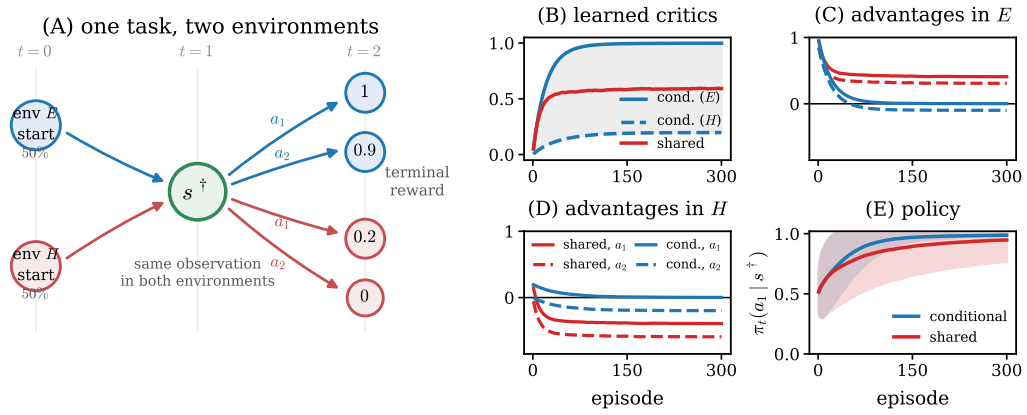

Figure 15: A minimal tabular example with *learned* critics (2,000 seeds, $\eta = 1$, critic learning rate $\beta = 0.1$, common random numbers; logits and both critics initialized at zero, so the initial policy is uniform). (A) Two environments E and H, drawn 50% each, share exactly one state $s^\dagger$; the same actions lead to endpoints worth $(1, 0.9)$ in E and $(0.2, 0)$ in H, so a_1 is optimal in both. (B) The conditional critic means track each environment; the shared critic mean approaches their average ($\to 0.6$). (C,D) Mean advantages in E and H: sharing promotes E's wrong action ($+0.3$) and suppresses H's correct action (-0.4) late in training, while the fitted conditional critics remove these mean sign reversals. (E) $\pi_t(a_1 \mid s^\dagger)$, mean ± 1 s.d.: the conditional process has a higher displayed mean after their common initialization in this example.

H.2 IDENTICAL VISIBLE INPUTS AND DIFFERENT FUTURES

Confirmed Procgen examples. Figure 16 reports two native Procgen pairs whose current RGB observations match exactly. Each policy is held fixed across the two environments, and its continuation value is estimated with 512 paired rollouts per condition at discount $\gamma = 0.99$. These are search selected existence examples, independently confirmed with fresh rollout randomness, not an estimate of how often value mismatch occurs.

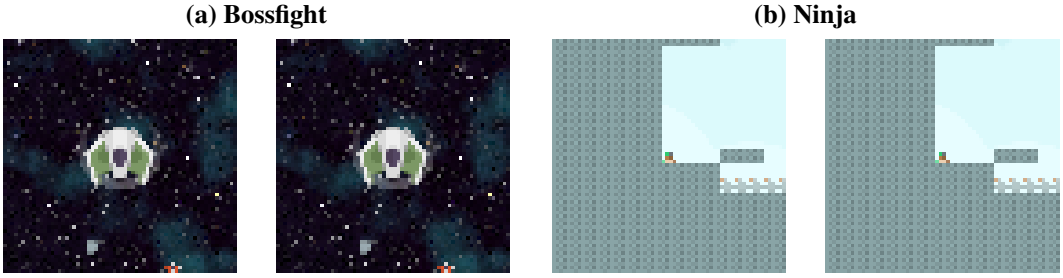

Figure 16: **Identical observations, different continuation values.** Within each pair, the initial 64×64 RGB observations are identical. Level 1/2 are display indices within each game, corresponding to Bossfight Easy/Hard at seed 181 and Ninja Hard seeds 6/9761, respectively. Values at these initial observations are estimated using 512 paired rollouts per condition with $\gamma = 0.99$ under multiple independent policies; each policy is held fixed across the two environments.

A constructed terrain illustration. Figure 17 illustrates a possible source of ambiguity in BipedalWalker. Panel (b) is the native Hardcore reset at seed 2; panel (a) removes only the original stump indicated in panel (b), leaving the robot, base ground, and all other terrain unchanged. The initial 24-dimensional observations, recast lidar fractions, and the rendered local region agree exactly.

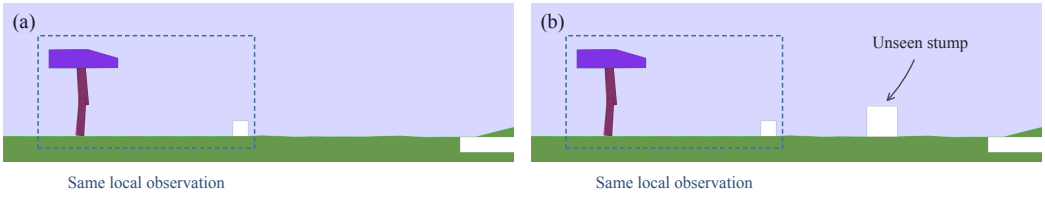

Figure 17: **Identical local observations with different distant terrain.** Panel (b) shows native BipedalWalker Hardcore terrain; panel (a) removes only the indicated stump. The robot state and current local observations are identical, because the stump lies beyond the lidar range. The different terrain can change subsequent failure risk and expected return under the same policy. These scenes illustrate a possible source of value mismatch, not measured returns.

H.3 Critic choice on a fixed rollout

We fit shared and conditional value heads on common data with the actor and encoder frozen, then recompute GAE on an independent common rollout. Procgen training and these matched updates use advantage normalization and PPO clipping. With the minibatch order and optimizer budget held fixed, the two critics yield opposite signs for $(27.1 \pm 1.8)\%$ of normalized advantages in Bossfight. After a complete PPO update, $(31.4 \pm 1.7)\%$ of sampled actions have probabilities that change in opposite directions relative to the initial actor (mean $\pm$ 1 s.d over 10 seeds). Aligning with our theories, changing the critic therefore changes both the advantages and the resulting policy update when the sampled experience is unchanged.