Drowning in Routine: Signal Dilution in Multi-Turn Agent Training

Yann Pernot [1 2] **Vi Retault** [2]

Abstract

Multi-turn agents interleave consequential decisions with routine execution: some actions change the downstream return distribution, while others are necessary but reward-equivalent. The cost of trajectory-level credit assignment, often attributed to long horizons, is in fact governed by *decision density* ρ: the fraction of turns whose actions affect the return. When decision density is low, routine turns create *signal dilution*: they add gradient variance to trajectory-level estimators such as GRPO without adding expected signal. Under explicit assumptions, the resulting turn-level to trajectory-level signal-to-noise ratio scales as $\rho^{-1/2}$, provided critic error remains controlled. The same analysis identifies the complementary regime: at high decision density, trajectory-level methods can remain competitive while avoiding the cost of a critic. In a controlled environment where ρ is exactly tunable, the predicted scaling is recovered with $R^2 = 0.999$, and the training-step gap widens significantly as $\rho \to 0$.

1. Introduction

Multi-turn reinforcement learning faces a credit-assignment problem: a single terminal reward must be allocated across many actions, only some of which actually determined the outcome (Minsky, 1961; Seo et al., 2019). The textbook remedy is to learn a value function and compute per-turn advantages, as PPO (Schulman et al., 2017) does. For LLM-scale policies, a learned critic is expensive: it doubles the model footprint and is hard to train accurately at scale (Shao et al., 2024). Critic-free trajectory-level methods such as Group Relative Policy Optimization (GRPO) (Shao et al., 2024) are therefore attractive in single-turn LLM post-training, but the same choice can become costly in multi-turn agentic settings. Recent work reports instability of direct multi-turn GRPO (Li et al., 2026) and gains from finer-grained credit assignment on multi-turn agent benchmarks (Feng et al., 2025; Wei et al., 2025). These results point to a recurring empirical gap without isolating its mechanism or the task properties that govern it. We address this by identifying a signal dilution mechanism behind the trajectory-level penalty and tying its strength to a structural quantity: the decision density.

Real multi-turn agents tend to spend most of their turns on actions that are necessary but not in themselves consequential (Zhou et al., 2024). Consider a coding agent debugging a repository (Yang et al., 2024). A handful of turns correspond to task-level commitments: deciding whether the bug is in the parser or the optimizer, choosing which file to patch, or selecting one fix rather than another. These turns significantly impact the rest of the trajectory and therefore the future return distribution. We call them *critical*. By contrast, some turns occur after such a subgoal has been fixed. If the agent has already decided to inspect the test failures, then running the full suite, running the failing test directly, or inspecting a cached log are different actions that serve the same purpose and tend to leave the downstream probability of solving the task nearly unchanged. We call them *routine*. We argue that the fraction of trajectory turns that are critical, which we call the *decision density* ρ, is a useful structural variable for reasoning about the cost of going critic-free.

Looking at the problem through this lens makes the failure mode of trajectory-level training precise. We formalize idealized routine states through distributional invariance of the episode return, and show that every routine turn injects variance into a trajectory-level gradient while contributing no signal in expectation; a disadvantage the turn-level estimator avoids by zeroing routine contributions pointwise. We call this mechanism *signal dilution*. We use signal-to-noise ratio (SNR) as a training-step efficiency proxy, following prior work that relates policy-gradient SNR to learning performance under a fixed evaluation budget (Roberts and Tedrake, 2008). Under non-degeneracy assumptions, the turn-level SNR advantage scales as $\Theta(\rho^{-1/2})$ in the perfect critic regime, yielding a theoretical upper envelope on the

[1]Mila - Québec AI Institute [2]Polytechnique Montréal. Correspondence to: Yann Pernot <yann.pernot@mila.quebec>.

Accepted at the FAGEN Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

gains a critic can provide. The analysis extends beyond the perfect critic case: turn-level gains survive only while critic-error variance stays below decision density, and saturate once it exceeds it.

We instantiate the assumptions in *Diluted Doors*, a controlled MDP whose decision density is exactly tunable; Exp. 1 quantifies the predicted gap, and Exp. 2 quantifies the resulting training-step efficiency gap. The predicted $\rho^{-1/2}$ law explains 99.9% of the variation in the empirical SNR ratio over a fifty-fold range of ρ, and the training-step gap between the two estimators widens significantly as $\rho \to 0$.

Contributions. We (i) formalize routine and critical turns and define decision density ρ; (ii) identify signal dilution as the mechanism behind the trajectory-level penalty: turn-level estimators cancel routine contributions pointwise, trajectory-level estimators only in expectation, leaving excess variance; (iii) derive the SNR scaling $\Theta(\rho^{-1/2})$ and extend it to imperfect critics; (iv) reframe trajectory-level cost as governed by decision density rather than horizon alone, with $\rho^{-1/2}$ as a theoretical upper envelope; (v) quantify in a controlled MDP both the predicted SNR gap and the training-step efficiency gap, which widens as $\rho \to 0$.

2. Preliminaries

2.1. Environment model

We model the environment the agent evolves in as a finite-horizon MDP with a terminal reward (Sutton and Barto, 2018), defined as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, T)$. $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the state transition kernel, $R : \mathcal{S}_T \to [0, 1]$ is the terminal reward, defined on states reached at horizon T, the number of turns. In this formulation, an LLM is not itself the MDP; it is the stochastic policy embedded in an interaction process whose state evolves over turns.

For notational simplicity, we write s_t for the agent-facing decision state, including all information needed by the policy.[1] The policy is assumed to be differentiable in θ for every (s, a) with $\pi_\theta(a \mid s) > 0$.

2.2. Trajectory-level and turn-level estimators

We compare two families of policy-gradient estimators (Sutton et al., 2000) that differ in how they assign credit to individual turns.

Consider a trajectory $\tau = (s_0, a_0, \ldots, s_{T-1}, a_{T-1})$.

[1] In an LLM agent, s_t can be taken to include the interaction history or rendered prompt, so the policy is written as $\pi_\theta(a_t \mid s_t)$. If the rendered history is only a partial observation of hidden environment state, the original problem is a POMDP; using the full history yields the standard history-state MDP view.

A *turn-level* policy-gradient estimator uses a state-value function V to compute a *per-turn advantage* $\hat{A}_t$ at every step (Schulman et al., 2015):

$$\hat{g}_{\text{turn}} = -\sum_{t=0}^{T-1} \hat{A}_t \, \nabla_\theta \log \pi_\theta(a_t \mid s_t), \tag{1}$$

With a perfect critic, $\hat{A}_t = A^{\pi_\theta}(s_t, a_t) = Q^{\pi_\theta}(s_t, a_t) - V^{\pi_\theta}(s_t)$, so the credit assigned to turn t directly reflects how much that action changed the expected return. PPO is an instantiation of this family (Schulman et al., 2017).

A *trajectory-level* policy-gradient estimator eliminates the value function by attaching a single scalar advantage $\hat{A}^{(i)}$ to every turn of trajectory $\tau^{(i)}$:

$$\hat{g}_{\text{traj}} = -\hat{A}^{(i)} \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t^{(i)} \mid s_t^{(i)}). \tag{2}$$

GRPO is an instantiation (Shao et al., 2024): it samples G trajectories from the same policy and uses the group-normalized return

$$\hat{A}^{(i)} = \frac{R(\tau^{(i)}) - \bar{R}}{\sigma_R + \epsilon} \tag{3}$$

where $\bar{R}$ and σ_R are the mean and standard deviation of returns within the group. Other trajectory-level estimators (e.g., REINFORCE with a baseline (Greensmith et al., 2004; Williams, 1992), RLOO (Ahmadian et al., 2024)) differ in how $\hat{A}^{(i)}$ is built from the group's returns but share the structure of Eq. (2): a single scalar advantage applied uniformly across the trajectory's scores.

The key structural contrast is that **the turn-level advantage $\hat{A}_t$ is specific to each turn**, while **the trajectory-level advantage $\hat{A}^{(i)}$ is a single scalar applied identically to every turn** of trajectory $\tau^{(i)}$.

In practice, instantiations such as PPO and GRPO optimize a clipped surrogate, and may include a KL penalty against a reference policy. We consider the simplified gradient for generality and to isolate the mechanism.

Throughout the analysis, we assume trajectory-level advantages are uniformly bounded, $|\hat{A}^{(i)}| \leq M_A < \infty$ a.s.

2.3. Gradient Signal-to-Noise Ratio

Policy-gradient estimators are stochastic as they are computed from sampled rollouts. For a fixed parameter coordinate k, the parameter-wise (coordinate-wise) SNR compares the expected update $\mathbb{E}[\hat{g}_k]$ (the signal) with the standard deviation of the sampled update $\hat{g}_k$ (the noise).

This is the square root of the parameter-wise GSNR used by Liu et al. (2020), and a coordinate-wise analogue of

the policy-gradient SNR studied by Roberts and Tedrake (2008), who showed empirically that SNR tracks learning performance under a fixed evaluation budget.

Definition 2.1 (Parameter-wise gradient SNR)**.** *For a gradient estimator $\hat{g}$ with $0 < \mathrm{Var}[\hat{g}_k] < \infty$, the signal-to-noise ratio of parameter k is*

$$\mathrm{SNR}_k(\hat{g}) \;=\; \frac{|\mathbb{E}[\hat{g}_k]|}{\sqrt{\mathrm{Var}[\hat{g}_k]}}.$$

The SNR is invariant to the sign of $\hat{g}$, so we omit the leading $-$ from Eqs. (1)–(2) hereafter.

The parameter-wise SNR is the object of the theoretical analysis. To summarize full-gradient estimates in experiments, we also use a pooled SNR, defined as the mean-gradient norm divided by the aggregate coordinate standard deviation:

Definition 2.2 (Pooled gradient SNR)**.** *For a gradient estimator $\hat{g}$ with $0 < \sum_k \mathrm{Var}[\hat{g}_k] < \infty$, the pooled (full-gradient) signal-to-noise ratio is*

$$\mathrm{SNR}(\hat{g}) \;=\; \frac{\|\mathbb{E}[\hat{g}]\|}{\sqrt{\sum_k \mathrm{Var}[\hat{g}_k]}}.$$

3. The Signal Dilution Problem

3.1. Critical States, Routine States, and Decision Density

In real trajectories, most turns offer action choices that are interchangeable with respect to the final return; routine states formalize that. [2]

Definition 3.1 (Routine state)**.** *Given a policy π, let $\mathcal{U}_\pi(s) = \{a \in \mathcal{A} : \pi(a \mid s) > 0\}$ denote its effective action support in state s. A state $s \in \mathcal{S}$ is* routine *(under π) if the conditional distribution*[3] *of the episode return is the same for every action in this support:*

$$\begin{aligned}&\mathcal{L}(R(\tau) \mid s_t = s,\, a_t = a,\, \pi)\\ &\quad = \mathcal{L}(R(\tau) \mid s_t = s,\, a_t = a',\, \pi)\\ &\qquad \forall a, a' \in \mathcal{U}_\pi(s).\end{aligned}$$

A state $s \in \mathcal{S}$ is critical *(under π) if it is not routine. We denote by $\mathcal{S}_R \subseteq \mathcal{S}$ the set of routine states under π and by $\mathcal{S}_C = \mathcal{S} \setminus \mathcal{S}_R$ the set of critical states under π.*

Definition 3.1 describes an idealization: in real environments, exact distributional invariance is rare and routineness lies on a spectrum. We adopt the exact partition because it admits clean statements and enables a controlled test of the prediction; we expect the scaling to hold approximately when states are approximately routine, and return to this point in the discussion.

Decision density summarizes how concentrated the learning-relevant decisions are along the trajectory.

Definition 3.2 (Decision density)**.** *Let $C^{(i)}$ denote the number of critical states visited by trajectory i, $\bar{C} = \mathbb{E}[C^{(i)}]$ the expected count. The* decision density *(under π) is $\rho = \bar{C}/T$.*

When ρ is high, most turns can affect the return; when ρ is low, many turns are routine and can dilute a trajectory-level learning signal.

3.2. Critical/Routine Gradient Decomposition

Let $u_t^{(i)} = \nabla_\theta \log \pi_\theta(a_t^{(i)} \mid s_t^{(i)})$ denote the score function at turn t of trajectory i. The trajectory-level gradient splits directly into critical-state and routine-state terms:

$$\hat{g}_{\mathrm{traj}} = \underbrace{\hat{A}^{(i)} \sum_{t:\, s_t^{(i)} \in \mathcal{S}_C} u_t^{(i)}}_{\hat{g}_C \text{ (critical)}} + \underbrace{\hat{A}^{(i)} \sum_{t:\, s_t^{(i)} \in \mathcal{S}_R} u_t^{(i)}}_{\hat{g}_R \text{ (routine)}}. \tag{4}$$

The critical term carries the learning signal. The routine term is the source of signal dilution: as we show next, it has zero expectation but nonzero variance for trajectory-level estimators.

3.3. Routine Signal Vanishing

The following proposition establishes that, for both estimators, the expected gradient receives contributions only from critical states.

We first establish a lemma that is the engine of the proof. The expectation throughout is taken over the G trajectories $\tau^{(1)}, \ldots, \tau^{(G)}$ sampled independently from π_θ.

Lemma 3.1 (Vanishing score expectation at routine states)**.** *Let $F^{(i)} = \psi_i(R^{(1)}, \ldots, R^{(G)})$ be a return-only trajectory scalar, with ψ_i and $F^{(i)} u_t^{(i)}$ integrable. Assume that the policy support is locally constant in θ.*

Then,

$$\mathbb{E}[F^{(i)} u_t^{(i)} \mid s_t^{(i)} = s] = \mathbf{0} \quad \forall s \in \mathcal{S}_R.$$

Proof. We prove the Lemma in the discrete action space setting.

Fix a rollout i, a turn t, and a state $s \in \mathcal{S}_R$. Write $\tau^{(-i)} = \{\tau^{(j)}\}_{j \neq i}$.

By the law of iterated expectations, $F^{(i)} u_t^{(i)}$ being inte-

[2] Distributional, rather than just expected, invariance is needed because trajectory-level estimators apply $\hat{A}^{(i)}$, a generally non-linear function of the realized return, to every score; equal expected returns across actions do not imply equal $\mathbb{E}[\hat{A} \mid s, a]$ when $\hat{A}$ is non-linear in R.

[3] $\mathcal{L}(X \mid \cdot)$ denotes the conditional distribution (law) of X.

grable,

$$\mathbb{E}\Big[F^{(i)}u_t^{(i)} \mid s_t^{(i)} = s\Big] = \mathbb{E}_{\tau^{(-i)}|s_t^{(i)}=s}\Big[\mathbb{E}\Big[F^{(i)}u_t^{(i)} \;\Big|\; s_t^{(i)} = s, \tau^{(-i)}\Big]\Big].$$

For fixed $\tau^{(-i)}$, the returns $R^{(j)}$ for $j \neq i$ are fixed. Define

$$\psi_{i,\tau^{(-i)}}(r) = \psi_i\Big(R^{(1)}, \ldots, R^{(i-1)}, r, R^{(i+1)}, \ldots, R^{(G)}\Big).$$

For any supported action a, set

$$m_a := \mathbb{E}\Big[\psi_{i,\tau^{(-i)}}(R^{(i)}) \;\Big|\; s_t^{(i)} = s, a_t^{(i)} = a, \tau^{(-i)}\Big].$$

Since s is routine, and $\tau^{(-i)}$ is independent of rollout i, Definition 3.1 gives

$$\mathcal{L}\Big(R^{(i)} \mid s_t^{(i)} = s, a_t^{(i)} = a, \tau^{(-i)}\Big) = \mathcal{L}\Big(R^{(i)} \mid s_t^{(i)} = s, a_t^{(i)} = a', \tau^{(-i)}\Big) \quad \forall a, a' \in \mathcal{U}_\pi(s)$$

Therefore, by integrability of ψ_i,

$$m_a = m_{a'} = c(s, \tau^{(-i)}) \quad \forall a, a' \in \mathcal{U}_\pi(s)$$

Then

$$\begin{aligned}
&\mathbb{E}\Big[F^{(i)}u_t^{(i)} \;\Big|\; s_t^{(i)} = s, \tau^{(-i)}\Big] \\
&= \sum_{a \in \mathcal{U}_\pi(s)} \Pr(a_t^{(i)} = a \mid s_t^{(i)} = s, \tau^{(-i)}) \\
&\quad \times \mathbb{E}\Big[F^{(i)}u_t^{(i)} \;\Big|\; s_t^{(i)} = s, a_t^{(i)} = a, \tau^{(-i)}\Big] \\
&= \sum_{a \in \mathcal{U}_\pi(s)} \pi_\theta(a \mid s) \nabla_\theta \log \pi_\theta(a \mid s) \\
&\quad \times \mathbb{E}\Big[F^{(i)} \;\Big|\; s_t^{(i)} = s, a_t^{(i)} = a, \tau^{(-i)}\Big] \\
&\quad (u_t^{(i)} \text{ fixed given } s, a) \\
&= \sum_{a \in \mathcal{U}_\pi(s)} \pi_\theta(a \mid s) \nabla_\theta \log \pi_\theta(a \mid s)\, m_a \\
&= c(s, \tau^{(-i)}) \sum_{a \in \mathcal{U}_\pi(s)} \nabla_\theta \pi_\theta(a \mid s) \\
&= c(s, \tau^{(-i)}) \nabla_\theta \sum_{a \in \mathcal{U}_\pi(s)} \pi_\theta(a \mid s) \\
&= \mathbf{0}. \\
&\quad \text{(locally constant support)}
\end{aligned}$$

Taking the outer expectation gives the claim. □

Lemma 3.1 can be directly used for the following proposition.

Proposition 3.1 (Routine Signal Vanishing)**.** *Under the assumptions of Lemma 3.1 (trajectory-level only), the expected gradient of both estimators receives contributions only from critical states:*

(i) ***Trajectory-level estimator.*** *The gradient contribution at routine states vanishes in expectation:*

$$\mathbb{E}\Big[\hat{A}^{(i)} \cdot u_t^{(i)} \;\Big|\; s_t^{(i)} = s\Big] = \mathbf{0} \quad \forall s \in \mathcal{S}_R,$$

$$\textit{so} \qquad \mathbb{E}[\hat{g}_{\text{traj}}] = -\sum_{t=0}^{T-1} \mathbb{E}\Big[\hat{A} \cdot u_t\, \mathbf{1}\{s_t \in \mathcal{S}_C\}\Big].$$

(ii) ***Turn-level estimator (perfect critic, $\hat{A}_t = A^{\pi_\theta}(s_t, a_t)$).*** *The gradient contribution at routine states vanishes deterministically:*

$$\hat{A}_t \cdot u_t = \mathbf{0} \quad \textit{a.s. for all } t \textit{ with } s_t \in \mathcal{S}_R,$$

$$\textit{so} \qquad \mathbb{E}[\hat{g}_{\text{turn}}] = -\sum_{t=0}^{T-1} \mathbb{E}\Big[\hat{A}_t \cdot u_t\, \mathbf{1}\{s_t \in \mathcal{S}_C\}\Big].$$

Proof. Part (i). Lemma 3.1 gives $\mathbb{E}[\hat{A}^{(i)}u_t^{(i)}] = \mathbf{0}$ at every routine step. The claim on $\mathbb{E}[\hat{g}_{\text{traj}}]$ follows by linearity of expectation.

Part (ii). By Definition 3.1, for $s_t \in \mathcal{S}_R$, the expected return is the same for every $a \in \mathcal{U}_\pi(s_t)$. Therefore $Q^{\pi_\theta}(s_t, a) = V^{\pi_\theta}(s_t)$ for all supported actions, so $A^{\pi_\theta}(s_t, a) = 0$. The claim on $\mathbb{E}[\hat{g}_{\text{turn}}]$ again follows by linearity of expectation. □

Interpretation. Proposition 3.1 pins down the **source of signal dilution**: for both estimators the expected gradient comes entirely from critical states, but the two families cancel routine contributions differently. The turn-level estimator zeros them *pointwise* ($\hat{A}_t u_t = \mathbf{0}$ at every routine step), while the trajectory-level estimator zeros them only *in expectation* ($\mathbb{E}[\hat{A}\, u_t] = \mathbf{0}$, but $\hat{A}\, u_t \neq \mathbf{0}$ on individual trajectories). The residual per-trajectory routine noise is what produces the trajectory-level estimator's excess variance, and it leads to a structural gap (Section 3.4).

3.4. Structural SNR Scaling

We now quantify the cost of the variance gap identified in Proposition 3.1. Throughout the section, we consider the fixed-$\bar{C}$ sweep $T = \bar{C}/\rho \to \infty$ (the number of critical turns is held fixed; routine turns are added). Under Assumptions 3.1–3.2 and a perfect critic, the turn-level to trajectory-level SNR ratio admits a closed-form scaling in the decision density ρ, which diverges as $\rho \to 0$.

The first proposition is coordinate-wise. Fix a parameter coordinate k and write

$$\Phi_{C,k} = \sum_{t:\, s_t \in \mathcal{S}_C} u_{t,k}, \quad \Phi_{R,k} = \sum_{t:\, s_t \in \mathcal{S}_R} u_{t,k}.$$

$$\Phi_k = \Phi_{C,k} + \Phi_{R,k}.$$

The result rests on two assumptions: a regularity condition keeping the critical sub-problem well-posed as T grows (A1), and a substantive non-degeneracy condition on routine exploration (A2).

Assumption 3.1 (A1 – Stable critical sub-problem)**.** *The critical-state contributions to the expected gradient are bounded: there exist constants $0 < m_\mu \le M_\mu < \infty$ such that*

$$m_\mu \;\le\; |\mathbb{E}[\hat{g}_{\mathrm{turn},C,k}]|,\; |\mathbb{E}[\hat{g}_{\mathrm{traj},C,k}]| \;\le\; M_\mu,$$

and the corresponding critical variances are bounded: there exists $0 < m_C \le M_C < \infty$ with

$$m_C \le \mathrm{Var}[\hat{g}_{\mathrm{turn},C,k}] \le M_C, \qquad \mathrm{Var}[\hat{A}\,\Phi_{C,k}] \le M_C.$$

A1 fixes the scale of the critical problem as routine turns are added: the reward-relevant signal neither vanishes nor blows up, and critical variance stays bounded for both estimators. The underlying critical task does not change as we scale T.

Assumption 3.2 (A2 – Non-degenerate routine exploration and score accumulation)**.** *Let N_R be the routine-state count of a trajectory. Let $\mathcal{H}_C$ be the sigma-field generated by N_R, the critical trajectory information, and the group's terminal returns, and assume $\hat{A}$ is $\mathcal{H}_C$-measurable. There exist constants $c_N, c_A, c_R > 0$ and $C_R < \infty$ such that*

(a) $\mathbb{E}[N_R] \ge c_N T$;

(b) $\mathbb{E}[\hat{A}^2 N_R] \ge c_A\, \mathbb{E}[N_R]$;

(c) $c_R\, N_R \;\le\; \mathbb{E}[\Phi_{R,k}^2 \mid \mathcal{H}_C] \;\le\; C_R\, N_R$ *a.s.*

A2 imposes three conditions: (a) the routine-state count grows linearly in T; (b) the trajectory advantage doesn't collapse on routine-heavy trajectories; (c) conditional routine score energy grows proportionally to the routine-state count. (a) is bookkeeping. The upper half of (c) rules out coherent long-range alignment of routine scores and rarely fails outside adversarial constructions. The load-bearing conditions are (b) and the lower half of (c). (b) is specific to group-relative advantages: it fails when all rollouts in a group share the same return, a known GRPO failure mode tied to task difficulty and group size (Le et al., 2026). The lower half of (c) is the broader concern: it fails when the policy becomes nearly deterministic on routine turns, which entropy collapse pushes toward in late training (Cui et al., 2025).

Lemma 3.2 (Critical/routine bounds at coordinate k)**.** *Assume a perfect critic and the standing bounded-advantage condition, and fix a coordinate k.*

(i) Under the upper-bound parts of A1 and the upper half of A2(c),

$$|\mathbb{E}[\hat{g}_{\mathrm{turn},k}]|,\; |\mathbb{E}[\hat{g}_{\mathrm{traj},k}]| = O(1),$$
$$\mathrm{Var}[\hat{g}_{\mathrm{turn},k}] = O(1), \quad \mathrm{Var}[\hat{g}_{\mathrm{traj},k}] = O(T).$$

(ii) If, in addition, the lower-bound parts of A1, A2(a), A2(b), and the lower half of A2(c) hold at k, the matching $\Omega(\cdot)$ bounds hold.

Proof in Appendix A.

Proposition 3.2 (Structural SNR scaling)**.** *Under the assumptions of Lemma 3.2(ii),*

$$\frac{\mathrm{SNR}_k(\hat{g}_{\mathrm{turn}})}{\mathrm{SNR}_k(\hat{g}_{\mathrm{traj}})} \;=\; \Theta(\rho^{-1/2}).$$

Proof in Appendix A.

Sketch. Both expected gradients receive contributions only from critical states (Prop. 3.1) and are bounded above and below by A1. The turn-level estimator zeros routine terms pointwise, so its variance stays $\Theta(1)$. The trajectory-level estimator instead pays $\hat{A}$ times the accumulated routine score, whose variance is $\Theta(T)$ under A2. Combining gives $\mathrm{SNR}_{\mathrm{turn}}/\mathrm{SNR}_{\mathrm{traj}} = \Theta(\rho^{-1/2})$.

Upper-envelope reading. The substantive parts of A2, A2(b) and the lower half of A2(c), enter only on the $\Omega(\rho^{-1/2})$ side. When they fail, $\mathrm{Var}[\hat{g}_{\mathrm{traj}}]$ grows slower than T and the realized ratio falls below $\rho^{-1/2}$, while the $O(\rho^{-1/2})$ upper bound still holds. Two further idealizations point the same way. Exact routine states let turn-level zero routine-turn contributions pointwise; approximate ones leave residual variance. A perfect critic avoids the routine-turn variance that critic error would otherwise reintroduce into turn-level (cf. Prop. 3.3). The $\rho^{-1/2}$ scaling can therefore be read as an upper envelope on the SNR gap achievable in real agentic tasks.

Decision density, not horizon. The cost of trajectory-level credit assignment relative to turn-level is often attributed to horizon length. The variance gap between estimators is generated by routine turns: by Proposition 3.1, routine turns add variance to the trajectory-level estimator only, while the turn-level estimator is unaffected. Trajectory length matters only insofar as it controls the routine-turn count.

The ratio diverges as $\rho \to 0$. The pooled SNR inherits the same scaling under weaker assumptions: the lower bounds need hold at only one coordinate.

Corollary 3.1 (Pooled SNR scaling)**.** *Assume a perfect critic, fixed parameter dimension d, and that A2(a)–(b) hold. Suppose the upper-bound parts of A1 and the upper half of A2(c) hold uniformly at every coordinate $k = 1, \dots, d$, and that the lower-bound parts of A1 together with the lower half of A2(c) hold at some coordinate $k^\star$. Then*

$$\frac{\mathrm{SNR}(\hat{g}_{\mathrm{turn}})}{\mathrm{SNR}(\hat{g}_{\mathrm{traj}})} = \Theta(\rho^{-1/2}).$$

Proof in Appendix A.1.

3.5. Imperfect Critic Regime

The previous scaling describes the best-case turn-level estimator, where a perfect critic removes routine-state contributions pointwise. We now ask what remains when the turn-level advantage is noisy. Write the imperfect critic estimate as $\widetilde{A}_t = A^{\pi_\theta}(s_t, a_t) + \epsilon_t$, and define

$$\hat{g}_{\mathrm{turn},\epsilon} = \sum_{t=0}^{T-1} \widetilde{A}_t u_t.$$

Equivalently,

$$\hat{g}_{\mathrm{turn},\epsilon} = \underbrace{\sum_{t=0}^{T-1} A^{\pi_\theta}(s_t, a_t) u_t}_{\hat{g}^\star_{\mathrm{turn}}} + \underbrace{\sum_{t=0}^{T-1} \epsilon_t u_t}_{Z_\epsilon},$$

where $\hat{g}^\star_{\mathrm{turn}}$ is the perfect critic turn-level estimator and Z_ϵ is the critic-error contribution.

We make the assumption directly on the gradient contribution of the critic error.

Assumption 3.3 (A3 – Gradient-unbiased critic error)**.** *Let $\mathcal{F}_\tau$ be the sigma-field generated by the sampled trajectory. The critic-error contribution is conditionally mean-zero,*

$$\mathbb{E}[Z_\epsilon \mid \mathcal{F}_\tau] = \mathbf{0},$$

and its coordinate-wise critic-noise level

$$\sigma^2_{\epsilon,k}(T) := \frac{1}{T}\,\mathbb{E}[\mathrm{Var}(Z_{\epsilon,k} \mid \mathcal{F}_\tau)]$$

is finite for every k.

Assumption 3.3 models imperfect critics through a variance-only error channel. The finite-variance condition is mild: it requires the critic-error contribution to grow at most linearly with horizon. The conditional mean-zero condition is the substantive part: it ensures that critic error preserves the expected policy-gradient signal and enters only as additional stochastic gradient noise around the perfect-critic update. A deterministic learned critic may also introduce systematic approximation bias; such bias is outside Proposition 3.3, and can misdirect the update rather than lowering its SNR.

Proposition 3.3 (Imperfect critic SNR scaling)**.** *Fix a coordinate k satisfying the standing bounded-advantage condition and Assumptions 3.1–3.3. Then*

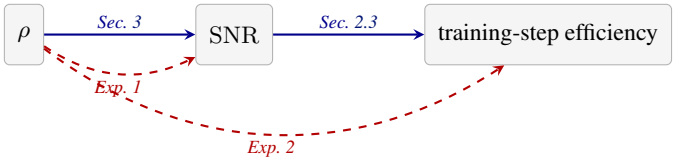
$$\frac{\mathrm{SNR}_k(\hat{g}_{\mathrm{turn},\epsilon})}{\mathrm{SNR}_k(\hat{g}_{\mathrm{traj}})} = \Theta\left(\frac{1}{\sqrt{\rho + \sigma^2_{\epsilon,k}(T)}}\right),$$

with constants depending on $\bar{C}$ and the constants in Assumptions 3.1–3.2. Proof in Appendix A.2.

Interpretation. Critic noise reintroduces routine-turn variance into the turn-level estimator: at routine states, $\widetilde{A}_t u_t = \epsilon_t u_t$ is mean-zero but no longer pointwise zero. The turn-level gain therefore depends on which dominates, ρ or $\sigma^2_{\epsilon,k}$. When $\sigma^2_{\epsilon,k} \ll \rho$, the perfect-critic scaling $\Theta(\rho^{-1/2})$ is recovered; when $\sigma^2_{\epsilon,k} \gg \rho$, the ratio flattens to a horizontal asymptote $\Theta(\sigma^{-1}_{\epsilon,k})$ set by the critic-error scale, independent of ρ. Low decision density makes turn-level credit assignment potentially valuable, but **only when critic error stays below the decision-density scale**.

4. Experimental Setup

Our theory predicts that ρ governs the gradient SNR ratio between turn-level and trajectory-level estimators (Section 3); prior work relates policy-gradient SNR to learning performance under fixed evaluation budgets (Roberts and Tedrake, 2008). Together they motivate two empirical predictions, summarized below.



We test these mechanism-level predictions in *Diluted Doors*, a deterministic MDP where ρ is set exactly and routine transitions are reward-equivalent by construction. Experiment 1 quantifies the gradient SNR ratio at initialization; Experiment 2 quantifies the resulting training-step efficiency gap as a function of ρ, with the SNR-to-training-step efficiency link taken from Roberts and Tedrake (2008). The policy is a small causal transformer (Vaswani et al., 2017); full details are in Appendix B.

Environment design. Diluted Doors is a balanced tree alternating C critical states (only one of K_C doors is correct, fixed across episodes) and L routine states between consecutive critical states. Every root-to-leaf trajectory has length $T = C(1 + L)$ and visits exactly C critical states, so $\rho = 1/(1 + L)$ is exact; varying L at fixed C sweeps ρ without changing the number of critical decisions. At each routine state, the K_R outgoing paths lead to isomorphic

subtrees, so the distribution of future returns is identical across paths and Definition 3.1 holds exactly. Any effect on the SNR ratio is therefore attributable to the routine-state mechanism rather than to hidden signal at routine states.

Experiment 1 — quantifying the predicted SNR gap. The two gradient estimators differ only in the advantage term: *trajectory-level* uses a group-relative trajectory scalar (GRPO-like), and *turn-level* uses a per-turn critic-based advantage; at initialization, V^{π_θ} admits a closed-form expression under the near-uniform policy, so the turn-level estimator uses this analytical critic. For each L, both estimators are evaluated on the same independently sampled batches across seeds, and the Bessel-corrected pooled SNR is estimated from these batch-gradient samples (Appendix B.6). We report the paired ratio $R_{\mathrm{SNR}}(\rho) = \mathrm{SNR}_{\mathrm{turn}}/\mathrm{SNR}_{\mathrm{traj}}$, aggregated across seeds by IQM with 95% bootstrap confidence intervals, and fit the power law $R(\rho) = C\rho^{\alpha}$ in log-log coordinates.

Experiment 2 — SNR-to-training-step efficiency link. The turn-level estimator uses a Monte Carlo critic with k rollouts per visited state. This does not fully achieve the perfect-critic regime, nor does it make critic quality constant across L, but it gives a controlled critic and avoids learned-critic training dynamics. We report iterations-to-threshold τ (the first iteration whose per-iteration mean reward reaches 0.9) and the paired-seed speedup $\tau_{\mathrm{traj}}/\tau_{\mathrm{turn}}$, aggregated as above. We fit a power law to the speedup to measure its scaling; $\alpha = -1/2$ **is not expected here**, as these are different quantities from the SNR. AUC results, sweep grids, and seed counts are in Appendices B.10 and B.11.

5. Results

5.1. Experiment 1: gradient SNR scaling

The initialization SNR ratio is reported in Figure 1. Fitting $R_{\mathrm{SNR}}(\rho) = C\rho^{\alpha}$ gives $\hat{\alpha} = -0.5084$ with 95% bootstrap confidence interval $[-0.5118, -0.5054]$ and log-space $R^2 = 0.9994$. The fitted exponent is within 0.0084 of the predicted value $-1/2$; constraining $\alpha = -1/2$ gives a nearly identical fit, with $R^2 = 0.9991$ (bootstrap test details in Appendix B.10). Corresponding curves under varying critic noise are reported in Appendix B.9.

Interpretation. Compared to the trajectory-level estimator, the perfect-critic turn-level estimator has 17.41x higher SNR at $\rho = 1/51$, and only 2.30x[4] SNR at $\rho = 1$. These results support a causal effect of decision density on the gradient SNR ratio in this environment.

[4]The ratio is not equal to 1 at $\rho = 1$ because the perfect-critic turn-level estimator has lower variance than the trajectory-level estimator even when no routine states are inserted.

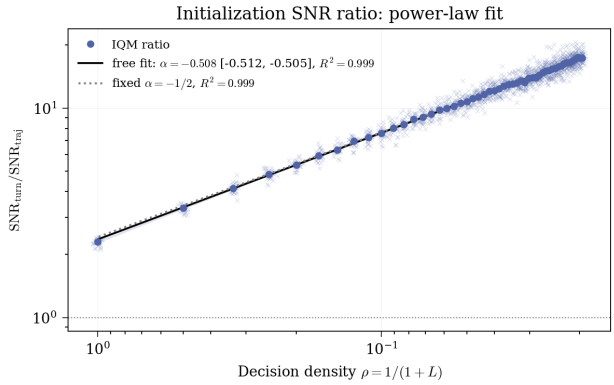


Figure 1. Paired initialization SNR ratio $\mathrm{SNR}_{\mathrm{turn}}/\mathrm{SNR}_{\mathrm{traj}}$ vs. decision density, log scale. Points are per-L IQMs.

5.2. Experiment 2: training-step efficiency

The threshold-speedup ratio as a function of ρ is reported in Figure 2. At $\rho = 1$, the two estimators reach the performance threshold in approximately the same number of training steps (paired-seed IQM 1.016, interval $[0.961, 1.081]$). At $\rho = 1/51$, the trajectory-level estimator takes roughly twice as many training steps as the turn-level estimator (paired-seed IQM 2.139, interval $[1.772, 2.525]$). The fit gives $\hat{\alpha} = -0.19$ ($R^2 = 0.948$), well separated from zero and confirming that training-step efficiency depends on ρ. AUC and individual iters curves are in App. B.8, B.10, B.7.

Interpretation. Reducing decision density while keeping the critical problem fixed widens the training-step efficiency gap between estimators, as predicted. These results support a causal effect of decision density on training-step efficiency in this environment.

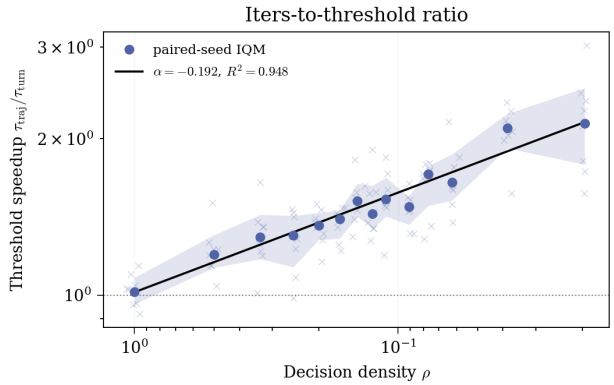


Figure 2. Threshold speedup $\tau_{\mathrm{traj}}/\tau_{\mathrm{turn}}$ vs. decision density, log scale.

6. Related Work

Trajectory-level optimization for LLMs. Standard LLM post-training methods often assign supervision at the level

of a complete response: RLHF for instruction following (Ouyang et al., 2022), DPO over preferred/dispreferred completions (Rafailov et al., 2023), and GRPO with group-normalized trajectory rewards (Shao et al., 2024). GRPO, introduced in DeepSeekMath, is especially relevant because it avoids training a value function by using group-level Monte Carlo baselines. Our work studies the statistical cost of trajectory-level credit assignment in multi-turn settings as a function of decision density.

Turn-level credit assignment for LLM agents. Recent work shows that multi-turn agent RL benefits from credit assignment below the trajectory level. GiGPO (Feng et al., 2025) adds critic-free step-level relative advantages via anchor-state grouping, while turn-level reward design (Wei et al., 2025) introduces dense turn-level rewards for multi-turn search agents. Turn-PPO (Li et al., 2026) further diagnoses direct multi-turn GRPO as unstable, attributing failures to uniform advantages across heterogeneous turns and high-variance sample-based estimates. These studies provide empirical evidence and algorithmic remedies for the limitations of trajectory-level credit assignment. Our contribution is complementary: we isolate *signal dilution* as a statistical mechanism behind this gap and show how its SNR cost scales with decision density ρ.

Gradient signal-to-noise ratio. Roberts and Tedrake (Roberts and Tedrake, 2008) argue that policy-gradient variance should be measured relative to the useful gradient signal, not in isolation. They show empirically that policy-gradient SNR predicts learning performance under fixed evaluation budgets. We use the same lens to study a different source of noise: routine turns that add trajectory-level gradient variance without adding expected signal. Our parameter-wise SNR is the square root of the parameter-wise GSNR used in supervised learning (Liu et al., 2020).

7. Discussion

We trace the turn-level versus trajectory-level gap to a single mechanism. Routine turns (Definition 3.1) contribute no expected signal: the perfect-critic turn-level estimator removes their contribution pointwise, while the trajectory-level estimator cancels it only in expectation (Proposition 3.1). Under Assumptions 3.1–3.2, the residual variance accumulates linearly in the routine-turn count, yielding the $\Theta(\rho^{-1/2})$ SNR advantage of Proposition 3.2. Proposition 3.3 shows that gradient-unbiased critic error reintroduces routine-turn variance into the turn-level estimator, so the ρ-scaling depends on the relative magnitudes of decision density and critic-error variance. Diluted Doors recovers the predicted scaling under an intervention on ρ, and training efficiency gaps widen as ρ decreases.

The $\rho^{-1/2}$ law should be read as an idealized upper envelope: real environments weaken the turn-level advantage through approximately routine states, near-deterministic policies on routine actions, and learned-critic error. The analysis nuances prior findings on turn-level credit assignment (Feng et al., 2025; Li et al., 2026; Wei et al., 2025): the gap should track decision density rather than horizon alone, so at high ρ, even on long horizons, the cost of a critic is unlikely to be justified. Decision density thus serves both as a diagnostic for when a critic is worth its cost and as a target for algorithm design that addresses signal dilution without paying the full critic price.

Three limitations bound these claims. First, no experiments are run on real LLM agents; the upper-envelope reading should carry over since the mechanism is environment-independent, but the realized gap and the magnitude of ρ in such systems have not been measured. Second, routine states are policy-dependent and not directly observable. Decision density can still be used conceptually, and a quantitative route is to probe a base agent through rollouts and estimate which action alternatives shift the downstream return distribution. Third, extending the definition to approximately routine states, e.g. by thresholding a distance between conditional return distributions, is the most important theoretical gap left open.

Several directions follow. The most direct is to estimate decision density, SNR ratios, and training gaps on agentic benchmarks with learned critics. A second is a soft-critical-state theory grounded in approximate distributional invariance, connected to the imperfect-critic regime. The mechanism also suggests critic-free variants that suppress routine-turn noise once routine states can be detected: lowering sampling temperature at these turns, or masking their score contributions.

8. Conclusion

Our findings support that multi-turn credit assignment is governed not by horizon alone, but by *decision density*: the fraction of turns whose actions affect the return distribution. When ρ is low, trajectory-level estimators attach a scalar advantage to many routine turns, adding variance without signal; turn-level estimators avoid this by cancelling routine contributions pointwise rather than only in expectation. This signal dilution mechanism yields an upper-envelope $\Theta(\rho^{-1/2})$ SNR advantage for turn-level credit assignment. The practical message is diagnostic: at high ρ, GRPO-like methods can avoid critic cost while remaining competitive; turn-level credit assignment becomes most useful when consequential decisions are sparse and critic error stays below the decision-density scale. A natural next step is to measure ρ in real agentic tasks and design cheaper ways to suppress routine-turn noise.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Agarwal, R., Schwarzer, M., Castro, P. S., Courville, A. C., and Bellemare, M. G. Deep reinforcement learning at the edge of the statistical precipice. *Advances in Neural Information Processing Systems*, 34, 2021.

Ahmadian, A., Cremer, C., Gallé, M., Fadaee, M., Kreutzer, J., Pietquin, O., Üstün, A., and Hooker, S. Back to basics: Revisiting REINFORCE-style optimization for learning from human feedback in LLMs. *ACL*, 2024.

Cui, G., Zhang, Y., Chen, J., et al. The entropy mechanism of reinforcement learning for reasoning language models. *arXiv preprint arXiv:2505.22617*, 2025.

Efron, B. and Tibshirani, R. J. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

Feng, L., et al. Group-in-Group Policy Optimization for LLM Agent Training. *arXiv preprint arXiv:2505.10978*, 2025.

Greensmith, E., Bartlett, P. L., and Baxter, J. Variance reduction techniques for gradient estimates in reinforcement learning. *Journal of Machine Learning Research*, 5:1471–1530, 2004.

Le, T.-L. V., Jeon, M., Vu, K., Lai, V., and Yang, E. No prompt left behind: Exploiting zero-variance prompts in LLM reinforcement learning via entropy-guided advantage shaping. *ICLR*, 2026.

Li, J., Zhou, P., Meng, R., Vadera, M. P., Li, L., and Li, Y. Turn-PPO: Turn-level advantage estimation with PPO for improved multi-turn RL in agentic LLMs. *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 6227–6243, 2026.

Liu, J., Bai, Y., Jiang, G., Chen, T., and Wang, H. Understanding why neural networks generalize well through GSNR of parameters. *ICLR*, 2020.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. *ICLR*, 2019.

Minsky, M. Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1):8–30, 1961.

Ouyang, L., Wu, J., Jiang, X., et al. Training language models to follow instructions with human feedback. *NeurIPS*, 2022.

Pascanu, R., Mikolov, T., and Bengio, Y. On the difficulty of training recurrent neural networks. *ICML*, 2013.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *NeurIPS*, 2023.

Roberts, J. and Tedrake, R. Signal-to-noise ratio analysis of policy gradient algorithms. *Advances in Neural Information Processing Systems*, 21, 2008.

Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Seo, M., Vecchietti, L. F., Lee, S., and Har, D. Rewards prediction-based credit assignment for reinforcement learning with sparse binary rewards. *IEEE Access*, 7:118776–118791, 2019.

Shao, Z., Wang, P., Zhu, Q., et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, 12:1057–1063, 2000.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, second edition, 2018.

Vaswani, A., Shazeer, N., Parmar, N., et al. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Wei, Q., Zeng, S., Li, C., et al. Reinforcing Multi-Turn Reasoning in LLM Agents via Turn-Level Reward Design. *arXiv preprint arXiv:2505.11821*, 2025.

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

Yang, J., Jimenez, C. E., Wettig, A., Lieret, K., Yao, S., Narasimhan, K., and Press, O. SWE-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems*, 37, 2024.

Zhou, S., Xu, F. F., Zhu, H., et al. WebArena: A realistic web environment for building autonomous agents. *ICLR*, 2024.

A. Structural existence of a gap

Proof of Lemma 3.2. Fix the coordinate k and suppress it. Decompose the trajectory-level estimator into critical and routine contributions:

$$\hat{g}_{\text{traj}} = \hat{A}\,\Phi_C + \hat{A}\,\Phi_R.$$

By Proposition 3.1, the routine part of the trajectory-level estimator has zero mean:

$$\mathbb{E}[\hat{A}\,\Phi_R] = 0.$$

By A2, $\hat{A}$ is $\mathcal{H}_C$-measurable, so $\hat{A}^2$ is $\mathcal{H}_C$-measurable. Using the tower property and the two-sided A2(c),

$$\begin{aligned}\mathrm{Var}[\hat{A}\,\Phi_R] &= \mathbb{E}[\hat{A}^2\Phi_R^2] \\ &= \mathbb{E}\Big[\hat{A}^2\,\mathbb{E}[\Phi_R^2 \mid \mathcal{H}_C]\Big],\end{aligned}$$

$$\begin{aligned}c_R\,\mathbb{E}[\hat{A}^2 N_R] &\le \mathrm{Var}[\hat{A}\,\Phi_R] \\ &\le C_R\,\mathbb{E}[\hat{A}^2 N_R].\end{aligned}$$

For the lower bound, A2(b) and A2(a) give $\mathbb{E}[\hat{A}^2 N_R] \ge c_A\,\mathbb{E}[N_R] \ge c_A c_N\,T$. For the upper bound, $|\hat{A}| \le M_A$ by the bounded-advantage assumption and $N_R \le T$ give $\mathbb{E}[\hat{A}^2 N_R] \le M_A^2\,T$. Hence

$$\begin{aligned}& c_R c_A c_N\,T \le \mathrm{Var}[\hat{A}\,\Phi_R] \le C_R M_A^2\,T, \\ \text{i.e.,}\quad & \mathrm{Var}[\hat{A}\,\Phi_R] = \Theta(T).\end{aligned}$$

By Prop. 3.1, the expected turn-level and trajectory-level gradients reduce to their critical-state contributions, which A1 bounds away from 0 and ∞:

$$m_\mu \le |\mathbb{E}[\hat{g}_{\text{turn}}]| \le M_\mu, \qquad m_\mu \le |\mathbb{E}[\hat{g}_{\text{traj}}]| \le M_\mu.$$

It remains to compare variances. By A1,

$$m_C \;\le\; \mathrm{Var}[\hat{g}_{\text{turn}}] \le M_C, \qquad \mathrm{Var}[\hat{A}\,\Phi_C] \le M_C,$$

while $\mathrm{Var}[\hat{A}\,\Phi_R] = \Theta(T)$ from the previous display. Writing the standard deviation of the sum as an L^2 norm of the centered variables,

$$\begin{aligned}&\sqrt{\mathrm{Var}[\hat{A}\,\Phi_C + \hat{A}\,\Phi_R]} \\ &\quad = \Big\|(\hat{A}\,\Phi_C - \mathbb{E}[\hat{A}\,\Phi_C]) + (\hat{A}\,\Phi_R - \mathbb{E}[\hat{A}\,\Phi_R])\Big\|_2,\end{aligned}$$

For sufficiently large T, the reverse and forward triangle inequalities, together with $c_R c_A c_N\,T \le \mathrm{Var}[\hat{A}\,\Phi_R] \le C_R M_A^2\,T$ and $\mathrm{Var}[\hat{A}\,\Phi_C] \le M_C$, bound it on both sides:

$$\begin{aligned}\sqrt{c_R c_A c_N\,T} - \sqrt{M_C} &\le \sqrt{\mathrm{Var}[\hat{A}\,\Phi_C + \hat{A}\,\Phi_R]} \\ &\le \sqrt{C_R M_A^2\,T} + \sqrt{M_C} \\ &= \Theta(\sqrt{T}).\end{aligned}$$

Hence $\mathrm{Var}[\hat{g}_{\text{traj}}] = \Theta(T)$. The one-coordinate bounds are therefore

$$\begin{aligned}|\mathbb{E}[\hat{g}_{\text{turn}}]| &= \Theta(1), \\ |\mathbb{E}[\hat{g}_{\text{traj}}]| &= \Theta(1), \\ \mathrm{Var}[\hat{g}_{\text{turn}}] &= \Theta(1), \\ \mathrm{Var}[\hat{g}_{\text{traj}}] &= \Theta(T).\end{aligned}$$

Inspecting the derivations above, the upper bounds use only the upper-bound parts of A1 and the upper half of A2(c), while the lower bounds additionally invoke A1's lower-bound parts together with A2(a), A2(b), and the lower half of A2(c). □

Proof of Proposition 3.2. Combining the bounds from Lemma 3.2,

$$\begin{aligned}\frac{\mathrm{SNR}_k(\hat{g}_{\text{turn}})}{\mathrm{SNR}_k(\hat{g}_{\text{traj}})} &= \frac{|\mathbb{E}[\hat{g}_{\text{turn}}]|}{|\mathbb{E}[\hat{g}_{\text{traj}}]|} \\ &\quad \times \sqrt{\frac{\mathrm{Var}[\hat{g}_{\text{traj}}]}{\mathrm{Var}[\hat{g}_{\text{turn}}]}} \\ &= \Theta(\sqrt{T}) = \Theta(\rho^{-1/2}).\end{aligned}$$

□

A.1. Pooled SNR scaling

Proof of Corollary 3.1. Let $\mathcal{K} = \{1, \ldots, d\}$ and write

$$\begin{aligned}\mu_{\bullet,k} &= \mathbb{E}[\hat{g}_{\bullet,k}], \\ \sigma^2_{\bullet,k} &= \mathrm{Var}[\hat{g}_{\bullet,k}], \\ \bullet &\in \{\text{turn}, \text{traj}\}.\end{aligned}$$

By Lemma 3.2, the uniform upper conditions give, for every $k \in \mathcal{K}$,

$$\begin{aligned}|\mu_{\bullet,k}| &= O(1), \\ \sigma^2_{\text{turn},k} &= O(1), \\ \sigma^2_{\text{traj},k} &= O(T),\end{aligned}$$

while the lower conditions at $k^\star$ give

$$\begin{aligned}|\mu_{\bullet,k^\star}| &= \Omega(1), \\ \sigma^2_{\text{turn},k^\star} &= \Omega(1), \\ \sigma^2_{\text{traj},k^\star} &= \Omega(T).\end{aligned}$$

Since d is fixed,

$$\begin{aligned}\|\mathbb{E}\hat{g}_\bullet\|^2 &= \sum_{k\in\mathcal{K}} \mu^2_{\bullet,k} = \Theta(1), \\ \sum_{k\in\mathcal{K}} \sigma^2_{\text{turn},k} &= \Theta(1), \\ \sum_{k\in\mathcal{K}} \sigma^2_{\text{traj},k} &= \Theta(T).\end{aligned}$$

Substituting into Definition 2.2,

$$\begin{aligned}\frac{\mathrm{SNR}(\hat{g}_{\mathrm{turn}})}{\mathrm{SNR}(\hat{g}_{\mathrm{traj}})} &= \frac{\|\mathbb{E}\hat{g}_{\mathrm{turn}}\|}{\|\mathbb{E}\hat{g}_{\mathrm{traj}}\|} \\ &\quad \times \sqrt{\frac{\sum_k \sigma^2_{\mathrm{traj},k}}{\sum_k \sigma^2_{\mathrm{turn},k}}} \\ &= \Theta(1) \cdot \sqrt{\frac{\Theta(T)}{\Theta(1)}} \\ &= \Theta(\sqrt{T}) = \Theta(\rho^{-1/2}).\end{aligned}$$

□

A.2. Imperfect critic proof

Proof of Proposition 3.3. Write

$$\hat{g}_{\mathrm{turn},\epsilon,k} = \hat{g}^{\star}_{\mathrm{turn},k} + Z_{\epsilon,k}.$$

By Assumption 3.3,

$$\mathbb{E}[Z_{\epsilon,k} \mid \mathcal{F}_\tau] = 0.$$

Since $\hat{g}^{\star}_{\mathrm{turn},k}$ is determined by the sampled trajectory,

$$\mathbb{E}[\hat{g}_{\mathrm{turn},\epsilon,k} \mid \mathcal{F}_\tau] = \hat{g}^{\star}_{\mathrm{turn},k}.$$

Taking expectation gives

$$\mathbb{E}[\hat{g}_{\mathrm{turn},\epsilon,k}] = \mathbb{E}[\hat{g}^{\star}_{\mathrm{turn},k}].$$

By Lemma 3.2, under Assumptions 3.1–3.2,

$$|\mathbb{E}[\hat{g}^{\star}_{\mathrm{turn},k}]| = \Theta(1),$$

so

$$|\mathbb{E}[\hat{g}_{\mathrm{turn},\epsilon,k}]| = \Theta(1).$$

For the variance, apply the law of total variance:

$$\begin{aligned}\mathrm{Var}[\hat{g}_{\mathrm{turn},\epsilon,k}] &= \mathrm{Var}\left(\mathbb{E}[\hat{g}_{\mathrm{turn},\epsilon,k} \mid \mathcal{F}_\tau]\right) \\ &\quad + \mathbb{E}\left[\mathrm{Var}\left(\hat{g}_{\mathrm{turn},\epsilon,k} \mid \mathcal{F}_\tau\right)\right].\end{aligned}$$

The first term is

$$\mathrm{Var}\left(\mathbb{E}[\hat{g}_{\mathrm{turn},\epsilon,k} \mid \mathcal{F}_\tau]\right) = \mathrm{Var}[\hat{g}^{\star}_{\mathrm{turn},k}].$$

Again by Lemma 3.2,

$$\mathrm{Var}[\hat{g}^{\star}_{\mathrm{turn},k}] = \Theta(1).$$

For the second term, $\hat{g}^{\star}_{\mathrm{turn},k}$ is fixed conditional on $\mathcal{F}_\tau$, so

$$\mathrm{Var}\left(\hat{g}_{\mathrm{turn},\epsilon,k} \mid \mathcal{F}_\tau\right) = \mathrm{Var}\left(Z_{\epsilon,k} \mid \mathcal{F}_\tau\right).$$

By the definition of $\sigma^2_{\epsilon,k}(T)$ in Assumption 3.3,

$$\mathbb{E}\left[\mathrm{Var}\left(Z_{\epsilon,k} \mid \mathcal{F}_\tau\right)\right] = T\sigma^2_{\epsilon,k}(T).$$

Therefore,

$$\mathrm{Var}[\hat{g}_{\mathrm{turn},\epsilon,k}] = \Theta(1) + T\sigma^2_{\epsilon,k}(T) = \Theta\big(1 + T\sigma^2_{\epsilon,k}(T)\big).$$

The existing structural result gives

$$|\mathbb{E}[\hat{g}_{\mathrm{traj},k}]| = \Theta(1), \qquad \mathrm{Var}[\hat{g}_{\mathrm{traj},k}] = \Theta(T).$$

Thus

$$\begin{aligned}\frac{\mathrm{SNR}_k(\hat{g}_{\mathrm{turn},\epsilon})}{\mathrm{SNR}_k(\hat{g}_{\mathrm{traj}})} &= \frac{|\mathbb{E}[\hat{g}_{\mathrm{turn},\epsilon,k}]|}{|\mathbb{E}[\hat{g}_{\mathrm{traj},k}]|}\sqrt{\frac{\mathrm{Var}[\hat{g}_{\mathrm{traj},k}]}{\mathrm{Var}[\hat{g}_{\mathrm{turn},\epsilon,k}]}} \\ &= \Theta(1)\sqrt{\frac{\Theta(T)}{\Theta(1 + T\sigma^2_{\epsilon,k}(T))}} \\ &= \Theta\left(\sqrt{\frac{T}{1 + T\sigma^2_{\epsilon,k}(T)}}\right).\end{aligned}$$

Finally, using $T = \bar{C}/\rho$,

$$\sqrt{\frac{T}{1 + T\sigma^2_{\epsilon,k}(T)}} = \sqrt{\frac{\bar{C}}{\rho + \bar{C}\sigma^2_{\epsilon,k}(T)}} = \Theta\left(\frac{1}{\sqrt{\rho + \sigma^2_{\epsilon,k}(T)}}\right),$$

since $\bar{C}$ is fixed. □

A.3. Asymptotic form

Under the hypotheses of Proposition 3.2,

$$\frac{\mathrm{SNR}_{\mathrm{turn}}}{\mathrm{SNR}_{\mathrm{traj}}} = \Theta(\sqrt{T}) = \Theta(\rho^{-1/2}), \tag{5}$$

in the limit $T \to \infty$ at fixed $\bar{C}$.

B. Diluted Doors environment and training details

This appendix expands the experimental setup of Section 4.

B.1. Environment design

Diluted Doors is a deterministic tree MDP. The agent observes its current state (position in the tree and available actions) and chooses one action per turn, selecting a door at choice states or a path at routine states. The tree structure and reward function are fixed across episodes; the agent must discover the reward structure through repeated interaction.

Reward. The reward is sparse and additive, revealed only at the end of the episode and equal to the fraction of correct choices at critical states:

$$R = \frac{1}{C}\sum_{c=1}^{C} \mathbf{1}[\text{agent chose the correct door at depth } c].$$

The additive form ensures that a wrong choice at depth c does not eliminate subsequent critical states: every trajectory visits exactly C critical states regardless of actions taken.

Tree structure. Every root-to-leaf path has length $T = C(1 + L)$ with C choice states and $C \cdot L$ routine states. Because the tree is balanced and routine subtrees are isomorphic, every trajectory visits exactly C critical states, so $\bar{C} = C$ and $\rho = 1/(1 + L)$ exactly.

B.2. Concrete episode

For $C = 2$, $K_C = 2$, $K_R = 2$, $L = 1$ ($T = 4$, $\rho = 1/2$), with correct doors $\{1, 2\}$ at depths $\{1, 2\}$:

Turn	Type	State shows	Agent picks
1	Critical	`door_1, door_2`	`door_1` (correct)
2	Routine	`path_1, path_2`	`path_2` (irrelevant)
3	Critical	`door_1, door_2`	`door_1` (wrong; correct is `door_2`)
4	Routine	`path_1, path_2`	`path_1` (irrelevant)

Terminal reward: $R = (1 + 0)/2 = 1/2$.

B.3. Tunable parameters

Parameter	
C	Number of critical states
L	Routine states per critical state
K_C	Doors per critical state
K_R	Paths per routine state
$T = C(1 + L)$	Total turns
$\rho = 1/(1 + L)$	Decision density

B.4. Policy class and initialization

The policy π_θ is a small causal Transformer (Vaswani et al., 2017) (30k parameters). Actions are represented as tokens: the vocabulary has $K+1$ entries, one per action plus a `START` symbol. At turn t, the input sequence is $[\texttt{START}, a_0, \ldots, a_{t-1}]$, and a linear head on the final position produces a distribution over the K next actions.

The network uses two pre-norm Transformer blocks with model dimension 32, two-head self-attention, and a feed-forward sub-block of hidden width 128. Token and positional embeddings are learned.

Weight initialization is Xavier-normal with gain 0.1 on linear layers and $\mathcal{N}(0, 0.02^2)$ on embeddings. At initialization, the policy is close to uniform across all states: the normalized action entropy $H(\pi(\cdot|s))/\log K$ is close to 1 and the mean per-step KL divergence to the uniform distribution is close to 0.

B.5. On-policy training loop

One rollout group of G trajectories drives one gradient step. We use the simplified gradients of Section 2.2 directly: no importance ratio, no clipping, and no KL penalty. The two estimators differ only in the advantage plugged into Eq. (1) or Eq. (2).

Trajectory-level estimator. We use the group-normalized advantage of Eq. (3), broadcast to every turn of trajectory i. A single gradient step is taken per rollout group; multiple epochs would require an importance-ratio correction whose interaction with the clip is a separate question from the gradient-form one we isolate.

Turn-level estimator with Monte Carlo V^π critic. For each visited state $s_t^{(i)}$, we draw k fresh rollouts under the current policy and set $\hat{V}(s_t^{(i)})$ to the mean of their returns. Under deterministic transitions and terminal reward, the TD(0) advantage

$$\begin{aligned} \hat{A}_t &= \hat{V}(s_{t+1}) - \hat{V}(s_t) \quad (t < T - 1), \\ \hat{A}_{T-1} &= R - \hat{V}(s_{T-1}) \end{aligned} \tag{6}$$

is consistent for A^{π_θ}, so at routine states $\hat{A}_t \to 0$ as $k \to \infty$, matching Proposition 3.1(ii). Since $\hat{V}(s_t)$ and $\hat{V}(s_{t+1})$ are computed from disjoint rollout batches, they are independent and the advantage variance scales as

$$\mathrm{Var}[\hat{A}_t] \;=\; \frac{\mathrm{Var}_\tau[R(\tau) \mid s_{t+1}, \pi_\theta] + \mathrm{Var}_\tau[R(\tau) \mid s_t, \pi_\theta]}{k}.$$

We use Monte Carlo rather than a learned critic so that critic quality is a known, controllable quantity (k); a learned critic would conflate the structural gap with critic-training dynamics.

Optimizer. AdamW (Loshchilov and Hutter, 2019) with learning rate 10^{-3}, linear warmup over the first 10% of iterations, gradient clipping at 1.0 (Pascanu et al., 2013), no entropy regularization. Each run comprises 500 iterations.

B.6. Evaluation statistics

For each L, scalar metrics are aggregated across seeds using the interquartile mean (IQM) (Agarwal et al., 2021). Confidence intervals are percentile bootstrap intervals with 5000 resamples (Efron and Tibshirani, 1993). For paired ratios, the ratio is computed seed-by-seed before aggregating across seeds.

Given per-L aggregate ratios $R(\rho)$, we fit

$$R(\rho) = C\rho^\alpha$$

by ordinary least squares. Reported R^2 values are computed in log-space. Confidence intervals for α are obtained by

bootstrapping seeds within each L, recomputing the per-L aggregate ratio, and refitting the power law.

For the association analysis, the common L grid between the initialization and training sweeps is used. We compute Pearson correlation between log initialization SNR ratios and log training ratios. Confidence intervals are bootstrap intervals over the common L values; permutation p-values are two-sided.

B.7. Supplementary association figures

Figure 3 reports the association between the initialization SNR ratio and the threshold-speedup ratio. Figure 4 reports the association between the initialization SNR ratio and the AUC ratio.

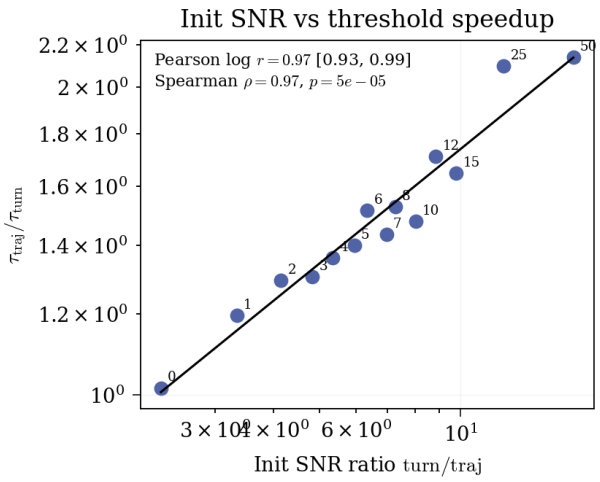


Figure 3. Association between the initialization SNR ratio and threshold-speedup ratio on the shared L values. Each point is one value of L.

B.8. Per-estimator iterations-to-threshold

Figure 5 reports iterations-to-threshold for each estimator separately, complementing the speedup ratio in the main text. At $\rho = 1$, both estimators reach the threshold in roughly 76–78 iterations; at $\rho = 1/51$, the trajectory-level estimator requires 225.25 iterations ($[186.0, 254.0]$) while the turn-level estimator requires 102.0 ($[94.75, 118.0]$).

B.9. Critic-noise frontier

The initialization SNR ratio under varying levels of critic noise empirically traces the imperfect-critic frontier of Proposition 3.3. Figure 6 shows the SNR ratio as a function of L for several critic-noise levels; Figure 7 shows the same data as a function of critic noise for several values of L.

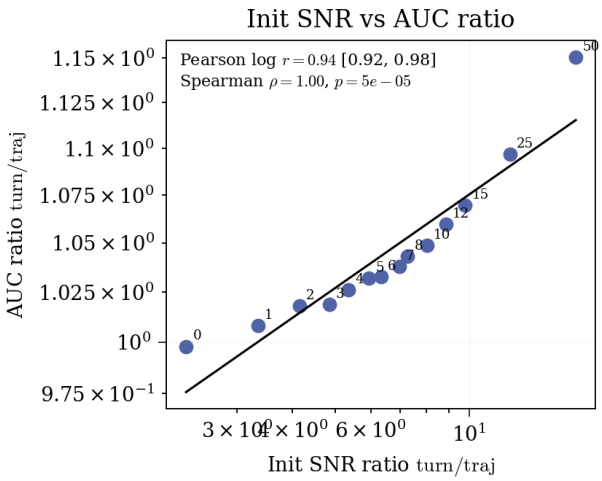


Figure 4. Association between the initialization SNR ratio and AUC ratio on the shared L values. Each point is one value of L.

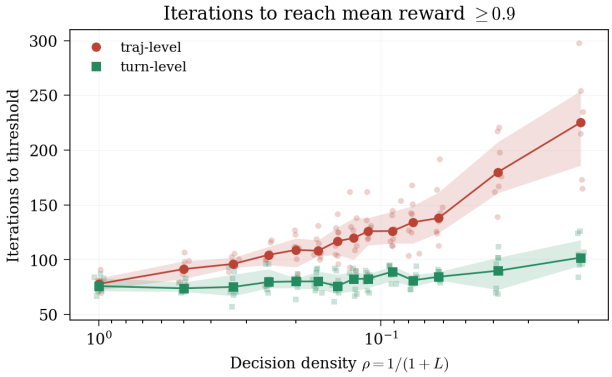


Figure 5. Iterations to reach mean reward at least 0.9 vs. decision density on a logarithmic ρ axis.

B.10. Supplementary AUC figures

AUC is retained as a whole-curve training summary, but the main analysis uses iterations-to-threshold. Figure 8 reports AUC by decision density, Figure 9 reports the paired AUC ratio, and Figure 4 reports its association with the initialization SNR ratio.

Bootstrap test for $\alpha = -1/2$ (Experiment 1). A bootstrap test of exact equality to $\alpha = -1/2$ rejects at $p = 0.0004$; the estimated deviation is small ($|\hat{\alpha} + 1/2| = 0.0084$), and the full confidence interval remains inside the practical equivalence band $[-0.52, -0.48]$.

B.11. Sweep grids

All experiments share $C = K_C = K_R = 4$, so random-policy reward is $1/K_C = 0.25$ and the routine action space is non-trivial ($K_R > 1$); whether the policy actually places mass on more than one routine action, and hence whether

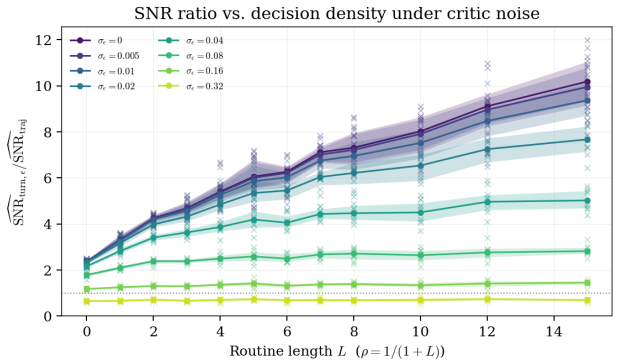


Figure 6. Initialization SNR ratio $\mathrm{SNR}_{\mathrm{turn},\epsilon}/\mathrm{SNR}_{\mathrm{traj}}$ vs. L at several critic-noise levels. Horizontal asymptotes form as we raise the critic error scale.

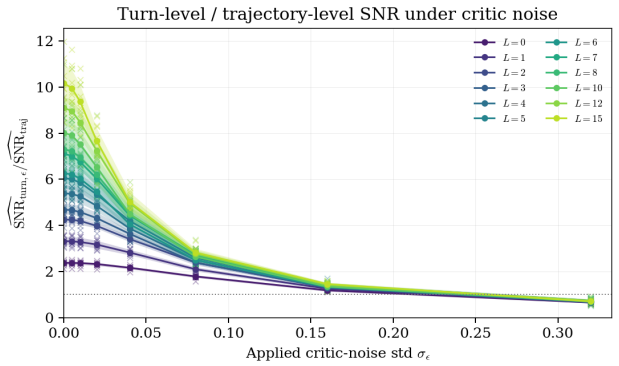


Figure 7. Initialization SNR ratio $\mathrm{SNR}_{\mathrm{turn},\epsilon}/\mathrm{SNR}_{\mathrm{traj}}$ vs. critic noise at several values of L.

routine score variance is nonzero, as required by Assumption 3.2(c), depends on π_θ and is not enforced by construction after initialization. Each experimental seed sets both the training run (weight initialization and on-policy rollout sampling) and the MDP instance (which of the K_C doors is correct at each critical depth), so cross-seed confidence intervals reflect robustness to both simultaneously.

Training-performance sweep	
L grid	$\{0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 12, 15, 25, 50\}$
Seeds / cell	8
Hyperparameters	Traj-level $G = 16$, Turn-level $k = 8$
SNR, init-probe	
L grid	$\{0, 1, \ldots, 50\}$
Seeds / cell	32
Hyperparameters	$G = 16$, $N = 1024$, analytical V

Table 1. Sweep grids. Decision density is $\rho = 1/(1 + L)$, so the $L \leq 50$ range covers $\rho \in [\frac{1}{51}, 1]$.

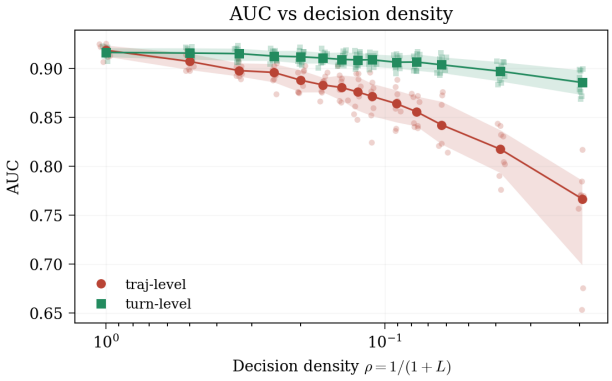


Figure 8. AUC vs. decision density on a logarithmic ρ axis.

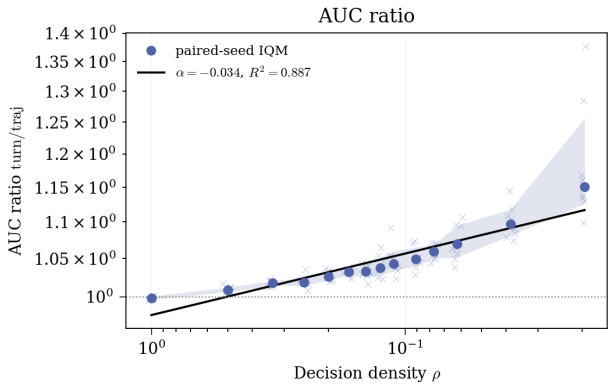


Figure 9. AUC ratio $\mathrm{AUC}_{\mathrm{turn}}/\mathrm{AUC}_{\mathrm{traj}}$ vs. decision density.